



HAL
open science

Going beyond the sentence: Contextual Machine Translation of Dialogue

Rachel Bawden

► **To cite this version:**

Rachel Bawden. Going beyond the sentence: Contextual Machine Translation of Dialogue. Computation and Language [cs.CL]. Université Paris Saclay (COMUE), 2018. English. NNT : 2018SACLS524 . tel-02004683

HAL Id: tel-02004683

<https://theses.hal.science/tel-02004683v1>

Submitted on 2 Feb 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Going beyond the sentence: Contextual Machine Translation of Dialogue

Thèse de doctorat de l'Université Paris-Saclay
préparée à l'Université Paris-Sud
au sein du Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur

Ecole doctorale n°580 Sciences et Technologies de l'Information et de la
Communication (STIC)
Spécialité de doctorat : Informatique

Thèse présentée et soutenue à Orsay, le 29 novembre 2018, par

RACHEL BAWDEN

Composition du Jury :

Nicolas Sabouret Professeur, LIMSI, CNRS, Univ. Paris-Sud, Université Paris-Saclay	Président
Jörg Tiedemann Professeur, Université d'Helsinki	Rapporteur
Loïc Barrault Maître de conférences, Université du Mans	Rapporteur
Lucia Specia Professeur, Université de Sheffield et Imperial College London	Examinatrice
Andrei Popescu-Belis Professeur, Haute École d'Ingénierie et de Gestion du Canton de Vaud	Examineur
Sophie Rosset Directrice de recherches, LIMSI, CNRS	Directrice de thèse
Thomas Lavergne Maître de conférences, LIMSI, CNRS, Univ. Paris-Sud, Université Paris-Saclay	Co-Directeur de thèse

Summary

While huge progress has been made in machine translation (MT) in recent years, the majority of MT systems still rely on the assumption that sentences can be translated in isolation. The result is that these MT models only have access to context within the current sentence; context from other sentences in the same text and information relevant to the scenario in which they are produced remain out of reach.

The aim of *contextual MT* is to overcome this limitation by providing ways of integrating extra-sentential context into the translation process. Context, concerning the other sentences in the text (*linguistic context*) and the scenario in which the text is produced (*extra-linguistic context*), is important for a variety of cases, such as discourse-level and other referential phenomena.

Successfully taking context into account in translation is challenging. Evaluating such strategies on their capacity to exploit context is also a challenge, standard evaluation metrics being inadequate and even misleading when it comes to assessing such improvement in contextual MT.

In this thesis, we propose a range of strategies to integrate both extra-linguistic and linguistic context into the translation process. We accompany our experiments with specifically designed evaluation methods, including new test sets and corpora. Our contextual strategies include *pre-processing* strategies designed to disambiguate the data on which MT models are trained, *post-processing* strategies to integrate context by post-editing MT outputs and strategies in which context is exploited *during translation* proper. We cover a range of different context-dependent phenomena, including anaphoric pronoun translation, lexical disambiguation, lexical cohesion and adaptation to properties of the scenario such as speaker gender and age. Our experiments for both phrase-based statistical MT and neural MT are applied in particular to the translation of English to French and focus specifically on the translation of informal written dialogues.

Résumé en français

Les systèmes de traduction automatique (TA) ont fait des progrès considérables ces dernières années. La majorité d'entre eux reposent pourtant sur l'hypothèse que les phrases peuvent être traduites indépendamment les unes des autres. Ces modèles de traduction ne s'appuient que sur les informations contenues dans la phrase à traduire. Ils n'ont accès ni aux informations présentes dans les phrases environnantes ni aux informations que pourrait fournir le contexte dans lequel ces phrases ont été produites.

La *TA contextuelle* a pour objectif de dépasser cette limitation en explorant différentes méthodes d'intégration du contexte extra-phrastique dans le processus de traduction. Les phrases environnantes (*contexte linguistique*) et le contexte de production des énoncés (*contexte extra-linguistique*) peuvent fournir des informations cruciales pour la traduction, notamment pour la prise en compte des phénomènes discursifs et des mécanismes référentiels.

La prise en compte du contexte est toutefois un défi pour la traduction automatique. Évaluer la capacité de telles stratégies à prendre réellement en compte le contexte et à améliorer ainsi la qualité de la traduction est également un problème délicat, les métriques d'évaluation usuelles étant pour cela inadaptées voire trompeuses.

Dans cette thèse, nous proposons plusieurs stratégies pour intégrer le contexte, tant linguistique qu'extra-linguistique, dans le processus de traduction. Nos expériences s'appuient sur des méthodes d'évaluation et des jeux de données que nous avons développés spécifiquement à cette fin. Nous explorons différents types de stratégies: les stratégies par *pré-traitement*, où l'on utilise le contexte pour désambiguïser les données fournies en entrée aux modèles; les stratégies par *post-traitement*, où l'on utilise le contexte pour modifier la sortie d'un modèle non-contextuel, et les stratégies où l'on exploite le contexte *pendant la traduction* proprement dite. Nous nous penchons sur de multiples phénomènes contextuels, et notamment sur la traduction des pronoms anaphoriques, la désambiguïstation lexicale, la cohésion lexicale et l'adaptation à des informations extra-linguistiques telles que l'âge ou le genre du locuteur. Nos expériences, qui relèvent pour certaines de la TA statistique et pour d'autres de la TA neuronale, concernent principalement la traduction de l'anglais vers le français, avec un intérêt particulier pour la traduction de dialogues spontanés.

Contents

1	Introduction and overview	1
1.1	Motivation for Contextual Machine Translation	1
1.2	Structure and detailed summary of this thesis	3
1.3	Publications related to this thesis	8
I	State of the Art: Contextual Machine Translation	11
2	The Role of Context	15
2.1	Ambiguity and the problem of translation	16
2.1.1	Source language ambiguity	16
2.1.2	Cross-lingual meaning transfer ambiguity	18
2.1.3	Target language ambiguity	19
2.1.4	Human versus machine translation	20
2.2	The importance of context in MT	20
2.2.1	What is context?	21
2.2.2	Nature and use of context	22
2.3	Conclusion	33
3	Sentence-level Machine Translation	35
3.1	Statistical Machine Translation (SMT)	37
3.1.1	Word alignments	39
3.1.2	Phrase-based translation models	41
3.1.3	Domain adaptation	46
3.1.4	Successes and Limitations of SMT	47

3.2	Neural Machine Translation (NMT)	49
3.2.1	Neural networks for NLP	49
3.2.2	Sequence-to-sequence NMT	56
3.2.3	Sequence-to-sequence NMT with attention	59
3.2.4	Recent advances in NMT	60
3.2.5	Successes and limitations	63
3.3	Evaluating Machine Translation	65
3.3.1	Issues in human evaluation of MT quality	66
3.3.2	Standard automatic evaluation metrics	67
3.3.3	Discussion	72
4	Contextual Machine Translation	73
4.1	Evaluating contextual MT	74
4.1.1	Problems associated with automatic evaluation of context	75
4.1.2	MT metrics augmented with discourse information	77
4.1.3	Conclusion	78
4.2	Modelling context for MT	79
4.2.1	Modelling context for SMT	80
4.2.2	Modelling context for NMT	81
4.3	Translation using structured linguistic context	82
4.3.1	Anaphoric pronouns	82
4.3.2	Lexical choice	87
4.3.3	Discourse connectives	91
4.3.4	Whole document decoding	92
4.4	Translation using unstructured linguistic context	93
4.5	Translation using extra-linguistic context	95
4.6	Conclusion on evaluating contextual MT	97

II Using contextual information for Machine Translation: strategies and evaluation **99**

5	Adapting translation to extra-linguistic context via pre-processing	103
5.1	Integrating speaker gender via domain adaptation	104
5.1.1	Annotating the <i>The Big Bang Theory</i> reproducible corpus	105
5.1.2	SMT models: baselines and adaptations	106
5.1.3	Manual analysis and discussion	111
5.1.4	Conclusion on data partitioning	113
5.2	Conclusion	114

6	Improving cohesion-based translation using post-processing	117
6.1	Preserving style in MT: generating English tag questions	118
6.1.1	Tag questions (TQs) and the difficulty for MT	119
6.1.2	Improving TQ generation in MT into English: our post-edition approach	123
6.1.3	Results, analysis and discussion	129
6.1.4	Conclusion to our tag-question experiments	137
6.2	Anaphoric pronoun translation with linguistically motivated features . .	138
6.2.1	Classification system: description and motivation	139
6.2.2	Results, analysis and discussion	147
6.2.3	Conclusion to pronoun translation via post-edition	151
6.3	General conclusion on post-edition approaches	152
7	Context-aware translation models	153
7.1	Translating discourse phenomena with unstructured linguistic context .	154
7.1.1	Hand-crafted test sets for contextual MT evaluation	155
7.1.2	Modifying the NMT architecture	160
7.1.3	Evaluation results and analysis	166
7.1.4	Conclusion and perspectives	172
7.2	Contextual NMT with extra-linguistic context	173
7.2.1	Creation of extra-linguistically annotated data	174
7.2.2	Contextual strategies	182
7.2.3	Experiments	185
7.2.4	BLEU score results	188
7.2.5	Targeted evaluation of speaker gender	192
7.2.6	Conclusion and perspectives	197
7.3	Conclusion	199
8	DIABLA: A corpus for the evaluation of contextual MT	201
8.1	Dialogue and human judgment collection protocol	202
8.1.1	Participants	202
8.1.2	Scenarios	204
8.1.3	Evaluation	205
8.1.4	MT systems and setup	207
8.2	Description of the corpus	209
8.2.1	Overview of translation successes and failures	210
8.2.2	Comparison with existing corpora	213
8.3	Evaluating contextual MT with the DIABLA corpus	215
8.3.1	Overall MT quality	215
8.3.2	Focus on a discourse-level phenomenon	218

8.4	Perspectives	219
8.4.1	Language analysis of MT-assisted interaction	219
8.4.2	MT evaluation	220
Conclusion and Perspectives		224
9	Conclusion and Perspectives	225
9.1	Conclusion	225
9.1.1	Trends in contextual MT and the impact on our work	225
9.1.2	Review of our aims and contributions	227
9.2	Perspectives	230
9.2.1	Evaluation of MT	230
9.2.2	Interpretability of contextual NMT strategies	231
9.2.3	Contextual MT for low resource language pairs	233
9.2.4	Contextual MT to Multimodal MT	235
9.2.5	Conclusion: To the future and beyond the sentence	236
Appendices		236
A	Context-aware translation models	237
A.1	Translating discourse phenomena with unstructured linguistic context	237
A.1.1	Training and decoding parameters	237
A.1.2	Visualisation of hierarchical attention weights	239
A.2	Contextual NMT with extra-linguistic context	240
A.2.1	Experimental setup	240
B	DIABLA: A corpus for the evaluation of contextual MT	241
B.1	Role-play scenarios	242
B.2	Dialogue collection: Final evaluation form	243
Bibliography		270
Résumé détaillé		271

1.1 Motivation for Contextual Machine Translation

The use of Machine Translation (MT) to translate everyday written exchanges is becoming increasingly commonplace; translation tools now regularly appear on chat applications and social networking sites to enable cross-lingual communication. Adapted MT systems must therefore be able to handle the type of language produced in these written yet informal contexts and deal with a wide variety of topics, styles and vocabularies. Importantly, the translation of conversation requires translating sentences coherently with respect to conversational flow in order for all aspects of the exchange, including speaker intent, attitude and style, to be correctly communicated.

Despite huge progress being made in MT year on year,¹ a number of phenomena remain difficult or impossible to translate using standard MT systems. One of the most striking approximations made by the majority of MT systems even today is the fact that sentences are translated in isolation from the other sentences within the same document. There are two main reasons for this: (i) translating long sequences of text presents computational difficulties (data-driven approaches to MT benefit from segmenting texts into smaller units for translation), and (ii) a majority of sentences do not need extra-sentential context to be translated correctly. Up until recently, there have often been easier gains in MT quality to be made by focusing on better modelling of local context within each sentence.

¹This can be seen with the ever increasing quality scores at the annual shared tasks, in particular for highly resourced language pairs (Bojar et al., 2016a, 2017).

However, as the quality of MT systems improves over time, particularly in high-resource settings, the insufficiencies of these sentence-level MT systems becomes increasingly apparent. A number of different phenomena require extra-sentential context to be correctly translated, and they remain impossible to translate using sentence-level MT systems (Hardmeier, 2012). Examples of such cases include *anaphoric pronoun translation* (Guillou, 2016), whereby the translated form of a pronoun is directly dependent on the grammatical gender of another element in the translated text (see Example (1)); *lexical disambiguation* (Carpuat and Wu, 2005; Vickrey et al., 2005), whereby the translation of an element is ambiguous and requires context to be disambiguated (see Example (2)); and *lexical cohesion* (Xiong et al., 2013; Guillou, 2013), whereby a word in the translation is dependent on the form of another translated element, for example when it must be identical to another form, as in Example (3). In each of the three examples illustrating these three cases, the correct translation of a context-dependent element (in bold) is dependent on *linguistic context* that appears outside of the current sentence (underlined), and therefore is inaccessible to standard sentence-level MT systems.

(1) *Anaphoric pronoun translation* (translation of *it*):

EN: She sat on the chair. But **it** broke.

FR: Elle s'est assise sur la chaise. Mais **elle** s'est cassée.

(2) *Lexical disambiguation* (translation of *legs*):

EN: The chair was in pieces. One of its **legs** had collapsed.

FR: La chaise était en morceaux. L'un de ses **pieds** avait lâché.

FR: #La chaise était en morceaux. L'une de ses **jambes** avait lâché.²

(3) *Lexical cohesion (repetition)* (translation of *tired*):

EN: Goldilocks was tired. Very **tired**.

FR: Boucle d'or était fatiguée. Très **fatiguée**.

FR: #Boucle d'or était fatiguée. Très **épuisée**.

The context that determines how the sentence should be translated is not restricted to linguistic context (the words of the text). It can also relate to the situation in which the text is produced (e.g. information concerning the speakers, their relationship, the topic of discussion, etc.), and therefore might not even appear in the text at all. An example of the use of *extra-linguistic* context such as this is speaker gender. In certain languages (e.g. French), certain wordforms (e.g. adjectives and past participles) agree in gender with the subject they qualify. When translating into these languages from a language for which this is not the case (e.g. English), in cases where this subject is the speaker, the gender of the speaker determines the correct translation. This is illustrated in Example (4), in which the translation of English *glad* is dependent on the gender of the speaker (*contente* for a female speaker and *content* for a male speaker).

²In this thesis, we use the character # to indicate incorrect translations from a discursive point of view.

(4) *Speaker gender:*

EN: I am so glad that I can lie down on this comfy bed.

FR_{FEM}: Je suis bien **contente**_{FEM} de pouvoir m'allonger dans ce lit douillet.

FR_{MASC}: Je suis bien **content**_{MASC} de pouvoir m'allonger dans ce lit douillet.

In this thesis, we aim to overcome this approximation made by sentence-level MT systems by reviewing and proposing different strategies to integrate information beyond the level of the sentence into the translation process. We refer to these strategies as *contextual MT*, and by *context* refer to any information outside of the segment of text being translated that could be useful for determining its correct translation. We will look at the integration of both *linguistic context* and *extra-linguistic context* using a variety of strategies. We will pay particular attention throughout this thesis to the evaluation of MT models with respect to the degree in which they succeed in using context.

1.2 Structure and detailed summary of this thesis

This thesis is structured in two main parts. The first part is dedicated to a discussion of our main research questions: what is contextual MT, why is it important and how have previous works sought to tackle it. We introduce the notions on which this thesis will be based, formalise the problem from a theoretical point of view and lay the grounds for our own contributions to contextual MT. In the second part, we present these contributions, in terms of the modelling of context, strategies to include context into MT and the evaluation of contextual MT models. We will discuss perspectives on individual works throughout the thesis, but nevertheless conclude with a final (unnumbered) part discussing other perspectives for future work.

Below we provide a detailed summary of the thesis by chapter.

Part I: State of the Art: Contextual Machine Translation

Before studying methods for the integration of context into MT, it is first important to lay down a certain number of fundamental notions: what is context, what are the current standard architectures for MT, and how has contextual MT been dealt with in the past. This first part of the thesis is dedicated to providing these foundations, which we present successively: why context is needed in translation in Chapter 2, standard MT architectures in Chapter 3 (both phrase-based statistical MT and neural MT) and finally approaches to contextual MT in Chapter 4. We aim to emphasise the difficulties associated with the approaches, as well as key conclusions from these past works. Evaluation of MT quality will also feature throughout last two chapters, as this will be an important factor in Part II of the thesis.

Chapter 2: The Role of Context This first chapter introduces the main research problem: the role of context in resolving the types of ambiguity that can arise during translation. We define and illustrate three crucial notions to the understanding of this thesis: *translation*, *ambiguity* and *context*.

We distinguish three different types of ambiguity that can arise during the translation process from a theoretical point of view: (i) *source language ambiguity*, (ii) *cross-lingual ambiguity* and (iii) *target language ambiguity*, depending on the point in the translation process at which ambiguity arises. We provide examples of each of these types of ambiguity and discuss the nature of the context that can be used to disambiguate them.

We define the role of context itself from two points of view. The first concerns the nature of the context, whether it concerns the words of the text being translated (which we refer to as *linguistic context*) or the situation in which the text is produced (which we refer to as *extra-linguistic context*). The second concerns the type of ambiguity that the context can resolve. From this point of view, we define and distinguish two types of contextual phenomena according to our own translation-orientated definitions: those concerning the *coherence* of a translation (the relations between the *underlying meanings* of a text) and those concerning the *cohesion* of a translation (the relations between the *forms* of a translation). We finish by describing in detail examples of coherence-based and cohesion-based contextual phenomena. Among coherence-based phenomena we review lexical coherence, the translation of discourse connectives and information structure. Among cohesion-based phenomena, we focus in particular on anaphora translation and lexical cohesion. Our contributions in Part II cover a certain number (although not all) of these phenomena and we will regularly refer back to this chapter later on in the thesis.

Chapter 3: Sentence-level Machine Translation Chapter 3 is dedicated to a presentation of MT itself: the general principles, the techniques used and the current state of the art.³ We look at two different MT architectures: phrase-based statistical MT (Koehn et al., 2007) and encoder-decoder neural MT with attention (Bahdanau et al., 2015; Sennrich et al., 2017). Both of these will be used for the experiments presented throughout this thesis. Importantly, this chapter will only focus on sentence-level (i.e. non-contextual) MT. We begin by describing statistical MT (SMT) (Section 3.1), the basics of the approach and the improvements which led to current implementations of phrase-based SMT. We then describe the second paradigm, neural MT (NMT) (Section 3.2), and how it differs from the first paradigm presented. Finally, we review standard methods of evaluating MT, with a particular focus on the commonly used metric BLEU (Papineni et al., 2002), which has dominated the domain.

³Naturally, the state of the art has progressed over the course of the three years spent on this topic. We will focus on MT architectures that were state-of-the-art during the first two years of this doctorate, and more recent advances in MT will be mentioned briefly at the end of the chapter.

Chapter 4: Contextual Machine Translation This third chapter unites the two previous ones by reviewing previously used strategies to integrate context into MT. We begin by looking at efforts to take into account document-level context in automatic evaluation metrics. We choose to begin with this aspect as results have been of limited impact. Alternative evaluation strategies will be mentioned throughout the rest of the chapter, but will be defined on a per-phenomenon basis. The different strategies to incorporate context will be enumerated according to the phenomena they are designed to tackle. We structure the different strategies according to how context is modelled (i.e. structured) prior to being integrated into the translation process and at what point in the translation process context is exploited. We draw a distinction between the use of *unstructured context*, which is raw, unprocessed data, which may or may not contain the context required for translation, and *structured context*, which is context that has been selected, annotated or processed to target a particular context-dependent phenomenon. There has been a shift in the way contextual strategies are designed: from techniques using highly structured context targeting a specific phenomenon to techniques using unstructured context with the aim of simultaneously resolving a range of context-dependent phenomena. We review these strategies, starting with translation using structured linguistic context (Section 4.3) and continuing with translation using unstructured linguistic context (Section 4.4). Finally, building on the strategies seen for linguistic context, we review techniques for integrating extra-linguistic context into MT (Section 4.5).

Part II: Using contextual information for Machine Translation: strategies and evaluation

Part II is dedicated to our contributions to contextual MT for both linguistic and extra-linguistic context and to the development of new evaluation methods and resources. The contributions presented in this thesis do not represent an exhaustive study of techniques. We adopt a broad view of strategies and types of context, our aim being to test a panorama of strategies and methods inspired by current techniques in the domain.

We regroup our contributions in terms of the moment at which context is exploited in the translation process: during *pre-processing* (Chapter 5), during *post-processing* (Chapter 6) or *during the translation process proper* (Chapter 7). For the classification into these three categories, we adopt the same distinctions as made in Chapter 4 when describing existing contextual MT strategies.

Chapter 5: Adapting translation to extra-linguistic context via pre-processing

The first contribution in this thesis is the adaption of phrase-based SMT to speaker gender using a pre-processing strategy. Our experiments for English-to-French translation test

a simple approach inspired by domain adaptation techniques (Foster and Kuhn, 2007; Pecina et al., 2012), consisting in partitioning data into subsets each, containing utterances produced by a single gender (male versus female). These partitioned datasets are then used to train separate gender-specific model components (phrase tables and language models), which are mixed and separately weighted. Our results show that the technique does not result in significant improvements according to standard automatic evaluation metrics, and that no significant improvements are seen in gender-specific phenomena. One of the main limitations is the lack of annotated data, which is exacerbated by the choice of strategy. We will come back to adaptation to speaker gender in Chapter 7 with an alternative strategy using NMT and the integration of context directly into the translation system, rather than in a pre-processing step.

Chapter 6: Improving cohesion-based translation using post-processing In this second contribution chapter, we review two different experiments to integrate context into automatically produced translations in a post-processing step. The experiments each target a particular phenomenon and use a classification strategy, the prediction of which is used to post-edit translations.

The first of these experiments (Section 6.1) introduces a new research problem, which has not previously been studied in detail in MT: the generation of English tag questions when translating into English. Tag questions (interrogative constructions such as *you've been sitting in my chair, haven't you?* and *sit there, ok?*) are used for a variety of communicative purposes, expressing a range of attitudes from politeness to aggressivity. Our aim is to study a method of improving the generation of tag questions when translating into English from languages for which tag questions are rarer constructions. We propose a sequence of very simple post-editing strategies for three source languages (Czech, German and French) into English, formulating the problem as a multi-classification problem. Our main focus however is to investigate how to evaluate such a problem, given its subjective nature and relative rareness in real corpora. We review a variety of different strategies, in an aim to explore what constitutes a good and natural translation.

The second section presents another post-edition classification strategy, this time for the translation of anaphoric pronouns from English to French (Section 6.2). Our classification system, which is designed for the 2016 shared task on cross-lingual pronoun prediction (Guillou et al., 2016), is inspired by linguistic intuitions to the problem of anaphoric pronoun translation. Aiming to test whether few, high-level linguistic features are sufficient to perform highly in the task, we find that although our system benefits from the use of highly structured linguistic context, the quality of the preprocessing tools create a bottleneck in performance and our features are not sufficient to resolve the task. This second experiment will contrast with experiments in the following chapter (Section 7.1), which seek to improve anaphoric pronoun translation using an alternative strategy.

Chapter 7: Context-aware translation models Chapter 7 presents the third of the different strategies to integrate context into MT: integrating context during the training of the translation model itself. We look at the integration of both linguistic context (Section 7.1) and extra-linguistic context (Section 7.2), using an NMT architecture in both cases. The two series of experiments share a common methodology: using pseudo-tokens (Sennrich et al., 2016a) and multi-encoder strategies (Zoph and Knight, 2016; Libovický and Helcl, 2017; Wang et al., 2017) to include context in the training of MT models. The idea behind both methods is to exploit the capacity of neural MT models to learn how to selectively use its input to produce the best possible translation. We compare the different methods used, and focus in both cases on how to correctly evaluate the models' use of context. In our first experiments on the use of linguistic context, we adopt an evaluation strategy whereby we use NMT model's capacity to score translations to test how well the model can rank a correct translation higher than an incorrect one. We design two contrastive discourse test sets to test the models' capacity to ensure correct anaphora translation and to ensure correct lexical choice. We present a novel strategy to incorporating linguistic context, but show that a lot of progress still needs to be made, particularly in terms of lexical choice. In our second experiments on the use of extra-linguistic context, we adopt the same strategies and adapt them to integrate a range of different extra-linguistic features (speaker gender, speaker age, film genre and film year). We assess the impact of order when prefixing the source sentence with tokens representing each feature, and test multi-encoder approaches for the inclusion of multiple features. In our evaluation, we target one of the features, speaker gender, and evaluate models on their capacity to ensure gender agreement in contexts where it is determined by speaker gender. We show that contrary to our expectations, the simple pseudo-token strategy proves effective in all settings, despite the stress tests we apply to it. Multi-encoder strategies also succeed in exploiting extra-linguistic features, but produce translations of a slightly lower overall quality.

Chapter 8: DIABLA: A corpus for the evaluation of contextual MT Having previously discussed automatic evaluation methods to target improvements in specific context-dependent phenomena, in this final chapter, we again focus on MT evaluation, but this time from a perspective of human evaluation. Our aim is two-fold: (i) to evaluate MT performance in its end setting, when used to mediate spontaneous, informal dialogues, and (ii) collect spontaneous bilingual dialogue data that can be used for analysis and evaluation in future work. We present the design of our data collection method, which consists in collecting written MT-mediated dialogues between native English and native French speakers. The participants evaluate the quality of the MT models used to mediate the dialogues from a monolingual perspective, giving us an original way of assessing the errors detected by participants and the impact that these errors may have on dialogue flow. Human judgments are collected for each translated sentence,

and the participants also provide fine-grained error judgments (e.g. errors in terms of style, grammar, word choice, etc.). By comparing two of the MT models from Section 7.1, the baseline model and a lightly contextual model, we show that the human judgments provide interesting global trends that can offer insights into MT quality. The resulting MT-mediated corpus, containing 5,748 sentences (and their automatic translations) will form the basis of future analysis, and we present our plans to extend the corpus by adding linguistic annotations (e.g. coreference chains, named entity recognition) and human post-editions and/or reference translations so that the corpus can be used as a test set for new MT models.

Conclusions and perspectives

In this final part of the thesis, we step back to review the different contributions presented and the way in which they fit into the general trends seen within the domain of MT, notably guided by the major shift from SMT approaches to NMT. With this change have come new ways of integrating context that do not have to rely on structuring and processing context prior to it being integrated into MT. In Section 9.1 we provide a summary of our work from different points of view, comparing our different contributions linked to speaker gender adaptation and to pronoun translation and the different methods of evaluation we have used throughout our work. Since we have discussed short-term perspectives at the end of each contribution in Part II, we discuss more global perspectives in Section 9.2. Taking four particularly active areas of MT (evaluation, interpretability of MT models, low-resource languages, multi-modal MT), we review possible perspectives of contextual MT within each area.

1.3 Publications related to this thesis

Several of the contributions presented in Part II of this thesis have appeared in peer-reviewed conference proceedings.⁴

Chapter 5

- Bawden, R., Wisniewski, G., and Maynard, H. (2016). Investigating gender adaptation for speech translation. In *Proceedings of the 23rd Conférence sur le Traitement Automatique des Langues Naturelles (TALN'16)*, pages 490–497, Paris, France

⁴The results and analyses presented in this thesis may differ slightly from the published forms, as experiments have been improved and analysed in more detail.

Chapter 6

- Bawden, R. (2016). Cross-lingual Pronoun Prediction with Linguistically Informed Features. In *Proceedings of the 1st Conference on Machine Translation (WMT'16)*, pages 564–570, Berlin, Germany
- Bawden, R. (2017b). Machine Translation of Speech-Like Texts: Strategies for the Inclusion of Context. In *Proceedings of the REcontres jeunes Chercheurs en Informatique pour le TAL (RECITAL'17)*, pages 1–14, Orléans, France
- Bawden, R. (2017a). Machine Translation, it's a question of style, innit? The case of English tag questions. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP'17)*, pages 2497–2502, Copenhagen, Denmark

Chapter 7

- Bawden, R., Sennrich, R., Birch, A., and Haddow, B. (2018b). Evaluating Discourse Phenomena in Neural Machine Translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics; Human Language Technologies (NAACL-HLT'18)*, pages 1304–1313, New Orleans, Louisiana, USA

Chapter 9 (Section 9.2.1)

- Bawden, R., Lavergne, T., and Rosset, S. (2018a). Detecting context-dependent sentences in parallel corpora. In *Proceedings of the 25th Conférence sur le Traitement Automatique des Langues Naturelles (TALN'18)*, pages 393–400, Rennes, France

Part I

State of the Art: Contextual Machine Translation

Introduction

The aim of this first part is to present the current methods and evaluation metrics for contextual machine translation (MT), with a particular focus on extra-sentential context (both linguistic and extra-linguistic). Studying how context can be integrated into the translation process and how context-aware models can be evaluated on their capacity to correctly use this context requires first laying down some fundamental ideas concerning contextual MT: firstly, what context is and why it is a problem for MT, and secondly what are the current state-of-the-art techniques in MT.

We therefore begin in Chapter 2 with a discussion on the role of context in both human and machine translation. In Chapter 3 we present standard sentence-level MT systems (both phrase-based statistical systems and neural systems) and standard evaluation metrics used by the community. Finally, in Chapter 4 we bring the discussion of the two previous chapters together to look at how context has previously been handled in MT, in terms of strategies, evaluation methods and resources.

2	The Role of Context	15
2.1	Ambiguity and the problem of translation	16
2.2	The importance of context in MT	20
2.3	Conclusion	33
3	Sentence-level Machine Translation	35
3.1	Statistical Machine Translation (SMT)	37
3.2	Neural Machine Translation (NMT)	49
3.3	Evaluating Machine Translation	65

4	Contextual Machine Translation	73
4.1	Evaluating contextual MT	74
4.2	Modelling context for MT	79
4.3	Translation using structured linguistic context	82
4.4	Translation using unstructured linguistic context	93
4.5	Translation using extra-linguistic context	95
4.6	Conclusion on evaluating contextual MT	97

CHAPTER 2

The Role of Context

When translating natural language, whether translation is performed by a human or by a machine, the main difficulty encountered is ambiguity. At first sight, this may not appear different from any task involving the analysis of natural language, such as part-of-speech (PoS) tagging, syntactic analysis or semantic analysis. After all, even from a monolingual perspective, the non-deterministic nature of natural language means that a single element can give rise to several interpretations. At the morpho-syntactic level, a single wordform can be associated with several different parts of speech, for example *spoke* ‘past tense of *to speak*’ (verb) versus ‘a wheel part’ (noun). At the syntactic level, structural attachment can also be ambiguous. Take for example the commonly cited case of prepositional attachment in the sentence *Jack saw Jill with a telescope*, in which the prepositional phrase *with a telescope* can either be associated with *Jack* or with *Jill*. With respect to semantics, different concepts can be represented by a single wordform by homonymy and polysemy, for example the homonymous *bank* meaning either ‘a financial institution’ or ‘the land beside a river’. However these types of ambiguity do not illustrate the entire range of challenges faced when translating from one language to another.

In this chapter, we discuss why ambiguity is the main problem encountered in translation, whether it is manual or automatic, and illustrate the crucial role that context has to play in providing information useful to the resolution of ambiguity. We shall focus in particular on the translation of informal texts and on the role of context related to (i) the linguistic content exchanged during the dialogue and (ii) the general situation in which the dialogue takes place. This chapter will serve to introduce the main ideas behind the contributions presented in this thesis, and examples of the phenomena that we will study.

2.1 Ambiguity and the problem of translation

To better understand why ambiguity is a problem, it is worth first taking a step back to reflect on what the process of translation involves from a theoretical point of view. Translation is the transfer of a segment of text from one language into another, preserving as best as possible the intended meaning of the original segment. Our definition of *meaning* is very wide and refers to the communicative intention and content of an utterance: its semantic and pragmatic content, speaker attitude, style, formality, etc. The size of the segment depends on the translation situation, and, if performed by a machine, on the computational and modelling capacity of the machine and of the method used: it may be a whole text, a paragraph, a sentence or even a single word. Whilst humans typically translate whole texts, we shall see in the next chapter that MT systems must work with much smaller segments for computational reasons.¹ Ambiguity arises in the translation process when there is a choice between several formulations for a given input segment. However large the segment is, the potential for ambiguity is always there. One way in which it can be reduced is by increasing the size of the segment being translated. This equates to adding more textual content that may provide some of the information necessary to disambiguate the ambiguous elements within the segment.

Translation is particularly difficult, because there are multiple stages at which ambiguity can arise, which require different types of context to be resolved. The reason for this is that translation involves two different language systems. Each language individually has the potential for ambiguity. However, additional ambiguity emerges between the two systems due to different conceptual mappings in the languages, as we shall discuss shortly. Figure 2.1 gives a simplified representation of these three types of ambiguity, which we will describe in more detail below: (i) source language ambiguity, (ii) cross-lingual meaning transfer ambiguity and (iii) target language ambiguity.

2.1.1 Source language ambiguity

The first type (step (i) in Figure 2.1) is similar to the ambiguity encountered by any NLP analysis task when dealing with a single language. It concerns the semantic interpretation of the source segment and the fact that a single source segment can have multiple interpretations and meanings. Examples include syntactic ambiguity, homonymy and polysemy, as mentioned above. Two such examples are given below in Examples (5)

¹As we shall see in Chapter 3, much of the progress in MT is the consequence of improvements of what is considered to be a minimal translation segment. The first models were word-based (translation probabilities were based on individual word alignments). Later on, in so-called phrase-based models, translation probabilities were based on short sequences of words. More recently, neural MT models take into account all words within a sentence. Importantly, although the basic units of translation have changed over time, the maximal translation unit has remained the sentence, and information beyond the sentence is inaccessible for the majority of MT systems today.

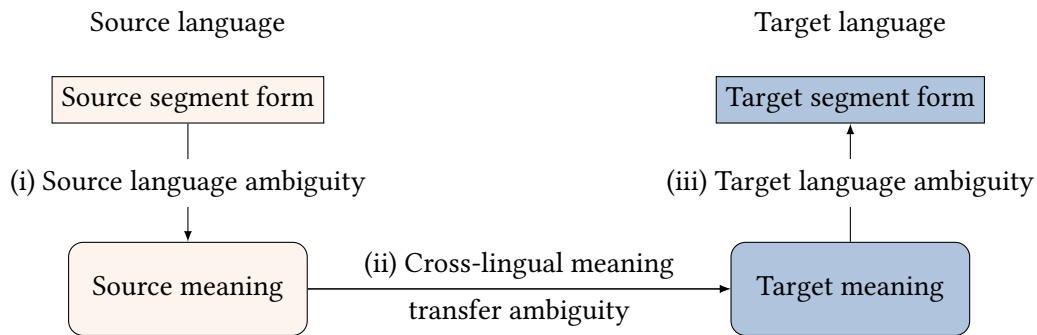


Figure 2.1: The translation problem from a theoretical point of view, with three points of potential ambiguity: (i) source-side, (ii) during semantic transfer, and (iii) target-side.

and (6), in which an inherent ambiguity in English must be resolved in the French translation: lexical semantic ambiguity in the first example and lexical semantic and syntactic ambiguity in the second.

(5) EN: It was far too **steep**

FR: C'était bien trop **cher/raide**.

Gloss: 'It was too **expensive/sharply inclined**.'

(6) EN: Christopher Robin **saw her duck**.

FR: **Christopher Robin a vu son canard** vs. **Christopher Robin l'a vu baisser la tête**.

Gloss: '**Christopher Robin saw the duck belonging to her** vs. **Christopher Robin saw her lower her head**.'

Unlike in NLP tasks concentrating on monolingual disambiguation, ambiguity present in the source language does not necessarily need to be resolved if the ambiguity can be preserved in the target language. This is highly dependent on the language pair involved in translation. Take for example the potential lexical disambiguation of the polysemous English word *glass*: *glass* can be translated into French using the same word *verre* whether the meaning is the solid, transparent material or the drinking receptacle. These two specific meanings therefore do not need to be disambiguated for this language pair. Similarly, the two separate meanings of the homonymous English word *crane* 'wading bird' or 'hoisting machine' can also be expressed using a single word *grue* in French. In cases such as these, the inherent ambiguity does not pose a problem for translation, because disambiguation, a choice between several target forms corresponding to each of the two meanings, is not necessary. The same cannot be said however for the translation of *crane* into Spanish, because of the necessity to disambiguate between the two forms *grúa* 'hoisting machine' and *grulla* 'wading bird', corresponding to the two meanings of *crane*. We will therefore focus only on ambiguity that is relevant in the translation process (specific to a language direction), which needs to be resolved.

2.1.2 Cross-lingual meaning transfer ambiguity

The second type of ambiguity (step (ii) in Figure 2.1) is specific to translation and concerns the passage from the meaning in the source language to the meaning in the target language. Ambiguity can arise during this transfer due to differences and mismatches in the conceptual spaces of the source and target languages. A simple example is the translation of English *owl* into French, which in everyday usage does not have a perfectly equivalent translation, there being instead two words *hibou* and *chouette* used to refer to two subspecies of owl. A similar example is given in Example (7), for the translation of the English word *river*, which in French must be translated as either *fleuve* or *rivière* depending on whether or not the river flows into the sea.

- (7) EN: They went swimming in the **river**.
FR: Ils ont nagé dans **le fleuve**/**la rivière**
Gloss: ‘They swam in **the river (flowing into the sea)**/**the rive (tributary of another river)**’

This type of ambiguity is famously seen in the differences in perception of colours and their naming conventions. The mapping of colours to colour names is not universal, as shown by the illustrations in Figure 2.2, adapted from (Regier et al., 2007). Translating colour names between languages whose conceptual mapping is different is therefore a complex feat requiring cultural knowledge.

Other common problems are linked to concepts that are highly associated to a particular culture or country, such that the concept is consequently language-specific. For example, there is no simple bijective semantic mapping between English *lawyer*, *solicitor* and *attorney* on the one hand and the French *avocat*, *juriste* and *notaire* on the other, since the functions are specific to the legal systems of the countries in which the languages are spoken.

Other, obvious examples of problematic elements for translation are national specialties such as *pasty*, *haggis* and *scone*, which do not have translation equivalents in most languages. As such, they do not pose a problem for ambiguity in the traditional sense, as it is more a case that there is no translation equivalent rather than several to choose from. However they do pose a problem of conceptual mapping, as in the previous examples. Social conventions involving politeness (which constitute useful information to be communicated), such as the use of honorifics (for example French *tu* ‘you_{INFORMAL}’ and *vous* ‘you_{FORMAL}’), also fit into this category. These do not always exist in the same form across languages. For example English *you* is used for all second person references regardless of familiarity or politeness, and on the other end of the scale, Japanese has a highly complex and hierarchical honorific system. Generating translations that correctly take such distinctions into account can be seen as essential for communicating the correct style and attitude.

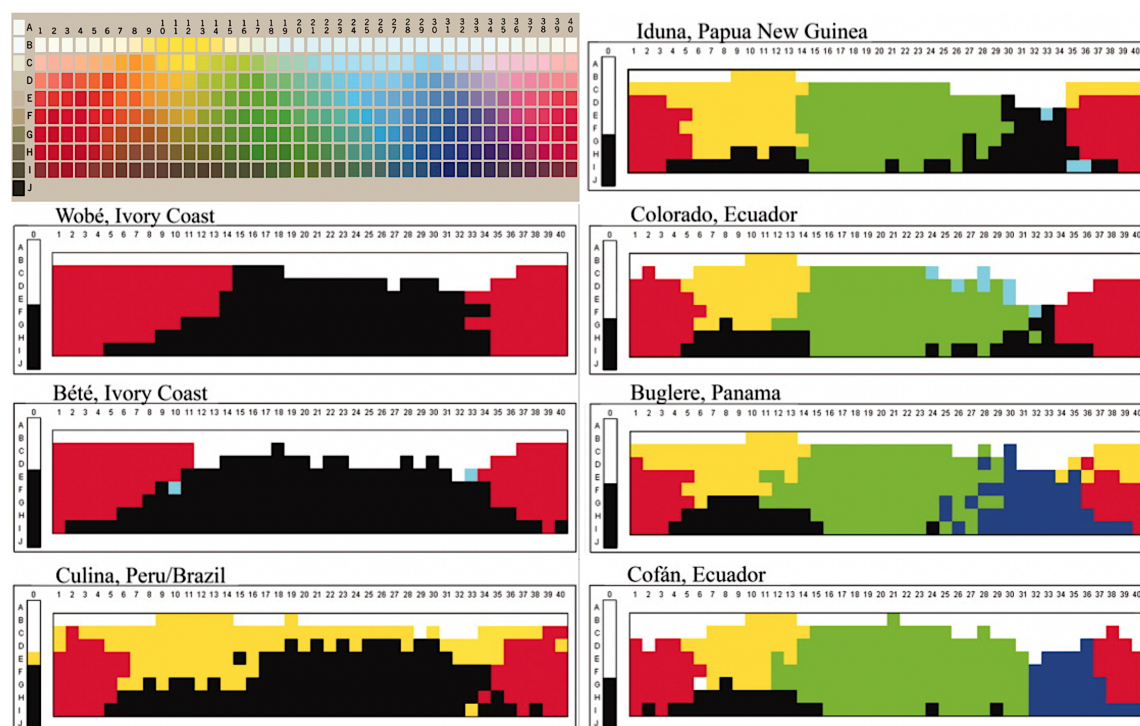


Figure 2.2: An illustration of the different mappings between the colour naming systems of various languages, showing that naming conventions cannot always be translated directly from one language to another. Figures from (Regier et al., 2007).

2.1.3 Target language ambiguity

The third type of ambiguity (step (iii) in Figure 2.1) concerns the mapping of the target meaning to its form in the target language, when information that is necessary to produce the correct form of the segment is not available in the meaning alone. Target-language ambiguity concerns the formal properties of a language, rather than the meaning of the utterances being translated. This notably concerns agreement phenomena, where the form of an element is determined by the form of another element, and the information is not otherwise available in the source segment. For example in French, which has grammatical gender agreement for nouns and pronouns, the correct choice of gender for the French equivalent of the anaphoric English pronoun *it* is determined by the grammatical gender of its coreferent, which is not always inferable from the meaning of the utterance or the source sentence if the coreferent does not appear within that sentence. This is the case in Example (8), where the translation of *it* must agree in gender with the translation of *egg*. We shall come back to the translation of anaphoric pronouns in more detail in Section 2.2.2.2.

(8) EN: The egg was on the wall. But suddenly it fell off.

FR: $\underline{L}'\text{œuf}_{\text{MASC}}$ était sur le mur. Mais tout d'un coup **il** est **tombé**_{MASC}.

FR: $\# \underline{L}'\text{œuf}_{\text{MASC}}$ était sur le mur. Mais tout d'un coup **elle** est **tombée**_{FEM}.

2.1.4 Human versus machine translation

What has been discussed so far is not specific *per se* to the problem of *machine* translation. A human translator is faced with much of the same ambiguity and must use all available information to correctly choose the correct interpretation of the source segment and the most adequate form in the target language. The added difficulty for an MT system is that the information available for translation is more restricted than for the human translator, who is equipped with social and cultural knowledge, and is more likely to have access to the entire text when translating. The most restricted scenario possible is the translation of words on an individual basis, without taking into account any surrounding context, and much of the difficulty in MT is also allowing access to surrounding context when translating. The idea of using neighbouring words to help disambiguation in word-based translation is far from new. In his 1949 memorandum, entitled *Translation*, Weaver (1955) writes:

...if... one can see not only the central word in question, but also say N words on either side, then, if N is large enough one can unambiguously decide the meaning of the central word.

As astutely remarked by Bar-Hillel (1960), this may be true for “*intelligent* readers” such as humans, but is insufficient for “electronic machines”, which lack the encyclopaedic knowledge necessary to use the context in a reasoned manner. Whilst humans can use context to guide their interpretation of a translation segment, the task is much more difficult for a machine, which must also be provided with a mechanism for using contextual information. MT systems must find ways of approximating the transfer of meaning from one language to another, which generally means learning a correspondence between the wordforms of the source and target segments. The way in which wordforms are modelled and mechanism by which the correspondence is learnt determine how expressive the models can be. As will be discussed in detail in Chapter 3, much of this progress has been achieved thanks to (i) a better form of representing wordforms, the minimal unit of translation, and (ii) changes within MT architectures, enabling an expansion of the size of the translation segment (from words to phrases and then to sentences) resulting in a better use of context *within* the sentence.

2.2 The importance of context in MT

A major drawback of most MT systems until now has been that the maximal translation unit has been the sentence, which amounts to translating sentences independently of each other. Beyond the level of the sentence, the *context of the sentences themselves* is most often ignored, both within the text (*linguistic context*) and outside of the text (*extra-*

linguistic context). For certain sentences, this means that the correct translation remains out of reach, however well the intra-sentential linguistic content is modelled. Take for instance Examples (9) and (10).

(9) EN: My sentence doesn't need context to be correctly translated.

FR: Ma phrase_{FEM} n'a pas besoin de contexte pour être traduite_{FEM} correctement.

(10) EN: But **mine** does.

FR: Mais la_{FEM} **mienne**_{FEM} si.

FR: #Mais le_{MASC} **mien**_{MASC} si.

Whereas the English source sentence in Example (9) can be correctly translated into French without the need for extra information, in Example (10), the correct French translation of *mine* requires knowing the grammatical gender of its antecedent (the French word *phrase* 'sentence') in order to choose the correct translation, the feminine variant *la mienne* 'mine_{FEM}', over the erroneous masculine variant, *le mien* 'mine_{MASC}' (marked with a # indicating that it is discursively inaccurate).² This example illustrates the fact that a text may be structured syntactically into sentences, but is above all a coherent unit, in which discourse phenomena and links span across sentence boundaries. Ambiguity within a sentence may be resolvable with intra-sentential context, but this is not always the case, and it is important to be able to look beyond the sentence to context within the surrounding sentences and even outside the text, to better guide translation.

2.2.1 What is context?

There has been a lot of debate concerning the exact definition of context. For many linguists, it is a concept that is "elusive of definition" (Widdowson, 2004). Mey (1993) refers to the general perception of context as "a rather undefined mass of factors that play a role in the production and consumption of utterances" and also "a notoriously hard concept to deal with". Our aim is not to provide a more accurate or complete definition of *context* than those already offered by the linguistic community, especially as our case specifically concerns the case of translation. However it is necessary to understand why context is important within our setting to understand how it can be used to improve MT. We will therefore suggest a general definition followed by an illustration of the utility of using context in MT through a selection of context-determined phenomena and examples.

The general role of linguistic context is to provide information to aid the interpretation of a segment of text that might not otherwise be correctly understood. As for the type of information that can make up context, our definition will remain relatively unrestrictive and therefore follow the definition offered by Hurford et al. (2007, p. 71):

²We will use the symbol # to indicate translations that incorrect in their discursive contexts, whether their incorrectness is due to grammatical reasons or other (for example pragmatic) reasons.

The **CONTEXT** of an utterance is a small subpart of the universe of discourse shared by speaker and hearer, and includes facts about the topic of the conversation in which the utterance occurs, and also facts about the situation in which the conversation itself takes place.

In other words, context includes any information shared between speakers that is necessary for the correct interpretation of a conversation (or text),³ or in our case, for its correct translation. However the degree to which linguists believe context to be necessary for the interpretation of a text varies considerably. Whereas traditionally MT simply disregards extra-sentential context, some viewpoints in linguistics go to the other extreme by suggesting that natural language can have no meaning whatsoever if taken out of context, such as for example Malinowski (1923, p. 307):

[A] word without linguistic context is a mere figment and stands for nothing by itself, so in the reality of a spoken living tongue, the utterance has no meaning except in the context of situation.

The viewpoint we adopt in this thesis will be somewhat more moderate than Malinowski's. Following attitudes such as that of Widdowson (2004), we acknowledge that not all sentences need context to be translated correctly. This seems to be supported by the fact that most MT systems are fairly successful even though they make the assumption that sentences can be translated independently of each other. As Widdowson states, opinions such as Malinowski's exaggerate the necessity of context for drawing conclusions about the meaning of a sentence and thus "undervalue the eliminating function of linguistic forms". What he means by this is that language is also highly conventionalised, and most of the time there is little ambiguity over what is meant by an utterance, even if theoretically there is a potential for ambiguity. This in no way diminishes the fact that certain phenomena *do* need context to be translated correctly, as we shall see in this chapter. Improving the translation of these elements can sometimes appear a thankless task. A majority of sentences do not necessarily require context to be correctly translated. The focal point of ambiguity is often a single word, whose correct translation does not result in large overall gains according to automatic evaluation metrics, yet can have a huge impact on the correctness and understandability of translation. In the following chapters, we will review evaluation metrics for MT, as well as strategies for specifically evaluating discourse-level phenomena.

2.2.2 Nature and use of context

Rather than simply listing different phenomena for which context is important for translation, we can study different angles of the problem by looking at two separate

³We shall be using *text* to refer to language productions, whether they are oral or written.

dimensions for the description of context: (i) the nature of contextual information and (ii) which sort of ambiguity it is used to resolve. In addition to being interesting from a theoretical point of view, we shall see later that these dimensions will also have an impact on how context is collected and how it can be integrated into MT systems.

Nature of context There are two main types of context that will be studied within this thesis: *linguistic context* and *extra-linguistic context*.⁴

The first, *linguistic context*, concerns any linguistic information found in the text itself and corresponds to what is commonly referred to as *discourse context*. In MT, linguistic context can be useful for a variety of different discourse phenomena such as coreference, grammatical agreement, lexical disambiguation, lexical cohesion, lexical repetition, discourse connectives, tag questions and other language devices that are not used in the same way in the two languages concerned by translation. We shall come back to these phenomena in more detail later in the chapter.

Extra-linguistic context concerns the “context of situation”, the relevant aspects of the setting, including those related to the speakers, time, place and culture in which the discourse is taking place. This can include anything from the attributes of the speaker and listeners (age, gender, upbringing, social status, their relationship (hierarchical, formal or informal), past events, whether they get on or not), to the situation in which the text is taking place (time, setting, purpose of the meeting), including common knowledge of the participants and events that may occur during the discussion and objects in the vicinity. Incidentally, information concerning the extra-linguistic context may be inferred from the linguistic context, if the information is expressed within the text. For example, a sentence *I am a woman* gives an indication that the speaker is female, as does the French sentence *Je suis heureuse_{FEM} de vous voir* ‘I am delighted_{FEM} to see you’, thanks to the feminine gender agreement of the adjective *heureux*. Extra-linguistic information may otherwise be inferred from raw data such as audio or video data, or made accessible as meta-information concerning the scenario and speakers.

Uses of context The role of context, as has previously been mentioned, is to provide information necessary to disambiguate and to guide the interpretation of ambiguous elements, in such a way that the meaning and style of the original text is retained in the translation.⁵ Context can help disambiguation at two levels: the text’s *coherence*

⁴There is a third type of context, *para-linguistic context*, that we shall not mention, concerning information communicated during the expression of the text, for example through prosody, gesture and facial expressions. We restrict study to written texts and therefore will not deal with this third context type.

⁵It is important to reiterate here that the aim of translation is to translate an entire *text*, and therefore also to ensure that the translation respects the logical and stylistic links between the elements of the text. This goes somewhat against the standard in MT, which usually consists in translating and evaluating at the sentence level, thereby ignoring the connective links between sentences.

and its *cohesion*. Widely studied in discourse analysis, these two aspects are commonly thought to be essential for defining what makes a text communicative. There is debate concerning the definitions of the aspects and their exact perimeter. We choose to follow the definitions of de Beaugrande and Dressler (1981, pp. 3-4):^{6,7}

- Coherence “concerns the ways in which the components of the textual world, i.e. the configuration of CONCEPTS and RELATIONS which *underlie* the surface text are *mutually accessible* and *relevant*.”
- Cohesion “concerns the ways in which the components of the SURFACE TEXT, i.e. the actual words we hear or see, are *mutually connected within a sequence*. The surface components **depend** upon each other according to grammatical forms and conventions, such that cohesion rests upon grammatical dependencies.”

In other words, there is a distinction between properties of the text concerning the surface forms (related to (iii) target ambiguity in Figure 2.1) and those concerning the underlying concepts (related to (i) source ambiguity and (ii) cross-lingual ambiguity in Figure 2.1).

Coherence therefore concerns the relevance and configuration of the ideas, concepts and relations underlying the text. A coherent structure is determined by the correct communication of the intended meaning, structured in a relevant and coherent way. Example (11) provides an illustration of a coherent versus an incoherent text. The first sequence of sentences is coherent, as it represents a logical sequence of events, whereas inverting the sentences, as in the second version, breaks this logical sequence. A second example of incoherent text is given in Example (12), in which the two sentences are unrelated and therefore do not form the unified whole necessary for a text to be coherent.

(11) EN: The apple fell from the tree. It then hit him on the head.
EN: #It then hit him on the head. The apple fell from the tree.

(12) EN: #The king was counting his money. Cats like cheese.

Cohesion on the other hand concerns the relationships of textual forms to one another, i.e. the *connectedness* of the text, specific to the way in which language is encoded by the language system. Examples of phenomena contributing to a text’s cohesive nature include those related to grammatical cohesion: anaphoric references, as in Example (13), ellipsis, as in Example (14), etc.) and those related to lexical cohesion (repetition of lexical items, as in Example (15), use of synonyms, collocations, etc.) (Halliday and Hasan, 1976).

⁶The typographical emphasis is preserved from the original citation.

⁷De Beaugrande and Dressler (1981) in fact distinguish seven *standards* to which a text must adhere to be defined as a “communicative occurrence”. Coherence and cohesion are what the authors refer to as *text-centred notions* and as such will be the object of our study. The other standards are *user-centred notions*: intentionality, acceptability, informativity, situationality and intertextuality. In MT, the first two *text-centred* notions are those that most directly concern the use of context to resolve ambiguity.

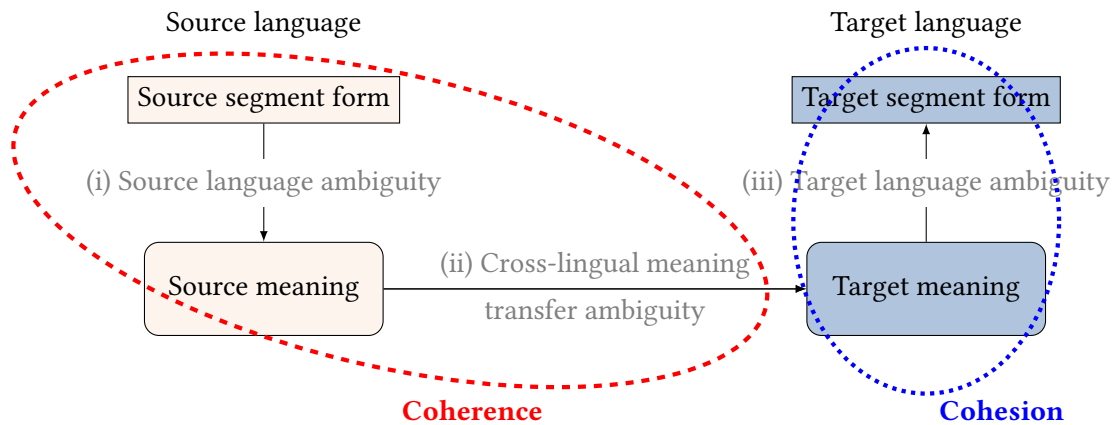


Figure 2.3: The translation problem from a theoretical point of view, with three points of potential ambiguity: (i) source-side, (ii) during semantic transfer, and (iii) target-side. We define the first two sources of ambiguity as corresponding to *coherence*-based ambiguity and the third source as corresponding to *cohesion*-based ambiguity.

- (13) FR: La pomme_{FEM} est **juteuse**_{FEM}. **Elle**_{FEM} est **prête**_{FEM} à être **mangée**_{FEM}.
 ‘The apple is juicy. It is ready to be eaten.’
- (14) EN: Have you checked the oven? I might **have**.
- (15) EN: The cat’s stuck up a tree. Which **tree**?
 EN: #The cat’s stuck up a tree. Which **oak**?

From the point of view of translation, context helping to resolve ambiguity arising on the source-side or cross-lingually can be equated to aiding the *coherence* of the text (choosing the correct words in context), whereas context helping to resolve ambiguity on the target-side can be seen as aiding the *cohesiveness* of the text (how well the textual elements of the translation are linked between each other). We show these interactions in Figure 2.3, superimposed onto the original translation schema in Figure 2.1, and give some detail examples of each of these types below. Note that we purposefully choose to separate the scope of these two notions, whereas in the linguistics literature, cohesion is sometimes considered to be a subpart or precondition to coherence (Halliday and Hasan, 1976; de Beaugrande and Dressler, 1981). In the remainder of this document, in relation to the process of translation in particular, we consider the two aspects to form part of a successive chain leading to a correct translation: *coherence* concerning the correct relation between the meaning of the original text and that of the target translation, and *cohesion* to be the correct coding of the meaning to linguistic form in the target language. For this reason, our definitions and classification of phenomena may differ somewhat from those cited in the literature. We give a more detailed description and discuss phenomena related to each of these concepts in the next two sections.

2.2.2.1 Coherence-based phenomena

A text is *coherent* when it is logical and semantically consistent, and the sentences within it form a logical succession of ideas that are relevant and well linked. When translating a text from one language to another, the text's coherence should be preserved regardless of the fact that translation may be performed on a segment smaller than the whole text, for example at the sentence level. In the specific case of translation, we shall assume that the source text is already a coherent text, and therefore the task of producing a coherent translation amounts to conserving the coherent nature of the text as best as possible.

The difficulty when translating is the fact that the language systems of the source and target languages differ, creating potential ambiguity, which, if unresolved, could lead to an incoherent translation. Here we discuss three different aspects contributing to discourse coherence: lexical coherence (concerning the semantically relevant word choice), the translation of discourse connectives, and information structure (how information within a sentence is packaged).

Lexical coherence (and word sense disambiguation) Lexical coherence concerns the semantic connections between the words of the text, and therefore how well a particular lexical choice fits semantically (and pragmatically) within the current discourse. Choosing the correct translation in keeping with lexical coherence means choosing target forms for words that together preserve their source meaning, despite possible ambiguity either in the source language or in the conceptual mapping between the source and target languages. In this respect, ensuring lexical coherence is treated within this thesis as equivalent to lexical disambiguation in the context of discourse. The cases on which we shall focus are therefore those in which the textual elements are ambiguous with respect to their translation, concerning both the first and second kinds of ambiguity in Figure 2.1.

At the very beginning of this chapter, we cited a number of examples of ambiguity types in natural language: morpho-syntactic ambiguity, syntactic ambiguity and semantic ambiguity. Yet here, in the context of translation, we only appear to discuss one of these types, lexical (semantic) ambiguity. The reason for this is that many forms of ambiguity are not present without there also being lexical ambiguity of some sort within the sentence. If this lexical ambiguity is resolved, this also often disambiguates the other forms of ambiguity. A clear example of this is the previously mentioned English sentence *I saw her duck*, which contains ambiguity on three linguistic levels: (i) morpho-syntactic ambiguity of *her* and *duck* (*her* as either an object pronoun or a possessive pronoun and *duck* as either a noun or a verb), (ii) syntactic ambiguity (*her duck* as the direct object of *saw* or *her* as the object of *saw* and subject of *duck*) and (iii) lexical ambiguity of the same two words (*duck* signifying either the bird or an action of lowering one's head). If we

were to translate this sentence into another language in which these ambiguities cannot be preserved, the choice between the two interpretations could in practice be made based uniquely on the disambiguation of the single lexical item *duck*. If an element of context enables us to ascertain that *duck* refers to the bird (or that it refers to lowering one's head), then the morphological and syntactic ambiguities are instantly resolved. It is often far easier to consider such examples in this light, because it simplifies the ways in which we perform disambiguation. As we shall see in the following chapter, the standard MT systems we will be using in this thesis do not rely on explicit morphological, syntactic or semantic analysis, and therefore all ambiguity comes down to a choice of the best sequence of translated wordforms given the other word choices within the sentence.

Discourse relations and discourse connectives An important part of a text's coherence is ensuring the logical links between sentences. Relations between sentences (or clauses) are known as discourse relations, and they can be explicitly rendered in the surface form of the text as discourse connectives.⁸ Discourse relations are associated with different senses depending on their function, such as causal (*because, since, as*), temporal (*as soon as, when*), conditional (*if*), purpose (*so that*), etc. (Prasad et al., 2008). Discourse connectives can be ambiguous, both in terms of having discourse and non-discourse uses, and in terms of having different functions (i.e. being associated with different discourse relations) depending on the context (Meyer et al., 2011; Roze et al., 2012). For example, in (16), the word *and* is ambiguous between the first instance which is simply a conjunction (with no discourse sense) and the second in which it represents a temporal discourse relation. In (17), the connective *while* has a discourse sense in both instances, but the first has a temporal sense whereas the second has a contrastive one.

(16) Little Jack Horner was eating Christmas pie *and*_{NONE} ice cream.
He put in his thumb *and*_{TEMPORAL} pulled out a plum.

(17) The king counted his money *while*_{TEMPORAL} he waited for his supper.
The king had lots of money *while*_{CONTRASTIVE} the pauper did not.

As well as this potential ambiguity, the use of discourse connectives is language-dependent. Discourse relations can even be implicit, in which case they are induced by the discourse context and not realised by a discourse connective at all, and the degree to which discourse relations are implicit depends on the language (Zufferey, 2016). The implicitation of discourse connectives in translation has been studied by Meyer and

⁸Contrarily to a number of authors who treat discourse connectives as an element pertaining to the lexical *cohesion* of a text (Halliday and Hasan, 1976; de Beaugrande and Dressler, 1981; Sim Smith, 2017), we consider, as per our earlier definition of lexical cohesion and lexical coherence, that the translation of discourse connectives concerns the lexical *coherence* of a text, as they encode the logical relations between sentences. Although their surfacic form (how they are rendered if at all) is dependent on the language, we consider that this is not directly determined by the surfacic forms of the other words in the text.

Webber (2013), who reveal that MT systems tend to overtranslate discourse connectives compared to human reference translations. They show that the human translations contain more zero-translations of discourse relations (where a discourse connective in the source language is not translated into the target language) than MT outputs for both English-to-French and English-to-German translation. This is a particularly difficult area of study, because the decision to translate (or not) a discourse connective can be one related to style and other subtle factors within the discourse context.

Information structure A third aspect, which is crucial for discourse coherence, is the way in which information is formally structured within sentences. The way the information is packaged is typically said to depend on properties of the concepts' roles within an utterance, such as *givenness*, *focus*, and *topic* (Chafe, 1976), which can be realised through different linguistic devices (e.g. intonation, active/passive, word order, clefting, dislocation, anaphora, etc.). The importance of information structure for translation lies in the fact that linguistic mechanisms are different from one language to another, and therefore it cannot be assumed that a piece of information will be packaged in the same way in the source and target languages. This involves a more holistic approach to translation than translating the elements of the source sentence individually, and a vision that includes the possibility of restructuring the information within the sentence on the target side.

To illustrate cross-lingual differences in information structuring, let us look at the example of left-dislocation in French. French, like English, is traditionally considered to be a language with SVO order, and therefore a canonical sentence would be of the form *Les chats aiment le lait* 'Cats like milk'. However, particularly in spoken French, there are other constructions commonly used other than this canonical sentence. One commonly used construction to introduce a new focus of discussion is dislocation, whereby the salient element is detached from the main clause and dislocated either to the left or the right and replaced in the main clause by a reduced form. *Les chats aiment le lait* could therefore become *Les chats, ils aiment le lait* '(lit.) The cats, they like milk'. More complex instances of dislocation are also possible, and are very common in spoken French (cf. Examples (18) and (19)).

- (18) Tu leur as donné à manger, aux chats ?
'(lit.) Did you give them_{DAT} [something] to eat, to the cats?'
Did you feed the cats?
- (19) Il l'a-t-il jamais attrapé, le gendarme, son voleur ?⁹
Did the police officer ever catch his thief?
'(lit.) Did he ever catch him, the police officer, his thief?'

⁹Example from (Vendryès, 1921, p. 103), cited by Queneau (1950, p. 15)

2.2.2.2 Cohesion-based phenomena

A text is *cohesive* when its surface forms are meaningfully connected to each other. In other words, cohesion concerns the relation between elements of a text, whereby one element's interpretation is dependent on another one. As described by Halliday and Hasan (1976), "the one presupposes the other, in the sense that it cannot be effectively decoded except by recourse to it". Contrarily to coherence-based phenomena discussed in the previous section, the element of context is a specific element, there being a well-defined cohesive link between the ambiguous dependent element and the determining context. This cohesive link between the dependent element and its context is linked either to grammatical agreement, for example reference phenomena, substitution and ellipsis, or to lexical repetition and collocation. The most apparent examples of such phenomena for translation are anaphoric phenomena (coreference, reference and substitution) and lexical cohesion, which we shall now discuss in a little more detail.

Anaphoric phenomena Within a text, words can refer to entities, objects or events that have previously been referred to, will be referred to later on, or are deducible from the situation in which the text occurs. It is common for at least some of these references to be represented by a reduced form, such as a pronoun, or by an ellipsis, when the referent is salient. Anaphora is a linking phenomenon by which these abbreviated referential expressions owe their interpretation (and their form) to the full expression they refer to, which we will refer to as the anaphor's *antecedent*.¹⁰ One of the most studied forms of anaphora is coreference, whereby expressions refer to the same entity. However a variety of different relationships can also be referred to as anaphoric. For English, they include possessives, relative pronouns, demonstrative pronouns, the use of indefinite pronouns, one-anaphora, nominal ellipsis and verbal ellipsis, of which we shall say more below.

The cohesive links between elements are not in themselves problematic for translation. The difficulty arises when the link is expressed through the target language's language system in a way that is not reflected in the source language. It is common for anaphoric elements to be morphologically marked for the attributes of their referent. However the way in which they are marked, if at all, is dependent on the language. For example, French, like many Romance languages, has grammatical gender marking for all nouns and personal pronouns, and determiners and adjectives agree in number and gender with the nouns they qualify. In English, only the personal pronouns *she* and *he* are gender-marked, and adjectives are gender-neutral. When translating anaphoric elements from English to French that are gender-neutral in English but gender-marked in French, the gender must be deduced from the translation of the anaphor's antecedent, a link that is often not

¹⁰The full form can also appear after the reduced form, in which case the reduced form is referred to as a cataphor rather than an anaphor and the full form the *postcedent*. However, in the following discussion, we will refer to both anaphora and cataphora under the umbrella term *anaphora*.

trivial to resolve automatically, especially when several possible candidate antecedents are available. The following examples illustrate the anaphora types mentioned in the previous paragraph and the difficulties they pose for English-to-French translation. In each case, the gender marking of the anaphor (in bold) is dependent on its antecedent (underlined) in the target language, information that is not otherwise present in the corresponding anaphor in the source language:

(20) Coreference:

EN: The moon shone in the sky. **It** looked like a big cheese.

FR: La lune_{FEM} brillait dans le ciel. **Elle**_{FEM} ressemblait à un grand fromage.

(21) One-anaphora:

EN: Templeton loved cheese. Especially **stinky ones**.

FR: Templeton adorait le fromage_{MASC}. Surtout **ceux**_{MASC} qui puaien.

(22) Possessive pronouns:

EN: He liked all cheese, and even made **his own**.

FR: Il adorait tous les fromages_{MASC} et fabriquait même **le sien**_{MASC}.

(23) Possessive pronouns

FR: Templeton_{MASC} en avait une collection. **Sa**_{FEM} maison_{FEM} en était pleine.

EN: Templeton_{MASC} had a collection. **His**_{MASC} house was full of them.

It is even possible for there to be no antecedent (or indeed postcedent) within the text. This case calls for a distinction to be made between *endophoric* expressions, those anaphors whose references can be found within the text (i.e. in the linguistic context), and *exophoric* expressions, those anaphors whose references are found outside of the text (i.e. in the extra-linguistic context). In the case of exophoric expressions, as the referential expression that determines the form of the anaphor does not appear in the text, it must be deduced from extra-linguistic context. One case in which this commonly occurs is when a speaker refers to themselves or to the interlocutor and uses an adjective or past participle that in the target language must agree in terms of gender with the referent. The following example illustrates this with the translation of English *happy* into French *content*_{MASC} or *contente*_{FEM} depending on the gender of the speaker. The information must be available in metadata to be used in the translation process, inferred from linguistic context, or extracted from raw extra-linguistic data such as audio, images or video.

(24) Speaker gender marking:

EN: **I** am very happy for him.

FR_{MASC}: **Je** suis très **content**_{MASC} pour lui.

FR_{FEM}: **Je** suis très **contente**_{FEM} pour lui.

Sometimes, the same type of phenomena cannot be used to translate into the target language, because the mechanism does not exist. For example, verbal ellipsis is common

in English, shown below with the auxiliary *have* referring back to ***have... made cheese before***. Such ellipsis with repetition of an auxiliary is not possible in French, and the translation must be rendered using an alternative expression, conveying the same meaning (*et toi ?* ‘And you?’).

(25) Verbal ellipsis:

EN: I have never made cheese before. **Have** you?

FR: Je n’ai jamais fabriqué de fromage. Et toi ?

In terms of the translation of anaphoric pronouns, the situation is in reality more complicated than the binary masculine/feminine ambiguity suggests. French also has singular neutral pronouns *ce/ceci*, *ça/cela*, which can be used as translations for *it*. Even the singular/plural distinction, which you would think would be the same in English as in French, is not always the same across languages. For example, in English, certain collective singular nouns such as *team* and *family* can be referred to using a plural *they*, whereas in French the use of a plural pronoun in such cases is not permitted, as illustrated in Example (26).

(26) EN: The team_{SG} were_{PL} getting ready. **They**_{PL} had a big match ahead of **them**_{PL}.

FR: L’équipe_{SG} se préparait_{SG}. **Elle**_{SG} avait un gros match devant **elle**_{SG}.

The correspondence of English and French subject pronouns is ambiguous in both directions, as can be seen in Figure 2.4, even when we do not consider the possibility of paraphrasing, which could result in a pronoun being translated by an element other than a pronoun.

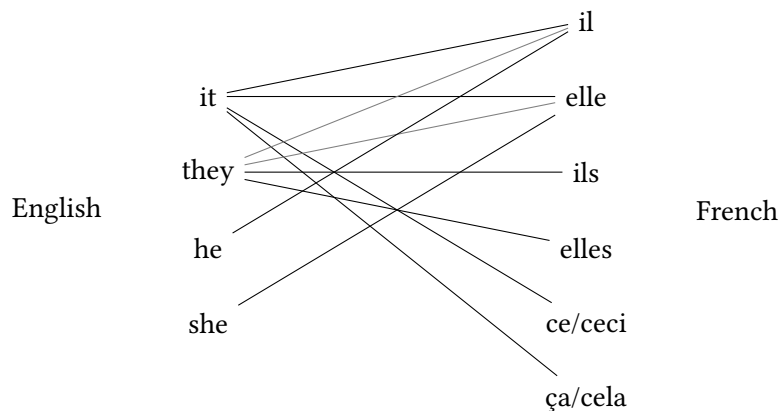


Figure 2.4: An illustration of the most common mappings between subject personal pronouns in English and French. The correspondence between *they* and *il* and *elle* is indicated in grey, because it mainly concerns collective nouns and gender-underspecified singular *they*.

Lexical cohesion As a form of cohesion, lexical cohesion also presents a dependency between a textual element and another element within the text, but in terms of lexical

choice rather than grammatical marking. Lexical cohesion represents two types of relation: *repetition* and *collocation* (Halliday and Hasan, 1976), which determine whether the relation is one in which the same word, a synonym or a superordinate wordform is imposed by the other element (in the case of repetition) or whether the word choice is guided by the fact that the elements frequently co-occur (in the case of collocation). Our definition of lexical cohesion concerns only those wordforms whose form is *dependent* on the choice of other wordforms in the text. Two wordforms being synonyms may be incidental and simply a property of a text, and we consider that this relation is only considered a contributor to lexical cohesion if the relationship of synonymy is formally necessary (as opposed to for example choosing exactly the same wordform in both cases).

An illustration of lexical repetition is given in Example (27), where the word *crazy* is repeated in the second sentence as a way of referring to the first instance in the previous sentence. For the French translation to communicate the same meaning as the source sentence, the same translation must be used in French for the three translations of *crazy*, creating an inter-dependency between the choice of forms used for the three translations of this word. The second French translation is therefore not cohesive and fails to conserve the intention of the source sentence.

(27) EN: The king is crazy. And **crazy** kings make **crazy** countries.

FR: Le roi est fou. Et les rois **fous** font des pays **fous**.

FR: #Le roi est fou. Et les rois **déments** font des pays dingues.

Repetition can also concern highly related words such as synonyms. An interesting case is when translating using two distinct synonyms (and not the exact same word) would be obligatory for the understanding of the text. Example (28) shows such a case, in which the repetition of *happy* (and its translation) is obligatory, but the use of a synonym for the third term *content* is also necessary to avoid the unhelpful and nonsensical translation given in the second translation indicated with a #.

(28) EN: The King was happy, if by **happy** you mean **content**.

FR: Le roi était heureux, si par **heureux** vous voulez dire **content**

FR: #Le roi était heureux, si par **heureux** vous voulez dire **heureux**

‘The king was happy, if by happy you mean happy.’

Lexical cohesion can also be observed between words that are semantically related, within a same lexical field. The use of certain terms contributes to the general cohesive nature of the text. It has been observed that speakers align their verbal behaviour to each other in order to show affinity, and the use of similar vocabulary plays a part in their communicative nature (cf. for example Giles et al. 1991).

The dependency between wordforms can also be evident in certain settings, where the phonetic properties of the words plays a major role. An example of this is poetry, in

which characteristics such as rhyme, alliteration and assonance all contribute to the cohesive nature of the text through the inter-dependency of wordform choices, based on the similarities between their phonetic properties.

2.3 Conclusion

Although not all types of context mentioned in this chapter will be tackled within this thesis, the definitions they help to illustrate provide us with a theoretical grounding on which the remainder of the thesis shall rest. Context is hugely important for translation, as has been illustrated through concrete examples of different context-dependent phenomena. We have focused in particular on the notion of *extra-sentential context*, the main focus of this thesis, since the integration of information beyond the sentence boundary is currently one of the major stumbling blocks of MT techniques, as we shall see in the remainder of this thesis.

The definitions we have provided will serve as a basis for discussion throughout our work, particularly concerning the description of strategies to include context and the evaluation of particular phenomena, and should help to better understand why certain strategic decisions are taken. We have presented a number of distinctions concerning the nature (*linguistic* versus *extra-linguistic context*) and uses of context (*lexical coherence* versus *lexical cohesion*), which will be referred to in the following chapters.

In the next chapter, we shall set extra-sentential context aside whilst we describe the sentence-level MT architectures that will be used as a basis for our experiments in Part II. We shall return to the notion of context beyond the sentence boundary in Chapter 4, in which we will review which strategies have been previously used to integrate context into MT.

CHAPTER 3

Sentence-level Machine Translation

Machine translation (MT) has undergone remarkable progress since the first theoretical formulation of the problem and the first implementations in the 1930s, 1940s and 1950s (Troyanskii, 1935; Shannon, 1948; Weaver, 1955). Much of the theory formulated in these early works is still contributes to the theoretical foundation of modern-day implementations. However the way in which translation is modelled in practice has changed over the years, both in the techniques adopted and in the degree to which context from within the sentence is used when making individual translation decisions.

As has been discussed in the previous chapter, one of the major challenges of translation is ambiguity. The simple approaches used in the early years of MT were insufficient for dealing with even simple cases of ambiguity (such as homonymy), since they largely relied on word-to-word translation, without sufficiently taking into account words' contexts, if at all. MT was dominated by *rule-based approaches* until at least the 1980s, which relied on expert linguistic knowledge to account for the vast number of exceptions and complexities of natural language. They relied on huge bases of linguistic rules (Vauquois et al., 1965), and therefore, despite being relatively robust provided that rule coverage was large, were unwieldy and inflexible. The theory behind methods of MT relying on statistical modelling (Shannon, 1948; Weaver, 1955) was developed early on, forming the basis of many current day implementations. However, the results were limited due to lack of data on which to estimate probabilities and a lack of processing power. Thanks to the development of resources and computational advances, *statistical approaches* to MT (SMT) became more widespread in the 1980s and 1990s and little by little replaced rule-based approaches. The focus shifted to developing methods of

automatically inferring linguistic information from corpora, which could then be used to find the most probable translation of a given source sentence. Despite the simplicity of initial implementations, the popularity of statistical methods in MT research eventually overtook rule-based methods and remained state-of-the-art until *neural MT* (NMT) began to rival and then to outperform SMT in the mid 2010s. In a situation parallel to conventional SMT, much of the theory and architecture behind NMT had existed for decades (Allen, 1987; Pollack, 1990), and it has been largely thanks to an increase in computing power (advances in GPUs) that these methods could finally be implemented (and improved) on a large scale. In terms of theoretical advances in translation modelling, much of the progress that has been seen in SMT and NMT has been in the way translation probabilities are modelled and the way in which the space of candidate translations is explored. In terms of translation ambiguity, an important factor in improving the expressivity of models has been in the increasing use of contextual information within the sentence to calculate translation probabilities. As we shall see in this chapter, word-based translation has progressively given way to phrase-based translation, and finally to translation on the sentence-level.

The existence of SMT and NMT, both *data-driven approaches*, was made possible thanks to the development and distribution of large corpora, known as *parallel corpora*, consisting of sentence-aligned translations from which translation probabilities could be estimated. In order to be able to generalise to unseen sentences, translation models require very large quantities of data to estimate sufficiently reliable probabilities, and are nevertheless often restricted in their modelling capacity in order to reduce the effect of data sparsity. Traditionally, large parallel corpora existed mainly in the news and parliamentary domains (e.g. the Hansard parallel corpus (Roukos et al., 1995) and EUROPARL (Koehn, 2005)), but today are also available for more informal domains such as TED talks (Tiedemann, 2012) and film subtitles, OPENSUBTITLES2016 (Lison and Tiedemann, 2016). The approaches studied within this thesis all rely on large parallel corpora for training, although they vary in the way MT models learn from such data.

In this chapter, we describe in more detail how MT has progressed over the years, from word-based and phrase-based SMT in Section 3.1 to NMT in Section 3.2. Despite the improvements seen in the use of linguistic context within the sentence, the translation architectures presented in this chapter follow standard practice in MT and take the sentence as the maximal translation unit, meaning that any context beyond the sentence is not taken into account. It is nevertheless important to lay down the foundations of these sentence-level architectures, before discussing in the following chapters ways in which extra-sentential context can be included. The discussion therefore focuses on the two different MT paradigms that will be used throughout this thesis: phrase-based SMT (Section 3.1.2) and sequence-to-sequence NMT with attention (Section 3.2). We conclude the chapter with a review of standard automatic MT evaluation, applied to the sentence-level, before addressing specific, contextual evaluation strategies in the next chapter.

3.1 Statistical Machine Translation (SMT)

As a data-driven method of MT, translation in SMT is modelled as a search for the most probable translation of a source sentence, whereby model parameters are learnt based on the properties of source sentences and their translations in large parallel corpora. The aim is to find a hypothesis translation \hat{t} of a source sentence s with the maximal probability of being the correct translation $P(t|s)$.¹ The difficulty lies in how this probability is calculated for a given candidate translation and how to explore the space of possible candidate translations without performing an intractable exhaustive search.

Decades before the first statistical implementations of MT were proposed in the 1980s (Brown et al., 1988), the foundations of statistical approaches to MT were already being laid out, the question of how to model the probability of a translation given a source sentence being largely inspired by Information Theory and the Noisy Channel Model (Shannon, 1948). The Noisy Channel Model is a simple model of communication, whereby a message between a sender and a receiver is distorted during transmission by a noisy communication channel, which has the effect of *encoding* the sender’s original message and producing a corrupt version at the receiver’s end. As illustrated in Figure 3.1, we consider the source sentence to be a corrupted version of the translation we wish to recover (as if this target language translation were the original uncorrupted message). The process of translation can be seen as attempting to undo this noise to find the translation of the observed source sentence.²

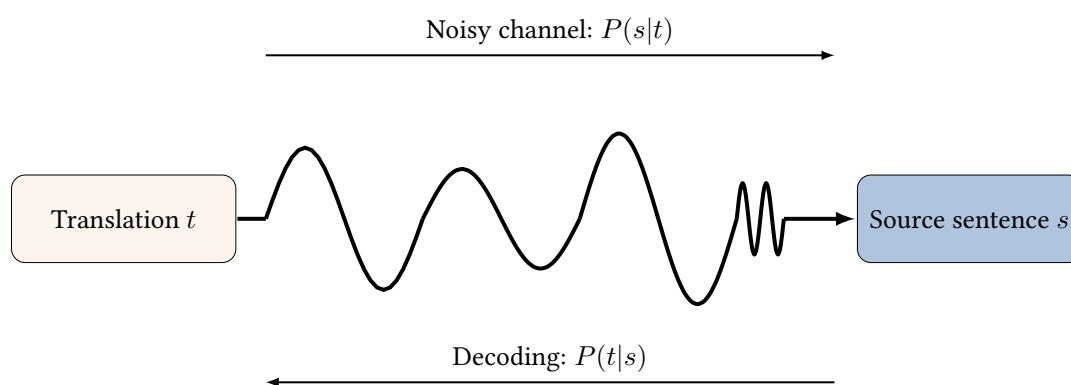


Figure 3.1: Illustration of the Noisy Channel Model of communication (Shannon, 1948), as applied to the translation (decoding) problem.

¹Traditionally in the literature the source and target languages are represented by the symbols f (for ‘foreign’) and e (for ‘English’), reflecting a somewhat anglo-centric view to translation. We choose instead to use the symbols s and t by virtue of their language-independent nature.

²A source of confusion when applying the noisy channel to MT is that the *target sentence* of the MT process (i.e. the sentence to be recovered) is modelled as the *source message*, and the original sentence of the MT process (the *source sentence*, in MT terms) as the *output of the noisy channel*. This can also be seen in other NLP applications to which the noisy channel approach is often applied, such as speech recognition or optical character recognition. The application of the model is more direct such as spell-checking, where the observed message is a poorly spelt sentence and the original message, which must be recovered, is the correctly spelt version.

This approach to SMT is generally formalised using a Bayesian reformulation, whereby, in order to find the translation maximising the probability $P(t|s)$, the probability $P(s|t)$ is used:

$$\hat{t} = \operatorname{argmax}_{t \in T} P(t|s) \quad (3.1)$$

$$= \operatorname{argmax}_{t \in T} \frac{P(s|t)P(t)}{P(s)} \quad (3.2)$$

$$= \operatorname{argmax}_{t \in T} P(s|t)P(t), \quad (3.3)$$

where t represents a translation and T the set of all possible translations. The passage between the two last steps is made possible since $P(s)$ is the same for all possible translations. The problem therefore becomes one in which, instead of directly maximising the probability of the translation given the source sentence, we seek to jointly maximise two probabilities: $P(s|t)$, representing the translation probability (calculated by a *translation model*), and $P(t)$, the probability that the translation sequence belongs to the target language (calculated by a *language model*). These two elements are customarily seen as reflecting the probability that the translation sequence is *faithful* to the source sequence and the probability that the translation sequence is *fluent* with respect to the target language. The advantage of such a formulation is that the original translation problem is broken down into two simpler sub-problems, which can be estimated separately, and the fluency and grammaticality of the hypothesis translation can be tackled directly with the use of a language model.

However, estimating these probabilities is not a trivial task. A maximum likelihood approach whereby the probabilities are estimated by calculating relative frequencies of *whole* sentences in a parallel corpus would suffer from data sparsity; probability estimations are likely to be poor and generalisation to new data impossible. Probabilities must therefore be estimated in a way that allows for a greater generalisation capacity, given the finite data available. For the language model, this typically means decomposing the probability into the probabilities of smaller sub-sequences, such as words or n -grams. For the translation model, the translation probability is decomposed into the probabilities of source sub-sequences and their translations. The first approaches to SMT calculated such probabilities at the word level, with independence assumptions concerning the translation of individual words. These approaches were then extended to sequences of words (known as phrases)³ and their translations. They remained state-of-the-art until full-scale neural MT approaches appeared from 2014 onwards.⁴

³Although commonly referred to as *phrases*, these sequences are often statistically rather than linguistically grounded and therefore should not be confused with the term *phrases* in the linguistic sense. Constraining the sequences of tokens to linguistically motivated phrases was found not to improve performance and can even have a negative impact on translation quality (Koehn et al., 2003).

⁴NMT approaches were especially successful for highly resourced language pairs, although today, various approaches have allowed NMT to also perform well on low-resource languages, in many

In this section, we present phrase-based SMT, with a particular focus on MOSES (Koehn et al., 2007), the phrase-based system that will be used in this thesis.⁵ We will begin our discussion with a description of word alignments and their automatic calculation, which were the basis of word-based SMT, but are also an important notion in phrase-based translation (Section 3.1.1). We will then describe phrase-based models themselves (Section 3.1.2), providing details about how the models are trained, tuned and used to decode new sentences. Finally, we shall briefly discuss ways in which domain adaptation techniques can be used to specialise models to new domains (Section 3.1.3).

3.1.1 Word alignments

An important concept in statistical MT is *word alignment*, a representation of the correspondence between source and target words in a parallel sentence, such that aligned words are translations of each other. The simplest case of word alignment is one-to-one alignment, whereby each word is aligned to one and only one word in the other language. However, in natural language, this is by no means systematic; a word can be translated by several words, and multiple words can be translated by one or multiple words, without there being a one-to-one mapping.⁶ This is the case for example of the translation of language-specific fixed expressions and idioms. An illustration of these different types of word alignment is given in Figure 3.2.

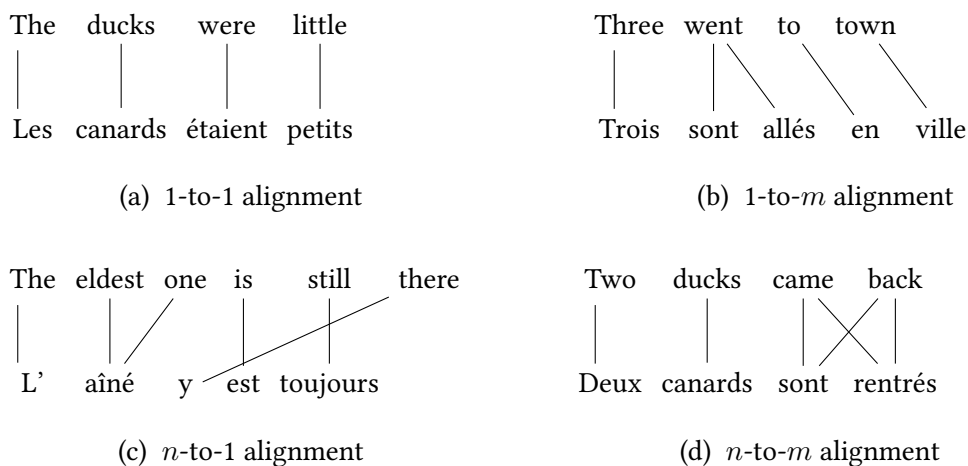


Figure 3.2: Examples of different types of word alignment.

The decomposition of the translation probability $P(s|t)$ into smaller, more reliable probabilities requires having some form of alignment of the words in the source and

cases outperforming SMT models. These approaches include transfer learning from better resourced language pairs (Zoph et al., 2016; Gu et al., 2018), data augmentation (Sennrich et al., 2016c) and unsupervised NMT (Artetxe et al., 2018; Lample et al., 2018).

⁵The neural system that will also be used is presented in Section 3.2.

⁶Sometimes words can even be aligned to no translated word at all. This is common in the case of discourse connectives for example, which can in some cases be translated implicitly (i.e. by nothing).

target sentence. Parallel corpora that have been automatically aligned at the sentence level are readily available (Koehn, 2005; Tiedemann, 2012; Lison and Tiedemann, 2016). However, automatic word alignments between source and target languages are much less trivial to produce, and the corpora are not manually word aligned. The first modern statistical models of the noisy channel framework for MT, produced and implemented in the 1980s and 1990s in the form of the IBM translation models (Brown et al., 1988, 1993) were therefore an important step for MT. These translation models, which were word-based models, brought with them techniques for automatically aligning parallel corpora on the word-level.

Each of the five IBM word-based models calculates the translation probability of a source sentence given a candidate translation based on *lexical translation probabilities*, individual translation probabilities associated with each source word. For example, according to IBM model 1, for a given alignment a between the source sentence s of length n and a candidate target sentence t of length m , the probability of s and a given t is calculated as the product of the lexical translation probabilities:

$$P(s, a|t) = \prod_{i=1}^n P(s_i|t_{a(i)}), \quad (3.4)$$

where $t_{a(i)}$ indicates the target word aligned to source word of index i . By marginalising over $A(s|t)$ all possible word alignments between s and t , the models provided the means of calculating the probability of the source sentence given a candidate translation:

$$P(s|t) = \sum_{a \in A(s,t)} P(s, a|t) \quad (3.5)$$

This importantly means that each source word must be associated with a single target word; the model can only express alignment type n -to-1, excluding alignment types 1-to- m and n -to- m . IBM model 2 introduces the notion of word position into the equation, IBM 3 the notion of *fertility*, modelling how many translation words could be produced for each source word, and IBM models 4 and 5 allow for more flexibility concerning word order differences through reordering by introducing a *distortion* probability. The word alignments themselves are induced from relative frequency counts in a raw parallel corpus and calculated iteratively using the expectation-maximisation algorithm (Dempster et al., 1977). Each of the models initiates its parameters with those obtained by the previous model, and the models are of increasing complexity, although the alignment type n -to- m remains out of reach by all five models.

When used for translation, the IBM models suffer from their strict independence assumptions concerning translation probabilities. The translation probabilities of source words are calculated independently of the other words in the sentence. A language model

that takes into account sequences of target words (see Section 3.1.2.2 for n -gram-based language models) can compensate slightly for this, but not entirely. In cases of ambiguity, for example if the source word has two different meanings that are translated differently in the target language, the most frequent translation is likely to be used, if the context disambiguating the possible translations is not in the very local context.

3.1.2 Phrase-based translation models

The phrase-based approach to MT (Och et al., 1999; Marcu and Wong, 2002; Zens et al., 2002; Koehn et al., 2003, 2007), state-of-the-art from the early 2000s to the revival of neural approaches from 2014 onwards, extends the word-based approach described above. A basic principle of phrase-based approaches is to calculate translation probabilities over short sequences of words rather than individual ones. The idea itself was not new; Warren Weaver had evoked the possibility of calculating probabilities over sequences of two tokens in his memorandum (Weaver, 1955, p. 14). However there were still many aspects left to be defined, both theoretically and experimentally: defining what constitutes a phrase, how phrase translation probabilities were to be calculated, and the integration of other components to further refine the scoring function.

Och et al.'s (1999) *alignment template* model was one of the first approaches to consider modelling sentences at the phrase level, although it still relied on lexical translation probabilities. The idea was to segment the sentence into phrases, which were translated word-by-word and the translated phrases were then reordered to allow for differences in word order between the source and target language. However, soon approaches were being proposed whereby the translation probabilities were defined directly between source and target phrases (Marcu and Wong, 2002; Zens et al., 2002; Koehn et al., 2003). These models not only out-performed word-based translation models, but also benefitted from the availability of open-source community-based toolkits, first PHARAOH (Koehn, 2004a) and then MOSES (Koehn et al., 2007), which, aside from being freely available state-of-the-art tools, undeniably also owe their popularity to their ease of use.⁷ Several other approaches to MT that are not typically grouped under the name *phrase-based* also sought to better model the dependency between tokens. Although they shall not be mentioned in detail in this thesis, syntax-based systems (Yamada and Knight, 2001), including hierarchical MT systems, which use synchronous context-free grammars for translation (Chiang, 2007), also provide a means of ensuring dependency between the translation of tokens, even non-contiguous ones.

A major advantage of frameworks such as MOSES is their flexibility and their ability to easily modify the scoring function for a candidate translation. This flexibility is the result of an earlier work by Och and Ney (2002), who reformulated the translation problem

⁷Among alternative systems we can mention Ncode (Crego et al., 2011)

as a discriminative learning problem (known as the *log-linear framework*), in which the translation probability $P(t|s)$ is modelled directly and expressed as the weighted sum of multiple models:

$$P(t|s) = \frac{\exp \sum_{m \in M} \lambda_m \phi_m(s, t)}{\sum_{t' \in T} \exp \sum_{m \in M} \lambda_m \phi_m(s, t)}, \quad (3.6)$$

where M represents all models, ϕ_m the application of the model as a function of s and t , λ_m the model's weight and T the possible translations of source sentence s . The aim of decoding being to find the t with the highest score, this formula can be simplified by removing the denominator, which is identical for every candidate translation of s :

$$\hat{t} = \operatorname{argmax}_{t \in T} \exp \sum_{m \in M} \lambda_m \phi_m(s, t) \quad (3.7)$$

The original noisy channel approach can be seen as a simplified version of this framework, in which the two models (the language model and the translation model) have an equal weighting. The advantage of the log-linear approach is that new models can be easily added to refine the scoring function and the models are weighted to give differing importance to each component. Once each component has been separately trained, the weights can be optimised as in any statistical linear model (cf. Section 3.1.2.3).

Within the MOSES framework, the scoring function is typically made up of four different models (See Equation 3.8). Each model can give several scores.

1. *Phrase translation model*, which determines the faithfulness of the translation with respect to the source.
2. *Language model*, which determines the fluency of the translation.
3. *Reordering (or distortion) score*, which measures the cost associated with reordering phrases, allowing for local flexibility in the ordering of phrases, particularly useful when the target language's word order differs from that of the source language. The reordering cost is calculated as the number of words skipped between the last word of the previous phrase and the first word of the current phrase, summed over all phrases in the translation.
4. *Word penalty*, which penalises translations that are either too long or too short. This ensures that translations are not under- or over-translated and are on average the correct length compared to the reference translations. The word penalty cost is simply $length(t)$.

$$p(t|s) \propto \exp\left(\underbrace{\lambda_{TM} \text{TM}(s|t)}_{\text{Phrase translation model}} + \underbrace{\lambda_{LM} \text{LM}(t)}_{\text{Language model}} + \underbrace{\lambda_{DM} \text{DM}(s, t)}_{\text{Distortion model}} + \underbrace{\lambda_{WP} \text{WP}(t)}_{\text{Word penalty}} \right) \quad (3.8)$$

The training of the first two models will be described in more detail below. The relative weight of each component is optimised during a *tuning step* (cf. Section 3.1.2.3).

3.1.2.1 Phrase translation model: Translation probabilities

The aim of the phrase translation model, as with the word-based model, is to evaluate the faithfulness of a translation with respect to a source utterance.

A first step in the training of the phrase translation model is to extract all possible phrases and their possible translations from a large parallel training corpus and to estimate the probability of the pair being translations of each other. The process of phrase extraction explicitly uses the principle of word alignment, as described in Section 3.1.1, to detect consecutively aligned sequences of words. One disadvantage of the alignment produced by IBM models is the restriction of alignments to *n*-to-1 alignments. A technique to allow *n*-to-*m* alignments is to perform *symmetrisation* of word-alignments, whereby word alignments are computed in each language direction and are then combined using heuristics (Och et al., 1999). A commonly used strategy, which shall be used in this thesis, is the *grow-diag-final-and* strategy described in (Koehn et al., 2003). To encourage a high recall for alignment,⁸ this strategy proceeds in three steps: (i) it first intersects the word alignments obtained by performing word alignments in both directions, (ii) it then adds alignments that neighbour those in the previous point and which appear in the union of the two alignments, and (iii) it finally adds further alignments if they have not already been added and appear in either of the two alignments.

Once phrases are extracted, translation log probabilities can be estimated through relative frequencies in a large parallel corpus. Equation 3.8 is a simplification, and in reality, the translation model score is not a single score, but is made up of four scores, each assigned a separate weight: phrase translation log probabilities $\log P_t(t|s)$, $\log P_t(s|t)$ and lexical translation log probabilities $\log P_{lex}(t|s)$ and $\log P_{lex}(s|t)$, based on the word-based translation probabilities of the words it contains. Lexical probabilities are particularly important for rare phrases, for which the probability estimations are less reliable. An extract of a phrase table, with each of these four probabilities, is shown in Table 3.1.

⁸A high recall is important for the phrase table to allow a maximum number of possible phrase pairs to be extracted. Noisy phrase pairs are not a problem for the phrase table (see Figure 3.1), because each pair is associated with its probability, so inaccurately extracted pairs will only receive a very small probability. It also allows for a wide coverage of possible translation candidates, increasing the model's cover in terms of vocabulary and its capacity to generalise to unseen sentences.

s	t	$\log P_t(t s)$	$\log P_t(s t)$	$\log P_{lex}(t s)$	$\log P_{lex}(s t)$
I like that .	j' aime ça !	0.25	0.03336	0.00689	7.79e-05
I like that .	j' aime ça , tiens .	1	0.01779	0.00344	2.51e-08
I like that .	je préfère ça .	0.0185	0.01348	0.00344	0.000136
I like that	ça , ça me plaît .	0.1	0.00175	0.00344	2.90e-08
I like that	ça me fait plaisir .	0.0277	0.00021	0.00344	4.10e-09
I like that	j' aime cette idée .	0.25	0.04379	0.00344	6.06e-08

Table 3.1: An extract of a phrase table, with associated translation probabilities.

3.1.2.2 Language models

The role of a language model is to evaluate how likely it is that a sequence of tokens belongs to a particular language. In SMT, the independence assumptions of the translation model mean that the translation probabilities of phrases are calculated independently of the other phrases in the sentence. The direct consequences of this are that translation candidates are not otherwise constrained to being grammatical or stylistically representative of the target language, and, importantly, translation choices are not made in the context of surrounding words, which is problematic for cases of ambiguity. The role of the language model is therefore to assign a score to possible candidates based on the compatibility of the candidate to the target language, taking into account more context than afforded by the translation model.

A widely used approach to language modelling is based on n -grams extracted from a large training corpus. It involves estimating the probability of each consecutive sequence of n tokens, based on relative frequency counts in a large, monolingual training corpus. The probability of a sentence according to the language model is simply the product of the probabilities of each n -gram appearing in the sequence, as illustrated for the following example, with the application of a 2-gram language model:

$$\begin{aligned}
 P(\text{Roses are red .}) &= P(\text{_BEGIN Roses}) \\
 &\quad \times P(\text{Roses are}) \\
 &\quad \times P(\text{are red}) \\
 &\quad \times P(\text{red .}) \\
 &\quad \times P(\text{. _END})
 \end{aligned} \tag{3.9}$$

Note that $n - 1$ tokens indicating the beginning and the end of the sequence are added into the equation, in order to correctly model the probabilities of the first and last words of the sequence.⁹

⁹In practice, to avoid the computational problem of dealing with very small probabilities, log probabilities

To account for words that were not seen during the training of the language model, a *smoothing* parameter is invariably added in the calculation of the probabilities, which assigns a small probability to unseen words or n -grams and adjusts the probabilities of other n -grams accordingly. The most simple strategy is to suppose that each n -gram appears at least once in the training data, and consequently to add one occurrence to all seen n -grams and an occurrence for all as yet unknown n -grams (known as *Laplace* or *add-one* smoothing). Other, more complex strategies can be used, such as linear interpolation or an estimation of the probability of unseen words based on the word classes to which they belong. Other commonly used approaches are Kneser-Ney smoothing (Chen and Goodman, 1996) and Good-Turing smoothing (Good, 1953).

In order for different translation candidates to be compared, the language model probabilities calculated must be comparable. The more tokens added to a sequence of words, the smaller the probability of the sentence becomes. The probabilities must therefore be normalised or compensated by a score relative to the respective length of the sentences.

3.1.2.3 Tuning

As seen above, the log-linear approach transforms the scoring function into a linear combination of different scoring functions, each assigned a weight, allowing components to have variable degrees of importance in the final scoring of a candidate translation. These weights are typically learnt during a step known as *tuning* by iteratively updating the parameters in order to maximise the translation quality on a small, held-out dataset, known as the *development* or *tuning set*. At each iteration, the model with its current parameters is used to decode (i.e. translate) the development set, the quality of the translation is estimated, and the parameters are updated in a way that is dependent on the chosen tuning algorithm.

Since translation quality is typically estimated using automatic evaluation metrics such as BLEU (Papineni et al., 2002), which will be described in more detail in Section 3.3, it is common to use BLEU as an estimation of the quality of translations during tuning. Given that BLEU is not a convex function, its optimisation is not trivial, and specialised optimisers have therefore been proposed.¹⁰ The most widely used optimisers within the Moses framework are Minimum Error Rate Training (MERT) (Och, 2003; Bertoldi et al., 2009), which is one of the first such optimisers and produces state-of-the-art results, and k -best MIRA (Cherry and Foster, 2012), which has been shown to scale better than MERT when more features are used and therefore to be more stable.

are typically used.

¹⁰See (Neubig and Watanabe, 2016) for a survey of optimisation techniques for SMT.

3.1.2.4 Decoding

Equipped with a scoring function for possible candidate translations, an MT system must be able to search among possible translation candidates in an efficient way. This is the task known as *decoding*. It goes without saying that exploring all possible translations of a given source sentence, which amounts to enumerating and scoring all possible sentences in the target language, is not feasible. The number of possible translations is theoretically infinite. Although in practice the number of candidates is limited by a finite vocabulary, a maximum translation length and the fact that translations are guided by the source words, it is still far too large in terms of memory capacity and time constraints to be explored in its entirety.

An approximation invariably made is therefore to limit the search space of possible candidate translations and to calculate the score only for the n -best partial translations at any one time by using a beam search (Tillmann, 2001). The decoder starts at the beginning of the source sentence and progressively advances, generating possible partial translations of the subsequence. These partial translations are scored and only the n -best scored are retained at that step. Once all source segments have been translated, the best scoring full translation \hat{t} is retained as the best candidate translation of the source sentence.

3.1.3 Domain adaptation

A downside of data-driven methods such as SMT (but which also applies to NMT) is the high dependency of the models on the parallel data on which they are trained. To translate a text from a specific domain, for which large quantities of data are not available, a generic model trained on another domain is likely to perform poorly, or at least less well than on the domain on which it was trained. Every aspect of the language in both the source and target sentences (vocabulary, sentence structure, style, etc.) is conditioned on the type of language seen at training time. However, there exist domain adaptation techniques used to exploit small amounts of in-domain data for parameter tuning, to better exploit the large quantities of *out-of-domain* data available, to artificially construct *in-domain* parallel data. This will notably be important for our experiments on gender adaptation in Section 5.1.

If in-domain parallel data is available, even in relatively small quantities, it can be used in various ways in conjunction with large amounts of out-of-domain data to adapt the model. One simple method is to tune a generic model that has been trained on a large out-of-domain parallel corpus using the in-domain data. Pecina et al. (2012) show that a model trained and tuned on parliamentary (EUROPARL) data results in BLEU scores that are much lower (up to -9 BLEU points for French-to-English) than if the same model is tuned on in-

domain medical data. Another strategy is to select “pseudo-in-domain” training sentences from a large out-of-domain corpus that most resemble the in-domain corpus, and to use these as a basis for training, a solution that can be more effective than using larger out-of-domain models (Moore and Lewis, 2010). For machine translation data selection, Moore and Lewis (2010) select sentences with the greatest difference between the cross-entropy obtained with an in-domain language model and with a generic language model trained on randomly selected sentences. The method, applied to only one side of parallel data, was extended by Axelrod et al. (2011) to select sentences based on both languages by summing the cross-entropy difference for the sentence in each language (known as the Modified Moore-Lewis (MML) technique). A third approach using in-domain parallel data is the mixture model approach, whereby separate models are trained on in-domain and out-of-domain data, and the models are then combined. The log-linear approach to SMT is well adapted to this approach, because models can simply be added as new scoring functions, which are assigned a weight during the tuning step or based on a distance metric in order to dynamically adapt the weights to new domains (Foster and Kuhn, 2007).

An alternative strategy is to exploit the availability of in-domain monolingual data, which is often much less rare than parallel corpora. If monolingual target data is available in large quantities, it can be automatically translated to produce a parallel corpus, in which the source sentences are machine translations. This data can then be used as additional parallel training data for translation models (Schwenk, 2008; Lambert et al., 2011; Sennrich et al., 2016c). Despite the noisiness of the data, the technique has proved useful in improving translation quality.

3.1.4 Successes and Limitations of SMT

A lot of progress has been made in SMT since the first implementations relying on word translation probabilities. The ability to model translation via word sequence probabilities means that word choice is performed using a greater degree of context, helping to disambiguate possible translations, encourage fluency and ensure a limited form of cohesion between the words of the translation. Within the log-linear framework, and thanks to the development of open-source software such as MOSES, adding new models and features to score translations is relatively easy. We shall see for example in the following chapter how new models have been designed within this standard framework to provide a scoring function that takes into account discourse context.

In spite of these advances, the gains seen in SMT translation quality quickly plateaued, and the limitations of the models proved difficult to overcome. One of the most visible problems of phrase-based models is a limited capacity to ensure long-distance dependencies within a sentence. They have a limited handling of grammatical agreement, word order differences and a limited context to ensure appropriate lexical choice. Why

do these models suffer from these problems? A number of works have focused on the analysis of translation errors in a bid to understand where the limitations comes from (Turchi et al., 2008; Wisniewski et al., 2010). The general consensus is that the limitations do not come from the fact that the correct translations are absent from the search space; the models are capable in most cases of producing the reference translation given the entries available in the phrase table. The problem is *reaching* the correct translation within the search space, which is a limitation of the current scoring functions available. Turchi et al. (2008) suggest that it is a problem of insufficient data for parameter estimation, in particular for infrequent words. However they also suggest that the quantities of data necessary to override this performance limitation in the current setup would be difficult to obtain. One of the main problems with the current scoring functions is that translation probabilities and language model probabilities are calculated based on the strong independence assumption that the probability of the target sequence can be decomposed into the probabilities of its constituent phrases. The maximum length of phrases within translation models is typically relatively long (up to 7 tokens for certain models), but the probabilities calculated even for phrases of more than four or five tokens are very unlikely to be reliable, due to the use of data of a finite size, therefore resulting in data sparsity. Likewise, n -gram models are typically limited to an n of at most four or five tokens due to insufficient data to calculate reliable probabilities for longer n -grams. The effect of this is that the scores provided by these models are based on local decisions in translation, and long-distance phenomena are therefore often disregarded by the scoring functions. The problem is particularly salient in morphologically rich languages, where a higher type/token ratio increases data sparsity.¹¹ However, it is also very relevant to the context-dependent phenomena presented in the previous chapter and which are the focus of this thesis.

Various strategies have been designed to overcome data sparsity and an inability to account for long-distance context. Factored translation models (Koehn and Hoang, 2007) were introduced, allowing source and target sentences to be decomposed into tiers of linguistic annotation levels, such as words, lemmas and PoS tags. The motivation behind this decomposition is to counter data sparsity by abstracting away from inflected forms and decomposing translation probabilities across each tier. However, the models proved computationally intractable, because of the increased decoding complexity caused by the multiplication of the number of mapping steps and phrase tables. In terms of dealing with long-distance phenomena, one of the strategies that has proved effective is to include more scoring functions that can provide better expressivity even at longer distances. The use of a neural language model as an extra component in phrase-based MT systems provided an opportunity to significantly improve translation quality (Schwenk et al., 2006; Hai Son et al., 2012; Vaswani et al., 2013). With a higher degree of expressivity

¹¹The type/token ratio is the number of different forms divided by the total number of forms. This is likely to increase in morphologically rich languages, as translation probabilities are spread out amongst the different inflected forms.

thanks to the continuous representation of words and a capacity to better exploit context within the sentence, neural language models provided extra information that the discrete n -gram language models could not. An increasing use of neural models within the phrase-based architecture eventually led to advances in pure neural MT systems (aided by computational advances), with the SMT paradigm being almost entirely replaced by a new one, NMT, which we shall now describe in the upcoming section.

3.2 Neural Machine Translation (NMT)

Since the mid-2010s, another translation paradigm, neural machine translation (NMT), has overtaken phrase-based SMT approaches to translation, and will also be used in this thesis. As in numerous NLP domains, for example part-of-speech (PoS) tagging (Plank et al., 2016), named entity recognition, semantic role labelling (cf. Collobert and Weston, 2008 for all three tasks) and syntactic parsing (Socher et al., 2013), neural networks have been successfully exploited in MT to give state-of-the-art performances. Their use presents a paradigm shift within the domain, with one major differentiating factor being the way basic textual units are represented. The use of continuous word representations, underpinning the use of neural networks for NLP, facilitates the combination of words, going some way to constructing a unified semantic representation of the input sentence and to some extent breaking the independence assumptions of phrase-based MT.¹²

In this section, we shall describe the neural approaches to MT, and in particular the sequence-to-sequence model with attention that is used and referenced in this thesis. We also aim to lay out some of the differences between this NMT architecture and the phrase-based architecture presented in the previous section, highlighting in particular the advantages offered by the neural approach, the new problems introduced, and the phenomena that remain unresolved.

3.2.1 Neural networks for NLP

3.2.1.1 Word representations

A major difference in the use of neural networks compared to the phrase-based approach is the representation of the basic units of a text. It is therefore worth discussing these word representations and more specifically the role that *word embedding* has to play in the use of neural networks for text processing. *Word embedding* refers to a variety of techniques used to map words to compact, continuous vectors (i.e. to *embed* them in a

¹²As shall be described in the current section, neither continuous representations nor neural networks are new to NLP, but have recently become more popular, leading to state-of-the-art implementations, thanks to advances in computational technology.

continuous space). This representation strategy, which has rapidly grown in popularity over the past decade, following research by Bengio et al. (2003), Schwenk (2007) and Collobert and Weston (2008), lies in opposition to the traditional strategy, at least in MT, of representing words as discrete, atomic units. Using these representations allows for greater model expressivity and a better potential to capture semantic similarity than with atomic units, making them particularly interesting for MT.

Traditionally in MT, words have been treated as discrete, atomic units. This is the case for example in n -gram language models and in the traditional phrase-based MT models introduced in the previous section. Such approaches are simple and relatively effective when trained on large amounts of data. However, a disadvantage of treating words as discrete units is an inability to model and exploit the semantic similarity between words, limiting the model's capacity to generalise and to handle rare words. These limitations can be overcome by using continuous word representations, which provide the necessary framework for a more fine-grained, semantic representation of words. Each word in the vocabulary is represented by a vector $w \in \mathbb{R}^D$, where D is the dimension fixed in advance. One of the major advantages of representing words as vectors is the fact that standard similarity measures such as cosine similarity or Euclidean distance can be used, enabling semantic distances to be calculated between words.

Contrary to what we may be led to think by the recent popularity surge for word embeddings, the use of compact, vectorial word representations is by no means new, and the theoretical underpinnings can be traced back at least to the 1950s and the theory of distributional semantics. The distributional hypothesis, the idea that you can define a “word by the company it keeps” (Harris, 1954), popularised in the 1950s by philosophers and linguists such as Harris (1954), Firth (1957) and Wittgenstein (1953), has been influential in the way textual input is represented in NLP and is the basis of many of the vectorial word representation strategies. The approach is to define a word's meaning, represented by the different values of the vector, based on the words appearing in the same context, rather than on the word's intrinsic properties. Words that appear in similar contexts are therefore expected to be semantically similar.

There are two classes of strategy for computing continuous word representations: *count-based* methods and *prediction-based* methods. *Count-based methods*, such as LSA (Deerwester et al., 1990), LDA (Blei et al., 2003) and HAL (Lund et al., 1995), were the earliest used methods for the calculation of vectorial representations. They are based on the use of word counts, usually weighted,¹³ as proxies for semantic feature representations. Importantly, they are also dimensionality reduction techniques, designed to produce compact representations, in which new latent features are induced from the original word counts. The new representations are more compact, making them

¹³Weights, such as tf-idf, are often applied to raw word counts to counterbalance the difference in frequency between words.

computationally more efficient for many learning algorithms, and denser in terms of the distribution of semantic information. *Prediction-based methods*, made popular in NLP by tools such as word2vec (Mikolov et al., 2013), learn distributed word representations, typically in a neural network framework, by learning to maximise the probability of the word’s context. These methods were commonly used in neural machine learning problems (cf. early work by Baldi and Hornik (1989)). They were first introduced in NLP much later by Bengio et al. (2003) in the context of neural, probabilistic language modelling and successfully applied to a number of NLP tasks, including PoS tagging, chunking, named-entity recognition, language model learning and semantic role labelling by Collobert and Weston (2008). These jointly learnt embeddings have been found to outperform classic count-based methods, as discussed in Baroni et al.’s (2014) paper aptly entitled “Don’t count, predict!”, although the more recent count-based method GloVe (Pennington et al., 2014) rivals the performance of prediction-based implementations such as word2vec.¹⁴

If the theoretical motivation and even certain methods for computing word embeddings have existed for many decades, what is the reason for the recent uptake in the use of word embeddings over the past decade? The main trigger is the generalised use of neural networks throughout NLP, for which embeddings present the ideal representation for input words. Moreover, the success of neural networks in the field of NLP has made prediction-based methods (learnt within a neural network) more popular, and this thanks in particular to the availability of large quantities of data and computational power, which meant neural networks, which had existed theoretically for decades, could finally be exploited in high-resource settings in NLP.

3.2.1.2 Neural networks

Before describing how neural networks can be used in the context of translation, we shall briefly provide some basic notions of neural networks. An artificial neural network is a type of machine learning architecture, in which a numerical, vectorised input x is mapped via a parametrised function f to a numerical, vectorised output y :

$$\hat{y} = f(x, \theta), \quad (3.10)$$

where \hat{y} is the predicted output produced by the function f with model parameters θ .

The basic unit of computation in a neural network is the *neurone* (shown in Figure 3.3), which receives inputs and produces outputs. The inputs to a neurone are weighted by

¹⁴In fact Levy and Goldberg (2014) showed that Mikolov et al.’s skip-gram with negative-sampling word embedding method is formally equivalent to a standard count-based method and produces similar results when computed in similar technical settings.

learned parameters, here represented by weights w_i for each input value x_i , which are linearly combined and to which a non-linear function, represented by the symbol f , is applied. For simplification purposes, in our graphical representations of neurones from here on, the bias value (shown in Figure 3.3) will be implicit.

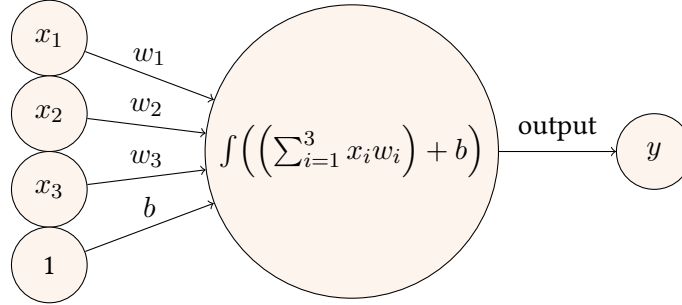


Figure 3.3: A single neurone with three inputs. f represents a non-linear function.

One of the simplest types of neural network is the *feed-forward neural network*, in which multiple neurones are arranged in layers, and transformations are successively applied. Each neurone receives weighted inputs from the neurones in the previous layer. An example of a simple feed-forward network, in which all neurones are *fully connected*, is shown in Figure 3.4, and of which the function can be summarised as follows:

$$\hat{y} = \sigma(W_a h_2 + b_3) \quad (3.11)$$

$$h_2 = f_2(W_b h_1 + b_2) \quad (3.12)$$

$$h_1 = f_1(W_c x + b_1), \quad (3.13)$$

where x is the input vector, \hat{y} the output vector, h_1 and h_2 the first and second hidden layers respectively, f_1 and f_2 are non-linear functions and all W 's and b 's are learned parameters (weights). σ represents a softmax function applied in the final output layer of the network, which is defined in Equation 3.14.

$$\sigma(z)_j = \frac{e^{(z_j)}}{\sum_{k=1}^K e^{(z_k)}} \quad (3.14)$$

The parameters of the model are learnt during training. We work in a supervised learning paradigm, in which we have access to the real outputs (the gold labels) y that we wish to reproduce. A loss function L is used to evaluate how close the predicted output \hat{y} , obtained by $f(x, \theta)$, is to the gold output. The parameters of the model are set to try to minimise (as much as possible) the loss over the training examples:

$$\hat{\theta} = \operatorname{argmin}_{\theta} \frac{1}{N} \sum_{i=1}^N L(f(x_i, \theta), y_i) \quad (3.15)$$

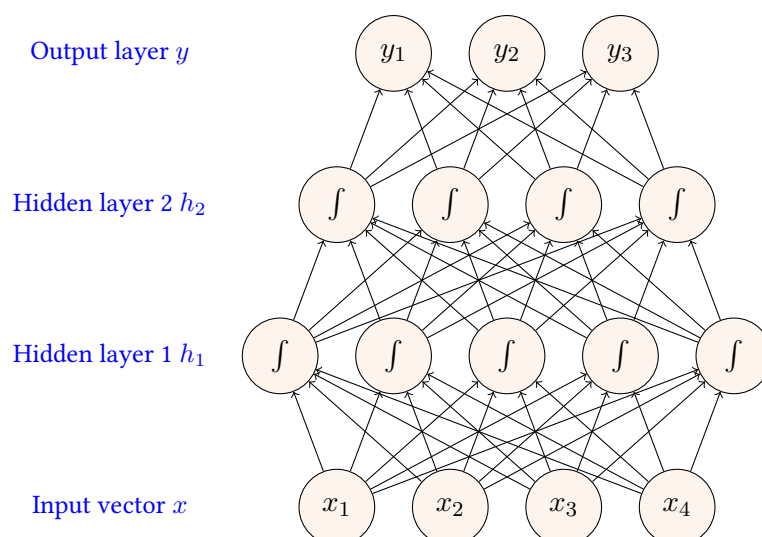


Figure 3.4: A feed-forward neural network with an input layer of dimension 4, a first hidden layer with 5 fully connected neurones, a second layer with 4 fully connected neurones and an output layer of dimension 3.

A commonly used learning technique for training neural networks is gradient-based optimisation, which involves iteratively updating the parameters θ by repeatedly calculating the loss estimation over the training examples, calculating the gradients of the parameters and modifying the parameters in the opposite direction to those gradients. Optimisation is stopped when a minimum is found for the function, or when another stopping criterion has been met (e.g. maximum number of updates reached). The function to be optimised is very often non-convex, making the possibility of reaching a local minimum (as opposed to a global one) high.¹⁵ In neural network architectures, gradients can be computed using a back-propagation algorithm (Rumelhart et al., 1986; Le Cun, 1988), designed to automatically calculate the distribution of loss over the different neurones of the network and to calculate the derivatives of the different parameters.

Embeddings in neural networks Embeddings can often be learnt jointly within a model trained for a specific NLP task, such that the word representations are adapted to the task at hand. This is certainly the case of most neural architectures for MT, for which the representation of input words is intuitively specific to the information necessary to translate into the target language.¹⁶ The strategy usually used in NMT is to learn the word representations within the neural framework as the first layer of the network for the source language words and as the final layer of the network for the target language words, as will be illustrated in Section 3.2.2. The representation of each input word to the NMT

¹⁵Initialisation of parameters is therefore important. It is common for a same experiment to be run multiple times with random initial parameters in order to select the best result of the multiple runs.

¹⁶Recent work on the learning of morphology by NMT shows that the amount of morphology learnt depends on the source and target languages (Belinkov et al., 2017; Vania et al., 2018).

system is a discrete representation, a so-called *one-hot vector*, of which the dimension is the size of the source vocabulary and all values are zero, except at the index of the word represented by the vector. This one-hot representation is mapped in the first layer of the network to an embedding layer of a dimension far inferior to the size of the input vocabulary, typically in the order of several hundred values, thus producing a compact word representation.

3.2.1.3 Recurrent neural networks

MT, as in other NLP tasks involving sequential data (e.g. PoS tagging and syntactic parsing), involves processing sequences of arbitrary length, i.e. producing a sequence of outputs rather than a single output. The feed-forward neural network described in the previous section, which is designed to predict a single output per input, is therefore not well suited to dealing with this problem; for sequential data this would mean making separate local decisions for each element of the sequence. Moreover, sequential prediction is known to benefit from optimising the prediction of the entire sequence rather than making local decisions, since natural language is highly contextual. For sequential text processing, the *recurrent neural network* (RNN) is a natural choice.

Introduced in the 1980s (Rumelhart et al., 1986), RNNs are explicitly designed to process sequences of arbitrary length using fixed-length vectors. Instead of there being a single input vector to the network, which is successively processed by the layers of the network, the input to an RNN is sequential. Input vectors are iteratively fed into the network over a series of timesteps, one after the other. Although the input vector is different at each timestep, the parameters used are shared over timesteps and the hidden state of the network is recomputed with each new input. Importantly, the RNN is characterised by the presence of direct cycles between units, connecting the unit's state with the state at the previous timestep, which means that information from the previous inputs can be used for later computations. Figure 3.5 provides a visualisation of this for an input sequence of length n . The computational graph is provided in its factorised form (left) and unfolded in time (right). This can also be represented in terms of the application of a recursive function:

$$h^{(i)} = \phi(x^{(i)}, h^{(i-1)}, \theta), \quad (3.16)$$

at timestep i , where the hidden state $h^{(i)}$ is a function of the input vector $x^{(i)}$ and the hidden state of the previous timestep $h^{(i-1)}$. The function ϕ in its simplest form is a linear transformation of $x^{(i)}$ and $h^{(i-1)}$, of which the parameters are shared for all timesteps, followed by a non-linear transformation. At each timestep, an output $y^{(i)}$ is produced. For a multi-class classification problem, the length of the output vector is equal to the number of possible classes. A softmax function σ is applied before the output layer, providing an

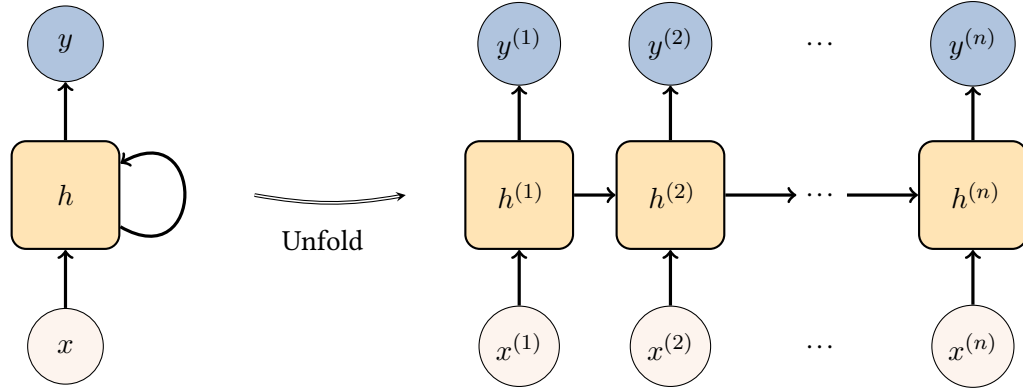


Figure 3.5: An RNN cell and its unfolded version for a sequence of n input vectors $x^{(1)}$ to $x^{(n)}$.

output vector in which the distribution of scores is similar to a probability distribution over the different possible classes:

$$y = \sigma(\phi(h^{(i)}, y^{(i)})) \quad (3.17)$$

The chosen class is therefore the index of y with the highest value.

The RNN's hidden state acts as a memory state, preserving information from one timestep to the next. The final state vector $h^{(t)}$ can potentially encode information from the entire input sequence. This ability to take into account the dependency of the different sequence positions is key to the success of RNNs for NLP tasks. In practice, the capacity of an RNN with a simple unit as described above is limited, in particular for longer sequences. More complex units have been proposed, with the aim of enhancing the RNN's expressivity and allowing longer distance dependencies. Some of the most successful ones are the Long Short-term Memory unit (LSTM), proposed by Hochreiter and Schmidhuber (1997) and the Gated Recurrent Unit (GRU) proposed by Cho et al. (2014), of which we shall not describe the details in this thesis.

In such a setup, each hidden state $h^{(i)}$ is calculated based on both the input $x^{(i)}$ and on the previous hidden state $h^{(i-1)}$, such that it is in fact influenced indirectly through its history by all previous timesteps $h^{(1)} \dots h^{(i-1)}$. However, it can be useful to also use information from the time steps that follow, in order to use the right context to make predictions. A solution is provided in what is known as the bi-directional RNN (bi-RNN), which consists of two RNNs, one used to encode the sequence from left to right (known as the forwards RNN) and the other to encode the sequence from right to left (known as the backwards RNN). At any position i , the forward state represents the sub-sequence $s_{\rightarrow}^{(1..i)}$ and the backward sequence the sub-sequence $s_{\leftarrow}^{(i..n)}$, with the result that together the two states represent the entire sequence $s^{(1..n)}$, but with a particular focus on position i . We call *annotation vectors* the concatenation of these states at each step of the encoder, which we refer to as $h^{(i)}$, thus redefining the term from our earlier definition. An illustration is given in Figure 3.6.

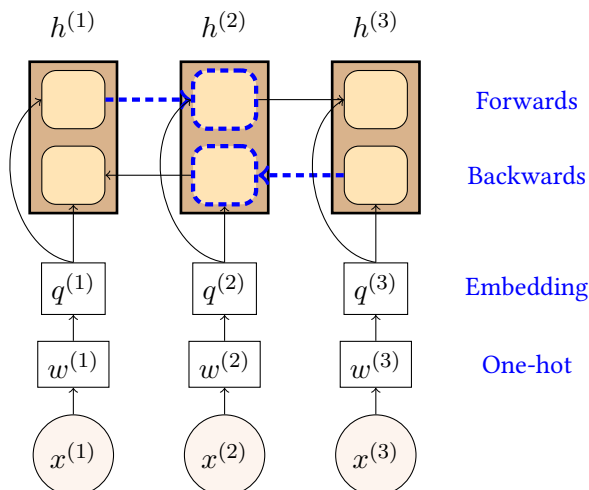


Figure 3.6: Bi-RNN encoder, in which the forward and backward states $s_{\rightarrow}^{(1..2)}$ and $s_{\leftarrow}^{(2..3)}$ at position 2 summarise the entire sequence $x^{(1..3)}$. The annotation vectors $h^{(i)}$ are the concatenation of the two states.

3.2.2 Sequence-to-sequence NMT

A major challenge for the early application of neural networks to MT was the fact that early RNNs require input and output sizes to be fixed in advance, with one output per input element in the sequence, as is the case for PoS tagging. For translation, in which the length of a source sentence is not guaranteed to be the same length as its translation, this is an unrealistic scenario. Sutskever et al. (2014) overcame this difficulty by using a two-tiered RNN, an *encoder-decoder* framework:¹⁷ the first tier recurrently *encodes* the source sequence into a fixed-size vector, and the second tier uses this input representation to recurrently *decode* the target sentence, which is only output once the entire sequence has been encoded (rather than at every encoding step as in sequence labelling tasks). From now on, we shall use the notations s and t for *source* and *target* instead of the generic symbols x and y used in our previous general description of neural networks.

An illustration of the framework is given in Figure 3.7, which is to be read from the bottom upwards, translating from French into English. As previously discussed, the individual words are first mapped to one-hot vectors $w^{(i)}$, which are subsequently mapped to learned dense word embeddings $q^{(i)}$, which are given as input to the encoder. The encoder's role is to compute a fixed-dimensional representation summarising the entire variable-length input sequence $s^{(1)}, \dots, s^{(n)}$. The encoder is an RNN, of which each hidden state $h^{(i)}$ is computed recursively as follows:

$$h^{(i)} = \phi(q^{(i)}, h^{(i-1)}), \quad (3.18)$$

¹⁷The idea of using an encoder-decoder framework was not new: Allen (1987) proposed a small neural encoder-decoder for natural language translation, and in other tasks, Pollack (1990) used the notion of *compressor* and *reconstructor* to encode and decode variable symbolic sequences and tree structures.

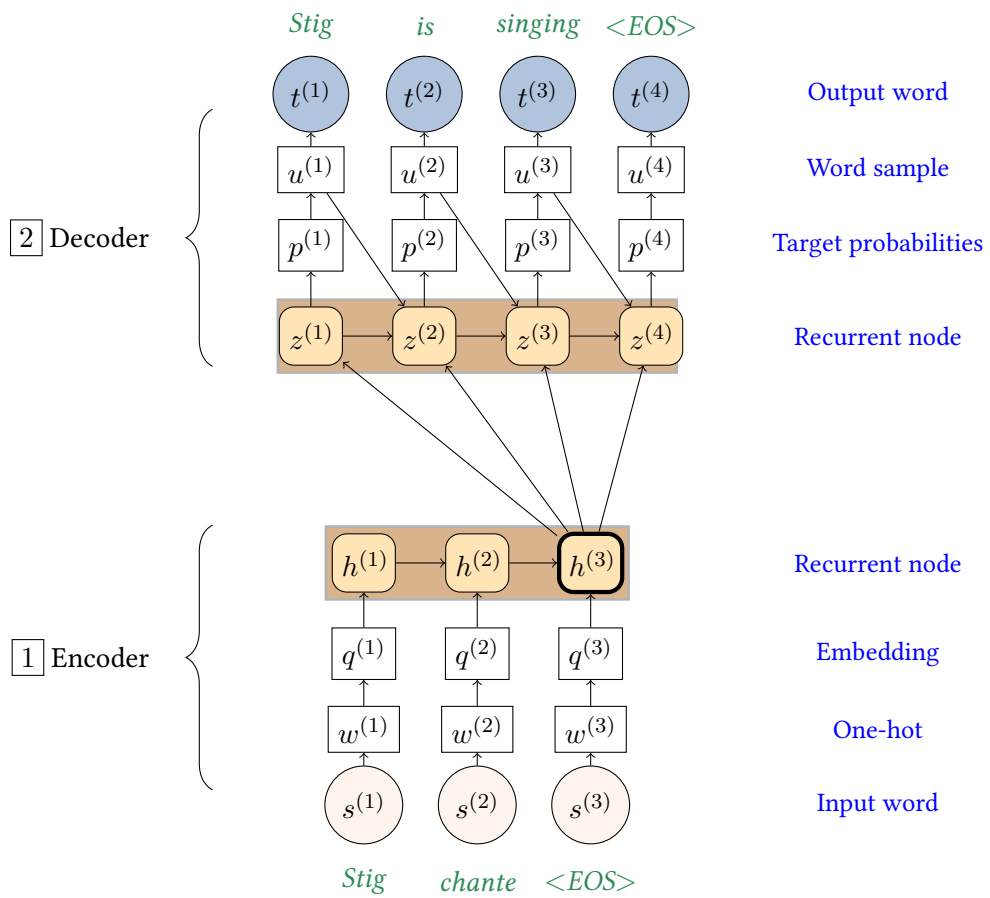


Figure 3.7: Sequence-to-sequence NMT (encoder-decoder framework).

where $h^{(i-1)}$ is the state of the hidden cell at the previous timestep, as described in Section 3.2.1.3. Each hidden state $h^{(i)}$ encodes the input sequence from the first word up to position i . The final node $h^{(n)}$ (marked in bold in Figure 3.7) therefore contains a dense representation of the entire input sequence, which we will refer to as the *context* (or *summary*) vector. Intuitively, the context vector should encode the semantics of the input sequence, specific to the chosen language pair.¹⁸ This is confirmed by Sutskever et al.’s (2014) visualisation of such contexts vectors using principal component analysis (PCA) to project the vectors to a 2-dimensional space (Figure 3.8). The plot shows that semantically similar source utterances indeed appear closer in the vector space than semantically dissimilar ones.

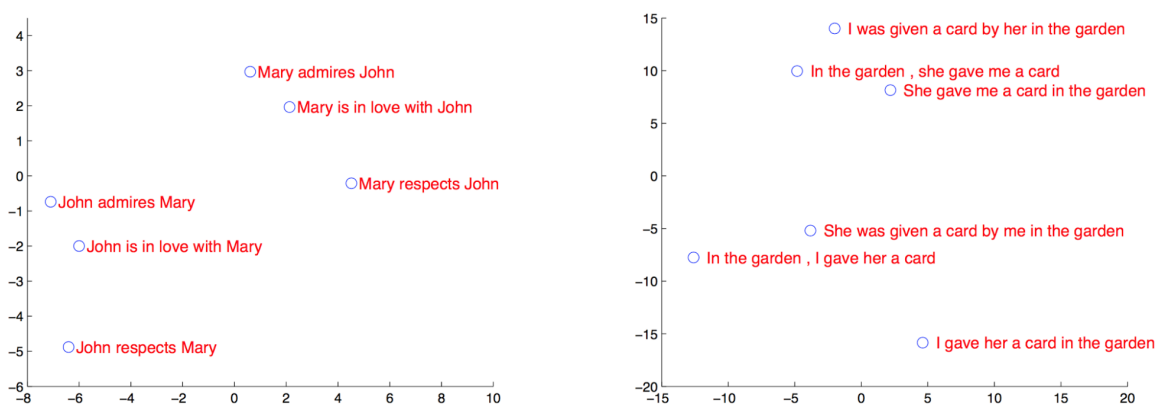


Figure 3.8: Visualisation of context vectors using PCA, in which sentences that are semantically similar are closer in the vector space. From (Sutskever et al., 2014).

The role of the decoder is to take this summary representation of the input sequence and produce the most probable output sequence. The probability of an output sequence $t^{(1)}, \dots, t^{(m)}$ is the product of word translation probabilities:

$$p(t|s) = \prod_{j=1}^m p(t^{(j)}|s, t^{(1..j)}),$$

Like the encoder, the decoder is an RNN, and it produces the translation sequentially, word by word. Each recurrent state $z^{(j)}$ is computed from the summary representation $h^{(n)}$, the previous decoder state $z^{(j-1)}$ and the word sample vector of the previous target word $u^{(j-1)}$:

$$z^{(j)} = \phi(h_n, u^{(j-1)}, z^{(j-1)})$$

¹⁸The way two language systems differ from each other depends on the language pair. A fixed representation of the input sequence should ideally encode an efficient semantic representation containing useful information for translation and this information is likely to differ according to the language pair, depending on how the two language systems are similar or different. A universal cross-lingual semantic representation of the input sequence, the long-sought-after *interlingua*, is a long-term goal of MT and has been the subject of a number of recent publications (see for instance Johnson et al., 2017). However at present, a language-specific representation performs systematically better in a high-resource setting.

At each step, the decoder outputs a vector of the dimension of the target vocabulary, with each position representing the score assigned to each word of the vocabulary. These scores are then probabilised using a softmax transformation to produce a vector of word probabilities $p^{(j)}$ over the target vocabulary.

$$e(v) = W_v^\top z^{(j)} + b_v \quad (3.19)$$

$$p(t^{(j)} = v | t^{(1..j-1)}, h_n) = \frac{\exp(e(v))}{\sum_{v' \in Y} \exp(e(v'))}, \quad (3.20)$$

The next word is selected by sampling from this probability distribution $u^{(j)}$ representing a one-hot vector of the sampled word, mapped to its textual form $t^{(j)}$. This decoding step is repeated, each time outputting a new translated word, until the end of sequence token $\langle \text{EOS} \rangle$ is produced.

3.2.3 Sequence-to-sequence NMT with attention

The major problem with the basic sequence-to-sequence model was the reliance on a fixed-length, static representation of the input sequence for all decoding steps.

This has several implications. Firstly, the translation quality was unsurprisingly shown to decrease as the source sentence length increased (Bahdanau et al., 2015); compressing a longer sequence into a fixed-length vector leads to a greater loss of information. Secondly, the use of the same context vector for each decoding step is suboptimal because the representation must represent the entire sequence, even though for a given decoding step, certain source words are more useful than others.

A solution was found in an alignment technique, inspired by a similar but more restricted method used for handwriting generation (Graves, 2013). Word alignment has always played a central role in MT and therefore the motivation behind the alignment technique is not surprising: at a given point in the translation process, certain input words are more important than others to select the translation of the next word. The technique, referred to as an *attention mechanism*, was first successfully applied to sequence-to-sequence translation models by Bahdanau et al. (2015), resulting in the first state-of-the-art NMT models to outperform phrase-based systems.

The attention mechanism is designed to assign weights to each of the annotation vectors $h^{(i)}$ produced by the RNN encoder. The weights are then used to calculate a weighted average of annotation vectors to produce a context vector $c^{(j)}$, representing the input sequence and specific to the decoding time step j . The attention mechanism is a simple neural network, which, for each decoding step j and for each position i of the input sequence, calculates an energy score $e^{(ij)}$ based on the previous decoder state $z^{(j-1)}$

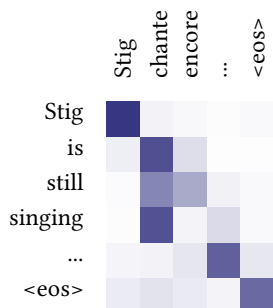


Figure 3.9: Visualisation of the attention weights $\alpha^{(ij)}$ for French to English translation.

and the annotation vector $h^{(i)}$. These scores are normalised to produce alpha weights $\alpha^{(ij)}$ representing a probability distribution over the annotation vectors $h^{(i)}$. These alpha weights are then used in the calculation of $c^{(j)}$, a weighted average of the annotation vectors:

$$e^{(ij)} = \tanh(W_e^\top z^{(j-1)} + U_e h^{(i)}) \quad (3.21)$$

$$\alpha^{(ij)} = \frac{\exp(e^{(ij)})}{\sum_{k=1}^N \exp(e^{(kj)})} \quad (3.22)$$

$$c^{(j)} = \sum_{i=1}^N \alpha^{(ij)} h^{(i)} \quad (3.23)$$

The entire schema for the sequence-to-sequence model with attention is shown in Figure 3.10 for the third decoding timestep.

This has the advantage of calculating more pertinent representations of the input sequence, specific to each decoding step, and remedies the performance drop previously seen for longer input sequences. It is a soft alignment technique, which predicts a probability distribution over the input sequence at each decoding step. A by-product of the strategy is that a soft alignment is automatically learnt between each decoded word and the sequence of inputs words. When the α weights are visualised in a matrix such as Figure 3.9, the alignment can even correspond to our intuitions about word alignment in translation, with higher weights for source words that are more likely to be the translation of or are useful for the translation of the target word. The weights are therefore sometimes used as proxies for word alignment probabilities. In reality, these attention weights do not always correspond to what would be expected from an alignment model; Koehn and Knowles (2017) show that attention weights can sometimes be concentrated on neighbouring words. Note that this is not the case in Figure 3.9.

3.2.4 Recent advances in NMT

New techniques and architectures are continually being developed for NMT, and although we shall not use all of these techniques in our own contributions in this thesis, it is worth

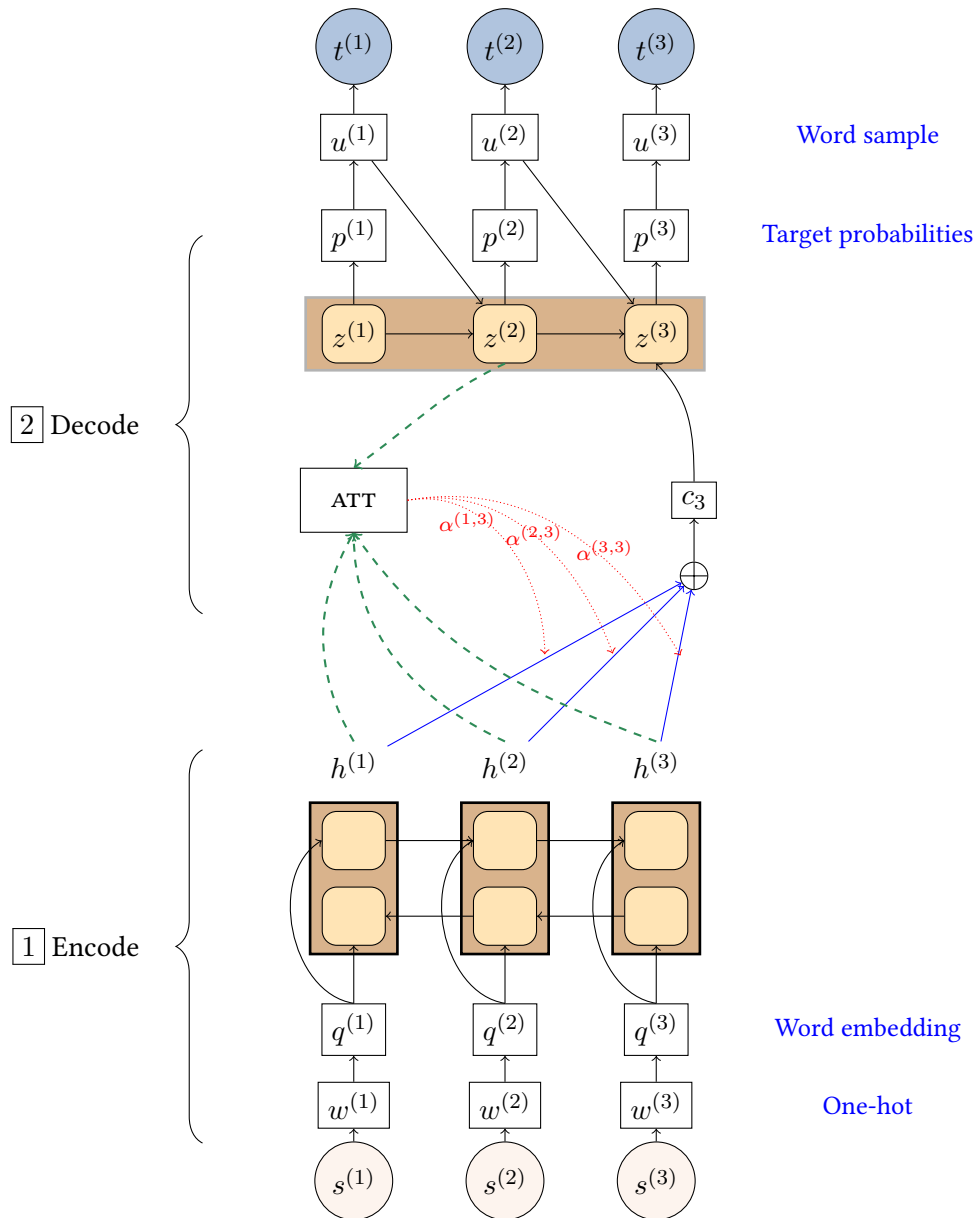


Figure 3.10: An illustration of the sequence-to-sequence neural MT architecture with attention at timestep 3 in the decoding process.

mentioning their existence. We choose to mention two recent developments: character-level NMT and the attention-based transformer model.

Character-level NMT One of the major problems with both SMT and NMT approaches has been the translation of words that do not appear in the training data. So called *out-of-vocabulary* words are either typically not translated or are translated as they appear in the source sentence. This solution is reasonable for certain named entities, which can be translated using the same word. However, this is not the ideal solution for words that should be translated using a word specific to the target language and were simply absent from the training data. Augmenting the amount of training data used is one solution to the problem, although it will never solve the problem entirely. The Zipfian distribution of words in a language makes it practically impossible for a (finite) MT vocabulary to cover all words that you may wish to translate. Moreover, increasing the vocabulary size linearly increases the complexity of training MT models and of decoding. The problem is especially apparent when translating into morphologically rich languages, as the type/token ratio is higher (the different morphological variants of the same lemma are encoded as separate items), a larger vocabulary is needed.

In the previous section, we introduced the notion of subword units, which are the result of segmenting words into smaller units prior to translation in a bid to increase the generalisation capacity of translation models (Sennrich et al., 2016d). The technique enables the vocabulary coverage to be wider, due to the fact that shorter sequences are more likely to be represented in the data, resulting in improved translation performances. Character-level NMT takes this principle further by supposing that instead of representing sentences as sequences of words, they can be represented as sequences of characters. Various approaches have shown that it is possible to learn to translate at this level. Luong and Manning (2016) adopt a hybrid strategy by using character-based translation for rare words only. They find that this strategy outperforms a pure word-based strategy and is capable of producing well-formed words in a morphologically rich setting. Other authors have gone further by proposing purely character-based strategies to NMT. Costa-Jussà and Fonollosa (2016) and Lee et al. (2017) both rely on convolutional neural network encoders to encode sequences of character embeddings. Whereas Costa-Jussà and Fonollosa (2016) still preserve word boundaries, and predict on a word-by-word basis, Lee et al. (2017) adopt a fully character-based approach, whereby no preliminary segmentation is performed. The systems achieve comparable results to those trained on words and subword units. The results are encouraging, and suggest that a greater generalisation capacity can be achieved through these models. However, challenges still remain, notably concerning sentence-level grammaticality, which appears to suffer somewhat in character-based models compared to those relying on larger translation units such as words or subwords (Sennrich, 2017).

Transformer (self-attention) NMT model Vaswani et al. (2017) propose an alternative to the recurrent encoder-decoder architecture. Their architecture relies on stacks of self-attention, rather than recurrent units, meaning that training can be better parallelised and training is therefore faster. Replacing recurrence completely is also beneficial in terms of modelling; since sequentiality is no longer encoded in the structure of the architecture, it enables attention to handle dependencies between words regardless of their distance in the source or target sentence. Without the use of recurrent neural networks, the model has to encode word positions explicitly, since a purely attention-driven model would otherwise be invariant to the order of words in the sequence. They encode positions by providing positional embeddings that are added to input embeddings, rather than making sequentiality an inherent part of the architecture.

The architecture is the new state of the art in MT and has been shown to achieve higher BLEU scores for a number of language pairs (Vaswani et al., 2017; Bojar et al., 2018). The architecture is too recent to be included in experiments within this thesis, but is certainly an option to be considered for future work. We shall mention this architecture in the perspectives to our experiments in Section 7.1, particularly for latent anaphora resolution during training.

3.2.5 Successes and limitations

Full-scale (implementable) NMT has undoubtedly led to huge improvements in translation quality compared to SMT, and this applies across a wide variety of translation quality criteria. As shown by a number of different authors (Bentivogli et al., 2016; Isabelle et al., 2017), NMT provides a higher level of grammatical agreement, particularly in long-distance phenomena, is more fluent and natural than SMT, displays better lexical choice and a better handling of word reordering.

This does not mean however that MT has been solved. There may be a better handling of many intra-sentential phenomena (Bentivogli et al., 2016), but some problems remain unsolved, and new ones have arisen. In both SMT and NMT paradigms, the vocabulary of the models is determined by data seen at training time, and words unseen during training (*OOVs* or *out-of-vocabulary* words) pose a problem for translation, because they cannot be translated. A common strategy is simply to translate unknown words using their corresponding source words, which in many cases is a reasonable strategy, in particular for proper nouns. A strategy frequently used in NMT to decrease the number of OOVs is to first segment tokens into smaller subword units based on the frequency of the subwords within the text. One such technique, which is commonly used, is BPE (for *byte-pair encoding*) (Sennrich et al., 2016d). It enables part of source words to be recognised by their subword units even if the whole word is not in the vocabulary and new target words to be constructed from several translated subwords. Whilst helping the problem, OOVs remain

computationally expensive, NMT models rely on this assumption to speed up training (by processing sentences in batches sorted by length) and to avoid having to translate segments of text that are too long (which would result in reduced translation quality). Therefore, despite the progress made in the modelling of context within the sentence, standard sentence-level NMT architectures suffer from the same limitation faced by the SMT models presented at the beginning of this chapter: a blindness to extra-sentential context, necessary for correctly translating context-dependent phenomena such as those discussed in Chapter 2. These same criticisms can also be found for the majority of evaluation metrics standardly used in the field, which also tend to overlook context-dependent phenomena, which in turn influence the directions taken by the MT research community. We shall take a look at these standard evaluation metrics in the next section, before discussing in Chapter 4 ways in which we *can* take into account context, in both MT and its evaluation.

3.3 Evaluating Machine Translation

Progress in MT relies on having a reliable way of measuring the quality of a translation. Given two different MT systems, particularly where one system presents a novel aspect over the other baseline system, it is important to have a way of measuring which one produces “better” translations. The standard way of testing this is to translate a *test set* of sentences with each system and to compare the two sets of outputs. If human reference translations are also provided, evaluation can be based on a comparison between the MT outputs and the human translations, and if not must be based on the source sentences alone. The comparison can be performed either manually (by human experts) or automatically, with both evaluation types complex to perform. Evaluation remains the Achilles heel of MT and a difficult task to perform automatically. To date, the most reliable technique for judging translation quality remains manual evaluation by human evaluators, the inevitable downside of which is that it is time-consuming and costly. Humans also tend to be subjective; each evaluator will have different attitudes to different types of error, and it therefore becomes difficult to compare evaluation scores across the literature, without redoing human evaluation for each new MT model trained. The development of automatic metrics has been instrumental in the development of MT architectures, as they enable MT outputs to be regularly compared at little cost and provide a deterministic and therefore reproducible way of evaluating translations. However, developing automatic metrics that mimic human quality evaluation is extremely complex due to the subtleties of natural language, and current metrics are far from being able to match the fidelity of human judgments, as we shall see below. Current automatic metrics for the global evaluation of translation quality are also inadequate in terms of evaluating contextual phenomena such as those

we will review in this thesis, because they are not designed to focus on particular aspects of translation quality. This will be the focus of Section 4.1.

In this section, we will first briefly present the challenges presented by the evaluation of translation (Section 3.3.1) to give a general view of the difficulty of the task. We shall then present automatic evaluation approaches, with a particular focus on the BLEU metric, which has been particularly influential for the domain of MT (Section 3.3.2). This section will also be a critique of standard automatic evaluation metrics including BLEU, which have been both a help and a hindrance in the domain. The issues in this section will provide the first step of a review of evaluation strategies in MT, which will be continued in the following chapter (Chapter 4) on the more specific topic of contextual information in MT, in which we will evoke alternative evaluation strategies.

3.3.1 Issues in human evaluation of MT quality

Evaluating translation quality is notoriously difficult, even for human evaluators, as the question of what makes a translation a good translation is difficult to formulate explicitly and in reproducibly. As a result, when humans are asked to evaluate sets of translations, the inter-annotator agreement is often low, and evaluations can even be inconsistent for a single evaluator (Turian et al., 2003). Evaluation is difficult because there are many possible translations for a same source sentence; many will be correct (they will be paraphrases of each other) and some will be better than others for a variety of different reasons and at varying degrees. See for example the different translations presented in (32). They differ in terms of their syntactic structure, word choice and style, yet all are good translations and could be judged differently according to evaluator preferences.

(32) EN: I like that a lot!

FR: J'aime beaucoup ça ! '(lit.) I like a lot that'

FR: Ça me plaît beaucoup ! '(lit.) That pleases me a lot'

FR: Ça me plaît énormément ! '(lit.) That pleases me enormously'

FR: Ah, c'est très bien, ça ! '(lit.) Ah, it's very good, that'

A major problem is that the quality of a translation can be judged on different levels, for example in terms of adequacy, fluency, word choice, style, naturalness and grammaticality, and if these criteria are not specified, different evaluators will put different degrees of emphasis on these different factors. Conversely, if multiple criteria are used, the evaluation procedure becomes more complex, requiring expert evaluators and more time for evaluation, and there is not always a clear consensus as to which set of translations has a higher overall quality compared to a second set of translations.

The evaluation strategy adopted depends on the ultimate aim of the evaluation. For example, to better understand the improvements gained through the use of a new

architecture, it may be useful to perform an evaluation with multiple error types. However, if the aim is to provide a clear ranking of the global quality of multiple systems and models, it is more convenient to have a simpler metric, which results in a clear and unique ranking of systems in terms of overall translation quality. This is the case of the annual shared task in news translation at the Workshop (now Conference) in Machine Translation. Several manual evaluation strategies have been tested since the first edition in 2006: (i) ranking on a five-point scale for adequacy and fluency (White et al., 1994), (ii) relative ranking of MT outputs, (iii) relative ranking of select constituents, (iv) a binary judgment of translation “acceptability” and (v) direct assessment (Bojar et al., 2016b). The current method used is direct assessment, which involves evaluators giving absolute quality measures, which are then normalised for individual variations. The method is highly correlated with relative ranking, used in preceding years, but enables evaluation to be crowdsourced (Graham et al., 2017). A downside is the large number of evaluations necessary to determine statistical significance between the different systems evaluated.

Major considerations in establishing an evaluation protocol are the ease of finding evaluators, the time necessary to evaluate and the ability to find a consensus from the set of final evaluations. One debate has been whether evaluators need to understand the source language of the translation outputs. Whilst bilingual evaluators, who are able to evaluate the quality based on both the source and the target language, are advocated by some (Dorr et al., 2011), other, more recent views recommend using monolingual speakers (of the target language) in order to (i) expand the potential pool of evaluators and (ii) obtain more consistent evaluations (Guzmán et al., 2015).²¹

3.3.2 Standard automatic evaluation metrics

The development of automatic metrics that provide a single, overall translation quality score is essential for the comparison of different MT architectures and models across authors and in time, and have been an important part of research in MT. Their development also meant that MT systems could be evaluated and compared without the need for human evaluators. Ideally, automatic evaluation scores should be highly correlated with human quality judgments, such that they can act as a proxy for overall translation quality. The idea of producing a single score is attractive from a practical point of view, in that we have a way of easily comparing the performance of two models, but as has been discussed in the previous section does not give us an idea of the quality of the separate dimensions of the translation quality. It is also key during the training and tuning of MT models. Having a single evaluation score enables model parameters to be tuned iteratively in order to optimise the score, as described in Section 3.1.2.3 for SMT.²²

²¹We shall come back to this in the context of our own data collection experiments in Chapter 8.

²²An automatic metric suitable for iterative tuning or training must also be relatively fast in order to realistically perform a large number of iterations of parameter optimisation.

The majority of automatic metrics rely on a comparison between the automatically produced translation and one or more human “reference” translations of the original source utterances.²³ Evaluation is usually performed on one or more *test sets*, a set of sentences that is disjoint from the set of sentences used for training and hyperparameter tuning, which is important for judging the MT models’ capacity to generalise to unseen sentences. Metrics differ in the way in which translations are compared, the simplest comparing exact matches of surface forms, and the more complex variants using strategies to approximate a comparison of the underlying semantics of the translations. Parallel to the progress made in MT models themselves, the basic units on which evaluation metrics base their scores began at the word level, progressed to the phrase level, and in more recent years rely on continuous representations.

An early metric used was *word error rate* (WER), a modified version of the Levenshtein distance used to calculate the distance between machine translated output and a human reference translation. Variants of this metric emerged, with slight variations such as a *position-independent WER* allowing for free reorderings and *translation edit rate* (TER). However, it was the development of the BLEU metric (Papineni et al., 2002) that has made the biggest mark on the domain. BLEU is surprisingly simple and, like the first metrics such as WER, relies on a comparison of the surface forms of the candidate and reference translations. BLEU is still today one of the most widely used metrics and, despite facing continual criticism, is practically ubiquitous in MT research. It is one of the metrics that we shall use throughout this thesis to evaluate overall translation quality, and we shall therefore describe its implementation and importantly its limitations in more detail.

3.3.2.1 BLEU

The automatic evaluation metric, BLEU (BiLingual Evaluation Understudy) (Papineni et al., 2002), has undoubtedly had a huge impact on MT research.²⁴ In spite of the criticism BLEU has met in terms of its inability to capture certain phenomena (we shall come back to this in more detail in Section 4.1 and for inter-sentential phenomena in Section 4.1.1), it is widely used and can be a useful tool in MT (despite its imperfections). Other than the fact that it is correlated with human judgment scores, BLEU owes its major success to its simplicity and speed, which makes it sufficiently inexpensive to be iteratively used as a tuning metric for SMT models compared to more advanced metrics that have since been developed (Section 3.3.2.2).

²³There has also been interest in methods of estimating translation quality, aptly named “Quality Estimation”, involving estimating translation quality without comparing to human translations. Quality estimation is useful for several purposes: indicating the confidence of a translation, or potential errors in a translation, particularly for post-editing, providing an estimation of the effort needed to post-edit, which can also be used to rescore potential translations and providing a means of judging whether a translation is of sufficient quality to be used as training data for MT (Specia et al., 2013).

²⁴The original paper recently won a “Test of Time Award” at NAACL 2018.

Calculation of BLEU The principle behind the BLEU score is to compare hypothesis translations with human reference translations of the same source sentences; the nearer the surface forms are, the higher the translation quality is deemed to be. BLEU is essentially a modified form of n -gram precision, based on the degree of overlap between the n -grams of a candidate translation and those of human reference translations. An n -gram precision is calculated for $n = 1, \dots, 4$ over the corpus of translated sentences as follows:

$$\text{precision}_n = \frac{\sum_{h \in H} \sum_{n\text{-gram} \in h} \#(n\text{-gram}, \text{ref}(h))}{\sum_{h' \in H} \sum_{n\text{-gram}' \in h'} \#(n\text{-gram}', h)}, \quad (3.24)$$

where H is the set of hypothesis translations, $\text{ref}(h)$ is the reference translation of translation h and $\#(n\text{-gram}, x)$ is the number of times an n -gram appears in sentence x . This number is in practice clipped (i.e. cannot exceed) the number of times the n -gram appears in the reference translation to avoid frequently repeated words n -grams in a translation resulted in an artificially high precision.

BLEU also integrates a brevity penalty (BP) to penalise short translations, which would otherwise be unduly rewarded by n -gram precisions.

$$\text{BP} = \left\{ \begin{array}{ll} 1 & \text{if } c > r \\ e^{(1-r)/c} & \text{if } c \leq r \end{array} \right\}, \quad (3.25)$$

where r is the length of the reference corpus and c the length of the machine translated text (both lengths in terms of numbers of words).

BLEU is then calculated as follows:

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^N w_n \log \text{precision}_n\right), \quad (3.26)$$

where weights w_n can be the same or set differently for each value of n . In practice it is standard for uniform weights to be used. The final BLEU score is between 0 and 1, where 1 is a perfect match between the hypothesis translation and the reference. Traditionally, BLEU is designed to evaluate translations of an entire corpus and is more reliable (and better correlated with human judgments) on larger corpora, as it is less sensitive to particularities of the test sentences chosen. The basic BLEU is inadapted for sentence-level evaluation, since the BLEU score can easily be equal to 0 if a higher order n -gram precision is 0. Sentence-level variants have also been developed, using smoothing techniques to ensure that scores do not default to 0 when one type of n -gram has a precision of 0 (Lin and Och, 2004; Chen and Cherry, 2014).

Criticism of BLEU Despite the fact that it is still practically universally cited across the literature, BLEU has suffered from a bad reputation, which can be attributed mainly to the community’s over-reliance on the metric, especially in settings in which it is inadapted to perform reliable comparisons. In the original article, Papineni et al. (2002) describe the metric as “an inexpensive automatic evaluation that is quick, language-independent, and correlates highly with human evaluation”²⁵ that can be used “to monitor the effect of daily changes to their systems in order to weed out bad ideas from good ideas.” The metric was designed to compare similar models for incremental, global changes on a same test set.²⁶ And yet BLEU is still being used to compare models across different architectures and is still too often used as the only way of evaluating an MT system. BLEU’s simplicity makes it extremely useful in its original role of providing guidelines for developers, but also detrimental to progress in MT when used incorrectly.

Callison-Burch et al. (2006) discuss some of the issues associated with the use of BLEU. They find that many different translation variants of a same source sentence are scored the same using BLEU, despite the fact that translation quality varies considerably between variants. One of the most limiting problems is the fact that BLEU is a geometric mean based on n -gram counts and as such does not distinguish between different permutations of the same n -grams. This means that word order, which has the potential to fundamentally change the meaning of a translation, is not taken into account. Callison-Burch et al. (2006) give the example of permuted bigram matches in two reference translations, Example (33) representing the same unigram and bigram matches (in their order of appearance in the translation) between the hypothesis translation and the reference translation, and Example (34) the permuted bigram matches. According to Callison-Burch et al.’s analysis, the two translations in Examples (33) and (34) would receive the same BLEU score, despite both translations being poor and the second being clearly nonsensical.

(33) Appeared calm | when | he was | taken | to the American plane | , | which will | to Miami , Florida .

(34) which will | he was | , | when | taken | Appeared calm | to the American plane | to Miami , Florida .

They mention other serious problems associated with BLEU that negatively impact the metric’s correlation with human judgments: the inability to account for surface form variation such as synonyms and paraphrases and the fact that all words contribute an

²⁵The fact that BLEU and other metrics are correlated with human judgments has been questioned by Turian et al. (2003), who show that correlation is low. They also remark however that correlation between human evaluators is also low.

²⁶BLEU does not always provide an equal evaluation across architectures. It is known to underestimate the quality of rule-based approaches (Callison-Burch et al., 2006), and to privilege statistical approaches (notably phrase-based approaches), since the basic units of the metric (n -grams) are also the basic units used in phrase-based systems.

equal weighting despite some words being more important than others. For example, a long translation in which the only mismatching token is the negation *not* can score highly, despite the meaning of the translation being the opposite of the reference translation because of the wrong use of negation. For example, in (35), the first translation, lacking the negation present in the original sentence and necessary to avoid reversing the meaning, is scored higher than the second (correct) translation according to BLEU.

- (35) EN: It is really not something that I am willing to discuss.
 FR_{REF.}: Ce n'est vraiment pas quelque chose que j'ai envie de discuter.
 FR₁: C'est vraiment quelque chose que j'ai envie de discuter.
 'It is really something that I want to discuss.'
 FR₂: Ce n'est vraiment pas quelque chose que je suis prête à discuter.
 'It is really not something that I am ready to discuss.'

The use of such a metric also requires honesty on behalf of its users; its formulation leaves it open to being artificially manipulated in order to give higher scores than warranted by the actual translation quality. One example of this concerns the brevity penalty, designed to penalise translations that are too short. If, at the level of an entire corpus, the translations are shorter than the reference translations, the brevity penalty will cause the score to be reduced. It can in some cases be possible to remedy this decrease in score simply by adding a word to the end of each sentence to avoid the penalty and therefore to increase the score. The modified translations will all be of lower quality than the original ones, since an arbitrary token has been added to the end of each sentence. However they would be unjustly rewarded by BLEU. It is for reasons such as this that BLEU should not be used as the *only* final evaluation of MT outputs, and should be used instead as a guide during development.

3.3.2.2 Other automatic metrics

Many other automatic metrics have been proposed since BLEU, with the aim of overcoming some of the problems associated with it, notably an inflexibility to the use of synonyms or in fact of any surface form variation (Callison-Burch et al., 2010; Ma et al., 2018). We shall mention just two of them here, of which the first will be mentioned in our experiments in Section 5.1: the use of flexible pattern matching and of character-based evaluation metrics.

The metric METEOR (Lavie and Denkowski, 2009) was explicitly designed to counterbalance certain drawbacks of BLEU. It is more flexible to surface form variation because of it does not rely on exact surface form matching, instead using pattern matching, based on stemming and a detection of synonyms using an external thesaurus. This allows translations to deviate from the formulation of reference translations as long as the words remain related (semantically or according to their form) to those in the reference translation. It

also uses the idea of automatic alignments between the reference translation and the hypothesis translation to better model the correspondence between the two translations, which can overcome some problems met with BLEU in terms of its lack of consideration for word order. More recently, character-based evaluation metrics such as CHRF (Popović, 2015) and CharacTER (Wang et al., 2016b) aim to allow more flexibility by taking into account variability using a different strategy: by relaxing the assumption that matches must be made between whole lexical items. These methods have proved better correlated to sentence-level human judgments for morphologically rich languages in particular.

The majority of existing metrics, despite aiming to overcome the problems faced with BLEU, still compare the surface forms of reference translations to those of hypothesis translations, and fail to capture certain phenomena, particularly those associated with cohesion, and those requiring context that is not found within the sentence. Despite alternative metrics being proposed, BLEU retains its place as the most used metric in the domain. This could be put down mainly to historical reasons, because of its success at the time (and its inertia in following years). Other metrics not being adopted in its stead could be seen as having a practical reason rather than a theoretical one: an alternative metric would have to be adopted across the domain and across authors in order for experiments to be comparable.

3.3.3 Discussion

Evaluation is a defining part of MT research. How we evaluate can determine on which aspects of language we concentrate, and consequently which direction our research takes. Given the difficulty even of human evaluation, it is inevitable that automatic evaluation metrics, so necessary for comparing different systems and architectures on a large scale, do not succeed in capturing all aspects of translating natural language. Choices must be made, as must compromises. Standard metrics such as BLEU are often used for a purpose for which they were not designed, that of providing the only evaluation of two systems. However, this does not mean that such metrics are not useful, when used in the correct setting. BLEU is simple and fast and tends to capture large differences in translation quality across two systems that share a similar architecture (e.g. a baseline system and an improved system). Throughout this thesis, we shall use BLEU to illustrate the overall translation quality of MT models, but mostly to give a general idea of quality. To understand what is really being improved by changes to a system, it is necessary to go into more detail. Human evaluation can be useful to understand what sort of errors are committed by the system, but is often too time-consuming to perform over many models and on a sufficiently large test set. An alternative solution is to design automatic metrics that target particular phenomena. For context-dependent phenomena, we shall see in the next chapter how evaluation metrics have evolved to evaluate contextual MT models.

Contextual Machine Translation

Ideally, MT should be able to take into account all relevant contextual information from within the document being translated, from the surrounding context and from world knowledge, as a human translator would. However, integrating context from outside the sentence has always been difficult, in SMT as well as NMT, in terms of complexity and in terms of the modelling capacity of MT architectures. It is a research topic that arose relatively late, due to the fact that focusing on the integration of extra-sentential context requires first having a translation of reasonable quality at the sentence level. Contextual MT has sparked interest from the community, particularly in terms of discourse translation, and most of the focus has been on anaphoric pronoun translation and lexical choice. However, there has recently also been work on incorporating extra-linguistic information into MT, spurred on by the availability of data for which additional information about the speakers or the topic of discussion are provided or can be inferred at a reasonable cost.

In this chapter, we will give an overview of the different strategies used in both SMT and NMT to integrate extra-sentential context and the different ways in which the strategies are evaluated. We choose to begin the chapter (Section 4.1) where we left off in the previous chapter with a discussion of automatic evaluation metrics. Whereas in the previous chapter we looked at traditional automatic MT evaluation, in this first section we review efforts to evaluate MT from a discursive (and importantly contextual) point of view. We present various methods that extend metrics to take into account a wider context but that still aim to produce a single, global MT evaluation score. As we shall see, this has its limitations, and efforts have not always been hugely successful.

Many of the alternative strategies to evaluate contextual MT are therefore specific to a particular phenomenon, and we shall present them alongside the phenomenon-specific MT strategies they are designed to evaluate in the remainder of the chapter. Given the plethora of different strategies adopted to integrate context into MT, we prefix our literature review by first discussing in Section 4.2 what it means to integrate context into MT. We structure our reflection around two questions:

1. How do we represent context before using it to improve MT? We will notably introduce a distinction between *structured* and *unstructured* context, which will be important in the remainder of this thesis.
2. At what point in the translation process can context be exploited?

These definitions will be useful both in the rest of the chapter and for our own contributions presented in Part II. The final three sections of the present chapter will be dedicated to reviewing MT strategies to take into account context, also providing relevant information on resources and evaluation. We will discuss the use of *structured linguistic context* for the translation of discourse phenomena (Section 4.3), before introducing new ways of using *unstructured linguistic context*, particularly in NMT (Section 4.4). Finally, we will review strategies specific to the integration of *extra-linguistic context* (Section 4.5), drawing on work previously presented for the inclusion of linguistic context.

4.1 Evaluating contextual MT: extending global metrics and the case for targeted evaluation

If the automatic evaluation of MT in general is a difficult task, automatically evaluating contextual MT is even more difficult. The widespread use of standard, sentence-level evaluation metrics such as BLEU has often led to the evaluation of the use of context being overlooked. There are several reasons of this, which are linked to the difficulty of the task itself and also to the fact that context-dependent phenomena may not be given the importance they are due in the calculation of an evaluation score.

Before looking at different strategies to integrate context into MT, we first briefly discuss efforts to render automatic evaluation metrics sensitive to discourse-level context. This is important to discuss, as it presents some of the challenges faced and justifies the later use of evaluation methods designed to target particular phenomena rather than to assess overall translation quality. In this section, we will therefore extend our discussion of evaluation metrics at the end of the previous chapter (Section 3.3) to the discourse-level, focusing uniquely on evaluation metrics providing an overall translation quality score. We begin by laying down some of the problems associated with automatic evaluation of

context (Section 4.1.1), before discussing several efforts to include discursive information in automatic metrics (Section 4.1.2).

4.1.1 Problems associated with automatic evaluation of context

We briefly look at three major difficulties associated with automatically evaluating contextual MT: (i) the context-dependent nature of surface forms, (ii) the rarity of context-dependent tokens and (iii) the difficulty of integrating extra-sentential information into MT metrics.

Context-dependent nature of surface forms Any metric that relies solely on comparing the surface form of a hypothesis translation with that of a reference translation is unlikely to perform well at evaluating the contextual phenomena that we wish to study. As we have seen in the previous chapter, the majority of evaluation metrics rely on a comparison of this sort. The use of synonyms or other forms of paraphrase can result in poor quality scores despite being correct, and inversely, incorrect translations may receive high scores due to a high level of matching tokens with the reference translations. Already a problem for sentence-level evaluation, this problem is fatal for cohesion-based phenomena, which necessarily concern a dependency between a form and another translation choice made elsewhere in the translated text. In Figure 4.1 for example, the correct French translation of the English pronoun *it* is entirely determined by the translation of the pronoun's antecedent. If the metric used does not account for this dependency between forms and simply compares the French pronoun to a reference translation, translations can be doubly penalised, or unduly rewarded:

- if the antecedent has been translated using a different choice from that of the reference translation (implying a change in gender), the translation could be doubly penalised: once for the lack of correspondence between the translated antecedent and the reference and again for the lack of mismatch between the translated pronoun and the reference pronoun.
- if the translated pronoun matches that of the reference translation but does not agree in gender with the antecedent in the MT-produced translation, it will be rewarded by the metric, despite resulting in a translation error.

The same problem occurs with lexical cohesion, which by definition concerns the lexical ties that exist within lexical elements, which may appear at a distance in the text (see Section 2.2.2.2).

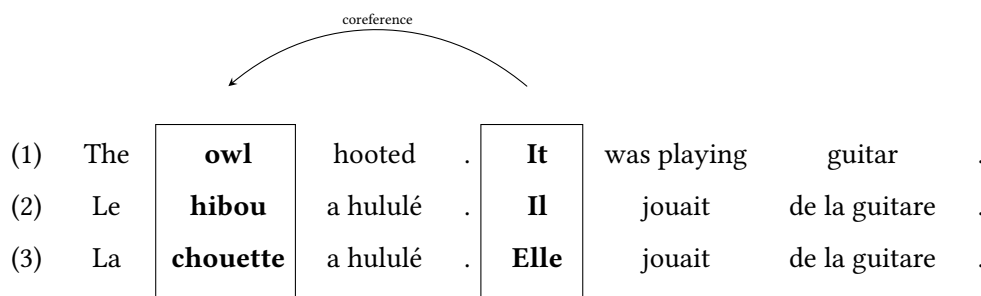


Figure 4.1: An example of two correct translations (2) and (3) of a same source sentence (1), in which the choice of the antecedent (“hibou” or “chouette”) determines the choice of the anaphoric pronoun (“il” or “elle” respectively).

Rarity of context-dependent tokens A major difficulty in evaluating contextual phenomena is that they are often rare relative to the total number of tokens in a corpus, despite them being crucial for good translation. The effect of this is that they can be overlooked by standard metrics whose aim is to provide an overall translation score, particularly if an equal weight is given to all tokens.

Take for example the translation of anaphoric pronouns. A poor translation can have the effect of breaking the coreference link and severely harming the understanding of the translation because of an inability to retrieve the correct reference behind the pronoun. However, this poor translation only concerns one token in the sentence, and sentences containing such pronouns are relatively few overall relative to the total number of sentences in a text.¹ It is therefore unlikely that automatic metrics that are designed to evaluate the overall translation quality will provide a score that will reflect the importance of correctly translating anaphoric pronouns.

Difficulty of adapting metrics to use extra-sentential context The third difficulty associated with evaluating contextual MT is the necessity to include information from beyond the sentence, particularly for cohesion-based phenomena. In Figure 4.1, the antecedent of the pronoun is found outside of the pronoun’s sentence. In order to assess whether the pronoun has been translated correctly, this antecedent (and its gender) must be first identified, which is itself not a trivial task. This problem is not unique to evaluation; it is also the main reason for the difficulty of designing contextual MT methods in general.

¹The frequency of occurrence of anaphoric pronouns depends on the style and genre of text, but remains relatively low overall. In her statistical corpus analysis, Guillou (2016) finds that English third person pronouns make up only approximately 3% of tokens in the TED talks corpora and 1% of tokens in the EU Bookshop corpus.

4.1.2 MT metrics augmented with discourse information

There have been propositions to take into account elements of discourse to improve global evaluation metrics, but with variable success. Several of these rely on external tools designed for discourse parsing, providing an automatically produced representation of discourse structure (which can account for inter-sentential links). An alternative strategy is to integrate discourse features into the metric that are not necessarily produced through a parsing strategy.

Using discourse parsing One direction has been to use discourse parsing to produce discourse representations of sentences, which can be used as a basis for evaluation. Comelles et al. (2010) use such a technique, basing their evaluation metric on a framework for representing the semantics of a text, Discourse Representation Theory (DRT) (Kamp and Reyle, 1993), with the aim of abstracting away from surface variation. The formalism also allows links across sentences, making it potentially useful for the evaluation of anaphoric pronouns, lexical cohesion and discourse connectives. They calculate the similarity between the representations of the hypothesis translation versus the reference translation using different methods: calculating the overlap of subpaths in the discourse trees or calculating the average lexical or morphological overlap in the tree structures. Although, the metrics are designed to capture document-level features, in practice they find that the metric does not succeed in capturing the quality of discourse-level phenomena such as anaphora. They also find that there was no improvement in the human correlation of the metric when compared to the sentence-level version of the same metric. Guzmán et al. (2014) also use discourse-level information for MT evaluation. They use a different formalism from Comelles et al. (2010), namely Rhetorical Structure Theory (RST), designed to make relations between textual spans explicit (rather than completely abstracting away as in DRT). They use the formalism to parse the hypothesis and reference translations and then calculate the similarity between the two trees, which is used as additional information on top of existing metrics. They test their resulting metrics on a larger quantity of data and a wider range of language pairs than Comelles et al. (2010) and, contrarily to Comelles et al. (2010), find that improvements can be seen; adding discourse information to existing metrics appears to be important complementary information.

Other discourse-level features Discourse-level features other than those obtained through discourse parsing can also be of use in metrics. Wong and Kit (2012) also find that extending existing metrics with specific document-level features related to lexical chains can improve the correlation of the metrics with human judgments. In an aim to capture lexical cohesion phenomena, they base their discourse features on the relative frequency of lexical cohesion devices (such as synonymy, near-synonymy, superordinate,

repetition and collocation), which involves detecting content words that are either repeated or collocated with other content words. They find a positive correlation between the number of lexical cohesion devices present in a translation and the quality of the translation, justifying their use of the features as a means of evaluating translation quality. However, this may be correct only up to a point, or specific to the dataset used, as contradictory results have been reported by Carpuat and Simard (2012). In their study of SMT consistency, they conclude that despite the fact that most systems do not look beyond the level of the sentence, consistency of translation is not always a problem and a high consistency of translation of specific terms does not necessarily indicate a higher translation quality. It is therefore difficult to conclude on whether this metric is robust to different scenarios. This will be rediscussed in Section 4.3.2.1, when looking at work on lexical consistency in MT.

4.1.3 Conclusion

The impact of discourse-level evaluation metrics on the MT community as a whole currently remains limited. It is encouraging to see more and more efforts to use discourse-level information to evaluate. However it is unclear whether the improvements seen in automatic metrics are sufficient to be adopted by the MT community. This is partly due to the huge influence and wide-scale adoption of BLEU (see Section 3.3.2.1). Another reason may be that the inclusion of document-level context often demands time-consuming discourse analysis and extraction of features, which give only slight improvements with respect to human judgments. Consequently, they are not typically metrics that can be used for optimisation of SMT systems, unlike BLEU which is fast to calculate by comparison. However, a second problem is the fact that global evaluation metrics by definition do not give us fine-grained breakdown of MT quality by different aspects. When testing new strategies to include contextual information into MT models, evaluation is not just about evaluating the overall quality of the translation, but also about evaluating whether the context is being used in the correct way. Especially if the contextual phenomenon is rare, global evaluation metrics are unlikely to capture any differences, or at least do not do so in a clear way.

An important part of the progress that can be seen in document-level translation is therefore in reality seen not through metrics of the overall translation quality but in phenomenon-specific evaluation metrics, which will be discussed in the following sections. These metrics can supply complementary information about the translation quality of specific phenomena, which is essential for validating that specific modifications to MT systems produce the desired improvements on these identifiable phenomena.

4.2 Modelling context for MT

Most of the previous work in contextual MT, mainly discourse translation, is associated with SMT, a comprehensive survey of which up to 2012 can be found in (Hardmeier, 2012). It has only been in recent years that such topics have been approached for NMT, which itself only became widespread from the mid-2015's. Many different strategies have been proposed to integrate context into MT, and although the problems themselves remain the same regardless of the architecture used, the way in which context is modelled and used by the translation system has been changing, especially with the rise of NMT. Whereas many early strategies seek to target a specific phenomenon with explicit modelling of contextual information (Carpuat and Wu, 2005; Le Nagard and Koehn, 2010; Meyer and Popescu-Belis, 2012), the current tendency is to provide the translation system with unstructured context for the system to learn how to use the information provided (Tiedemann, 2010; Gong et al., 2011; Libovický and Helcl, 2017; Wang et al., 2017), made possible largely by NMT models' memorisation capacity. To better understand why this might be, and the motivations behind the different strategies used, it may be useful to consider two questions:

1. How is context modelled prior to translation?
2. How is context used to improve translation?

The first question concerns if and how context is modelled prior to it being used in the translation process. We consider that context can either be given in a raw, unprocessed format, which we refer to as *unstructured context*, or can be processed (i.e. structured, annotated or selected to target a particular phenomenon), in which case we refer to it as *structured context*. To illustrate this distinction for the case of lexical disambiguation, unstructured linguistic context could refer to all the words of neighbouring sentences (which may or not contain disambiguating context), and structured linguistic context could be *specific* words of the neighbouring sentences that have been selected because of their pertinence for disambiguation. The type of linguistic context (source or target information) that will be most useful will depend on the type of phenomena targeted. This is very much linked to the distinction we previously made between coherence-based and cohesion-based phenomena: coherence-based phenomena may benefit from either source or target-side information, whereas cohesion-based phenomena necessarily require at least target context to be correctly handled. Extra-linguistic information, such as speaker gender and conversation topic, are already considered as structured information, since they have been encoded as discrete labels. However extra-linguistic information *can* be unstructured, for example in the form of raw images or videos of the dialogue scenario and the objects in the setting, used for instance in multi-modal MT (Specia et al., 2016; Calixto et al., 2017).

The second question concerns how (and when) the context, which has either been structured or left unstructured, is exploited in the translation process. One solution is to design a new MT architecture capable of using contextual information in a new way. However, designing a new architecture is not trivial, and it is often easier to adapt existing architectures. There are three ways in which context can be used in conjunction with existing MT architectures:

1. *Pre-processing*: modify the data on which the MT system is trained using contextual information (and therefore the data to be translated too) to remove the ambiguity otherwise present, and apply the standard MT architecture to the pre-processed data (e.g. Le Nagard and Koehn, 2010),
2. *Post-processing*: use the contextual information to post-edit existing MT outputs as produced by a standard MT model, thereby avoiding any modification to the existing MT architecture (e.g. Stymne, 2016; Bawden, 2016; Luotolahti et al., 2016),
3. *Within the translation model*: integrate contextual information into the current functions of the MT system, or modify the architecture so that it is capable of using contextual information (e.g. Hardmeier and Federico, 2010; Hardmeier et al., 2013; Sennrich et al., 2016a; Libovický and Helcl, 2017; Wang et al., 2017).

For those approaches that rely on pre- or post-processing to integrate contextual information, the type of MT architecture used for translation (SMT or NMT) does not affect the design of the approach. However, this is not the case when context is directly exploited within the translation model; this strategy is necessarily dependent on the constraints imposed by the MT architecture and the representation of the sentence to be translated. As we have seen in Chapter 3, SMT and NMT differ greatly in their way of handling text and their way of modelling the translation process, so before looking at specific contextual strategies, we first briefly review each of the paradigms in terms of their capacity to model external context.

4.2.1 Modelling context for SMT

Standard phrase-based models represent words and phrases as discrete units, whose mapping between the two languages is explicitly modelled using a probabilistic phrase table (Section 3.1.2.1). This has a direct impact on the way context can be used in such systems.

To integrate context into the SMT translation process itself, there are at least two options: (i) add a new customised scoring model designed to take into account contextual information, which therefore contributes to the final score of hypothesis translations, or (ii) modify the existing components (e.g. language model or translation model) to

take into account more context. This second option may require changing the way in which words or phrases are represented to allow for richer information to be used to calculate translation probabilities. Two different examples of this are *factored translation models* (Koehn and Hoang, 2007), in which words are represented by different tiers of annotations, and *cache-based models* (Tiedemann, 2010; Gong et al., 2011), designed to calculate probabilities dynamically based on the recent history. Adding new components to phrase-based systems such as MOSES is relatively easy. However, a major difficulty when integrating new context-designed models has been in achieving systematic and significant improvements in translation quality that can overcome the limitations of the standard models in phrase-based systems (Section 3.1.4). Efforts to modify the pre-existing model components have been of variable success. For example, cache-based models for the integration of context appeared to result in translation gains and improve lexical consistency. However, other models such as the factored model architecture become less tractable at a larger scale and therefore are difficult to apply to high-resource settings. One notable effort to create a new architecture in which to incorporate contextual features is DOCENT, a document-wide decoder, designed to integrate features from throughout the document (Hardmeier et al., 2013), which will be discussed in some more detail later in this chapter (Section 4.3.4).

4.2.2 Modelling context for NMT

As seen in Section 3.2.1.1, in NMT, words and sentences are represented by continuous representations. The very idea of modelling translation as a successive encoding of a sequence of words (Section 3.2.1.3) is based on the idea that new information can easily be added and memorised, as long as it is also encoded as a continuous representation. This makes it theoretically easier to incorporate context into the translation model, which can learn to use it where appropriate, without having to explicitly model the way in which the information should be used. With NMT, it is possible to provide arbitrary information in the form of extra tokens within the source sentence and for them to be available at every decoding step. Since the mapping from source sentence to target sentence is performed via intermediate representations of the sentence, rather than direct mappings between source and target phrases as in SMT, extra information in the source sentence can be taken into account by the model and used for the translation of any of the target words. This adds more flexibility as to how the extra contextual information can be used. Importantly, since the neural model is designed to select important information and ignore other less important information, the task of modelling context can be transferred from the researcher to the model. This is the case of a simple contextual NMT strategy, whereby the surrounding linguistic context can be prepended to the current source sentence in order to guide translation (Tiedemann and Scherrer, 2017). This strategy, which is simple and efficient, is nonetheless limited in the amount of contextual information it

can realistically encode, and also by the fact that we have little control over how the model uses the information, particularly if the relevant context is situated very far away from the ambiguous element. In the past few years, new architectures for contextual MT have also been proposed, designed to better model the interaction between the context and the current sentences. Many of the current solutions concern *multi-encoder models* (Libovický and Helcl, 2017; Wang et al., 2017; Jean et al., 2017a), whereby the context is encoded separately from the source sentence and is combined later within the model.

4.3 Translation using structured linguistic context

Much of the work in contextual translation has been carried out on a per-phenomenon basis, which has been important for drawing attention to particular discourse phenomena that require specific types of context to be correctly translated. Particular emphasis has been placed on two phenomena in particular, anaphoric pronouns and lexical choice (regrouping lexical cohesion and coherence), with a wide range of different strategies tested. A third phenomenon, discourse connectives, has also received some interest. The aim to tackle a specific phenomenon often leads to strategies in which a very specific type of context is used in translation, one that has been selected and structured (often with the help of external tools) to help the translation of the phenomenon in particular. In this section we shall review such approaches to contextual translation using structured linguistic context, with a particular emphasis on specific phenomena: anaphoric pronouns (Section 4.3.1), lexical choice (Section 4.3.2) and discourse connectives (Section 4.3.3). Finally, we shall evoke a translation architecture designed to use structured context from the whole document, which has previously been used in a number of experiments to aid the translation of specific discourse-level phenomena (Section 4.3.4).

4.3.1 Anaphoric pronouns

The translation of anaphoric pronouns has been the object of considerable interest in the MT community (Guillou et al., 2016; Loáiciga et al., 2017); its impact on the quality of a translation is evident (cf. Section 2.2.2.2), and it also benefits from the fact that the textual elements involved, the anaphoric pronoun and its referent, are theoretically well identified (see (Guillou, 2016) and (Loaiciga Sanchez, 2017) for in-depth studies of the problem). Research into anaphoric pronoun translation provides good examples of all three different strategies described in the introduction to the current chapter: post-processing, pre-processing and modifications to the translation process. As it is one of the more studied aspects of contextual MT, it has been approached from many angles, and a comparison of the strategies reveals the general tendency to replace explicit modelling

of the phenomenon with implicit modelling by the translation architecture itself; a high reliance on external tools and resources, particularly for SMT models, is replaced by a more expressive NMT architecture designed to learn how to use the information provided, even when it is unstructured.

4.3.1.1 Strategies

As anaphora translation is dependent on target-side agreement, necessitating access to the translation of the pronoun's antecedent (or postcedent), post-processing has certainly been the strategy of choice for tackling the problem. The majority of the submissions to the two evaluation campaigns for anaphoric pronoun translation in 2016 (Guillou et al., 2016) and 2017 (Loáiciga et al., 2017) follow this strategy, since the task was formulated as such. The systems differ largely with respect to the type of information and resources used for prediction. Whereas the baseline system (a language model) and Luong and Popescu-Belis (2016) only use target-side information, other systems choose to exploit both source and target information (Novák, 2016; Stymne, 2016). Interestingly, there was no clear indication that the use of external resources such as lexica or tagging tools worked better than simply exploiting the context within the text itself. Neural classification systems (Luotolahti et al., 2016; Stymne et al., 2017; Hardmeier, 2017) generally performed very well, and became the dominant system type for the 2017 evaluation campaign. There remains much improvement to be made, however. Neural systems with access only to sentence-internal information perform surprisingly well on pronoun prediction (cf. the winning system in the 2016 task (Luotolahti et al., 2016)), showing that the overall quality of anaphoric pronoun prediction is sufficiently low for systems to make substantial gains by concentrating just on those pronouns whose antecedent is within the same sentence. There are at least two downsides of using a post-processing approach. The first is that the effectiveness of the approach is dependent on the quality of the MT outputs, which are difficult to modify or improve once they have been produced without remodelling the translation problem. The second, which is linked, is that the post-processing effort does not necessarily have access to the modelling capacities of the MT system, which could provide useful information.

Despite the fact that anaphoric pronoun translation is a cohesion-based (target language) phenomenon, some efforts have been made to approach the problem using a pre-processing-like strategy. In one of the first efforts to tackle the problem, Le Nagard and Koehn (2010) design a two-pass translation system for English to French, whereby a baseline translation system is first used to produce translations from which the relevant context (the gender of the pronoun's antecedent) can be extracted. The system relies on coreference resolution on the English source side to identify the anaphor's nominal antecedent, and a morphological lexicon to identify the gender of the noun's translation (in the target language). This identified gender is then used to annotate the English

pronoun (e.g. *it-fem*, *it-masc*) in the training data. A new SMT system is then trained on the annotated data, with the aim of estimating better probabilities for the translation of the anaphoric pronouns thanks to their disambiguation in the source language. Whilst being relatively simple, the method has three disadvantages: (i) the possibility that the antecedent be translated differently in a second pass, (ii) the necessity to translate all sentences twice, and (iii), as remarked by the authors, its reliance on the quality of coreference resolution, which in this case was insufficient for the task (only 56% of pronouns were correctly labelled by the coreference resolution system). Their results therefore show almost identical results to a baseline system (at 68% and 69% precision respectively), despite the fact that the newly estimated translation probabilities for the translated anaphoric pronouns were slightly shifted towards the expected translations.

A logical step to avoid translating twice was to design a method of integrating anaphora resolution into the translation process. Following Le Nagard and Koehn's (2010) two-pass approach, Hardmeier and Federico (2010) and Hardmeier et al. (2011) also choose to rely on explicit coreference resolution, but design a one-pass system for English to German and English to French, whereby coreference resolution is integrated directly into the decoding process via an additional scoring model. They use the BART coreference resolution tool (Broscheit et al., 2010) to annotate anaphoric links, and use their additional model to provide the probability of a target pronoun given its previously translated antecedent in the target language. Unlike Le Nagard and Koehn (2010), they do see slight improvements according to precision and recall metrics targeting the evaluation of anaphoric pronouns in particular, but the gains are modest. It is clear that a method using external tools such as a coreference resolution system is reliant on the quality of such tools, and there is high potential for error propagation. A crucial limitation of the use of external coreference tools is that they are not optimised (or evaluated) for their use in translation systems, for which only certain coreferential links are useful (between ambiguous anaphoric pronouns and their nominal antecedent), and therefore even state-of-the-art coreference resolvers may be of insufficient quality for translation. Is there a way around using external tools to provide translation systems with the necessary context? For standard SMT architectures, this appears unlikely, due to their inherent inability to take into account long-term textual dependencies (Section 3.1.4). However, more recently, NMT architectures have made promising progress in anaphoric pronoun translation, without requiring explicit modelling of coreference. Jean et al. (2017b), who additionally include the previous sentence as an extra input, compare three ways of combining the input of the current sentence with that of the previous one, inspired by work on contextual language modelling (Wang and Cho, 2016) and multi-way NMT (Firat et al., 2016). They show that their system rivals post-processing strategies explicitly designed to post-edit anaphoric pronouns, without the use of external tools and resources, and without targeting coreference in particular.

4.3.1.2 Resources and evaluation

Evaluation is a crucial aspect of improving translation quality of anaphoric pronouns. As often remarked (Le Nagard and Koehn, 2010; Hardmeier and Federico, 2010; Hardmeier, 2012), standard evaluation metrics such as BLEU do not always reflect improvements in anaphoric pronoun translation, and alternative evaluation methods must therefore be used. Anaphoric pronoun translation is one of the most straightforward context-dependent phenomena to evaluate, because many of the pronoun instances can be evaluated as either correct or incorrect, depending on the gender of the nominal antecedent. It is therefore common to opt for an evaluation strategy based on standard metrics such as precision, recall and F-score, which is possible with manual evaluation. Despite the problem being well defined, *automatic* evaluation is not without its difficulties. Two main difficulties arise when comparing the output of different translation systems. The first is the fact that the pronoun’s translation will be dependent on the translation of the nominal antecedent (see the discussion in Section 2.2.2.2). The second is that translations are not guaranteed to translate all source anaphoric pronouns as pronouns in the target language. Comparing separate MT outputs may therefore mean evaluating over different sets of pronominal instances, making comparison of multiple outputs difficult.

A solution to these problems was found by organisers of the WMT and DiscoMT shared tasks on cross-lingual pronoun translation in 2016 and 2017 (Guillou et al., 2016; Loáiciga et al., 2017),² who decided to formulate the problem as a post-edition problem, whereby the aim of the task was to correctly choose the form of certain anaphoric pronouns in the target language,³ removed from the text and replaced by a placeholder. By choosing to have participants post-edit pronoun instances based on the same set of target sentences, evaluation could be performed on a comparable basis, using standard precision and recall metrics. The choice was made not to use MT-produced target sentences, but the reference translations themselves, thus facilitating evaluation by making it possible to directly compare pronoun predictions with the pronoun forms in the reference translations. The choice not to work from MT outputs did have its downsides, the main one being that target sentences had to be modified so that they did not include agreement phenomena automatically providing the correct gender of the pronoun. Whereas the source sentences were provided in tokenised form, the target sentences were therefore provided as lemma-PoS-tag pairs, without the original word forms.⁴ An example of three sentences, of which

²Four language directions were available in 2016: English-to-French, French-to-English, English-to-German, German-to-English, and in 2017, French-to-English was replaced by Spanish-to-English. The training and development data provided was parallel data from EUROPARL (Koehn, 2005), News-Commentary (Nc-v9) (Tiedemann, 2012), and the TED corpus (Cettolo et al., 2012).

³Only ambiguous subject pronouns were studied in this task.

⁴This decision was taken following a first shared task in 2015 (Hardmeier et al., 2015), in which all systems submitted were beaten by an n -gram based language model, due to the fact that much of the morphological information necessary was available within the target sentence itself. For example,

4.3. Translation using structured linguistic context

the second contains a placeholder for the prediction of a pronominal form, from the English-to-French training data is shown in Table 4.1.

Pron.	Source sentence	Target sentence	Word alignments
	A lot of arguments against it .	il _{PRON} y _{PRON} avoir _{VER} beaucoup _{ADV} de _{PRP} argument _{NOM} contre _{PRP} ce _{PRON} pratique _{NOM} .	0-3 1-3 2-4 3-5 4-6 5-7 5-8 6-9
elles	Fish farms pollute , most of them do anyway , and they 're inefficient , take tuna .	le _{DET} ferme _{NOM} de _{PRP} aquaculture _{NOM} polluer _{VER} , _{PUN} du _{PRP} moins _{ADV} le _{DET} plupart _{NOM} de _{PRP} entre _{PRP} elle _{PRON} , _{PUN} et _{KON} [] être _{VER} inefficace _{ADJ} , _{PUN} prendre _{VER} par _{PRP} exemple _{NOM} le _{DET} thon _{NOM} .	0-0 1-1 1-3 2-4 3-5 4-9 5-10 6-11 6-12 7-12 8-7 9-13 10-14 11-15 12-16 13-17 14-18 15-19 16-23 17-24
	I almost never cook with it .	je _{PRON} ne _{ADV} la _{PRON} garder _{VER} presque _{ADV} jamais _{ADV} en _{PRP} cuisine _{NOM} .	0-0 1-4 2-5 3-7 4-3 4-6 5-2 6-8

Table 4.1: An extract of three sentences from the English-to-French data provided for the shared task, taken from the TED corpus. Source sentences are tokenised, but target sentences are tokenised, tagged and lemmatised (the original forms are not given). Automatic word alignments are given between the source and target sentences. The anaphoric pronoun to predict in the example is the feminine plural *elles*, which corefers with the French lemma *ferme* ‘farm’.

Whilst the shared task has had an undoubtedly positive impact on research in discourse MT, the formulation of the task makes the scenario different from the more realistic scenario of post-editing real MT outputs. Another downside of the task, a consequence of the natural distribution of pronouns in the data, is the very uneven distribution of pronoun types within the corpora, in particular the test set (see Table 4.2 for the distribution of French pronoun classes). To discourage participants from seeking gains on the most frequent pronoun types (such as *il* and *ce* for French), and instead encourage them to resolve the least probable types (such as *on*, *elle* and *elles* for French), the official metric used was macro-average recall, thus giving more weight to instances of rarer classes. The knock-on effect was the high sensitivity of the metric to very minor differences, exacerbated by the very small numbers of the rarest classes (e.g. *on*, *elle* and *elles* for French).

An alternative evaluation strategy, designed to evaluate MT outputs, was developed by Guillou and Hardmeier (2016). Their PROTEST test suite includes an English-to-French test set (the test set used for the 2015 DiscoMT pronoun prediction shared task), annotated for pronoun types, and an automatic evaluation script to compare MT outputs to the annotated reference translations. There are a total of 4,732 pronoun types annotated in the corpus, 644 of which represent anaphoric pronouns whose antecedent is within the same sentence, and 761 anaphoric pronouns whose antecedent is outside of the current

a target sentence [] *est affamée*_{FEM}. ‘[] is starving’, directly provides the pronoun’s gender due to grammatical agreement, which would not be available in machine-translated output.

sentence. The test set is annotated in the style of the ParCor parallel corpus (Guillou et al., 2014), a parallel corpus annotated for coreferential links and link types for pronouns, developed to aid MT research into anaphoric pronouns. Several possible translations of nominal antecedents are provided, where appropriate, in order to allow for variation in translations, and the evaluation script allows for manual evaluation to be done for those translations not yet covered by the test set. The suite is designed to evaluate both incremental changes of a same system, for which the automatic part of the evaluation may suffice to gauge performance differences between systems, and to compare system performances by partly relying on manual evaluation.

4.3.2 Lexical choice

A second highly studied aspect for MT is lexical choice. Both lexical coherence and lexical cohesion can fall under the category of *lexical choice*, when they concern the choice of which words to use in context. It will be useful to group these two terms for two reasons. The first is that the line between coherence and cohesion can sometimes be blurred. Take for example the case of obligatory lexical repetition of *de poivre* ‘the pepper’ in Example (36). English *pepper* is ambiguous between the spice (Fr. *poivre*) and the fruit (Fr. *poivron*), and so the lexical disambiguation of the second instance of *the pepper* can simultaneously be seen as a lexical disambiguation problem (thus concerning coherence) and a cohesion problem.

- (36) EN: Sprinkle the pepper into the pan. *The pepper?*
 FR: Ajoute une pincée de poivre dans la poêle. *De poivre ?*

The second reason is that approaches designed to deal with one of the two phenomena sometimes inevitably touch upon the other when using contextual target-side information. As we have seen in Section 2.2.2, lexical disambiguation, as a coherence-based phenomenon, can use context from either the source or the target (since ideally the target should reflect the same semantic content as the source), whereas lexical cohesion chiefly concerns target-side context.

Dataset	Total	ce	cela	elle	elles	il	ils	on	OTHER
EUROPARL	494,110	51,170	13,202	48,460	18,387	168,579	45,603	9,452	139,257
Nc-v9	35,226	2,822	1,027	4,224	1,918	8,248	7,451	566	8,970
IWSLT15	69,487	16,415	6,908	3,286	3,276	9,682	17,145	1,549	11,226
TED (dev)	563	151	63	25	15	57	140	10	102
TED (test)	363	68	31	23	25	61	71	9	75

Table 4.2: The distribution of pronoun classes for the English-to-French task in the data provided for the 2016 task (Guillou et al., 2016). The final datasets represent the development and test sets respectively.

Strategies that seek to model lexical choice explicitly (by targeting particular words and injecting the relevant structured context) can be often seen as adhering to two different angles of attack: (i) seeking to improve lexical consistency, and (ii) aiming to perform word sense disambiguation (WSD). We shall review these two sets of strategies here and also discuss targeted evaluation methods designed to assess MT models on their ability to make correct lexical decisions.

4.3.2.1 Lexical consistency

One angle under which lexical cohesion and lexical disambiguation have been studied and evaluated is that of lexical consistency, concerning both cohesion and coherence. A number of studies are dedicated, particularly in SMT, to improving the level of consistency in translations, under the hypothesis that greater consistency results in better translation quality, especially in terms of reducing ambiguity. Consistency is typically evaluated by using standard metrics such as BLEU, despite them being mostly unadapted⁵ or by using manual analyses relying on a reduction in the number of translation errors.

Inspired by Gale et al.'s (1992) hypothesis that “one meaning per discourse” can lead to better performance and consistency in monolingual WSD tasks, Carpuat (2009) applied the same principle to SMT, leading to small gains in translation quality as measured by BLEU and METEOR. A similar strategy was used by Xiao et al. (2011), who replace translations of ambiguous words by the most frequently used translation using a two-pass approach, resulting in a 25% reduction in translation errors linked to inconsistency. A softer approach was introduced by Ture et al. (2012) and later by Pu et al. (2017) to encourage rather than to impose lexical consistency across sentences. They also used a two-pass approach by way of passing document-level features linked to lexical consistency based on the translation produced in the initial pass.

Despite these initial positive results, there has been some debate as to whether lexical consistency is actually correlated with translation quality. It is quite clear that for certain domains, in particular those requiring a specific, controlled vocabulary, such as technical or legal documents, lexical consistency could be required, and in certain cases of lexical cohesion, such as in obligatory lexical repetition as in Example (27) on page 32, consistency is necessary to convey the correct meaning. However, it is unclear whether document-wide lexical consistency outside of these contexts is something that should be sought in MT. The debate was introduced by Carpuat and Simard (2012), who study the relationship between lexical consistency and translation quality in SMT, and by Guillou (2013), who analyses when lexical consistency is deemed to be good, also comparing human translations to SMT outputs. Their consensus is that a high level of

⁵Unless the reference translation display the same consistency sought after and use the same terminology as the MT outputs.

consistency does not necessarily equate to an improvement in translation quality. SMT systems proved to be relatively consistent in the choice of vocabulary, even more so than human translators in some cases (Guillou, 2013). However the degree to which increased consistency resulted in improvements in translation quality appeared to depend on a variety of factors: the overall translation quality, the type of element concerned and the genre and style of text. Carpuat and Simard (2012) remark that weaker SMT systems, which are usually trained on less data than stronger systems, tend to have fewer translation options in general, and therefore tend to translate consistently, but not out of choice. Guillou (2013) finds that the importance of consistency can also depend on the morphological category and the rarity of the word: keeping consistency for light verbs was not found to be helpful, whereas consistency seemed more important for nouns. Importantly, the level of consistency was found to be very dependent on the style of the author and the genre of the text and therefore not a criterion that can be directly equated to translation quality without going into further detail.

4.3.2.2 Lexical choice via word sense disambiguation (WSD)

A second vision of lexical choice is as a word sense disambiguation problem, which can potentially be handled using pre-existing WSD techniques. Applying these systems to MT is a natural development. As mentioned by Carpuat and Wu (2005), it is a test of the hypothesis often made in WSD research that the systems can be used in downstream applications such as MT. Carpuat and Wu (2005) actually question this hypothesis, by showing that applying standard WSD to SMT is not beneficial and actually harms the translation performance. They found that a major stumbling block was the fact that very few words overall needed disambiguation, and sometimes word choice was degraded by the WSD model. A second major problem was the way in which the WSD was integrated into the SMT model: they achieved negative results when constraining the decoding to certain candidates and when modifying the ambiguous items through post-processing. Their method was both detrimental to performance and costly in terms of effort. Vickrey et al. (2005) achieve slightly more promising results, but they do not test the impact of WSD on a full translation task. They instead perform WSD on word translation using classification with contextual features. It was not until sense disambiguation was applied to larger sequence of texts, such as phrases, rather than to individual words that more promising results were seen (Carpuat and Wu, 2007; Chan et al., 2007). Specia et al. (2008) use a different strategy and choose to rerank hypothesis translations using rich WSD features. This technique has the advantage of being able to use a large number of expressive features than can be integrated directly into an SMT system. They see improvements for a select number of frequent ambiguous words. A major problem with the use of out-of-the-box WSD systems for translation is that they are not optimised for the translation task, and therefore are of a lesser utility than their intrinsic evaluation

on monolingual data would lead to believe. Whereas monolingual WSD systems aim to distinguish between many different fine-grained senses, these sense differences do not necessarily result in a different translation per sense. Errors in the WSD can easily lead to error propagation later in translation, particularly for rarer terms.

For NMT, the stakes have changed slightly, and intra-sentential lexical ambiguity can pose less of a problem if clear disambiguating context is present within the sentence. The memory capacity of recurrent encoder-decoder NMT models means that models are designed to learn to use information from the previously translated words, and the attention mechanism means that a representation of the entire source sentence is used at each decoding step. This memory capacity (and the access to the entire source sentence) means that they are at an advantage compared with phrase-based SMT systems in their ability to handle long-distance context. These improvements in intra-sentential modelling for translation does not however prevent lexical ambiguity from being a problem across sentences, and word sense disambiguation remains a relevant issue. As previously mentioned, inputting additional features in NMT is relatively easy, as they can be included as input tokens in the source sentence and remain accessible throughout the translation of the sentence. It has proved effective for a range of linguistic features using factored translation, by which each word can be represented as the concatenation of a certain number of features (Sennrich and Haddow, 2016). Rios Gonzales et al. (2017) use this method to include sense labels as additional input features in an NMT system, by first mapping words to sense embeddings computed with SENSEGRAM (Plevina et al., 2016) based on their surrounding context. They compare this with a second method based on the construction of lexical chains (series of words within a text that are semantically related). They use the sense embeddings from their first method and construct their lexical chains using the method presented in (Mascarell, 2017). They then annotate each word with the sense embeddings contained within its lexical chain, and input these as additional features. Despite the ease of including the features, they do not see systematic gains for the two language pairs tested (German-to-English and German-to-French) for either of the two methods. The method nevertheless remains promising for future work due to the ease of integrating features.

4.3.2.3 Evaluation

Developing tractable methods of evaluation for lexical choice is a must to overcome using unadapted automatic metrics such as BLEU. With the aim of testing the translation of particularly ambiguous words, a blank-filling tasking for the MT of ambiguous words was introduced by Vickrey et al. (2005). However, this was centred only around certain words and does not represent a full translation task. In a similar fashion, Mihalcea et al. (2010) introduce a cross-lingual version of the lexical substitution challenge: a task to predict several possible translations of a word in context. It differs from the blank-filling

task in that the target sentence is not given, and the aim is instead to predict the multiple possible translation of an English source word in context. Again, this method, although interesting, does not test an MT model's capacity to translate whole sentences in context. The difficulty lies in how to evaluate whole translations that may differ widely between different systems. Rios Gonzales et al. (2017) use a different technique, namely using an MT model's capacity to score translations. Inspired by a technique used to evaluate grammaticality of translations (Sennrich, 2017), they evaluate models on their ability to provide a higher score to a correct translation than incorrect ones, where the contrastive translations differs only in the translation of the ambiguous word. Their test sets are automatically created by artificially creating lexical errors and are on a large scale (6,500+ sentence pairs for each of the language pairs German-to-English and German-to-French).

4.3.3 Discourse connectives

A third discourse phenomenon that has previously been studied in the context of MT, but less so than the two previous aspects, is that of discourse connectives (Meyer et al., 2012; Meyer and Webber, 2013; Meyer and Popescu-Belis, 2012; Steele and Specia, 2016). As described in Section 2.2.2.1, the translation of discourse connectives can be problematic because of their frequent ambiguity with respect to their discourse function and because discourse relations are not always explicitly marked with a discourse connective. This second problem is challenging in both translation directions: when translating a discourse connective, a decision must be taken as to whether or not to translate it, and when translating a sentence with an implicit discourse relation, it may be necessary to translate using a discourse connective, despite there being none present in the source sentence.

Meyer and Popescu-Belis (2012) succeed in improving the translation of discourse connectives in an English-to-French SMT system, by first training a classifier to predict discourse senses and then integrating the information into an SMT system. They test two strategies to make use of the predict senses: (i) modifying the phrase table after training and tuning by labelling source connectives for their senses based on their translations, and (ii) training a new model on training data in which discourse connectives are labelled for their senses. They see limited changes in terms of BLEU score, which is unsurprising, but do see improvements to discourse connective translation according to a manual annotation, based on the number of translations degraded or improved. Modifying the SMT phrase table had the greatest effect with 34% of connectives improved against 20% degraded. However, training a new model on the labelled data also led to minor improvements (18% improved versus 14% degraded). It is interesting to note that improvements are only seen when connective sense classification uses features from the candidate translations as well as source features, showing that information from within the target language is vital for the translation of discourse connectives.

In similar experiments, Meyer and Poláková (2013) improve English-to-Czech SMT by training a new system directly on large-scale manually annotated parallel data (Popescu-Belis et al., 2012), rather than using a classifier. A manual evaluation revealed an improvement of between 8 and 20% in the translation of discourse connectives, when not including the frequent connective *but*, which was systematically translated with the unnatural translation *jenžé*, due to the presence of this translation in the training data. They do find however some problems with very rare connectives, exacerbated by the fact that labelling the connectives increases data sparsity.

4.3.4 Whole document decoding

Translating using information from the entire document has always been a difficulty for MT, and is the reason for the sentence being considered the maximal translation unit in many cases. The two-pass approach previously mentioned for both anaphoric pronoun translation and lexical disambiguation is popular because of its simplicity, enabling a second system, such as a classifier to have access to a first set of translations from which to extract relevant features to aid in the production of the final translations. Hardmeier et al.'s (2012) whole document decoder, DOCENT, functions in such a way. The principle is to construct a framework enabling the inclusion of new features from the entire document to help the translation of discourse phenomena. The framework requires a baseline translation from which features can be extracted, and then performs local searches (hill climbing) to incrementally improve the document's translation. Inspired by work in greedy decoding (Langlais et al., 2007), the idea of the system is to incrementally modify a baseline translation by using the highest scoring transformation operator amongst the possible modifications “swap”, “move” and “replace”. The system is flexible, allowing for custom actions to be introduced, and the scoring system itself enables features from the entire document, ensuring that cross-sentence dependencies could also be taken into account.

Despite the model's flexibility and potential for increased complexity, gains have remained relatively modest. Several authors have used the framework to tackle particular discourse phenomena. For example, Mascarell (2017) integrates a feature function scoring the semantic similarity of lexical chains as automatically detected within the translations. The aim was to favour translations from which the automatically detected lexical chains contained items that were maximally related semantically. García-Martínez et al. (2017) also focus on lexical consistency, using DOCENT to integrate a feature function based on the consistency of a new lexical item given its context, and introduce a new “change” operation to explore multiple lexical changes in a single step. Again, they see non-significant changes to the BLEU score, but promising results via manual evaluation.

4.4 Translation using unstructured linguistic context

The previously described works focus on improving the translation of particular phenomena by explicitly providing the context required for translation. The approach they use is to tackle specific problems of ambiguity by modelling the context upstream of translation, and therefore providing structured or inferred context to the translation process itself. There is another angle that can be taken to use context in translation, which consists in exploiting the translation architecture's way of modelling translation, such that structure within the context is learnt by the translation model itself, without the need for upstream modelling. There are a number of advantages associated with this approach. By jointly learning the structure of the context at the same time as learning to translate, the information learnt from the context is more likely to be directly relevant for translation. There is also a reduced risk of error propagation brought about by the use external tools and resources, which may introduce erroneous or partial information. The difficulty is finding methods well adapted to how the translation model is learnt, bearing in mind the constraints imposed in the modelling process.

For phrase-based SMT systems, one way of learning how to use context is to rely on dynamic translation and language models, whose probabilities change according to recent vocabulary choices. Inspired by domain adaptation techniques, these *cache-based* models are designed to give more weight to recent linguistic history by assigning weights to context that decay over time. Inspired by cache-based language modelling for acoustic models (Kuhn and De Mori, 1990), Tiedemann (2010) experiments with additional, dynamic models for SMT with a decaying cache. He finds that whilst the use of a dynamic language model improves the overall quality as estimated by BLEU for a majority of documents, the use of a translation model cache does not lead to significant gains, which the author puts down to the noise of the cache and the propagation of errors from past translations. Gong et al. (2011) extend this idea by combining dynamic models with static models at the document level, which they refer to as “topic cache”, based on similar documents. They too find slight gains in overall translation quality, but again find this to be dependent on the document, resulting for some documents in degradation of translation quality. Apart from the weaknesses cited by Tiedemann (2010), it is somewhat unclear why these methods were not more successful, and notably why performance was so degraded for certain documents. One possible explanation offered is that the cache is over-simplified (due to data sparsity) and therefore over-simplistic in the way it can bring new information, also making it more prone to error propagation.

In NMT, the same types of cache-based models can be implemented in a less simplistic way. Instead of manipulating discrete units as with phrase-based systems, there is the possibility of storing richer representations of recent history that may help to avoid error propagation or learn ways of deciding whether using recent history is indeed

useful. Kuang et al. (2018) implement in NMT a very similar strategy to the one used by Gong et al. (2011) for SMT, using both a dynamic cache to store recent word from target hypotheses and a topic cache storing semantically similar words from the current document. They use both caches to produce a cache-based probability score of the next word and combine it with the probability obtained from the NMT decoder via a gating mechanism, allowing flexibility as to how much weight the cache has in the final translation of a word. They achieve slight gains in overall translation quality and improvement in the coherence of translations, based on the semantic similarity of words in successive sentences.

There has been a lot of work recently in different ways of exploiting previous unstructured linguistic context in NMT. Using multiple encoders, whereby linguistic context is encoded using a separate encoder from the source sentence, is one option. Inspired by work on dialogue generation (Serban et al., 2016) and multi-source translation (Zoph and Knight, 2016), the idea is that a separate representation of the linguistic context can be learnt, and then used within the model to supply secondary information. It can be used to initialise the RNN encoder or decoder (Wang et al., 2017) of the main NMT model. It can be combined via a gating mechanism with the decoder hidden state representations (Wang et al., 2017) in order to influence the translation of each word. It can also be integrated at an earlier stage within the model, by for example using a second attention mechanism to compute a secondary context vector (as per the terminology of Section 3.2), which can then be combined with the context vector representing the source sentence at each decoding step. Libovický and Helcl (2017) use such a strategy, and control the importance of the representation of the previous context by adding a secondary, hierarchical attention mechanism over the two context vectors, thus assigning a hierarchical attention weight to each context representation.

The NMT architecture makes it easy to integrate new representations into the translation system, but the real problem is designing a way in which the linguistic context is used *well*, resulting in visible improvements in translation quality. Many of the works up until present have focused on better ways to use context from the previous sentence, rather than aiming to include a wider history. Although the representations of context within NMT are rich, they are often represented by a fixed-size vector within the model. Given that the representation is learnt by the model, the researcher has little say over what type of contextual information can be stored in this vector. Given its fixed size, it is also incapable of storing an infinite linguistic history. But even before looking to design models with infinite history, it is clear that there is still progress to be made to integrate even very recent history such as the previous sentence. In Jean et al.'s (2017a) contextual multi-attention model, they evaluate using both BLEU and a targeted evaluation of anaphoric pronouns (as per the evaluation in the 2016 shared task on pronoun prediction (Guillou et al., 2016)). Their results show that their contextual model gives moderate gains according to both types of evaluation, but only when the training data size is small (fewer

than one million parallel sentence pairs), and the gains rapidly disappear, finally resulting in degraded performance, when more data is used. Lacking an appropriate way of telling how the contextual information is being used by the systems, it is very difficult to draw conclusions about whether these contextual models are actually providing improvements or not, or whether the additional parameters in the model (or the additional data input in the form of the linguistic context) is helping the models to train better, where adding more sentence-level data would also provide gains.

4.5 Translation using extra-linguistic context

Although the tendency to use unstructured context rather than structured context has not spread to all types of extra-linguistic context as much it has done for linguistic context, the passage of MT to neural architectures has provided new opportunities for the integration of this kind of context. The situation is different from linguistic context for a number of reasons: (i) extra-linguistic context is not the same type of data as the source sentence being translated, and therefore should not necessarily be encoded in the model in the same way, (ii) certain types of extra-linguistic context apply to a whole sentence or even a whole document rather than to a specific word or sentence (e.g. topic or scenario), and (iii) there is often less data available that is accompanied by appropriate extra-linguistic context.

The third point is probably the main reason for there not yet being as much research into the integration of this type of context as linguistic context, for which the text itself can provide its own context. However resources are becoming more widely available, for different types of extra-linguistic context. Certain types of texts are accompanied by meta-data indicating information, providing structured information about the context in which the text was produced. For example, parallel corpora such as OPENSUBTITLES2016 (Lison and Tiedemann, 2016) are accompanied by information about the genre, year and country of origin for a large number of films, and are linked to their imdb numbers, which can potentially be used to retrieve extra information about the films, such as summaries and descriptions of settings. A subset of films in the same corpus can also be automatically aligned to online film transcripts (Wang et al., 2016a; van der Wees et al., 2016). Since transcripts are annotated for speakers, more detailed extra-linguistic information can be automatically extracted from actor profile sites, provided extra-linguistically annotated parallel data. Providing unstructured extra-linguistic context is less easy than it is for linguistic context, which is by our definition present within the text. One way in which it has recently been provided in an unstructured way is in the form of images, to be used to aid the translation of captions or descriptions into a target language (Elliott et al., 2016, 2017). However such data is expensive to produce and therefore is not yet available in the same quantities as raw parallel text.

There are various different strategies used to exploit extra-linguistic context in MT, depending largely on whether the context is structured (as annotations to the text) or unstructured (in the case of visual data for example). Many of the techniques resemble those used for linguistic context, but with the added complication that the context to be included is not of the same nature as the source text. Many of the methods currently used are surprisingly simple, and yet effective, as we shall see in more detail below. However, techniques are starting to be developed to exploit richer types of extra-linguistic context, either by implicitly learning them within a model ((Michel and Neubig, 2018) for speaker personality) or by inputting rich visual information (Specia et al., 2016).

Structured extra-linguistic context One way of viewing certain types of structured extra-linguistic context is as different classes, or to use familiar terminology, domains. Information such as the genre of a film, or the topic of conversation is a domain in the traditional sense, but we can also consider information such as the gender or the age of the speaker as being a domain insofar as they define a sub-language on which we may want to adapt an MT model. One of the simplest ways to use these types of context in a domain adaptation setting (regardless of the MT architecture) is to partition all data based on context labels (for example to create a subcorpus for each genre of film containing only sentences associated with that genre) and then to train separate model components for each genre. These components can be interpolated with larger, generic models and then tuned to the class label in order to create specific MT models for each context value. This approach has been used in various forms for various speaker traits such as gender and formality (van der Wees et al., 2016), as well as topic and genre (Foster and Kuhn, 2007). This simple method becomes less adapted when multiple extra-linguistic traits are to be used, as this would result in multiplication of the number of data partitions necessary, increasing the problem of data sparsity.

When training an NMT models, having large quantities of training data is important. It is therefore important to envisage methods that do not reduce the quantity of data available. A very simple technique that has proved effective in NMT models is to include contextual values as extra tokens within the sentence itself, for example as a pseudo-token at the beginning of the sentence, which can be used throughout translation. Such a method has been used to control the use of polite and familiar pronouns for the translation of English into German (Sennrich et al., 2016a) and has proved highly successful, despite its simplicity. What is yet unknown is whether this technique can scale up to multiple pseudo-tokens. The approach appears not to lend itself to multiple contextual values, because it would treat them as part of the ordered sequence of words of the current sentence, despite order of the contextual values being completely arbitrary. An interesting recent work proposes to learn speaker traits implicitly within the model. Michel and Neubig (2018) compare the pseudo-token approach, whereby they provide the speaker identity as a single pseudo-token, to an approach where they introduce an

additional bias vector before the final softmax computation. They compare the use of a speaker-specific bias vector with a factored bias approach whereby parameters can be shared between speakers. They achieve slight gains in BLEU score with the use of bias vectors, but very similar results to the pseudo-token approach for some language pairs. However, they also show that their method also allows them to better predict the author of a sentence, showing that the method is capable of learning some implicit speaker traits.

Unstructured extra-linguistic context Although we shall not be presenting contributions on the use of unstructured extra-linguistic context within this thesis, it is worth mentioning the growing interest within the community surrounding multi-modal NMT (Elliott et al., 2015; HITSCHLER et al., 2016). This relatively new research domain concentrates on how to exploit raw extra-linguistic context in the form of images to aid translation of captions or descriptions, the idea being that in many cases, the image will provide extra, richer information not present in the source sentence that can be useful for disambiguation (Specia et al., 2016; Elliott et al., 2017). A number of strategies that have been used to exploit the extra visual information actually resemble the more recent strategies used to integrate unstructured linguistic information evoked in Section 4.4. Having encoded the visual information separately from the source sentence (but nevertheless within the same NMT system), the encoded representation can be combined with the source sentence representation via a hierarchical attention mechanism (Helcl and Libovický, 2017), via a shared attention mechanism (Caglayan et al., 2016a) or used to initialise the encoder and/or decoder (Calixto et al., 2017). Many alternative strategies are now emerging, including efforts to first extract structured information from the images that can be directly provided to the MT system (Huang et al., 2016). Correctly exploiting the image data is however a huge challenge, as shown by the fact that in the 2016 shared tasks all systems using visual information were outperformed by the textual baseline MT system.

4.6 Conclusion on evaluating contextual MT

A wide variety of strategies have been tried and tested for taking into account both linguistic and extra-linguistic context. The methods used have changed with the changing MT architectures, and we now have more possibilities for using extra information in translation, both structured and unstructured. It appears with the introduction of NMT that the integration of context (at least of neighbouring sentences) is simpler than was the case for SMT. Many of the approaches seen before NMT became popular relied heavily on external tools and resources to provide extra information that was otherwise difficult to obtain. Phrase-based SMT models could not learn such information for themselves, and also had a limited memory capacity during translation. However there remains a lot of progress to be made. In Part II of this thesis, we aim to

go some way to providing new methods of handling context for different phenomena. We shall see the same trends in contextual methods as presented in this first part, with our early methods relying on richly structured context and external tools and our later experiments relying on the design of the translation models to learn how to exploit context.

In terms of evaluation, there remain some fundamental problems in the way we test whether or not they give us real improvements by using context. It is pretty well established that BLEU scores are a bad reflection of the translation quality of discourse-level phenomena, or even phenomena requiring extra-linguistic context. We have seen in Sections 3.3 and 4.1 the problems associated with BLEU if it is to be used as an end metric (as opposed to a tool during the development phase). And yet many of the contextual strategies still use increases in BLEU score to justify that their models exploit context well, without relying on a secondary metric or analysis to back this up. As suggested by Jean et al.'s (2017a) paper entitled 'Does Neural Machine Translation Benefit from Larger Context?', it is not clear whether increases in overall translation quality actually mean that the context is being used appropriately in many of these models, or whether it means that too little data was used to train the model.

Evaluation strategies do exist for specific phenomena, as shown by the shared tasks on anaphoric pronoun translation (Guillou et al., 2016), the Protest test suite (Guillou and Hardmeier, 2016) and contrastive test sets for WSD (Rios Gonzales et al., 2017), and it is in the interest of progress that we develop new strategies and datasets for evaluating new models.

Part II

Using contextual information for Machine Translation: strategies and evaluation

Introduction

We dedicated the first part of this thesis to laying out why it is important to integrate context into MT, how standard sentence-level MT systems work and which strategies have previously been used to integrate context into MT. In this second part, we will present a range of our contributions to contextual MT, both in terms of strategies for integrating context and in terms of methods for evaluating how well the context is exploited.

We do not aim to present an exhaustive study of different strategy types, instead choosing to present and compare a range of different methods, applied to different types of context, inspired by trends in the community and new architectural advances.

Our contributions to contextual strategies are structured according to the categorisation of strategies we presented in Section 4.2, which is determined by the moment at which context is exploited in the translation in the process: (i) in a pre-processing step, (ii) in post-processing of MT outputs or (iii) in the MT model itself (learnt during training). The strategies presented reflect the advances in MT over the last couple of years, our first experiments dealing with phrase-based SMT models and the final experiments concerning NMT. This has an inevitable impact on the choice of strategies we choose to adopt. Our pre-processing experiments in Chapter 5 and our post-processing experiments in Chapter 6 are designed to integrate the context at a moment other than translation *per se*. The advantage of this is that the methods are *a priori* architecture-agnostic and do not require modifying the translation architecture nor having to adapt the method to the constraints imposed by it. In our third set of experiments in Chapter 7, we exploit the flexibility afforded by NMT architectures to explore methods of integrating context in the learning of the translation model itself. We use one of these contextual NMT architecture, alongside a baseline NMT architecture, to train two different models to

mediate spontaneous bilingual dialogues in Chapter 8. The aim of this chapter is to apply the translation models in a real-life setting for both evaluation purposes and to collect real dialogue data for further research. Throughout the four chapters, we explore the issues in contextual MT in more detail, provide solutions and provide the foundations for future work.

5	Adapting translation to extra-linguistic context via pre-processing	103
5.1	Integrating speaker gender via domain adaptation	104
5.2	Conclusion	114
6	Improving cohesion-based translation using post-processing	117
6.1	Preserving style in MT: generating English tag questions	118
6.2	Anaphoric pronoun translation with linguistically motivated features . .	138
6.3	General conclusion on post-edition approaches	152
7	Context-aware translation models	153
7.1	Translating discourse phenomena with unstructured linguistic context .	154
7.2	Contextual NMT with extra-linguistic context	173
7.3	Conclusion	199
8	DIABLA: A corpus for the evaluation of contextual MT	201
8.1	Dialogue and human judgment collection protocol	202
8.2	Description of the corpus	209
8.3	Evaluating contextual MT with the DIABLA corpus	215
8.4	Perspectives	219

Adapting translation to extra-linguistic context via pre-processing

Related publications: (Bawden et al., 2016) and (Bawden, 2017b)

We begin our series of contributions by focusing on the integration of context upstream of translation (i.e. in a pre-processing step). This approach can only be used with certain types of context. For instance, it would be inappropriate for cohesion-based context, where the context is based on the formal nature of the target translation. However, for extra-linguistic context, which is of an external nature, such an approach is well suited.

We choose to look at one particular strategy of pre-processing, inspired by domain adaptation techniques. We apply the strategy to the adaptation of SMT models to speaker gender for the translation of television subtitles from English to French. As one of the early contributions in this thesis, chronologically speaking, the architecture used for translation is an SMT architecture, implemented using Moses (Koehn et al., 2007).¹ We shall revisit speaker gender in Section 7.2 using alternative strategies specific to NMT, in an aim to overcome some of the difficulties faced in this preliminary work.

¹The same technique is also compatible with an NMT architecture.

5.1 Integrating speaker gender via domain adaptation

We have previously seen that extra-linguistic context can be important for translation (cf. Section 4.5). In this section, we take one type of extra-linguistic context, speaker gender, and investigate how it can be used to improve translation of English to French. Such information can be necessary to translate certain constructions, for which the translations are morphologically marked for speaker gender. This is the case in Examples (37-39), where speaker gender determines the form of certain self-referential adjectives, past participles and nouns.

(37) EN: I am **surprised** and **shocked**.

FR_{MASC}: Je suis **surpris** et **choqué**.

FR_{FEM}: Je suis **surprise** et **choquée**.

(38) EN: She **saw** me coming.

FR_{MASC}: Elle m'a **vu** venir.

FR_{FEM}: Elle m'a **vue** venir.

(39) EN: I wanted to be a **singer**.

FR_{MASC}: Je voulais être **chanteur**.

FR_{FEM}: Je voulais être **chanteuse**.

This is the most obvious type of influence that speaker gender may have. However it is also possible for speaker gender to influence other aspects of language that are much more subtle. The question has long been studied in linguistics (Cameron and Coates, 1985; Goddard and Meân Patterson, 2000). It has been almost invariably shown that female language is more conservative (less non-standard), and strays less from the prestige norm than it does for male speakers. Whatever the sociological or cultural reasons for this divergence, this difference in language could well have an impact on NLP processing of utterances of either male or female speakers. Van der Wees et al. (2016) find that there are differences in translation quality between male and female speakers (worse translation for male speakers), which they link to the fact that male speakers tend to produce more vulgar utterances, which are harder to translate. Hovy (2015) finds that integrating demographic factors including speaker gender into a number of different classification tasks also improves performances. He finds for a sentiment analysis task that male reviewers tend to be more negative and offer fewer positive comments than female reviewers. Using information about the speaker in NLP tasks can have a positive effect. Wherever there are differences between subparts of a text, it can be useful to adapt automatic data-driven processing techniques to take into account those differences, since they are dependent on the data on which they are trained and tuned. It is therefore intuitive to use domain adaptation techniques to integrate such information, and to treat the different values (e.g. male and female) as different *domains*.

The techniques we propose to use here are domain adaptation techniques, ways of adapting otherwise *out-of-domain* SMT models to a particular domain (in our case to speaker gender). Even the simple technique of tuning a generic SMT model on in-domain data has shown to improve translation quality (Pecina et al., 2012). Another simple way of adapting models is to train and tune model components on in-domain data that is partitioned into classes. Foster and Kuhn (2007) and Finch et al. (2009) show that mixing models trained on the different partitions and tuning them to specific classes can help performance for topic-dependent and sentence-type-dependent data respectively.

The technique we use here is similar to that of Foster and Kuhn (2007) and Finch et al. (2009). Starting from a parallel corpus of television subtitles, annotated on the sentence level for speaker gender, we divide the corpus into two subsets, one containing utterances spoken by male characters and the other utterances spoken by female characters. We experiment with different ways of using these partitions of the original corpus to create two gender-adapted SMT models, one for male speakers and one for female speakers, which improve the translation quality of a non-adapted SMT baseline. Independent experiments using the same techniques for speaker traits including speaker gender were conducted in parallel by Wang et al. (2016a) and van der Wees et al. (2016).

We begin by describing the annotation of a parallel in-domain corpus for speaker gender (Section 5.1.1). This data is used as in-domain data to train gender-adapted SMT components, to tune models and for our test set. In Section 5.1.2 we describe the different translation models, both the baseline non-adapted models and the multiple gender-adapted variants, and the results of our experiments. In Section 5.1.3 we analyse these results using manual evaluation and propose a discussion of the method and the evaluation metrics used. We shall offer a critical viewpoint on the approach, particularly concerning the feasibility of its extension to cover multiple contextual traits.

5.1.1 Annotating the *The Big Bang Theory* reproducible corpus

Taking into account speaker gender in translation requires having access to parallel data for which sentences are annotated for speaker gender. This type of data is not always readily available. For example, much of the available subtitle data does not typically include sentence-level information other than the timestamps of the subtitles, as this type of information is not necessary for their primary use. In this work, we rely on a corpus of annotated data produced using the TVD plugin (Roy et al., 2014), which is designed to produce reproducible datasets, exploiting visual, auditory and textual data directly extracted from DVDs and freely available web sources.² We choose to use the first two seasons of the American television series *The Big Bang Theory* (BBT), which is already available in the plugin. The corpus contains several layers of information: aligned video,

²Due to copyright restrictions, the corpus cannot be distributed, but is reproducible using the plugin.

audio, OCR-extracted multi-lingual subtitles³ and fan-produced transcripts, which have been aligned to the audio signal (Bredin et al., 2014).⁴

We extract a parallel corpus by aligning French and English subtitles using the associated timestamps. Since there is not always a one-to-one mapping and a perfect temporal alignment between French and English subtitles, we use heuristics to concatenate subtitles where necessary. In an effort to keep as many subtitles as possible, we eliminate subtitles only when there are no corresponding subtitles in the other language. The transcripts, aligned to the audio signal, provide speaker names, which are automatically transferred to the English subtitles. Finally, speaker gender is assigned to each subtitle by manually mapping speaker names to their corresponding gender. We leave the task of using automatically predicted gender to future work.⁵

We divide the corpus into three datasets: BBT-train (the first forty episodes), used to train translation and language models, BBT-dev (the next six episodes), used for tuning, and BBT-test (the last six episodes), used for evaluation. As per the strategy used, we partition each set into two subsets, one for each gender, male and female.⁶ Basic corpora statistics can be found in Table 5.1. Note that there is a strong class imbalance towards male speakers, who produce approximately three quarters of all test sentences. Subtitles corresponding to female speakers are also on average shorter than those for male speakers, and the percentage of out-of-vocabulary tokens compared to the two subtitle corpora (OPENSUBTITLES2016 and BBT-train) is much smaller for female than male speakers, perhaps indicating a less heterogeneous use of vocabulary within this corpus, or a consequence of the smaller size of the dataset. This will prove important for the interpretation of results.

5.1.2 SMT models: baselines and adaptations

Given the small size of the gender-annotated training data, it cannot be used in isolation to train MT models. The data would provide an insufficient coverage of vocabulary and probabilities would be insufficiently reliable. This gender-adapted data is instead used on top of a strong generic baseline model to bias it towards the in-domain (gender-specific) language. Our baseline models rely on two pre-existing parallel corpora, EUROPARL (Koehn, 2005) and OPENSUBTITLES2016 (Lison and Tiedemann, 2016), which will be described in more detail below. Our experimental setup is simple. We use the MOSES

³Subtitles were automatically extracted from the image using Tesseract (Smith, 2007) and VobSub2SRT (<https://github.com/ruediger/VobSub2SRT>).

⁴<http://bigbangtrans.wordpress.com>

⁵Gender identification is a standard part of speaker diarisation systems (e.g. Barras et al., 2006), and we can therefore assume that if the information is not provided as meta-information, it can be obtained through processing of the audio signal, if translating oral transcripts.

⁶We ignore the small number of subtitles that are associated with speakers of unknown gender (189 sentences out of 9,592 for BBT-train).

SMT architecture (Koehn et al., 2007), which allows new components to be easily added into the scoring function, as described in Section 3.1.2. Each generic training corpus is used to train a phrase table and a language model, which are used as the basis for the SMT model. The gender-annotated data is then used to adapt the model in a number of non-mutually exclusive ways as tuning data, as training data for a new in-domain language model, and as training data for a new in-domain phrase table.

5.1.2.1 Additional (non-adapted) training data

For our baseline model, we use the EUROPARL parallel corpus of parliamentary speeches (Koehn, 2005) and, more adapted to our domain, the film and television subtitle corpus OPENSUBTITLES2016 (Lison and Tiedemann, 2016).⁷ As this second corpus is very large, we use the Modified Moore-Lewis (MML) filtering algorithm⁸ (Axelrod et al., 2011). We use this filtering technique to create a smaller version of the corpus, which is used to train a phrase table and avoid the computational difficulties of training on such a huge amount of data. This filtered version, which we refer to as OPENSUBS-mml, contains the 8% of OPENSUBTITLES2016 sentences most similar to our in-domain training set (BBT-train).⁹ We nevertheless still use OPENSUBTITLES2016 to train a French language model.

A comparison of the characteristics of the different corpora, including the BBT corpus, is given in Table 5.1. In terms of the type of language used, OPENSUBTITLES2016 is the most similar to the in-domain BBT data. The average sentence length is most similar (9.4 tokens for OPENSUBTITLES2016 and 9.0 for BBT-train), and much shorter than EUROPARL (27.1 tokens). One side-effect of filtering OPENSUBTITLES2016 is the drop in average length between the corpus before and after filtering, most likely an effect of the fact that longer sentences are more different from each other in the two subtitle corpora. This results in the fact that the average sentence length in OPENSUBS-mml (5.9 tokens) is much shorter than the other corpora. Note also the relative sentence lengths of source and target sentences. Whereas French translations are generally longer than the corresponding English ones (as is the case with EUROPARL), the opposite is seen for the subtitle corpora, most probably linked to a shortening during subtitling due to subtitle-specific space constraints and approximations made in translation (e.g. the use of non-literal translations and partial translations).

OPENSUBTITLES2016 and OPENSUBS-mml are most similar to our in-domain data in terms of the vocabulary used. The corpora also contain a similar proportion of the token *je* ‘I’ (6.2% of tokens for OPENSUBTITLES2016, 4.8% for OPENSUBS-mml and 3.0% for BBT-train), compared to 0.9% for EUROPARL), which importantly indicates that the similar dialogic

⁷We remove all episodes from *The Big Bang Theory* from this second corpus to avoid any bias.

⁸See the discussion on domain adaptation techniques in Section 3.1.3.

⁹The 8% was chosen to ensure a resulting dataset of approximately 2 million sentences, which we judge to be a sufficient quantity of data to train a phrase table.

5.1. Integrating speaker gender via domain adaptation

nature of the corpora. We calculate the number of out-of-vocabulary (OOV) tokens on the English side of our in-domain data with respect to the three corpora to be used to train phrase tables (EUROPARL, OPENSUBS-mml and BBT-train). OPENSUBS-mml unsurprisingly appears to be the most adapted corpus vocabulary-wise to our BBT data, resulting in the fewest OOV tokens.

Corpus	#sents.	Ave. #toks./sent		%OOVs with respect to...		
		EN	FR	EUROPARL	OPENSUBS-mml	BBT-train
BBT-train	9,597	9.0	8.4	4.6	2.0	0
BBT-train _{MASC}	7,462	9.0	8.5	4.6	2.1	0
BBT-train _{FEM}	1,941	8.9	8.1	4.2	1.3	0
BBT-dev	2,089	9.1	8.1	4.2	1.7	7.3
BBT-dev _{MASC}	1,506	9.2	8.1	4.4	1.8	7.8
BBT-dev _{FEM}	428	8.7	7.6	3.7	1.1	5.0
BBT-test	1,941	9.2	8.2	4.3	1.7	7.2
BBT-test _{MASC}	1,438	9.4	8.4	4.3	2.0	8.9
BBT-test _{FEM}	354	8.8	8.0	4.3	1.0	4.7
EUROPARL	1,969,224	27.1	30.0			
OPENSUBTITLES2016	27,737,442	9.4	8.9			
OPENSUBS-mml	2,218,997	5.9	5.4			

Table 5.1: Corpora statistics of all corpora used. Out-of-vocabulary (OOV) percentages are calculated for the English source-side tokens with respect to each of the corpora shown.

5.1.2.2 Choosing a high quality baseline

We begin by choosing a sufficiently good baseline SMT model trained on non-gender-specific data. This is necessary for both comparison purposes and as a basis on which to add gender-adapted components. As candidate models, we train a series of translation models based on each of the training corpora cited above and on combinations of the corpora, and we compare their translation quality. All models are trained using the MOSES toolkit (Koehn et al., 2007). We compute word alignments over all training data used (BBT-train, EUROPARL and OPENSUBTITLES2016), and parameters are estimated using KBMIRA to optimise the BLEU score. Each corpus is used to train a phrase table and a language model,¹⁰ which are used as scoring components in the final translation model. When several corpora are cited (e.g. EUROPARL + BBT-train), multiple phrase tables and multiple language models are used (one per corpus). For EUROPARL and BBT-train, the same corpus is used to train both the phrase table and the language model. For OPENSUBTITLES data, OPENSUBS-mml data is used to train the phrase table (for computational reasons as mentioned above), whereas the full corpus (OPENSUBTITLES2016) is used for the language model. We shall refer to this model as OPENSUBS from now on. Multiple language and phrase tables are combined using MOSES's

¹⁰Language models are 4-gram models, trained using KenLM (Heafield et al., 2013).

default model combination approach,¹¹ and are assigned weights during the tuning step. A single reordering model is used for each system, based on the largest corpora used.

Model data	BBT-dev			BBT-test		
	BLEU _{ORIG}	BLEU	METEOR	BLEU _{ORIG}	BLEU	METEOR
BBT-train	13.21	11.83	32.7	13.54	12.08	32.6
EUROPARL	13.99	11.46	32.8	13.84	11.79	32.6
OPENSUBS	23.07	19.73	41.9	23.44	20.48	43.2
EUROPARL + BBT-train	16.66	14.08	36.2	16.62	14.40	36.8
OPENSUBS + BBT-train	23.23	20.08	41.9	23.87	21.08	43.4
OPENSUBS + EUROPARL	24.39	20.78	43.0	24.45	21.30	44.1
OPENSUBS + EUROPARL + BBT-train	24.45	21.12	42.8	24.31	21.42	43.7

Table 5.2: Automatic evaluation of baseline models using both BLEU and METEOR. The top three scoring models in each column are highlighted in green, the shade indicating their rank (darkest=best).

The translation quality of the candidate baseline models according to automatic evaluation metrics is shown in Table 5.2. Given the inadequacy of automatic metrics (cf. Section 3.3.2), we evaluate using two of them: BLEU and METEOR. Contrarily to the results presented in (Bawden et al., 2016), we provide two BLEU evaluation scores, each corresponding to a different tokenisation of the translations. The original BLEU score (BLEU_{ORIG}) corresponds to BLEU calculated on MOSES-tokenised texts,¹² and the second BLEU score (simply marked as BLEU) uses the more standardised BLEU, as provided for the WMT shared task evaluation, which includes a simple, internal tokenisation.¹³ Both BLEU scores are calculated on cased versions of the text. This second BLEU score has the advantage of being agnostic to tokenisation across systems, and will be used throughout this thesis. The first three rows show the results of models trained on each of the corpora in turn (one phrase table and one language model per model), and the four following rows show the results of models using combinations of corpora (several phrase tables and language models). It is unsurprising that both the BBT-train and EUROPARL models generalise poorly; the first lacks coverage because of its small size and the second because it is ill-adapted to speech-like data. However adding EUROPARL to OPENSUBS, by far the best-adapted dataset, does improve the scores. Similar results are achieved on the test set. As the models are tuned with BLEU, we judge the best model combination to be the one with the highest scores on the dev set as obtained with the second metric METEOR. This model is the one trained on OPENSUBS + EUROPARL, which we use as our baseline from now on.

¹¹Multiple phrase tables are combined using Moses’s *either* strategy.

¹²This tokenisation was used for BLEU evaluation in (Bawden et al., 2016). The scores here differ slightly from those cited in the article because we report cased rather than lowercased BLEU, in order to be consistent with other results presented in this thesis.

¹³We use the `multi-bleu-detok.perl` script distributed as part of the MOSES toolkit.

5.1.2.3 Gender-specific adaptations

Taking our chosen baseline system (OPENSUBS + EUROPARL), we experiment with a series of adaptations: (i) using gender-specific tuning data, (ii) adding a gender-specific language model (+LM_{FEM/MASC}), (iii) adding a gender-specific phrase table (+TM_{FEM/MASC}), and (iv) adding both a gender-specific phrase table and language model. The additional gender-specific models are trained using the gender-specific subsets of BBT-train. We distinguish three types of tuning data: *all* (BBT-dev), *female* (BBT-dev_{FEM}) and *male* (BBT-dev_{MASC}). We test each of the models individually on BBT-test_{FEM} and BBT-test_{MASC}.

Model adaptation	Tuning data	BBT-test _{MASC}		BBT-test _{FEM}	
		BLEU _{ORIG}	BLEU	BLEU _{ORIG}	BLEU
<i>Baseline</i>					
∅	all	23.65	20.33	25.01	23.21
<i>(i) Gender-specific tuning data</i>					
∅	male	23.84	20.33	25.50	23.69
∅	female	23.37	20.00	25.06	23.24
<i>(ii) Addition of a gender-specific language model</i>					
+LM _{MASC}	all	23.91	20.63	24.56	22.76
+LM _{FEM}	all	23.12	19.95	23.92	21.92
+LM _{MASC}	male	23.69	20.38	25.18	23.32
+LM _{FEM}	female	23.73	20.34	25.99	23.24
<i>(iii) Addition of a gender-specific phrase table</i>					
+TM _{MASC}	all	23.70	20.47	24.92	23.32
+TM _{FEM}	all	23.39	20.06	24.67	22.80
+TM _{MASC}	male	23.60	20.47	25.08	23.75
+TM _{FEM}	female	23.26	20.01	25.11	23.06
<i>(iv) Addition of a gender-specific language model and phrase table</i>					
+LM _{MASC} +TM _{MASC}	all	23.82	20.65	25.09	23.63
+LM _{FEM} +TM _{FEM}	all	23.32	20.16	24.24	22.25
+LM _{MASC} +TM _{MASC}	male	23.95	20.72	25.53	23.86
+LM _{FEM} +TM _{FEM}	female	22.30	19.74	24.56	22.69

Table 5.3: Translation performance after adaptation of the OPENSUBS + EUROPARL baseline model. As before, the colours indicate the best three systems in each column, with the darkest shade of green indicating the best system.

The results (Table 5.3) show that methods exploiting speaker gender vary between showing very slight gains and degrading translation quality. The gains seen in are in fact not significant (tested using paired bootstrap resampling (Koehn, 2004b)), shedding doubt on the presence of improvements at all. Small gains can be seen for both genders for at least one configuration compared to the baseline (first row). The highest score seen for the male test set is the combination of a specific language model, a specific translation model and BBT-dev_{MASC} for tuning, with a BLEU_{ORIG} score of 23.95, representing a +0.30

BLEU_{ORIG} improvement over the baseline score. The improvement between the baseline and adapted model is greater for female speakers, with an improvement of +0.98 BLEU_{ORIG} (to give a BLEU_{ORIG} score of 25.99) for the configuration (+LM_{FEM} tuned on BBT-dev_{FEM}). However the overall picture is not clear. The pattern of results is different depending on the tokenisation used (cf. the results for BLEU vs. BLEU_{ORIG}), despite the scores using the same metric. For example, the highest scoring model for the female test set changes according to which evaluation is being performed. Moreover, contrary to expectations, a high scoring model for the BBT-test_{FEM} set is the male-adapted model +LM_{MASC}+TM_{MASC}, tuned to BBT-dev_{MASC}. This result suggests that any improvements do not necessarily stem from contextual gender information, but are most likely to be a by-product of using different data for training and tuning. This is particularly noticeable for the female-adapted model, because there is far less female-specific data on which to train SMT components; the partitioning of data results in a division in the amount of training data available. Models that are trained on more data are therefore likely to produce higher results on this test set, regardless of whether the model is gender-adapted.

5.1.3 Manual analysis and discussion

In light of this, do any of these slight gains in the automatic metrics really lead to an improved use of gender information? We have already discussed in Part I the problems associated with standard automatic evaluation metrics, and so we shall not repeat them here.

We therefore look for explicit signs of improvement, by analysing the translation of gender-marked adjectives referring to the speaker, as in (37), repeated here in (40). Given the small size of our dataset, the number of sentences containing gender marking of this type are too rare; we manually identify 11 cases in the female test set (out of 354 sentences) which, given the lexical choice, could have resulted in a correction of gender, only one of which actually results in a correction. We even identify a case of reported speech uttered by a female speaker, in which the gender is erroneously corrected: *The man said 'I am a physicist'*, translated as *L'homme a dit 'je suis physicienne'* with a feminine suffix *-ienne*.

(40) EN: I am **surprised** and **shocked**.

FR_{MASC}: Je suis **surpris** et **choqué**.

FR_{FEM}: Je suis **surprise** et **choquée**.

To dig further to see if there are any notable differences between the outputs of the baseline and adapted systems, we perform a small manual evaluation. We compare the translations of both BBT-test_{MASC} and BBT-test_{FEM} produced by the baseline and by the best model for each gender: +LM_{MASC}+TM_{MASC} tuned on BBT-dev_{MASC} (for male

5.1. Integrating speaker gender via domain adaptation

Source and translations	Quality comparison
EN: ... and the American spirit ? FR _{baseline} : et l' esprit ? 'and the spirit' FR _{adapted} : et l' esprit américain ? 'and the American spirit'	Better
EN: I am on the horns of a dilemma . FR _{baseline} : je suis dans un dilemme . 'I am in a dilemma' FR _{adapted} : je suis à un dilemme . 'I am at a dilemma'	Worse
EN: ... T describing your current circumstances ... FR _{baseline} : décrivez votre circonstances actuelles ... T ... 'describe _{2.pl} your current circumstances ... T ... FR _{adapted} : ... T décris ta situation actuelle ... '... T describe _{2.sg} your current situation ...	Better
EN: what do you want ? FR _{baseline} : tu veux quoi ? 'what do you want?' FR _{adapted} : qu' est-ce que tu veux ? 'what do you want?'	Neutral

Table 5.4: Some examples of the manual comparison of the baseline and gender-adapted models. We indicate whether the gender-adapted model is *Better*, *Worse* or *Neutral* (equally (in-)correct), and annotate the differences between the MT outputs.

speakers) and +LM_{FEM} tuned on BBT-dev_{FEM} (for female speakers).¹⁴ For each test set, we annotate the differences between the non-identical translations of the baseline the gender-adapted model in terms of seven difference types: addition, deletion, reordering, lexical choice, tense, gender agreement and *tu/vous* distinction. Several difference types may be observed for a single translation. We also indicate whether the gender-adapted translation is *better*, *worse* or *neutral* (equally good or equally poor). We annotate all 114 of the female utterances that differ between the baseline and adapted model, and we randomly select 200 of the 523 differing sentences for male speakers. Some examples of these annotations are given in Table 5.4.

	Number of differing translations			
	Total	Better	Worse	Neutral
Male	200	76 (38%)	64 (32%)	60 (30%)
Female	114	50 (44%)	32 (28%)	32 (28%)

Table 5.5: Manually annotated quality differences between the outputs of the gender-adapted models compared to the baseline.

¹⁴We choose the models that are the highest scoring according to BLEU_{orig} on the test set.

	% of differing sentences that contain a change						
	Add.	Del.	Reord.	Lex. choice	Tense	Gdr. agr.	<i>tu/vous</i>
Male	28.5	17.5	14	58	4	0.5	5
Female	35	15	9.5	55	5	3.5	9.5

Table 5.6: Manual analysis of differences between baseline and the gender-adapted models.

Results are provided in Tables 5.5 and 5.6. The most common difference for both genders is lexical choice, followed by additions and deletions. A change in lexical choice is more often associated with an improved translation than a degraded one (38% vs. 31% for both genders), but is far from being consistently an improvement in the gender-adapted model. Changes most linked to an improvement are additions, and conversely, changes most linked to a degradation are deletions; for male speakers, 73% of sentences whose only difference is an addition are improved (82% for females), and 93% of sentences whose only difference is a deletion are degraded (60% for females). These observations suggest that the difference in BLEU score might result from differences in sentence length; the BLEU metric heavily penalises translation hypotheses that are shorter than the reference and, as shown in Table 5.1, female utterances are on average shorter than male utterances. It turns out that the baseline model produces translations that are 99.3% of the length of the reference translations for male speakers and 97% for female speakers, whereas the adapted models produce translations for which the relative shortening is lesser: 99.8% for male speakers and 98.8% for female speakers.

What these results tell us is that most improvements seen in the translations by the adapted models are not linked to gender-marking. There are far too few occurrences to be able to test whether there is any effect. Errors linked to other factors such as translation length and wrong lexical choice have far more impact on the quality of translation, according to the outcome of this manual analysis. Domain adaptation appears to have an impact, but not necessarily in the way expected. Each dataset has its own distinct lexical properties, and it is possible that the improvements are simply due to minor lexical specificities of the two datasets and not necessarily due to a real gender bias. Data sparsity, in particular for female speakers, is a problem with this method, which could in part explain why the male-adapted model performs highly on the female-specific test set.

5.1.4 Conclusion on data partitioning

Despite the generally inconclusive results of these experiments, they nevertheless enable us to draw a certain number of conclusions about the strategy used and about the adequacy of evaluation metrics. Similar experiments performed in the same period as our own (Wang et al., 2016a; van der Wees et al., 2016) also show the same small gains in automatic evaluation scores, but variable results across-the-board. They perform the

experiments in slightly different settings. Van der Wees et al. (2016) employ the same method based on sentence type (questions, declaratives, exclamations) and formality (vulgar, colloquial, neutral), with most significant gains being seen for vulgar utterances following adaptation. Wang et al. (2016a) use a specific language model, as we do in this work, to adapt translation to speaker gender, and see gains of the same order of magnitude as our own (approximately +0.5 BLEU). We conduct a wider range of experiments than Wang et al. (2016a), and although we do see slight gains in BLEU score according to certain configurations, they are not significant. Importantly, the strategy itself does not lead to systematic improvements across the adaptation methods used, and the results are sometimes counter-intuitive in terms of which model has the highest BLEU score for a particular gender. These results point to an inadequacy in the choice of methodology, namely data sparsity. The method itself relies on data partitioning, which inevitably leads to the use of smaller datasets than the original data available. MT in its current state is a domain in which having large amounts of data is a must to produce a robust and high-performing model, so the decision to reduce the size of the data used is not one that should be taken lightly. The few improvements seen are not clearly linked to the element of context we focus on, showing that while the method does provide some sort of bias towards the different training data used, it is not the type of bias we hope to achieve.

It is difficult to imagine this method being extended to integrate many sentence-level features. In theory, the method is simple, only requiring annotating data according to the set of features (e.g. sentence type, formality, topic). However, in practice, partitioning data depending on the set of class labels, whose number is multiplied each time a new feature is added, would result in increasingly smaller datasets on which to train and/or tune models, which would inevitably lead to degradation in translation performance.

5.2 Conclusion

The problem of pre-processing approaches to integrating context is that there is no guarantee that the context will be used in the way in which it is meant to be used. The techniques that can be used depend highly on the modelling capacity of the MT architecture. The experiment shown in this chapter is an example of this: although the context provided was exact each time (the gender of the speaker), the phrase-based SMT models were unable to use this information correctly, as quantity of data on which the models are trained is limited. If the masculine and feminine variants of gender-marked adjectives are not present enough times in the gender-annotated data then the translation probabilities will be insufficiently reliable to make an informed distinction during translation. As discovered in this experiment, the partitioning of data makes the overall translation quality suffer, and the highest scoring models are sometimes simply those trained on most data.

There are alternative solutions to this simple strategy designed to overcome the problem we encountered with data sparsity, although they tend not to scale to larger training data sizes and more open domains. A solution has been proposed by Saluja et al. (2011). It involves using all data for all domains, rather than partitioning data. They weight the different training sentences according to their similarity to the test sentence (based on their set of class labels), therefore giving more importance to similar domains. Their experiments shows some improvements, but in a limited domain, and has not been shown to work in a more open domain such as ours. Another solution proposed is the use of factored translation models (Koehn and Hoang, 2007), which allow extra tiers of linguistic annotations to be provided, factorising the information. Unfortunately, the method suffers from computational problems when scaling up to large datasets, and has yet to be used in a very large-scale setting.

Since these solutions have been proposed, MT has undergone the shift from SMT to NMT, and new possibilities have become available. Data partitioning could theoretically still be used for NMT, but is unlikely to be used extensively; NMT functions particularly well for high resource languages, and training many individual models would reduce the amount of available training data. NMT presents new, simple ways of integrating context, exploiting the capacity of the model to learn how to exploit the context. One such strategy, which we shall explore in Section 7.2, is to add a pseudo-token at the beginning of the source sentence containing the value of the context to be used (e.g. GENDER-male or GENDER-female). Unlike SMT, NMT enables the decoder to use this information at any point during the translation process, and not necessarily to translate the token itself. Its presence simply adds an element of context on which the translation can be based. This technique has been used effectively to encourage the consistent use of familiar and honorary pronouns (*tu* vs. *vous* by Sennrich et al. (2016a)). The advantage over the approach presented in this chapter is that there is no longer the need to partition the data and therefore reduce the amount of training data available.

We shall come back to speaker gender in Section 7.2. We take this pseudo-token strategy a little further, looking at multiple types of extra-linguistic context. We compare the strategy to an alternative strategy, relying on changing the NMT architecture used. We will also address the problem of evaluation in this later section, adopting a more targeted strategy for estimating whether the context is being used as we hope it should be.

Improving cohesion-based translation using post-processing

Related publications: (Bawden, 2016), (Bawden, 2017a) and (Bawden, 2017b)

Having seen in the previous chapter an approach involving pre-processing data prior to training and translating, in this chapter we take a look at approaches designed to improve translation by exploiting context at the other end of the process, in a post-processing step. Many of the previous works on contextual MT (see Chapter 4) rely on the post-edition of MT outputs to take advantage of linguistic context. When target language context is necessary for translation, post-edition is an attractive solution, as it is applied downstream of the MT process itself, and MT outputs, from which linguistic context can be extracted, are therefore available. Post-editing is a system-agnostic strategy that can be applied to any MT output, whether it is from an SMT or an NMT system (although the error types may differ depending on the MT system used).

This chapter is dedicated to two contributions relying on post-editing to improve the translation of two contextual phenomena: *tag questions* when translating into English from French, Czech and German (Section 6.1) and *anaphoric pronouns* when translating from English to French (Section 6.2). Both phenomena are related to cohesion and require using linguistic context from within the MT outputs, making post-editing a sensible strategy. We choose to use a similar method for both phenomena, using linguistic features in a statistical classification framework. However the particularities of each of the phenomena lead us to choose slightly different strategies concerning the degree to which linguistic context is structured. For anaphoric pronoun translation, we choose to use few, highly structured features based on linguistic intuitions. For the generation of

tag questions, we choose to use a wide range of features based on structured context. Our aim, as elsewhere in Part II although perhaps to a greater extent in this case, is to reflect on how strategies could be developed to tackle the problem and on the potential pitfalls of evaluation.

6.1 Preserving style in MT: generating English tag questions

When it comes to the MT of discourse, revisiting the question of what constitutes a high quality translation is essential: which aspects of language should be tackled and how to evaluate them. Amongst the aspects that are important to study are those that can be considered stylistic; translating is about preserving as best as possible all communicative aspects of the text being translated, including any indications of speaker stance and attitude. We choose to study one particular aspect, whose use is related to speaker attitude and style: the English tag question (hereafter TQ), utterances such as *catchy, ain't it?* and *it wasn't him, was it?*. The problem we focus on is the translation of TQs into English, since few other languages have such a systematic use of TQs.¹ Learning where to generate them when an equivalent is not present in the source sentence is notably a real challenge. Their correct translation is both a question of coherence (in terms of communicating speaker attitude) and one of cohesion, ensuring the appropriate form is chosen given the rest of the translation.

We propose to study the new task of generating English TQs in translation. It poses theoretical problems about what it means to preserve style, how to evaluate what makes a good, stylistically appropriate translation, and what is the best way to deal with a discursive aspect that is quite language-specific and not very common at the level of an entire text. We build on preliminary experiments presented in (Bawden, 2017a) to further explore what it means to improve the generation of TQs, and why evaluation remains an open problem. In our experiments, the target language is always English, and we experiment with three source languages: Czech (CS), French (FR) and German (DE), chosen to reflect a range of different high-resource scenarios. Each of our experiments is formulated as a multi-class classification task using features from both source sentences and machine translated outputs. The prediction is then used to post-edit state-of-the-art MT outputs.

We begin this section by first defining TQs and discussing their various communicative functions (Section 6.1.1). In Section 6.1.2, we present our classification approaches, which we use to make predictions over how to post-edit the MT baseline output. In our results

¹Translating TQs from English into another languages is also challenging, but the challenge is slightly different: largely concerning when not to translate TQs into the target language.

section (Section 6.1.3), we discuss how to evaluate such approaches. We present a series of different evaluations, discussing what they mean in terms of translation quality. We also provide some preliminary analyses of the classification approaches themselves, in an attempt to learn more about the use of TQs.

6.1.1 Tag questions (TQs) and the difficulty for MT

TQs are common constructions in spoken English, formed of a main (or *host*) clause, which is typically declarative, followed by a peripheral interrogative element, known as the *question tag*. Two examples of TQs are given in (41) and (42). Note that we will distinguish the terms *tag question* (TQ), referring to the entire sentence (including the anchor clause, indicated in italics), and *question tag* (indicated in bold).

(41) *You don't want that, right?*

(42) *You're not panicking, are you?*

TQs have a variety of complex communicative functions. As a type of interrogative, the question tag has the effect of inviting the interlocutor's attention and a response. Soars and Soars (2000) describe this as a way of "keeping the conversation going" and "inviting listeners to communicate". However, the communicative impact is more than just encouraging the other speaker to reply (although this can sometimes be the case). In many cases, the communicative function of the TQ would be lost if the question tag were removed (or if a paraphrase representing the same communicative function were not used). The function of TQs are complex and communicate a range of information about speaker attitude, tone, the relationship between dialogue participants, common ground and dialogue flow (McGregor, 1995), all of which are important to preserve when translating into another language. Holmes (1983, 1984, 1995) considers tags to be *hedges*, softening or avoidance strategies, expressing doubt about something that is otherwise assumed. To illustrate this, we provide possible analytic glosses for the communicative function of four TQs in Examples (43-46).²

(43) *You are going to the party, aren't you?*

'I believe that you are going to the party, or would like you to be going to the party and wish for you to confirm that this is indeed the case.'

(44) *You're not from New York, right?*

'I suspect that you are not from New York, but am not certain (and would like confirmation)'

or 'I would like to know more about where you are from.'

²The following examples are adapted from attested examples in OPENSUBTITLES2016 (Lison and Tiedemann, 2016).

(45) *Not very nice, is it.*³

‘Now that you have experienced the same thing, you can see that it is not nice when such a thing happens to you. So what do you say now?’

(46) *So you think you’re the boss, do you?*

‘You are acting out of place (above your station), so I challenge you to say that you think you have the authority to act that way.’

There are two main classes of TQ: the *grammatical TQ* and the *lexical TQ*. In its canonical form, the English question tag is of the grammatical type, formed of an auxiliary verb, which can be negated, followed by a pronoun. Two such examples are given in Examples (47) and (48), in which the question tag is shown in bold and the anchor clause in italics. The question tag’s auxiliary verb and pronoun typically parallel the verb and subject of the anchor clause (underlined), the verb in terms of tense, and the pronoun in terms of number, gender and person.

(47) *You do believe in happy endings, don’t you?*

(48) *He can’t do that, can he?*

The grammatical structure of TQs and agreement between the anchor clause and the question tag is what gives these types of TQs the name *grammatical TQs*. Although in theory their form is systematic, their attested usage is in fact more complex. For example, the subject and verb of the main clause can typically be elided, as in Example (49).⁴ It is also possible to have imperative rather than declarative anchor clauses, as in (50), non-canonical uses of morphology as in (51) and a change of interlocutor mid-utterance as in (52).

(49) \emptyset *So typical, isn’t it?*

\emptyset *So clumsy, aren’t I?*

(50) *Let’s roll back the clock, shall we?*

(51) *Everything just falls apart, don’t it?*

(52) *Oh, he loves it, don’t you, George?*

³The intonation can give an indication of the speaker’s belief about the probability of the anchor clause’s veracity (rising tone indicating more doubt). In texts, this can be represented typographically by a distinction between a question mark (rising tone) and a full stop (falling tone). In this thesis, we do not attempt to predict this distinction, and make the simplification that all predicted question tags are followed by a question mark.

⁴This leaves the question tag as the only indicator of the subject of the sentence. In these examples, the subject and verb of the anchor clause are marked by the symbol \emptyset . The use of the form *are* with the subject *I*, although not strictly grammatically correct, is widespread and accepted in questions with subject-auxiliary inversion.

The second type of TQ, the *lexical TQ*, is formed of a word or phrase that is invariant to the subject and verb of the anchor clause, as in (53) and (54). In (53), the question tag *innit* is a lexicalised version of the grammatical question tag *isn't it*, which has become invariant, and as a result can be used after any declarative clause as an interrogative marker. In (54), the question tag is formed of a word whose meaning is veridical in nature.

(53) *He's a proper bad man, innit?*

(54) *There's got to be a cure, right?*

Cross-linguistic use of tag questions Lexical TQs are common cross-linguistically (Axelsson, 2011). Many languages have some form of lexical TQ, as a means of “interrogativising” utterances, often using veridical words for example, German *nicht wahr* ‘not true’, *nicht* ‘no’ and *stimmt’s* ‘is it true’, French *non* ‘no’, *pas vrai* ‘not true’ and Czech *ne* ‘no’ and *ano* ‘yes’. Other types of expressions commonly used are those putting doubt on the previous clause (German *oder* ‘or’), and those appealing to the perception of the interlocutor (English *don't you think*, French *tu vois* ‘you see’ and Czech *vidte* ‘you_{PL/POLITE} see’.)

However, few languages have such a systematic use of TQs in general, and grammatical TQs are often seen as very particular to English, especially to British English.⁵ For MT, the situation is especially complex when translating into English. There are two main difficulties. The first involves deciding whether or not to translate using a question tag, especially when there is no such question tag in the source sentence. This is similar to the problem faced by translating implicit discourse relations into a language with more explicit marking of discourse connectives (Meyer and Webber, 2013). The second difficulty concerns the correct choice of form of TQs, in particular for grammatical TQs. As in anaphor resolution, the form of grammatical question tags must be grammatically cohesive with the rest of the translation; the form of the pronoun and the auxiliary are dependent on the main clause. As with pronoun prediction, evaluation is complicated, because the correct tag form is dependent on the rest of the translation, and may not have the same form as the reference translation. Consider the German sentence *Sie lebt noch, nicht wahr?* and its English translation *She's alive, isn't she?*. The choice of the question tag *isn't she?* is dependent on the subject and verb of the anchor clause. Had the translation been *She still lives*, the correct question tag would have been *doesn't she?*

⁵The frequency of usage of different types of TQ is dependent on the language dialect and the level of formality. Tottie and Hoffmann (2006) find that grammatical TQs are more frequent in British English than American English (up to nine times more frequent). They also find that the age of the speaker also has an impact on the frequency of usage, younger speakers tending to use fewer grammatical TQs and more lexical TQs than older speakers.

Preparing our experiments: corpus annotation for English TQs If we want to improve the translation of TQs when translating into English, it is first useful to annotate English TQs in parallel data. This will give us a better idea of their distribution, and also serve to create training data to be used in our post-edition approach to improving their translation (Section 6.1.2). We decide to base our annotations on film subtitles from the OPENSUBTITLES2016 parallel corpus, due to the large quantities of available data and the fact that many of the films represent a more informal and speech-like style of language. We automatically annotate each subtitle of the English side of the corpus for the presence/absence of a TQ, and where applicable the question tag form used (*right, ok, isn't it, are you, etc.*).

We first automatically clean the subtitles using heuristics. Some of the subtitles have been extracted from the subtitle image of films using optical character recognition (OCR), which is sometimes imperfect. Our heuristic therefore include identifying common OCR (optical character recognition) errors and correcting encoding problems. To help our identification of TQs, we first tokenise the texts, truecase them to remove sentence-initial capital letters, tokenise them using the MElt tokeniser (Denis and Sagot, 2012) and truecase them using the MOSES truecaser (Koehn et al., 2007). We detect if a sentence contains a TQ, and if it does which question tag form is present (e.g. *is it?, right?*), by applying robust, manually defined lexical rules, which are briefly described below.⁶ It is important for these rules to have a high coverage but also a high precision; it is easy to over-detect TQs with rules that are too lax, since question tag forms can frequently appear within sentences that are not TQs (cf. our manual evaluation below). Although TQs can appear mid-sentence, we choose to only detect utterance-final TQs, and only those whose anchor clause is within the same sentence. This is important to distinguish TQs from *echo questions* (e.g. *Are you?*), which have the same form as question tags, but are questions in their own right, mirroring previous utterances.

For grammatical TQs, our rules are designed to detect patterns of auxiliary verbs (*am, are, may, shall, etc.*), pronouns (*I, s/he, there, etc.*) and negative elements (*n't, not, nae, etc.*), that appear in the correct environment. Possible verbs, pronouns and negative elements are enumerated, and the correct environment is determined mainly by the presence of punctuation (such as a following question mark or full stop) and the presence of the end of the sentence.⁷ For lexical TQs, which do not exhibit a systematic internal structure, our solution is to enumerate as many lexical TQs as possible, inspired by the linguistics

⁶The complete list of rules (regular expressions) can be found in the code available online at <https://github.com/rbawden/tag-questions>.

⁷For example one of our simplest rules (to detect positive grammatical TQs) is to detect whether a sentence contains a final auxiliary verb followed by a grammatically acceptable pronoun, optionally followed by punctuation marks and then necessarily followed by the end of the sentence. Excluded from the previous rule are those candidate TQs whose anchor appears in a list of identified exceptions (*yes, no, ok, etc.*) or which do not have an anchor clause at all.

literature, native intuitions and a corpus analysis of OPENSUBTITLES2016.⁸

A manual evaluation on a random subset of 500 grammatical TQs and 500 lexical TQs shows that TQ detection is near perfect (accuracy of $\approx 98\%$ for all TQs and recall of 100% on sentence-final grammatical TQs whose anchor clause is in the same subtitle). Recall for lexical TQs cannot be accurately measured, as identification relies on a closed list of forms, and any recall measure would therefore be biased.

The parallel corpora we annotated this way are of varying sizes depending on the language pair: 15.1M sentences for CS \rightarrow EN, 6.2M sentences for DE \rightarrow EN and 15.1M sentences for FR \rightarrow EN. We divide each parallel corpus into three sets: TRAIN (2/3), DEV (1/6) and TEST (1/6). The frequency and distribution of TQs is shown in Table 6.1. TQs make up approximately 1% of subtitles, which makes them a relatively rare phenomenon, but they are common among questions ($\approx 20\%$). There are between 270 and 315 distinct English question tags, depending on the language pair. The most frequent question tag is *right?* ($\approx 20\%$), followed by *ok?* ($\approx 16\%$). The distribution is Zipfian (See Figure 6.1), and the majority of labels are grammatical question tags, the most frequent grammatical tag (*isn't it?*) representing only $\approx 3\%$ of all TQs, revealing the huge class imbalance that exists between the different question tags. We include a label *none* for those sentences that are not considered to be TQs (NONE).

	#sents.	#English TQs			#labels	
		ALL	GRAM	LEX	GRAM	LEX
CS \rightarrow EN	15.1M	169,300	51,740	117,560	315	20
DE \rightarrow EN	6.2M	60,330	21,291	39,039	270	17
FR \rightarrow EN	15.1M	149,847	44,651	105,196	276	19

Table 6.1: TQ distribution for each language pair.

6.1.2 Improving TQ generation in MT into English: our post-edition approach

We consider it important to seek to improve the translation of all aspects of a translation, including those linked to style and attitude, such as TQs. Our aim is therefore to study whether it is possible to improve the generation of TQs when translating into English, such that the MT outputs resemble the utterances of a native English speaker with respect to the use of TQs. We would expect them to be used in the correct circumstances, and with a grammatically and pragmatically acceptable question tag form.

⁸ The list of lexical question tags we use (with some leeway for spelling variations) is the following: *innit*, *right*, *eh*, *see*, *remember*, *you know*, *yer know*, *remember*, *or what*, *yeah*, *aye*, *you see*, *like*, *ok okay*, *don't you think*, *don't yer think*, *correct*, *all right*, *alright*.

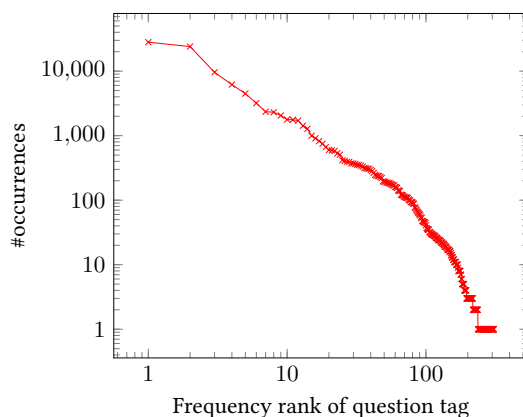


Figure 6.1: The Zipfian distribution of question tags in the CS→EN training corpus (here represented using a log-log scale).

We choose to use a post-edition strategy, through which we modify MT outputs depending on predictions made about whether the translated sentence should be a TQ, and if so which question tag it should contain. Making predictions in a post-processing step (once MT has been performed) gives the advantage of providing access to all the source and target translations (including of the previous and following sentences) when making a prediction. However, it has the inevitable disadvantage of being dependent on the MT baseline output, however incorrect the translation may be.

We test several strategies for the prediction of which question tag form to use for a given source sentence and baseline translation. Possible labels are the different question tag forms (e.g. *isn't it*, *ok*, etc.) and a label indicating that the translation is not a TQ at all (NONE). Predicted tags are either used to replace the question tag already present in the MT output or appended to it if one is not present. Presenting this task as a classification task whereby a single label is considered correct for a given TQ is of course a huge simplification; it is often the case that several possible question tag forms are acceptable for a given communicative function. It is therefore possible for there to be several correct question tags. We nevertheless retain this simplification and rely partly on statistical classifiers to attempt to uncover statistical patterns in the data.

6.1.2.1 Question tag classification

As can be seen in Table 6.1, there are a large number of possible classes, if we are to treat the problem as a classification problem. Moreover, the distribution of classes is very unequal, a majority of sentences being non-TQs (99%), and the tag question labels being distributed very unevenly too. This can be problematic for statistical classification, on which we nevertheless want to rely due to its capacity to find patterns in the data

that are not necessarily simple.⁹ We therefore propose three classification strategies, of increasing complexity, designed to address this label distribution problem. The strategies are illustrated schematically in Figure 6.2 on page 127.

CL1: This statistical classifier is trained to directly predict the question tag form (or the label `NONE`) (See Figure 6.2a). It is therefore a multi-class classifier with approximately 300 different classes (depending on the language pair). Given the unequal distribution of labels, we expect this strategy to struggle to predict the variety of question tag forms that we would expect from human-produced texts.

CL2: This strategy uses two classifiers, the second one introduced to specifically handle the prediction of the numerous and often sparsely distributed grammatical question tags (See Figure 6.2b):

1. a first statistical classifier ($CL2_{INIT2}$) to predict the question tag forms of lexical TQs, a label `NONE` for non-TQs and a single label `GRAM` for all grammatical TQs.
2. a second rule-based classifier ($CL2_{GRAM}$) to predict the form of grammatical question tags (as predicted by $CL2_{INIT2}$), based on heuristics. If a form cannot be found simply, then the classifier predicts the label `NONE`.

CL3: This third strategy continues the idea of sequentially predicting question tag forms introduced in CL2 by further decomposing the prediction process into three classifiers (See Figure 6.2c):

1. a first statistical classifier ($CL3_{INIT3}$) to predict coarse-grained labels `LEX` and `GRAM` (regrouping all lexical and grammatical TQs respectively) and a label `NONE` for non-TQs
2. a second statistical classifier ($CL3_{LEX}$) to predict the forms of those sentences classified as lexical TQs by $CL3_{INIT3}$. To account for potential errors produced in the first step, sentences can also be classified as either `GRAM` and `NONE`.
3. a third rule-based classifier ($CL3_{GRAM}$, identical to $CL2_{GRAM}$) to predict the question tag forms of sentences classified as `GRAM` by the first and second classifiers. As before, this classifier can decide to predict `NONE` if no form is easily found.

The decision to use statistical classifiers to predict coarse-grained labels `GRAM`, `LEX` and

⁹We also envisaged a classification approach relying on neural networks, but chose instead to rely a simpler and more interpretable approach for our initial experiments. As we shall see in the rest of this chapter, the issue of how to evaluate is one that is fundamentally more important than the choice of architecture (as it is a prerequisite for any experimentation).

NONE and lexical question tag forms is motivated by the fact that it is difficult to devise clear rules to define when to use one class or form rather than another. We therefore use statistical classifiers in an aim to simulate the distribution of natural speech (as represented by the training data). The choice to use a rule-based approach for the grammatical question tags is based on the systematic and highly predictable nature of grammatical question tag forms based on the anchor clause¹⁰ and by the fact that there are a very large number of possible combinations of grammatical tag elements, of which certain are very sparsely represented in the data. Using a rule-based approach even enables us to predict grammatical tags that do not appear in the training data. More details about the types of features used for statistical classification, the statistical system setup and the rule-based classification are given below.

“Baseline” MT outputs For Czech and German, we use the top systems at WMT16, both attentional encoder-decoder NMT models (Sennrich et al., 2016b). For French, we trained a phrase-based model with MOSES (Koehn et al., 2007), as there was no pre-trained NMT system for French-to-English available at the time.¹¹ Baseline predictions are automatically extracted from the MT outputs using the English TQ identification rules that we developed for corpus annotation. These same baseline MT outputs are used as the basis for post-edition; the new predictions from the classification system are used to update the baseline translations.

Statistical experimental setup All statistical classifiers are linear classifiers trained using Vowpal Wabbit (Langford et al., 2009).¹² The learning strategy is a linear regression, with squared error loss, generalised to multi-class classification using the *one against all* strategy.¹³ To account for the huge class imbalance, during training we weight examples based on the relative frequency of their associated classes based on the training data. The degree to which weighting is used is optimised on the DEV set; we vary weights from equal for all examples (i.e. no weighting) to weights that fully counterbalance the class distribution.

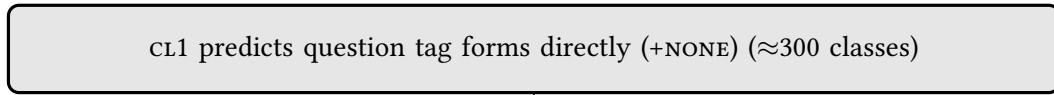
Statistical features All statistical classifiers are trained on the same set of features. We use both automatically and manually defined lexical feature templates, in an attempt

¹⁰As long as the subject and verb are identified (which is not always very simple when the anchor clause is complex).

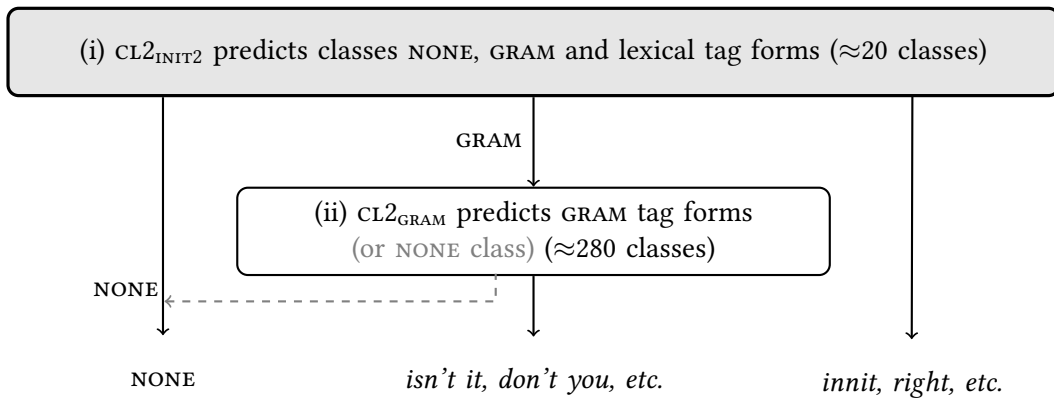
¹¹We use a combination of three phrase tables and three 4-gram KenLM language models (Heafield et al., 2013), trained on EUROPARL, TED Talks and 3M-sentence subtitles, tuned using KBMIRA on a disjoint 2.5K-sentence subset.

¹²<http://hunch.net/~vw/>

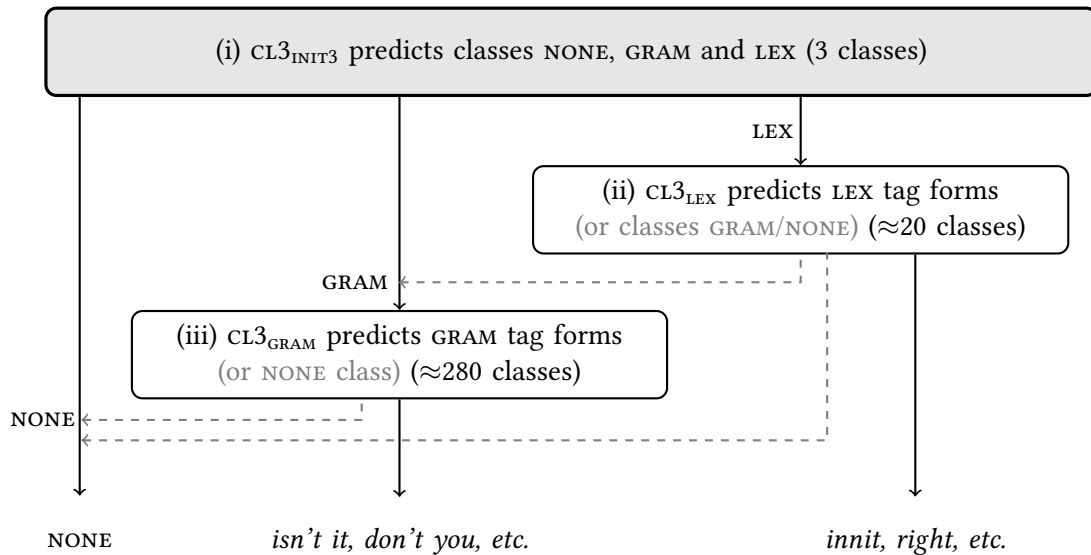
¹³We use FTRL-proximal optimisation, L2 regularisation ($\lambda = 10^{-6}$), quadratic features and a bit precision of 24.



(a) CL1 classification: direct prediction of question tag forms (+NONE).



(b) CL2 classification: (i) NONE and GRAM classes and lexical question tag forms, (ii) grammatical tag forms.



(c) CL3 classification: (i) three coarse-grained classes (NONE, GRAM and LEX), (ii) lexical tag question forms, and (iii) grammatical tag question forms.

Figure 6.2: Our three classification strategies of increasing complexity: CL1, CL2 and CL3. Each strategy introduces an additional classifier to deal with the prediction of forms of certain classes of question tag separately.

to provide a sufficiently diverse range of pertinent information. Unless indicated, the features templates are applied to both the source sentence and the baseline MT output.

- The first set of features are automatically identified bag-of-words, which represent the 500 uni-, bi- and tri-grams most associated with the presence of a TQ (as opposed to a non-TQ), as calculated using a g-test statistical significance test.
- The second set of features are manually defined, based on language-specific question-response patterns and recognisable lexical clues. They include:
 - the presence of a question tag (and its form), detected through lists of language-specific lexical tag forms,
 - the presence of a final question mark,
 - the following subtitle contains a specific response (from a predefined list of replies such as *OK*, *yes*, *no*, etc.),
 - the first words of the MT output (1–4 gram), the last auxiliary, the last pronoun and the last pronoun–auxiliary pair,
 - (Czech and German only) whether the following subtitle contains a verb that appears in the current subtitle (and if so, we include as a feature the verb type and the preceding word in both the current and following subtitles).¹⁴

Rule-based grammatical TQ prediction Our rule-based approach is designed to predict which grammatical tag should be appended to a given translation, for which the label GRAM was assigned in previous classification steps. Based on the MT output alone, we attempt to automatically detect the pronoun and appropriate auxiliary that could be used in the question tag, using simple heuristics. The rules consist in the identification of certain lexical cues from the translated anchor clause. For instance, utterance-initial words can be a good indicator of the use of a particular question tag: imperatives such as *let's ...* (indicative of the tag *shall we*), and claims about the interlocutor's perception such as *you think...* or *you know...* (indicative of the tag *don't you*). When there is a single auxiliary and subject, these are directly used to construct a question tag, using, as a simplification, the opposite polarity to that of the anchor clause, which is the most common polarity pattern in TQs. We include several rules to account for complex clauses and perform grammatical checking between the subject and auxiliary of the question tag.

¹⁴In German and Czech, it is common for a reply to a yes/no question to repeat the verb of the question, e.g. *Poslala jsi mu to?* 'Did you send it to him?' – *Poslala jsem* 'Yes, I did' (lit. 'send (I)_did') (Gruet-Skrabalova, 2013). We added these features to help distinguish classical yes/no questions from potential TQs.

6.1.3 Results, analysis and discussion

As discussed in Section 2.2.2, evaluating cohesion-related MT phenomena can be problematic. A question tag can be a correct choice without matching the question tag form in the reference translation, making traditional metrics involving lexical comparison (including all standard MT evaluation metrics) ill-adapted to the task. This is all the more relevant given that we choose to apply post-editing techniques to translation machine outputs, making an automatic comparison with the reference question tags unsuitable. To illustrate this difficulty, we provide a range of different ways to evaluate how well our systems handle the generation of question tags, and compare their suitability (Section 6.1.3.1). We also provide an analysis of the classification models in Section 6.1.3.2 to see what can be learnt from these strategies.

6.1.3.1 Evaluation

To evaluate the different approaches, we look at the results of each of our classification strategies as compared to the baseline using (i) traditional metrics (precision, recall, F-score), (ii) a manual comparison of a subset of predictions and (iii) a measure of the divergence of the distribution of question tag predictions with respect to the distribution in reference translations. We aim to show that the metrics can be misleading when taken out of context, and are all insufficient to varying degrees for evaluating the translation quality of English TQs.

Traditional evaluation metrics In Table 6.2, we provide results using traditional metrics (precision, recall and F-score), which we will compare with a manual evaluation. In an attempt to overcome the problem of comparing exact predicted labels to those in the reference translation, we calculate scores based on the labels' coarse-grained category (GRAM, LEX and NONE), obtained by mapping each predicted label to the associated category. We hope to be able to see certain trends as to the prediction within these categories. All scores must be viewed together to get a global picture. We calculate precision (P), recall (R), F-score (F) for each of these three coarse-grained classes, without looking at the exact question tag forms used. Within each coarse-grained class, we also provide labelling precision (P*), corresponding to the percentage of question tags within that coarse-grained class for which the correct question tag form was predicted, with respect to the tag used in the reference translation. Labelling precision is also given overall (for all test sentences and all coarse-grained classes) in the final column.

Overall labelling precision (Overall P*) is significantly improved for all language pairs when using CL2 and CL3 over both the baseline (the non-post-edited MT outputs) and CL1. The strategy of separating the classification process into multiple classifiers appears

Lang. pair	Gram TQs					Lex TQs					Non-TQs					Overall	
	P	R	F	P*	F	P	R	F	P*	F	P	R	F	P*	F	P*	
CS→EN	baseline	54.91	45.66	49.86	36.02	53.79	57.61	55.63	43.36	99.64	99.65	99.64	99.65	99.64	99.64	99.00	
	CL1	67.31	13.57	22.59	50.0	61.7	39.00	47.79	51.37	99.32	99.88	99.60	99.88	99.60	99.05	99.05	
	CL2	60.62	37.61	46.42	41.03	68.75	43.89	53.58	57.23	99.44	99.85	99.65	99.85	99.65	99.11	99.11	
	CL3	58.97	40.53	48.04	39.40	68.21	47.44	55.96	56.23	99.49	99.83	99.66	99.83	99.66	99.12	99.12	
DE→EN	baseline	51.65	28.00	36.32	9.97	77.20	44.67	56.59	66.59	99.43	99.88	99.65	99.88	99.65	99.13	99.13	
	CL1	72.53	1.88	3.66	54.95	75.44	18.90	30.22	64.3	99.11	99.97	99.54	99.97	99.54	99.05	99.05	
	CL2	60.67	44.02	51.02	31.57	80.5	43.56	56.53	72.08	99.48	99.9	99.69	99.9	99.69	99.21	99.21	
	CL3	57.33	50.09	53.46	29.22	80.80	44.13	57.08	72.11	99.52	99.89	99.7	99.89	99.7	99.21	99.21	
FR→EN	baseline	41.15	47.18	43.96	12.95	57.63	38.52	46.18	52.55	99.53	99.72	99.62	99.72	99.62	99.03	99.03	
	CL1	66.30	9.36	16.41	44.87	55.05	28.80	37.81	51.73	99.32	99.89	99.60	99.89	99.60	99.12	99.12	
	CL2	58.74	35.46	44.22	38.61	63.99	37.01	46.9	60.52	99.45	99.86	99.65	99.86	99.65	99.19	99.19	
	CL3	59.40	35.21	44.21	38.87	63.07	39.39	48.50	59.66	99.47	99.85	99.66	99.85	99.66	99.20	99.20	

Table 6.2: Precision (P), Recall (R), F-score (F) and fine-grained labelling precision (P*) for the TEST set on each language pair. Results are given for each coarse-grained TQ class (GRAM, LEX and NONE). Labelling precision is calculated on the subtitles with the corresponding predicted coarse-grained label. Marked in green are the cells containing the best F-scores for coarse-grained label groupings and the overall labelling precision (for fine-grained classes).

Lang.	Source	Reference	Baseline	CL1	CL2	CL3
FR	Les parents sont parfois si autoritaires, pas vrai ?	Parents can be so bossy sometimes can't they?	The parents are sometimes so bossy, isn't it? (X)	none (X)	right (✓)	right (✓)
DE	Rufst du uns bitte zurück, wenn du das abhörst?	Would you just please give us a call when you get this?	Call us when you get this message, ok? (✓)	ok (✓)	none (X)	none (X)
FR	Mm, ça fait toujours mal, n'est-ce pas, héros ?	Still stings, doesn't it?	Does it still hurt, right? (X)	right (X)	does it (X)	does it (X)
CS	Královni jezdci jsou venkovské síly, ne?	The Queen's Rangers are provincial forces, aren't they? (✓)	The king's riders are rural power, aren't they? (✓)	none (X)	right (✓)	none (X)

Table 6.3: Examples from the manual comparison of the system predictions (including the baseline MT output). We indicate when each of the systems produces a question tag that can be considered correct given the other translation choices made in the MT output (correct=✓ and incorrect=X). There can be several correct predictions for a given sentence.

to be validated by the fact that CL2 and CL3 have the highest overall labelling precision and, apart from CS→EN, have the highest F-score per coarse-grained category. The high overall precision is mainly due to a prediction of non-TQs that better matches the reference translations, represented by the high corresponding F-scores for CL3 for all three language pairs.

For grammatical and lexical TQs, the recall of CL2 and CL3 is systematically higher than that of CL1. CL1 suffers from the sparsity of the question tag labels and predicts few TQs with respect to our multi-tier strategy of predicting grammatical and lexical tags separately. For CL1, this results in high precision but very low recall (as low as 1.88% for DE→EN, the language direction with the least training data). The drop in recall between CL3 and CL1 is much higher for grammatical TQs than lexical TQs. This is due to the huge class imbalance in the different question tags (270 grammatical tags vs. 17 lexical tags), which causes the purely statistical one-pass system to favour the more frequent lexical tags and struggle to predict the wide range of rarer grammatical tags. This does suggest that our two sequential classifiers (CL2 and CL3) can produce more grammatical TQs, aided most likely by the initial prediction of coarse-grained categories. Globally, the automatic metrics show that CL3 gives marginally higher results than CL2, aided possibly by the extra division of the task into three classifiers rather than two.

Confusion matrices of the predictions To give another perspective of some of the differences between the predictions of our systems compared to the baseline translations, we show the distribution of the labels among the three coarse-grained classes in Table 6.4. A perfect categorisation according to the reference translation (gold labels) would be all occurrences concentrated along the diagonal. Here we only show the results of the third system, CL3.

		predicted (baseline)			predicted (CL3)			
		GRAM	LEX	NONE	GRAM	LEX	NONE	
CS→EN	gold	GRAM	3687	2722	1916	3268	1422	3635
		LEX	1399	11459	6965	1242	9450	9131
		NONE	1751	7083	2484366	1113	3011	2489076
DE→EN	gold	GRAM	957	209	2245	1750	126	1535
		LEX	459	3221	3625	781	3178	3346
		NONE	487	739	1018590	629	539	1018648
FR→EN	gold	GRAM	3181	715	2981	2430	1021	3426
		LEX	1952	6829	8807	853	6941	9794
		NONE	2828	4242	2487538	860	2949	2490799

Table 6.4: Confusion matrix of predicted versus gold tags (when question tags are grouped into their three coarse-grained classes) for baseline models and CL3 models for each language pair. The majority label is marked in green.

The results vary depending on the language pair. A notable difference is the fact that our model often predicts more non-TQs than the baseline model. The most common classification errors by our model are therefore over-predicting non-TQs when one was expected. The effect is most evident for our highest resource language direction, CS→EN, for which there is a huge increase in the number of non-TQs predicted with respect to the baseline predictions. In light of the results in Table 6.2, this appears to show that our high results according to traditional metrics are due to a high prediction of non-TQs. According to these three confusion matrices, the distribution of question tags when compared to the reference translations appears to be often better according to the baseline model than when our classification system is used for both CS→EN and FR→EN. The situation is slightly different for the language direction DE→EN, where the predictions of CL3 are slightly more in line with the labels of the reference translations, predicting a higher number of grammatical tags.

Manual analysis The surest way of evaluating the predictions is to perform a manual analysis. For each language, we therefore manually evaluate a subset of examples in order to compare the different predictions. We randomly select 150 examples for which the baseline prediction is different from the predictions of CL3. For each example, we note which of the predictions (of the baseline, CL1, CL2 and CL3) are correct TQs with respect to how the sentence is translated in the baseline translation, as shown by the three examples in Table 6.3. Note that several different predictions may be correct for a single translation. For those translations which are either too poorly to make TQ prediction meaningful, we label the translation as *odd* and do not provide a comparison of the different predictions.

	#correct predictions				#odd
	baseline	CL1	CL2	CL3	
CS→EN	124	48	55	53	3
DE→EN	21	15	69	78	10
FR→EN	71	51	69	71	18

Table 6.5: Results of the manual analysis of 150 predictions, showing the number of acceptable predictions for each system type. The last column indicates the number of translations that were too strange for a TQ prediction to be made. The first, second and third best scores are marked with increasingly light shades of green.

The results of the manual analysis show an interesting difference from the automatic analysis (Table 6.5). For DE→EN, the models CL1, CL2 and CL3 increasingly improve on the baseline system, which only produces an adequate question tag in 14% of the sentences selected, compared to approximately 50% for CL3. However, for the other two language pairs, the post-edition approaches show at best an equally good prediction (for FR→EN) and at worst, a degradation of results (for CS→EN). The best translations are seen for the CS→EN model, which is an NMT model trained in a high resource setting.

The higher quality can be seen by the relatively low number of odd translations in our sample, compared to the other two language directions. The DE→EN model, although neural, is trained on little data in comparison, and therefore has a much lower translation quality, particularly for question tags. The FR→EN SMT model has a reasonably high score for question tags, but a relatively high number of odd translations overall. These results appear to suggest that our classification approach helps in settings where the basic translation system is of a lower quality, but does not help in high-resource NMT situations such as for CS→EN.

An important observation from this manual analysis is that developing a strategy that appears to produce better question tags than the baseline translation is in fact trivial, if we use two simple rules: (i) always predicting a question tag when one appears in the baseline (to ensure that this part of the prediction is at least as good as the baseline), and (ii) using ‘safe’ and very frequent question tags, that are likely to be acceptable in most cases. In the majority of cases where a frequent lexical tag such as *right* or *ok* is predicted, it is deemed acceptable according to the manual analysis, and choosing a lexical tag over a grammatical tag has the advantage of not having to ensure the grammatical agreement of the form with the anchor. However, this type of prediction does not satisfy our aim to simulate natural use of TQs; it does not necessarily produce very satisfactory translations if there are several TQs close to each other, and the same question tag is systematically used. Our classification strategies do not follow these two rules, and are therefore at a disadvantage. They under-predict the presence of TQs, which is in part due to the insufficiencies of our rule-based approach in assigning a grammatical question tag form. The use of a statistical classifier trained on the reference data means that their aim is also to reproduce the distribution of question tags seen in the original English data.

Distribution of question tags If the aim of MT is to produce translations that resemble as best as possible translations produced by humans, one way in which we can evaluate our system is to look at the distribution of question tags in the MT outputs. Ideally, the distribution among the different types of question tags should be as close as possible as to those in the reference translations. For those sentences assigned a question tags, we compare the distributions of tag forms for each of the four MT outputs (baseline, CL1, CL2 and CL3) to that of the reference translations (Table 6.5).

The numbers of TQs predicted are different for each of the sets of predictions. CL1 systematically produces fewer TQs than the baseline and the other two classification systems. CL3 produces slightly more TQs overall than CL2, but all produce fewer TQs than the baseline MT outputs. Amongst those sentences marked as TQs by each of our systems, we compare how the question tags are distributed amongst the different possible forms and compare the distribution to the distribution in the reference translations. We measure the divergence of each distribution from the gold distribution between by calculating the

Kullback-Leibler (KL) divergence (Kullback and Leibler, 1951) between the probability distributions amongst the potential question tags.

	KL divergence from gold distribution			
	baseline	CL1	CL2	CL3
CS→EN	0.26	2.85	1.82	1.73
DE→EN	1.40	1.72	0.99	1.02
FR→EN	0.52	1.58	0.08	0.27

Table 6.6: The Kullback-Leibler (KL) divergence between the question tag distributions produced by each system and the question tag distribution of the reference translations. The first, second and third lowest divergences are marked in increasingly light shades of green.

The results suggest that for DE→EN and FR→EN, our CL2 and CL3 systems improve the distribution of question tags that are assigned, with slightly lower KL divergence than the baseline model (0.99 for CL2 vs. 1.40 for DE→EN and 0.52 for CL2 vs. 0.52 for FR→EN). CL1 shows a very poor distribution when compared to the reference distribution, systematically giving a more divergent distribution than the baseline distribution. Again, we see that the language direction CS→EN shows poorer results for our classification models, the baseline distribution being far more similar to the reference distribution (KL score of 0.26 compared to 1.73 for the model CL3). The smaller the training data available, the more the systems (including the baseline) tend to over-predict the more common question tags, such as *right* and *ok* with respect to the reference translations. For example, for DE→EN, *right* makes up 69% of question tags predicted by CL1 and 41% of those in the baseline translations (compared to 27% in the gold translations). This falls to 47% and 32% respectively for CS→EN. The large data size of CS→EN coupled with the fact that the baseline is trained with a neural MT model could explain why the baseline predictions are more adapted for this language direction.

Conclusion on evaluation metrics Two types of conclusion can be made from the results shown in this section.

The first concerns the adequacy of the evaluation metrics to judge whether improvements have been made and which system produces the best predictions. As they stand, none of the evaluation methods appear to be adequate by themselves to judge prediction quality. Traditional metrics appear to be the most unreliable of all, as is expected. The imbalance of the different classes, in particular with respect to the label NONE for non-TQ, mean that systems that over-predict the dominant label are unduly rewarded. Our manual analysis shows that this is the case. However the manual analysis too shows a bias to predicting ‘safe’ labels, rather than using the range of different question tags available. The analysis of the question tag distributions enable us to see whether the systems are close or not to natural use of English in real texts. Using manual analysis in conjunction

with an analysis of distributions does not provide us with a single evaluation score that can be easily interpreted, but does provide us with more meaningful information about the adequacy of predictions.

In light of this, the second conclusion concerns the evaluation of the systems themselves: which strategy performs best? Other than the fact that cL1 performs poorly for all language pairs, this appears to depend largely on the language pair tested, on the quantity of data available and on the quality of the MT system used. For example, the language pair with the most available data and the use of an NMT system of high quality is CS→EN, for which our strategies do not improve on the baseline output, and even degrade it. This can be seen by both the manual analysis and the more divergent question tag distribution compared to the baseline predictions. The system for the second most highly resourced language pair, FR→EN, is a phrase-based MT system. Our classification systems appear to equal the quality of the baseline predictions according to the manual analysis, but are far better in terms of the question tag distribution. Finally, our multi-classifier systems appear to greatly outperform the baseline predictions for our lower-resource language direction, DE→EN, in terms of the manual analysis and the label distribution.

6.1.3.2 Can we learn anything from the models?

One of the aims of these experiments was to learn more about the situations in which TQs occur, and how to predict when to translate using a TQ when the source sentence is not one itself. Given the variable quality of our classification systems, the analysis can only be limited. However, we provide some preliminary analyses for these two points here, by analysing some aspects of the models and of the predictions themselves.

Which contextual features are most useful? A first indication of which contextual features may be useful can be found in the list of n -grams that were found to be most indicative of the presence of a TQ as identified by the statistical g -test (see the description of statistical features above). Unsurprisingly, the most prominent n -grams are those containing question tags in the source language, for example for Czech, the unigrams [*vid'*], [*vid'te*] and [*žejo*], the bigrams [*že ?*], [*ne ?*], [*jo ?*], [*že jo*], etc. and the trigrams [, *že ?*], [, *ne ?*] and [, *dobře ?*], etc. The significant n -grams found in the MT outputs are similarly question tags such as [, *right ?*], [, *ok ?*], [*do you ?*]. Interestingly, the lists also include a number of sentence-initial adjectives in the English (target) sentences, such as *Crazy*, *Ironic* and *Exciting*, which is evocative of TQs such as *Crazy, isn't it?*. The lists quickly show signs of overfitting and therefore do not reveal any other obvious patterns that are indicative of a TQ.

How often is an English TQ generated from a non-TQ in the source language?

Since TQs are commonly associated with English and less so with most other languages, we would expect there to be a relatively high number of parallel sentences for which the source sentence is not a TQ but the English target sentence is. We can also expect this effect to be slightly reduced in the OPENSUBTITLES2016 dataset, due to the fact that many of the film subtitles are originally written in English and many of the foreign subtitles could be largely inspired by the English subtitle and therefore more likely to adopt a similar formulation. We can expect more TQs in the non-English subtitles than would be expected in independently produced texts.¹⁵ In spite of this, it is nevertheless the case that many of the English TQs are not TQs in the source language. For example, this is the case for 2,811 out of 10,708 TQs in the DE→EN test set, 6,070 out of 24,286 TQs in the CS→EN test set and 9,257 out of 24,453 TQs in the FR→EN test set.

Are our models capable of predicting the presence of a TQ even when one is not present in the source language? We calculate the number of times this occurs in the test sets for each language. Out of our three classifiers, CL3 predicts the most question tags in this setting: 5 for DE→EN, 13,189 for CS→EN and 1,796 for FR→EN. The most frequent predictions correspond to the most common tags (*right, you know, ok, etc.*), and their distribution is representative of the distribution over the entire test set. However, the baseline translated using a question tag for all of these sentences, and therefore the classifiers do not provide any more information than provided by the MT model. The very low number of such examples for DE→EN is most likely a reflection of the classifier's dependence on the presence of a TQ in the German source sentence, exacerbated by the small quantity of training data on which to learn more complicated patterns.

6.1.4 Conclusion to our tag-question experiments

Focusing on the particular aspect of generating English TQs in MT, has led us to explore the difficulties of designing an effective method to improve their translation, and has led us to uncover the difficulties of choosing an appropriate evaluation method to compare translations. TQs are a particularly difficult phenomenon to evaluate, as they are largely determined by stylistic factors. There can also be several possible correct question tags for a single anchor clause, and the set of correct question tags is determined by the translation of the anchor clause. Our experiments have shown us that a single metric is not ideal, since it can be easily manipulated by choosing a simple and trivial strategy, but which does not respect the natural distribution of question tags as produced by English speakers. We therefore propose to also compare the distribution of predicted question tags with that of the reference translations.

¹⁵The sources of the subtitles are not indicated in the corpus, so there is no way of confirming the origin of the subtitles. Some are reproductions of the official subtitles (sometimes extracted by OCR), others are translations of the English subtitles, or even transcriptions of the dubbed version of the film.

Our choice of post-editing strategy improves question tag prediction for our lowest resourced language direction, DE→EN, however appears to equal the results of our baseline system for FR→EN and degrades it for CS→EN, the language direction with the MT system of the highest quality. The choice to use a series of classifiers rather than a single multi-class classifier was shown to be a good strategy, as it enables us to decompose the problem into several smaller and more manageable problems, limiting the effect of highly imbalanced class labels. For now, our analyses do not provide us with any extra information about the scenarios in which TQs are used, and this could be interesting to study in future work, with the use of higher quality systems or the analysis of an NMT system itself.

The choice to use post-processing to tackle this aspect of translation was made for several reasons. It enables us to have access to the target translations to ensure grammatical cohesion. However, it is especially a strategy that is practical, because it does not require making changes to an MT system, which would be time-consuming during development. Having access to MT outputs and designing a lightweight system to post-edit them is a faster way of improving translation. However is it the most effective? One of the difficulties we face in our experiments is the fact that the use of TQs is determined by subtle stylistic factors that are difficult to model using simple linguistic features guided by intuitions. Although grammatical question tags are largely dependent on the form of the anchor clause, the choice of lexical tag can appear at times arbitrary. Given this seemingly arbitrary nature, it could be interesting to seek to improve TQ translation through integrating unstructured context into the MT system itself, and to let statistical probability determine the outcome. We leave this strategy for future work.

6.2 Anaphoric pronoun translation with linguistically motivated features

As mentioned in Section 4.3.1, one of the most frequently used strategies to improve the translation of anaphoric pronouns is to post-edit MT outputs. As part of the WMT shared task on cross-lingual pronoun prediction (Guillou et al., 2016), introduced in Section 4.3.1, we designed such a system for the prediction of the French translations of the English subject pronouns *it* and *they*. Our system is based on a statistical feature-based classification approach, for which the features are motivated linguistically and are specifically selected to tackle particular difficulties of the task. This motivation therefore puts to the test the idea that the MT of anaphoric pronouns can be resolved through the use of structured context, selected according to linguistic intuitions about the problem. As in a number of previous works for the same problem, we therefore rely heavily on NLP resources and tools. As in (Le Nagard and Koehn, 2010; Hardmeier and Federico, 2010),

we choose to model coreference explicitly by using a coreference resolver (de Marneffe et al., 2015) and a morphological lexicon (Sagot, 2010), and to explicitly detect impersonal pronouns.

Task description The general setting of the task has previously been described in Section 4.3.1. However we choose to restate certain points of the task setting here for the purpose of clarity.

The aim of the task is to correctly predict the French translation of previously identified English subject pronouns *it* and *they*. The occurrences have been previously identified automatically in the data, and the translations replaced by placeholders, indicating where a prediction must be made. Multiple training corpora are available: EUROPARL (Koehn, 2005), IWSLT data and TED talks (Tiedemann, 2012), which we used as development data. The test set itself is from TED Talks, containing monolingual speeches, which can differ greatly in their degree of spontaneity, formality and topic. Data sizes and the distribution of pronouns are given in Table 6.7, copied here from Section 4.3.1. Although the purpose of a post-editing task is to modify MT outputs, the task is formulated differently, to ensure that evaluation can be comparable across all systems, and that it can be performed automatically. The pronouns must therefore be predicted in the reference translation themselves, rather than in MT outputs. Importantly, these reference translations are provided in an ‘underspecified format’ to avoid the task being artificially trivial: the translations are tokenised and the tokens are replaced by the corresponding lemmas and PoS tags, as shown in Table 6.8, again copied from Section 4.3.1.

The possible translations of the two English pronouns are grouped into 8 classes. We have previously seen how the correspondence between the English and French pronouns is ambiguous. The possible pronominal translations include anaphoric pronouns that agree in gender with the antecedent (IL, ELLE, ILS and ELLES), the impersonal (and therefore invariable) pronoun IL, used in impersonal constructions such as *il pleut* ‘it is raining’ and *il faut...* ‘it is necessary...’, the demonstrative pronouns CE (weak form) and CELA/ÇA (strong form) and finally the indefinite pronoun ON ‘they/we/you/one’. There is also a default category OTHER, regrouping all other translations that are aligned with the English pronoun. This can be due to rephrasing, errors or the pronoun being untranslated.

6.2.1 Classification system: description and motivation

Our classifier is a random forest classifier, implemented in SCIKIT-LEARN (Pedregosa et al., 2011).¹⁶ The choice of learning algorithm is based partly on random forests’ ability to account for class imbalance and outliers. The algorithm also has the advantage of being

¹⁶We use Gini as the optimising criterion, 250 estimators, a maximum depth of 500 and a minimum number of leaf samples of 1. All other parameters are those provided by default.

6.2. Anaphoric pronoun translation with linguistically motivated features

Dataset	Total	CE	CELA	ELLE	ELLES	IL	ILS	ON	OTHER
EUROPARL	494,110	51,170	13,202	48,460	18,387	168,579	45,603	9,452	139,257
Nc-v9	35,226	2,822	1,027	4,224	1,918	8,248	7,451	566	8,970
IWSLT15	69,487	16,415	6,908	3,286	3,276	9,682	17,145	1,549	11,226
TED (dev)	563	151	63	25	15	57	140	10	102
TED (test)	363	68	31	23	25	61	71	9	75

Table 6.7: Copy of Table 4.2. The distribution of pronoun classes for the English-to-French task in the data provided. The final datasets represent the development and test sets respectively. The pronouns correspond to translations of English subject pronouns *it* or *they*.

Pron.	Source sentence	Target sentence	Word alignments
	A lot of arguments against it .	il _{PRON} y _{PRON} avoir _{VER} beaucoup _{ADV} de _{PRP} argument _{NOM} contre _{PRP} ce _{PRON} pratique _{NOM} .	0-3 1-3 2-4 3-5 4-6 5-7 5-8 6-9
elles	Fish farms pollute , most of them do anyway , and they 're inefficient , take tuna .	le _{DET} ferme _{NOM} de _{PRP} aquaculture _{NOM} polluer _{VER} , _{PUN} du _{PRP} moins _{ADV} le _{DET} plupart _{NOM} de _{PRP} entre _{PRP} elle _{PRON} , _{PUN} et _{KON} [] être _{VER} inefficace _{ADJ} , _{PUN} prendre _{VER} par _{PRP} exemple _{NOM} le _{DET} thon _{NOM} .	0-0 1-1 1-3 2-4 3-5 4-9 5-10 6-11 6-12 7-12 8-7 9-13 10-14 11-15 12-16 13-17 14-18 15-19 16-23 17-24
	I almost never cook with it .	je _{PRON} ne _{ADV} la _{PRON} garder _{VER} presque _{ADV} jamais _{ADV} en _{PRP} cuisine _{NOM} .	0-0 1-4 2-5 3-7 4-3 4-6 5-2 6-8

Table 6.8: Copy of Table 4.1. An extract of three sentences from the English-to-French data provided for the shared task, taken from the TED corpus. Source sentences are tokenised, but target sentences are tokenised, tagged and lemmatised (the original forms are not given). Automatic word alignments are given between the source and target sentences. The anaphoric pronoun to predict in the example is the feminine plural ELLES, which corefers with the French lemma *ferme* ‘farm’.

non-linear, increasing the chances of findings patterns in the data with a relatively small number of features. We split the task into separate classifiers for each of the English pronouns *it* and *they*; a preliminary comparative study suggested that this produces slightly better results than training a single classifier for all source pronouns.

A possible way of trying to resolve the task would be to bombard a classifier with many weakly motivated features in the aim that the classifier learns how to use them. We choose the opposite strategy, privileging the use of linguistic tools and resources to provide a relatively small number of strongly motivated linguistic features, in order to see whether using interpretable features and linguistic intuitions is sufficient for the task. In terms of our description of methods previously used in the literature in Section 4, this tests how adapted external tools and resources are to provide sufficiently good structured context.

6.2.1.1 Data pre-processing

To extract features, we first pre-process the data to provide various linguistic annotations: PoS tags and dependency parses for both languages, coreference resolution for English and morphological analysis for French. English annotations are all produced using the Stanford Core-NLP toolkit (Manning et al., 2014). For parsing of French sentences, standard, pre-trained parsing models cannot be used on the lemma-based French sentences, and we therefore re-train a parsing model solely based on lemmas and PoS-tags, using the Mate graph-based transition parser (Bohnet and Nivre, 2012) and the French training data for the 2014 SPMRL shared task (Seddah et al., 2014). Some pre-processing is necessary to create a compatible tagset between the SPMRL data and the task training data.¹⁷ We enrich the French annotations using a morphological and syntactic lexicon, the *Lefff* (Sagot, 2010), to include noun gender by mapping lemmas to their genders (allowing for ambiguity). We also use the lexicon to provide information about impersonal verbs and adjectives, which will be described in more detail below.

6.2.1.2 Baseline features

The baseline system to which we compare our own classifier is based on the predictions of a 5-gram language model, provided by the task organisers. The language model can provide the most likely candidate translation of a given placeholder, based on the surrounding tokens, and can also provide an estimation of the probability that each of the possible pronouns be the correct translation. We use a combination of these features as our baseline system:

1. the most probable pronoun class
2. the concatenation of the two most probable pronoun classes
- 3-5. the most probable class if its probability is superior to 90%, 80% and 50% respectively

6.2.1.3 Linguistically motivated features

We model the context necessary for pronoun prediction using three types of linguistic feature. Each of the types is designed to cover one or more aspects of the translation problem, in order to cover as best as possible all potential translations of each English pronoun: anaphoric, impersonal, indefinite, demonstrative or non-pronominal/absent translations.

¹⁷An analysis of the quality of the syntactic annotations using the SPMRL test set and scorer gives an unlabelled attachment score of 89.83%.

1. **Coreference resolution features and impersonal pronoun detection:** These features are relevant for the translation of anaphoric pronouns with a nominal antecedent (or postcedent) in the text, of which the translation is *IL*, *ELLE*, *ILS* or *ELLES*. We deal with impersonal pronouns in this step too, in order to distinguish between personal and impersonal uses of the French pronoun *IL*.
2. **Local, syntax-based features:** These are features designed to help distinguish the use of *CE* ‘it/that’, *CELA* ‘it/that’ and *IL* as a translation of *it*, and to distinguish the use of *ILS* ‘they’ and *ON* ‘they/we/you/one’ as a translation of *they*.
3. **Identification of particularly discriminative contexts:** The detection of particular constructions and contexts is designed in particular to distinguish the use of *CE* or *CELA* instead of *IL*, and for the detection of environments in which the identification of a pronoun as a translation is probably erroneous (corresponding to the pronoun class *OTHER*).

Coreference resolution features and impersonal pronoun detection The main difficulty of translating the personal pronouns *it* and *they* into French is the fact that personal pronouns are gender-marked, the correct gender depending on the pronoun’s antecedent. In addition, although in a majority of cases, the pronoun’s number (singular or plural) does not change between the English and French translation, rephrasing in the translation can result in a change of number, and as described in Section 2.2.2.2, there can also be a mismatch between the use of plural *they* in English, where a French singular *IL* or *ELLE* is required. We choose to use explicit coreference resolution to provide two features, number and gender, that in theory can be used to resolve a majority of cases, where *it* and *they* are personal pronouns and translated with either *il/elle* or *ils/elles*.

We use the Stanford coreference resolver (de Marneffe et al., 2015) to identify the pronoun’s antecedent. The tool is available for English and in any case would apply poorly to the processed target sentences in French. We therefore apply the tool on the English sentence, and use the automatic alignments provided in the task to identify the nominal head of each pronoun’s antecedent in the French sentence. We identify the gender of the noun from its entry in the morphological lexicon *Lefff* (Sagot, 2010). In case of ambiguity, we include both genders. Since the French sentences are lemmatised, any number information is lost from the words, and therefore must be sought elsewhere. We compare two different strategies: (i) taking the number from the nominal head of the English antecedent as predicted by PoS tags of the English sentence, and (ii) taking the number of the English pronoun itself. A schema representing this process is given in Figure 6.3. Coreference chains can span across sentence boundaries, and mentions can span several words, in which case we take information associated with the mention’s head, as identified through simple language-specific rules.

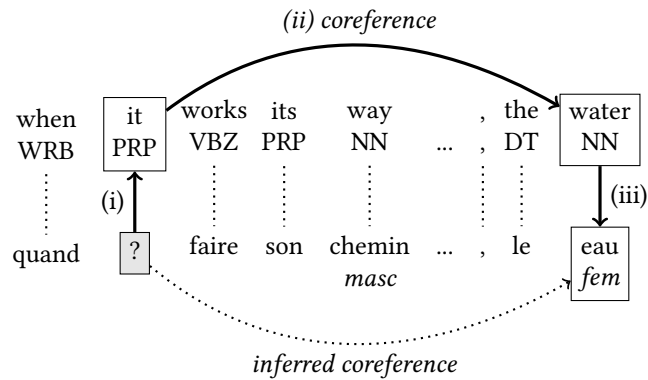


Figure 6.3: Use of coreference chains to determine gender and number of anaphoric pronouns. This example actually shows a cataphor, where the full noun appears after the corresponding pronoun. The value of the gender feature is determined by the noun *eau*'s entry in the *Lefff*, and its number is determined either by that of the word *water* or of the English pronoun *it* depending on which method is chosen. In this case, the result is singular in both cases.

In the case of perfect coreference resolution, and assuming no ambiguity in terms of the gender of French nouns, these two features alone could be expected to resolve the translation of the majority of the occurrences of *it* and *they* that are anaphoric and have a nominal antecedent in the text. In practice, there are several points at which the process may provide only partial or erroneous information. The accuracy of the features depends on the ability of the coreference tool to detect accurate and complete coreference chains, the quality of the automatic word alignments, the accuracy of the PoS tags (for the prediction of number) and the coverage of the lexicon (for French noun gender). The part of the process that poses the biggest problem is the coreference resolution itself. It is part of the process that requires the most linguistic modelling and also provides a bottleneck for the following steps. We evaluate the quality of the coreference tool (and the alignments) on the development set (TED-dev) by manually annotating the French pronouns for their full noun coreferent, and comparing this with the coreferent predicted via our method presented above. Of the 237 pronouns of the form *IL*, *ELLE*, *ILS* or *ELLES* in the set, 194 are anaphoric with a textual referent. The correct coreferent is identified in only 53% of cases, and the majority of those found are associated with the masculine plural class *ILS*. Moreover, 32% of these pronouns are linked only to other pronouns, therefore with no explicit referent (in particular for the feminine plural *ELLES*). The tool also often fails to predict impersonal pronouns, erroneously supplying coreference chains for 18 impersonal pronouns out of 25. Coreference resolution is a difficult NLP task, and tools are often not specifically adapted for those cases that are important for translation. They are typically evaluated on their capacity to construct as much of the chains as possible, but in translation, the only link that we are interested in is between anaphoric pronouns and the head noun of the chain. The lack of optimisation of the tool for these links could explain why so few pronouns are associated with their coreferent in our task.

Given the insufficiencies of the coreference tool, we decide to apply two modifications to the feature values obtained using the process previously described. The first is a back-off method for anaphora resolution in cases where no antecedent is found by the coreference tool. It consists in considering that the nearest preceding noun phrase in the previous sentence is the pronoun's referent, and using the same principles described above to attribute the number and gender values based on this antecedent. Although likely to add noise, this method provides more data points than the coreference tool, which under-detects referents of anaphoric pronouns. The second method we use, which also compensates for the over-generation of values (for non-anaphoric pronouns), is the detection of impersonal pronouns. We apply heuristic rules¹⁸ to detect impersonal instances of the French pronoun *IL*, and in the case of finding one, add the value *impersonal* to the number and gender values (or replace the values assigned in previous steps if this is the case). We consider a pronoun to be an impersonal *IL* when it is in an impersonal construction (containing an impersonal verb or adjective), information provided by the *Lefff*. Certain cases of non-ambiguous impersonals such as *il faut le faire* 'it must be done' are easily dealt with.¹⁹ Ambiguous cases, where the adjective or verb can be used both personally and impersonally, can often be disambiguated by the context, for example by the presence of a following *de* 'to' for verbs and adjectives or *que* 'that' for verbs. For example, *il est intéressant*. 'it/he is interesting' vs. *il est intéressant de...* 'it_{IMP} is interesting to...' or *il est intéressant que tu fasses...* 'it_{IMP} would be interesting for you to do...'.

Local syntactic features The choice of pronoun can sometimes be determined by its local context, for example the type of construction used or the verb of which it is the subject. The local context may be useful for detecting if the English pronoun is translated by something other than a French pronoun, or not translated at all, or even for helping to detect impersonal *IL*, if the case is not covered by the previously mentioned strategy designed for this aim. We also expect this to help in the detection of indefinite pronoun *ON*, as it is frequently used with certain verbs (e.g. *devoir* 'must'), and in situations where in English an impersonal construction would be used, e.g. *On devrait, par exemple, signaler que...* 'it should be pointed out that...'. The features include the form of the English aligned token (raw and lowercased), the form, PoS tag and lemma of the syntactic governor of the English aligned token and the PoS tag and lemma of the syntactic governor of the French pronoun. Finally, we include a boolean feature indicating whether or not the pronoun is found at the beginning of the sentence.

Discriminative context feature The most frequent category is *OTHER*, representing a host of different translations. To help in particular this class (plus the detection of

¹⁸Tools do exist for impersonal detection, however they are designed to process tokens and not lemmas.

¹⁹Such a construction is provided in the shared task data as [] *falloir/VER le/PRO faire/VER*.

impersonal expressions), we produce a single, strong feature, whose value is the class (if any) to which the pronoun’s context indicates that it is particularly likely to be associated. We calculate this value by analysing the target pronoun’s wider and richer context, and identifying whether this context has been seen very often with one particular pronoun class, and very little with other pronoun classes. If such a class is found, we assign this class as the feature value.

A first step is to extract contexts around all target pronouns and to count how often each context appears with each pronoun class. The contexts we consider are defined by values at certain positions relative to the target pronoun. We store the lemmas and PoS tags of the words at the following positions: (i) 2 following, (ii) 1 preceding and 2 following, (iii) 1 preceding and 3 following, (iv) the governor, (v) the governor and the function, (vi) the governor and its governor, and (vii) the preceding token, the pronoun’s governor and this governor’s function.²⁰ Each context value is linked with a certain pronoun class, with which it is particularly well associated. We measure the associativeness (or discriminativeness) of a context to a class by using a heuristic calculation based on the relative occurrences of the context with the pronoun class and the other pronouns, calculated on the training and development sets, as per Equation 6.1:

$$score(\langle c, y \rangle) = \frac{\text{occ}(\langle c, y \rangle)}{\sum_{y' \in Y} \text{occ}(\langle c, y' \rangle)} \sqrt{\text{occ}(\langle c, y \rangle)} \quad (6.1)$$

where c is a given context, y a given class and Y is the set of possible classes. The score is designed to rank highly $\langle \text{context}, \text{class} \rangle$ pairs that have both a high percentage of occurrences of the context with the class and a high frequency of occurrence. It is reasonable compromise between the probability of the context being associated with the given class and their frequency of co-occurrence. Although not normalised, the score, which is greater for a more relevant pair, has the advantage of being constant for a given probability and frequency count, and is therefore not dependent on the rarity of either the class or the context, unlike similar measures such as the log-likelihood ratio.

We select the 10,000 top-ranked pairs and further filter to only keep pairs where the context is associated with the class more than 95% of the time.²¹ If there is a match with one of the contexts identified, then the associated pronoun class (e.g. OTHER, IL) is used as the feature value. Otherwise a default value of *NA* is attributed. A total of 5,003 contexts are retained: 2,658 for OTHER, 1,987 for IL, 347 for CE, 9 for ON and 2 for CELA. Some of the identified contexts are given in Table 6.9, and corresponding examples from the training data are given in Table 6.10.

²⁰These templates are not symmetrical with respect to left and right context in order to privilege the context concerning the governing verb of the pronoun, which is more likely to be situated in the right context than the left, and may not be the token directly following the pronoun.

²¹We tested several values in preliminary experiments on the development set and found these values to be a good compromise between score optimisation and training time.

6.2. Anaphoric pronoun translation with linguistically motivated features

	Relative position					class	Num.	%
	-1	+1	+2	+3	gov.			
(1)		un	NOM			OTHER	1503	99
(2)	VER				NOM _{det}	OTHER	1003	97
(3)	le				VER _{subj}	ON	478	96
(4)	,	être	ADJ	que		IL	4131	98
(5)	PUN	être	ADJ	de		IL	5239	95

Table 6.9: Examples of contexts with their associated classes. We also give the percentage of occurrences of the context with the associated class and their frequency of co-occurrence. Corresponding examples are given in Table 6.10.

(1)	OTHER	EN: ...to bring it down to a more human level...
		FR: ...à _{PRP} [] un _{DET} échelle _{NOM} plus _{ADV} humain _{ADJ} ...
(2)	OTHER	EN: People react by voting with their feet and go where they can earn a crust .
		FR: le _{DET} gens _{NOM} réagir _{VER} en _{PRP} “PUN voter _{VER} avec _{PRP} leur _{DET} pied _{NOM} “PUN ,PUN en _{PRON} quitter _{VER} [] région _{NOM} dans _{PRP} le _{DET} espoir _{NOM} de _{PRP} trouver _{VER} de _{PRP} meilleur _{ADJ} revenu _{NOM} .
(3)	ON	EN: ...risks that had previously not even been considered .
		FR: ...risque _{NOM} que _{PRON} le _{DET} [] ne _{ADV} avoir _{VER} pas _{ADV} calculer _{VER} .
(4)	IL	EN: On the other hand , it is crucial that...
		FR: par _{PRP} ailleurs _{ADV} ,PUN [] être _{VER} important _{ADJ} que _{KON} ...
(5)	IL	EN: It will certainly be necessary to...
		FR: à _{PRP} ce _{PRON} égard _{NOM} ,PUN [] être _{VER} certainement _{ADV} nécessaire _{ADJ} de _{PRP} ...

Table 6.10: Examples from the training data corresponding to contexts matching each of the contexts in Table 6.9.

The identified contexts are particularly useful for detecting the OTHER class, which includes empty instances (the pronoun is untranslated) and translations other than the seven target pronoun classes. For example, there is a high degree of association between the OTHER class and having a context in which the target placeholder is directly followed by determiner *un* and a noun (first example in Table 6.9); this is a strange place to have a pronoun, and therefore models the fact that this is likely to be a context in which a pronoun will not be used. They can be especially useful in cases of alignment problems or anomalous predictions, and also for detecting certain collocations, for example *il est ADJ de* ‘it is ADJ to’ and *il est ADJ que* ‘it is ADJ that’.

6.2.2 Results, analysis and discussion

The main metric used to evaluate the systems is macro-averaged recall, designed to encourage participants to concentrate on the rarer classes. Results as compared to the baseline system are shown in Table 6.11. The two official baselines, provided by the task organisers, are $\text{baseline}_{\text{WMT-1}}$ and $\text{baseline}_{\text{WMT-2}}$. They are two variants of the same 5-gram language model trained on the task training data, on which we base our features in Section 6.2.1.2. The systems vary only in the optimisation of a parameter designed to counterbalance the fact that n -gram language models tend to provide higher probabilities to shorter strings; $\text{baseline}_{\text{WMT-1}}$ uses an unoptimised parameter whereas $\text{baseline}_{\text{WMT-2}}$ uses a parameter value obtained through optimisation on the dev set. We also provide two extra baselines: $\text{baseline}_{\text{mostFreqPro}}$, which predicts the most frequent class for each English pronoun (masc. sg. *IL* for *it* and masc. pl. *ILS* for *they*) and $\text{baseline}_{\text{LM}}$, which uses as features the form of the English pronoun (*it* or *they*) and the language model features described in Section 6.2.1.2. All scores are produced using the official evaluation script. We provide several results of our system.²² The two different versions (suffixed 1 and 2) correspond to the two different methods of providing the number value of the coreference features (see Section 6.2.1.3): the first method takes the number of the last referent identified by the coreference tool, and the second from the form of the aligned English pronoun. For comparative purposes, we also look at the scores of two additional variants for each version. *NoLM* variants do not use language model features, whereas *SimpleCR* variants only rely on the Stanford tool for coreference resolution, excluding our back-off method for these features.

6.2.2.1 Discussion

The evaluation metric for the task (macro-averaged recall) is such that very sparse classes hold a huge weight in the final evaluation. For example, correctly predicting an additional occurrence of *ON* improves the overall score by more than 1%. There are also vast differences in classification quality between the datasets, as illustrated by the systematic percentage point increase in score (up to 6 points) between the development and test set. This highlights the fact that the heterogeneity of data (also linked to data sparsity) should be taken into account when designing a system.

There is no significant difference between the two variants of our system. Compared to the four baselines, the linguistically rich systems perform systematically better. The much lower scores of $\text{baseline}_{\text{LM}}$ compared to S1 and S2 show that adding our linguistic

²²The scores differ slightly from those reported in (Bawden, 2016). A minor implementation issue was found concerning the use of the context, which nevertheless did not have a huge impact on the final results; the official primary submission resulted in a macro-averaged recall of 56% on the development set and 59% on the test set.

System	Macro-avg. Recall (%)		Acc. (%)
	Dev	Test	Test
baseline _{WMT-1}	41	47	52
baseline _{WMT-2}	-	51	53
baseline _{mostFreqPro}	24	24	35
baseline _{LM}	49	55	66
S1	56	61	69
S2	55	59	68
S1 _{NoLM}	52	54	63
S2 _{NoLM}	51	55	64
S1 _{SimpleCR}	55	61	71
S2_{SimpleCR}	56	61	71

Table 6.11: Comparative results of baseline systems and our systems S1 and S2, rounded to the nearest percent. The first, second and best scoring systems are marked with increasingly light shades of green. See the text for explanations about the differences between the models.

features provides extra and different information from the language model features. A slightly disconcerting observation is that if we remove the language model features (S1_{NoLM} and S2_{NoLM}), the score compared to baseline_{LM} is up to 3 percentage points higher on the development set, but only equivalent on the test set, suggesting that the information needed to predict the pronouns in the test set is probably mostly local, requiring less linguistic knowledge, another effect of the different natures of the sets and their small sizes. The experiments with simple coreference give comparable scores on the development set and higher scores on the test set (up to 61% macro-averaged recall for S1_{SimpleCR}). It is difficult to draw any conclusions about which method of gender and number induction is best, although our back-off method appears to be detrimental overall, adding noise to the data.

To look closer into the performance of our best system, S2_{SimpleCR}, we can look at the test results per pronoun class in the classification matrix, shown in Table 6.12 on the development set. Unsurprisingly, the most problematic classes are ELLE and ELLES, for which the only means of correctly predicting the gender is to have access to the pronoun’s textual referent and its gender. Although a majority of feminine pronouns are classified as having the correct number, only 3 out of 25 occurrences of ELLES are assigned the correct class. The other two classes for which the system performs less well are CELA (often confused with IL) and ON (confused with ILS and OTHER). These are all the least frequent pronoun classes, which therefore have a large impact on the overall score because of the macro-averaged metric. The classes which are best predicted are CE, with a high precision of 92% (85% F-score), OTHER with a high recall of 88% (82% F-score) and ILS with a recall of 79% (77% F-score).

	Classified as								Sum	P (%)	R (%)	F (%)
	CE	ELLE	ELLES	IL	ILS	CELA	ON	OTHER				
CE	54	1	0	11	0	0	0	2	68	92	79	85
ELLE	0	13	1	6	0	2	0	1	23	42	57	48
ELLES	1	2	3	1	13	1	0	4	25	23	12	16
IL	2	7	0	44	1	2	1	4	61	62	72	67
ILS	0	1	9	0	56	0	0	5	71	76	79	77
CELA	0	5	0	7	0	13	1	5	31	72	42	53
ON	0	0	0	0	2	0	5	2	9	56	56	56
OTHER	2	2	0	2	2	0	2	75	85	77	88	82
Sum	59	31	13	71	74	18	9	98				
Micro-averaged										71	71	71
Macro-averaged										62	61	60

Table 6.12: A breakdown of results for the system $S2_{SimpleCR}$ on the test set.

Oracle coreference resolver We predicted in Section 6.2.1.3 that coreference resolution would be the step at which most errors are made. The low percentage scores for the feminine pronouns confirm this. In order to assess the performance of our system independently of this specific tool, we imagine a scenario in which we have access to perfect impersonal pronoun detection and coreference resolution and can therefore correctly predict all instances of IL, ILS, ELLE and ELLES. We can therefore see how perfect coreference resolution would effect our results and what cases still remain unsolved.

We simulate perfect coreference resolution by supplying the expected gender and number features for all anaphoric pronouns in our dev set. This involves mapping the target class to the correct gender and number values (e.g. ELLE to FEM and SG, ILS to MASC and PL). In order not to confuse anaphoric IL with impersonal IL, we also detect impersonals. To supply the gold values, we detect all impersonal IL pronouns using the tool *ilimp* (Danlos, 2005), as applied to the French target sentences in the test set. The tool is designed to detect impersonal IL in normal French sentences (i.e. not those that have been converted into lemma and PoS tag pairs). Luckily, the unprocessed French sentences (not converted into lemmas and PoS tags) are available for the French-to-English version of the same task, and we therefore apply the tool to these French tokenised sentences which correspond to the same ones available for the English-to-French task.

The result (for the development set) when using oracle coreference resolution is a macro-averaged recall of 85% (vs. 56% when using coreference detection method described in Section 6.2.1.3). This show that if the anaphoric pronouns are predicted with 100% precision and recall, there are still errors, notably for the label ON, for which the precision is 57% and the recall only 40%, due to 6 out of 10 occurrences being classified as OTHER. The other class with low recall (although a high precision of 97%) is CELA, for which 25 out of 63 occurrences are incorrectly classified as OTHER. This suggests that there is a

positive bias towards the OTHER class, which is the third most frequent. We speculate that the over-prediction of this class could be due to the discriminative context feature (cf. the paragraph of the same name on page 144), which is geared to predict the OTHER class. Having such a statistically strong feature, with contexts highly related to a certain class does not allow for exceptions to the rule.

The oracle result shows that there is room for improvement for the other pronouns, even with perfect coreference resolution. To improve the use of discriminative contexts, there are two options. Firstly, the thresholds for the identification of contexts could be revised; they could either be increased to reinforce the feature's strength, or decreased to allow for more noise, enabling other features to counterbalance it in some cases. Secondly, more well-designed features that allow for a greater decomposition of decisions could be used, rather than relying on a single feature that does not allow any deviation from the rule.

The results obtained are promising and do suggest that highly structured context is useful for the task. However, pronoun translation is far from perfect, even when the antecedent appears within the same sentence. We have shown through our analysis that while coreference tools are at present inadequate for this task, there remain errors, linked to the inadequacy of our chosen features. The errors concern in particular the pronouns ON and CELA, whose usage is highly dependent on the translation structure used, of which we possibly do not supply enough information.

6.2.2.2 Comparison with other methods

Out of the nine participants in the English-to-French task, our systems is placed sixth, only slightly behind the systems in fourth and fifth place (less than 2%). The other systems use a range of different methods, some choosing to model coreference explicitly as in our method (Hardmeier, 2016; Tiedemann, 2016; Luotolahti et al., 2016), and others not (Stymne, 2016). As previously mentioned (cf. Section 4.3.1), the winning system for this edition (with a macro-averaged recall of 66%, vs. 61% for our classifier) is a neural classifier (Luotolahti et al., 2016) that does not rely on any extra-sentential information. The fact that this is the strongest system, even over systems using extra-sentential information, shows that there are still gains to be made in modelling anaphoric pronoun translation within the sentence. The second highest system is a linear classifier by Stymne (2016). The submission is interesting as it uses a range of linguistic features and relies greatly on external tools, not unlike our own system. Stymne does not directly model anaphora resolution, however, and the high score (macro-averaged recall of 65%) obtained compared to our own system indicate that externally resolving coreference may not be the best strategy. If supplied with enough linguistic features (the opposite of our strategy), the classifier can learn to choose the correct pronoun regardless. Stymne (2016) performs feature exploration, and like us, finds that local context (the three words preceding and

following the source and target pronoun) is particularly useful. She also uses language model features and dependency parsing, as we do, and finds these to be particularly useful too. She finds that enlarging the local context window does not actually help for this language direction (English-to-French), contrarily to other language directions tested, supporting the fact that the final dev and test sets may contain a relatively high number of pronouns for whom the antecedent is in a very local context. Stymne also performs two-step classification, by which she first distinguishes between the dominant OTHER class and all other pronoun classes, before distinguishing between the individual pronoun classes in a second classification step. This appears to be the major reason for the higher performances of her system with respect to ours, as her comparative one-pass classification system achieved similar scores to our own system.

6.2.3 Conclusion to pronoun translation via post-edition

The shared task setting is one that certainly encourages research into this type of phenomena, but remains artificial in the way it is set up. The choice to render French translations underspecified by tagging and lemmatising sentences removes the triviality of the task, but creates a situation that is very different from post-editing MT outputs. The systems had to counteract possible errors in PoS tagging and lemmatisation that could have led to errors (e.g. CE ‘this’ as a determiner is often wrongly tagged as a pronoun in the input provided by the task organisers), and would not have posed the same difficulty in a real translation setup. The OTHER class is also in some ways a by-product of the task setup, representing a host of different ways of translating pronouns, but also a consequence of noise in the alignment of English pronouns to French translations. Had the systems worked directly on MT outputs, it is likely that only French pronouns would have been corrected, eliminating the need for such a class, which nevertheless plays a big role in the task.

The tide is changing, and the submissions to the 2017 edition of the same shared task presented a dramatic change in the techniques used. Neural classification systems were the dominant systems used. Jean et al. (2017b) even present a contextual neural translation system designed not for the post-editing task but for translation itself. They use an additional encoder and attention mechanism to encode the previous sentence (in a way that is dependent on the current sentence), and then combine the representation obtained within that of the current sentence. Their results are comparable with some of the top systems on the task, showing that it is possible to tackle the problem of pronoun translation from a different angle: providing unstructured context rather than structured context, and changing the translation architecture itself. This change is positive, as it also reflects the fact that pronoun translation can be integrated into the translation process, removing the need for dedicated post-editing systems. Chapter 7 will be dedicated to such methods, for the integration of linguistic and extra-linguistic context.

6.3 General conclusion on post-edition approaches

Many previous works have, like our own, adopted post-editing strategies to integrate context into translation. From a practical point of view, it is a light-weight development strategy, as new post-editing models can be developed and fine-tuned without having to retrain an MT model. The context is available in both the source and target sentences and can be used by the post-editing models, which is essential for cohesion-based phenomena. We have seen in this chapter two such phenomena: TQs, which are cohesion-based with respect to the choice of question tag form (in particular for grammatical question tags), and anaphoric pronouns, whose forms are entirely dependent on the translation of their antecedents.

However, there are downsides to tackling these problems downstream of translation. We have shown two separate systems here, each designed to tackle a specific phenomena. The features used in each classification system are chosen and optimised specifically for the task, requiring pre-processing, feature extraction and training of a classifier. The processing chain is therefore costly, and the idea of tackling multiple such phenomena separately in such a way is not particularly satisfactory in terms of the use of resources and time. The gains we have seen in both experiments are relatively slight, despite the structuring of contextual information and the reliance on external resources of linguistic knowledge, making us question whether this is the best way to improve translation of these phenomena. There are at least two disadvantages of handling them in a separate step from translation itself: (i) we are not taking advantage of the information learnt by the MT model, and are therefore obliged to construct specific features from scratch, and (ii) we alter the translations according to a single specific phenomena targeted, and therefore have a limited ability to correct other aspects of the translation. This second point is particularly pertinent for tag questions, where the anchor clause could benefit from being made at the same time as the translation of the question tag in a global decision.

In the next chapter, we shall therefore pursue a third approach to integrating context, involving changing the MT architecture to accommodate for contextual information. As a consequence, the approach in the next chapter, in stark comparison with the two just presented, will not rely on structuring of linguistic context. The methods used will instead rely on the MT models learning to use context during translation.

Context-aware translation models

Related publications: (Bawden et al., 2018b)

We have seen in the previous two chapters two different approaches to applying context in MT. The first (Chapter 5) sought to exploit context in a pre-processing step, and the second to exploit via post-processing. Both of these approaches were MT-architecture-agnostic, and they bypassed the issue of exploiting context in translation by either modifying the data on which the MT models are trained or by correcting MT outputs. In this chapter, we explore a contextual strategy that involves modifying the MT architecture such that it is the role of the MT model itself to learn how and when to use context.

Our focus will be on NMT and the encoder-decoder model presented in Section 3.2. The approach we will be studying involves encoding context via an additional encoder. The resulting context representation is then combined with the representation of the current source sentence within the network. We rely on the NMT model learning how to best exploit the context (if at all) in order to optimise translation quality. In comparison with the two post-edition approaches of the previous chapter, the degree to which context is modelled prior to translation is minimal here. We present two experiments, the first (in Section 7.1) for the integration of previous linguistic context, represented by the unstructured sequence of words contained in the previous sentence, and the second (in Section 7.2) for the integration of extra-linguistic context related to speaker identity and topic. We shall focus heavily on evaluation in both sections, in an effort to understand whether the given context is truly being used to improve the translation of specific context-dependent phenomena.

7.1 Translating discourse phenomena with unstructured linguistic context

Our previous efforts to integrate linguistic context (cf. Chapter 6) have relied on structuring context prior to it being used to improve MT outputs. The basis of our strategy was to target particularly pertinent types of information to improve the translation of particular phenomena. This involved the processing, analysis and extraction of particular features. However, we have seen that such “power-house” strategies are limited: the gains in translation quality are minor, despite the approaches themselves being costly in terms of time, resources and human effort in terms of explicit linguistic modelling.

Here we shall look at different ways of modifying the MT architecture to integrate unstructured linguistic context into the decoder itself. For these experiments, we shall be concentrating on the NMT architecture presented in Section 3.2, and we shall test a variety of strategies using an additional encoder to encode linguistic context (Section 7.1.2), inspired by previous work on multi-encoder NMT strategies (Zoph and Knight, 2016; Libovický and Helcl, 2017; Jean et al., 2017a; Wang et al., 2017).

We put a particular focus on evaluation, insisting on the fact that it is important to test whether or not the context is really being used to solve some of the context-dependent problems presented in Chapter 2. The evaluation strategy we use to compare the models is inspired by previous work by Sennrich (2017) for MT grammaticality and Rios Gonzales et al. (2017) for word sense disambiguation. It consists in using the capacity of NMT models to rescore translations to compare the scores assigned by the model to contrastive sentence pairs, consisting of a correct and an incorrect translation of a source sentence in context.

We evaluate the use of linguistic context for *anaphoric pronoun translation* and *lexical choice* (regrouping both lexical disambiguation and lexical cohesion), to improve translation of sentences such as those in (55-57).

(55) *Anaphoric pronoun translation:*

EN: The bee is busy. // **It** is making honey.

FR: L'abeille_{FEM} est occupée. // **Elle**_{FEM}/#il_{MASC} fait du miel.

(56) *Lexical disambiguation:*

EN: And the code? // Still some **bugs**...

FR: Et le code ? // Encore quelques **bugs**/#insectes...

(57) *Lexical cohesion:*

EN: Do you fancy some soup? // **Some soup**?

FR: Tu veux de la soupe? // **De la soupe**/#du potage?

We begin this chapter by describing our hand-crafted test sets and how they can be used to evaluate contextual MT models. We describe the various contextual strategies we test in Section 7.1.2, and provide an evaluation of the different methods in Section 7.1.3. We use two different evaluation methods: an overall evaluation of translation quality using BLEU, and a targeted evaluation using our two discourse test sets.

7.1.1 Hand-crafted test sets for contextual MT evaluation

It is important to evaluate whether linguistic context is being meaningfully exploited by the MT model. The previous methods we have explored in this thesis have been limited by the evaluation methods chosen, and we have seen that traditional metrics that evaluate the overall translation quality are not well adapted to judging specific changes. In Section 6.2, we saw an alternative strategy for evaluating pronoun prediction used for the evaluation campaign on cross-lingual pronoun prediction, relying on evaluating pronoun translation in noisy reference translations. One of the problems with this evaluation method is that the pronouns' referents are not guaranteed to be outside of the current sentence, and therefore the test set does not specifically test how well extra-sentential context is being used. Challenges sets have previously been used to target the evaluation of particular phenomena (Isabelle et al., 2017), but evaluation is manual and therefore time-consuming. Guillou (2016) propose an alternative solution relying partly on automatic evaluation and partly on manual evaluation when the behaviour of the MT system has not been forseen. However, the set is not designed to specifically test the use of context.

The strategy we adopt is to construct a test set containing contrastive translation pairs, each pair containing a correct and an incorrect translation of the same source sentence. Instead of evaluating a translation produced by the NMT model, the models can be used to score the two translations. We can then evaluate the models on their capacity to score the correct translation higher than the incorrect translation. The idea has previously been used to test the grammaticality of NMT models (Sennrich, 2017) and for word sense disambiguation in translation (Rios Gonzales et al., 2017). We adapt the idea to evaluate contextual MT models' capacity to handle (i) translation of anaphoric elements and (ii) lexical choice (regrouping coherence- and cohesion-based phenomena), by constructing two such contrastive test sets.¹ Our test sets differ from previous ones in that examples necessarily need the previous context (source and/or target-side) for the translations to be correctly ranked. Unlike the evaluation test set used for the cross-lingual pronoun prediction task (see Section 6.2), the ambiguous pronouns' antecedents are guaranteed not to appear within the current sentence, meaning that, for MT systems to score highly, they must use extra-sentential context. Each of our examples includes the

¹The test sets are freely available at <https://diamt.limsi.fr/eval.html>.

previous sentence,² which contains the (source- and/or target-side) context necessary to disambiguate the correct and incorrect translations.³ Contrastive pairs differ by only as few words as necessary. All examples in the test sets are hand-crafted but closely inspired by real examples from OpenSubtitles2016 (Lison and Tiedemann, 2016) to ensure that they are credible and that vocabulary and syntactic structures are varied. Each of the test sets contains 200 contrastive pairs and is designed such that a non-contextual baseline system would achieve 50% accuracy.

Anaphora test set The set contains 50 example blocks, each containing four contrastive translation pairs (see the two example blocks in Table 7.1). The test set’s aim is to test the integration of target-side linguistic context. Each block is defined by a source sentence containing an occurrence of the anaphoric element (including personal pronouns *it* or *they*, possessive pronouns (e.g. *mine*) and one-anaphora) and its preceding context, containing the anaphoric element’s nominal antecedent. Four contrastive translation pairs of the previous and current source sentence are given, each with a different translation of the nominal antecedent, of which two are feminine and two are masculine per block. Each pair contains a correct translation of the current sentence, in which the pronoun’s gender is correct with respect to the antecedent’s translation, and a contrastive (incorrect) translation, in which the pronoun’s gender is inverted (along with agreement linked to the pronoun choice). Within an example block, there are only two different translations of the current sentence, each one successively corresponding to the correct and incorrect translations depending on the preceding context.

Two of the pairs contain what we refer to as a “contextually correct” translation of the current sentence instead of a “correct” one, for which the antecedent in the previous sentence is strangely or wrongly translated (e.g. *flies* translated as *araignées* “spiders” and *papillons* “butterflies” in Table 7.1). In the “contextually-correct” translation, the pronoun, whose translation is wholly dependent on the translated antecedent, is coherent with this translation choice. These contextually correct examples assess the use of target-side context, taking into account previous translation choices.

Among the 200 contrastive pairs (50 example blocks in total), target pronouns are evenly distributed according to number and gender with 50 examples (25 correct and 25 contextually correct) for each of the pronoun types (m.sg, f.sg, m.pl and f.pl). Since there are only two possible translations of the current sentence per example block, an MT

²This previous sentence can actually be made up of one sentence or more if the sentences are short. When this is the case, it is treated as a single sequence (i.e. a single input) by the MT systems, rather than as a multiple input sentences. The data we use also contains examples such as this, due to the automatic alignment of subtitles in the parallel corpus.

³We acknowledge that in reality, the disambiguating context is not guaranteed to be in the previous sentence (cf. Guillou (2016, p. 161), for the distribution of intra- and inter-sentential anaphoric pronouns). However it is important to first judge in a controlled way whether or not models are actually capable of using extra-sentential linguistic context at all, before investigating longer distance context.

Source:	
Context	Oh, I hate <u>flies</u> . Look there's another one!
Current sentence	Don't worry, I'll kill it for you.
Target:	
1 context:	Oh je déteste les <u>mouches</u> . Regarde, il y en a <u>une</u> autre !
correct:	T'inquiète, je la tuerai pour toi.
incorrect:	T'inquiète, je le tuerai pour toi.
2 context:	Oh je déteste les <u>mouchérons</u> . Regarde, il y en a <u>un</u> autre !
correct:	T'inquiète, je le tuerai pour toi.
incorrect:	T'inquiète, je la tuerai pour toi.
3 context:	Oh je déteste les <u>araignées</u> . Regarde, il y en a <u>une</u> autre !
contextually correct:	T'inquiète, je la tuerai pour toi.
incorrect:	T'inquiète, je le tuerai pour toi.
4 context:	Oh je déteste les <u>papillons</u> . Regarde, il y en a <u>un</u> autre !
contextually correct:	T'inquiète, je le tuerai pour toi.
incorrect:	T'inquiète, je la tuerai pour toi.
Source:	
Context	Can you authenticate these <u>letters</u> , please?
Current sentence	Yes, they're mine .
Target:	
1 context:	Pourriez-vous authentifier ces <u>lettres</u> , s'il vous plaît ?
correct:	Oui, ce sont les miennes .
incorrect:	Oui, ce sont les miens .
2 context:	Pourriez-vous authentifier ces <u>courriers</u> , s'il vous plaît ?
correct:	Oui, ce sont les miens .
incorrect:	Oui, ce sont les miennes .
3 context:	Pourriez-vous authentifier ces <u>documents</u> , s'il vous plaît ?
contextually correct:	Oui, ce sont les miens .
incorrect:	Oui, ce sont les miennes .
4 context:	Pourriez-vous authentifier ces <u>signatures</u> , s'il vous plaît ?
contextually correct:	Oui, ce sont les miennes .
incorrect:	Oui, ce sont les miens .

Table 7.1: Two blocks of examples from the anaphoric reference test set. The anaphora test set is made up of fifty such example blocks, therefore containing two hundred contrastive translation pairs. In each example, the previous sentence (context) disambiguates the two contrastive translations.

system can only score all examples within a block correctly if it correctly disambiguates, and a non-contextual baseline system is guaranteed to score 50%.

Including contextually correct examples has two advantages. As described above, they effectively allow us to really test how well the target context is being used, taking into account previous translations choice. However, they are also useful for ensuring that our test sets are balanced for gender, each example set containing two male anaphoric elements and two female anaphoric elements. Since it is difficult to ensure having two completely correct translations of every antecedent of which one is feminine and the other masculine, including two contextually correct examples allows us to keep this gender balance, by distributing the genders among the contextually correct and correct sentences.

Lexical choice test set The lexical choice test set contains 100 example blocks, each containing two contrastive pairs (see Table 7.2). Each of the blocks is constructed such that there is a single ambiguous source sentence, with two possible translations provided. The use of one translation over the other is determined by disambiguating context found in the previous sentence. The context may be found on the source side, the target side or both. In each contrastive pair, the incorrect translation of the current sentence corresponds to the correct translation of the other pair, such that the block can only be entirely correct if the disambiguating context is correctly used.

All test set examples have in common that the current English sentence is ambiguous and that its correct translation into French relies on context in the previous sentence. In some cases, the correct translation is determined more by cohesion, for example the necessity to respect alignment or repetition (second example block in Table 7.2). This means that despite two translations of an English source word being synonyms (e.g. *dingue* and *fou*, “crazy”), they are not interchangeable in a discourse context, given that the chosen formulation (alignment) requires repetition of the word of the previous sentence. In other cases, lexical choice is determined more by coherence, for example by a general semantic context provided by the previous sentence, in a more classic disambiguation setting as in the first example block in Table 7.2, where the English *steeper* is ambiguous between French *cher* “more expensive” and *raide* “sharply sloped”. However, these types are not mutually exclusive and the distinction is not always so clear.⁴

⁴For example, several examples contain a lexically ambiguous word, for which the context necessary for disambiguation provides the basis to disambiguate the word semantically, but is also a case of lexical cohesion (e.g. repetition). E.g. *I am not mad* ‘Je ne suis pas folle’ vs. ‘Je ne suis pas fâchée’ avec les phrases précédentes *You must be completely bad if you think it’s possible.* ‘Tu dois être complètement folle si tu penses que c’est possible.’ and *Oh no, now you’re mad at me.* ‘Oh non, maintenant tu es fâchée contre moi.’

Source:		
	context:	So what do you say to £50?
	current sentence	It's a little steeper than I was expecting.
Target:		
	context:	Qu'est-ce que vous pensez de 50£ ?
	correct:	C'est un peu plus cher que ce que je pensais.
	incorrect:	C'est un peu plus raide que ce que je pensais.
Source:		
	context:	How are your feet holding up?
	current sentence:	It's a little steeper than I was expecting.
Target:		
	context:	Comment vont tes pieds ?
	correct:	C'est un peu plus raide que ce que je pensais.
	incorrect:	C'est un peu plus cher que ce que je pensais.
Source:		
	context:	What's crazy about me?
	current sentence:	Is this crazy ?
Target:		
	context:	Qu'est-ce qu'il y a de dingue chez moi ?
	correct:	Est-ce que ça c'est dingue ?
	incorrect:	Est-ce que ça c'est fou ?
Source:		
	context:	What's crazy about me?
	current sentence;	Is this crazy ?
Target:		
	context:	Qu'est-ce qu'il y a de fou chez moi ?
	correct:	Est-ce que ça c'est fou ?
	incorrect:	Est-ce que ça c'est dingue ?

Table 7.2: Two example blocks, the first illustrating a problem of lexical ambiguity, and the second a problem of lexical cohesion (repetition). The lexical choice test set is made up of one hundred such example blocks, containing two hundred contrastive translation pairs.

7.1.2 Modifying the NMT architecture

NMT has provided new ways of integrating context. The use of dense vectors to represent sequences of text has made it possible to easily combine them through the application of functions within the NMT architecture. An intuitive way of integrating linguistic context into the NMT process is to compute a dense representation of the linguistic context, as is done with the source sentence to be translated, and to combine the two representations using an appropriate function. Such an approach requires changing the architecture to include an additional encoder to encode the linguistic context, and a mechanism to combine the two representations. The idea has previously been used for a variety of different aims including multi-source translation (Zoph and Knight, 2016) and multi-modal translation (Caglayan et al., 2016b; Huang et al., 2016). Unlike multi-modal translation, which typically uses two complementary representations of the main input, for example a textual description and an image, linguistically contextual NMT has focused on exploiting the previous linguistic context as auxiliary input alongside the current sentence to be translated (Libovický and Helcl, 2017; Wang et al., 2017), and this is the strategy that we shall adopt here.

Here, we test a variety of multi-encoder strategies for incorporating linguistic context into NMT. Our hypothesis is that providing the previous context (albeit limited to the previous sentence) to the NMT system will help to improve the translation of context-dependent phenomena such as lexical disambiguation, lexical cohesion and anaphora translation. We compare different ways of combining the representations of the context and the source sentence, as inspired by techniques from the literature. We also experiment with the inclusion of both source- and/or target-side context. We expect target-side context to be particularly useful for resolving cohesion-based phenomena such as anaphora translation, which relies on knowing the form of the determining word in the translation. For lexical disambiguation, we could expect either source or target context to be useful, since both may contain the necessary disambiguating context equally (if the translation is correct). The fact that we limit the context to the previous sentence is of course a huge simplification; the context necessary to help translation could appear elsewhere in the text, before the previous sentence or even after the current sentence. However, it is important to start with a simpler setting in order to assess the feasibility of the techniques before extending them to more challenging scenarios.

We begin by describing some simple single-encoder strategies for including context into NMT (Section 7.1.2.1). These will be used as a comparison for the multi-encoder strategies described in Section 7.1.2.2, and will also be as inspiration for our novel, hybrid strategy, presented in Section 7.1.2.3. The different strategies are shown schematically in Figure 7.1.

7.1.2.1 Single-encoder models

We train three single-source models: a baseline model and two contextual models. The baseline model translates sentences independently of each other (Figure 7.1a), using the model described in Section 3.2. The two contextual models, which are described in (Tiedemann and Scherrer, 2017), are designed to incorporate the preceding sentence by prepending it to the current one, separated by a <CONCAT> token (Figure 7.1b). The first method, which we refer to as 2-TO-2, is trained on concatenated source and target sentences, such that the previous and current sentence are translated together. The translation of the current sentence is obtained by extracting the tokens following the translated concatenation token and discarding preceding tokens. Although the non-translation of the concatenation symbol is possible, in practice this was rare (<0.02%). If this occurs, the whole translation is kept. The second method, 2-TO-1, follows the same principle, except that only source (and not target) training sentences undergo concatenation; the model is trained to directly produce the translation of the current sentence. The comparison of these two methods allows us to assess the impact of translating the context in improving translation.

7.1.2.2 Multi-encoder models

We compare the single-encoder models to different multi-encoder models. We encode the previous (source and/or target) sentence using a separate encoder (with separate parameters) to produce a context vector of the auxiliary input in a parallel fashion to the current source sentence. Note that this means that a second separate attention mechanism is used to produce the additional context vector at each decoding timestep j . The encoding of the previous sentence and the current sentence at timestep j results in two context vectors (as per the terminology introduced in Section 3.2), which we call $c_1^{(j)}$ and $c_2^{(j)}$ respectively. These two vectors are then combined to form a single context vector $c^{(j)}$ to be used for decoding (see Figure 7.1c). We study three combination strategies: concatenation, the use of an attention gate and hierarchical attention, each taken from previous work in the literature.⁵ References to these works will be provided below, as we describe each strategy in more detail. As in Section 3.2, all W 's, U 's and b 's are learned parameters. We refer to the recurrent state of the decoder at timestep j as $z^{(j)}$.

Attention concatenation The two context vectors $c_1^{(j)}$ and $c_2^{(j)}$ are concatenated and the resulting vector undergoes a linear transformation in order to return it to its original dimension to produce $c^{(j)}$ (similar to work by Zoph and Knight (2016)).

⁵We also tested using the auxiliary context to initialise the decoder, similar to Wang et al. (2017), which was ineffective in our experiments.

$$c^{(j)} = W_c[c_1^{(j)}; c_2^{(j)}] + b_c \quad (7.1)$$

Attention gate A gate $r^{(j)}$ is learnt between the two vectors to give differing importance to the elements of each context vector, similar to Wang et al.’s 2017 strategy.

$$r^{(j)} = \tanh\left(W_r c_1^{(j)} + W_s c_2^{(j)} + b_r\right) \quad (7.2)$$

$$c^{(j)} = r^{(j)} \odot \left(W_t c_1^{(j)}\right) + (1 - r^{(j)}) \odot \left(W_u c_2^{(j)}\right) \quad (7.3)$$

Hierarchical attention An additional (hierarchical) attention mechanism (Libovický and Helcl, 2017) is introduced to assign a weight to each encoder’s context vector (designed for an arbitrary number of encoders).

$$e_k^{(j)} = v_b^\top \tanh\left(W_b z^{(j-1)} + U_k c_k^{(j)} + b_e\right) \quad (7.4)$$

$$\beta_k^{(j)} = \frac{\exp\left(e_k^{(j)}\right)}{\sum_{k'=1}^K \exp\left(e_{k'}^{(j)}\right)} \quad (7.5)$$

$$c^{(j)} = \sum_{k=1}^K \beta_k^{(j)} V_k c_k^{(j)} \quad (7.6)$$

where $c_k^{(j)}$ is the context vector of the k^{th} encoder (K is equal to 2 or 3 in our experiments) at decoding step j , $z^{(j-1)}$ is the previous decoder state, and W_b , U_k and V_k are learned weight matrices, U_k and V_k being specific to each encoder number k .

7.1.2.3 Novel strategy of hierarchical attention and context decoding

We also test a novel strategy of combining multiple encoders and decoding of both the previous and current sentence. We use separate, multiple encoders to encode the previous and current sentence and combine the context vectors using hierarchical attention. We train the model to produce the concatenation of the previous and current target sentences, of which the second part is kept, as in the contextual single encoder model 2-to-2. A schema representing this strategy is given in Figure 7.2. The advantage of this approach is that it combines the two strategies of including context: the use of multiple encoders, which creates a separate and specific context vector for the previous sentence, and the decoding of the previous sentence, which means that the target context is contained in the decoder’s history while it is translating the current sentence.

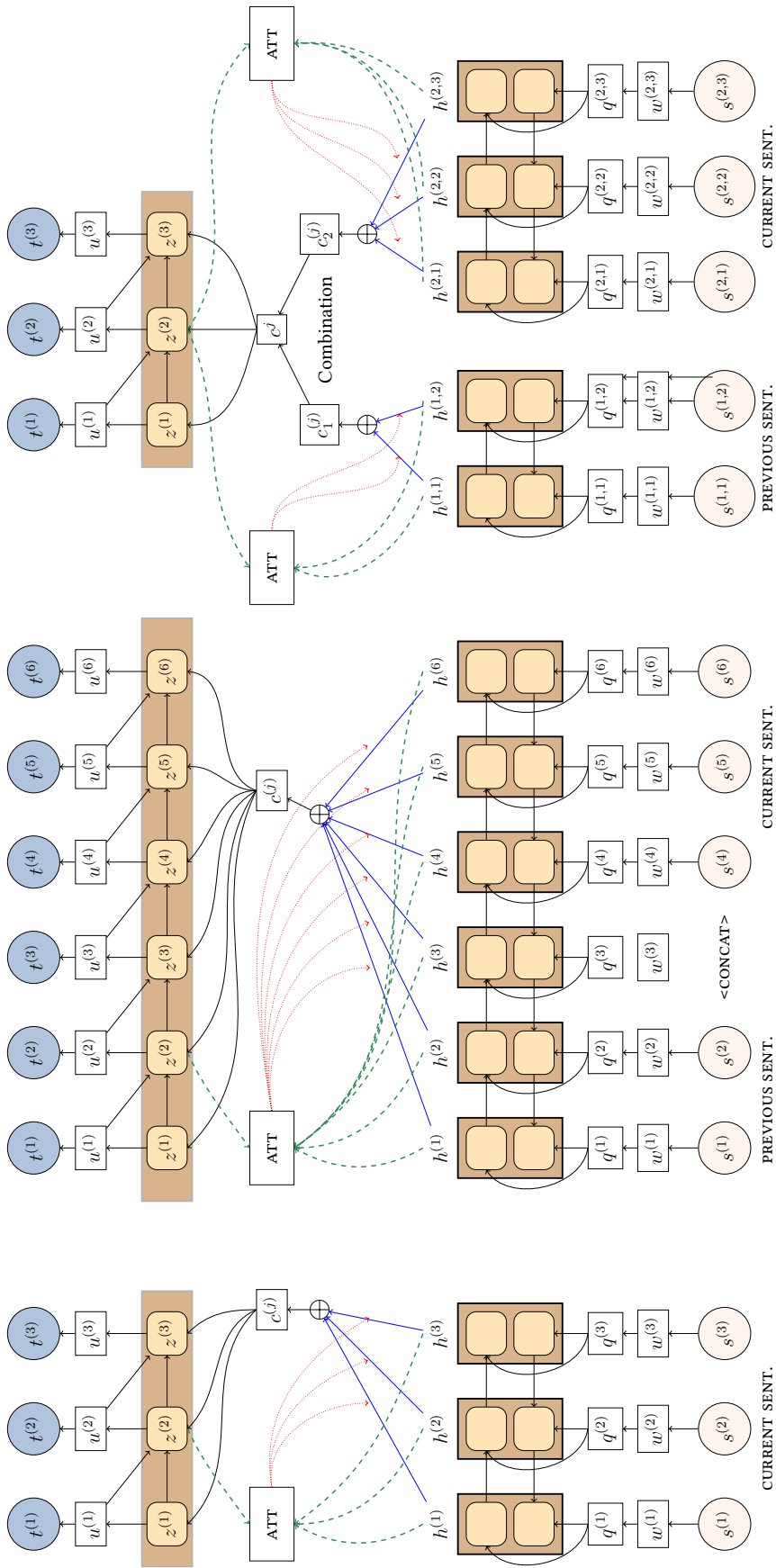


Figure 7.1: The baseline model and the two contextual strategies tested (single and multi-encoder).

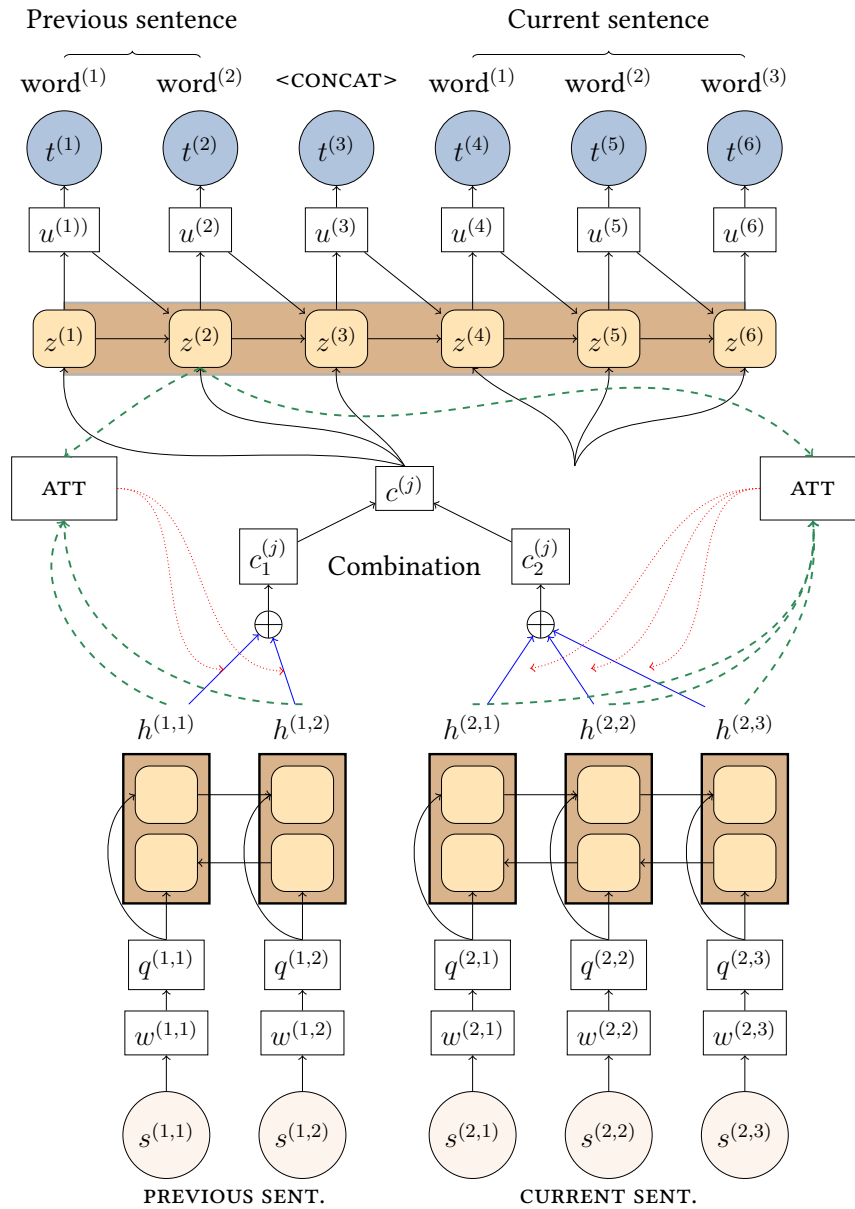


Figure 7.2: Our hybrid strategy combines the multiple-encoder approach with the idea of decoding the concatenation of the previous and current sentence. We refer to this strategy, as per our naming conventions, as HIER-TO-2

7.1.2.4 Experimental setup

Each of the multi-encoder strategies is tested using the previous source and target sentences as an additional input (prefixed as *s-* and *t-* respectively) in order to test which is the most useful disambiguating context. Two additional models tested are triple-encoder models, which use both the previous source and target (prefixed as *s-t-*). We abbreviate the attention strategies to *CONCAT*, *GATE* and *HIER*.

7.1.2.5 Data

Models are trained and tested on English-to-French parallel subtitles from OpenSubtitles2016⁶ (Lison and Tiedemann, 2016). The data is first corrected using heuristics (e.g. minor corrections of OCR and encoding errors). It is then tokenised, further cleaned (keeping subtitles ≤ 80 tokens) and truecased using the Moses toolkit (Koehn et al., 2007) and finally split into subword units using BPE (Sennrich et al., 2016d).⁷ We run all experiments in a high-resource setting, with a training set of ≈ 29 M parallel sentences, with vocabulary sizes of ≈ 55 k for English and ≈ 60 k for French.

The choice to split certain words into subwords rather than whole words potentially interacts with the idea of correctly translating anaphoric phenomena, since the antecedents themselves may be spread over several subword units. In practice, this segmentation could actually be seen more as an advantage than as a disadvantage, as it may enable the gender of unknown words to be identified through productive suffixes that are separated by subword segmentation. Moreover, since the aim is not to produce anaphora resolution but simply to ensure correct translation, other linguistic cues such as the gender of a preceding determiner may provide the necessary information, and therefore we do not need to pay as much attention to retaining linguistically motivated word units.

7.1.2.6 Training details

All models are sequence-to-sequence models with attention (Bahdanau et al., 2015), implemented in *NEMATUS* (Sennrich et al., 2017). The final models used for translation and scoring are ensembles of the last three check-pointed models during training.⁸ This not only results in a better scoring final model, but also ensures a greater degree of stability with respect to the results. Full details concerning hyperparameters, data and the experimental setup are provided in Appendix A.1.1.

⁶<http://www.opensubtitles.org>

⁷90,000 joint merge operations with a minimum threshold of 50.

⁸Ensembling is performed as per its implementation in *NEMATUS*. The geometric average of all individual models' probability distribution is calculated to produce the distribution of the ensembled model.

Models that use the previous target sentence are trained using the previous reference translation rather than the MT output of the previous sentence. This is necessary to avoid having to either translate in two passes or to alter the way in which batches of shuffled sentences are used to train. However this strategy is clearly suboptimal since the training setting is different from the one in which it will be applied – when using the model to translate, only the MT output (rather than a reference translation) of the previous sentence is available. However, when translating new sentences (i.e. when not training), we use baseline translations of the previous sentences as input to the model.⁹ For the targeted evaluation, the problem does not occur since we simply rescore translations that are already provided, rather than require the MT model to translate.

The auxiliary input for the first sentence (for which there is no previous sentence) is set as the pseudo-token *<BEGIN>* (both during training and testing). Although the training and testing data is split into films, we choose to ignore film boundaries. There relative few number of sentences affected by this simplification (the same number as the number of films) and this strategy could also make the model more robust to noise.

7.1.3 Evaluation results and analysis

We evaluate the different models (described in Section 7.1.2 and of which the names are defined in Section 7.1.2.4) using two types of evaluation method. The initial evaluation is of overall translation using the BLEU (Section 7.1.3.1). In Section 7.1.3.2, we then evaluate and compare the models in a more targeted fashion using the hand-crafted test sets we described in Section 7.1.1.

7.1.3.1 Overall performance

Overall translation quality is evaluated using the traditional automatic metric BLEU (Papineni et al., 2002). Despite the many criticisms of BLEU (see Sections 3.3 and 4.1), it can be useful to give an overall idea of the quality of a translation model, as a sanity check for translation quality. Any large, significant changes in BLEU (especially if they are calculated over a large test set) can also be a good indicator of which methods work best. We calculate BLEU scores on four reasonably large different test sets, each regrouping three films of a specific genre: comedy, crime, fantasy and horror.¹⁰

The results are provided in Table 7.3. The models are described in the first half of the table: *#In* is the number of input sentences, the type of auxiliary input of which (previous

⁹These are produced by the non-contextual encoder-decoder model that only take as input the current sentence.

¹⁰The size of the test sets vary from 2,158 sentences for horror and 2,790 sentences for fantasy to 4,227 for crime and 4,490 sentences for comedy.

source or target) is indicated by *Aux.*, *#Out* is the number of sentences translated, and *#Enc* is the number of encoders used to encode the input sentences. When there is a single encoder and more than one input, the input sentences are concatenated, with a special separator between them, to form a single input to the encoder.

	System Description				BLEU			
	Aux.	#In	#Out	#Enc.	Comedy	Crime	Fantasy	Horror
<i>Single-encoder, non-contextual model</i>								
BASELINE	—	1	1	1	19.52	22.07	26.30	33.05
<i>Single-encoder with concatenated input</i>								
2-TO-2	src	2	2	1	20.09	22.93	26.60	33.59
2-TO-1	src	2	1	1	19.51	21.81	26.78	34.37
<i>Multi-encoder models (+previous target sentence)</i>								
T-CONCAT	trg	2	1	2	18.33	20.90	24.36	32.90
T-HIER	trg	2	1	2	17.89	20.77	25.42	31.93
T-GATE	trg	2	1	2	18.25	20.76	25.55	32.64
<i>Multi-encoder models (+previous source sentence)</i>								
S-CONCAT	src	2	1	2	19.35	22.41	26.50	33.67
S-HIER	src	2	1	2	20.22	21.90	26.81	34.04
S-GATE	src	2	1	2	19.89	22.80	26.87	33.81
S-T-HIER	src, trg	3	1	3	19.53	22.53	26.87	33.24
<i>Multi-encoder with concatenated output</i>								
S-HIER-TO-2	src	2	2	2	20.85	22.81	27.17	34.62
S-T-HIER-TO-2	src, trg	3	2	3	18.80	21.18	27.68	33.33

Table 7.3: Results (de-tokenised, cased BLEU)¹¹ of the models on four different test sets, each containing three films from each film genre. The best, second- and third-best results are highlighted by decreasingly dark shades of green.

Scores vary dramatically depending on the genre, and the best model is not always the same for each of the genres, showing that it is important to evaluate models on several test sets rather than a single one. Contrary to intuition, using the previous target sentence as an auxiliary input (prefix τ -) degrades the overall performance considerably. We also experiment with using reference translations of the previous translation at test time (rather than the MT outputs). However this does not significantly improve this result, suggesting that it is unlikely to be a case of overfitting during training. The highest performing model is our novel S-HIER-TO-2 model with more than +1 over the baseline BLEU on almost all test sets. There is no clear second best model, since performance depends strongly on the test set used, although the single-encoder contextual strategy 2-TO-2 performs surprisingly well given its simplicity. However, these results tell us little about how well context is being used. To verify the use of context, we turn to our targeted evaluation.

¹¹These results are calculated using MULTI-BLEU.PERL. An updated version of the BLEU scores using MULTI-

7.1.3.2 Targeted discourse evaluation

Table 7.4 shows the results on the two discourse test sets.

Anaphora The basic multi-encoder models do not perform well on the anaphora test set; all multi-encoder models giving at best random accuracy, as with the baseline. This set is designed to test the model’s capacity to exploit previous target context. It is therefore unsurprising that multi-encoder models using just the previous source sentence perform poorly. It is possible that certain pronouns could be correctly predicted from the source antecedents, if the antecedent only has one possible translation. However, this non-robust way of translating pronouns is not tested by the test set. More surprisingly, the multi-encoder models using the previous target sentence also perform poorly on the test set. An explanation could be that the target sentence is not being encoded sufficiently well in this framework, resulting in poor learning. This hypothesis is supported by the low overall translation performance shown in Table 7.3.

Two models perform well on the test set: 2-TO-2 and our S-HIER-TO-2. The high scores, particularly on the less common feminine pronouns, which can only be achieved through using contextual linguistic information, show that these models are capable of using previous linguistic context to disambiguate pronouns. The progressively high performance of these models can be seen in Figure 7.3, which illustrates the training progress of these models. The S-T-HIER-TO-2 model (which uses the previous target sentence as a second auxiliary input) performs much worse than S-HIER-TO-2, showing that the addition of the previous target sentence is detrimental to performance. Whilst the results for the “correct” examples (CORR.) are almost always higher than the “contextually-correct” examples (CONTEXT), for which the antecedent is strangely translated, the TO-2 models also give improved results on these examples, showing that the target context is necessarily being exploited during decoding.

These results show that the translation of the previous sentence is the most important factor in the efficient use of linguistic context. Combining the S-HIER model with decoding of the previous target sentence (S-HIER-TO-2) produces some of the best results across all pronoun types, and the 2-TO-2 model performs almost always second best.

Lexical choice Much less variation in scores can be seen here, suggesting that these examples are more challenging and that there is room for improvement. Unlike the coreference examples, the multi-encoder strategies exploiting the previous source sentences perform better than the baseline (up to 53.5% for S-CONCAT). Yet again, using the previous target sentence as an auxiliary input achieves near random accuracy. 2-TO-2

BLEU-DETOK.PEEL will be provided in the final version of this thesis.

	Anaphora (%)							Lex. choice (%)
	ALL	M.SG.	F.SG.	M.PL.	F.PL.	CORR.	CONTEXT	ALL
BASELINE	50.0	80.0	20.0	80.0	20.0	53.0	47.0	50.0
2-TO-2	63.5	92.0	50.0	84.0	28.0	68.0	59.0	52.0
2-TO-1	52.0	72.0	28.0	84.0	24.0	54.0	50.0	53.0
T-CONCAT	49.0	88.0	8.0	96.0	4.0	50.0	48.0	51.5
T-HIER	47.0	78.0	10.0	90.0	10.0	47.0	47.0	50.5
T-GATE	47.0	80.0	6.0	82.0	20.0	45.0	49.0	49.0
S-CONCAT	50.0	68.0	32.0	88.0	12.0	55.0	45.0	53.5
S-HIER	50.0	64.0	36.0	80.0	20.0	55.0	45.0	53.0
S-GATE	50.0	68.0	32.0	84.0	16.0	55.0	45.0	51.5
S-T-HIER	49.5	94.0	4.0	88.0	12.0	53.0	46.0	53.0
S-HIER-TO-2	72.5	100.0	40.0	90.0	36.0	77.0	68.0	57.0
S-T-HIER-TO-2	56.5	84.0	36.0	86.0	20.0	55.0	58.0	51.5

Table 7.4: Results on the discourse test sets (% correct). For the anaphora set, results are also given for each pronoun class. CORR. and CONT. correspond to the “correct” and “contextually correct” examples. The best, second- and third-best results are highlighted by increasingly light shades of green.

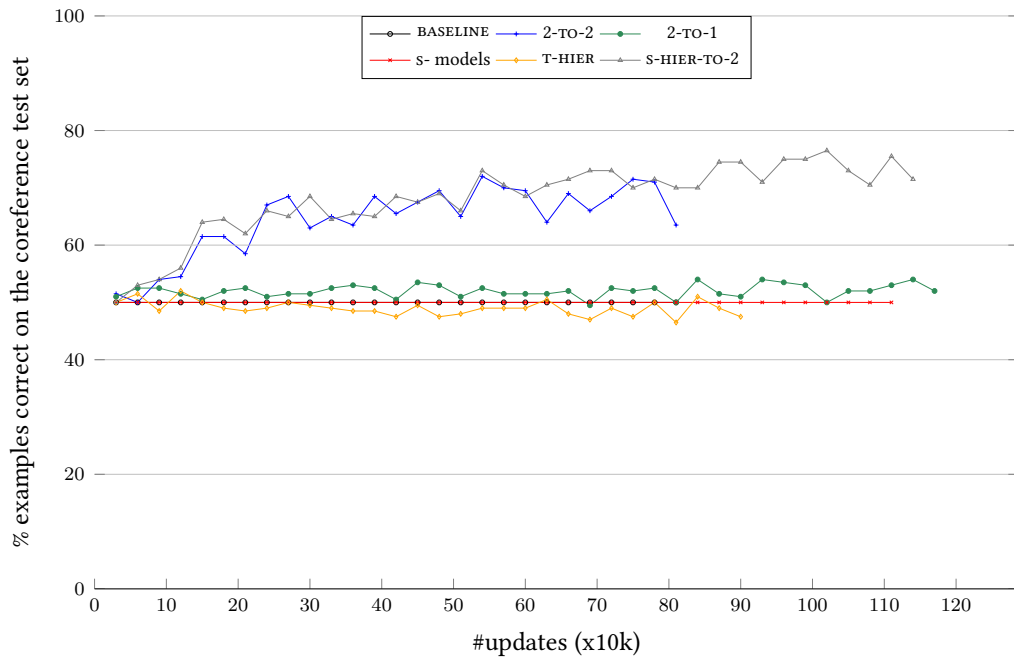


Figure 7.3: Progression of the percentage of correctly ranked examples (from the anaphora test set) during training.

and 2-TO-1 achieve similarly low scores (52% and 53%), suggesting that if concatenated input is used, decoding the previous sentence does not add more information.

However, combining multi-encoding with the decoding of the previous and the current sentences (S-HIER-TO-2) greatly improves the handling of the ambiguous translations, improving the accuracy to 57%. Extending this same model to also exploit the previous target sentence (S-T-HIER-TO-2) degrades this result, giving very similar scores to T-HIER, and is therefore not illustrated in Figure 7.3. This provides further support for the idea that the target sentence is not encoded efficiently as an auxiliary input and adds noise to the model, whereas exploiting the target context as a bias in the recurrent decoder is more effective.

7.1.3.3 How much is the context being used?

Looking at the attention weights can sometimes offer insights into which input elements are being attended to at each step. For coreference resolution, we would expect the decoder to attend to the pronoun's antecedent. The effect is most expected when the previous target sentence is used, but it could also apply for the previous source sentence when the antecedent has only one possible translation. Unlike Tiedemann and Scherrer (2017), we do not observe increased attention between a translated pronoun and its source antecedent. Given the discourse test set results, which can only give high scores when target-side context is used, the contextual information of the type studied in this paper seems to be best exploited when channelled through the recurrent decoder rather than when encoded through the input. This could explain why coreference is not easily seen via attention weights; the crucial information is encoded on the decoder-side rather than in the encoder.

We nevertheless take a look at the hierarchical attention weights for two of our models: S-HIER and S-HIER-TO-2. Hierarchical attention weights are those assigned to each context vector at each decoding timestep, giving for each translated word the relative importance (a weight between 0 and 1) of the context vector representing the previous sentence with respect to the context vector representing the source sentence.

To get an idea to what extent the previous context encoding is attended to by each of the models, we calculate the average attention weight calculated for this previous source context encoding per decoding step. To ensure that our comparison is fair, we calculate this using our two test sets. We can see from Table 7.5 that the average attention weight assigned to the context representation of the previous sentence vary according to which model is used. Whereas the S-HIER model assigns an average weight of between 0.26 and 0.27 to the previous context representation per target word (depending on the test set), the S-HIER-TO-2 model assigns a smaller weight of 0.17 to 0.18, a decrease of 0.09 in each case. This drop in the attention weight can be explained by the difference between the

Model	Average weight per decoding step		#steps with higher weight	
	Anaphora	Lexical choice	Anaphora	Lexical choice
S-HIER	0.26	0.27	3328 (81.4%)	2825 (83.1%)
S-HIER-TO-2	0.17	0.18	758 (18.6%)	573 (16.9%)
Difference	0.09	0.09	2570	2252

Table 7.5: Analysis of the hierarchical attention weight assigned to the *previous* source context for models S-HIER and S-HIER-TO-2. For each test set, this table gives the average weight per decoding step (i.e. per translated word) computed for the previous source context and the number (and percentage) of decoding steps for which the model assigns a higher weight than the other model for the same target word.

models. The second model, S-HIER-TO-2, which decodes both the previous and current sentence, has access to the previous context via the decoder’s representation, whereas the first one, S-HIER, does not. Intuitively, S-HIER-TO-2 depends less on this previous context from the encoder, as information concerning the previous sentence is already available through the recurrent decoder, whereas S-HIER is prone to using the previous context representation produced by the additional encoder, as it does not have access to the information otherwise. This difference in the values of the hierarchical attention of the two models can also be seen by a pair-wise comparison of weights for each target word (again compared for the same target sentences), shown on the right side of Table 7.5. Although in a majority of instances, S-HIER assigns a higher weight to the previous source context than S-HIER-TO-2 for the same word (of the same sentence), this is by no means systematic. For the anaphora test set, the weight assigned by S-HIER-TO-2 to the previous source context is higher than the corresponding weight (in the same decoding step) assigned by S-HIER in 18.6% of cases. For the lexical choice set, the figures are similar, with a slightly lower percentage of 16.9% decoding steps presenting higher weights for S-HIER-TO-2 than S-HIER. The target words corresponding to these decoding steps for which S-HIER-TO-2 assigns a higher weight than in the equivalent decoding step for S-HIER do not appear to correspond to a clear pattern. For both test sets, the most frequent such target words are among the most frequent words in the set (e.g. *de* ‘of/from’, *le* ‘it/the’, the comma and full stop and *ne* ‘not’). Normalising the occurrences for their frequency does not lead to clearly interpretable results, as the resulting most frequent words are dominated by hapaxes. Setting a minimum frequency threshold results in a list of words whose occurrences are just above the threshold defined, and therefore are of limited interpretability. We include two visualisations of these hierarchical attention weights for two different sentences in Appendix A.1.2.

7.1.4 Conclusion and perspectives

Further exploration into the interpretation of neural networks is necessary to understand how the information is transferred in the network, and to better understand the role of the encoder and of the decoder. Such interpretation is still in its early days. These experiments have nevertheless enabled us to compare the potential of different NMT strategies to integrate context. It is unsurprising that enlarging the segment of text to be translated as in the 2-TO-2 method works better than a baseline model, since the context is being directly provided within the input sequence. However such a method is not tractable for longer distance context. For encoder-decoder models, we have been able to show that decoding the context is an important factor in improving the use of context, as shown by the higher scores of our novel strategy *S-HIER-TO-2*, which uses multiple encoders, but also decodes the context as a prefix to the translation of the current sentence. Unlike Tiedemann and Scherrer (2017), in such a model, we have been unable to see any clear signs that the attention mechanism is attending in particular to specific elements of encoded context. If the information is better passed through the decoder, then this could partly explain why.

A natural development for this work is to extend it to new NMT architectures, including the Transformer NMT model (Vaswani et al., 2017) (cf. Section 3.2.4), which does not include any recurrent units, instead relying on layers of self-attention. This perspective has in fact already been carried out, with promising results. Recent experiments by Voita et al. (2018) for English to Russian using the Transformer do show that such a model attends to the coreferent of anaphoric pronouns. The choice of model, in which recurrent units are replaced by layers of self-attention, increases the interpretability of the model, since the attention weights can be studied to identify to which contextual words the model attended the most. They show that model can perform latent anaphora resolution. Although not specifically designed for the task, they show that the model gives a 72% precision in agreement links for a subset of sentences for which Stanford CoreNLP coreference resolver (Manning et al., 2014) identified an anaphoric link. By comparison, the Stanford resolver, designed specifically for the task, had a score of 77% on the same anaphoric pronouns.

Future work in contextual MT will with no doubt look into exploiting context from a much wider field than the previous sentence. However it is important for progress to be made incrementally. Recent work has built on our work to study the effect of adding the following sentence, several previous sentences and combinations of these (Agrawal et al., 2018). Their experiments, also using the Transformer model, show that higher BLEU scores (and better TER scores) can be achieved when using more context. However they do not provide an evaluation targeting how well the extra-sentential context added is used specifically by their models.

Any perspectives in the long term are very likely to be dependent on new advances in NMT architectures. Exploiting context correctly, whether it is from within a sentence or across sentences is key to improving MT. New ways of exploiting the context within sentences could therefore certainly be used to improve the exploitation of context further afield. However, it could also be worthwhile investigating how coupling the strategies seen in this section with the inclusion of additional structured information, such as richer linguistic features, could affect how models are able to use the context given.

We have shown through our experiments that even when the context is necessarily in the previous sentence, there is still a lot of progress to be made in exploiting this limited context correctly, in particular for the lexical choice test set, for which accuracy reached only 57% for our best model. Evaluation is key to tracking progress. The use of test sets such as ours is fundamental for understanding the limits of our models, and whether they are actually capable of using context to improve the translation of context-dependent phenomena, rather than simply improving the general overall translation quality.

7.2 Contextual NMT with extra-linguistic context

In the previous section we saw several ways to integrate linguistic context into NMT. Without having to explicitly structure the context provided to the MT system, we saw an improved use of the context as the neural network learns how to use such information. In this section, we seek to discover whether the same techniques (or adaptations of these techniques) can also be effective for the inclusion of different types of *extra-linguistic* context, again for the translation of subtitles. The extra-linguistic context we aim to integrate is film genre, film year, speaker gender and speaker age.

In the first chapter of Part II (Section 5.1), we attempted to adapt MT models to extra-linguistic context (and specifically to speaker gender). The method used, consisting of partitioning the datasets into male and female subsets according to the gender of the speaker, resulted in data sparsity issues, which would be further aggravated if multiple types of context had been taken into account. We therefore revisit the same phenomenon here using alternative strategies that do not have the same drawbacks; contrarily to these previous experiments for the integration of extra-linguistic context, the context will be exploited by the translation model itself, rather than being exploited prior to translation. This offers more flexibility and the possibility for the model to learn how to use the information given, without affecting the amount of training data available. The strategies we test are heavily inspired by those tested in the previous section for the integration of linguistic context (Section 7.1). Our aim is to test whether these strategies also provide an adequate solution for extra-linguistic context and to see how they fare when including multiple types of such context.

A major stumbling block of our previous work on speaker gender (Section 5.1) was the very limited ability to specifically evaluate how well models' were using context. Traditional metrics such as BLEU can indicate gains in other aspects of translation (such as sentence length, general word choice, etc.), without there being an explicit correlation with the degree to which context is effectively exploited. Furthermore, it is often extremely difficult to objectively evaluate how well MT models use the majority of the context types we will study (film genre, film year and speaker age). We therefore choose to evaluate all models on their capacity to correctly adapt to speaker gender, a feature which has an explicit and objective effect on morphological gender marking.

We begin this section by describing the creation of the extra-linguistically annotated data that will be used for our training and test data (Section 7.2.1). As in our previous experiments in Section 5.1, we shall be working on subtitle data enriched with extra-linguistic information obtained from film transcripts, but on a much larger scale. In the next section we review the contextual strategies that will be tested and discuss the implications these strategies have for incorporating extra-linguistic context as opposed to raw (and therefore sequential) linguistic context (Section 7.2.2). The experiments, which make up the remainder of the chapter are described, along with our evaluation methods, in Section 7.2.3. We report our results using two different evaluation strategies: the first using BLEU to evaluate the overall translation quality of the translations produced by each model (Section 7.2.4) and the second using an evaluation strategy targeted the use of speaker gender (Section 7.2.5).

7.2.1 Creation of extra-linguistically annotated data

As previously mentioned, one of the limiting factors of including extra-linguistic context in MT is having the necessary resources to train such models. Not all datasets come with such information, or at least not in sufficiently high quantities to be able to train high quality translation models. In this work, we rely on both unannotated and annotated data to train our models. We exploit the large quantities of unannotated data to pre-train high quality MT models, and then continue the training of these models using contextually annotated data. Our training data is the English-French OPENSUBTITLES2016 parallel corpus, of which we obtain annotations for a subset, comprising four types of extra-linguistic contextual information: film genre, film year, speaker gender and speaker age. In this section, we describe our method of obtaining the contextual annotations, with the use of automatically aligned film transcripts in Section 7.2.1.1. We evaluate the alignments between film transcripts and the subtitles in Section 7.2.1.2, and finally, we provide details and statistics on the distribution of the contextual values in the data (Section 7.2.1.5).

7.2.1.1 Obtaining contextual annotations for subtitle data

The OPENSUBTITLES2016 corpus is already annotated with basic meta-data at the film level: film genres, the release year, and the imdb number, which links the film to other information in the official film database. However the resource does not contain subtitle-level annotations concerning the speakers. For a subset of these films, additional resources exist online that enable us to obtain such annotations. For example, van der Wees et al. (2016), Wang et al. (2016a) and Lison and Meena (2016) all use online fan-produced film transcriptions to provide information about the speakers (their names and gender and where the turn boundaries lie in a dialogue). By automatically aligning the transcriptions found online with the subtitles, such information can be transferred to the subtitles and used as extra-linguistic context for MT. This technique is in fact the same used in our gender adaptation experiments presented in Section 5.1, but on a much larger scale. We use the same principle as these past works to annotate part of our data for information concerning the speakers. We do not use the annotations of previous publications due to copyright restrictions, or because the annotations are not available for our working language pair English-French. However, our extraction process is highly influenced by Lison and Meena (2016) and van der Wees et al. (2016), from whom we borrow certain tools, scripts and resources. The annotation process is made up of the following steps:

1. Collection of structured transcripts:
 - Web scraping to obtain English film and television series transcriptions
 - Extraction of structured transcriptions (speaker names and associated utterances) from HTML
 - Labelling of film and series transcripts with their imdb numbers
 - Mapping of character names to actors names as displayed on the imdb profile of the film or series, and extraction of additional information about the actor playing the character (e.g. gender and date of birth). If the character names are too generic to be mapped to an actor name (e.g. *boy* or *old man*) or cannot be linked to their actor, then we use heuristics to map the names at least to a probable gender. We use lists of gender-specific terms such as *woman*, *man*, *policeman*, etc. and lists of male and female names.
2. Automatic alignment of the transcriptions with the English side of the OPENSUBTITLES2016 parallel corpus using the Champollion sentence aligner (Ma, 2006)
3. Transfer of the speaker-related information (name, gender, age) to the parallel subtitles corpus via the automatic alignments. This preliminary information will

serve as the basis for the construction of our dataset, as described below in Sections 7.2.1.3 and 7.2.1.4.

We use the scraped and structured transcripts provided by Lison and Meena’s extraction process, but choose to re-align the transcriptions and the films. This enables us to know how much of the film could be aligned and therefore to filter out any films that are not sufficiently aligned (a threshold we define at 70%).¹² We shall evaluate the quality of our re-alignment in Section 7.2.1.2.

We divide the remaining films into train, dev and test sets, keeping the best aligned films for the test set. We provide statistics for the annotated datasets in Table 7.6. For comparison purposes, we also provide statistics for the unannotated training data, which is used to pre-train models. A noticeable difference between the unannotated and annotated datasets is the average sentence length: the annotated data (for which online transcriptions exist) has sentences which are on average much longer than in the much larger dataset.¹³ The French subtitles are also marginally shorter than the English subtitles for the annotated data, which is not the case for the unannotated data.¹⁴ These are two factors that models will learn to adapt to when training is continued using the annotated data.

	English			French	
	#sents	#tokens	#tokens/#sents	#tokens	#tokens/#sents
<i>Unannotated OPENSUBTITLES2016 data (for pre-training)</i>					
pre-train	24,140,225	174,593,562	7.2	175,432,942	7.3
<i>Annotated OPENSUBTITLES2016 data (for adapted training)</i>					
train	1,696,040	13,462,830	7.9	13,268,188	7.8
dev	3,000	24,131	8.0	23,524	7.8
test	50,000	399864	8.0	394212	7.9

Table 7.6: Corpus statistics for each dataset, including the unannotated data used for pre-training (pre-train) and the annotated data described in this section (train, dev and test). Token numbers are calculated on pre-processed sentences (to which subword segmentation has been applied), but do not contain any additional tokens indicating contextual values.

It should be noted that not all subtitles are annotated for each feature type. Film information is only available for certain films. Speaker information is almost systematically

¹²Insufficient alignment can be due to a number of reasons: an unpredictable HTML structure for the transcript, leading to difficulties extracting the speakers and the utterances, difficulty separating scene directions from the spoken content of the transcripts, errors in the film subtitles, etc.

¹³This is likely to be due to the nature of the films for which annotations are available. These films are those for which transcriptions are available online, and there therefore is a potential bias towards films with more elaborate dialogues, as opposed to series and soaps.

¹⁴This is potentially due to the limited screen space available for subtitles, which imposes a maximum subtitle length, resulting sometimes in slightly shortened translations.

partial, for a number of reasons, including errors in the alignment process, untranscribed subtitles or an ability to map the character name to an imdb profile and therefore additional information about the speaker. We do not consider this a problem: previous work on the use of pseudo-tokens for contextual information suggests that not systematically using contextual information is useful to avoid over-reliance on the presence of tokens, enabling the model to also be used in situations where the information is not available (Sennrich et al., 2016a).

7.2.1.2 Evaluating the alignments between subtitles and transcripts

To ensure that the data is of sufficient quality, we performed a manual annotation of the quality of the automatic alignments and the speaker values assigned. We randomly selected fifty films from the entire annotated corpus (train, dev or test) and then randomly selected a block of twenty consecutive subtitles from each film, which we evaluated in terms of the quality of the alignment with the film transcript and the values of the speaker information associated with each subtitle. Out of the 1,000 subtitles, 847 are well aligned, 104 are unaligned and 49 are poorly aligned. Out of those that are well aligned, the correct speaker gender is supplied in 729 cases, no gender is supplied in 66 cases and the wrong gender in 52 cases. Although there is a little noise in the data (5% of sentences will receive the wrong gender), a vast majority of sentences that are annotated with gender are assigned the correct one.

7.2.1.3 Film genre, film year and speaker age

We process the annotations obtained by previously described method to add additional values, correct others and agglomerate certain values to avoid sparse labels, as described in this section for film genre, film year and speaker age. We separately discuss values for speaker gender in the next section, as these will be the focus of a particular processing methodology. Table 7.8 provides an example of all contextual values for a sequence of subtitles.

Film genre Film genre values are those provided in OPENSUBTITLES2016. A film can (and often does) have several genre values, or may be associated with no genre at all. There are twenty-one different values of which the most common are *drama* (312,225 films in the train set), *comedy* (168,158 films), *crime* (113,676 films) and *thriller* (103,194). The least common genres are *musical* (2,151 films) and *film noir* (4,746). In order to simplify our experiments and allow for a more controlled analysis, we reduce multiple genre labels to one genre per subtitle, by randomly selecting one of them.

7.2. Contextual NMT with extra-linguistic context

Aligned subtitle (Sub.) and transcription (Trans.)		Correct?
Sub.	They had the girls of our community preparing for months.	✓
Trans.	They had the girls of our community preparing for months.	
Sub.	4 would be chosen for the harvest.	✓
Trans.	Four would be chosen for the Harvest.	
Sub.	They said that it was an honor, that they were special.	✓
Trans.	They said that it was an honor, that they were special.	
Sub.	I thought it was a myth.	✓
Trans.	I thought it was a myth .	
Sub.	Was it?	✓
Trans.	Was it ?	
Sub.	Marcel, bit early in the day for you, isn't it?	✓
Trans.	Marcel . – Bit early in the day for you, isn't it ?	
Sub.	I know.	✓
Trans.	I know , I make this look easy , but I still have an empire to run .	
Sub.	I make this look easy, but I still have an empire to run.	✓
Trans.	I know , I make this look easy, but I still have an empire to run.	

Table 7.7: Examples of the manual evaluation of the automatic alignment between subtitles and transcripts. Since the alignment is only used to transfer the annotations from the transcripts, the segmentation does not need to be identical for the subtitles and transcripts. Note that the final sentence of the transcript in these examples is split over two subtitles.

Film year All films are associated with a release year, and this is mapped to its decade (e.g. 1996 to 1990s) to create more coarse-grained categories. There are 9 different possible decades, from the 1930s to the 2010s.

Speaker age We use the date of birth to calculate the age of the actor at the time the film was released, and we map the ages to discrete categories: infant (<4), child (<10), teen (<20), 20-something (<30), adult (<50), older-adult (<65), elderly (≥ 65). We make the assumption that actor age corresponds approximately to character age.

7.2.1.4 Augmenting speaker gender annotation

For the gender of the speaker, there are five possible values: *male singular*, *female singular*, *male plural*, *female plural* and *female and male plural*; there can be multiple speakers for a same subtitle, although these are relatively rare (cf. Table 7.11). We nevertheless keep them in our training data, although we will not test such values in our targeted evaluation, opting instead to concentrate on the distinction between singular male and female values.

Source/Target	Speaker		Year	Film Genre
	Gender	Age		
EN: You must relax. FR: Vous devez vous détendre .	Male	Older adult	1990s	Adventure
EN: Relax? FR: Je n'ai jamais été aussi détendu !	Male	Adult	1990s	Adventure
EN: If I were any more relaxed, I'd be dead! FR: Plus détendu, tu meurs !	Male	Adult	1990s	Animation

Table 7.8: Examples of the annotation values.

As described previously, annotations obtained using the previously described method are partial, due to non-alignment of subtitles or the character name being difficult to associate with a gender. As we will put a special emphasis on studying how our MT models account for speaker gender, it is in our interest for there to be as many values as possible. We therefore extend and correct gender annotations based on the content of the sentences themselves.

Rule-based speaker gender detection in French sentences Our method of augmenting the number (and quality) of speaker gender annotations is a rule-based approach, applied to the French (target) side of the parallel corpus. We take advantage of the fact that in French, this information can be marked explicitly. When translating from English into French, certain words that are used to qualify the speaker (which are not marked for gender in English) are morphologically marked for gender in French. For example, in (58), the gender-neutral English words *happy* and *nurse* are necessarily gender-marked in French, agreeing with the gender of the speaker.

(58) EN: I am a very happy nurse.

FR_{MASC}: Je suis **un infirmier** très **content**.

FR_{FEM}: Je suis **une infirmière** très **contente**.

Our automatic method detects such gender-marked sentences in French texts and predicts the gender of the speaker from the gender-marked words they contain. This method will also be the basis for our targeted evaluation method, which aims to specifically assess the ability of our models to correctly take into account speaker gender information (cf. Section 7.2.3.2). The method targets a number of different French constructions that can contain gender marking determined by speaker gender. We use pattern matching to identify (i) sentences containing such constructions and (ii) the gender for which they are marked (male, female or underspecified).

Our aim is to cover a sufficient number of cases, without over-predicting. We seek to identify the following cases (each followed by a canonical example):

1. Adjectival agreement

Je suis content(e) “I am happy”

2. Nominal agreement

Je suis votre voisin(e) “I am your neighbour”

3. Past participle with auxiliary être

Je suis allé(e) “I went”

4. Past participle with auxiliary avoir and preceding direct object

Il m’a grondé(e) “He told me off”

We identify these cases using simple, manually defined rules, which are translated into regular expressions. The French text to which the script is applied is first tokenised and PoS-tagged using the MELt tagger (Denis and Sagot, 2012). The rules specify sequences of tokens depending on their properties (wordforms, lemmas and/or PoS-tags).¹⁵

Once a sentence containing one of these speaker-gender-dependent constructions is identified, we detect which gender marking is used: male, female or underspecified (this third category is for adjectives, participles or nouns that are underspecified in terms of gender marking, such as *sociable* ‘sociable’ and *neutre* ‘neutral’). We use a morphological lexicon, the *Lefff* (Sagot, 2010), to identify the gender of the word indicated in bold in each construction, and if no gender can be found, we use heuristics to determine the gender.¹⁶

We evaluate the identification process by randomly selecting 500 sentences detected as containing a gender marking, and manually counting how many times the predicted gender is correct. 490 of the 500 predictions were correct (98%), the breakdown of which is provided in Table 7.9. A majority of errors concern sentences incorrectly identified

¹⁵We illustrate here how we designed these rules with several examples. In the rules shown below, *je* ‘I’, *moi*, ‘me’, *qui* ‘who’, *désolé* ‘sorry’, *avoir* ‘have’ and *être* ‘be’ are lemmas, “past participle”, “adjective”, “noun”, “determiner” are PoS tags and “gap of n - m ” signifies a sequence of words of minimum length n and maximum length m . We mark in bold the token that carries the gender marking that should match that of the speaker:

- *je*, gap of 0-2, *être*, gap of 0-3, **past participle or adjective**
- *je*, gap of 0-2, *être*, gap of 0-3, determiner, **noun**
- *moi*, *qui*, gap of 0-2, *être*, gap of 0-3, **past participle**
- *moi*, *qui*, gap of 0-2, *être*, gap of 0-3, determiner, **noun**
- *me*, *avoir*, gap of 0-2, **past participle**
- **désolé**

Note that the choice to include this final template, containing just the word *désolé(e)* ‘sorry’ was made in order to cover the very large number of occurrences in which this was the only word uttered. It is a case which almost always agrees with the speaker’s gender, which is not the case of other adjectives that appear in sentence-initial position, or alone in an utterance.

¹⁶These heuristics consist in making a decision based on the suffix of the words, identifying final *e* for feminine words, for example. If several (incompatible) gender markings are found in the same sentence, we make the approximation of taking the first one found by one of our templates.

as containing an underspecified participle, adjective or noun, where the sentence in fact contained a male or female gender-marked word. This is in itself encouraging, as it shows that there are very few errors concerning a mix-up between male and female genders. The most common error was an over-prediction of our rules linked to the flexibility of allowing several intermediary words (marked as gaps in our templates). For example *Je suis en rouge et noir, vous devez changer* ‘I am in red and black, you should change’ is incorrectly identified as containing male gender marking due to the word *noir* ‘black’, which is a noun, but in other circumstances can be a masculine adjective. Despite these errors, we consider the identification to be sufficiently accurate for our needs.

Predicted	#Manually identified as...				Total
	MASC	FEM	UNDERSP.	none	
MASC	244	1	2	3	250
FEM	1	199	4	0	204
UNDERSP.	5	4	37	0	46

Table 7.9: A confusion matrix showing the evaluation of our gender marking identification rules, based on a random sample of 500 sentences identified using our method.

Handling mismatches One problem encountered with speaker gender is that the French subtitles often contain agreement errors with respect to gender marking, due to the fact that fan-produced subtitles (and sometimes even official subtitles) contain grammatical errors. When feminine gender agreement only impacts the adjective or participle orthographically and is not reflected in the pronunciation (e.g. *amical*_{MASC} vs. *amicale*_{FEM} ‘friendly’ /amikal/ and *fâché*_{MASC} vs. *fâchée*_{FEM} ‘angry’ /faje/), a common error is to use the masculine variant instead of the feminine one (and the opposite sometimes occurs too). This is problematic for us if we want to learn to produce the correctly marked versions in translation. Instead of correcting these errors in the subtitles themselves, we modify the gender labels to reflect the gender expressed in the subtitle, as identified using the above-described script.

Final speaker gender annotation Applying our speaker gender detection method to our annotated data leads to an increase in the number of annotated labels, and a slight redistribution amongst gender labels, as shown in Table 7.10. The final distribution of speaker gender per dataset is given in Table 7.11.

7.2.1.5 Properties of the final annotated corpus

Table 7.12 shows the number of annotations of each type across the three sets. While film year is available for all sentences, other extra-linguistic features are less available,

7.2. Contextual NMT with extra-linguistic context

Gender	#subtitles in train set		
	before	after	% change
MASC SG	811,743	815,951	+0.5
FEM SG	417,612	423,249	+1.3
MASC PL	1,434	1,416	-1.3
FEM PL	2,539	2,485	-2.1
MASC/FEM PL	20,179	19,657	-2.6
Unannotated	442,533	433,282	-2.1

Table 7.10: The distribution of gender labels in the training set before and after augmenting and modifying the labels based on the linguistic content of the French subtitles.

	#subtitles				
	MASC SG	FEM SG	MASC PL	FEM PL	MASC/FEM PL
train	815,951 (48.1%)	423,249 (25.0%)	1,416 (0.0%)	2,485 (0.1%)	19,657 (1.2%)
dev	1,143 (38.1%)	587 (19.6%)	0 (0.0%)	0 (0.0%)	29 (1.0%)
test	24,440 (48.9%)	13,064 (25.1%)	51 (0.1%)	96 (0.2%)	958 (1.9%)

Table 7.11: The distribution of the gender labels in the train, dev and test sets. Percentages of the total number of subtitles per dataset are given in brackets. Note that speaker gender annotations are not available for all subtitles.

notably speaker age and film genre. The dev set is only used to validate the training progress, and therefore is small compared to the other two sets – its lacks of certain annotations therefore does not pose a problem in terms of the model learning how to use contextual information.

Set	#sents	film genre	#sents annotated for...		
			film year	speaker gender	speaker age
train	1,696,040	508,928	1,696,040	1,262,758	241,439
dev	3,000	0	3,000	1,759	0
test	50,000	15,004	50,000	38,609	7,283

Table 7.12: Corpus statistics per dataset: numbers of sentences and numbers of annotated sentences for each type of extra-linguistic context.

7.2.2 Contextual strategies

We compare two different strategies of including context into NMT. Both strategies have been seen in the previous section for the integration of linguistic context (Section 7.1), with extra-linguistic context assuming the role that the previous sentence occupied in our previous experiments. The first involves prefixing the sentence to be translated with *pseudo-tokens* indicating the values of the contextual information as per Sennrich et al.

(2016a). The second involves encoding the contextual information separately from the source sentence using a secondary encoder and attention mechanism.

7.2.2.1 Pseudo-token strategy

We prepend contextual information as pseudo-tokens to sentences, as initial tokens in the sequence. The extra pseudo-tokens serve to influence the translation, and the approach relies on the MT model learning to incorporate the information when necessary. Examples of data with a single token (indicating speaker gender) are given in Table (7.13).

Source	Target
GENDER-m I should leave .	je dois y aller .
GENDER-m and you found me .	et tu m' as trouvé .
GENDER-f I' m so sorry .	je suis désolée .
GENDER-f I don't know ... what to say.	je ne sais pas quoi dire .

Table 7.13: Examples of training data in which source sentences are prefixed with a contextual token (here indicating speaker gender).

The gender-token strategy has been used previously in the literature for a variety of different contextual features, and variants of the strategy exist concerning the position of the pseudo-tokens in the sentence. Whereas in some works the pseudo-tokens are appended to the sentence (as side-constraints) (Sennrich et al., 2016a), others prepend the token to the sentence (Jehl and Riezler, 2017), as we propose to do here. A comparative study by Takeno et al. (2017) of the two approaches suggests that the prefix strategy provides more flexibility, although the difference does not appear significant. We choose to adopt the prefix approach to remain consistent with our experiments in the previous section. The strategy has been previously used to control politeness in the translation of English *you* into German *du* (informal) and *Sie* (polite) (Sennrich et al., 2016a), by first predicting which form should be used, and then encoding the value as a *pseudo-token* at the end of the source sentence. They find that the method is very effective in controlling which form is used, and if the gold formality labels are used, significant improvements can also be seen to the BLEU score. Similar experiments have been conducted for various other types of information: sentence length by Takeno et al. (2017) and topic information for patent translation by Jehl and Riezler (2017). The most similar experiment to ours is the integration of speaker gender information for English-to-Arabic translation by Elaraby et al. (2018). They see improvements in BLEU score, particularly for those sentences that are gender-marked. However we aim to go further here, providing a more detailed analysis of our results, and also looking at the interaction of several labels.

We test the pseudo-token strategy for the integration of various types of extra-linguistic context (speaker gender, speaker age, film genre and film year). This is an intuitive way to

encode the unstructured linguistic context (cf. 2-TO-1 and 2-TO-2 models of Section 7.1), since the current sentence is a natural continuation of the previous sentence and is composed of sequential tokens. However the same cannot be said for a set of extra-linguistic features, for which there is no inherent notion of order (and does not constitute a real sentence). The use of this method is therefore somewhat counter-intuitive with respect to the inclusion of multiple tokens. One of our aims will therefore be to observe how the order of pseudo-tokens can affect how well contextual information is exploited.

In light of our observations concerning the important role of the decoder in exploiting contextual information in Section 7.1, we test two variants of each model, corresponding respectively to the 2-TO-1 and 2-TO-2 strategies of Section 7.1. In other words, for each information type, we train two models, one for which the pseudo-tokens are prefixed just to the source side of the data (*src*), and the other for which the pseudo-tokens are included on both the source and target sides (*src+trg*). The second model type is therefore trained to translate the pseudo-token(s) in the MT output, which is then post-processed to remove the contextual tokens. An example of the training data for the *src+trg* model is given in Table 7.14.

Source	Target
GENDER-m I should leave .	GENDER-m je dois y aller .
GENDER-m and you found me .	GENDER-m et tu m' as trouvé .
GENDER-f I' m so sorry .	GENDER-f je suis désolée .
GENDER-f I don't know ... what to say.	GENDER-f je ne sais pas quoi dire .

Table 7.14: Examples of training data in which source and target sentences are prefixed with a contextual token (here indicating speaker gender).

7.2.2.2 Multiple encoders

An alternative strategy is the use of an additional encoder and attention mechanism to encode the extra-linguistic information separately from the source sentence. This strategy could have the advantage of not increasing the length of the sentence being translated, and of treating context as a separate input, with its own specific learnt parameters.

We reuse the HIER-TO-2 model from Section 7.1, which consists of two separate RNN encoders (each with its own attention mechanism), the first one used to encode the contextual information and the second to encode the sentence to be translated. This model, which we refer to as *multi-seq-gender-age-year-genre*, retains the sequential nature of the context being encoded, since the order of the tokens must be fixed in advance. We therefore compare this to a similar architecture, in which we remove the recurrent element from the first encoder (whose role is to encode the extra-linguistic context),

the aim being to remove the counter-intuitive notion of order from the context. We refer to this architecture, shown in Figure 7.4 as *multi-nonseq-gender-age-year-genre*. As before (with both s-HIER and *multi-seq-gender-age-year-genre*), an additional attention mechanism calculates a context vector $c_1^{(j)}$, designed to represent the extra-linguistic context at each decoding step j , but this is calculated over projections of the embeddings of the contextual features rather than over annotation vectors as with the source sentence. All notation is the same as presented in Section 3.2 and Section 7.1 and all W 's, U 's and b 's are learned parameters.

Therefore, at each decoding step j , the context vector $c_1^{(j)}$ is calculated as follows:

$$e_1^{(i)} = \tanh(W_a^\top z^{(j)} + U_a h^{(1,i)}) \quad (7.7)$$

$$\alpha_1^{(ij)} = \frac{\exp(e_1^{(i)})}{\sum_{i'}^N \exp(e_1^{(i')})} \quad (7.8)$$

$$c_1^{(j)} = \sum_{i=1}^K \alpha_1^{(ij)} h^{(1,i)} \quad (7.9)$$

The representations of each token $h^{(1,i)}$ are calculated as follows:

$$h^{(1,i)} = \tanh(W_h q^{(1,i)} + b_h) \quad (7.10)$$

The context vector $c_2^{(j)}$ for the main source sentence is calculated as described in Section 3.2 (using a recurrent encoder). The two context vectors are then combined using a hierarchical attention mechanism (Libovický and Helcl, 2017), as follows (copied from Section 7.1), where K is the number of encoders (here equal to 2):

$$e_k^{(j)} = v_b^\top \tanh(W_b z^{(j-1)} + U_k c_k^{(j)} + b_e) \quad (7.11)$$

$$\beta_k^{(j)} = \frac{\exp(e_k^{(j)})}{\sum_{k'=1}^K \exp(e_{k'}^{(j)})} \quad (7.12)$$

$$c^{(j)} = \sum_{k=1}^K \beta_k^{(j)} V_k c_k^{(j)} \quad (7.13)$$

7.2.3 Experiments

7.2.3.1 Experimental setup

We train all systems using NEMATUS (Sennrich et al., 2017), using the same hyper-parameters as in Section 7.1 (detailed in Appendix A.1.1). We pre-train all models using the large unannotated OPENSUBTITLES2016 training set, which is first filtered to remove

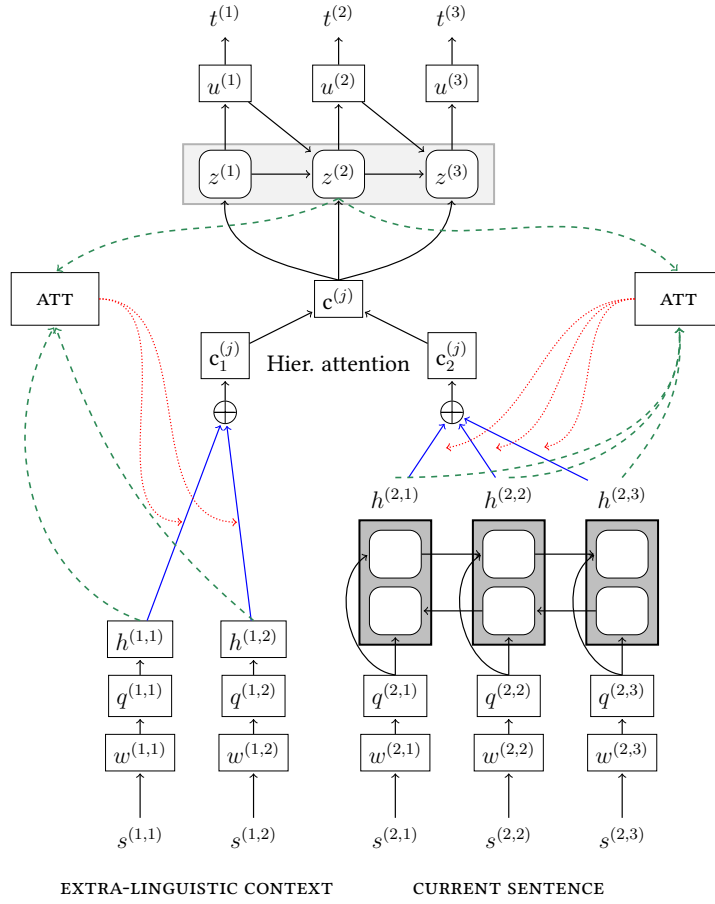


Figure 7.4: Our proposition for a multi-encoder model in which the recurrent unit of the first encoder (for the encoding of contextual tokens) is removed. The attention mechanism applies to projected versions of the token embeddings in order to calculate $c_1^{(j)}$.

poorly or partially translated parallel sentences, and then continue training of this model using the specific annotated dataset. Additional details about the experimental setup can be found in Appendix A.2.1.

As described in Section 7.2.2, we compare two different strategies for including the information: the pseudo-token approach and the use of a separate secondary encoder. The pseudo-token approach uses a baseline NMT architecture with a single encoder. The contextual information is integrated by prepending it to the beginning of the training sentences to which it corresponds. We test several different models using this approach, trained on different numbers and different types of contextual information: (i) single-feature models to test the effect of each of the extra-linguistic context types individually, (ii) speaker gender and film genre models to test the effect of adding two features (and in different orders), and (iii) multiple feature models (successively including film year and speaker age) to test how adding additional information can affect the performance. The multi-encoder models are tested for the integration of all four types of extra-linguistic information. As mentioned in the previous section, we compare two strategies, one

for which the additional encoder is an RNN encoder identical in structure to the main sentence encoder (*multi-seq-gender-age-year-genre*), and a second for which the recurrent part of the additional encoder is removed, removing the notion of order from the tokens (*multi-nonseq-gender-age-year-genre*).

For the pseudo-token approach, the pre-trained model is a single-encoder model trained without any contextual values. For the multi-encoder approach, the pre-trained model is a multi-encoder HIER model, whose contextual encoder is trained on a sequence of four distinct tokens that act as placeholders for each of the four contextual tokens that will be provided in the second training step. This multi-encoder pre-trained model is not trained to reproduce the pseudo-tokens in the target sentences. We continue the training of the model using our annotated data, systematically providing four contextual tokens for each sentence. Whenever a sentence lacks one or more contextual values, we replace them by the corresponding placeholders used in the pre-training step.

We have seen in Table 7.6 that the annotated datasets have slightly different properties compared to the large unannotated data used for pre-training, to which our models will have to adapt. To ensure a fair comparison, we continue training our pre-trained model using the same data as for our contextual models (but excluding any contextual features). The result will be our “baseline” model, although we shall also compare our results to those of the pre-trained model.

Each approach can be tested in two scenarios, one in which the source sentence only is prefixed with the token (*src*) and the other in which the model is also trained to reproduce the token in the target sentence (*src+trg*). Our initial experiments (i.e. our single feature models) use both strategies in order to compare them. For our later experiments, we only use the *src+trg* strategy, as it appears to produce the better results on these first experiments.

7.2.3.2 Evaluation methods

We evaluate the contextual models using two methods of evaluation. The first is an evaluation of overall translation quality using BLEU. This enables us to see any significant changes in the translation quality when adding specific information in a certain way. The second evaluation is a targeted evaluation of speaker gender in order to assess how well each model exploits speaker gender information.

We choose to provide this targeted evaluation as it provides us with interpretable results about the way context is being exploited. This differs from the case of standard metrics such as BLEU, which offer relatively little insight into how context is being used. Ideally, it would be useful to provide a targeted evaluation of each of the different extra-linguistic feature types. However for many types of extra-linguistic context (genre, style,

politeness, era, speaker age), it is not clear what the desired influence of the context may be.¹⁷ Intuitively, they are likely to influence aspects such as word choice or syntactic formulations. However it may be difficult to objectively analyse such improvements, particularly in terms of their adequacy with respect to the extra-linguistic context provided. For example, evaluating the adequacy of a translation with respect to the age of the speaker is very subjective: which type of language is adapted to each age group? Do we even expect to see a difference in the style depending on the age of a speaker? The answer to these questions is neither very clear nor safe from subjectivity and individual preferences. However, speaker gender is different, as it is at least partially possible to objectively evaluate how much an MT model succeeds in using this information when translating into a language with explicit speaker gender marking. It is an aspect that we attempted to evaluate in Section 5.1, with the downside that in our previous experiments there was an insufficient number of sentences containing such gender marking.¹⁸

Our targeted evaluation relies on the same principle as our data augmentation method described in Section 7.2.1.4. We rely on the fact that in the French translations, certain words are marked for speaker gender. By focusing uniquely on these sentence types, we can study how effectively MT models can be biased in favour of either male or female gender marking when the different context values are provided as input. Each model is used to decode two versions of the test set, one in which speaker gender is systematically set as male for all speakers and the other in which speaker gender is systematically set as female. To each of the sets of translations we apply the same speaker gender detection and identification script described in Section 7.2.1.4. From these automatically identified genders, we can compare the percentage of sentences for which the model successfully uses gender marking that matches the speaker gender assigned as input.

7.2.4 BLEU score results

7.2.4.1 First impressions: improvements in BLEU

We begin by evaluating each model using BLEU, applied to the translated test set, to get an overall picture of the translation quality of our models as assessed by this metric.¹⁹

The results are shown in Table 7.15. The BLEU score of our pre-trained model on which the other models are based is 30.17, which shows that it is a high performing model. Each of the single-feature contextual models improves upon this score. Film year achieves the

¹⁷It should be noted that when dealing with translations, the identity of the translator can inevitably have an impact on the choice of language used. This is however not an aspect that we have the luxury of studying, as it is rare to have any let alone detailed information about the translators.

¹⁸Given the problems in Section 5.1, we do not evaluate these early approaches on the larger dataset.

¹⁹The BLEU score is calculated on cased translations using the `multi-bleu-detok.perl` script from the Moses toolkit.

best BLEU scores for the *src* setup, followed by speaker gender. In the *src+trg* scenario, the two features achieve comparably high BLEU scores of 30.65 and 30.67 respectively. Given the very large size of our test set (50,000 sentences), even small gains can be significant, as calculated using a paired bootstrap resampling significance test (Koehn, 2004b). We indicate those results for which improvements in BLEU are statistically significantly better than the baseline model (* for $p < 0.05$ and ** for $p < 0.01$).

As we saw in Section 7.1 for the integration of linguistic context, decoding the context appears to improve BLEU scores. There are systematic increases in BLEU score (ranging from between +0.2 to +0.32) for the *src+trg* models when compared to their corresponding *src* model. As a result, we only train multi-feature pseudo-token models in *src+trg* mode.

The BLEU score of the baseline model also improves on that of the pre-trained model, with similar improvements to those seen when adding the contextual tokens in the *src* setup, confirming that gains cannot be entirely linked to the use of contextual information, and are also due to other effects, such as training on more adapted data.

The pseudo-token approaches integrating several different tokens also achieve BLEU scores significantly better than the baseline model, although the quality varies according to which tokens are used. However, the BLEU score of the two multi-encoder models (*multi-seq-gender-age-year-genre* and *multi-nonseq-gender-age-year-genre*) are significantly lower than the other models. This is most likely linked to the way in which the models are pre-trained: the parameters of the additional encoder for the contextual information must be initialised during pre-training, and we choose to do this using tokens indicating that each of the features was not present. Given that these tokens were the same for all sentences, it is unlikely that the parameters were initialised well during pre-training, thus penalising the final models produced. If the multi-encoder strategy is to be used in a real-life scenario to integrate context, an alternative strategy for initialising the parameters should be used.

7.2.4.2 Further analysis of BLEU improvements

Are these gains entirely attributable to the better exploitation of context? We provide a first test of this by training three additional models, whose training data is artificially manipulated to include extra informationless tokens or to add noise to the data:

- **SAME-GENDER-ALL:** each sentence in the data is systematically prefixed with the same pseudo-token, which is therefore semantically void.
- **SAME-GENDER-SUBSET:** only the sentences that were annotated for gender are systematically prefixed with the same pseudo-token. The length of the corpus is therefore identical to the one used when gender is being integrated.

	BLEU	
	src	src+trg
gender	30.35	30.67**
age	30.27	30.44*
genre	30.25	30.45*
year	30.42	30.65**
gender		30.67**
gender-genre		30.71**
genre-gender		30.61**
gender-year-genre		30.52*
gender-age-year-genre		30.81**
multi-seq-gender-age-year-genre		28.26
multi-nonseq-gender-age-year-genre		28.63
pre-trained	30.17	
baseline	30.31	

Table 7.15: BLEU scores on our test set when prefixing each sentence with each of the specified features (as per the pseudo-token approach). *src* indicates that the model is only trained on data in which the sources sentences are prefixed with the token, whereas *src+trg* indicates that the model is also trained to translate the prepended feature in the target translation. The first-, second- and third-best results in each section of the table are indicated in increasingly light shades of green. Results that are statistically significant from that of the baseline model are indicated (* for $p < 0.05$ and ** for $p < 0.01$).

- RANDOM-GENDER-SUBSET: speaker gender tokens are randomly assigned amongst the sentences that were originally annotated for speaker gender. The same distribution of genders is kept as when gender is being integrated, but their association with the sentences is shuffled.

The results (Table 7.16) show that all three strategies result in insignificant gains in BLEU score compared to the pre-trained model when the tokens are only included on the source side of the data (*src*). However, they all systematically result in significant improvements when the model also learns to translate the tokens (*src+trg*). The largest gain is seen when an extra token is systematically added to all source and target sentences. The effect of also decoding the token (*src+trg*) is just as great (+0.29) as seen when semantically loaded contextual tokens are used. These results give us some food for thought, and also put into question the evaluation strategies of many previous experiments in the literature. Our results show that BLEU alone cannot be used to evaluate how models exploit context, as it is subject to other factors linked to the approach used.

So what part of the approach led to this increase in BLEU score? One possibility is that the gains seen are an artefact of changing the length of the source and target sentences during training. An indication that this could be the case is the effect that changing the sentence length (through prefixing) has on the final length (in tokens) of the translations.

Prefix...	src	src+trg
SAME-GENDER-ALL	30.27	30.56**
SAME-GENDER-SUBSET	30.28	30.45*
RANDOM-GENDER-SUBSET	30.20	30.46*
pre-trained	30.17	
baseline	30.31	
gender	30.35	30.67**

Table 7.16: See caption of Table 7.15. This result corresponds to models trained on data that is artificially prefixed with different informationless tokens or to which noise has been added. We include the results of the pre-trained, baseline and gender experiments for comparison purposes. Results that are statistically significant from that of the baseline model are indicated (* for $p < 0.05$ and ** for $p < 0.01$).

	BLEU		$l_{\text{hyp}} - l_{\text{ref}}$	
	src	src+trg	src	src+trg
gender	30.35	30.67**	+6,038	+625
age	30.27	30.44*	+5,941	+3,427
genre	30.25	30.45*	+5,237	+1,263
year	30.42	30.65**	+4,632	-1,699
pre-trained	30.17		+20,150	
baseline	30.31		+5,116	
SAME-GENDER-ALL	30.27	30.56**	+4,859	+932

Table 7.17: The effect of each feature and mode (src or src+trg) on BLEU score achieved by single-feature models (cf. Table 7.15), and the length difference between the reference test set ($l_{\text{ref}} = 361,377$ tokens) and its translation by each model (l_{hyp}). The three smallest length differences are indicated by increasingly light shades of green. Results that are statistically significant from that of the baseline model are indicated (* for $p < 0.05$ and ** for $p < 0.01$).

Table 7.17 shows the length difference between the reference translations of the test set and the translations produced by each single-feature model. All translations have been previously post-processed and therefore do not include any pseudo-tokens. With the exception of the model trained using a *year* prefix, all models produce translations that are longer than the reference translations. The pre-trained model produces the longest translations (+20,150 tokens compared to the reference translations), which is largely reduced when the training of the model is continued. An interesting observation is that when the model is trained to translate the pseudo-token in the target sentences, the resulting translations (once the pseudo-token has been removed) are systematically shorter than when the model is not trained to translate the token.

So how does this affect the BLEU score? The length difference actually has little direct difference on the BLEU score, since in most cases the hypothesis translations remain longer than the reference translations – no brevity penalty is applied to the scores. The

only exception is the *year-src+trg* model, which is slightly penalised for producing short translations. The BLEU scores are systematically higher for these *src+trg* models, and this holds even when the token is identical for every single sentence (baseline+token). Even when the score is penalised due to short translations (the case of *year-src+trg*), the BLEU score is higher than its *src* equivalent. This increase in BLEU score is due in each case to marginally higher unigram, bigram, trigram and quadrigram precisions (as calculated by BLEU). To some extent, this may suggest that translating the token enables the model to produce translations that are shorter, using formulations that are slightly more similar to the reference translations. A hypothesis could be that, given that the French subtitles are globally shorter than the English subtitles, adding a pseudo-token to both sides of the data during training reduces the length ratio between the source and target sentences, enabling the model to more easily learn to produce translations of a more appropriate length. Although we leave further investigations to future work, this enables us to again confirm that the BLEU score is not an appropriate measure of how well context is being used. Results can easily be misleading and even significant gains in BLEU score can be a side-effect of an initial design choice. Further evaluation is therefore necessary in order to ascertain whether the context is being used by the translation model.

7.2.5 Targeted evaluation of speaker gender

To see whether and how the context really has an impact on the use of context in translation, we apply our targeted evaluation strategy (described in Section 7.2.3.2) to a range of different scenarios. We use each of the models to translate two versions of the test set, one in which all speakers are set to male and one in which all speakers are set to female. We use the automatic gender-marking-detection script described in Section 7.2.1.4 to estimate from each set of translations the percentage of sentences that correspond to the assigned gender. A translation model that exploits the information perfectly is expected to contain gender-marked sentences that are all of the assigned gender for the test set or are underspecified. Since the set of gender-marked sentences is not guaranteed to be the same for each translation model, we report results only on the 926 sentences that are gender-marked for all models.²⁰

For comparison purposes, we show in Table 7.18 the results of the pre-trained model and the baseline model (with continued training). Note that neither model has access to speaker gender information.

²⁰This is likely to introduce a bias towards certain types of sentences for which producing a gender-marked construction is one of the few options available, as opposed to produce a paraphrased version of the same translation that does not contain such a construction.

	MASC	FEM	UNDERSP.
pre-trained	79.5	13.6	6.9
baseline	58.9	34.2	6.9

Table 7.18: The percentages of gender-marked translations produced by the pre-trained and baseline models.

7.2.5.1 Impact of translating the pseudo-token

In our previous experiments in Section 7.1, learning to translate the contextual elements (i.e. reproducing them in the target sentence) increased the capacity of the model to exploit context correctly. We have seen here that learning to translate the tokens (*src+trg*) gives higher BLEU scores than when they are not translate (*src*), but that part of the increase in BLEU score is likely due to factors concerning translation length. We therefore test here whether there are real improvements in the way in which context is exploited. Looking only at the model incorporating speaker gender, we compare the two strategies, *src* and *src+trg*, using our targeted evaluation method.

Feature	src				src+trg			
	%MASC	%FEM	%UNDERSP.	%match	%MASC	%FEM	%UNDERSP.	%match
<i>All speakers are set as male</i>								
gender	87.4	5.6	7.0	94.4	89.2	4.1	6.7	95.9
<i>All speakers are set as female</i>								
gender	10.0	83.3	6.7	90.0	8.6	84.7	6.7	91.4

Table 7.19: The result of the targeted speaker gender-based evaluation. Results are given as percentages of sentences containing a potentially gender-marked construction that are marked for each of the genders (MASC for masculine, FEM for feminine and UNDERSP. for underspecified). All these results are significantly better than the baseline predictions at $p < 0.01$. The *src+trg* models are significantly better than the *src* variants at $p < 0.01$ for male speakers and $p < 0.05$ for female speakers.

The results (Table 7.19) show that both methods work well in re-producing the correct gender markings when all speakers are indicated as either male or female.²¹ When all speakers are set to male, between 87.4% and 89.2% of sentences marked as gender-marked are identified as being male-gendered-marked. When those that are underspecified are also included (we refer to this total score as %match), the percentage of sentences compatible with male speakers goes to 94.4% and 95.9%. The higher results are obtained for the *src+trg* model. The same pattern is seen when all speakers are set to female, although the results are slightly lower: 90.0% of gender-marked sentences match for

²¹In this table and in the following tables, statistical significance is measured using the McNemar mid-p test (Fagerland et al., 2013).

src and 91.4% for *src+tg*. The second method, *src+trg* again gives higher percentages of matching genders, confirming that this method does appear to help the use of context.

7.2.5.2 Two-feature models and the impact of pseudo-token order

Table 7.20 shows that adding an extra token indicating the film genre next to the token indicating speaker gender leads to little change in how gender information is being learnt by the model. Interestingly, there is also little significant difference between the performance of the models when the order of the tokens is changed (*genre-gender* vs. *gender-genre*), indicating that changing the order when two tokens are used is not problematic. This goes against our intuitions that imposing different orders on the tokens would have a significant impact on how the context is exploited.

7.2.5.3 Three- and four-feature models

The results in Table 7.21 show that adding additional tokens does not lead to significant differences in the models' ability to exploit speaker gender, despite there being a variable number of tokens to handle. Although the differences between the models are marginal, it is interesting to note that the model including gender, year and genre tokens appears to be the best scoring of all models, a result that is counter-intuitive and not easily explained.

7.2.5.4 Multi-encoder models

Taking the HIER-TO-2 architecture from Section 7.1, we test how extra-linguistic context can be exploited when encoded using an additional RNN encoder (we refer to this model as *multi-seq-gender-age-year-genre*). The results, shown in Table 7.22, show that the multi-encoder strategy works well in encoding context, producing scores that are comparable to the best scores obtained using the pseudo-token strategy, and even slightly higher.

We next look at the scores produced by our variant of this model, presented in Section 7.2.2. This model removes the sequential component of the contextual encoder, whose presence is non-intuitive when dealing with non-ordered sets of contextual tokens. The results obtained for this model (97.3% for male speakers and 93.2% for female speakers) are the highest results of all the models we test, suggesting that this way of encoding the information is as good, if not better than the other two methods tested: the pseudo-token approach and the multi-RNN-encoder approach.

Feature	src+trg			%match
	%MASC	%FEM	%UNDERSP.	
<i>All speakers are set as male</i>				
gender	89.2	4.1	6.7	95.9
gender-genre	88.8	4.5	6.7	95.5
genre-gender	89.2	3.9	6.9	96.1
<i>All speakers are set as female</i>				
gender	8.6	84.7	6.7	91.4
gender-genre	8.6	85.1	6.3	91.4
genre-gender	9.2	84.2	6.6	90.8

Table 7.20: Results of the targeted evaluation of speaker gender, with the addition of an extra token (film genre). The first-, second- and third-best results are indicated by increasingly light shades of green. The difference between the three models is not significant (with $p < 0.05$).

Feature	src+trg			%match
	%MASC	%FEM	%UNDERSP.	
<i>All speakers are set as male</i>				
gender	89.2	4.1	6.7	95.9
gender-genre	88.8	4.5	6.7	95.5
genre-gender	89.2	3.9	6.9	96.1
gender-year-genre	90.1	3.5	6.5	96.5
gender-age-year-genre	89.1	4.0	6.9	96.0
<i>All speakers are set as female</i>				
gender	8.6	84.7	6.7	91.4
gender-genre	8.6	85.1	6.3	91.4
genre-gender	9.2	84.2	6.6	90.8
gender-year-genre	8.4	85.3	6.3	91.6
gender-age-year-genre	8.9	84.7	6.5	91.1

Table 7.21: Results of the targeted evaluation of speaker gender, with the addition of an additional tokens (genre, year and speaker age). The first-, second- and third-best results are indicated by increasingly light shades of green. The difference between the five models is not significant (with $p < 0.05$).

Feature	src+trg			%match
	%MASC	%FEM	%UNDERSP.	
<i>All speakers are set as male</i>				
gender	89.2	4.1	6.7	95.9
gender-genre	88.8	4.5	6.7	95.5
genre-gender	89.2	3.9	6.9	96.1
gender-year-genre	90.1	3.5	6.5	96.5
gender-age-year-genre	89.1	4.0	6.9	96.0
multi-seq-gender-age-year-genre	89.6	3.7	6.7	96.3
multi-nonseq-gender-age-year-genre	91.0	2.7	6.3	97.3
<i>All speakers are set as female</i>				
gender	8.6	84.7	6.7	91.4
gender-genre	8.6	85.1	6.3	91.4
genre-gender	9.2	84.2	6.6	90.8
gender-year-genre	8.4	85.3	6.3	91.6
gender-age-year-genre	8.9	84.7	6.5	91.1
multi-seq-gender-age-year-genre	7.5	86.2	6.4	92.5
multi-nonseq-gender-age-year-genre	6.8	86.6	6.6	93.2

Table 7.22: Results of the targeted evaluation of speaker gender. A comparison with the multi-encoder strategies. The first-, second- and third-best results are indicated by increasingly light shades of green. The two multi-encoder models are not significantly different between each other for female speakers but are marginally significant for male speakers (with $p < 0.05$). For *multi-seq-gender-age-year-genre*, it is only significantly better than *genre-gender* for female speakers (with $p < 0.05$). However, *multi-nonseq-gender-age-year-genre* is significantly better than all pseudo-token approaches except *gender-year-genre* (with $p < 0.05$).

7.2.5.5 Comparison of predictions on gold data

In Table 7.23 we provide some summarising results, as calculated on the test set using the gold labels. For each model we provide the percentage of gender-marked translations that match the gold label in our data (or contains one of our constructions that is invariably marked). This percentage is calculated on the subset of sentences for which all models (including the reference translations) are gender-marked – a total of 529 sentences. We also calculate the BLEU score for this subset of sentences, and the BLEU score for the overall test set, containing 50,000 sentences.

In terms of their ability to produce gender agreement that matches the gold labels, the models are relatively consistent with respect to the previous tables of results produced. The best scoring model is *gender-year-genre*, and there is little difference between the other models. One model does stand out: the multi-encoder model performs less well on the gold labels than in previous results, which is surprising, and is possibly a consequence of different translation choices being made (and of the subset of sentences evaluated here).

	Gendered sents.		All
	%match	BLEU	BLEU
pre-trained	59.7	30.54	30.17
baseline	62.6	30.75	30.31
src+trg models			
gender	95.8	35.50	30.67
gender-genre	96.0	35.81	30.71
genre-gender	95.8	35.54	30.61
gender-year-genre	96.4	36.31	30.52
gender-age-year-genre	96.2	36.49	30.81
multi-seq-gender-age-year-genre	94.7	33.11	28.26
multi-nonseq-gender-age-year-genre	97.4	35.74	28.63

Table 7.23: The results of the models using the gold data. %match indicates the percentage of gender-marked sentences that match the gold labels (calculated on the same subset for all models), and the BLEU score is given for the gender-marked subset and for the whole test set.

As for the models using the pseudo-token approach, contrary to our predictions, the models do not always suffer when extra tokens are added – this appears to depend on the type of context being integrated, although there is no clear pattern as to which type of context is detrimental or beneficial to performance. Speaker gender adaptation appears to be more successful when the *year* and/or *age* tokens are included, producing some of the highest results. The highest scoring model according to BLEU is *gender-age-year-genre* (*src+trg*), with a score of 36.49 on the gendered subset and 30.81 on the test set overall. This represents a +0.64 improvement over the pre-trained model, +0.50 over the baseline and +0.25 over the best non-contextual model (SAME-GENDER-ALL). The order of tokens does appear to have a slight effect on the BLEU scores, but yet again, this does not appear to be systematic. Despite the *gender-genre* and *genre-gender* models having very similar results in terms of their adaptation to speaker gender, the second model appears to suffer marginally in terms of BLEU score. This leads to a question for future work: is this slight difference the result of the order difference in the tokens or is it down to chance?

7.2.6 Conclusion and perspectives

These experiments confirm that the contextual NMT strategies introduced in Section 7.1 are well adapted to incorporating extra-linguistic information. Unlike our experiments in Section 5.1, the translation models presented in this section show their capacity to exploit speaker gender, and do so at high rates. The flexibility of the NMT architecture allows us to include the information without reducing the overall translation quality due to data sparsity. The pseudo-token strategy, which we considered to be counter-intuitive with respect to the inclusion of non-sequential extra-linguistic context, works in fact

surprisingly well in all scenarios, resisting the stress tests we provided. The performance of the multi-encoder models is more variable. Although they score highly in their ability to adapt to speaker gender, their overall translation quality is significantly lower than the other models, and the baseline model. This shows that the correct exploitation of context is not always directly linked to BLEU scores. As previously described in Section 7.2.4, this lower BLEU score could be linked to the way in which the models are pre-trained and an alternative strategy will have to be found to better initialise the parameters of the additional encoder, if this setup is to be kept.

We are able to show tangible differences between model performances depending on the strategy used and the type of context used by the model. By targeting one of the contextual features, speaker gender, we are able to provide an objective evaluation of the models' ability to use context, and observe the effect of perturbing the model by adding additional contextual tokens. Importantly, we have further questioned the role of BLEU, but also rely on it as a verification of global translation quality, enabling a compromise between correctly exploiting context and ensuring high translation quality.

A number of questions have been brought to light in this section, and we have yet to provide conclusive answers for them. Whilst we have seen that pseudo-token order does not have a huge effect on how context is used, it can lead to fluctuations in overall translation quality, as shown by BLEU. It would be interesting in future work to piece apart why this is, and in what situations it occurs. Contrarily to what we expected, adding extra tokens does not necessarily harm performance, and can even be beneficial to both the exploitation of context and the overall translation quality. Again, we have as yet been unable to show why adding extra contextual tokens may help the exploitation of one of these tokens in particular, and this could be useful to study. Finally, if multiple encoder strategies are to be considered as a viable alternative for encoding extra-linguistic information, a different strategy must be envisaged for the initialisation of the parameters in our pre-trained model.

There are potential improvements that could be made to such models in the future, concerning the way that the extra-linguistic context is provided to the model. In the current setting, there is a potential for noise in the data, due to the fact that speaker gender is provided even for sentences for which it is objectively not necessary for gender markings in the target sentence. It could therefore be useful to see what the effect would be to only include gender in sentences for which it is likely to be necessary. The advantage would be marginally faster training and translation, but especially could lead to a better association of gender to certain constructions. On the other hand, as mentioned at the beginning of this chapter, extra-linguistic context (including speaker gender) is likely to have an effect on some of the translation choices made that is less interpretable than gender marking. To gauge how great the effect is, and importantly which aspects of the translation are most affected by each type of context, it would be interesting to provide

an in-depth analysis of the impact of these factors on reference translations. These two opposing directions might provide us with some insight into how useful the different extra-linguistic features may be, and if so, why.

7.3 Conclusion

The advances in MT technology have allowed us to explore several new methods of integrating context within the translation model itself, rather than resorting to pre-processing and post-processing techniques. Given that the exploitation of context is carried out by the MT model itself, we can afford to provide linguistic context in an unstructured manner, either by integrating it as part of the source sequence or encoding it in parallel using a similar encoder to the one used for translation. We have seen in this chapter that the approaches used (concatenation of the context to the source sentence and the use of an additional encoder) were effective for the exploitation of both linguistic and extra-linguistic context. The results presented in Section 7.2 go against our intuitions that providing the context sequentially as if they formed part of the input sequence would be problematic for their effective exploitation; all pseudo-token approaches produced good results for both overall translation quality according to BLEU and a targeted evaluation of speaker gender adaptation.

The relative effectiveness of the approaches despite the limited structure in the contextual information provided to the model does not mean that adding additional structure would not be beneficial, in particular for linguistic context. This may enable us to integrate context from a wider field, and to better exploit the local context that we have already tried to integrate into our models.

The integration of linguistic and of extra-linguistic context is not mutually exclusive. In a real-life setting, in which MT is used to mediate conversations between speakers of two different languages, we would ideally want to provide translations that are both adapted in terms of the linguistic dialogue context and in terms of the scenario and properties of the speakers (cf. the setting of our experiments in the following chapter).²² Future work could therefore look at effectively ways of combining these different types of context.

²²In our following experiments for human evaluation of MT collected during bilingual dialogues, a variety of extra-linguistic information (gender, age and topic information) as well as previous linguistic dialogue history are both available. The models used in the following chapter are not adapted to include all such information, although this is certainly something that should be done in the future.

DIABLA: A corpus for the evaluation of contextual MT

Evaluating MT outputs and analysing errors is fundamental to guiding improvements in MT system design. In the previous chapter, we isolated particular phenomena to better understand how well each of the models uses contextual information. We looked at both linguistic context (for anaphora translation and lexical choice) and extra-linguistic context (for the adaptation of translation into French to speaker gender) in controlled settings. Another way of evaluating MT models, which can be complementary to any other evaluation techniques used, is to apply the MT models in one of their end settings (i.e. in our case to mediate informal written dialogues), and collect judgments from the users as to how well they perform. By comparing two models in the same setting, human judgments can be compared in order to evaluate the impact of the differences in design of the two models. Moreover, the dialogues and accompanying human judgments collected constitute a useful resource for further analysis and research in MT.

In this direction, in this chapter we provide an evaluation method relying on human evaluators participating in bilingual MT-mediated dialogues, and compare a baseline NMT model to one of our previously studied linguistically contextual models. The spontaneous MT-mediated dialogues produced are collected, along with their sentence-level human evaluations to produce a corpus of 144 dialogues containing 5,748 sentences. The resulting corpus, which we name DIABLA (for *Dialogue BiLingue* ‘BiLingual Dialogue’), provides us with an alternative strategy for evaluating our MT systems. It also importantly provides us with spontaneously produced informal data, which will be useful for future research into language analysis and MT errors.¹

¹The corpus will be freely distributed under a CC BY-SA 3.0 licence on publication of the work.

We begin by describing our methodology for collecting dialogues and human judgments (Section 8.1). The design of the protocol is instrumental to successfully collecting dialogues that are natural, spontaneous and anonymous as well as evaluation judgments that will enable us to draw conclusions about the quality of the MT models used. Our aim is for this protocol to be reused in the future to produce similar evaluation datasets for new languages and new MT models. In Section 8.2, we describe some properties of the dialogues collected, with examples of the types of language used, before providing an analysis of the sentence-level human evaluations in Section 8.3.

8.1 Dialogue and human judgment collection protocol

Our main aim is to collect spontaneously produced bilingual dialogues with sentence-level human judgments that will provide us with insight into the quality of the MT systems used to mediate the dialogues. For the resulting corpus to also be of use to the MT research community, we aim to release the collected dialogues as a challenge set. It is important to collect a sufficient number of dialogues to be able to conduct a reliable evaluation and to later constitute a reasonably sized test set. It is also important to vary the dialogue scenarios and to ensure a balance in terms of the number of dialogues mediated by each MT model. Our protocol is designed to make dialoguing easy and enjoyable to encourage a high level of participation. We aim to guide participants by providing role-play scenarios, especially to provide inspiration, without constraining too much the types of utterances produced.

We designed and implemented a dedicated web interface for dialogue collection, through which participants can register, log in and chat to other participants (an example of which is shown in Figure 8.1). Multiple users can be online at any one time, and they choose with whom they wish to dialogue via a central page, displaying available speakers of the opposite language. One of twelve role-play scenarios (cf. Section 8.1.2) is also randomly selected at the beginning of each dialogue.

8.1.1 Participants

Our participants are all unpaid, adult volunteers and are solicited both directly (colleagues, family and friends) and indirectly (via a public announcement on social media). They are encouraged to take part in several dialogues, and the interface is therefore designed to enable them to register an account, which they can use to log in each time they wish to dialogue. When registering, users provide basic personal information that could be useful for future analysis: their age bracket, gender, English and French language speaking ability, other languages spoken and whether they work in research

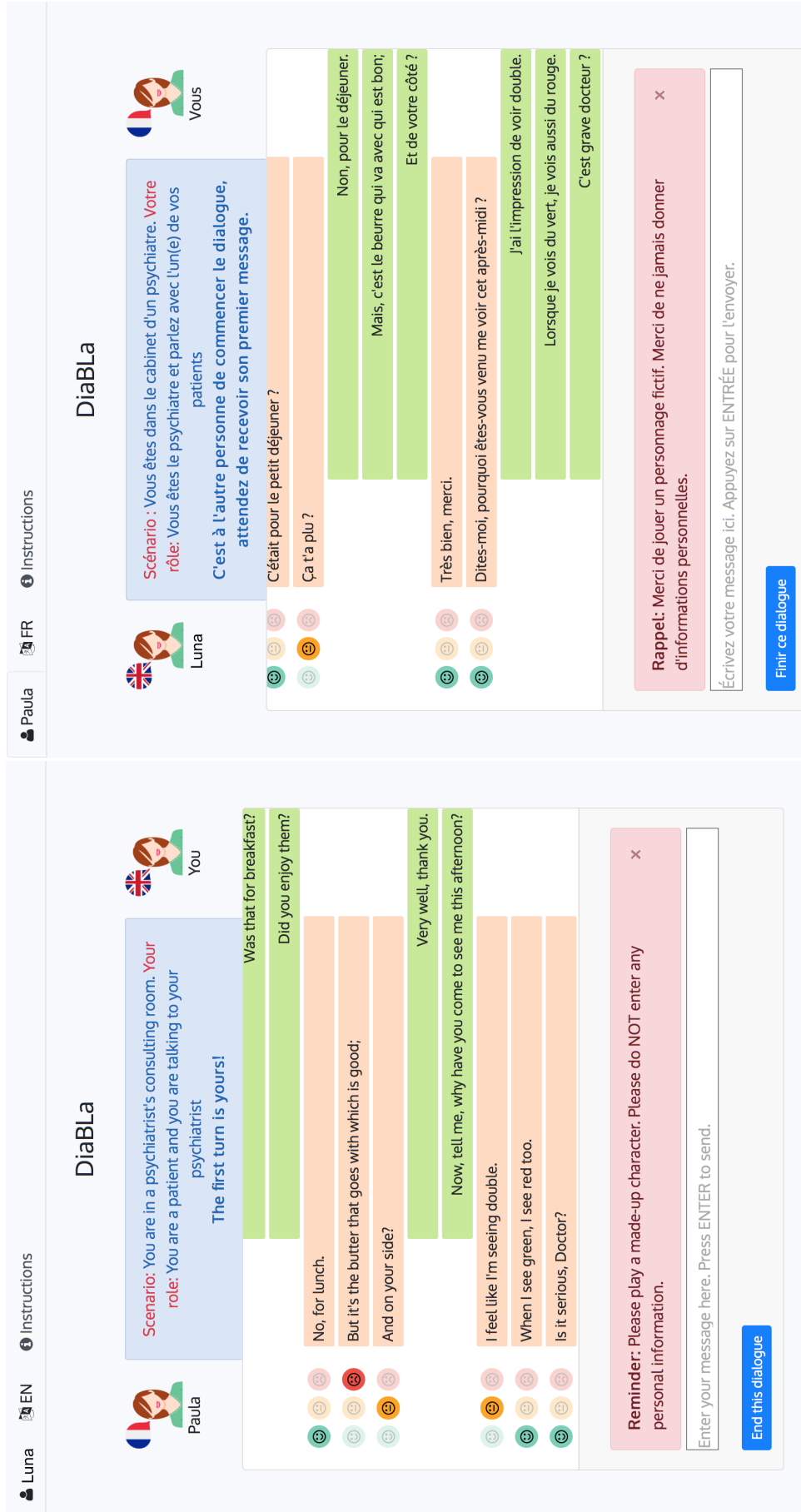


Figure 8.1: An example of a dialogue from the perspective of each of the participants: the English participant, Luna, on the left, and the French participant, Paula on the right. Each participant must evaluate the translation of the other participant's utterances using the three-smiley system. Not shown in this figure is the finer-grained evaluation participants can complete when they select either *medium* or *poor* quality. An example of this is given in Figure 8.3.

or Natural Language Processing. Once registered, the only information that is visible is their username and their gender, in the form of an icon (male or female), which can potentially be used as a source of information during the dialogue (for example for gender agreement in French). Although users log in using a username that will be displayed to others, they are discouraged from using these usernames in the dialogues. To ensure that this is truly the case, we perform manual anonymisation, in which any usernames written in the dialogues are replaced by an alternative first name. These replacement names (which do not appear elsewhere in the corpus) are indicated in a file distributed with the corpus.

8.1.2 Scenarios

A difficult part of collecting real dialogues from human participants is to encourage natural, spontaneous language behaviour. It is particularly important for participants not to be lost for words, be too embarrassed to participate, or, conversely, to offer personal details that cannot be freely distributed. To overcome these potential pitfalls, we choose to impose role-play scenarios in which participants assume fictional roles assigned to them at the beginning of the dialogue. This has the advantage of offering anonymity to participants and providing inspiration, whilst also restricting topics to a controlled list, enabling us to perform a more controlled comparison of dialogues as to the type of language and vocabulary used.

We propose twelve scenarios, chosen to reflect a range of different settings and speaker relationships, but which could also plausibly be reflected by the types of scenarios present in the corpus used to train our models, OPENSUBTITLES2016 (Lison and Tiedemann, 2016). Two scenarios are shown in Figure 8.2 and the full list is given in Appendix B.1. At the beginning of each dialogue, a scenario is randomly selected, and roles are randomly assigned to each speaker.² The first turn is also assigned randomly to one speaker to help to initiate the dialogue.

In the protocol provided to participants, we indicate that the dialogues should read like written drama transcriptions rather than chat messages, and that the scenarios should be treated like improvised drama role-play scenarios. The use of emoticons, SMS speech and usernames is therefore strongly discouraged, and the use of correct spelling, grammar and punctuation encouraged. This also alleviates potential problems linked to the disparities of the ages of participants between French and English speakers (see Section 8.2), since they are less likely to use generation-dependent spellings, abbreviations and SMS speech (e.g. *LOL*, *tfn*, etc.). Although it would also be interesting to study the translation of noisy non-canonical texts such as those found on social networking sites, the aim of our work

²In practice, we varied the strategy of scenario selection towards the end of data collection to ensure that the corpus contained the same number of dialogues per scenario.

<p>You are both stuck in a lift at work. Role 1: You are an employee and you are with your boss. Role 2: You are the boss and are with an employee.</p> <p>You are in a retirement home. Role 1: You are visiting and talking to an old friend. Role 2: You are a resident and you are talking with an old friend who is visiting you.</p>
--

Figure 8.2: Two of the twelve scenarios and the associated roles used to guide the dialogues. The full list is given in Appendix B.1.

is the translation of informal dialogues, particularly in terms of the use of context, and we therefore choose at present not to work on both aspects simultaneously.

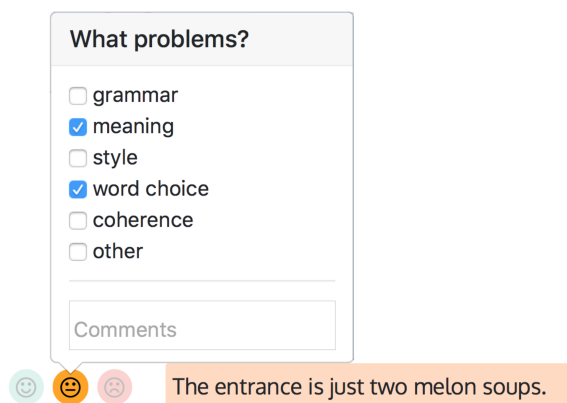
8.1.3 Evaluation

The translation models used to mediate the dialogues are evaluated by the participants themselves. Since each participant only sees the translated version of their partner’s utterances, the evaluation is performed from a monolingual point of view. As mentioned in the discussion on human evaluation techniques in Section 3.3.1, although bilingual evaluators are seen to be the gold standard of MT evaluation, the use of monolingual evaluators is recommended as a way of expanding the pool of people able to perform human evaluation (Dorr et al., 2011) and even of obtaining more consistent evaluations in some cases (Guzmán et al., 2015). However our main reason for using monolingual evaluators is that they enable us to evaluate the MT models based on the perception of the end users of an MT-mediated dialogue system, who do not necessarily understand the language of the other speaker. Collecting the evaluation judgments during the dialogues enables us to have first-hand information about how translation errors affect the participants’ ability to dialogue.³

Human evaluation judgments are collected in two phases: during the dialogues in real time (at the sentence-level) and once the dialogue is finished (an overall quality judgment). Each of these is described in more detail below.

Sentence-level human judgments The first phase is a sentence-level human evaluation of all translated utterances. Participants evaluate in real time and are encouraged to evaluate utterances as they wait for the other person to write a message. This reduces the risk of evaluation disrupting the dialogue flow and also ensures that

³It is possible that some translation mistakes go unnoticed if the translation appears not to contain any mistakes. However, this will generally result in errors in terms of dialogue coherence, which may be picked up by the participants.



What problems?

grammar

meaning

style

word choice

coherence

other

Comments

😊 😐 😞 The entrance is just two melon soups.

Figure 8.3: The sentence-level evaluation form. The original French sentence was *L'entrée c'est juste deux soupes de melon*. “The starter is just two melon soups.”

evaluations are more thorough and relevant to the perception of the dialogue participants. Timestamps of modifications are recorded, allowing us to record any changes in decision that may be made over the course of the dialogue.⁴

Participants indicate sentence translation quality by selecting one of three icons: a green smiley face for a *perfect* translation, an orange neutral face for *medium* quality translation and a red sad face for a *poor* translation. They may optionally indicate in which way the translations were imperfect by selecting one or more error types: *grammar*, *meaning*, *style*, *word choice*, *coherence* and *other*. Finally, they may also write a free comment further detailing errors and explaining their evaluation choice. An example is shown in Figure 8.3, and a screenshot of the participants’ dialogue screens are given in Figure 8.1.

Examples of each of the error types are provided in the protocol that participants are encouraged to read before dialoguing and to which they have access during dialogues. The examples are purposefully not exhaustive so that participants evaluate naturally according to their intuitions. Below is a list of some of the examples given to participants:

- Grammar (the sentence is grammatically incorrect)
 - Wrong number agreement: *The boy **are** there*
 - Missing articles: *I want \emptyset dog*
- Meaning (The sentence (or an element in the sentence) does not make sense)
 - *I was told by my **avocado** that a sentence was likely.*
- Style (the level of formality is inconsistent or language usage is strange)
 - Strange/unnatural utterances
 - Wrong level of formality: “*What’s up*” in a job interview

⁴Information from following utterances may help them better understand the nature of errors and therefore change their previous judgments or add additional comments.

- Word Choice (a poor word choice is made)
 - *I did you a chocolate cake* (instead of *I made you a chocolate cake.*)
 - *He took an attempt.* (instead of *He made an attempt.*)
- Coherence (lack of consistency with previous utterances or the context)
 - Wrong pronoun used that refers to something previously mentioned
 - (For French) Inconsistent use of *tu* and *vous*
 - Word choice is inconsistent with what was previously said (e.g. - *I'm angry!* - *What do you mean by 'upset'?*)

Final, overall evaluation Once the dialogue is finished, each participant is asked to complete the second evaluation phase, a short questionnaire indicating their general perception of the dialogue. They are first asked to communicate any technical problems. If no technical problems arose, they are asked to provide an evaluation of the translation quality according to the same set of criteria as in the online evaluation (*grammar, meaning, style, word choice, coherence* and *other*), each according to a five-point scale (*excellent, good, average, poor* and *very poor*). They also give specific comments on the overall quality of the dialogue, on particular aspects poorly translated and on the interface itself, and finally indicate whether they would use such an interface in the real world. A screenshot of this final evaluation form can be found in Appendix B.2.

8.1.4 MT systems and setup

Training data and pre-processing All models are trained using OPENSUBTITLES2016 (Lison and Tiedemann, 2016). The data is cleaned, tokenised and true-cased using the MOSES toolkit (Koehn et al., 2007) and tokens are split into subword units using BPE (Sennrich et al., 2016d). As in our experiments in Section 7.2, but contrarily to the experiments in Section 7.1, the data is filtered to exclude poorly aligned or truncated sentences⁵, resulting in a training set of 24,140,225 sentences. This proved necessary after initial testing showed that translations were often truncated with respect to the original source sentence, most likely due to the fact that many target subtitles are only partial translations of the original English source sentence. At test time (during dialogues), participants' text is split into sentences,⁶ tokenised, true-cased and split into subword units, before being translated.

⁵We automatically align the subtitles on the word level using FASTALIGN (Dyer et al., 2013), with symmetrised and the *grow-diag-final-and* strategy. We then filter out those sentences for which fewer than 80% of either source or target tokens are not aligned.

⁶We use a rule-based sentence splitter, which is an improved (and not yet published) version of the one included in the SxPipe shallow processing pipeline (Sagot and Boullier, 2008).

MT systems We compare two types of MT model for MT quality (see Section 8.3),⁷ corresponding to the baseline and 2-TO-2 models of Section 7.1, used in our experiments for the integration of linguistic context for discourse translation. We choose the lightweight 2-TO-2 model as opposed to our highest scoring model to ensure that translation times are not too different between each of the two models compared.

The systems are trained in the same way as the baseline and 2-TO-2 systems in Section 7.1, but on the filtered training data to avoid truncation. Both models are neural encoder-decoder models with attention (Bahdanau et al., 2015), implemented using NEMATUS (Sennrich et al., 2017). The first model (BASELINE) is trained to translate sentences in isolation. The second model (2-TO-2), is trained to translate sentences in the context of the previous sentence following (Tiedemann and Scherrer, 2017) and (Bawden et al., 2018b). This is done by simply concatenating each sentence with its previous sentence, separated by a special token, and translating both sentences at once. In a post-processing step, only the current sentence is kept. The concatenation method relies on the previous sentence being in the same language as the current sentence. When the previous sentence is produced by the same speaker, the original version of the previous sentence can be concatenated to the current sentence. However, when the previous sentence is produced by the other speaker, the MT version of the sentence is concatenated to the current sentence, as shown in Table 8.1. Translation itself is performed using MARIAN (Junczys-Dowmunt et al., 2018), designed to be able to translate NEMATUS models, but which decodes much faster than NEMATUS itself. Contrarily to our work in Section 7.2, neither of our models integrates extra-linguistic context, although this would be useful for future work.

Web interface and display The web interface is implemented in HTML using javascript and a python-based server using the FLASK package. We started out from a simple chat interface template available online⁸, and heavily adapted and extended it to our needs (to include account creation, logging in, a welcome room, linking to our MT servers, management of scenarios as well as providing the interfaces needed for evaluation).

Utterances are always displayed sentence by sentence, even if several sentences are sent at once, so that they can be evaluated individually. The participants have access to all past utterances (in their native language) and can write a new message whenever they want (they do not have to strictly alternate with the other person). The interface indicates when the other person is writing and when a message has been sent and is currently being translated. Participants are encouraged to evaluate the MT quality of the messages already received whilst the other person is typing. The delay between a message being

⁷Within a dialogue, the same type of model is always used for both language directions.

⁸<https://github.com/miguelgrinberg/Flask-SocketIO-Chat>

	Original	MT input	Raw MT output	Postprocessed
	...			
EN	I've got some chocolate biscuits left too.	yes please , that would be great . I've got some chocolate biscuits left too.	yes please , that would be great . <CONCAT> j' ai aussi des biscuits au chocolat .	J'ai aussi des biscuits au chocolat.
FR	Super !	j' ai aussi des biscuits au chocolat . <CONCAT> super !	I also have chocolate biscuits . <CONCAT> great!	Great!
FR	Aurais-tu du sucre ?	super ! <CONCAT> aurais -tu du sucre ?	great ! <CONCAT> do you have any sugar ?	Do you have any sugar?

Table 8.1: Illustration of the MT input/output and the final postprocessed translation for the 2-ro-2 model. Note that when the speaker of the previous sentence is not the same as the current speaker (row 2), the postprocessed MT output (in red) is used as the context, which is concatenated to the current sentence. When the speaker of the previous sentence is the same as the current speaker (row 3), the original message (in blue) is concatenated to the current sentence.

sent and it being received is variable (depending on the load of the machine and on the length of the segment to be translated. On average the delay is just under 4.5 seconds, most messages taking either three or four seconds to be sent, pre-processed, translated and post-processed. To reduce the effect of this delay, we only display a user's message once the other participant has also received it. This means that each participant is aware of the delay taken and is less inclined to be frustrated that the other person has not yet replied. We encourage users to type a minimum of fifteen sentences each per dialogue, although participants can choose to end the dialogue at any time, or continue for longer if they want.

8.2 Description of the corpus

We conducted the dialogue experiment for a period of approximately a month between native French- and native English-speaking volunteers, of whom some characteristics are given in Table 8.2. There were 75 different active accounts, equally spread between French and English speakers. French speakers are twice as likely to be involved in NLP or in research than English speakers, and there is also a difference in terms of the average ages of the speaker per language (the French speakers are on average younger than the English ones). However, we have alleviated some of the generation-dependent traits of writing by discouraging the use of SMS-speech, smileys and abbreviations. We have yet to see if these differences in speaker characteristics have an impact on the language used or

their way of evaluating.⁹ The information about participants is provided with the corpus, and so further investigation is possible. We consider that it should not influence our evaluation too much inasmuch as we compare the two different types of model (baseline vs. 2-TO-2 for each language direction separately).

	EN	FR	All
Total number	37	38	75
#Researchers	7	17	24
#Experience in NLP	6	14	20
Min. age bracket	18-24	18-24	18-24
Max. age bracket	65-74	65-74	65-74
Median age bracket	55-64	25-34	35-44
Modal age bracket	55-64	25-34	25-34

Table 8.2: Basic characteristics concerning the participants.

The corpus contains a total of 144 dialogues, with a minimum number of 10 utterances each. In reality, most dialogues contain more sentences than this, with 75.7% of dialogues containing 35 sentences or more. Each dialogue scenario is presented an equal number of times for each MT model type, amounting to twelve dialogues per scenario (six for each model type). Some basic corpus characteristics are given in Table 8.3. The corpus contains a total of 56,622 tokens (27,727 English tokens and 28,895 French tokens) over a total of 5,748 sentences.

8.2.1 Overview of translation successes and failures

An extract of dialogue, representing a fictional argument, is given in Figure 8.4, providing an example of the type of language used by participants. The language used is colloquial and contains a number of fixed expressions (e.g. *get off your intellectual high-horse...*, *...Mr Fancy pants*), which can prove difficult for MT, as is the case in this example.

The MT models are sometimes sufficiently robust to handle certain spelling and grammatical errors, although they are not initially designed to do so. In the extract in Figure 8.4, there are two spelling/grammatical errors: *qui ne penses* ‘who think_{2SG}’ is written instead of *qui ne pense* ‘who thinks_{3SG}’ and *rality* instead of *reality*. Both are translated adequately, the first translated into English using the correct word *thinks* and the second being translated into French as *ralité* instead of *réalité*, thus conserving the spelling error in translation. This is possible thanks to subword segmentation, which enables the suffix of the word to be correctly translated into its French variant, whilst the unseen first part of the word is conserved.

⁹As users are assigned roles in the scenario, their age can be somewhat compensated by their role’s stereotypical age (e.g. teacher vs. student scenario).

EN: Why do get the impression this is the first time you've mentioned anything going wrong between us? <i>MT: Pourquoi pensez-vous que c'est la première fois que vous dites que quelque chose va mal entre nous ?</i> <i>Eval:</i> 😊
FR: Je suis comme Louis Jovet dans les enfants du Paradis, quand il dit à Arletty qu'il a besoin de changer d'air. <i>MT: I'm like Louis Jovet in the children of Heaven, when he tells Arletty he needs a change of air.</i> <i>Eval:</i> 😊
FR: Tu te rappelles ce film magnifique de Marcel Carné , nous l'avons vu en français <i>MT: You remember that beautiful Marcel Carborn, we saw it in French.</i> <i>Eval:</i> 😊 (meaning) 'that'... could probably do without it'
EN: i have no idea what you're talking about. <i>MT: Je ne sais pas de quoi tu parles.</i> <i>Eval:</i> 😊
EN: But, i'm going to go with you on this one and assume that it has just slipped my mind. <i>MT: Mais je vais y aller avec toi sur ce coup-là et assumer que ça m'a échappé.</i> <i>Eval:</i> 😊
EN: so, yeah, sure, i remember your stupid film. <i>MT: Donc, oui, bien sûr, je me souviens de ton stupide film.</i> <i>Eval:</i> 😊
EN: it probably sucked. <i>MT: C'était probablement naze.</i> <i>Eval:</i> 😊
...
FR: J'ai besoin de stimulation intellectuelle, émotionnelle, pas de vivre à côté d'un geek attardé et boutonneux qui ne penses que à ses jeux vidéos et à ses mangas. <i>MT: I need intellectual stimulation, emotional stimulation, not to live next to a retarded and pimply-faced geek who thinks only of his video games and its Manga.</i> <i>Eval:</i> 😊
EN: well look at you, Mr Fancy pants! <i>MT: Eh bien, regardez-moi ça, Mr.</i> <i>Eval:</i> 😊 'the Mr isn't useful'
...
EN: how about you get off your intellectual high-horse and get back to reality for a bit <i>MT: Que diriez-vous de sortir de votre cheval intellectuel et de retourner à la ralité un moment ?</i> <i>Eval:</i> 😊 (meaning) 'cheval intellectuel doesn't exists'

Figure 8.4: A dialogue extract with baseline MT output (grey and prefixed with “*MT:*”) and human evaluation (prefixed with “*Eval:*”). The participants never see the original text written by the other participant. Text in bold is discussed in the text. Typographical errors and omitted words were present in the original dialogue.

8.2. Description of the corpus

	EN	FR	All
#Turns			
Total	1,067	1,089	2,156
Mean per dialogue	7.4	7.6	15.0
#Sentences			
Total	2,865	2,883	5,748
Mean per dialogue	19.9	20.0	39.9
Min. and max. per dialogue	5-42	5-60	10-102
Mean per turn	2.7	2.6	2.7
Min. and max. per turn	1-9	1-10	1-10
#Tokens (original messages)			
Total	28,137	29,275	57,412
Total unique	3,471	4,058	-
Mean per dialogue	195.4	203.3	398.7
Min. and max. per dialogue	67-612	54-664	121-1,276
Mean /sentence	9.8	10.2	10.0
Min. and max. per sentence	1-46	1-52	1-52
#Tokens (translated messages)			
Total	27,727	28,895	56,622
Total unique	3,116	3,728	-
Mean per dialogue	192.5	200.7	393.2
Min. and max. per dialogue	54-634	69-633	123-1,267
Mean /sentence	9.6	10.1	9.9
Min. and max per sentence	1-47	1-74	1-74

Table 8.3: DIABLA corpus characteristics

The dialogues also contain cultural references, such as references to films and actors. In many cases named entities are well conserved, although sometimes cause problems. For example, *Marcel Carné* is translated as *Marcel Carborn*. This is also due to subword segmentation, which segments *Carné* into two subwords, *Car* and *né* ‘born’, of which the second element is translated into English.

The MT-mediated nature of our corpus also gives rise to certain formulations that are specific to the type of interaction found in cross-lingual dialogues and are therefore less likely to be found in existing corpora. An example of this are meta-discussions about the previously translated utterance when MT errors lead to a misunderstanding, which appear several times in the corpus. For example, in Figure 8.5, the second French utterance is incorrectly translated into English as *Or the Thai limit?*, which leads to a slight breakdown in understanding. The English speaker asks what the French speaker means by *the Thai limit* to help clarify the situation. For this utterance to be translated correctly, the quoted segment of text should be translated as it was in the original utterance (i.e. as *la limite thai*). It is instead translated as *la limite thailandaise*, which

is probably understandable given the context, but does not respect the cohesive nature of the original text. Taking into account such formulations would require a special case, as it is unlikely that such cases arise in corpora that are not inherently bilingual and are not mediated by MT.

FR: Tu connais un restau indonésien ?
<i>MT: You know an Indonesian restaurant?</i>
FR: Ou à la limite thaï ?
(gloss) ‘Or at a push, Thai’
<i>MT: Or the Thai limit?</i>
EN: What do you mean by the Thai limit ?
<i>MT: Qu’est-ce que tu veux dire par la limite thaïlandaise ?</i>

Figure 8.5: An example of a poorly translated meta-exchange resulting from a translation error. The translation of the English speaker’s sentence, which quotes the previous translation, formulates the quoted sequence differently from the original sentence: *la limite thaïlandaise* instead of *la limite thaï*.

8.2.2 Comparison with existing corpora

Aside its utility for the evaluation of our contextual MT models, the DIABLA corpus is a useful resource in itself for future research into MT-mediated dialogue. To our knowledge, DIABLA is the first corpus of its kind. A number of corpora are similar in nature, in terms of their setup or the type of language used. However they either cover different domains or are not designed with the same aim in mind.

Parallel corpora of informal text OPENSUBTITLES2016 (Lison and Tiedemann, 2016), used throughout this thesis, is similar to our corpus in terms of the style of language used and the heterogeneity of conversation topics. The corpus is on a much larger scale than our own, and the film domains are more diverse. However the conversations remain scripted rather than being spontaneous exchanges, and are translations of monolingual texts, rather than being bilingual conversations.

Speech corpora are an alternative resource that can be adapted for translation purposes. They provide conversational data, which, if transcribed and then translated, provide written data for training MT models.¹⁰ Post et al.’s (2013) Fisher and Callhome Spanish-English Speech Translation Corpus is an extension of the Fisher (Graff et al., 2010)

¹⁰The corpora can be used for speech translation, whereby speech recognition systems, or at least spectrogram feature extraction must be first used prior to translation. This step is often noisy and can result in further MT errors downstream.

and Callhome (Canavan and Zipperlen, 1996) speech corpora, with English translations provided through crowdsourcing. The resource provides 171,254 parallel sentences of conversational data, and the presence of transcriptions means that the corpus can also be used for informal text translation. The corpus contains real spontaneous conversations.

However, both OPENSUBTITLES2016 and the Fisher and Callhome Speech Translation Corpus are translations of monolingual texts, rather than bilingual conversations, and do not therefore include the interaction MT-mediated setup that is central to our corpus.

Translation-mediated bilingual corpora Several specifically bilingual corpora do exist, although their setup and domains differ from ours. The MSLT corpus (Federmann and Lewis, 2016) is a bilingual speech corpus of conversation held over Skype, which are transcribed and developed as a test set for MT. The participants are all bilingual speakers, who therefore understand all of their partner’s utterances without the need for mediation. They discuss as a potential option the mediation of dialogues between speakers of different languages, but decide not to pursue the option as speech recognition errors had a negative impact on MT and mutual understanding was impaired. In our setting, we do not encounter such problems, as our participants type their exchanges, and we therefore have no reliance on speech recognition software. They also mention that clarification questions during the dialogue impacts dialogue flow and slows down the conversation. As our exchanges are written, the speed of conversation is slower than a spoken conversation. We also purposefully choose to collect data from a real setting in order to analyse how well the dialogue can flow when mediated with such MT systems, an aspect that the MSLT corpus does not aim to illustrate.

A number of other bilingual corpora have been developed over the year in particular for the travel and hospitality domains, and are based on oral conversations with transcriptions. They are collected in a variety of scenarios, from mediation by human interpreters to the use of MT systems. The largest of these corpora is the SLDB (Spoken Language DataBase) (Morimoto et al., 1994), a collection of transcripts of spoken English-Japanese dialogues in the hotellery domain, containing 618 dialogues and approximately 22,000 sentences (21,769 Japanese sentences and 22,928 English sentences). To avoid speech recognition and MT errors, they choose to adopt a different strategy to the creators of the MSLT corpus; the dialogues are instead mediated by professional (human) interpreters. The MAD (Machine-aided Dialogs) corpus (Takezawa and Kikui, 2003) is constructed in similar way but dialogues are mediated by machines (human typists for the transcript and then MT systems). Also designed for the travel domain, the corpus contains five role-play scenarios of varying complexity, containing between 1,437 and 3,022 utterances each. Finally, the Field Experiment Data of Takezawa et al. (2007) also consists of role-play scenarios but designed to be more natural than the previous MAD corpus. The resulting corpus is relatively small (608 utterances from Japanese to English,

660 utterances from English to Japanese, 344 utterances from Japanese to Chinese and 484 utterances from Chinese to Japanese). The corpus is the most similar to our own, because of the use of MT to mediate the dialogues, and the participants are also asked to evaluate the overall quality of the system once the dialogue is finished (by assigning one of four subjective scores: perfect, good, fair and nonsense/none).

The majority of the bilingual corpora available are not mediated by MT systems (the case of the MSLT corpus and SLDB corpus). The Field Experiment Data is most similar in terms of its setting, but is not accompanied by sentence-level human judgements about the quality of the translation systems. Such annotations are an important advantage of our corpus, as they enable us to analyse what types of errors the MT system commits and how this could hinder communication in a way that overall translation scores (a single judgment at the end of a dialogue) cannot. The domain of the corpus is also very different to ours: the restriction of their system to the travel domain makes the scenarios more formalised and predictable, whereas our corpus contains a range of different everyday scenarios, with unpredictable language structures and varied vocabulary.

8.3 Evaluating contextual MT with the DIABLA corpus

Although the corpus is useful as a resource in itself, the primary objective of our corpus is to be able to evaluate MT systems in real dialogue contexts. We therefore take a look here at the human judgments provided for each of the MT models tested: the non-contextual baseline model and the lightly contextual 2-TO-2 from Section 7.1. We look here at global trends in evaluation. Given that the evaluation is performed by different participants on different subsets of sentences, comparisons should only be made when strong tendencies occur. We aim to test whether such an evaluation can give us reliable and consistent information about the quality of the translations produced by the two different types of system. The evaluations nevertheless provide a rich source of information about how MT mistakes are perceived by different native speakers.

8.3.1 Overall MT quality

We begin by reviewing the judgments of the overall MT quality, as calculated on the sentence-level human evaluations. Differences between models are shown in Figure 8.6. They unsurprisingly show that MT quality is dependent on the language direction; translation into English is perceived as better than into French, approximately half of all EN→FR sentences being annotated as *medium* or *poor*.¹¹ There is little difference

¹¹It should be noted that these categories are used as soon as there is any kind of problem noticed in the sentence, however minor the problem may be or whatever the nature of the error.

in perceived quality between the `BASELINE` and `2-TO-2` for `FR→EN`. This contrasts with `EN→FR`, for which the number of sentences marked as *perfect* is higher by +4% for `2-TO-2` than for `BASELINE`.

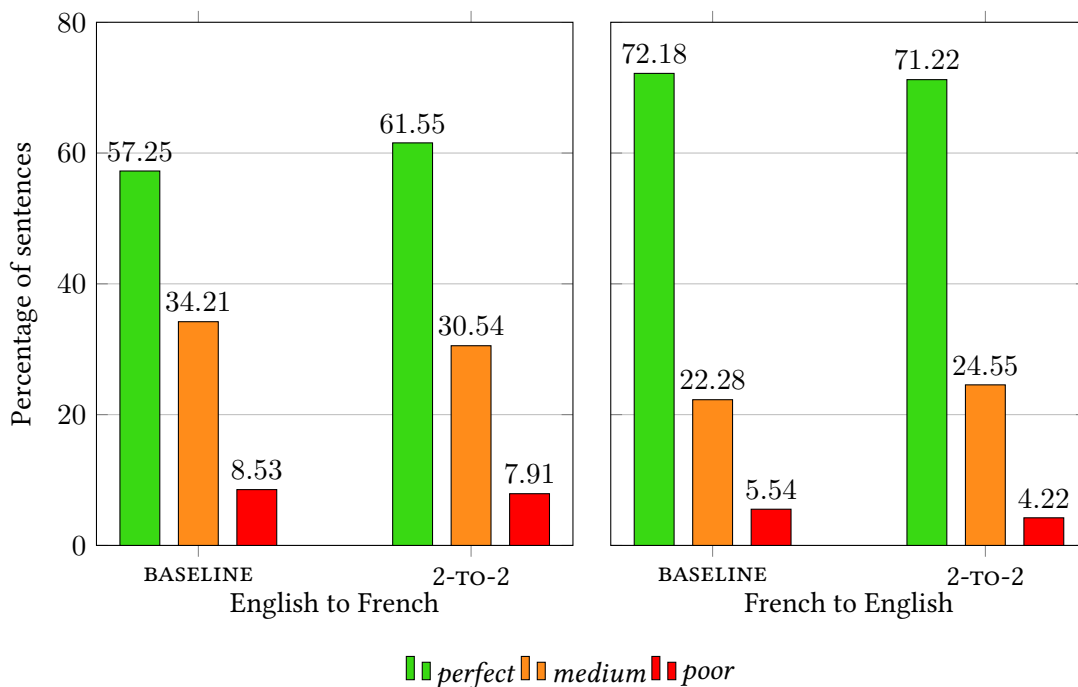


Figure 8.6: Percentage of sentences for each language direction and model marked as *perfect*, *medium* and *poor*.

The detailed evaluation results shown in Figure 8.7 give evaluation scores for each problem type. The few number of problems classed as *other* indicates that our categorisation of MT errors was sufficiently well chosen. The most salient errors for all language directions and models are in *word choice*, especially when translating into French, with approximately 16% of sentences deemed to contain a word choice error. As with the overall evaluations, there are few differences between `BASELINE` and `2-TO-2` for `FR→EN`, but significant differences are seen for `EN→FR`: `2-TO-2` models perform better, with fewer errors in most problem types, except *word choice*. A notable difference is the lower frequency of *coherence*-related errors for `2-TO-2`. The types of coherence errors also appear to be less serious, as there is a lower percentage of these translations which are labelled as *poor* (as opposed to *medium*). These results are encouraging, as they show that our data collection method is a viable way to collect human judgments, and that such judgments can reveal fine-grained differences in MT systems, even when evaluations are performed on different sentence sets. They also appear to suggest that our lightly contextual model is better at ensuring the coherence of translations.

In spite of the errors, the translation quality is in general good, especially into English, and participant feedback is excellent concerning intelligibility and dialogue flow. As well as the sentence-level judgments, participants indicated the overall quality of the

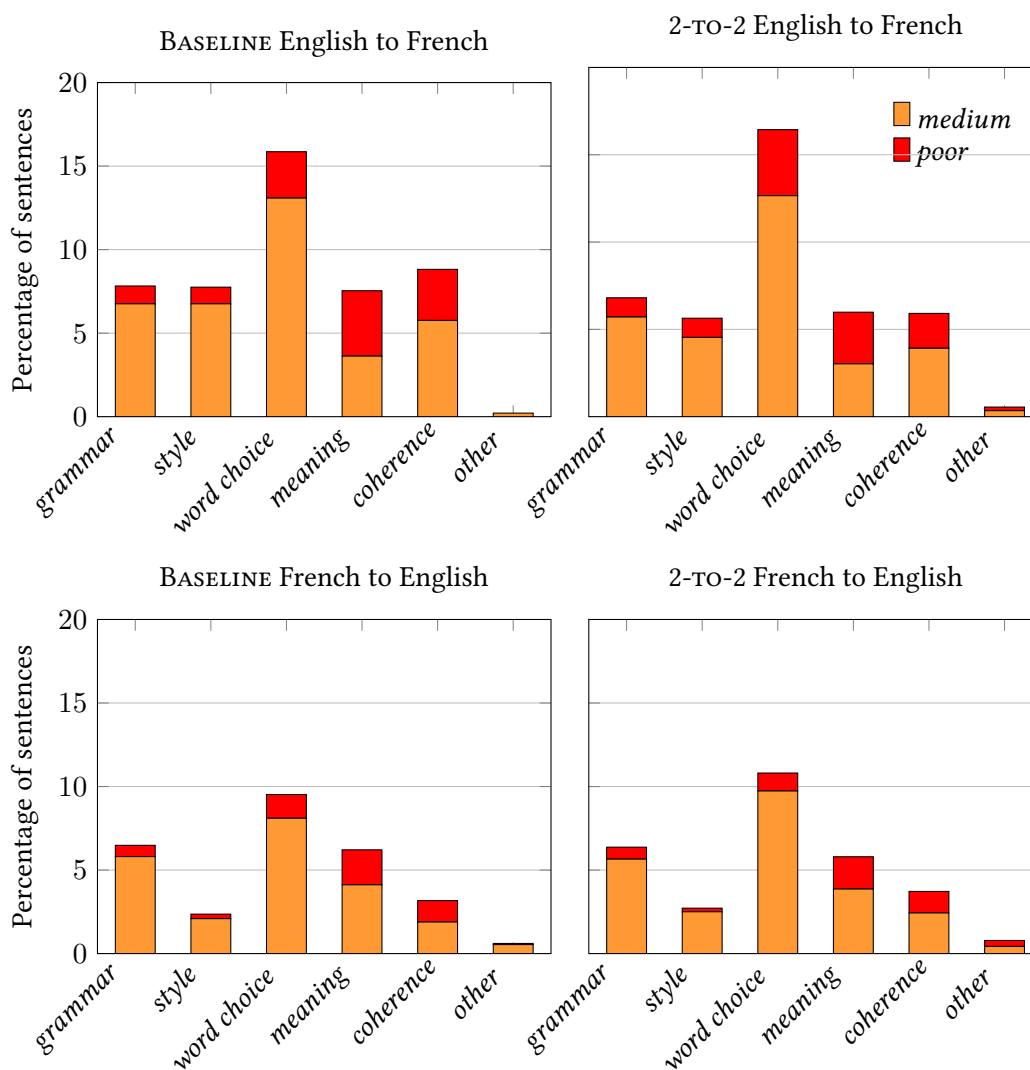


Figure 8.7: Percentage of all sentences for each model type and language direction marked as containing each problem. A sentence can contain several problem types. Bars are cumulative and indicate the percentage for sentences marked as *medium* (orange) and *poor* (red).

dialogue once the dialogue was complete. Participants indicated that they would use such a system to communicate with a speaker of another language 89% of the time. In 81% of dialogues, grammaticality was indicated as either *good* or *excellent*. Coherence, style and meaning were all indicated as being *good* or *excellent* between 76% and 79% of the time. As a confirmation of the sentence-level evaluations, word choice was seen as the most problematic error type, indicated in only 56% of dialogues as being *good* or *excellent* (40% of dialogues had *average* word choice, leaving a very small percentage in which it was perceived as *poor*).

8.3.2 Focus on a discourse-level phenomenon

While we leave a thorough analysis of the corpus to future work, we provide a preliminary analysis of the effect of adding a previous sentence, by focusing on a particular discourse-level phenomenon: the translation of English *you* into French.

When translating from English into French, the English second person pronoun *you* is ambiguous between the forms *tu* ‘you_{sg} (informal)’ and *vous* ‘you_{sg} (formal) or you_{pl}’. In our dialogues, the form *you* almost invariably applies to the other speaker, and therefore to a singular form of the pronoun, rather than a plural one, resulting in ambiguity between *tu* (singular informal) and *vous* (singular formal). Most of the scenarios were such that we would expect the same pronoun to be used by each speaker, and a consistent use of the pronouns throughout the dialogue. The inconsistent use of these two forms was one of the most commented problems by French speakers. The importance of using honorific pronouns consistently is the reason behind the strategy proposed by Sennrich et al. (2016a) using side-constraints.

Neither of our models explicitly handles the choice between *tu* and *vous*. In a real-life translation scenario, such a choice could be set at the beginning of the dialogue based on the scenario type as determined by the most probable pronouns for the roles in the scenario. For example, the appropriate pronouns could be *vous* between a teacher and a student and *tu* between spouses.

However, our lightly contextual 2-TO-2 does take into account pairs of consecutive sentences, and therefore could be expected to have more consistent use of the pronouns across neighbouring sentences, especially in symmetrical scenarios in which participants use the same pronoun. As a proxy for their ability to account for lexical cohesion, we look at the two models’ ability to ensure consistent translation of the pronouns across consecutive sentences. For each model, we take translated sentences in which *tu* or *vous* appear, and for which the previous sentence also contains either *tu* or *vous*. By comparing the number of times the current sentence contains the same pronoun as the previous sentence (see Table 8.4), we can estimate the degree of translation consistency for this particular aspect. Although the absolute figures are too low to provide statistical

significance, we see a general trend that the 2-TO-2 model shows greater consistency in the use of the pronouns over the baseline model, with +9% in the consistency use of *tu* and +6% in the consistent use of *vous*.

Prev. \ Curr.	BASELINE		2TO2	
	<i>tu</i>	<i>vous</i>	<i>tu</i>	<i>vous</i>
<i>tu</i>	49	31	55	23
<i>vous</i>	27	27	22	28

Table 8.4: For each model, the number of times each model translates using *tu* and *vous* and either of the forms *tu* and *vous* also appears in the previous sentence.

8.4 Perspectives

We have shown through a preliminary analysis that our collected human judgments provide a viable form of MT evaluation and can be further analysed to give us more insight into MT of dialogue. In the future, we aim to analyse the corpus in greater depth, particularly with the help of additional annotations. The corpus, which will be made freely available, offers many opportunities for future work. Here, we choose to focus on two possible future perspectives: the analysis of the type of language used in MT-mediated dialogues (Section 8.4.1) and the evaluation of future MT models (Section 8.4.2).

8.4.1 Language analysis of MT-assisted interaction

As MT systems are becoming more common online and in social media, it is important for MT systems to take into account the type of language that may be used and the way in which user behaviour may affect MT performance. Non-canonical syntactic structures, spelling and typing errors, text mimicking speech, including pauses and reformulations, must also be taken into account if MT systems are to be used for successful communication in more informal environments. The language used in our corpus is relatively clean in terms of spelling. However spelling errors and grammatical errors do exist in the corpus, and do have an impact on the quality of the translation, although this has yet to be analysed in detail. It would certainly be useful to study how spelling errors as well as informal reformulations impact translation understanding.

We intend to further extend the English-French corpus in future work and annotate it with discourse-level information, which will pave the way for future phenomenon-specific evaluation: how they are handled by the different MT systems and evaluated by the participants. In this direction, we have manually annotated anaphoric phenomena in 27 dialogues (anaphoric pronouns, event coreference, possessives...). Despite the small

size of this sample, it already displays interesting characteristics, which could provide a strong basis for future work. Anaphoric references are common in the sample annotated: 250 anaphoric pronouns, 34 possessive pronouns, and 117 instances of event coreference. Their incorrect translation was often a cause of communication problems (see Figure 8.8). The impact of these problems will be investigated further.

EN _{ORIG.} : Can I lie on this <u>couch</u> ?
...
FR _{TRANS.} : I don't want to bother, he looks clean and new

Figure 8.8: An example of mistranslated coreference with the incorrect translation of French *il* as *he*, referring to *canapé* ‘couch’.

Another interesting aspect of human-MT interaction would be to study how users themselves adapt to using such a tool during the dialogues. How do they deal with translation errors, particularly those that make the dialogue incoherent? Do they adjust their language over time, and how do they indicate when they have not understood correctly? An interesting line of research would be to use the corpus to study users’ communication strategies, for example by studying breakdowns in communication as in (Higashinaka et al., 2016).

8.4.2 MT evaluation

Human evaluation remains the most accurate form of MT evaluation, especially for understanding which aspects of language pose difficulties for translation. While hand-crafted examples and challenge sets provide the means to test particular phenomena (King and Falkedal, 1990; Isabelle et al., 2017; Bawden et al., 2018b), it is also important to observe and evaluate the quality of translation on real, spontaneously produced texts. Our corpus provides this opportunity, as it contains spontaneous productions by human participants and is richly annotated for MT quality by its end users (see Section 8.3).

The corpus has enabled us to evaluate our contextual MT model with respect to a baseline model, and the evaluation has proved viable for showing global trends in MT quality. We will further extend this analysis in the near future. As a resource for evaluating future models, the corpus can be used in its current state as a challenge set for new MT models. It provides an example of spontaneously produced productions in a real, unscripted setting. The sentence-level human judgments provided with the corpus can be used as an indicator as to which sentences were the most challenging for MT, and used to target these sentences when performing manual evaluation of new translations of the same sentences. For example, manual evaluation of new translations of our test set can be guided towards those sentences whose translations were marked as *poor*, to provide

an informed idea of the quality of the new models on these difficult examples, and to encourage development for particularly challenging phenomena.

Finally, we intend to provide post-editions and/or reference translations of the MT outputs in the near future, so that the corpus can also be used as a test set for the automatic evaluation of new MT models. This will also serve as an extra annotation that could be valuable for evaluating and analysing the monolingual human judgments that were produced during the dialogues.

Conclusion and Perspectives

Conclusion and Perspectives

9.1 Conclusion

Our goal of improving the use of extra-sentential context in MT has led us to review and implement a wide range of contextual MT strategies and to investigate a variety of evaluation methods to target particular aspects of contextual translation. Over the three years spent studying the topic, we have been lucky enough to benefit from the high level of interest that the research community has invested in contextual MT. Along with advances in standard MT architectures, in particular the emergence of a new state of the art with the development of neural approaches to MT, this interest in contextual MT has offered us many opportunities to discover new methods of handling context and to see tangible progress in the degree to which it can be exploited by MT systems.

9.1.1 Trends in contextual MT and the impact on our work

In Chapter 4 of this thesis, we laid down a number of distinctions that were to structure the way in which we viewed different contextual strategies. The first one concerned whether context is structured or left unstructured prior to being integrated into translation, and the second concerned the point in the translation process at which the context is applied (during pre-processing, post-processing or during the training of the MT model). In the different strategies we presented in Part II, the two aspects have been highly linked. Although this did not necessarily have to be the case, it does show a

general trend in the way methods have been evolving over time. The experiments we have presented incidentally follow a roughly chronological order, with our initial experiments involving pre-processing for gender adaptation (Chapter 5) applied to SMT models, and our final experiments, which involve modifying the MT architecture applied to NMT (Chapter 7). Between these two chapters, our experiments dedicated to post-processing methods (Chapter 6), which are architecture-agnostic, are applied to the outputs of both SMT and NMT or to noisy reference translations (as proxies for MT outputs).

While many of the early strategies to integrate context into MT concentrated on targeting individual phenomena using highly structured linguistic features (cf. Sections 6.1 and 6.2), more recently the tendency has been to delegate the task of exploiting context to the translation model itself, to be jointly learnt during the translation process. As we saw in Chapter 7, this means that context can be exploited without the need for explicit modelling. For linguistic context (limited to the previous sentence), no explicit modelling was performed and the context was input into the translation model as a prefix to the source sentence or as a separate input (Section 7.1). For extra-linguistic context, contextual information was treated in much the same way as linguistic context, encoded as tokens, and the same strategies proved effective despite extra-linguistic context not displaying the same sequential properties as text (Section 7.2). The linguist could be taken aback by the method's simplicity and the lack of linguistic modelling — at least on behalf of the researcher. This simplification in the way context is modelled *prior to its integration* does not necessarily represent a step back. The complexity is transferred instead to the MT model itself, made possible by the expressivity of NMT architectures. This does not mean that, in the future, additional structured information cannot also be supplied as context.

This change in methodology is a consequence of the shift from SMT to NMT. In phrase-based SMT, the elementary units are phrases and in NMT they are words. However, the two architectures differ in the way these units are represented. In phrase-based SMT, the units are categorical, atomic units, whereas in NMT they are represented in a continuous space. As a result, whereas memory in SMT is limited (approximated on a very local level by n -gram language models), it is a fundamental part of how NMT models function, with the use of mathematical functions to alter intermediate continuous representations with extra context. Integrating context into SMT is more unwieldy, because it does not inherently have such a memory function. Integrating context into the functions of the MT model requires altering the elementary units by replacing them with another discrete unit, or modifying which units are selected by changing their probabilities and/or the function used to score hypotheses. Providing an effective new scoring model requires a good modelling of the hypothesis and the context. Some of the more effective scoring components in SMT have been neural language models, thanks to their ability to take into account long-distance dependencies within the sentence (Bengio et al., 2003), through their ability to memorise information between timesteps.

9.1.2 Review of our aims and contributions

The aims set out at the beginning of this thesis were to review a panorama of different techniques for the inclusion of both linguistic and extra-linguistic context and to propose new ways of evaluating how well the context is taken into account.

We began our work in Part I with an explanation of why context is so important for MT (Section 2), an overview of the current state of the art in MT (Section 3), and a review of the strategies previously used for contextual MT (Section 4). We concentrated on what it means to integrate context from a theoretical point of view, basing our discussion on the distinction between three points in the translation process at which ambiguity can arise: (i) source language ambiguity, (ii) cross-lingual meaning transfer and (iii) target language ambiguity. The point at which an ambiguity takes place can have an impact on which type of context can be used to resolve it and the point in the translation process at which the context can be integrated. In the context of translation, this led us to define two classes of context-dependent phenomena, those related to coherence (concerning the transfer of the intended meaning of the original segment) and those related to cohesion (concerning the formal encoding of the target sentence). We later used these distinctions in both our description of previously used strategies for contextual MT and to structure our own contributions in Part II.

The previous work on contextual MT has tackled a wide range of contextual phenomena, but with a particular focus on anaphoric pronoun prediction. Despite the interest in integrating context, these previous methods have often involved a lot of effort in terms of pre-processing of data in order to provide structured context, and they result in only small gains in translation quality. These approaches have targeted particular phenomena rather than several at the same time, meaning that for all such phenomena to be handled in translation, a cascade of these methods would have to be used, multiplying the time and effort necessary for translation. Over time, the types of techniques used have been heavily influenced by the “neural revolution” that has swept over the domain of NLP, which has led to the change in strategy described above, from structured to unstructured context.

In Part II we proposed our own methods of integrating context accompanied by either adapted evaluation strategies or discussions concerning the problems associated with evaluation. Our experiments represented each of the three strategies presented in Part I:

- Chapter 5 was dedicated to a method of adapting translation to speaker gender through *pre-processing*.
- In Chapter 6 we presented two experiments relying on *post-processing* strategies to integrate context relevant to two different phenomena, tag questions (TQs) (Section 6.1) and anaphoric pronouns (Section 6.2).

- In Chapter 7, we adapted neural MT architectures to integrate context during the training of MT models, and tested the strategies for the integration of both linguistic and extra-linguistic context.
- Finally, in Chapter 8, we presented a contribution of a different nature: a corpus of bilingual dialogues using two of the models presented in Section 7.1, designed for future analysis and evaluation.

To illustrate the progress made in contextual MT and its evaluation, let us consider our contributions from the point of view of three aspects, which have been recurring features of our work: (i) integration of extra-linguistic context (applied to speaker gender), (ii) integration of linguistic context (applied to pronominal anaphora), and (iii) evaluation of contextual phenomena, studied in all contributions.

Extra-linguistic context: speaker gender Adaptation to speaker gender is handled twice, once in Section 5.1 using pre-processing of the training data for a phrase-based SMT model, and again in Section 7.2 using adaptations of a neural architecture. The data sparsity issues encountered with the first method are overcome in the second experiment thanks to two advances: the availability of more resources, and the use of a method that does not exacerbate data sparsity. Although the experiments are not strictly comparable due to a difference in the quantity of data used, the design of the approaches is fundamentally different: whereas the use of data partitioning in Section 5.1 reduces the quantity of data available for training, our approach in Section 7.2 avoids this issue entirely. Our results are a reflection of the progress made in our capacity to use the context provided. Whereas in Section 5.1, we saw no improvement in the use of gender, in Section 7.2, our methods proved highly successful, as shown through a targeted evaluation of gender-marked French translations.

Linguistic context: anaphoric pronouns As one of the most widely studied topics in contextual MT, it is unsurprising that we too focus on the translation of anaphoric pronouns. Again, we present experiments on each side of the architectural shift: our initial experiments in Section 6.2 using a statistical classifier to make use of highly structured linguistic features, motivated by linguistic intuitions, and our experiments in Section 7.1 relying on a very different approach using an NMT architecture, by which no prior structuring of linguistic context is done before it is used in translation. It had previously been shown by Jean et al. (2017b) that training an NMT to use context, without specifically targeting pronoun translation, could result in comparable results to methods specifically designed for that purpose. Our approach continues this line of reasoning, and we show in Section 7.1 that this way of integrating in context is effective for a range of contextual phenomena, particularly as no specific structuration of the context was

performed prior to its integration. In addition to testing previously proposed contextual strategies, we also present a novel contextual strategy, which outperforms the previous approaches in terms of overall translation quality, pronoun translation and lexical choice.

Evaluation of contextual MT Evaluation has been a major concern of this thesis. Our criticism of the automatic evaluation metrics that are frequently used as the only method to show performance gains has been present from the beginning of the thesis (Section 3.3). The main problem with the use of such metrics for the evaluation of contextual MT is that the better use of extra-sentential context is not always reflected by the scores, and gains seen can also be a simple side-effect of the approach used, rather than a consequence of a better use of context. Nothing shows this better than our observation in Section 7.2 that simply adding an additional token to each source and target sentence gives systematic improvements in BLEU score, even when the token is semantically void. Throughout the thesis, we have continually aimed to show that BLEU scores are insufficient for proving that context is being effectively exploited. We nevertheless use the metric to show that our models do not degrade overall translation quality.

In the first two chapters of Part II, we concentrate more on providing a critique of existing metrics than of proposing metrics that we find sufficient for showing contextual gains. In our experiments on gender adaptation in Section 7.2, we showed that BLEU scores are unreliable, but we are not yet capable of providing an effective alternative solution. In Section 6.1, we discuss how we might evaluate the generation of TQs when translating into English. Given that this is a new topic and no previous efforts have been made to analyse such a stylistic phenomenon, we experiment with several types of evaluation and compare the merits of each of them. Our general conclusion is that a range of different criteria would have to be taken into account to be able to even envisage evaluating such a phenomenon.

In Chapters 7 and 8, based on our past experiences for evaluation, we look at three different evaluation strategies that each prove effective in demonstrating the different models' use of context. The first evaluation method uses contrastive test sets for the evaluation of anaphoric phenomena and lexical choice (Section 7.1). Inspired by similar evaluation methods (Sennrich, 2017; Rios Gonzales et al., 2017), the test sets contain pairs of incorrect and correct translations of a source sentence, which are to be scored by the different MT models to be evaluated. A model that is able to provide a higher score to the most correct translations (as opposed to the incorrect ones), is deemed to be better than the other models. Our test sets are designed to specifically test the models' capacity to use extra-sentential context, which is the only factor differentiating the correct and incorrect sentences. The second evaluation method we use in Section 7.2 is a targeted evaluation of speaker gender. We rely on the availability of a large test set among which we automatically identify translations that adhere to certain criteria: in our case the

presence of a construction in which words are potentially marked for the gender of the speaker. By comparing the percentage of sentences for which we are able to bias the gender marking either towards masculine or feminine, we show that the approaches tested are very effective in adapting to speaker gender. Finally, in Chapter 8, we present DiaBLa, a bilingual MT evaluation corpus, containing sentence-level human evaluations. We design a protocol for collecting human evaluations of MT quality and compare two different models, a non-contextual baseline system and a lightly contextual model (from Section 7.1). Despite the fact that evaluation is performed uniquely from a monolingual perspective, we show that tangible differences can be seen between the two different models. The corpus can also provide us with rich information for further study.

9.2 Perspectives

We have provided critiques and short-term perspectives at the end of each section in Part II to conclude on the individual experiments and approaches. We will not repeat these perspectives and instead choose to provide a reflection on the role contextual MT may have to play in four particularly active areas of MT: evaluation, the interpretation of NMT, MT for low resource language pairs and multi-modal MT.

9.2.1 Evaluation of MT

Evaluation methods have been at the heart of this thesis, but remain crucial for the future of MT. We have explored several different methods of evaluating, including the use of contrastive test sets (Bawden et al., 2018b) (Section 7.1), automatic evaluation of gender-marking as determined by speaker gender (Section 7.2), and the development of a test set of spontaneously produced, bilingual data (Section 8).

The creation of new test sets and challenge sets is a continual necessity in order to cover new phenomena and new language pairs. Challenge sets to target particular phenomena of varying difficulty provide us with the means to understand the limits of current models and to provide ways of tackling them in the future (Isabelle et al., 2017; Burlot and Yvon, 2017; Isabelle and Kuhn, 2018). Larger test suites focusing on individual phenomena make it possible to analyse models' performance on a slightly larger scale (Guillou and Hardmeier, 2016; Rios Gonzales et al., 2017).

While current evaluation methods for contextual MT are highly useful, producing specific challenge sets is often time-consuming. One current gap in the evaluation literature for contextual MT is an evaluation set of examples containing context-dependent phenomena that are representative of real language use. Since manually identifying real sentences is very time-consuming, a long-term goal would be to automatically construct such a

test set, including examples that are context-dependent but are not generally found in manually constructed sets. We provided some preliminary theoretical reflections on how such a test set could be produced in (Bawden et al., 2018a). Ideally a method to detect such sentences would adhere to three key principles (i) the unbiased nature of the test set (with respect to a particular MT architecture), (ii) diversity and a large coverage of the phenomena detected and (iii) easy transferability to other language pairs. In (Bawden et al., 2018a), we weighed up the advantages and disadvantages of two potential methods, the first relying on sentence embeddings (e.g. DOC2VEC; Mikolov et al., 2013) and the second on contextual word embeddings (e.g. CONTEXT2VEC; Melamud et al., 2016). Both methods were designed to test whether a ‘better’ representation can be achieved when extra-sentential context is used over the representation obtained using information from the current sentence alone. In preliminary experiments, we observed three main limitations that would have to be surmounted for such a detection method to be viable: (i) the intrinsic probability of words, as determined by their frequency, has a very large effect on their probability in context, making it very complicated to assess the effect of adding context, (ii) the capacity of generic language models to model complex and structured problems such as coreference chains is insufficient, even for simple, short utterances, and (iii) in light of the second limitation, all context, even if not directly relevant to the translation of the ambiguous word, has an effect on the representation of a word or a sentence. We have little control over which information is considered important by the model, particularly if we wish to keep the approach as general as possible. However, recent progress made in techniques for sentence embedding (Subramanian et al., 2018; Cer et al., 2018), in particular in multi-lingual settings (Singla et al., 2018), could provide an opportunity to develop new methods in the future.

9.2.2 Interpretability of contextual NMT strategies

A criticism of NMT is the lack of transparency with respect to its interpretability. In comparison with phrase-based models, which are easier to interpret thanks to a direct modelling between discrete units, analysing and interpreting neural approaches to MT is far more difficult, and remains an open challenge (Koehn and Knowles, 2017). At this point in time, the analysis of neural networks is gaining traction (cf. the Blackbox NLP workshop¹ organised in 2018) and an important part of future work will be to provide a more detailed analysis of how context is integrated in NMT models, in order to better understand the limitations of current approaches.

In our experiments for the integration of linguistic context (Section 7.1), we saw that much of the context needed for cohesion-based phenomena was channelled through the RNN decoder. Unlike the attention mechanism, which can be interpreted as an

¹<https://blackboxnlp.github.io>

alignment between source elements and each decoding step, the decoder is much less easily analysed. However, more work is being dedicated to analysing and visualising models in order to better understand how information is propagated throughout the network. For example, Voita et al. (2018) are able to analyse the attention mechanism of their contextual neural architecture, since no recurrent components were used. Going beyond attention, Ding et al. (2017) provide an analysis of the contribution of source and target words through a visualisation of their respective contributions at each decoding step, based on layer-wise relevance propagation. They show that the contributions of words differs greatly between layers, and that there can be more insights to be found than those offers through the attention mechanism. Gaining insights into how information is propagated could offer clues about how best to use context to influence translations and how much context can be realistically kept in memory. One of our future aims is to extend methods to a wider context. Within this work we saw that including the previous sentence in translation did not lead to perfect scores, and it is unlikely that the same technique could be used to integrate context from a document of indefinite length, since information is likely to be lost over time. Knowing how information is memorised would offer some information about how far back we can expect to go when taking into account document-level linguistic context.

In addition to analysing the internal functions of the models, more knowledge about the models we have been using can be sought through studying the behaviour of the models. This can be done in two ways: by comparing MT quality in different settings (e.g. with variable amounts of training data, different sentence lengths, etc.), or by studying the usefulness of representations learnt by the architectures for other tasks (through transfer learning). Testing different settings could help us to understand our observation in Section 7.2 that overall translation quality (as measured by BLEU) is affected by the addition of a semantically void token at the beginning of each sentence. We have not yet been able to confirm why this technique leads to gains, or whether the gains are translated into tangible quality improvements. Moreover, it would be interesting to test the effect on different language pairs and on test sets of different average sentence lengths to see whether or not the effect persists. It may also be important to investigate this further, to offer potential tips to improve translation quality, and/or to potentially offer warnings about the cause of certain quality gains that are reported in the literature. In terms of transfer learning, a better idea of how information is represented within NMT models is to use the representations in another setting. This has recently become popular for a number of NLP tasks (Chrupała et al., 2017; Peters et al., 2017). For NMT, there are similar such works. Belinkov et al. (2017) provide an extrinsic evaluation of the representations learnt during training of the NMT model by using them in separate NLP tasks (PoS and morphological tagging). They show that character-based models in particular are better at representing morphological information, and help in particular for unseen words. They also show that the lower layers of the network contain the most useful morphological

information. For word sense disambiguation, Marvin and Koehn (2018) analyse the word embeddings of the deeper layers of the NMT encoder to test their ability to disambiguate word senses. Their study shows that the method is promising and could provide more insight if applied on a larger scale. There is no reason why similar techniques cannot be used for contextual MT, in particular for the evaluation of the effect that adding context has on the hidden representations of sentences translated by the NMT model.

9.2.3 Contextual MT for low resource language pairs

A research domain that has been gaining popularity is MT for low resource language pairs (i.e. in our case those for which there exists little parallel data). As techniques progress, the community is expanding their interest to a wider range of language pairs than those that were initially used. This can be seen for example in the choice of language pairs for the annual translation shared tasks at WMT, which in recent years have included languages such as Turkish and Romanian, to the exclusion of some highly resourced languages such as French (Bojar et al., 2016a, 2017).

In our experiments within this thesis, we have concentrated specifically on language pairs for which large quantities of data are available.² We have considered it important to work on high resourced language pairs for two reasons, both linked to the overall translation quality that can be achieved. Firstly, concentrating on better taking into account context only realistically makes sense if the quality of translation is adequate in the first place. For example, correctly translating an anaphoric pronoun is likely only to have a positive impact on the translation if the translation itself makes sense, and is sufficiently understandable. While integrating extra-sentential context is important, in terms of priority, it comes after ensuring that intra-sentential context is first modelled correctly. Secondly, our aim was to specifically study and evaluate approaches to integrating context. This is much easier to do in a setting in which translation is already of a reasonable quality, and, importantly, provides us with a scenario in which it becomes necessary not to confuse an improvement in the way context is exploited with gains in overall translation quality.

It has been shown by Jean et al. (2017a) that certain architectures designed to integrate context bring about improvements in BLEU over a baseline model only in low resource settings. They designed a contextual architecture to exploit inter-sentential links between the current and previous sentences. Both the model's capacity to exploit context and its overall translation quality (as measured by BLEU) increased as the quantity of training data used was increased. However, the difference in BLEU score between the contextual

²Our TQ experiments for the translation of German to English in Section 6.1 were those trained on the least data. We observed in these experiments that the quality of the translations has an impact on the degree to which our post-processing improved the translation, which is unsurprising since the baseline translations were of an inferior quality to the other models (better resourced) language pairs.

model and a baseline one gradually diminished as more training data was added, resulting in a lower BLEU score for the contextual model once all training data was used. It is therefore important not just to look at improvements in BLEU score, but also to provide a more targeted evaluation. On the other hand, it is also important to look at the impact adding context has on the overall translation quality and to avoid reducing the overall quality. This can only be really tested in optimal settings (i.e. using the best experimental settings, including sufficient training data).

Designing MT models for languages for which less data is available poses different challenges. Many of the recent NMT architectures require large quantities of data to avoid underfitting parameters, and the models perform less well on smaller quantities of data. However, there have recently been advances in techniques to handle these types of settings. These include unsupervised MT (Artetxe et al., 2018; Lample et al., 2018), alternative techniques to better exploit available monolingual data (Gulcehre et al., 2015; Sennrich et al., 2016c; Zhang and Zong, 2016; Di Gangi and Federico, 2017; Currey et al., 2017) and transfer techniques between close languages (Zoph et al., 2016; Passban et al., 2017). Many of the more recent techniques, in particular for unsupervised MT, rely on strategies that consist in simulating a scenario in which we have reasonable quantities of data. For example, both Artetxe et al. (2018) and Lample et al. (2018) rely on techniques consisting of learning NMT models on monolingual data by iteratively translating monolingual corpora in both source and target languages using the current model parameters and then using the corpora and the translations produced as an artificial parallel corpus to train the model at the next iteration. Therefore, in some sense, this can be seen as a way of adapting NMT architectures to produce artificial data, in the same way that back-translated data has been used to augment the quantity of data available, in particular for domain adaptation (Sennrich et al., 2016c). This contrasts with another possible way of approaching low resource MT, which would be to concentrate on providing rich annotations of the data available (Li et al., 2017, 2018), or using external resources (Arthur et al., 2016). The two are not necessarily incompatible, and this will most certainly figure in future work in the domain.

In light of these developments in low resource MT, what does this mean for how contextual MT could be applied in this setting? Given that a reasonable baseline translation quality is necessary before reasonably targeting contextual phenomena, the future of contextual MT for low-resource languages may well depend on the progress made. If a reasonable translation quality can be reached using these techniques, the same techniques as used in high resource settings could then be used. The alternative would be to use similar strategies to those used at the beginning of this thesis: i.e. using highly structured linguistic context that is applied either in a pre-processing or post-processing step. However, as seen in Section 6.2, the efficiency of such methods is heavily dependent on the quality of the processing tools available. For lower resourced languages, such tools are less likely to be available.

9.2.4 Contextual MT to Multimodal MT

Another area of MT that has received heightened interest over the last couple of years has been multi-modal NMT, particular in terms of the joint processing of images and text for caption generation and translation (Specia et al., 2016; Elliott et al., 2017). Several tasks have been designed to test the exploitation of images and text, including multi-modal caption translation (Hitschler et al., 2016) and image description generation in several languages (Elliott et al., 2015). Images can be seen as context in themselves (more specifically as unstructured extra-linguistic context), as they can provide information that is not present in the sentence to be translated and help to resolve ambiguity that would otherwise arise during translation. The information is however of a different nature from the contextual types seen in this thesis. We have explored methods for integrating sequential linguistic context and discrete labels indicating extra-linguistic context, whereas images contain fine-grained information, structured in a different way. Some of the strategies previously used to integrate visual context have nevertheless been very similar to those seen in Chapter 7. For example, Libovický and Helcl (2017), who introduce the idea of using hierarchical attention for NMT and apply it to linguistic context, extend this idea to multi-modal translation in (Helcl and Libovický, 2017). They change the additional encoder to a convolutional network trained to represent image data, rather than use the bi-directional RNN encoder previously used for linguistic context. This is an area of research that would be fruitful to explore, and could even provide us with new ways of exploiting the other types of context explored in this thesis.

A second type of multi-modal MT is the use of acoustic information to help the translation of speech. Within this thesis, we made the choice to work on the translation of *written* dialogues, rather than of *oral* dialogues. This enabled us to set aside the complications brought about by dealing with a speech signal, which can be noisy and lead to error propagation downstream. We were therefore able to concentrate on the contextual aspect of our problem, making use of the large quantity of written parallel sentences available. However, in future work it would be interesting to also take acoustic information into account. Some work has already been done in this direction: Deena et al. (2017) study the impact of including acoustic embeddings to aid NMT, and show provide complementary information to topic-related embeddings and show promise for future work. We chose in this current work to limit our study of context to linguistic and extra-linguistic context. However a third type of context, para-linguistic context, which we mentioned in Chapter 2, could also be worthwhile studying. Para-linguistic context could include acoustic information, linked for example to intonation and prosody, which could provide us with key information about the attitude of the speaker and help to disambiguate certain constructions, particularly those containing syntactic ambiguity. This is an element that could have helped our study of TQs in Section 6.1, in terms of their detection and in identifying the attitude conveyed by the speaker.

9.2.5 Conclusion: To the future and beyond the sentence

Within this thesis, we have explored a range of contextual MT methods and new ways of evaluating them. Our contributions have been guided by ongoing trends in contextual MT research and have followed the shift in state-of-the-art MT techniques, in particular the shift from SMT to NMT. This is reflected in the different ways in which we have structured linguistic context, and the approaches we have used to exploit both linguistic and extra-linguistic context. In addition to new methods of integrating context, we also provided new evaluation protocols and new evaluation resources, and through our discussions and analyses we raised a number of open questions for the future.

Integrating context into MT remains an important aim of the MT community, and is likely to receive a lot of interest in coming years. Having recently seen a move towards the use of unstructured linguistic context, we may expect to see efforts to integrate richer types of contextual information in order to better exploit the context currently available. However, it is also likely that advances in MT architectures will also lead to new ways of information being learnt within MT models themselves.

Context-aware translation models

This appendix contains the full description of training and decoding parameters and other information relative to the experimental setups of the experiments in Sections 7.1 and 7.2.

A.1 Translating discourse phenomena with unstructured linguistic context

A.1.1 Training and decoding parameters

Training and testing data Models are trained and tested on English-to-French parallel subtitles from OpenSubtitles2016¹ (Lison and Tiedemann, 2016). The data is first corrected using heuristics (e.g. minor corrections of OCR and encoding errors). It is then tokenised, further cleaned (keeping subtitles ≤ 80 tokens) and truecased using the Moses toolkit (Koehn et al., 2007) and finally split into subword units using BPE (Sennrich et al., 2016d).² We run all experiments in a high-resource setting, with a training set of ≈ 29 M parallel sentences, with vocabulary sizes of ≈ 55 k for English and ≈ 60 k for French.

Training parameters All models are implemented in NEMATUS (Sennrich et al., 2017), of which the modified code is available at <https://diamt.limsi.fr/eval.html>. The baseline,

¹<http://www.opensubtitles.org>

²90,000 merge operations with a minimum threshold of 50.

2-TO-2 and 2-TO-1 models correspond to the baseline encoder-decoder architecture with attention (Bahdanau et al., 2015). The multi-encoder models are modifications of the original architecture to include a secondary (or two additional) encoders, which are identical in terms of their structure to the original encoder. However parameters are not shared across encoders.

Training is performed using the Adam optimiser with a learning rate of 0.0001 until convergence. We use embedding layers of dimension 512 and hidden layers of 1024. For training, the maximum sentence length is 50 subwords.³ We use batch sizes of 80, tied decoder embeddings and layer normalisation. The hyper-parameters are the same for all models and are the same as those used for the University of Edinburgh submissions to the news translation shared task at WMT16 and WMT17. Validation on our held-out dev set of 3000 sentences is performed every 10,000 updates (using BLEU) and checkpointed models produced every 30,000 updates.

Decoding setup The final models used for translation and scoring are ensembles of the last three check-pointed models during training.⁴ This not only results in a better scoring final model, but also ensures a greater degree of stability with respect to the results. We use the option `-SUPPRESS_UNK`, which suppresses hypotheses containing the UNK token during decoding. We use normalisation of scores for sentence length and a beam size of 12.

Sentences to be translated are processed in the same way as the training and validation data (except for the removal of longer sentences). For models requiring concatenation of the source and/or target sentences, the sentences are concatenated with their previous sentence once the other pre-processing steps have been applied. The concatenation symbol `<CONCAT>` is added to the vocabulary, replacing a very rare word.

Once translated, sentences are post-processed to remove the effects of subword splitting and then detruccased and detokenised using the Moses scripts. All BLEU scores are calculated using the `MULTI-BLEU-DETOK.PERL` script on detokenised and cased translations.

Scoring Source and target sentences can be scored using NEMATUS. These scores correspond to cross-entropy scores calculated by the model, which are normalised for sentence length. These can be provided per sentence (normalised) or per word. We use these scores to determine for a pair of contrastive sentences which sentence is ranked higher by the model (i.e. has a lower cross entropy score).

³76 when source sentences are concatenated to the previous sentence in order to keep the same percentage of training sentences as for other models.

⁴Ensembling is performed as per its implementation in NEMATUS. The geometric average of all individual models' probability distributed is calculated to produce the distribution of the ensembled model.

A.1.2 Visualisation of hierarchical attention weights

Table A.1 shows the previous source context’s hierarchical attention weight at each decoding step for two hierarchical models: s-HIER and s-HIER-TO-2.

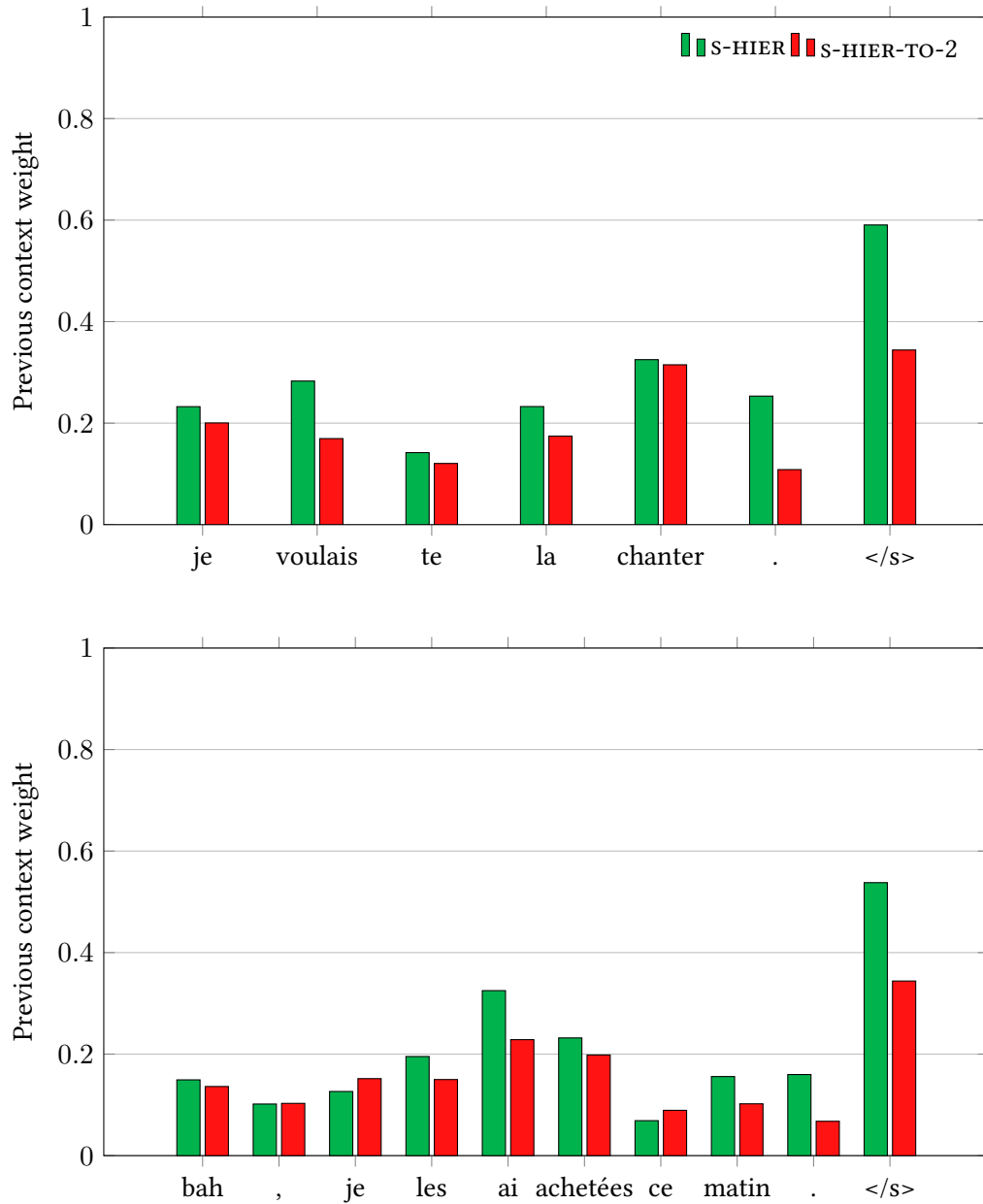


Figure A.1: A comparative visualisation of hierarchical attention weights associated with the previous context representation at each decoding step (shown for each word decoded). The weights are shown for the two models s-HIER and s-HIER-TO-2 for two sentences of our anaphora test set.

A.2 Contextual NMT with extra-linguistic context

A.2.1 Experimental setup

Training and decoding parameters The setup of these experiments is the same as presented in Appendix A.1.1 in terms of the hyper-parameters and software used.

Pre-training data The annotated data used to train our contextually adapted models are described within Section 7.2. The data used to pre-train our models is OPENSUBTITLES2016, minus the parallel sentences included in the annotated data, which is also used for testing.

Contrarily to our experiments in Section 7.1, we first filter the corpus to remove parallel sentences that are poorly aligned or only partially translated. Whereas this was less problematic for our previous experiments, as we evaluated using contrastive scoring, the noisiness of the data has an impact on our targeted evaluation in this section. Occasional truncation of translations (i.e. the source sentence is not entirely translated) leads to fewer sentences containing gender-marked constructions that can be used for our evaluation. We automatically align the subtitles on the word level using FASTALIGN (Dyer et al., 2013), with symmetrised and the *grow-diag-final-and* strategy. We then filter out those sentence for which fewer than 80% of either source or target tokens are not aligned. The choice of 80% is a compromise between conserving large quantities of data and limiting noise in the data. The final training data contains 24,140,225 parallel sentences.

APPENDIX B

DIABLA: A corpus for the evaluation of contextual MT

This appendix contains information relative to Chapter 8, our corpus of bilingual dialogues. In Appendix B.1 we provide the full list of scenarios and roles used in dialogue collection. In Appendix B.2, we present the final evaluation form the participants fill in once the dialogue is complete.

B.1 Role-play scenarios

<p>You are both lost in a forest. Roles: N/A</p>
<p>You are chefs preparing a meal. Role 1: You are the head chef and you are talking to your subordinate. Role 2: You are the subordinate chef and you are talking to the head chef.</p>
<p>You are in a classroom. Role 1: You are the teacher and you are talking to a student. Role 2: You are the student and you are talking to your teacher.</p>
<p>You are feeding the ducks by the pond. Roles: N/A</p>
<p>You are both organising a party. Role 1: It's your party. Role 2: It's their party.</p>
<p>You are both stuck in a lift at work. Role 1: You are an employee and you are with your boss. Role 2: You are the boss and are with an employee.</p>
<p>You are in a retirement home. Role 1: You are visiting and talking to an old friend. Role 2: You are a resident and you are talking with an old friend who is visiting you.</p>
<p>You are in a bar. Role 1: You are the bartender and talking to a customer. Role 2: You are a customer and are talking to the bartender.</p>
<p>You are in an aeroplane. Role 1: You are scared and are speaking to the person sitting next to you. Role 2: You are speaking to the person next to you, who is scared.</p>
<p>You are at home in the evening. Role 1: You are telling your spouse about the awful day you had. Role 2: You are listening to your spouse telling you about the awful day they had.</p>
<p>You are in a psychiatrist's consulting room. Role 1: You are the psychiatrist and are with your patient. Role 2: You are a patient and you are talking to your psychiatrist.</p>
<p>You are on holiday by the pool. Role 1: You are trying to relax and the other person wants to do something else. Role 2: You want to do something else and the other person is trying to relax.</p>

Figure B.1: The twelve scenarios and roles chosen to guide the dialogues.

B.2 Dialogue collection: Final evaluation form

Overall Evaluation

Please take moments to answer some very general questions on the dialogue you just had. [See the dialogue again here \(opens in another window\)](#)

Were there any practical problems with the dialogue?

- The other person was absent
- The other person did not follow the instructions
- There was a technical problem with DiaBLa itself
(the interface, the machine translation system...)
- No, none of the above

Please rate the following aspects of your partner's translated speech:

	Excellent	Good	Average	Poor	Very poor
Grammar ^(?)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Meaning ^(?)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Word choice ^(?)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Style ^(?)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Coherence ^(?)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Specific comments on the overall quality of the dialogue

What particular aspects were poorly translated?

Specific comments on the DiaBLa chat interface

If I were to communicate by instant message with somebody with whom I do not share a common language, I would gladly use an interface such as DiaBLa, as the quality of the translations is good enough.

Ok

Figure B.2: The evaluation form presented to participants at the end of a dialogue.

Bibliography

- Agrawal, R., Turchi, M., and Negri, M. (2018). Contextual Handling in Neural Machine Translation: Look Behind, Ahead and on Both Sides. In *Proceedings of the 21st Annual Conference of the European Association for Machine Translation (EAMT'18)*, pages 11–20, Alicante, Spain.
- Allen, R. (1987). Several Studies on Natural Language and Back-Propagation. In *Proceedings of the IEEE 1st International Conference on Neural Networks*, pages 335–341, San Diego, California, USA.
- Artetxe, M., Labaka, G., Agirre, E., and Cho, K. (2018). Unsupervised Neural Machine Translation. In *Proceedings of the 6th International Conference on Learning Representations (ICLR'18)*, Vancouver, Canada.
- Arthur, P., Neubig, G., and Nakamura, S. (2016). Incorporating Discrete Translation Lexicons into Neural Machine Translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP'16)*, pages 1557–1567, Austin, Texas, USA.
- Axelrod, A., He, X., and Gao, J. (2011). Domain Adaptation via Pseudo In-Domain Data Selection. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP'11)*, pages 355–362, Edinburgh, UK.
- Axelsson, K. (2011). A cross-linguistic study of grammatically-dependent question tags. *Studies in Language*, 35(4):793–851.
- Bahdanau, D., Cho, K., and Bengio, Y. (2015). Neural Machine Translation by Jointly Learning to Align and Translate. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR'15)*, San Diego, California, USA.

- Baldi, P. and Hornik, K. (1989). Neural Networks and Principal Component Analysis: Learning from Examples Without Local Minima. *Neural Networks*, 2:53–58.
- Bar-Hillel, Y. (1960). The Present Status of Automatic Translation of Languages. *Advances in Computers*, 1:91–163.
- Baroni, M., Dinu, G., and Kruszewski, G. (2014). Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL'14)*, pages 238–247, Baltimore, Maryland, USA.
- Barras, C., Zhu, X., Meignier, S., and Gauvain, J.-L. (2006). Multistage speaker diarization of broadcast news. *IEEE Transactions on Audio, Speech and Language Processing*, 14(5):1505–1512.
- Bawden, R. (2016). Cross-lingual Pronoun Prediction with Linguistically Informed Features. In *Proceedings of the 1st Conference on Machine Translation (WMT'16)*, pages 564–570, Berlin, Germany.
- Bawden, R. (2017a). Machine Translation, it's a question of style, innit? The case of English tag questions. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP'17)*, pages 2497–2502, Copenhagen, Denmark.
- Bawden, R. (2017b). Machine Translation of Speech-Like Texts: Strategies for the Inclusion of Context. In *Proceedings of the REcontres jeunes Chercheurs en Informatique pour le TAL (RECITAL'17)*, pages 1–14, Orléans, France.
- Bawden, R., Lavergne, T., and Rosset, S. (2018a). Detecting context-dependent sentences in parallel corpora. In *Proceedings of the 25th Conférence sur le Traitement Automatique des Langues Naturelles (TALN'18)*, pages 393–400, Rennes, France.
- Bawden, R., Sennrich, R., Birch, A., and Haddow, B. (2018b). Evaluating Discourse Phenomena in Neural Machine Translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics; Human Language Technologies (NAACL-HLT'18)*, pages 1304–1313, New Orleans, Louisiana, USA.
- Bawden, R., Wisniewski, G., and Maynard, H. (2016). Investigating gender adaptation for speech translation. In *Proceedings of the 23rd Conférence sur le Traitement Automatique des Langues Naturelles (TALN'16)*, pages 490–497, Paris, France.
- Belinkov, Y., Durrani, N., Dalvi, F., Sajjad, H., and Glass, J. (2017). What do Neural Machine Translation Models Learn about Morphology? In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL'17)*, pages 861–872, Vancouver, Canada.

- Bengio, Y., Ducharme, R., Vincent, P., and Janvin, C. (2003). A Neural Probabilistic Language Model. *Journal of Machine Learning Research*, 3:1137–1155.
- Bentivogli, L., Bisazza, A., Cettolo, M., and Federico, M. (2016). Neural versus Phrase-Based Machine Translation Quality: a Case Study. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP'16)*, pages 257–267, Austin, Texas, USA.
- Bertoldi, N., Haddow, B., and Fouet, J.-B. (2009). Improved Minimum Error Rate Training in Moses. *The Prague Bulletin of Mathematical Linguistics*, 91:7–16.
- Blei, D., Ng, A., and Jordan, M. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022.
- Bohnet, B. and Nivre, J. (2012). A Transition-Based System for Joint Part-of-Speech Tagging and Labeled Non-Projective Dependency Parsing. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL'12)*, pages 1455–1465, Jeju Island, South Korea.
- Bojar, O., Chatterjee, R., Federmann, C., Graham, Y., Haddow, B., Huang, S., Huck, M., Koehn, P., Liu, Q., Logacheva, V., Monz, C., Negri, M., Post, M., Rubino, R., Specia, L., and Turchi, M. (2017). Findings of the 2017 Conference on Machine Translation (WMT17). In *Proceedings of the 2nd Conference on Machine Translation (WMT'17)*, pages 169–214, Copenhagen, Denmark.
- Bojar, O., Chatterjee, R., Federmann, C., Graham, Y., Haddow, B., Huck, M., Jimeno Yepes, A., Koehn, P., Logacheva, V., Monz, C., Negri, M., Névól, A., Neves, M., Popel, M., Post, M., Rubino, R., Scarton, C., Specia, L., Turchi, M., Vespoor, K., and Zampieri, M. (2016a). Findings of the 2016 Conference on Machine Translation (WMT16). In *Proceedings of the 1st Conference on Machine Translation (WMT'16)*, pages 131–198, Berlin, Germany.
- Bojar, O., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Huck, M., Koehn, P., and Monz, C. (2018). Findings of the 2018 conference on machine translation. In *Proceedings of the 3rd Conference on Machine Translation (WMT'18)*, pages 272–307.
- Bojar, O., Federmann, C., Haddow, B., Koehn, P., Post, M., and Specia, L. (2016b). Ten Years of WMT Evaluation Campaigns: Lessons Learnt. In *Proceedings of the LREC 2016 Workshop "Translation Evaluation – From Fragmented Tools and Data Sets to an Integrated Ecosystem"*, Portorož, Slovenia.
- Bredin, H., Roy, A., Pêcheux, N., and Allauzen, A. (2014). "Sheldon speaking, bonjour!" - Leveraging Multilingual Tracks for (Weakly) Supervised Speaker Identification. In *Proceedings of the 22nd ACM International Conference on Multimedia (ACM-MM'14)*, pages 137–146, Orlando, Florida, USA.

- Broscheit, S., Poesio, M., Ponzetto, S. P., Rodriguez, K. J., Romano, L., Uryupina, O., Versley, Y., and Zanolini, R. (2010). BART: A Multilingual Anaphora Resolution System. In *Proceedings of the 5th International Workshop on Semantic Evaluation (SemEval'10)*, pages 104–107, Uppsala, Sweden.
- Brown, P. F., Cocke, J., Della Pietra, S. A., Della Pietra, V. J., Jelinek, F., Mercer, R. L., and Roossin, P. S. (1988). A Statistical Approach to Language Translation. In *Proceedings of the 12th International Conference on Computational Linguistics*, pages 71–76, Budapest, Hungary.
- Brown, P. F., Della Pietra, S. A., Della Pietra, V. J., and Mercer, R. L. (1993). The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, 19(2).
- Burlot, F. and Yvon, F. (2017). Evaluating the morphological competence of Machine Translation Systems. In *Proceedings of the Second Conference on Machine Translation (WMT'17)*, pages 43–55, Copenhagen, Denmark.
- Caglayan, O., Barrault, L., and Bougares, F. (2016a). Does Multimodality Help Human and Machine for Translation and Image Captioning? In *Proceedings of the 1st Conference on Machine Translation (WMT'16)*, pages 627–633, Berlin, Germany.
- Caglayan, O., Barrault, L., and Bougares, F. (2016b). Multimodal Attention for Neural Machine Translation. In *arXiv:1609.03976*.
- Calixto, I., Liu, Q., and Campbell, N. (2017). Incorporating global visual features into attention-based neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP'17)*, pages 992–1003, Copenhagen, Denmark.
- Callison-Burch, C., Koehn, P., Monz, C., Peterson, K., Przybocki, M., and Zaidan, O. F. (2010). Findings of the 2010 joint workshop on statistical machine translation and metrics for machine translation. In *Proceedings of the Joint 5th Workshop on Statistical Machine Translation and Metrics (WMT-MetricsMATR'10)*, pages 17–53, Uppsala, Sweden.
- Callison-Burch, C., Osborne, M., and Koehn, P. (2006). Re-evaluating the Role of BLEU in Machine Translation Research. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL'06)*, pages 249–256, Trento, Italy.
- Cameron, D. and Coates, J. (1985). Some problems in the sociolinguistic explanation of sex differences. *Language and Communication*, 5(3):143–151.
- Canavan, A. and Zipperlen, G. (1996). *CALLHOME Spanish Speech LDC96S35*. Linguistic Data Consortium, Philadelphia, Pennsylvania, USA.

- Carpuat, M. (2009). One translation per discourse. In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions (SEW'09)*, pages 19–27, Boulder, Colorado, USA.
- Carpuat, M. and Simard, M. (2012). The Trouble with SMT Consistency. In *Proceedings of the 7th Workshop on Statistical Machine Translation (WMT'12)*, pages 442–449, Montreal, Canada.
- Carpuat, M. and Wu, D. (2005). Word Sense Disambiguation vs. Statistical Machine Translation. In *Proceedings of the 43rd Meeting of the Association for Computational Linguistics (ACL'05)*, pages 387–394, Ann Arbor, Michigan, USA.
- Carpuat, M. and Wu, D. (2007). How Phrase Sense Disambiguation outperforms Word Sense Disambiguation for Statistical Machine Translation. In *Proceedings of the 11th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI'07)*, pages 43–52, Skövde, Sweden.
- Cer, D., Yang, Y., Kong, S.-y., Hua, N., Limtiaco, N., John, R. S., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., Sung, Y.-H., Strophe, B., and Kurzweil, R. (2018). Universal Sentence Encoder. *arXiv:1803.11175 [cs]*.
- Cettolo, M., Girardi, C., and Federico, M. (2012). WIT³: Web Inventory of Transcribed and Translated Talks. In *Proceedings of the 16th Conference of the European Association for Machine Translation (EAMT'12)*, pages 261–268, Trento, Italy.
- Chafe, W. (1976). Givenness, Contrastiveness, Definiteness, Subjects, Topics, and Point of View. In Li, C., editor, *Subject and Topic*, pages 22–55. Academic Press, New York City, New York, USA.
- Chan, Y. S., Ng, H. T., and Chiang, D. (2007). Word Sense Disambiguation Improves Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL'07)*, pages 33–40, Prague, Czech Republic.
- Chen, B. and Cherry, C. (2014). A Systematic Comparison of Smoothing Techniques for Sentence-Level BLEU. In *Proceedings of the 9th Workshop on Statistical Machine Translation (WMT'14)*, pages 362–367, Baltimore, Maryland, USA.
- Chen, S. F. and Goodman, J. (1996). An Empirical Study of Smoothing Techniques for Language Modeling. In *Proceedings of the 34th annual meeting on Association for Computational Linguistics (ACL'96)*, pages 310–318, Santa Cruz, California, USA.
- Cherry, C. and Foster, G. (2012). Batch Tuning Strategies for Statistical Machine Translation. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT'12)*, pages 427–436, Montreal, Canada.

- Chiang, D. (2007). Hierarchical Phrase-Based Translation. *Computational Linguistics*, 33(2):201–228.
- Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP'14)*, pages 1724–1734, Doha, Qatar.
- Chrupała, G., Gelderloos, L., and Alishahi, A. (2017). Representations of language in a model of visually grounded speech signal. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL'17)*, pages 613–622, Vancouver, Canada.
- Collobert, R. and Weston, J. (2008). A Unified Architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning. In *Proceedings of the 25th International Conference on Machine Learning (ICML'08)*, pages 160–167, Helsinki, Finland.
- Comelles, E., Gimenez, J., Màrquez, L., Castellon, I., and Arranz, V. (2010). Document-level Automatic MT Evaluation based on Discourse Representations. In *Proceedings of the Joint 5th Workshop on Statistical Machine Translation and MetricsMATR*, pages 333–338, Uppsala, Sweden.
- Costa-Jussà, M. R. and Fonollosa, J. A. R. (2016). Character-based Neural Machine Translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL'16)*, pages 357–361, Berlin, Germany.
- Crego, J. M., Yvon, F., and Mariño, J. B. (2011). Ncode: an Open Source Bilingual N-gram SMT Toolkit. *The Prague Bulletin of Mathematical Linguistics*, 96(1):49–58.
- Currey, A., Valerio, A., and Heafield, K. (2017). Copied Monolingual Data Improves Low-Resource Neural Machine Translation. In *Proceedings of the 2nd Conference on Machine Translation (WMT'17)*, pages 148–156, Copenhagen, Denmark.
- Danlos, L. (2005). Automatic recognition of French expletive pronoun occurrences. In *Proceedings of the 2nd International Joint Conference on Natural Language Processing (IJCNLP'05)*, pages 73–78, Jeju Island, South Korea.
- de Beaugrande, R. and Dressler, W. (1981). *Introduction to Text Linguistics*. Longman, London, UK.
- de Marneffe, M.-C., Recasens, M., and Potts, C. (2015). Modeling the Lifespan of Discourse Entities with Application to Coreference Resolution. *Journal of Artificial Intelligence Research*, 52:445–475.

- Deena, S., Ng, R., Madhyastha, P., Specia, L., and Hain, T. (2017). Exploring the use of acoustic embeddings in neural machine translation. In *Proceedings of the Automatic Speech Recognition and Understanding Workshop (ASRU'17)*, pages 450–457, Okinawa, Japan.
- Deerwester, S., Dumais, S., Furnas, G., and Landauer, T. (1990). Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6).
- Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society. Series B (methodological)*, 39(1):1–38.
- Denis, P. and Sagot, B. (2012). Coupling an annotated corpus and a lexicon for state-of-the-art POS tagging. *Language Resources and Evaluation*, 46(4):721–736.
- Di Gangi, M. A. and Federico, M. (2017). Monolingual Embeddings for Low Resourced Neural Machine Translation. In *Proceedings of the 14th International Workshop on Spoken Language Translation (IWSLT'17)*, pages 97–104, Tokyo, Japan.
- Ding, Y., Liu, Y., Luan, H., and Sun, M. (2017). Visualizing and Understanding Neural Machine Translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL'17)*, pages 1150–1159, Vancouver, Canada.
- Dorr, B., Snover, M., and Madnani, N. (2011). Introduction. In Olive, J., Christianson, C., and McCary, J., editors, *Handbook of Natural Language Processing and Machine Translation*, pages 745–758. Springer-Verlag, New York City, New York, USA.
- Dyer, C., Chahuneau, V., and Smith, N. A. (2013). A simple, fast, and effective reparameterization of IBM Model 2. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT'13)*, pages 644–648, Denver, Colorado, USA.
- Elaraby, M., Tawfik, A. Y., Khaled, M., Hassan, H., and Osama, A. (2018). Gender Aware Spoken Language Translation Applied to English-Arabic. In *Proceedings of the 2nd International Conference on Natural Language and Speech Processing (ICNLSP'18)*, Algiers, Algeria.
- Elliott, D., Frank, S., Barrault, L., Bougares, F., and Specia, L. (2017). Findings of the Second Shared Task on Multimodal Machine Translation and Multilingual Image Description. In *Proceedings of the 2nd Conference on Machine Translation (WMT'17)*, pages 215–233, Copenhagen, Denmark.
- Elliott, D., Frank, S., and Hasler, E. (2015). Multilingual Image Description with Neural Sequence Models. *arXiv:1510.04709 [cs]*.

- Elliott, D., Frank, S., Khalil, S., and Specia, L. (2016). Multi30k: Multilingual English-German Image Descriptions. In *Proceedings of the 5th Workshop on Vision and Language*, pages 70–74, Berlin, Germany.
- Fagerland, M. W., Lydersen, S., and Laake, P. (2013). The McNemar test for binary matched-pairs data: mid-p and asymptotic are better than exact conditional. *BMC Medical Research Methodology*, 13(91).
- Federmann, C. and Lewis, W. (2016). Microsoft Speech Language Translation (MSLT) Corpus: The IWSLT 2016 release for English, French and German. In *Proceedings of the 13th International Workshop on Spoken Language Translation (IWSLT'16)*, Seattle, Washington, USA.
- Finch, A., Sumita, E., and Nakamura, S. (2009). Class-Dependent Modeling for Dialog Translation. *IEICE Transactions on Information and Systems*, 92(12):2469–2477.
- Firat, O., Cho, K., and Bengio, Y. (2016). Multi-way, multilingual neural machine translation with a shared attention mechanism. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT'16)*, pages 866–875, San Diego, California, USA.
- Firth, J. (1957). A synopsis of linguistic theory 1930-55. *Studies in linguistic analysis*, pages 1–32.
- Foster, G. and Kuhn, R. (2007). Mixture-Model Adaptation for SMT. In *Proceedings of the 2nd Workshop on Statistical Machine Translation (WMT'07)*, pages 128–135, Prague, Czech Republic.
- Gale, W. A., Church, K. W., and Yarowsky, D. (1992). One sense per discourse. In *Proceedings of the 4th DARPA Speech and Natural Language Workshop*, pages 233–237, Harriman, New York, USA.
- García-Martínez, E., Creus, C., and España-Bonet, C. (2017). Using Word Embeddings to Enforce Document-Level Lexical Consistency in Machine Translation. *The Prague Bulletin of Mathematical Linguistics*, 108(1):85–96.
- Giles, H., Coupland, N., and Coupland, J. (1991). Accommodation theory: Communication, context, and consequences. In *Contexts of accommodation: Developments in applied sociolinguistics*. Cambridge University Press, Cambridge, UK.
- Goddard, A. and Meân Patterson, L. (2000). *Language and gender*. Routledge, London, UK.
- Gong, Z., Zhang, M., and Guodong, Z. (2011). Cache-based Document-level Statistical Machine Translation. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP'11)*, pages 909–919, Edinburgh, UK.

- Good, I. J. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika*, 40(3–4):237–264.
- Graff, D., Huang, S., Cartagena, I., Walker, K., and Cieri, C. (2010). *Fisher Spanish Speech LDC2010S01*. Linguistic Data Consortium, Philadelphia, Pennsylvania, USA.
- Graham, Y., Baldwin, T., Moffat, A., and Zobel, J. (2017). Can machine translation systems be evaluated by the crowd alone. *Natural Language Engineering*, 23(1):3–30.
- Graves, A. (2013). Generating Sequences With Recurrent Neural Networks. In *arXiv:1308.0850 [cs.NE]*.
- Gruet-Skrabalova, H. (2013). Verbs and particles in minimal answers to yes-no questions in Czech. In *Formal Description of Slavic Languages 10* (Slavic Grammar from a Formal Perspective, the 10th anniversary FDSL conference), pages 197–215, Leipzig, Germany.
- Gu, J., Hassan, H., Devlin, J., and Li, V. O. (2018). Universal Neural Machine Translation for Extremely Low Resource Languages. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics; Human Language Technologies (NAACL-HLT’18)*, pages 344–354, New Orleans, Louisiana, USA.
- Guillou, L. (2013). Analysing Lexical Consistency in Translation. In *Proceedings of the 1st Workshop on Discourse In Machine Translation (DISCOMT’13)*, pages 10–18, Sofia, Bulgaria.
- Guillou, L. (2016). *Incorporating Pronoun Function into Statistical Machine Translation*. PhD Thesis, School of Informatics. University of Edinburgh.
- Guillou, L. and Hardmeier, C. (2016). PROTEST: A Test Suite for Evaluating Pronouns in Machine Translation. In *Proceedings of the 10th Language Resources and Evaluation Conference (LREC’16)*, pages 636–643, Portorož, Slovenia.
- Guillou, L., Hardmeier, C., Nakov, P., Stymne, S., Tiedemann, J., Versley, Y., Cettolo, M., Webber, B., and Popescu-Belis, A. (2016). Findings of the 2016 WMT Shared Task on Cross-lingual Pronoun Prediction. In *Proceedings of the 1st Conference on Machine Translation (WMT’16)*, pages 525–542, Berlin, Germany.
- Guillou, L., Hardmeier, C., Smith, A., Tiedemann, J., and Webber, B. (2014). ParCor 1.0: A Parallel Pronoun-Coreference Corpus to Support Statistical MT. In *Proceedings of the 9th Language Resources and Evaluation Conference (LREC’14)*, pages 3191–3198, Reykjavik, Iceland.
- Gulcehre, C., Firat, O., Xu, K., Cho, K., Barrault, L., Lin, H.-C., Bougares, F., Schwenk, H., and Bengio, Y. (2015). On Using Monolingual Corpora in Neural Machine Translation. In *arXiv:1503.03535*.

- Guzmán, F., Abdelali, A., Temnikova, I., Sajjad, H., and Vogel, S. (2015). How do Humans Evaluate Machine Translation. In *Proceedings of the 10th Workshop on Statistical Machine Translation (WMT'15)*, pages 457–466, Lisbon, Portugal.
- Guzmán, F., Joty, S., Màrquez, L., and Nakov, P. (2014). Using Discourse Structure Improves Machine Translation Evaluation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL'14)*, pages 687–698, Baltimore, Maryland, USA.
- Hai Son, L., Allauzen, A., and Yvon, F. (2012). Continuous space translation models with neural networks. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT'12)*, pages 39–48, Montreal, Canada.
- Halliday, M. and Hasan, R. (1976). *Cohesion in English*. Number 9 in English Language Series. Longman, London, UK.
- Hardmeier, C. (2012). Discourse in statistical machine translation. a survey and a case study. *Discours*, 11.
- Hardmeier, C. (2016). Pronoun prediction with latent anaphora resolution. In *Proceedings of the 1st Conference on Machine Translation (WMT'16)*, Berlin, Germany.
- Hardmeier, C. (2017). Predicting pronouns with a convolutional network and an n-gram model. In *Proceedings of the 3rd Workshop on Discourse in Machine Translation (DISCOMT'17)*, pages 58–62, Copenhagen, Denmark.
- Hardmeier, C. and Federico, M. (2010). Modelling pronominal anaphora in statistical machine translation. In *Proceedings of the 7th International Workshop on Spoken Language Translation (IWSLT'10)*, pages 283–289, Paris, France.
- Hardmeier, C., Nakov, P., Stymne, S., Tiedemann, J., Versley, Y., and Cettolo, M. (2015). Pronoun-focused MT and cross-lingual pronoun prediction: Findings of the 2015 DiscoMT shared task on pronoun translation. In *Proceedings of the 2nd Workshop on Discourse in Machine Translation (DISCOMT'15)*, pages 1–16, Lisbon, Portugal.
- Hardmeier, C., Nivre, J., and Tiedemann, J. (2012). Document-Wide Decoding for Phrase-Based Statistical Machine Translation. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL'12)*, pages 1179–1190, Jeju Island, South Korea.
- Hardmeier, C., Saers, M., Federico, M., and Prashant, M. (2011). The Uppsala-FBK Systems at WMT 2011. In *Proceedings of the 6th Workshop on Statistical Machine Translation (WMT'11)*, pages 372–378, Edinburgh, UK.

- Hardmeier, C., Stymne, S., Tiedemann, J., and Nivre, J. (2013). Docent: A Document-Level Decoder for Phrase-Based Statistical Machine Translation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL'13)*, pages 193–198, Sofia, Bulgaria.
- Harris, Z. (1954). Distributional structure. *Word*, 10(23):146–162.
- Heafield, K., Pouzyrevsky, I., Clark, J. H., and Koehn, P. (2013). Scalable Modified Kneser-Ney Language Model Estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL'13)*, pages 690–696, Sofia, Bulgaria.
- Helcl, J. and Libovický, J. (2017). CUNI System for the WMT17 Multimodal Translation Task. In *Proceedings of the 2nd Conference on Machine Translation (WMT'17)*, pages 450–457, Copenhagen, Denmark.
- Higashinaka, R., Funakoshi, K., Kobayashi, Y., and Inaba, M. (2016). The Dialogue Breakdown Detection Challenge: Task Description, Datasets, and Evaluation Metrics. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC'16)*, pages 3146–3150, Portorož, Slovenia.
- Hitschler, J., Schamoni, S., and Riezler, S. (2016). Multimodal Pivots for Image Caption Translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL'16)*, pages 2399–2409, Berlin, Germany.
- Hochreiter, S. and Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780.
- Holmes, J. (1983). The Functions of Tag Questions. *English Language Research Journal*, 3:40–65.
- Holmes, J. (1984). Hedging your bets and sitting on the fence: Some evidence for hedges as support structures. *Te Reo*, 27:47–62.
- Holmes, J. (1995). *Women, Men and Politeness*. Longman, London, UK.
- Hovy, D. (2015). Demographic Factors Improve Classification Performance. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (ACL-IJCNLP'15)*, pages 752–762, Beijing, China.
- Huang, P.-Y., Liu, F., Shiang, S.-R., Oh, J., and Dyer, C. (2016). Attention-based Multimodal Neural Machine Translation. In *Proceedings of the 1st Conference on Machine Translation*, volume 2: of *WMT'16* (WMT'16), pages 639–645, Berlin, Germany.
- Hurford, J., Heasley, B., and Smith, M. (2007). *Semantics: A Coursebook*. Cambridge University Press, Cambridge, UK.

- Imankulova, A., Sato, T., and Komachi, M. (2017). Improving Low-Resource Neural Machine Translation with Filtered Pseudo-parallel Corpus. In *Proceedings of the 4th Workshop on Asian Translation (AFNLP'17)*, pages 70–78, Taipei, Taiwan.
- Isabelle, P., Cherry, C., and Foster, G. (2017). A Challenge Set Approach to Evaluating Machine Translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP'17)*, pages 2476–2486, Copenhagen, Denmark.
- Isabelle, P. and Kuhn, R. (2018). A Challenge Set for French –\textgreater English Machine Translation. *arXiv:1806.02725 [cs]*.
- Jean, S., Lauly, S., Firat, O., and Cho, K. (2017a). Does Neural Machine Translation Benefit from Larger Context? In *arXiv:1704.05135*.
- Jean, S., Lauly, S., Firat, O., and Cho, K. (2017b). Neural Machine Translation for Cross-Lingual Pronoun Prediction. In *Proceedings of the 3rd Workshop on Discourse in Machine Translation (DISCOMT'17)*, pages 54–57, Copenhagen, Denmark.
- Jehl, L. and Riezler, S. (2017). Document Information as Side Constraints for Improved Neural Patent Translation. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (AMTA'18)*, pages 1–12, Boston, Massachusetts, USA.
- Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F., Wattenberg, M., Corrado, G., Hughes, M., and Dean, J. (2017). Google’s Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.
- Junczys-Dowmunt, M., Grundkiewicz, R., Dwojak, T., Hoang, H., Heafield, K., Necker-mann, T., Seide, F., Germann, U., Aji, A. F., Bogoychev, N., Martins, A. F. T., and Birch, A. (2018). Marian: Fast Neural Machine Translation in C++. *arXiv:1804.00344 [cs]*.
- Kamp, H. and Reyle, U. (1993). *From Discourse to Logic: An Introduction to Modeltheoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory*. Kluwer, Dordrecht, Netherlands.
- King, M. and Falkedal, K. (1990). Using test suites in evaluation of machine translation systems. In *Proceedings of the 1990 Conference on Computational Linguistics (COLING'90)*, pages 211–216, Helsinki, Finland.
- Koehn, P. (2004a). Pharaoh: A Beam Search Decoder for Phrase-Based Statistical Machine Translation Models. In *Proceedings of the 6th Conference of the Association for Machine Translation in the Americas (AMTA'04)*, pages 115–124, Washington, District of Columbia, USA.

- Koehn, P. (2004b). Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP'14)*, pages 388–395, Barcelona, Spain.
- Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proceedings of the 10th Machine Translation Summit*, pages 79–86, Phuket, Thailand.
- Koehn, P. and Hoang, H. (2007). Factored translation models. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP'07)*, pages 868–876, Prague, Czech Republic.
- Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL'07)*, pages 177–180, Prague, Czech Republic.
- Koehn, P. and Knowles, R. (2017). Six Challenges for Neural Machine Translation. In *Proceedings of the 1st Workshop on Neural Machine Translation*, pages 28–39, Vancouver, Canada.
- Koehn, P., Och, F. J., and Marcu, D. (2003). Statistical Phrase-Based Translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT'03)*, pages 48–54, Edmonton, Canada.
- Kuang, S., Xiong, D., Luo, W., and Zhou, G. (2018). Modeling Coherence for Neural Machine Translation with Dynamic and Topic Caches. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING'18)*, Santa Fe, New Mexico, USA.
- Kuhn, R. and De Mori, R. (1990). Cache-based natural language model for speech recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(6):570–583.
- Kullback, S. and Leibler, R. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86.
- Lambert, P., Schwenk, H., Servan, C., and Abdul-Rauf, S. (2011). Investigations on Translation Model Adaptation Using Monolingual Data. In *Proceedings of the 6th Workshop on Statistical Machine Translation (WMT'11)*, pages 284–293, Edinburgh, UK.
- Lample, G., Conneau, A., Denoyer, L., and Ranzato, M. (2018). Unsupervised Machine Translation using monolingual corpora only. In *Proceedings of the 6th International Conference on Learning Representations (ICLR'18)*, Vancouver, Canada.

- Langford, J., Li, L., and Zhang, T. (2009). Sparse Online Learning via Truncated Gradient. *The Journal of Machine Learning Research*, pages 777–801.
- Langlais, P., Patry, A., and Gotti, F. (2007). A Greedy Decoder for Phrase-Based Statistical Machine Translation. In *Proceedings of the 11th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI'07)*, pages 104–113, Skövde, Sweden.
- Lavie, A. and Denkowski, M. J. (2009). The Meteor metric for automatic evaluation of machine translation. *Machine Translation*, 23(2-3):105–115.
- Le Cun, Y. (1988). A Theoretical Framework for Back-Propagation. In *Proceedings of the 1988 Connectionist Models Summer School*, pages 21–28, San Mateo, California, USA.
- Le Nagard, R. and Koehn, P. (2010). Aiding pronoun translation with co-reference resolution. In *Proceedings of the 5th Workshop on Statistical Machine Translation (WMT'10)*, pages 252–261, Uppsala, Sweden.
- Lee, J., Cho, K., and Hofmann, T. (2017). Fully Character-Level Neural Machine Translation without Explicit Segmentation. *Transactions of the Association for Computational Linguistics*, 5:365–378.
- Levy, O. and Goldberg, Y. (2014). Neural Word Embedding as Implicit Matrix Factorization. In *Proceedings of the 27th International Conference on Neural Information Processing Systems (NIPS'14)*, pages 2177–2185, Montreal, Canada.
- Li, J., Xiong, D., Tu, Z., Zhu, M., Zhang, M., and Zhou, G. (2017). Modeling Source Syntax for Neural Machine Translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL'17)*, pages 688–697, Vancouver, Canada.
- Li, Q., Wong, D. F., Chao, L. S., Zhu, M., Xiao, T., Zhu, J., and Zhang, M. (2018). Linguistic Knowledge-Aware Neural Machine Translation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(12):2341–2354.
- Libovický, J. and Helcl, J. (2017). Attention Strategies for Multi-Source Sequence-to-Sequence Learning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL'17)*, pages 196–202, Vancouver, Canada.
- Lin, C.-Y. and Och, F. J. (2004). Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL'04)*, pages 605–612, Barcelona, Spain.

- Lison, P. and Meena, R. (2016). Automatic turn segmentation for movie and TV subtitles. In *Proceedings of the 2016 Spoken Language Technology Workshop*, pages 245–252, San Diego, California, USA.
- Lison, P. and Tiedemann, J. (2016). OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles. In *Proceedings of the 10th Language Resources and Evaluation Conference (LREC'16)*, pages 923–929, Portorož, Slovenia.
- Loaiciga Sanchez, S. (2017). *Pronominal anaphora and verbal tenses in machine translation*. PhD Thesis, University of Geneva.
- Loáiciga, S., Stymne, S., Nakov, P., Hardmeier, C., Tiedemann, J., Cettolo, M., and Versley, Y. (2017). Findings of the 2017 DiscoMT Shared Task on Cross-lingual Pronoun Prediction. In *Proceedings of the 3rd Workshop on Discourse in Machine Translation (DISCOMT'17)*, pages 1–16, Copenhagen, Denmark.
- Lund, K., Burgess, C., and Atchley, R. A. (1995). Semantic and Associative Priming in a High-Dimensional Semantic Space. In *Proceedings of the 17th Annual Conference of the Cognitive Science Society*, volume 17, pages 660–665, Pittsburgh, Pennsylvania, USA.
- Luong, M.-T. and Manning, C. D. (2016). Achieving open vocabulary neural machine translation with hybrid word-character models. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL'16)*, pages 1054–1063, Berlin, Germany.
- Luong, N. Q. and Popescu-Belis, A. (2016). Pronoun language model and grammatical heuristics for aiding pronoun prediction. In *Proceedings of the 1st Conference on Machine Translation (WMT'16)*, pages 589–595, Berlin, Germany.
- Luotolahti, J., Kanerva, J., and Ginter, F. (2016). Cross-lingual pronoun prediction with deep recurrent neural networks. In *Proceedings of the 1st Conference on Machine Translation (WMT'16)*, pages 596–601, Berlin, Germany.
- Ma, Q., Bojar, O., and Graham, Y. (2018). Results of the WMT18 Metrics Shared Task: Both characters and embeddings achieve good performance. In *Proceedings of the 3rd Conference on Machine Translation (WMT'18)*, pages 671–688.
- Ma, X. (2006). Champollion: A robust parallel text sentence aligner. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'06)*, pages 489–492, Genoa, Italy.
- Malinowski, B. (1923). The problem of meaning in primitive languages. In Ogden, C. K. and Richards, I. A., editors, *The Meaning of Meaning*. Routledge & Kegan Paul, London, UK.

- Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., and McClosky, D. (2014). The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL'14)*, pages 55–60, Baltimore, Maryland, USA.
- Marcu, D. and Wong, W. (2002). A Phrase-Based, Joint Probability Model for Statistical Machine Translation. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP'02)*, pages 133–139, Philadelphia, Pennsylvania, USA.
- Marvin, R. and Koehn, P. (2018). Exploring Word Sense Disambiguation Abilities of Neural Machine Translation Systems. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (AMTA'18)*, pages 125–131, Boston, Massachusetts, USA.
- Mascarell, L. (2017). Lexical Chains meet Word Embeddings in Document-level Statistical Machine Translation. In *Proceedings of the 3rd Workshop on Discourse in Machine Translation (DISCOMT'17)*, pages 99–109, Copenhagen, Denmark.
- McGregor, W. (1995). The English ‘tag question’: A new analysis, is(n't) it? *On Subject and theme: A discourse functional perspective*, 118:91–121.
- Melamud, O., Goldberger, J., and Dagan, I. (2016). context2vec: Learning Generic Context Embedding with Bidirectional LSTM. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning (CoNLL'16)*, pages 51–61, Berlin, Germany.
- Mey, J. (1993). *Pragmatics: An Introduction*. Blackwell, Oxford, UK.
- Meyer, T. and Poláková, L. (2013). Machine Translation with Many Manually Labeled Discourse Connectives. In *Proceedings of the 9th Workshop on Statistical Machine Translation (WMT'13)*, pages 43–50, Sofia, Bulgaria.
- Meyer, T. and Popescu-Belis, A. (2012). Using sense-labeled discourse connectives for statistical machine translation. In *Proceedings of the Joint Workshop on Exploiting Synergies between Information Retrieval and Machine Translation and Hybrid Approaches to Machine Translation (ESIRMT-HyTra'12)*, pages 129–138, Avignon, France.
- Meyer, T., Popescu-Belis, A., and Hajlaoui, N. (2012). Machine Translation of Labeled Discourse Connectives. In *Proceedings of the 10th Biennial Conference of the Association for Machine Translation in the Americas (AMTA'12)*, pages 129–138, San Diego, California, USA.
- Meyer, T., Popescu-Belis, A., Zufferey, S., and Cartoni, B. (2011). Multilingual Annotation and Disambiguation of Discourse Connectives for Machine Translation. In *Proceedings of the 12th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 194–203, Portland, Oregon, USA.

- Meyer, T. and Webber, B. (2013). Implication of Discourse Connectives in (Machine) Translation. In *Proceedings of the 1st Workshop on Discourse in Machine Translation (DISCOMT'13)*, pages 19–26, Sofia, Bulgaria.
- Michel, P. and Neubig, G. (2018). Extreme Adaptation for Personalized Neural Machine Translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL'18)*, Melbourne, Australia.
- Mihalcea, R., Sinha, R., and McCarthy, D. (2010). SemEval-2010 Task 2: Cross-Lingual Lexical Substitution. In *Proceedings of the 5th International Workshop on Semantic Evaluation (SemEval'10)*, pages 9–14, Uppsala, Sweden.
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th Annual Conference on Neural Information Processing Systems (NIPS'13)*, pages 3111–3119, Lake Tahoe, Nevada, USA.
- Moore, R. C. and Lewis, W. (2010). Intelligent Selection of Language Model Training Data. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL'10)*, pages 220–224, Columbus, Ohio, USA.
- Morimoto, T., Uratani, N., Takezawa, T., Furuse, O., Sobashima, Y., Iida, H., Nakamura, A., Sagisaka, Y., Higuchi, N., and Yamazaki, Y. (1994). A Speech and Language Database for Speech Translation Research. In *Proceedings of the 3rd International Conference on Spoken Language Processing (ICSLP'94)*, pages 1791–1794, Yokohama, Japan.
- Neubig, G. and Watanabe, T. (2016). Optimization for Statistical Machine Translation: A Survey. *Computational Linguistics*, 42(1):1–54.
- Novák, M. (2016). Pronoun prediction with linguistic features and example weighing. In *Proceedings of the 1st Conference on Machine Translation (WMT'16)*, Berlin, Germany.
- Och, F. J. (2003). Minimum Error Rate Training in Statistical Machine Translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics (ACL'03)*, pages 160–167, Sapporo, Japan.
- Och, F. J. and Ney, H. (2002). Discriminative training and maximum entropy models for statistical machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL'02)*, pages 295–302, Philadelphia, Pennsylvania, USA.
- Och, F. J., Tillmann, C., and Ney, H. (1999). Improved Alignment Models for Statistical Machine Translation. In *Proceeding of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP-VLC'99)*, pages 20–28, College Park, Maryland, USA.

- Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL'02)*, pages 311–318, Philadelphia, Pennsylvania, USA.
- Passban, P., Liu, Q., and Way, A. (2017). Translating Low-Resource Languages by Vocabulary Adaptation from Close Counterparts. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 16(4).
- Pecina, P., Toral, A., and Van Genabith, J. (2012). Simple and Effective Parameter Tuning for Domain Adaptation of Statistical Machine Translation. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING'12)*, pages 2209–2224, Mumbai, India.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.
- Pelevina, M., Arefiev, N., Biemann, C., and Panchenko, A. (2016). Making Sense of Word Embeddings. In *Proceedings of the 1st Workshop on Representation Learning for NLP*, pages 174–183, Berlin, Germany.
- Pennington, J., Socher, R., and Manning, C. D. (2014). GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP'14)*, pages 1532–1543, Doha, Qatar.
- Peters, M. E., Ammar, W., Bhagavatula, C., and Power, R. (2017). Semi-supervised sequence tagging with bidirectional language models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL'17)*, pages 1756–1765, Vancouver, Canada.
- Plank, B., Søgaard, A., and Goldberg, Y. (2016). Multilingual Part-of-Speech Tagging with Bidirectional Long Short-Term Memory Models and Auxiliary Loss. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL'16)*, pages 412–418, Berlin, Germany.
- Pollack, J. (1990). Recursive Distributed Representations. *Artificial Intelligence*, 46(1):77–105.
- Popescu-Belis, A., Meyer, T., Liyanapathirana, J., Cartoni, B., and Zufferey, S. (2012). Discourse-level Annotation over Europarl for Machine Translation: Connectives and Pronouns. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'12)*, pages 2716–2720, Istanbul, Turkey.

- Popović, M. (2015). CHRF: Character n-gram F-Score for Automatic MT Evaluation. In *Proceedings of the 10th Workshop on Statistical Machine Translation (WMT'15)*, pages 392–395, Lisbon, Portugal.
- Post, M., Kumar, G., Lopez, A., Karakos, D., Callison-Burch, C., and Khudanpur, S. (2013). Improved Speech-to-Text Translation with the Fisher and Callhome Spanish-English Speech Translation Corpus. In *Proceedings of the 10th International Workshop on Spoken Language Translation (IWSLT'13)*, Heidelberg, Germany.
- Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A., and Webber, B. (2008). The Penn Discourse TreeBank 2.0. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC'08)*, pages 2961–2968, Marrakech, Morocco.
- Pu, X., Mascarell, L., and Popescu-Belis, A. (2017). Consistent Translation of Repeated Nouns using Syntactic and Semantic Cues. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL'17)*, pages 948–957, Valencia, Spain.
- Queneau, R. (1950). *Bâtons, Chiffres et Lettres*. Gallimard, Paris, France.
- Regier, T., Kay, P., and Khertarpal, N. (2007). Color naming reflects optimal partitions of color space. *Proceedings of the National Academy of Sciences of the United States of America*, 104(4):1436–1441.
- Rios Gonzales, A., Mascarell, L., and Sennrich, R. (2017). Improving Word Sense Disambiguation in Neural Machine Translation with Sense Embeddings. In *Proceedings of the 2nd Conference on Machine Translation (WMT'17)*, pages 11–19, Copenhagen, Denmark.
- Roukos, S., Graff, D., and Melamed, D. (1995). *Hansard French/English LDC95T20*. Linguistic Data Consortium, Philadelphia, Pennsylvania, USA.
- Roy, A., Guinaudeau, C., Bredin, H., and Barras, C. (2014). TVD: A Reproducible and Multiply Aligned TV Series Dataset. In *Proceedings of the 9th Language Resources and Evaluation Conference (LREC'14)*, pages 418–425, Reykjavik, Iceland.
- Roze, C., Danlos, L., and Muller, P. (2012). LEXCONN: A French Lexicon of Discourse Connectives. *Discours*, 10(1).
- Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning internal representations by error propagation. *Nature*, 323:533–536.
- Sagot, B. (2010). The *Lefff*, a Freely Available and Large-coverage Morphological and Syntactic Lexicon for French. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC'10)*, pages 2744–2751, Valletta, Malta.

- Sagot, B. and Boullier, P. (2008). SxPipe 2: architecture pour le traitement pré-syntaxique de corpus bruts. *Traitement Automatique des Langues*, 49(2):155–188.
- Saluja, A., Lane, I., and Zhang, Y. (2011). Context-aware Language Modeling for Conversational Speech Translation. In *Proceedings of the 13th Machine Translation Summit*, pages 97–104, Xiamen, China.
- Schwenk, H. (2007). Continuous space language models. *Computer Speech and Language*, 21(3):492–518.
- Schwenk, H. (2008). Investigations on Large-Scale Lightly-Supervised Training for Statistical Machine Translation. In *Proceedings of the 5th International Workshop on Spoken Language Translation (IWSLT’08)*, pages 182–189, Honolulu, Hawaii, USA.
- Schwenk, H., Dchelotte, D., and Gauvain, J.-L. (2006). Continuous-space language models for statistical machine translation. In *Proceedings of the Joint 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING-ACL’06)*, pages 723–730, Sydney, Australia.
- Seddah, D., Kübler, S., and Tsarfaty, R. (2014). Introducing the SPMRL 2014 shared task on parsing morphologically-rich languages. In *Proceedings of the 1st Joint Workshop on Statistical Parsing of Morphologically Rich Languages and Syntactic Analysis of Non-Canonical Languages (SPMRL-SANCL’14)*, pages 103–109, Dublin, Ireland.
- Sennrich, R. (2017). How Grammatical is Character-level Neural Machine Translation? In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL’17)*, pages 376–382, Valencia, Spain.
- Sennrich, R., Firat, O., Cho, K., Birch, A., Haddow, B., Hitschler, J., Junczys-Dowmunt, M., Läubli, S., Valerio, A., Barone, M., Mokry, J., and Nădejde, M. (2017). Nematus: a Toolkit for Neural Machine Translation. In *Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL’17)*, pages 65–68, Valencia, Spain.
- Sennrich, R. and Haddow, B. (2016). Linguistic Input Features Improve Neural Machine Translation. In *Proceedings of the 1st Conference on Machine Translation (WMT’16)*, pages 83–91, Berlin, Germany.
- Sennrich, R., Haddow, B., and Birch, A. (2016a). Controlling Politeness in Neural Machine Translation via Side Constraints. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT’16)*, pages 35–40, San Diego, California, USA.
- Sennrich, R., Haddow, B., and Birch, A. (2016b). Edinburgh Neural Machine Translation Systems for WMT 16. In *Proceedings of the 1st Conference on Machine Translation (WMT’16)*, pages 368–373, Berlin, Germany.

- Sennrich, R., Haddow, B., and Birch, A. (2016c). Improving Neural Machine Translation Models with Monolingual Data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL'16)*, pages 86–96, Berlin, Germany.
- Sennrich, R., Haddow, B., and Birch, A. (2016d). Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL'16)*, pages 1715–1725, Berlin, Germany.
- Serban, I., Sordoni, A., Bengio, Y., Courville, A., and Pineau, J. (2016). Building end-to-end dialogue systems using generative hierarchical neural network models. In *Proceedings of the 30th AAAI Conference on Artificial Intelligence (AAAI'16)*, pages 3776–3783, Phoenix, Arizona, USA.
- Shannon, C. (1948). A Mathematical Theory of Communication. *The Bell System Technical Journal*, 27(3):379–423.
- Sim Smith, K. (2017). *Coherence in Machine Translation*. PhD Thesis, University of Sheffield.
- Singla, K., Can, D., and Narayanan, S. (2018). A Multi-task Approach to Learning Multilingual Representations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL'18)*, pages 214–220, Melbourne, Australia.
- Smith, R. (2007). An Overview of the Tesseract OCR Engine. In *Proceedings of the 9th International Conference on Document Analysis and Recognition (ICDAR'07)*, pages 629–633, Washington, District of Columbia, USA.
- Soars, J. and Soars, L. (2000). *New headway English course. Pre-intermediate student's book*. Oxford University Press, Oxford, UK, 2nd edition.
- Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C., Ng, A., and Potts, C. (2013). Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP'13)*, pages 1631–1642, Seattle, Washington, USA.
- Specia, L., Frank, S., Sima'an, K., and Elliott, D. (2016). A shared task on multi-modal machine translation and crosslingual image description. In *Proceedings of the 1st Conference on Machine Translation (WMT'16)*, pages 543–553, Berlin, Germany.
- Specia, L., Sankaran, B., and das Graças Volpe Nunes, M. (2008). n-Best Reranking for the Efficient Integration of Word Sense Disambiguation and Statistical Machine Translation. In *Proceedings of the 9th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing'08)*, pages 339–410, Haifa, Israel.

- Specia, L., Shah, K., de Souza, J. G. C., and Cohn, T. (2013). QuEst - A translation quality estimation framework. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL'13)*, pages 79–84, Sofia, Bulgaria.
- Steele, D. and Specia, L. (2016). Predicting and Using Implicit Discourse Elements in Chinese-English Translation. *Baltic Journal of Modern Computing*, 4(2):305–317.
- Stymne, S. (2016). Feature exploration for cross-lingual pronoun prediction. In *Proceedings of the 1st Conference on Machine Translation (WMT'16)*, pages 609–615, Berlin, Germany.
- Stymne, S., Loáiciga, S., and Cap, F. (2017). A BiLSTM-based system for cross-lingual pronoun prediction. In *Proceedings of the 3rd Workshop on Discourse in Machine Translation (DISCOMT'17)*, pages 47–53, Copenhagen, Denmark.
- Subramanian, S., Trischler, A., Bengio, Y., and Pal, C. J. (2018). Learning General Purpose Distributed Sentence Representations via Large Scale Multi-task Learning. In *Proceedings of the 6th International Conference on Learning Representations (ICLR'18)*, Vancouver, Canada.
- Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to Sequence Learning with Neural Networks. In *Proceedings of the 27th Annual Conference on Neural Information Processing Systems (NIPS'14)*, pages 3104–3112, Montreal, Canada.
- Takeo, S., Nagata, M., and Yamamoto, K. (2017). Controlling Target Features in Neural Machine Translation via Prefix Constraints. In *Proceedings of the 4th Workshop on Asian Translation (AFNLP'17)*, pages 55–63, Taipei, Taiwan.
- Takezawa, T. and Kikui, G. (2003). Collecting Machine-Translation-Aided Bilingual Dialogues for Corpus-Based Speech Translation. In *Proceedings of the 8th European Conference on Speech Communication and Technology (ESPEECH'03)*, pages 2757–2760, Geneva, Switzerland.
- Takezawa, T., Kikui, G., Mizushima, M., and Sumita, E. (2007). Multilingual Spoken Language Corpus Development for Communication Research. *Computational Linguistics and Chinese Language Processing*, 12(3):303–324.
- Tiedemann, J. (2010). Context Adaptation in Statistical Machine Translation Using Models with Exponentially Decaying Cache. In *Proceedings of the 2010 Workshop on Domain Adaptation for Natural Language Processing*, pages 8–15, Uppsala, Sweden.
- Tiedemann, J. (2012). Parallel Data, Tools and Interfaces in OPUS. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey.

- Tiedemann, J. (2016). A linear baseline classifier for cross-lingual pronoun prediction. In *Proceedings of the 1st Conference on Machine Translation (WMT'16)*, pages 616–619, Berlin, Germany.
- Tiedemann, J. and Scherrer, Y. (2017). Neural Machine Translation with Extended Context. In *Proceedings of the 3rd Workshop on Discourse in Machine Translation (DISCOMT'17)*, pages 82–92, Copenhagen, Denmark.
- Tillmann, C. (2001). *Word Re-Ordering and Dynamic Programming based Search Algorithm for Statistical Machine Translation*. PhD Thesis, RWTH Aachen University, Aachen, Germany.
- Tottie, G. and Hoffmann, S. (2006). Tag questions in British and American English. *Journal of English Linguistics*, 34(4):283–311.
- Troyanskii, P. P. (1935). *Mashina dlya podbora i pechataniya slov pri perevode s odnogo yazyka na drugoy ili na neskol'kikh yazykakh odnovremenno (Machine for selecting and typing words when translating from one language to another or across multiple languages simultaneously)*.
- Tu, Z., Liu, Y., Shang, L., Liu, X., and Li, H. (2017). Neural Machine Translation with Reconstruction. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence (AAAI'17)*, pages 3097–3103, San Francisco, California, USA.
- Turchi, M., De Bie, T., and Cristianini, N. (2008). Learning performance of a machine translation system: a statistical and computational analysis. In *Proceedings of the 3rd Workshop on Statistical Machine Translation*, pages 35–43, Columbus, Ohio, USA.
- Ture, F., Oard, D. W., and Resnik, P. (2012). Encouraging Consistent Translation Choices. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT'12)*, pages 417–426, Montreal, Canada.
- Turian, J., Shen, L., and Melamed, I. D. (2003). Evaluation of Machine Translation and its Evaluation. In *Proceedings of the 9th Machine Translation Summit*, pages 386–393, New Orleans, Louisiana, USA.
- van der Wees, M., Bisazza, A., and Monz, C. (2016). Measuring the Effect of Conversational Aspects on Machine Translation Quality. In *Proceedings of the 26th International Conference on Computational Linguistics (COLING'16)*, pages 2571–2581, Osaka, Japan.
- Vania, C., Grivas, A., and Lopez, A. (2018). What do character-level models learn about morphology? The case of dependency parsing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP'18)*, pages 2573–2583, Brussels, Belgium.

- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is All you Need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.
- Vaswani, A., Zhao, Y., and Fossum, V. (2013). Decoding with Large-Scale Neural Language Models Improves Translation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP'13)*, pages 1387–1392, Seattle, Washington, USA.
- Vauquois, B., Veillon, G., and Veyrunes, J. (1965). Application des grammaires formelles aux modèles linguistiques en traduction automatique. *Kybernetika*, 1(3):281–289.
- Vendryès, J. (1921). *Le Langage, Introduction linguistique à l'histoire*. La Renaissance du livre, Paris, France.
- Vickrey, D., Biewald, L., Teyssier, M., and Koller, D. (2005). Word-Sense Disambiguation for Machine Translation. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT-EMNLP'05)*, pages 771–778, Vancouver, Canada.
- Voita, E., Serdyukov, P., Sennrich, R., and Titov, I. (2018). Context-Aware Neural Machine Translation Learns Anaphora Resolution. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL'18)*, pages 1264–1274, Melbourne, Australia.
- Wang, L., Tu, Z., Way, A., and Qun Liu (2017). Exploiting Cross-Sentence Context for Neural Machine Translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP'17)*, pages 2816–2821, Copenhagen, Denmark.
- Wang, L., Zhang, X., Tuy, Z., Way, A., and Liu, Q. (2016a). Automatic construction of discourse corpora for dialogue translation. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC'16)*, pages 2748–2754, Portorož, Slovenia.
- Wang, T. and Cho, K. (2016). Larger-Context Language Modelling with Recurrent Neural Network. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL'16)*, pages 1319–1329, Berlin, Germany.
- Wang, W., Peter, J.-T., Rosendahl, H., and Ney, H. (2016b). CharacTER: Translation Edit Rate on Character Level. In *Proceedings of the 1st Conference on Machine Translation (WMT'16)*, pages 505–510, Berlin, Germany.
- Weaver, W. (1955). Translation. In *Machine Translation of Languages*. MIT Press.

- White, J. S., O'Connell, T., and O'Mara, F. (1994). The ARPA MT Evaluation Methodologies: Evolution, Lessons and Future Approaches. In *Proceedings of the 10th Biennial Conference of the Association for Machine Translation in the Americas (AMTA'94)*, pages 193–205.
- Widdowson, H. G. (2004). *Text, Context, Pretext: Critical Issues in Discourse Analysis*. Number 35 in *Language in society*. Blackwell, Oxford, UK.
- Wisniewski, G., Allauzen, A., and Yvon, F. (2010). Assessing Phrase-Based Translation Models with Oracle Decoding. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP'10)*, pages 933–943, Cambridge, Massachusetts, USA.
- Wittgenstein, L. (1953). *Philosophical Investigations*. Blackwell, Oxford, UK.
- Wong, B. T. M. and Kit, C. (2012). Extending Machine Translation Evaluation Metrics with Lexical Cohesion To Document Level. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL'12)*, pages 1060–1068, Jeju Island, South Korea.
- Xiao, T., Zhu, J., Yao, S., and Zhang, H. (2011). Document-level consistency verification in machine translation. In *Proceedings of the 13th Machine Translation Summit*, pages 131–138, Xiamen, China.
- Xiong, D., Ding, Y., Zhang, M., and Lim Tan, C. (2013). Lexical Chain Based Cohesion Models for Document-Level Statistical Machine Translation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP'13)*, pages 1563–1573, Seattle, Washington, USA.
- Yamada, K. and Knight, K. (2001). A Syntax-based Statistical Translation Model. In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics (ACL'01)*, pages 523–530, Toulouse, France.
- Zens, R., Och, F. J., and Ney, H. (2002). Phrase-based statistical machine translation. In *Proceedings of the 25th German Conference on Artificial Intelligence (KI'02)*, pages 18–32, Aachen, Germany.
- Zhang, J. and Zong, C. (2016). Exploiting source-side monolingual data in neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP'16)*, pages 1535–1545, Austin, Texas, USA.
- Zoph, B. and Knight, K. (2016). Multi-source Neural Translation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT'16)*, pages 30–34, San Diego, California, USA.

- Zoph, B., Yuret, D., May, J., and Knight, K. (2016). Transfer Learning for Low-Resource Neural Machine Translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP'16)*, pages 1568–1575, Austin, Texas, USA.
- Zufferey, S. (2016). Discourse connectives across languages: Factors influencing their explicit or implicit translation. *Languages in Contrast*, 16(2):264–279.

Résumé détaillé

La traduction automatique (TA) est de plus en plus utilisée pour traduire des dialogues informels écrits. Les outils de traduction apparaissent régulièrement sur des applications de tchat et sur les réseaux sociaux pour permettre des communications cross-lingues. Pour cela, il est nécessaire que les systèmes de TA soient adaptés au type de langage utilisé dans ces contextes écrits mais informels, qui sont associés à une grande variété de thèmes de discussion, de styles et de vocabulaires. Plus important encore, la traduction des conversations nécessite de traduire des phrases de façon cohérente par rapport au flux conversationnel, pour que tous les aspects de l'échange, y compris l'intention du locuteur, l'attitude et le style soient correctement restitués.

Malgré les progrès importants dont ont bénéficié les techniques de TA au fil des années, un certain nombre de phénomènes sont encore difficiles voire impossibles à traduire par les systèmes de TA standards. Une des approximations les plus remarquables faites encore aujourd'hui par la majorité des systèmes de TA consiste à traduire les phrases au sein d'un même document indépendamment les unes des autres. Il y a deux raisons principales pour cette approximation: (i) traduire de longues séquences de texte présente des difficultés computationnelles (les approches de TA guidées par les données requièrent la segmentation des textes en unités de traduction plus petites), et (ii) une majorité de phrases n'ont pas besoin de contexte extra-phrastique pour être correctement traduites. Jusqu'à récemment, il était souvent plus facile d'améliorer la qualité de la TA en se concentrant sur une meilleure modélisation du contexte local à l'intérieur de chaque phrase.

Cependant, l'amélioration de la qualité des systèmes de TA, en particulier pour des paires de langues bien dotées, rend encore plus évidentes les insuffisances des systèmes de TA qui traduisent au niveau de la phrase. De multiples phénomènes nécessitent de prendre en compte le contexte extra-phrastique pour être correctement traduits et ne peuvent être traduits avec les systèmes de TA standards (Hardmeier, 2012). Il en est ainsi de la *traduction des pronoms anaphoriques* (Guillou, 2016), où la forme traduite d'un pronom est directement dépendante du genre grammatical d'un autre élément textuel (voir l'exemple (59)), de la *désambiguïisation lexicale* (Carpuat and Wu, 2005; Vickrey et al., 2005), où la traduction d'un élément ambigu nécessite des informations contextuelles pour qu'il soit désambiguïé (voir l'exemple (60)), et de la *cohésion lexicale* (Xiong et al., 2013; Guillou, 2013), où le choix d'un mot dans traduction dépend de la forme d'un autre élément traduit, par exemple quand il doit être identique à un autre mot, comme dans l'exemple (61). Dans chacun de ces trois exemples, la traduction correcte de chaque élément dépendant du contexte (en gras) dépend du *contexte linguistique* (souligné) qui apparaît en dehors de la phrase courante. Elle reste donc inaccessible par les systèmes de TA standards qui opèrent au niveau de la phrase.

(59) *Traduction de pronoms anaphoriques* (traduction de *it*):

EN: She sat on the chair. But **it** broke.

FR: Elle s'est assise sur la chaise. Mais **elle** s'est cassée.

(60) *Désambiguïisation lexicale* (traduction de *legs*):

EN: The chair was in pieces. One of its **legs** had collapsed.

FR: La chaise était en morceaux. L'un de ses **pieds** avait lâché.

FR: #La chaise était en morceaux. L'une de ses **jambes** avait lâché.¹

(61) *Cohésion lexicale (répétition)* (traduction de *tired*):

EN: Goldilocks was tired. Very **tired**.

FR: Boucle d'or était fatiguée. Très **fatiguée**.

FR: #Boucle d'or était fatiguée. Très **épuisée**.

Les informations contextuelles qui déterminent la façon dont une phrase doit être traduite ne se limitent pas au contexte linguistique (les mots du texte, y compris les phrases précédentes). Elles peuvent également relever de la situation dans laquelle le texte est produit (par exemple les informations concernant les locuteurs, la relation entre eux, le thème de discussion, etc.) et peut ainsi ne pas apparaître du tout dans le texte lui-même. Le genre du locuteur est un exemple d'un tel contexte *extra-linguistique*. Dans certaines langues comme le français, certains mots (par exemple les adjectifs et les participes passés) s'accordent ainsi en genre avec le sujet qu'ils qualifient. Lorsque l'on traduit vers ces langues à partir d'une langue pour laquelle ceci n'est pas le cas (comme l'anglais) et lorsque le sujet est le locuteur, le genre du locuteur détermine la traduction correcte. Ceci est illustré par l'exemple (62).

(62) *Genre du locuteur*:

EN: I am so glad that I can lie down on this comfy bed.

FR_{FEM}: Je suis bien **contente**_{FEM} de pouvoir m'allonger dans ce lit douillet.

FR_{MASC}: Je suis bien **content**_{MASC} de pouvoir m'allonger dans ce lit douillet.

Dans cette thèse, notre objectif est de dépasser cette approximation faite par les systèmes de TA standards en étudiant et en proposant différentes stratégies pour intégrer dans le processus de traduction des informations dont l'origine est au-delà de la phrase courante. Nous qualifions de telles approches de *TA contextuelle*, et nous appelons *contexte* des informations dont l'origine se situe à l'extérieur du segment textuel à traduire et qui peuvent être utiles pour produire une traduction correcte. Nous étudions l'intégration du *contexte linguistique* et du *contexte extra-linguistique* en nous servant de différentes stratégies. Nous prêtons une attention particulière à l'évaluation des modèles de TA quant à leur capacité à exploiter ces informations contextuelles.

La thèse est structurée en deux parties. La première présente la TA contextuelle et les travaux antérieurs dans ce domaine. La deuxième est dédiée à nos propres contributions.

¹Nous indiquons les traductions contextuellement incorrectes avec le caractère #.

La première partie répond aux questions suivantes: qu'est-ce que la TA contextuelle, pourquoi est-elle importante et comment certains travaux précédents ont essayé de l'aborder. Nous présentons les notions sur lesquelles repose cette thèse, nous formalisons la problématique d'un point de vue théorique et établissons les fondements de nos propres contributions à la TA contextuelle. Nous définissons et illustrons ainsi trois notions cruciales, celles de *traduction*, d'*ambiguïté* et de *contexte*. Nous distinguons trois types d'ambiguïté qui peuvent apparaître pendant le processus de traduction analysé sous un angle théorique: (i) l'*ambiguïté de la langue source*, (ii) l'*ambiguïté cross-lingue* et (iii) l'*ambiguïté de la langue cible*, en fonction du point d'apparition de l'ambiguïté dans le processus de traduction. Nous fournissons des exemples de chaque type d'ambiguïté et discutons de la nature du contexte qui peut être utilisé pour les résoudre.

Dans la deuxième partie, nous présentons nos contributions à la modélisation du contexte, aux stratégies pour intégrer le contexte dans la TA et à l'évaluation de modèles de TA contextuels. La présentation de ces contributions suit un ordre chronologique, représentant les évolutions dans les techniques utilisées, suivant l'évolution du domaine, notamment par rapport à l'apparition de modèles neuronaux au détriment des modèles statistiques qui avait précédemment dominé l'état de l'art. Tandis que nos premières contributions suivent une méthodologie qui permet l'exploitation du contexte d'avoir lieu soit en amont du processus de traduction (dans une phase de pré-traitement) soit en aval du processus de traduction (dans une phase de post-édition), nos expériences plus récentes s'attaquent à la problématique de l'entraînement de modèles de traduction capables d'apprendre directement comment utiliser le contexte. Nous comparons plusieurs stratégies différentes, notamment avec des architectures neuronales multi-encodeurs, et proposons également de nouvelles architectures, que nous appliquons tant à l'utilisation de contexte linguistique qu'à l'exploitation du contexte extra-linguistique.

Tout au long de la thèse, nous adoptons un œil critique par rapport à l'évaluation, en mettant en question les métriques automatiques classiques et en mettant en avant des stratégies plus ciblées. Par exemple, nous discutons de la problématique de l'utilisation des métriques standards pour l'évaluation de l'utilisation d'éléments stylistiques (application à la génération de "tag questions" en anglais). Nous décrivons également la construction de jeux de tests contrastifs pour évaluer la capacité des modèles contextuels à exploiter le contexte linguistique pour traduire correctement certains éléments discursifs. Enfin, nous montrons comment nous avons construit un corpus de test formé de dialogues spontanés entre locuteurs anglais et français médiés par la TA, dans le but d'évaluer nos modèles contextuels dans un cadre écologique mais également de fournir une ressource utile pour évaluer de futurs modèles et mieux analyser les caractéristiques linguistiques des énoncés produits dans de telles circonstances.

Titre: Au-delà de la phrase : traduction automatique de dialogue en contexte

Mots clés: traduction automatique; contexte; dialogue; discours

Résumé:

Les systèmes de traduction automatique (TA) ont fait des progrès considérables ces dernières années. La majorité d'entre eux reposent pourtant sur l'hypothèse que les phrases peuvent être traduites indépendamment les unes des autres. Ces modèles de traduction ne s'appuient que sur les informations contenues dans la phrase à traduire. Ils n'ont accès ni aux informations présentes dans les phrases environnantes ni aux informations que pourrait fournir le contexte dans lequel ces phrases ont été produites.

La *TA contextuelle* a pour objectif de dépasser cette limitation en explorant différentes méthodes d'intégration du contexte extra-phrastique dans le processus de traduction. Les phrases environnantes (*contexte linguistique*) et le contexte de production des énoncés (*contexte extra-linguistique*) peuvent fournir des informations cruciales pour la traduction, notamment pour la prise en compte des phénomènes discursifs et des mécanismes référentiels.

La prise en compte du contexte est toutefois un défi pour la traduction automatique. Évaluer la capacité de telles stratégie à prendre réellement en compte le contexte et à améliorer ainsi la qualité de la traduction est également un problème délicat, les métriques

d'évaluation usuelles étant pour cela inadaptées voire trompeuses.

Dans cette thèse, nous proposons plusieurs stratégies pour intégrer le contexte, tant linguistique qu'extra-linguistique, dans le processus de traduction. Nos expériences s'appuient sur des méthodes d'évaluation et des jeux de données que nous avons développés spécifiquement à cette fin. Nous explorons différents types de stratégies: les stratégies par *pré-traitement*, où l'on utilise le contexte pour désambiguïser les données fournies en entrée aux modèles; les stratégies par *post-traitement*, où l'on utilise le contexte pour modifier la sortie d'un modèle non-contextuel, et les stratégies où l'on exploite le contexte *pendant la traduction* proprement dite. Nous nous penchons sur de multiples phénomènes contextuels, et notamment sur la traduction des pronoms anaphoriques, la désambiguïstation lexicale, la cohésion lexicale et l'adaptation à des informations extra-linguistiques telles que l'âge ou le genre du locuteur. Nos expériences, qui relèvent pour certaines de la TA statistique et pour d'autres de la TA neuronale, concernent principalement la traduction de l'anglais vers le français, avec un intérêt particulier pour la traduction de dialogues spontanés.

Title: Going beyond the sentence: Contextual Machine Translation of Dialogue

Keywords: machine translation, context, dialogue, discourse

Abstract:

While huge progress has been made in machine translation (MT) in recent years, the majority of MT systems still rely on the assumption that sentences can be translated in isolation. The result is that these MT models only have access to context within the current sentence; context from other sentences in the same text and information relevant to the scenario in which they are produced remain out of reach.

The aim of *contextual MT* is to overcome this limitation by providing ways of integrating extra-sentential context into the translation process. Context, concerning the other sentences in the text (*linguistic context*) and the scenario in which the text is produced (*extra-linguistic context*), is important for a variety of cases, such as discourse-level and other referential phenomena.

Successfully taking context into account in translation is challenging. Evaluating such strategies on their capacity to exploit context is also a challenge, standard evaluation metrics being inadequate and

even misleading when it comes to assessing such improvement in contextual MT.

In this thesis, we propose a range of strategies to integrate both extra-linguistic context and linguistic context into the translation process. We accompany our experiments with specifically designed evaluation methods, including new test sets and corpora. Our contextual strategies include *pre-processing* strategies designed to disambiguate the data on which MT models are trained, *post-processing* strategies to integrate context by post-editing MT outputs and strategies in which context is exploited *during translation* proper. We cover a range of different context-dependent phenomena, including anaphoric pronoun translation, lexical disambiguation, lexical cohesion and adaptation to properties of the scenario such as speaker gender and age. Our experiments for both phrase-based statistical MT and neural MT are applied in particular to the translation of English to French and focus specifically on the translation of informal written dialogues.

