Towards a Better Human-Machine Collaboration in Statistical Translation : Example of Systematic Medical Reviews

Julia Ive

NNT: 2017SACLS225

Towards a Better Human-Machine Collaboration in Statistical Translation: Example of Systematic Medical Reviews

Thèse de doctorat de l'Université Paris-Saclay
préparée à l'Université Paris-Sud

École 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 1er septembre 2017, par

Julia Ive

Composition du Jury :

M. Nicolas Sabouret Professeur, Université Paris Sud	Président
Mme Pierrette Bouillon Professeure, Université de Genève	Rapporteur
M. Marco Turchi Chercheur, Fondazione Bruno Kessler	Rapporteur
M. Emmanuel Planas Maître de conférences, Université Catholique de l'Ouest	Examinateur
M. Philippe Ravaud Professeur, Université Descartes	Examinateur
M. François Yvon Professeur, Université Paris Sud	Directeur de thèse
M. Aurélien Max Maître de conférences, Université Paris Sud	Co-encadrant de thèse

Thèse de doctorat

Acknowledgments

I would like first of all to thank my both thesis advisers François Yvon and Aurélien Max for their help and support during all those years. François was a rigorous reviewer of all my scientific writing, which helped me to master the style. I am personally grateful to him for that. My enriching discussions with him helped me to gain a broad overview of the domain and its methods. François always managed to find a minute for me in spite of his busy schedule, always having a solution to any problem. Aurélien was always there for inspiring discussions and new creative ideas, as well for support in all the circumstances. He was an insatiable source of enthusiasm and inspiration.

I would also like to thank Philippe Ravaud for welcoming me at Cochrane France. Without his financial support this thesis would not have been possible.

I sincerely thank the members of my thesis jury. I am very grateful to my both reviewers Pierrette Bouillon and Marco Turchi for their insightful comments and suggestions. Marco was particularly patient to annotate in details my thesis with a pen. I also thank the other members of the jury Emmanuel Planas, Philippe Ravaud and Nicolas Sabouret, who also presided the jury, for having carefully read my thesis and providing valuable comments.

I would like to thank all my wonderful colleagues at both LIMSI-CNRS and Cochrane France for creating a productive and friendly working environment. My special thanks are to Hanna, who shared the office with me in Cochrane, for being particularly patient about technical server issues. It was my great pleasure to work with her. I also thank my colleagues at LIMSI Cyril, Thierry, Elena, Rachel, and later Charlotte for being excellent office mates. I also want to thank, among others, Franck, Lauriane, Mathieu and Thomas for always being there to help.

Finally, I thank my tiny dogs, Mars, Poupshik and Hermès, for never asking more than I could give them during the busy years and creating a cosy, lively and loving atmosphere at home. I would also like to thank my mother for never trying to stop me, my friend Kim for encouraging me to contact François for the first time and my friend Vincent for organizing my defense party in his own way.

Contents

Abbreviations

PBSMT Phrase-Based Statistical Machine Translation

MT-SEG Machine-Oriented Segmentation

SYNT-SEG Syntactically-Motivated Segmentation

WORD-SEG Word-Level Segmentation

`DT` Difficult-to-Translate

`ET` Easy-to-Translate

`kb-mira` K-Best Margin-Infused Relaxed Algorithm

`mert` Minimum Error Rate Training

ABS Scientific Abstract of a Cochrane Review Abstract

ADJ Adjective

ADJP Adjective Phrase

ADV Adverb

ADVP Adverb Phrase

AL Active Learning

amod Adjectival Modifier

APE Automatic Post-Editing

API Application Programming Interface

AR Arabic

BLEU BiLingual Evaluation Understudy

case Case Marking

CAT Computer-Assisted Translation

CC Coordination Conjuction

CE Confidence Estimation

CL Controlled Language

CONJ Conjuction

CRF Conditional Random Fields

D Deletion (Post-edit operation)

DA Domain Adaptation

DET Determiner (for French)

det Determiner (Universal Dependency)

DT Determiner (for English)

EBMT Example-Based Machine Translation

EM Expectation-Maximization (Algorithm)

EN English

ES Spanish

FFNN Feedforward Neural Networks

FR French

GWT Google Web Toolkit

HAMT Human-Assisted Machine Translation

HMM Hidden Markov Model

HTER Human-Targeted Translation Edit Rate

HUME Human UCCA-based Machine Translation Evaluation

I Insertion (Post-edit operation)

IMT Interactive Machine Translation

IN Preposition or Subordination Conjunction

JJ Adjective

LM Language Model

M Match (between Machine Translation and Reference)

MDA Multi-Domain Adaptation

METEOR Metric for Evaluation of Translation with Explicit ORdering

MIRA Margin-Infused Relaxed Algorithm

MQM Multidimensional Quality Metrics

MT Machine Translation

MTL Multi-Task Learning

N Noun

NC Common Noun

NLP Natural Language Processing

nmod Nominal Modifier

NMT Neural Machine Translation

NN Noun, Singular or Mass

NNS Noun, Plural

NP Noun Phrase

O Outside (Not belonging to any chunk)

OL Online Learning

OOV Out-of-Vocabulary

P Preposition

PAR Paraphrase Match (between Machine Translation and Reference)

PE Post-Edition

PLS Plain Language Summary of a Cochrane Review Abstract

POS Part-of-Speech

PP Prepositional Phrase

PRE Pre-Edition

PREP Preposition

PRIMT Pick-Revise Interactive Machine Translation

PRP Personal Pronoun

PT Phrase-Table

PUNC Punctiation

QE Quality Estimation

RF Random Forests

RM Reordering Model

RU Russian

S Substitution (Post-edit operation)

SA Suffix Array

SBAR Clause introduced by a Subordinating Conjunction

SH Shift (Post-edit operation)

SMT Statistical Machine Translation

SVM Support Vector Machines

T Stem Match (between Machine Translation and Reference)

TER Translation Edit Rate

TM Translation Model

UCCA Universal Conceptual Cognitive Annotation

UN United Nations

V Verb

VBD Verb, Past Tense

VBG Verb, Gerund or Present Participle

VBN Verb, Past Participle

VBZ Verb, 3rd Person Singular Present

VP Verb Phrase

VPP Verb Past Participle

WMT Workshop on Machine Translation

Y Synonym Match (between Machine Translation and Reference)

1 Introduction

1.1 Human-Machine Collaboration in Machine Translation

Translation plays a major role in the intercultural communication. Traditionally, it involves complex and demanding cognitive activities performed by bilingual human specialists. Numerous technological advances seek to reduce the cost of this process and make it more accessible to a wider range of contexts by continuously improving **Machine Translation** (MT).

Today MT has progressed enough to be efficiently used in many situations, including professional translation environments and production scenarios. These past few years research in MT is being switched from the **Statistical Machine Translation** (SMT) approach [Brown et al., 1990; Koehn et al., 2003] to the **Neural Machine Translation** (NMT) approach [Sutskever et al., 2014; Bahdanau et al., 2015]. The latter has received most of the attention due to the significant improvement in translation quality it offers.

However, even high-quality translations produced by modern translation systems are not of a publishable quality. Since a human intervention is required, one of the potential ways to improve MT is to ensure an optimized human-machine collaboration.

Human interactions with MT systems commonly involve two steps: (a) human intervention and injection of human knowledge into the MT process; (b) exploitation of the obtained human knowledge by MT.

Human intervention into the MT process typically takes place at the end of this process, and optionally at the beginning of it. The former intervention is referred to **Post-Edition** (PE), the process of manually correcting MT to achieve a publishable quality; the latter usually consists in processing the source text in order to prepare it for MT, and is referred to as **Pre-Edition** (PRE). Exploitation of the obtained human knowledge is commonly performed by means of regular updates to MT system trainable components.

A substantial part of the research in MT today is dedicated to studying PE and how to make this process less demanding in terms of human effort [i.a., Federico et al., 2013; Koehn et al., 2013; Denkowski et al., 2014; Koponen, 2015; Knowles and Koehn, 2016; Sanchez-Torron and Koehn, 2016; Chatterjee et al., 2017a]. PRE scenarios tend to receive less attention [i.a., Mohit and Hwa, 2007; Resnik et al., 2010; Marie and Max, 2015].

Human cognitive load during both PE and PRE can be significantly reduced if the machine indicates which target units require PE, or which source units (difficult-to-translate units) need to be pre-translated. Detection of those units can be performed using Quality Estimation (QE) techniques.

Concrete implementation choices for human-machine collaboration in professional production scenarios depend on many factors such as MT type, its quality, translation domain, qualities and skills of human experts, etc.

1.2 Towards a New Protocol for Improved Human-Machine Collaboration

In this work, we study a pre-edition scenario, hereinafter referred to as **Human-Assisted MT (HAMT) protocol**, that involves the following three steps: (1) automatic detection of fragments of the source text that could be problematic for the MT system; (2) resolution of these difficulties by a human expert, who provides the system with the expected translations of these segments (pre-translation); (3) exploitation of this information by the MT system to improve its output. Our protocol mainly targets the resolution of difficulties at the document level in a batch setting mode.

Steps (1)-(3) reproduce the typical behavior of a human translator: he or she will first analyze the source text, detect its parts that will be difficult to translate, then consult any external source of information to resolve difficulties before he or she starts translating. This suggests that our protocol may meet individual preferences of some translators and can be proposed as an alternative or a complement to PE [Kay, 1997].

There are other arguments in favor of PRE in general, of pre-translation in particular: it gives the human expert more control over the MT process, as it guarantees that some erroneous lexical choices or false senses will not appear in the target text.[1] Furthermore, when difficult

[1] As it can potentially happen if PRE is performed monolingually by means of source rewriting [Resnik et al., 2010; Seretan et al., 2014].

segments are repeated across a long document, their correct translation will be entered only once, where PE would imply multiple corrections of the same segment.[2] Finally, it is expected that constraining the translation process with human suggestions will also improve the machine output in the neighborhood of these correct segments, resulting in indirect improvements of quality that happen, as it were, for free. This will result in the reduction of human effort involved in final PE, which may be required depending on the translation purpose.

Our protocol can be particularly beneficial for production contexts, where PE is performed by mainly non-professionals (e.g., domain specialists). Non-professionals tend to introduce errors in their post-edits, as well to leave MT errors uncorrected. Among those errors, lexical errors risk to be the most harmful, as they more likely can lead to the incorrect understanding of text. Thanks to the pre-translation procedure of our protocol, which can be performed by professionals, more lexical errors in MT output can be prevented. Additionally, the final PE load, left to non-professionals, will be thus reduced to a minimum, at the same time reducing the risk of potential errors they may introduce.

The approach becomes even more beneficial in a multilingual setting, where common difficulties can be resolved once for many language pairs under the condition of availability of multilingual experts. Alternatively, translations into some languages can help to resolve difficulties in other languages.

To better assess the real potential of our protocol, we seek to answer in this work the following questions: (a) Can translation difficulties be reliably identified in step (1)? Should difficulties be detected at the level of words or phrases? (b) Can the human translations provided for difficult-to-translate units be successfully exploited by an MT system? How significant are indirect improvements? What is their nature? (c) How realistic is the PRE protocol in terms of the human effort involved? A last question, that may be worth asking in multi-source translation scenarios, finally is: (d) Are some source difficulties common to several target languages? Or are they specific to each language pair?

To answer these questions, we have decided to cast the task of difficulty detection as an automatic binary classification task. This solution is in line with state-of-the-art solutions to obtain targeted human help in MT [Mohit and Hwa, 2007; Bojar et al., 2017a]. We make the machine exploit the human help by means of *constrained decoding* and study the results of this exploitation by tracing quality improvements in second-pass MT. Indirect effects of re-translation

[2] A problem that is addressed by PE when complemented with adaptive learning [Mathur et al., 2013; Denkowski et al., 2014].

are studied as compared to the first-pass MT output, where we replace initial translations of difficulties by their pre-translations. Finally, we assess the human effort involved in pre-translation by comparing it to the one involved in PE. We also analyze translation difficulties common for several target languages sharing the same source in English.

1.3 Automatic Translation of Cochrane Medical Review Abstracts

Having presented the focus of our work, we will now present its context.

Cochrane Collaboration[3] is an international non-profit organization that regularly publishes high-quality evidence-based research review abstracts in medicine as a part of its mission to spread medical knowledge. These abstracts, originally written in English, are translated into 16 languages including French, Spanish, Japanese, and traditional Chinese.

For the English-French translation of abstracts, SMT combined with PE was introduced in Cochrane France (a part of Cochrane Collaboration) in 2013. For this purpose, we developed a specialized high-quality **Phrase-Based SMT** (PBSMT) system. PE of abstracts is systematically performed by mainly volunteers, specialists in the medical domain, who tend to introduce/leave errors in their post-edits related to the consistency of Cochrane terminology. Those post-edits are published online after limited quality control.

Thus, the proposed human-machine collaboration protocol is particularly beneficial for the Cochrane context as:

- it represents a control scenario, necessary to maintain the quality of post-edited abstracts. Resolution of source translation difficulties can be performed by professional translators with expertise in Cochrane terminology;

- a final PE step, as systematically required given the sensibility of medical information, can be performed by volunteer domain professionals. The PE effort reduction as a potential side effect of PRE is particularly important to decrease their workload.

Cochrane also proposes a multilingual scenario for difficulty resolution with English as the source language. At the same time the Cochrane context, in particular, in-domain MT, post-edited by domain specialists, multi-target translation with English as the source language, is

[3] `http://www.cochrane.org`

quite representative for MT settings in other domains, and we believe that all our results are generalizable and can be adapted to other configurations.

1.4 Contributions

Our first main contribution is **a system-independent methodology for translation difficulty detection**. We define the notion of **subsentential source-side translation difficulty**: difficult-to-translate segments are segments for which an MT system makes erroneous predictions. We show that, using this methodology, difficulties can be reliably detected both at the word level and at the phrase level, using a simple set of features without access to system-specific information.

We also study the problem of **detecting translation difficulties in a multilingual scenario**, the first attempt of this kind to our knowledge. Our study has allowed us to conclude that translation difficulties depend on the language pair, rather than solely on the source language or on the target language, which opens a range of new perspectives where several source texts in different languages are exploited (*multi-source translation*).

Our second main contribution consists in a proposal of a **HAMT protocol** that accommodates the results of our translation difficulty detection procedure and enables resolution of those difficulties by pre-translation. Our protocol mainly targets the resolution of difficulties at the document level. We assess our proposal in a simulated setting and provide some preliminary results of a real-life assessment. We show that pre-translating source difficulties could be both realistic in terms of the human effort involved and beneficial for the final MT quality. Indeed, the machine can successfully exploit human suggestions to improve the automatic translation of neighboring words.

1.5 Thesis Outline

This manuscript is organized as follows.

Chapter 2 gives an overview of SMT and its basics, as well an overview of automatic evaluation methods for MT.

Chapter 3 is focused on human-machine collaboration in MT. We describe the nature and the principles of human intervention into the MT process (PE and PRE), as well as the principles of exploitation of the obtained human knowledge by MT. We target the peculiarities of those

processes for SMT. We then detail our motivations to focus our work on the source-oriented targeted interaction that involves resolution of difficult-to-translate segments.

In Chapter 4 we describe the operational context, which led Cochrane France to invest in the production of high-quality English-French MT of medical research reviews. We assess the quality of post-edits performed to this MT by mainly volunteer domain specialists. We then present our fine-grained methodology for diagnosing this translation process. Our analysis showed that an SMT system in this type of configuration faces mainly text-specific contextual difficulties, which makes their automatic detection indispensable.

Chapter 5 presents our methodology for the automatic detection of source-side translation difficulties. We experiment with several segmentation types, different classification algorithms, as well as various sets of features, including system-independent and system-dependent features. We show that in an intrinsic evaluation translation difficulties can be reliably detected both at the word level and at the phrase level, using only a simple set of features. Furthermore, we study translation difficulty detection in a multilingual scenario.

We introduce our HAMT protocol in Chapter 6. This protocol accommodates our difficulty detection methodology and allows resolving difficulties by pre-translation. At first, we evaluate this difficulty resolution procedure in terms of its cost/benefit trade-off in a simulated sentence-level setting and conclude on its efficiency. We then provide promising results of our preliminary document-level HAMT experiments in a real-life setting.

We close the manuscript with a summary of the performed work and a discussion of some perspectives.

2 | A Statistical Machine Translation Primer

Contents

Machine Translation (MT) is the process of automatically translating from one natural language into another.

The first attempts of MT date back to the middle of the 20th century. These attempts tried to search for patterns of morphological, syntactic and semantic transfer rules, as well as to find a universal basis of the language. For instance, Oswald and Fletcher [1951] proposed to identify "noun blocks" and "verb blocks" in the source, and to determine which blocks are to be reordered before word-for-word translation into another language.[1]

Those attempts found their continuation in **Rule-Based MT**, which was the dominating approach during the 1970s (an approach adopted, for example, by Systran[2] which was one of the most popular commercial rule-based translation products). Those systems involved human linguistic knowledge to create language-specific transfer rules. However, the creation of such systems was very costly.

[1] For more on the early history of MT see. Hutchins [1997]

[2] `http://www.systransoft.com`

In the 1980s and 1990s the search for semantic *universalia* continued. **Example-Based** MT (EBMT) systems of that time searched for a sentence similar to the input sentence in a stored corpus of source sentences and their translations. Output contained the matched translation modified in a relevant way [Nagao, 1984].

In the 1990s **Statistical Translation** (SMT) emerged, as introduced by [Brown et al., 1990]. This new approach formalized MT as a statistical optimization problem. It relies on machine learning methods and requires a *parallel text* (bitext) as the main resource. In a bitext, each sentence in the source language is associated with a sentence in the target language or sentences in many target languages (see Table 2.1). Supported by the rapid development of computational power, as well by increasing quantities of openly accessible information on the Internet this approach became dominant at the turn of the century.

Recently, **Neural MT** (NMT) [Sutskever et al., 2014; Bahdanau et al., 2015] has become computationally manageable and has gained a lot of attention, offering a major qualitative improvement over statistical approaches.

This chapter will introduce the basics of SMT, focusing on the concepts that are necessary to understand the contents of this study.[3]

2.1 Basic Principles and System Types

SMT sees translation as the process of finding the most probable target sentence given a source sentence (see section 2.2). The process exploits probabilities that are estimated automatically by training statistical models using a parallel corpus and word alignments (see sections 2.3.1, 2.3, 2.4). The most probable translation is thus computed using the weighted combination of different model scores (see sections 2.5, 2.6). The process is routinely diagnosed by means of automatic MT evaluation (see section 2.7).

With regard to translation units, SMT can be divided into three groups: word-based, phrase-based and syntax-based SMT. The work presented in this thesis is framed within the phrase-based approach. The remainder of this chapter will provide a brief introduction to SMT using **Phrase-Based SMT** (PBSMT) as an example.[4]

[3] For a detailed introduction to SMT see e.g. [Koehn, 2010a]

[4] Descriptions of other SMT approaches can be found in Koehn et al. [2003]; Chiang [2005]; Lopez [2008a].

src.	trg.
Take me to Robin, quick!	*Bring mich zu Robin, schnell!*
Much, what's happened?	*Much, was ist geschehen?*
King Richard's in England. In Sherwood!	*Richard ist im Wald von Sherwood.*
What?	*Was?*
Prince John sent Dickon to Kent Road Tavern last night to kill the king.	*Prinz John schickte Dickon gestern zur Schänke, um den König zu töten.*

Table 2.1 – Example of an English-German parallel corpus (taken from the OpenSubtitles2016 corpus [Lison and Tiedemann, 2016])

2.2 Phrase-Based Statistical Machine Translation

The main rationale behind the introduction of phrase-based SMT was the fact that the word can not be considered as a universal minimal sense-bearing unit: groups of words often form a sense that is not deducible from them separately. This can be illustrated by numerous English phrasal verbs, which are verbs followed by a preposition or an adverb that changes the meaning of the verb: e.g., "get together" means "meet", whereas "get" means "obtain". Using phrases not only allows to take context into account, but also allows to pack syntactic groups or collocations into one single unit, enabling the system, for instance, to record some agreement phenomena.

The notion of a *phrase* defined by the approach has no linguistic foundation: in this context a "phrase" is just a word or a contiguous sequence of words.

The basic translation process can be decomposed into the following steps (Figure 2.1):

1. the source sentence $\mathbf{f}$ is split into I phrases $\bar{f}_1, \bar{f}_2, \ldots, \bar{f}_I$;
2. each $\bar{f}_i$ is translated independently (*independence assumption*);
3. the resulting target phrases are reordered into a target sentence $\mathbf{e} = \bar{e}_1, \bar{e}_2, \ldots, \bar{e}_J$.

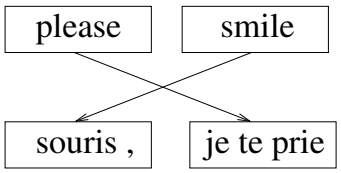


Figure 2.1 – Illustration of PBSMT

In the next sections we will describe the process more formally and give details on how translations and reorderings are computed and evaluated.

$\mathbf{f}$	source sentence
$\mathbf{e}$	target sentence
$\mathbf{e}^1$	the best scored MT hypothesis corresponding to $\mathbf{f}$ (1-best hypothesis)
$\hat{\mathbf{e}}$	reference human translation of $\mathbf{f}$ (possibly post-edited, i.e., corrected MT)
$\mathbf{e}^r$	MT hypothesis closest to $\hat{\mathbf{e}}$ as measured by an automatic metric
$f_1, f_2, \ldots, f_P$	source words
$e_1, e_2, \ldots, e_M$	target words
$\mathbf{a}$	an alignment between the words of $\mathbf{f}$ and the words of $\mathbf{e}$
$\mathbf{a}^1$	an alignment between the words of $\mathbf{f}$ and the words of $\mathbf{e}^1$ as produced by the decoder
$\bar{f}_1, \bar{f}_2, \ldots, \bar{f}_I$	source phrases as defined by PBSMT
$\bar{e}_1, \bar{e}_2, \ldots, \bar{e}_J$	target phrases as defined by PBSMT
$f_k, \ldots f_b = f_{[k:b]}$	arbitrary $\mathbf{f}$ segment
$e^1_r, \ldots e^1_g = e^1_{[r:g]}$	the best scored partial MT hypothesis corresponding to $f_{[k:b]}$

Table 2.2 – Notations used in the remainder of the thesis

Table 2.2 presents the notations that will be used in this chapter, as well as in the remainder of the thesis.

2.2.1 Formal Definition

We will now describe PBSMT more formally. Statistical approaches apply the *noisy channel model* to translation [Shannon, 1948]. They consider $\mathbf{f}$ as $\mathbf{e}$ distorted by some transmission noise. Thus, translation reduces to finding $\mathbf{e}$ that maximizes the product of its prior probability $p(\mathbf{e})$ and the conditional probability $p(\mathbf{f}|\mathbf{e})$, according to the Bayes rule:

$$\mathbf{e}^1 = \underset{\mathbf{e}}{\operatorname{argmax}}\, p(\mathbf{e}|\mathbf{f}) = argmax_{\mathbf{e}}\, p(\mathbf{f}|\mathbf{e})\, p_{LM}(\mathbf{e}), \tag{2.1}$$

where $p(\mathbf{f}|\mathbf{e})$ is computed by the *translation model* (TM) and $p_{LM}(\mathbf{e})$ is computed by the *language model* (LM). In practice the TM consists of a set of models that can include translation, reordering and distortion probabilities learned from a parallel corpus. $p_{LM}(\mathbf{e})$ models the probability that $\mathbf{e}$ exists in the target language, assigning higher scores to frequent words or word sequences, thereby providing a proxy to grammatical correctness and typicality.

We will now explain how to build those models in their simplest version.

2.3 The Translation Model

A simple TM consists of a *phrase-table* (PT) and a *reordering model* (RM).

The creation of a PT involves: (1) creating a word alignment **a** between pairs of parallel sentences (**f**,**e**) in a bitext; (2) extracting parallel phrases (bi-phrases ($\bar{f}$, $\bar{e}$)) consistent with **a**.

2.3.1 Word Alignments

A word alignment **a** encodes possible word-level correspondences between two languages. This assumes that each source word can be translated to zero, one or several word(s). Given a source sentence $\mathbf{f} = f_1, f_2, \ldots, f_P$ and a parallel target sentence $\mathbf{e} = e_1, e_2, \ldots, e_M$, the most intuitive presentation of an alignment between them is a matrix $A \in B_{M,P}(\{0,1\})$, where each cell $a_{m,p}$ indicates whether f_p is aligned to e_m or not. To avoid computational problems an asymmetric model is typically used, where each word of one sentence is labeled with the position of a word in another sentence (Figure 2.2).

Thus, the alignment function **a** maps each e at position m to one f at position p:

$$a : m \rightarrow p \tag{2.2}$$

When a word e_m has no relation to any word in **f**, it is considered to be aligned to the word f_0 =`NULL`. The introduction of the `NULL` word extends **f** by one word $(P + 1)$.

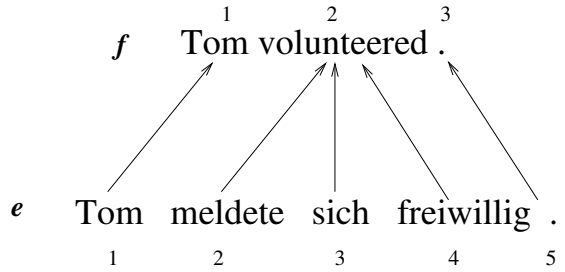


Figure 2.2 – Example of aligning German words with positions in the English sentence

IBM Models are the most commonly used alignment models [Brown et al., 1993].

IBM Model 1 is the simplest lexical translation model, in which each position in the target has the same probability to align to any position in the source: in other words, the probability of a link (e_m, f_p) does not depend on the positions m and p:

$$p(\mathbf{e}, \mathbf{a}|\mathbf{f}) = \frac{\epsilon}{(P+1)^M} \prod_{m=1}^{M} t(e_m|f_{\mathbf{a}(m)}), \tag{2.3}$$

where ϵ is a normalization constant.

More complex IBM word alignment models search to overcome deficiencies of *IBM Model 1*. One of these deficiencies is the assumption that the probability of a link (e_m, f_p) does not depend on the positions m and p, even though a source word at the beginning of a sentence is rarely aligned to a target word at the end of a sentence, which is especially true for related languages. Thus, *IBM Model 2* introduces a dependency between the absolute position m of the aligned word and the absolute position p. It is also important to consider that some words tend to have no translation into another language, or on the contrary be translated using several words (e.g., the English auxiliary word "do" often has no direct counterpart). *IBM Model 3* models this tendency of target words to received more or less than a single link (e_m, f_p) and explicitly includes a fertility model. *IBM Model 4* adds a relative distortion model, in which the position of the translation of an input word is based on the position of the translation of the previous input word. This model also takes word classes into account to deal with the data sparsity issue.[5]

Another commonly used alignment model is the *Hidden Markov Model* (HMM) model proposed by Vogel et al. [1996]. This model makes a link (e_m, f_p) explicitly dependent on the preceding link, rather than on absolute positions of aligned words. This model offers a series of attractive properties and serves as the basis for numerous extensions [Toutanova et al., 2002; Deng and Byrne, 2005; He, 2007; Graça et al., 2010].

In practice alignments are obtained by using a cascade of models of increasing complexity, where simpler models are used to initialize more complex ones.

Alignments produced in both directions are merged into a unified representation using various symmetrization heuristics (usually the union of alignment links is taken) [Och and Ney, 2003; Koehn, 2010a].

Word alignments are quite costly to create manually and can not be readily deduced from the parallel data. This problem is considered as a problem of *incomplete data* and is solved by the unsupervised *Expectation-Maximization* (EM) algorithm [Dempster et al., 1977].

[5] For more information on IBM Models see [Brown et al., 1993; Och and Ney, 2003]

2.3.2 Phrase-Table Building

Assuming we have an alignment $\mathbf{a}$ generated by a combination of the means described above, the next step is to extract and score the bi-phrases $(\bar{f}, \bar{e})$ consistent with $\mathbf{a}$.

More formally, a bi-phrase $(\bar{f}, \bar{e})$ is consistent with the alignment $\mathbf{a}$ if all the words in $\bar{f}$ have alignment points in $\mathbf{a} = \{(p, m) \subset \{1 \dots P\} \times \{1 \dots M\}\}$ with the words in $\bar{e}$ and *vice versa*:

$$\begin{aligned} (\bar{f}, \bar{e}) \text{ consistent with } \mathbf{a} \Leftrightarrow \\ \forall f_p \in \bar{f} : (f_p, e_m) \in \mathbf{a} \Rightarrow e_m \in \bar{e} \\ AND\ \forall e_m \in \bar{e} : (f_p, e_m) \in \mathbf{a} \Rightarrow f_p \in \bar{f} \\ AND\ \exists f_p \in \bar{f}, e_m \in \bar{e} : (f_p, e_m) \in \mathbf{a} \end{aligned} \tag{2.4}$$

After all the bi-phrases are extracted, it is possible to estimate how many sentence pairs include $(\bar{f}, \bar{e})$ just by counting. Finally, the translation probability $\phi(\bar{e}|\bar{f})$ is estimated as relative frequency:

$$\phi(\bar{e}|\bar{f}) = \frac{count(\bar{f}, \bar{e})}{\sum_{\bar{e}_i} count(\bar{f}, \bar{e}_i)} \tag{2.5}$$

Inverse translation probabilities $\phi(\bar{f}|\bar{e})$ are estimated in a similar way.

However, infrequent bi-phrases can cause problems. If we meet both $\bar{f}$ and $\bar{e}$ only once in a corpus, then $\phi(\bar{f}|\bar{e}) = \phi(\bar{e}|\bar{f}) = 1$, which is clearly an overestimation. To overcome this problem, it can be estimated how lexically coherent are word translations within a bi-phrase by means of a smoothing method called *lexical weighting* [Koehn et al., 2003]:

$$lex(\bar{e}|\bar{f}, \mathbf{a}) = \prod_{m=m_{start}}^{m_{end}} \frac{1}{|\{p|(p, m) \in \mathbf{a}\}|} \sum_{\forall (p,m) \in \mathbf{a}} w(e_m|f_p), \tag{2.6}$$

where the lexical translation probability $w(e_m|f_p)$ is estimated as relative frequency from the word-aligned corpus:

$$w(e|f) = \frac{count(f, e)}{\sum_{e'} count(f, e')} \tag{2.7}$$

The lexical weighting is also computed in both directions: $lex(\bar{e}|\bar{f}, \mathbf{a})$ and $lex(\bar{f}|\bar{e}, \mathbf{a})$.

The resulting probabilities are stored in a phrase-table and used while translating $\bar{f}$. Even if the training corpus (a parallel corpus used for phrase and corresponding statistics extraction) is

$\bar{e}$	$\phi(\bar{e}\|\bar{f})$	$lex(\bar{e}\|\bar{f},\mathbf{a})$	$\phi(\bar{f}\|\bar{e})$	$lex(\bar{f}\|\bar{e},\mathbf{a})$
implants zygomatiques ,	0.33	0.036	0.33	0.8
implants zygomatiques	0.33	0.8	0.33	0.8
les implants zygomatiques	0.33	0.16	0.5	0.4

Table 2.3 – PT entries for $\bar{f}$ "zygomatic implants"

very large, test data can contain words unknown to the model (*out-of-vocabulary*, OOV). They are most commonly just copied to output, which can significantly decrease its quality.

The example in Table 2.3 illustrates PT entries storing the above mentioned probabilities. We can see that some of those entries contain noise: for instance, associating "zygomatic implants" with "*implants zygomatiques ,*" receives the same direct translation probability score as the other bi-phrases ($\phi(\bar{e}|\bar{f}) = 0.33$), as each of them is seen in the training data only once. This score is balanced by a relatively low lexical weighting score of $lex(\bar{e}|\bar{f},\mathbf{a}) = 0.036$, since the target phrase "*implants zygomatiques ,*" contains a token (comma in this case) that has no high-probability alignment link to any of the words in the source phrase "zygomatic implants" (including the `NULL` word).

2.3.3 Reordering Models

Reordering models (RMs) are necessary to evaluate the order changes between the source and the target sentences.

The simplest *distance-based RM* considers target phrase reordering relative to the previous phrase in the source language [Koehn et al., 2003]. Let $start_i$ and end_i be the positions of the first and last words of $\bar{f}_i$ that translates into $\bar{e}_j$. A reordering distance is computed: $start_i - end_i - 1$. The reordering distance is the number of words skipped (either forward or backward) when taking source words out of sequence. This model is non-selectively applied to all source phrases. However, some phrases are reordered more frequently than others, suggesting more sophisticated models.

The basic model is usually extended by a *lexicalized RM* [Tillmann, 2004], which predicts the orientation type o given an actual bi-phrase: $p_o(o|\bar{f},\bar{e})$.

The following reordering types are commonly considered:

- *monotone orientation* (Mon) – a bi-phrase directly follows the preceding bi-phrase;
- *swap orientation* (Sw) – a bi-phrase is moved before the preceding bi-phrase;

- *discontinuous orientation* (Dis) – a bi-phrase is not adjacent to the preceding bi-phrase.

Figure 2.3 illustrates the three reordering types. For instance, the orientation of "*Mir*" translating "to me" is of type "swap" (Sw) as it comes before the translation of the preceding source group ("It seems").

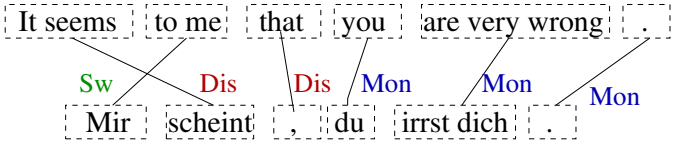


Figure 2.3 – Illustration of the reordering types for an English-German translation (taken from the Tatoeba corpus (`https://tatoeba.org/eng`))

The probabilities are estimated over a corpus as follows:

$$p_o(o|\bar{f},\bar{e}) = \frac{count(o,\bar{f},\bar{e})}{\sum_o count(o,\bar{f},\bar{e})} \tag{2.8}$$

Other commonly used reordering models are *hierarchical* models [Galley and Manning, 2008] that improve reordering by taking syntactic information into account.

2.4 The Language Model

A Language Model (LM) estimates how likely a sentence is, assessing mostly the correctness of the particular juxtaposition of words it contains. More fluent sentences are thus scored higher. LMs are usually built using some Markov assumption (n-gram[6] models). The probability of a sequence $p(e_1^M)$ is decomposed as follows:

$$p(e_1^M) = p(e_1)p(e_2|e_1)\dots p(e_m|e_1,e_2,\dots,e_{m-1}) \tag{2.9}$$

Thus, $p(e_1^M)$ is the product of $p(e)$ given the history of the preceding words. Usually this history is limited to the $n-1$ previous words. For computational reasons, the value of n is small and rarely exceeds 5 or 6.

LM probabilities are estimated over a large target corpus. For instance, the probability of the 3-gram $p(e_3|e_1,e_2)$ is computed as follows:

[6] An n-gram is a contiguous sequence of n words.

$$p(e_3|e_1,e_2) = \frac{count(e_1,e_2,e_3)}{\sum_{e'_3} count(e_1,e_2,e'_3)} \tag{2.10}$$

And the LM probability of a sentence for a 3-gram model is computed as follows (usually computed as log-probability for computational stability):

$$\begin{aligned}\log p(Er, schloss, die, Hochschule, ab) = &\log p(Er| < s >, < s >)\\ &+ \log p(schloss| < s >, Er)\\ &+ \log p(die|Er, schloss)\\ &+ \log p(Hochschule|schloss, die)\\ &+ \log p(ab|die, Hochschule)\\ &+ \log p(< /s > |Hochschule, ab),\end{aligned} \tag{2.11}$$

where $< s >$ and $< /s >$ respectively denote the beginning and the end of a sentence.

To face the problem of sparse data a series of *smoothing* techniques is used in language modeling [Witten and Bell, 1991; Kneser and Ney, 1995; Chen and Goodman, 1996; Brants et al., 2007]. Those techniques assign a non-zero probability to unseen events using the statistics of lower order n-grams to estimate parameters of higher order models.

2.5 Scoring

While searching for the best translation $\mathbf{e}^1$ according to Equation (2.1) different components of the model can become more or less important and take different weights.

Thus, for computational convenience Equation (2.1) is usually presented in a form of a *log-linear model*, i.e., transformed in a linear combination of the logarithms of the model parameters:

$$\mathbf{e}^1 = \underset{\mathbf{e}}{\operatorname{argmax}} \sum_{d=0}^{t} \lambda_d h_d(\mathbf{f}, \mathbf{e}, \mathbf{a}), \tag{2.12}$$

where h_d are, for example, the following feature functions:

$$
\begin{aligned}
h_1(\mathbf{f}, \mathbf{e}, \mathbf{a}) &= \sum_{i=0}^{I} \log \phi(\bar{f}_i|\bar{e}_i) \\
h_2(\mathbf{f}, \mathbf{e}, \mathbf{a}) &= \sum_{i=0}^{I} \log d(start_i - end_{i-1} - 1) \\
h_3(\mathbf{f}, \mathbf{e}, \mathbf{a}) &= \log p_{LM}(\mathbf{e}),
\end{aligned}
\tag{2.13}
$$

where $\phi(\bar{f}_i|\bar{e}_i)$ represents the probability given by the TM, e_i is a target phrase in the translation with its corresponding source phrase f_i; $d(start_i - end_{i-1} - 1)$ is the distance-based RM; and $p_{LM}(\mathbf{e})$ is the LM probability.

Weights tuning is performed by translating a smaller parallel corpus (*development set*, usually less than 5K lines), excluded from the training data. This set is supposed to be an appropriate approximation of the test data.

A system translates a development set with initial parameters, then the weights are updated so as to improve the model score of the best/most promising translation hypotheses in a so-called *n-best list* (the top n translations found according to the model), where translation quality is measured by any automatic metric (see sections 2.6, 2.7). The process is repeated until convergence (i.e., when no translation quality improvement is observed).

Weight updates are usually performed either by using `mert` (Minimum Error Rate Training) [Och, 2003] or `kb-mira` (a variant of the Margin Infused Relaxed Algorithm) [Cherry and Foster, 2012] algorithms.[7]

2.6 Decoding

The process of finding $\mathbf{e}^1$ according to Equation (2.12) is called *decoding*. This problem is known to be NP-complete [Knight, 1999]. This is because an exact decoding will need to explore a combinatorial number of segmentations, permutations and translations of the source sentence. Consequently, *heuristic search* methods are used to find a solution close to the optimal in a reasonable time.

A commonly used heuristics is the *beam search algorithm* [Jelinek, 1997]. An output sentence is generated left to right by extending an existing partial translation (hypothesis) with translations of yet untranslated phrases $\bar{f}$. Hypotheses are stored in data structures that permit

[7] For more on the optimization of the system parameters see e.g. Neubig and Watanabe [2016]

efficient pruning of similar or low-probability hypotheses [Koehn et al., 2003].

The search space is often presented in the from of a search graph (Figure 2.4). The n-best scored paths in this graph form the *n-best list*. Each hypothesis, including $\mathbf{e}^1$, is usually provided with word and phrase alignments from the PT.

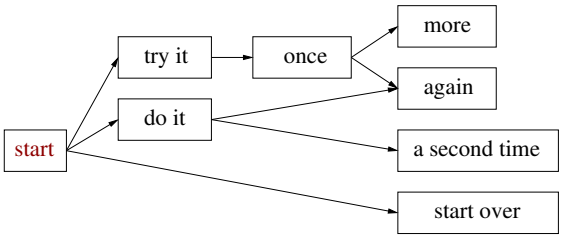


Figure 2.4 – Example of an output search graph for the source sentence "*Essaye encore une fois*". $\bar{e}$ are represented by rectangles, a path from the start to any non-final state forms a partial hypothesis.

2.7 Automatic Evaluation

As shown above, evaluating the quality of MT systems is an important part of the development process. Since human evaluation is rather expensive and time-consuming, usually an automatic evaluation is used. It compares MT hypotheses to human translations (*references*). The most widely used automatic metrics are BLEU, TER and METEOR.

BLEU (BiLingual Evaluation Understudy) computes the geometric mean of modified n-gram precisions between hypotheses and references [Papineni et al., 2002].[8] It is defined as follows:

$$BLEU \;=\; BP \;\cdot exp\left(\sum_{n=1}^{N} w_n \; log \; PR_n\right), \tag{2.14}$$

where the brevity penalty (BP) is computed according to:

$$BP = \begin{cases} 1 & \text{if } l_{hyp} > l_{ref} \\ e^{(1-l_{ref}/l_{hyp})} & \text{if } l_{hyp} \leq l_{ref}, \end{cases} \tag{2.15}$$

where PR_n is the modified n-gram precision, the ratio between n-grams of a hypothesis matching reference n-grams and the total number of hypothesis n-grams. The brevity penalty penalizes short translations. The default configuration uses $N = 4$ and uniform weights $w_n = 1/N$.

[8] Precision estimates how many units generated by the system are correct.

What a pity!
It's a pity.
What a shame!
Too bad.
That's too bad.
What a pity.

Table 2.4 – Examples of multiple translation into English of the German sentence "*Schade*" (taken from the Tatoeba corpus)

TER (Translation Edit Rate) is defined as the minimum number of edits (*#op*) required to change a hypothesis so that it exactly matches a reference [Snover et al., 2006]:

$$TER = \frac{\#op}{l_{ref}} \tag{2.16}$$

The basic set of edit operations comprises the following four operations: substitution (S), insertion (I), deletion (D) and shift (SH). As optimal calculation of edit-distance with shift operations is NP-complete [Shapira and Storer, 2007], the metric uses an approximation. At the first step, it employs a greedy search to find an optimal set of shifts that minimizes the number of edit operations. Then, the remaining edit distance is found by means of dynamic programming techniques. As a by-product the metric computes a monolingual alignment $\mathbf{e}^1 \rightarrow \hat{\mathbf{e}}$. In this alignment each word-to-word connection is labeled with an operation type or a match (M) applied to e^1 to obtain $\hat{e}$.

The HTER (Human-Targeted TER [Snover et al., 2006]) variant is measured using actual post-edited MT as the reference translation.

Those standard metrics have received a lot of criticism for (a) giving the same importance to all the words in a sentence (however, only a certain set of key words in a sentence is crucial for its understanding); (b) matching words exactly, since sometimes the grammatical form is less important to convey a meaning; not taking translation variants into account, etc. (e.g., the criticism of BLEU by Callison-Burch et al. [2006]; Lommel [2016], etc., see Table 2.4).

Some of those criticisms are answered by METEOR (Metric for Evaluation of Translation with Explicit ORdering) [Denkowski and Lavie, 2014]. The metric creates a hypothesis $\rightarrow$ reference alignment (a mapping between unigrams). It matches synonyms, morphological variants and paraphrases. The final score is computed as follows:

$$METEOR \;=\; F_{mean} \cdot (1 - penalty), \tag{2.17}$$

where F_{mean} is equal to:

$$F_{mean} = \frac{10 \cdot PR_{uni} \cdot RC_{uni}}{RC_{uni} + 9 \cdot PR_{uni}}, \tag{2.18}$$

where PR_{uni} is the unigram precision, RC_{uni} is the unigram recall.[9]

The penalty is computed as follows:

$$penalty = 0.5 \cdot (\frac{\#chunks}{\#unigrams\_matched})^3, \tag{2.19}$$

where the *chunk* is a maximal group of adjacent words that are aligned to adjacent positions in the reference.

Similarly to METEOR, the TERp extension of TER uses a broader notion of a match and takes synonyms (Y), paraphrases (PAR) and stem[10] matches (T) into account [Snover et al., 2009a].

To answer the criticism on using a single translation variant as the reference, BLEU and TER can be extended to accommodate multiple references (hence, for instance, this extended version of BLEU is referred to as multi-BLEU).

Other metrics take other aspects of translation quality into account: e.g., discourse structures (DiscoTK [Joty et al., 2014]), dependency relations (DPMF [Yu et al., 2015]), semantic entailment relationships (MEANT [Lo and Wu, 2013]).

Since 2008 the annual Workshop on Machine Translation (WMT) tasks perform evaluations of new automatic metrics [Machacek and Bojar, 2014; Stanojević et al., 2015; Bojar et al., 2016d, 2017b]. According to these evaluations BLEU is still competitive, and competes mainly with character-based metrics [Bojar et al., 2016c].

An example of a character-based metric is CharacTer [Wang et al., 2016]. It can be viewed as a character-level version of TER with a normalization over the hypothesis sentence length:

$$\text{CharacTer} = \frac{\text{shift cost} + \text{edit distance}}{\#\text{hyp characters}} \tag{2.20}$$

Such a normalization permits to score longer hypotheses higher, which was found to better correlate with human judgements.

Other metrics that compete with BLEU and TER are tuned combination metrics. They comprise many features and multiple metrics (e.g., BEER [Stanojević and Sima'an, 2015], DPM-

[9] Recall computes how many of the units that should be generated are correctly produced.
[10] Stem is the constant part of an inflected word.

Fcomb [Yu et al., 2015], RATATOUILLE [Marie and Apidianaki, 2015]).

2.8 Summary

This chapter has presented the basics of SMT using the example of its phrase-based version (PBSMT).

SMT searches to find the most probable translation given a source language sentence. This type of MT is based on the strong assumption that each *phrase* (sequence of words) can be independently translated into another language. The resulting target phrases are then reordered. The extraction of phrases, their translations and probabilities is performed using a parallel corpus. Each sentence of the source language in a parallel corpus is associated with a translation into the target language. The extraction is based on *word alignments* (correspondences) between words in parallel sentences.

The independence assumption is counterbalanced by the *language model* (LM) that computes the probability of a certain word given its preceding context, often limited to 3-4 words. Thus, the whole model better approaches the real-life conditions where words of a sentence form a common context and are moreover connected in the context of a text.

Different SMT models are not equally important and are usually provided with weights, which are optimized on a small representative part of parallel data (development set).

All the possible phrase segmentations of each source sentence, their possible translations and reordering possibilities form a complex search space. Heuristic search methods are needed to find a solution close to the optimal in a reasonable time (e.g., *beam search algorithm*).

Various automatic metrics, that compare automatic hypotheses to human references, are used for cheap and quick evaluation of SMT systems during the development phase.

The *vanilla* PBSMT architecture is implemented in the `Moses` toolkit [Koehn et al., 2007], which will be used in all the MT-related experiments in this work.

At the time of its creation SMT offered a range of advantages compared to alternative approaches (rule-based, example-based): it is language-independent, and SMT systems can be built quickly, fully automatically and at a low cost (only a parallel corpus is needed). Those advantages permitted SMT to gain significant popularity and build state-of-the-art MT systems for almost 20 years.

Nevertheless, SMT is prone to numerous deficiencies. Almost all of its aspects have been subject to continuous improvement during those years. For instance, due to its intrinsic limitations

src.	A patient with **dermatofibrosarcoma protuberans** on the left thigh, **a rare sarcoma** of the soft parts, **is presented**.
hyp.	*Un patient avec un* ***le dermatofibrosarcome*** *sur la cuisse gauche,* ***une rare sarcome*** *des tissus mous,* ***est présentée.*** 'A patient with a **the dermatofibrosarcoma** on the left thigh, **a (Fem., Sg.) rare sarcoma** of soft tissues, **is presented (Fem., Sg.)**.'
ref.	*Un patient atteint* ***d'un dermatofibrosarcome protubérant*** *sur la cuisse gauche,* ***un rare sarcome*** *des parties molles,* ***est présenté.*** 'A patient with **a dermatofibrosarcoma protuberans** on the left thigh, **a (Masc., Sg.) rare sarcoma** of the soft parts, **is presented (Masc., Sg.)**.'

Table 2.5 – Example of SMT deficiencies (English-French)

which include, among others, the generation of translations by mere concatenation, SMT is often unable to resolve long-distance semantic and syntactic relations between sentence components, even for closely-related languages, like English and French. The example in Figure 2.5 shows that an SMT system failed to translate the rare term "dermatofibrosarcoma protuberans", as well as failed to choose a correct gender form for the article "a" separated from its head word "sarcoma" by an adjective. Also note the agreement error between the past participle form "*est présenté*" 'is presented' and its masculine head noun "*patient*" 'patient'.

Recently, *Neural MT* (NMT) has emerged as a promising alternative architecture and defined new state of the art that outperforms SMT systems in fluency of translation (output is more correct from the point of view of a language). NMT models take larger source and target contexts into account. However, SMT translations are still sometimes more adequate (better preserve source content) [Bentivogli et al., 2016; Bojar et al., 2016a]. They were also shown to be more competitive in narrow in-domain scenarios [Farajian et al., 2017]. PBSMT is also more transparent and offers ways, through the continuous extension of the PT, to improve its knowledge base with immediate impact on the translation quality.

We are inclined to believe that SMT, including PBSMT, will continue to be state of the art for a while in certain production conditions. Such a use case is presented in this work: PBSMT provides high-quality English-French translation in a specialized medical domain, for which we have access to the training data that closely match the test data.

3 | Human-Machine Collaboration in Statistical Machine Translation

Contents

Today even high-quality MT does not produce output of a publishable quality. Since a human intervention is required, one of the potential ways to improve MT is to ensure an optimized human-machine collaboration.

MT-related human-machine collaboration commonly involves two aspects: (a) human intervention and injection of human knowledge into the MT process (see section 3.1); (b) exploitation

of the obtained human knowledge by MT (see section 3.3).

The idea of human-in-the-loop MT emerged after researchers started to realize the impossibility to create a universal language or to produce "perfect" fully-automatic translations [Bar-Hillel, 1951; Weaver, 1955; Kay, 1997].[1] Since then, the most common type of human intervention into MT is the *ex-post* intervention, which is usually referred to as **post-edition** (PE) (see section 3.1.1). During PE the human corrects MT errors to achieve a publishable quality. Additional human intervention can take place *ex-ante*. It consists in normalizing the input to make it more easily processable by the machine, and is usually referred to as **pre-edition** (PRE) (see section 3.1.2). Both processes can be guided by the machine, which anticipates where human intervention is needed using, for instance, **Quality Estimation** (QE) techniques (see section 3.1.3). Another way to help humans in the laborious PE and PRE tasks is proposed by **Interactive Machine Translation** (IMT), which predicts sequences that will be typed in by the user (section 3.2). In our description we will focus our attention on discussing peculiarities of PE and PRE of SMT.

In section 3.3 we will present the ways of exploiting the obtained human knowledge, which mainly consist in online updates of models, again focusing on SMT. In section 3.4 we will briefly describe **Computer-Assisted Translation** (CAT) environments, which accommodate PE and PRE, ensure the interactive component, feedback exploitation, as well as offer an extended translator help.

All those approaches to human-machine collaboration will be presented together with production scenarios they fit in. We close the chapter by providing our reasons to focus this study on targeted PRE scenarios using source-side QE.

3.1 Injection of Human Knowledge

The idea of PE and PRE as the processes of human knowledge injection into MT was first conceptualized in the works of Reifler [1950]; Bar-Hillel [1951]; Kay [1973]. In these works human intervention took place at the end, and optionally at the beginning, of the MT process. The role of a *post-editor* (a human who performs PE), who was supposed to know the target language, was to resolve residual semantic ambiguities (unresolved by the machine), as well as to perform stylistic smoothing. A *pre-editor* (a human who performs PRE), who knows the source language, should eliminate morphological and syntactical ambiguities, as well as reorder source

[1] "For those targets in which high accuracy is a conditio sine que non, pure MT has to be given up in favor of a mixed MT, i.e., a translation process in which a human brain intervenes." [Bar-Hillel, 1951]

words according to the target language word order using a set of instructions. PRE can also be performed by the machine without human help [Bar-Hillel, 1951]. With the passing of time peculiarities of PE and PRE have been changing with MT paradigm shifts, whereas the nature and points of human intervention stay unchanged (see Figure 3.1).

As for the types of knowledge necessary to perform PRE and PE, they have been questioned as well [Koehn, 2010b; Hu et al., 2011; Schwartz et al., 2014]. Human translation traditionally requires fluent knowledge of the source language and native knowledge of the target language. But can a text be prepared for translation without any knowledge of the target language? And can PE be performed without any knowledge of the source language? As for today, those questions are solved in different ways depending on each particular task.

In the following sections we will describe the main tasks of a post-editor dealing with SMT output (see section 3.1.1). We will highlight the main strategies of PRE for SMT, such as source normalization, which aims to make the test samples more alike to the training data (see section 3.1.2). In section 3.1.3 we will give an introduction to QE that predicts MT quality. Target-side QE can guide the PE process by indicating which parts of MT need to be post-edited. Source-side QE can be used to indicate where targeted human help is needed during PRE. This targeted help can then be exploited by the decoder to generate a second-pass MT output of a better quality.

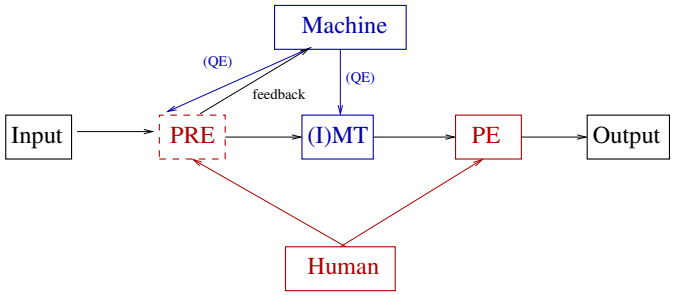


Figure 3.1 – Human-machine collaboration (dotted boxes denote optional steps)

3.1.1 Post-Edition

PE at the current stage of MT development is categorized according to the amount of the human effort involved: *surface or light PE*, with general purpose to make MT understandable, and *full PE*, which is supposed to bring the resulting text to the level of a full-fledged human translation.

In practice, even if guidelines are provided,[2] the amount and the nature of corrections in both cases tend to stay undefined. They depend on the target language, and more importantly on the subjective language perception of a post-editor. Different studies report different conclusions about the preferred amount of PE: even trained professionals can tend to change either "too much" or "too little" [e.g., Depraetere, 2010; Koponen, 2015], and only a very low percentage of sentences is post-edited by different post-editors in the same way [e.g., Wisniewski et al., 2013]. However, recent research tend to agree on the fact that PE results are of comparable quality to the ones produced by human translation [e.g., Fiederer and O'Brien, 2009; García, 2011; Guerberof Arenas, 2014].

As hiring trained professionals is costly, often the help of non-professional (lay) post-editors, domain specialists is used. Such post-editors tend make lexical errors in their post-edits, as well as other errors related to translation accuracy (e.g., missing information). They also systematically leave some MT errors uncorrected, especially those that are related to literal translation or syntactic errors [Mitchell, 2015]. Different studies suggest providing non-professional post-editors with professional PE help on request [Schwartz et al., 2014; Mitchell, 2015].

Another side of the problem is the amount of the human effort involved in PE: cognitive load to analyze poorly-translated MT output, the effort of making repetitive changes to systematic errors in this output, etc. Taking the complexity of the activity into account, measuring this effort is challenging. It can be done using standard automatic metrics (see section 2.7), keyboard/mouse event rates (e.g., *actual edit rate* [Och et al., 2003; Barrachina et al., 2008; Sanchez-Torron and Koehn, 2016]), post-editing time, eye-tracking [Bojar et al., 2016b], different fine-grained methodologies [Temnikova, 2010; Lacruz et al., 2014; Popovic et al., 2014], etc.

A certain number of recent works confirm that PE is less costly than human translation in terms of the effort involved [e.g., Plitt and Masselot, 2010; Green et al., 2013; Elming and Carl, 2014; Guerberof Arenas, 2014]. This effort, however, as well as the final translation quality tend to depend on the quality of the post-edited MT output [e.g., O'Brien, 2011; Sanchez-Torron and Koehn, 2016]. Thus, IMT seeks to help the human by proposing completions to partial translations he or she has typed in and to save eventually a certain amount of keystrokes (see section 3.2). An MT system can also constantly improve its output by adapting itself to post-edited feedback (see section 3.3.1), which may help to avoid repetitive corrections of this output by the human. Finally, only MT output of a certain quality could be presented to a post-editor

[2] For instance, see the guidelines provided by the TAUS resource center: `https://www.taus.net/academy/best-practices/postedit-best-practices/machine-translation-post-editing-guidelines`

(which is "worth post-editing"). This problem is addressed by QE (see section 3.1.3).

3.1.2 Pre-Edition

PRE for SMT usually consists in normalizing the source to make the test samples more alike to the training data. Such PRE modifications are related to the ones applied by controlled languages (CLs) [Kuhn, 2014]. CL is intended to restrict text creation using a certain set of predefined lexical and syntactic patterns under the condition that principal natural properties of the base language are preserved (not always the case for PRE). The usage of CLs has generally positive impact on the resulting quality of MT output. However, the implementation of such languages, involving their creation and special training for users, risks to be costly and is almost reserved to large-scale industrial projects [Aikawa et al., 2007; Temnikova, 2010].

PRE modification of SMT can be *human-* or *machine-oriented*. Human-oriented changes target improving MT while improving input from the linguistic point of view: for instance, resolving casing and spelling issues, correcting punctuation, resolving homophone confusion issues (e.g., "plain/plane"), grammatical issues (an incorrect verb form or an incorrect preposition), etc. These changes are applied to reduce noise in user-generated content [Jiang et al., 2012; Seretan et al., 2014], or to improve input readability [Miyata et al., 2015]. This type of PRE is costly, as it requires the manual creation of rules. Additionally, those rules risk not to be portable to other types of texts or language pairs.

Machine-oriented PRE targets improving only MT and allows all possible changes to input, up to reducing its readability. Typical tasks involve source reordering to bring its word order closer to the one of the target language, handling register mismatches, paraphrasing or simplification according to the MT training data, etc. In a fully automatic mode, a PRE strategy can be applied to all source units: for instance, applying pre-reordering rules for all input sentences [Xia and McCord, 2004; Crego and Mariño, 2006; Ceausu and Hunsicker, 2014; Pal et al., 2014], presenting all paraphrased variants of a sentence as an input graph [Du et al., 2010; Onishi et al., 2010], etc. Applying such strategies results in stable positive effects on MT mostly for borderline cases: for instance, for language pairs with significant and systematic syntactic differences, like English and Hindi [Pal et al., 2014], or for resource-limited language pairs [Du et al., 2010]. For less extreme cases, like English-French translation, such strategies may require systematic evaluation, so that positive changes in final MT would not be neutralized by negative effects. Thus, for instance, in an attempt to address register mismatches between the training and the test data for English-French translation, Rayner et al. [2012] show that the rule of systematic

rewriting of second person singular into second person plural verb forms with the corresponding changes of personal pronouns tends to improve MT output. However, systematic rewriting of "*est-ce que*" 'is it that' in informal questions into formal style can even slightly decrease the quality of output, since translations of questions are highly context-dependent.

Targeted contextual application of PRE strategies can be more efficient [Morita and Ishida, 2009; Resnik et al., 2010; Mirkin et al., 2013; Marie and Max, 2015]. For instance, [Resnik et al., 2010] study a scenario, where problematic source units are detected by comparing the original source to the back-translated one.[3] Back-translated units that do not match the corresponding original source units are then paraphrased by monolingual users. Other targeted PRE scenarios may involve QE (see section 3.1.3).

However, all the above-mentioned monolingual PRE strategies do not guarantee positive improvements, and more importantly correct lexical choices in final MT output. A bilingual pre-translation strategy can address this problem and help to secure translation choices in this output. Pre-translations can be provided by the human [Mohit and Hwa, 2007] or obtained from pre-defined dictionaries [Arcan et al., 2014; Cheng et al., 2016]. They can be then used as constraints while decoding. This results in stable improvements in MT output due to the presence of "correct" pre-translations, as well as in possible positive changes in the translation of their contexts.

An original approach to provide pre-translations is proposed by Marie and Max [2015], where the user helps the machine by indicating "correct" MT choices in first-pass output. The machine then exploits the information on the boundaries of these "correct" translations: while re-decoding it can filter its models to prioritize "good" phrases and neglect "bad" phrases.

3.1.3 Quality Estimation and its Role in Post- and Pre-Edition

Quality Estimation (QE) [Blatz et al., 2004; Specia et al., 2009] targets the prediction of MT quality in the absence of reference translations. To be more precise, given features extracted from a source unit and its corresponding MT unit, completed with features related to the translation process itself, a trained QE model predicts an automatic metric score for this MT unit.[4]

Various source units, such as the document, paragraph, sentence and word have been studied

[3] Back-translation consists in automatic translation of MT back into the source language.

[4] We distinguish QE from Confidence Estimation (CE), which is a term often used to refer to QE [Specia and Giménez, 2010]. We consider CE to be the system internal assessment of its output [Ueffing and Ney, 2005; de Gispert et al., 2013], as opposed to an external assessment provided by automatic evaluation or predicted by QE. CE results can be used as input features in QE.

in QE for some years now. Phrase-level QE has become a focus of research only recently [Logacheva and Specia, 2015].

QE features are traditionally characterized as *black-box* (system-independent) or *glass-box* (system-dependent, extracted from the translation process): e.g., at the word level we can distinguish the POS and the lemma of a word as system-independent features, and the system posterior probability of producing a certain word in a certain position (a measure of Confidence Estimation) as a glass-box feature. Recently, various neural features (e.g., word embeddings) have raised a lot of interest [Abdelsalam et al., 2016; Kim and Lee, 2016; Blain et al., 2017]. Pseudo-reference scores (comparisons of outputs of different systems for the same input sentence) can be important features as well [Bojar et al., 2016c].

Whereas at the document, paragraph or sentence levels QE predicts automatic scores (e.g., BLEU, TER, etc.), at the word and phrase levels predictions are often binary: *OK* or *BAD* [Bach et al., 2011; Bojar et al., 2015, 2016a]. This is because sentence-level QE scores help rank sentences that are worth post-editing, while word-level QE aims to spot words that need to be changed during PE.

For the document, paragraph and sentence levels QE models are trained using various regression algorithms: e.g., Support Vector Machines (SVMs) [Cortes and Vapnik, 1995], Multilayer Perceptron [Rosenblatt, 1957] and Gaussian Process [Rasmussen and Williams, 2005]. For the word and phrase levels algorithms such as Conditional Random Fields (CRFs) [Lafferty et al., 2001] or Random Forests [Breiman, 2001] are used. Training data can be labeled using automatic metrics and references, binary labels for word- and phrase-level QE are produced using TER or METEOR alignments.

QE results are actively used for PE. They indicate if an MT unit is "worth post-editing". In general, sentence-level QE was shown to be the most useful [Bertoldi et al., 2013b; Hunsicker and Ceausu, 2014; Turchi et al., 2015]. For instance, Turchi et al. [2015] demonstrate that sentence-level QE is efficient for medium-size sentences and for high-quality MT. The practical usefulness of word-level QE is more arguable, and some studies have even found that it could be distracting for post-editors [Raybaud, 2012; Alabau et al., 2013; Sanchis-Trilles et al., 2013].

Alternatively, QE can be source-side. Here, target-side quality indicators can be projected onto source units to reveal their translation quality, also referred to as translation difficulty in this context. This **source-side QE** approach has received little attention [Mohit and Hwa, 2007; Cheng et al., 2016].

Mohit and Hwa [2007] introduce the notion of subsentential *translation difficulty* as a measure

relative to an MT system. They cast the task of detecting difficult segments as a binary classification task (easy or difficult to translate) at the phrase level (syntactically-motivated segments). A phrase is marked as difficult-to-translate if the removal of its translation from a hypothesis has a positive impact on the resulting document-level BLEU score (calculated against a correspondingly modified reference). Mohit and Hwa [2007] consider parse tree constituents whose string span is between 25% and 75% of the full sentence length, and use SVMs as the classification algorithm.

Cheng et al. [2016] consider a source phrase as difficult-to-translate, if it is translated "incorrectly" and if its "correct" translation crucially improves the translation of the rest of the sentence. The authors study difficulties at the level of SMT phrases (see section 2.2). A phrase is marked as difficult-to-translate if, after constraining its translation, the quality of a re-translated sentence (measured in BLEU) is significantly improved as compared to results obtained for other phrases of this sentence. The authors experiment with Maximum Entropy classification models [Ratnaparkhi, 1997], SVMs and FFNNs [Rosenblatt, 1957]. They show the best performance for the last model.

3.2 Interactive Machine Translation

Interactive Machine Translation (IMT) anticipates the sequence that will be typed in by the user (the *suffix*, denoted $\mathbf{e}_{sf}$), knowing the already validated part of $\mathbf{e}$ (the *prefix*, denoted $\mathbf{e}_{pr}$). IMT systems typically have to work with real time constraints. The prediction task can be formulated using again the noisy channel approach:

$$\mathbf{e}^1 = \underset{\mathbf{e}}{\operatorname{argmax}}\, p(\mathbf{f}|\mathbf{e}_{pr}, \mathbf{e}_{sf})\; p_{LM}(\mathbf{e}_{sf}|\mathbf{e}_{pr}) \tag{3.1}$$

The first IMT systems, e.g, TransType [Langlais and Lapalme, 2002], used a very simple model where the TM predicts only one word at a time (for word-based SMT). Longer suggestions were taken from user lexicons. Suggestions were usually displayed in a drop-down list after each sequence typed in by the user (see Figure 3.2).

The web-based Caira system [Koehn, 2009] started a next generation of IMT using PBSMT and made translation suggestions at the phrase level.

Modern IMT is more sophisticated and does predictions by choosing corresponding partial hypotheses from decoding search graphs or by means of *prefix-constrained decoding*, whereby a

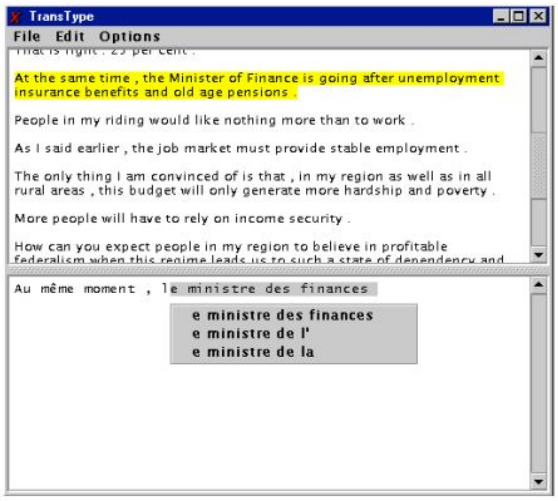


Figure 3.2 – Illustration of the translation process in the TransType system (reproduced from [Langlais and Lapalme, 2002])

new partial hypothesis is regenerated each time the prefix is $\mathbf{e}_{pr}$ is revised or extended.

The first solution matches $\mathbf{e}_{pr}$ with a partial path in the search graph until a node a is reached.[5] Outgoing paths from node a are proposed as suggestions. In the cases where $\mathbf{e}_{pr}$ does not exist in the MT search space, the best partial path until a is selected based on the shortest string edit distance between this path and $\mathbf{e}_{pr}$ [Och et al., 2003].

This approach was extended by Koehn et al. [2014] to introduce an improved case-insensitive, stemmed and approximate matching, and to give more importance to matching the last word in $\mathbf{e}_{pr}$.

Regeneration of new hypotheses by prefix-constrained decoding is usually optimized to satisfy real-time requirements: e.g., Bender et al. [2009] regenerate the search graph only if an existing graph is not able to produce a completion; Wuebker et al. [2016] and Ortiz-Martínez et al. [2009] align $\mathbf{e}_{pr}$ to $\mathbf{f}$, so that only unaligned words are re-translated.

With the change of paradigm, IMT has been re-implemented for NMT. The standard objective during NMT decoding is to predict one word at a time given the previously generated words [Sutskever et al., 2014]. Thus, the model naturally accommodates prefix-constrained decoding: $\mathbf{e}_{pr}$ can be directly used to condition the next steps of the decoding process.

Predictive NMT is in general more accurate than predictive PBSMT [Knowles and Koehn,

[5]Here, for computational reasons, word graphs are typically used. A word graph is a weighted directed acyclic graph, in which each node represents a partial translation hypothesis and each edge is labeled with a word of the target sentence. For a more detailed description of word hypotheses graphs see Ueffing et al. [2002].

2016; Wuebker et al., 2016]. Recently, predictive NMT has been improved to satisfy real-time constraints: Knowles and Koehn [2016] suggest to partially re-translate a source sentence and combine this re-translation with the initial hypothesis.

Figure 3.3 illustrates the predictive procedure of the commercial Lilt[6] IMT system. In this system the presentation of predictive suggestions in a drop-down list (often found to be distracting) was replaced by simply displaying them in a separate line. We can see that the initial MT for the sentence "Spasticity, which is an increase in muscle tone, is the most common difficulty with movement seen in children with cerebral palsy." is too literal.[7] The prefix "*Chez les enfants*" 'In the children' is thus entered by the user, allowing a plausible continuation to be proposed by the system and validated by the user.[8]

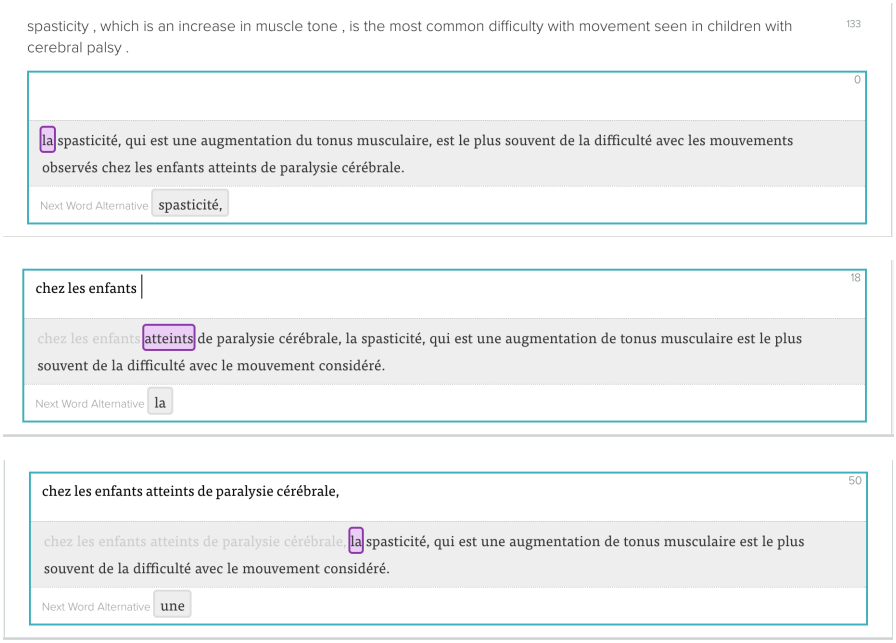


Figure 3.3 – Illustration of predictive translation in Lilt for the prefix "*Chez les enfants*" 'In the children'

Recent studies tend to agree on the fact that IMT increases PE time and that special training

[6] `https://lilt.com/`

[7] "*La spasticité, qui est une augmentation du tonus musculaire, est le plus souvent de la difficulté avec les mouvements observés chez les enfants atteints de paralysie cérébrale.*" 'The spasticity, which is an increase in muscle tone, is the most often of difficulty with the movements observed in the children with cerebral palsy.'

[8] Final hypothesis: "*Chez les enfants atteints de paralysie cérébrale, la spasticité, qui est une augmentation de tonus musculaire est le plus souvent de la difficulté avec le mouvement considéré.*" 'In the children with cerebral palsy, the spasticity, which is an increase in muscle tone is the most often of the difficulty with the movement considered.' Ref.: "*Chez les enfants atteints de paralysie cérébrale, on observe une spasticité, une augmentation du tonus musculaire, qui est la difficulté la plus fréquente lorsqu'ils bougent.*" 'In the children with cerebral palsy, we observe a spasticity, an increase in muscle tone, which is the most frequent difficulty when they move.'

and motivation are necessary to make IMT more efficient [Green et al., 2014; Sanchis-Trilles et al., 2014; Underwood et al., 2014]. Nevertheless, IMT can increase the usability of PE for online adaptation of MT models. For instance, Green et al. [2014] compare classical PE to interactive PE for two language pairs (French-English, English-German) and three domains. Experiments are performed with professional translators. Interactive PE was found to be 20% slower than PE. But with interactive PE subjects produced translations that were closer to refined MT suggestions, and hence more useful for online updates.

3.3 Exploitation of Human Knowledge

Once PE is finished, an MT system receives the feedback and exploits it: performs an online adaptation of its models according to the user new inputs. In the case of SMT this adaptation can be performed to model weights, as well as to the model itself (see section 3.3.1). The next sections will briefly describe technical details of those updates that vary depending on the amount of feedback. We will also present ways of exploiting PRE feedback in SMT. Then, we will turn our attention to the problem of compatibility of those updates with real-life PE scenarios. We will discuss the issue of the permanency of updates and of the selectiveness towards this feedback (see section 3.3.2). We will also describe some scenarios of using online updates for domain adaptation (see section 3.3.3), as well as the benefits of performing online updates to an automatic PE system, rather than to an MT system (see section 3.3.4).

3.3.1 Online Adaptation

The adaptation of system weights is usually performed by means of *online learning* (OL) mechanisms. OL, as opposed to *batch learning*, updates models on a per-example basis instead of going through the whole example set. OL takes the following steps:

1. OL receives an instance;
2. predicts its label;
3. receives the true label;
4. performs an update of the model.

Such online training schedules perfectly fit into the PE scenario, where the human-machine interaction usually happens on a per-sentence or per-document basis:

1. MT system translates a sentence/text;
2. the user post-edits the MT output;
3. the corrected MT is sent back to the system;
4. the system updates its models accordingly.

Online Adaptation of Weights

Online adaptation of weights adjusts the weights of a model, seeking to increase the score of hypotheses that are close to the post-edited translation, and accordingly to reduce the score of competing outputs. Various *online algorithms* are used for this purpose.

The most common online algorithm is the *Simple (Single-layer) Perceptron* [Rosenblatt, 1957]. The basic learning mechanism is as follows: if $\mathbf{e}^1$ is not the closest hypothesis to the reference from the n-best list ($\mathbf{e}^r$, as estimated by automatic metrics, see section 2.7), we update the weight of each feature with the difference between $\mathbf{e}^r$ and $\mathbf{e}^1$ feature values:

$$\vec{w} \leftarrow \vec{w} + \vec{\phi}(\mathbf{f}, \mathbf{e}^r) - \vec{\phi}(\mathbf{f}, \mathbf{e}^1), \tag{3.2}$$

where $\vec{\phi}(\mathbf{f}, \mathbf{e})$ is an arbitrary feature vector quantifying various aspects of the relationship between $\mathbf{f}$ and $\mathbf{e}$.

Another online algorithm widely used in the MT community is the *Margin-infused relaxed algorithm* (MIRA) [Crammer et al., 2006; Watanabe et al., 2007]. It aims to minimize the *structured hinge loss* function:

$$\ell(\vec{w}) = \max_{\mathbf{e} \in \varepsilon} \left[\Delta(\mathbf{e}) + \vec{w}\big(\vec{\phi}(\mathbf{f}, \mathbf{e}) - \vec{\phi}(\mathbf{f}, \mathbf{e}^r)\big) \right], \tag{3.3}$$

where $\Delta(\mathbf{e}) = BLEU(\mathbf{e}^r) - BLEU(\mathbf{e})$ is the cost of choosing $\mathbf{e}$ instead of $\mathbf{e}^r$, usually measured in BLEU.

The loss is 0 only if $\vec{w}$ separates each $\mathbf{e}$ in the space of translation hypotheses from $\mathbf{e}^r$ by a margin proportional to their difference in BLEU: $\vec{w} \cdot \vec{\phi}(\mathbf{f}, \mathbf{e}^r) > \vec{w} \cdot \vec{\phi}(\mathbf{f}, \mathbf{e}) + \Delta(\mathbf{e}), \forall \mathbf{e} \in \varepsilon$.

As the BLEU metric is not adapted to the evaluation of isolated sentences, either the score is computed with $n = 1$ or a smoothing technique is used [Lin and Och, 2004; Chen and Cherry, 2014] (see section 2.7).[9]

[9] As BLEU computes a geometric mean of n-gram precisions, if a higher order n-gram precision (e.g., $n = 4$) of a sentence is 0, then the BLEU score of the entire sentence is 0, even if some lower order n-grams are matched.

At each step s an update makes the smallest change to $\vec{w}$ (subject to the regularization parameter C) that separates $\mathbf{e}^r$ from negative hypotheses.

Let $\mathbf{e}'$ be a "fear" hypothesis that maximizes $\ell(w)$. The update is then performed as follows:

$$\vec{w}_{s+1} = \vec{w}_s + \eta_s\big(\vec{\phi}(\mathbf{f}, \mathbf{e}^r) - \vec{\phi}(\mathbf{f}, \mathbf{e}')\big), \tag{3.4}$$

where

$$\eta_s = \min\Big[C, \frac{\ell(\vec{w}_s)}{\|\vec{\phi}(\mathbf{f}, \mathbf{e}^r) - \vec{\phi}(\mathbf{f}, \mathbf{e}')\|^2}\Big] \tag{3.5}$$

MIRA and Perceptron variants are widely used for online weight updates by different SMT systems [i.a., Martínez-Gómez et al., 2011; Mathur et al., 2013; Denkowski et al., 2014].

Other online algorithms used for updates include *discriminative ridge regression* [Martínez-Gómez et al., 2012], *online subgradient AdaGrad* [Wuebker et al., 2015], etc. Their description and comparison of their characteristics are beyond the scope of this work.

However, updates of weights do not add new information to the search space, therefore they are usually combined with updates of such SMT components as the TM and the LM.

Similar to the standard TM training procedure (see section 2.3), TM updates are preceded by the word alignment procedure between the source and the newly post-edited target.

Online Word Alignments

To satisfy real-time constraints, the word alignment procedure for feedback data is commonly performed using online EM procedures (as opposed to the time-consuming batch EM word alignment) [i.a., Neal and Hinton, 1998; Cappé and Moulines, 2009]. For instance, the commonly used *stepwise online EM* [Cappé and Moulines, 2009] gathers statistics over new data and interpolates them with existing statistics until stabilization.

Forced alignment is another solution to the on-the-fly word alignment. Here old models are used to align new data. This procedure does not collect new statistics and does not improve old models. In practice it is efficient to align a small quantity of new data, since statistics computed on a small amount of data are not likely to influence the previous distribution. For instance, in an online scenario Denkowski et al. [2014] use this procedure to align a post-edited sentence to the corresponding source.

A more efficient alternative to standard forced alignment approaches can be, for instance, an online version of the associative sub-sentential `Anymalign` alignment method [Lardilleux et al.,

2012; Gong, 2014], which performs better for the alignment of rare words. This method relies on comparisons of source and target word occurrence distributions over randomly sampled sub-corpora.

Other online alignment solutions include, for example, obtaining the alignment $\mathbf{f} \rightarrow \hat{\mathbf{e}}$ directly from the 3-way alignment $\mathbf{f} \rightarrow \mathbf{e}^1 \rightarrow \hat{\mathbf{e}}$. Knowing the alignment between $\mathbf{f}$ and $\mathbf{e}^1$, in most of the cases produced by the decoder, the alignment $\mathbf{e}^1 \rightarrow \hat{\mathbf{e}}$ between the initial hypothesis and the post-edited translation is computed as a by-product of automatic evaluation (e.g., using METEOR, TER, etc.) [Blain et al., 2012].

Online Adaptation of the TM

As soon as a word alignment is produced, the TM can be updated. However, recalculation of the probabilities for the updated training corpus would be time-consuming and would not fit the online update scenario. Consequently, specific techniques are used. Their implementations depend on the update frequency:

- updates can be performed after each post-edited sentence. They target reducing the burden of repetitive changes within a document. For such updates the models should be manipulated with the help of efficient data storage structures to ensure the update speed (measured in seconds);
- updates can also be performed when a more significant amount of post-edited data becomes available (one or more documents). These updates target adapting a system to new test data. They are usually launched as background processes while post-editors are still working. In this case, standard models are updated using special techniques.

For both scenarios, a common practice consists in maintaining two models: a smaller model for immediate updates according to human feedback that is paired to a larger general model. Direct updates to general models are more rare and increase the risk of degrading the quality of the initial MT system.

Immediate Online Adaptation of the TM One of the most commonly used data structures for per-sentence updates is the *suffix array* (SA) [Manber and Myers, 1990] data structure (see Figure 3.4). The structure indexes source and target data, and their alignment. An SA over the source corpus $f_1, \ldots, f_T$ is an array $1, \ldots, T$ of all the token positions in the corpus. Once alphabetically sorted it serves to efficiently find all occurrences of $\bar{f}$ with a complexity of

$O(2\log(F))$. The indexed alignment serves to extract bi-phrases and estimate their probabilities. However, for frequent source phrases this process still risks to be time-consuming. To overcome the issue occurrences of $\bar{f}$ can be sampled [Callison-Burch et al., 2005; Lopez, 2008b].

Other data structures used for efficient TM manipulation include, for example, the *asymmetrical double trie* structure [Ortiz-Martínez, 2011].

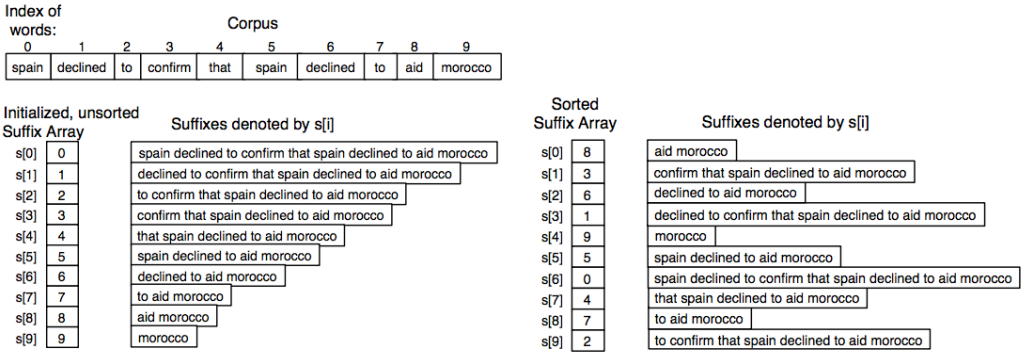


Figure 3.4 – Illustration of a SA for a very small corpus (reproduced from [Callison-Burch et al., 2005])

The common architecture of the TMs for online updates includes two structures: one that indexes the initial training data and another one that stores the feedback data [Denkowski et al., 2014; Germann, 2015]. For instance, Denkowski et al. [2014] maintain a dynamic lookup table, storing bi-phrases and counts generated from user feedback. A new sentence is translated using a sentence-specific TM. This TM is generated by sampling from a static SA and updated according to the lookup table.

Updates can also be performed directly to general models using, for example, the *dynamic SA* structure [Salson et al., 2010], which allows deletions and insertions to a statically computed SA [Levenberg et al., 2010], or to counts stored in a trie [Ortiz-Martínez, 2011].

Delayed Adaptation of the TM Delayed TM updates are performed when a larger amount of post-edited data (usually, one or more documents, in a production scenario, a day of PE) becomes available. Their main purpose is to adapt a system to new test data.

Here, again, for safer updates, a smaller model is created for human feedback. This model can be given priority over a general model: for instance, the adapted model is consulted first, and the general model is only queried for bi-phrases that are not observed in the former model.[10]

[10]This procedure is implemented by the `back-off` mode of the `Moses` toolkit.

The models can also compete. In this case their respective weights are set during the standard training procedures (log-linear combination of TMs) [Hardt and Elming, 2010; Blain et al., 2012; Gong et al., 2014].

Adapted and general TMs can also be merged by means of the `fill-up` technique [Nakov, 2008; Bisazza et al., 2011]. In this case only the entries from an adapted model absent from a general model are added to the latter. Entries of this merged TM are additionally marked with the provenance feature.

The LM adaptation is rarer than the TM adaptation and is less important. Adding a previously unknown word to a TM will have a greater impact on MT than adding it to an LM. In the case of the immediate LM adaptation corresponding statistics are stored in the memory and updated accordingly with feedback [Ortiz-Martínez, 2011; Denkowski et al., 2014]. For delayed updates, a linear interpolation of the scores of an adapted LM and a general LM is performed. However, such an interpolation requires optimization of mixture weights [e.g., Koehn and Schroeder, 2007].

The above-mentioned implementations consider the necessity of permanent changes to models. Often they additionally apply certain strategies to select the feedback crucial for improving MT quality (see sections 3.3.2, 3.3.3). In contrast, the cache-based update approach described below considers feedback to be subjective or document-dependent and implements task-specific models.

Cache-Based Models The *cache* technology is commonly used in Computer Science. Its main principle is to store recently accessed data in a separate storage with easier access to save time for future requests. A cache systematically filters out data that are no longer in use (according to the age parameter, decaying factor, which is the preferred term in NLP [Clarkson and Robinson, 1997]).

Cache-based models were introduced to MT by Nepveu et al. [2004]. The main motivation is the fact that people tend to use a limited vocabulary in speech and writing. Cache-based models are "single-use" models created for a specific user or a specific task. They use a limited amount of features (usually only the decaying factor) and are usually combined with static models [Bertoldi et al., 2013a]. Cache models are particularly beneficial for the translation of repetitive texts (manuals, instructions, etc.).

Bertoldi et al. [2013a] also introduce a cache LM updated with n-grams that contain at least one content word. This LM rewards target n-grams in MT output if they were previously present

in user feedback.

Constrained Decoding As for the exploitation of pre-translation feedback, the most straightforward solution consists in constraining the PBSMT decoder segmentation and providing translations for target phrase (see section 2.6) [Mohit and Hwa, 2007; Cheng et al., 2016]. An implementation of this procedure is proposed by the `exclusive xml-mode` of the `Moses` toolkit.[11] It allows to mark a segment with its desirable translation: e.g., `<np translation="Le volume LO">LE volume</np> was not significant`. However, this method is very restrictive and does not use potentially good translations from PT bi-phrases ($\bar{f}$, $\bar{e}$) that overlap with the marked span: e.g., an overlapping bi-phrase for the running example "LE volume was ||| *le volume LO était*".

Note that similar constrained decoding solutions, guaranteeing the presence of lexical constraints in output, was recently proposed, for instance, by Hokamp and Liu [2017] and Chatterjee et al. [2017b] for NMT.

Alternatively, any other online method from the methods above can be used to update models before translation will take place. The usage of those methods is less restrictive, since, for instance, for cache-based models, the decoder will have freedom to use either bi-phrases from the dynamic TM or bi-phrases from the static TM prioritized by the cache LM. As source segments may repeat but be pre-translated or not, source tokens can be made unique for more precise updates.

In general, the success of online updates depends on the domain (they are more efficient for repetitive texts, for instance, manuals, instructions, etc.), as well as on the type and the parameters of updates [Bertoldi et al., 2013b; Cettolo et al., 2013].

3.3.2 Selectiveness towards Human Feedback (Active Learning)

The influence of additional data on MT quality is a well-studied issue [Turchi et al., 2008; Gascó et al., 2012; Haddow and Koehn, 2012]. In the absence of real-time constraints, the utility of new data can be verified through the time-consuming process of system re-training and testing. Regular minor online updates make such verifications almost impossible. In a real-life setting, the

[11] `http://www.statmt.org/moses/?n=Advanced.Hybrid#ntoc1`;The `inclusive xml-mode` mode allows PT entries to compete with specified pre-translations. This configuration accepts pre-defined probabilities for provided translations and leaves the final translation choice to the target LM. However, correct values of probabilities are hard to configure and competing is not necessarily the best option if the presence of pre-translation should be guaranteed.

process risks to become even more complex with the presence of multiple post-editors providing their feedback at the same time.

One of the approaches that addresses the issue is **Active Learning** (AL) [Settles, 2009]. It relates a group of methods for choosing the training samples that are most likely to improve a system. The initial goal of these techniques is to save human effort by proposing to annotate less data, for MT followed by PE the focus is usually switched to being selective towards the data used for updates.

AL is commonly applied to SMT at the sentence level. The main intuition is to choose the most informative sentences for updates that contain a maximum amount of new information [Eck et al., 2005; Haffari and Sarkar, 2009; Haffari et al., 2009; González-Rubio et al., 2012; Du et al., 2015]. For instance, Eck et al. [2005] weight sentences according to the quantity of previously unseen frequent n-grams:

$$w(\mathbf{e}) = \sum_{n=1}^{N} \Big[\sum_{\text{unseen } n\text{-gram}} fr(n\text{-gram}) \Big] \tag{3.6}$$

Other sentence-level AL strategies measure informativeness using various n-gram related scores, MT confidence scores, sentence perplexity values (the degree of uncertainty as measured by an LM),[12] representativeness scores (the rationale behind is that only sentences, which are "representative" of the underlying distribution are useful for updates) [González-Rubio et al., 2012; González-Rubio and Casacuberta, 2014; Du et al., 2015], etc.

More traditional cost-sensitive AL for MT searches to maximize quality gains while minimizing the user PE effort [Bloodgood and Callison-Burch, 2010; González-Rubio and Casacuberta, 2014]. For instance, Bloodgood and Callison-Burch [2010] propose that users only translate frequent n-grams not covered by the training data.

Selectiveness can also have as its goal to reduce feedback noise, which can be present, for instance, in a multiple-user PE scenario, where different post-editors can not be equally trusted.

To address this issue and mitigate feedback bias, Mathur et al. [2014] propose to use *multi-task learning* (MTL) [Cavallanti et al., 2010]. MTL training is done for several tasks X with the goal to improve generalization over all of them by exploiting potential relations between them. We will explain the process for the Perceptron algorithm (Equation 3.2). The overall goal of MTL is to learn the X weight vectors simultaneously, one for each post-editor in an online fashion. Weight update for a post-editor x after each feedback is performed as follows:

[12] The intuition behind those strategies is obvious: the model can learn a lot from correct labels of samples in whose labels it is uncertain.

$$\vec{w}_x \leftarrow \vec{w}_x + \vec{w}_x\vec{\phi}_x(\mathbf{f}, \mathbf{e}^r) \cdot (A \otimes L_Y)^{-1}_{x,*} \cdot \vec{\Phi}_x(\mathbf{f}, \mathbf{e}^1), \tag{3.7}$$

where

$$\vec{\Phi}_x(\mathbf{f}, \mathbf{e}^1) = (\underbrace{0, \cdots, 0}_{\text{(x-1)Y times}}, \vec{\phi}_x(\mathbf{f}, \mathbf{e}^1), \underbrace{0, \cdots, 0}_{\text{(X-x)Y times}}) \in \mathbb{R}^{XY}, \tag{3.8}$$

where $A \otimes L_Y$ is the Kronecker product of the interaction matrix A^{-1} of dimensions $X \times X$ and the identity matrix L_Y of dimensions $Y \times Y$. A^{-1} defines relations between post-editors. These relations can be defined by ranking or PE similarity scores.

3.3.3 Domain Adaptation

The problem of adapting to human feedback can also be viewed as a problem of *Domain Adaptation* (DA). DA chooses data that will help a system to resolve the mismatch between its training data and test data. This mismatch can be domain-related (e.g., news → medical), style-related (e.g., medical patent → medical research review), register-related (e.g., medical popular scientific style → medical scientific style), etc., or any combination of these.

For static models the task consists in efficiently choosing and integrating new domain-specific data with current models [i.a., Daumé III and Jagarlamudi, 2011; Banerjee et al., 2012; Sennrich, 2012; Carpuat et al., 2013; Cuong and Sima'an, 2014; Chen et al., 2016].

Within a collaborative scenario, DA is used to adapt a general-domain MT system to project-specific translation when some post-edited data are available [Gong et al., 2012; Cettolo et al., 2014; Blain et al., 2015; Wuebker et al., 2015]. This is a typical scenario for delayed updates. For instance, Cettolo et al. [2014] perform domain-specific data selection using the traditional cross-entropy method proposed by Moore and Lewis [2010]. In language modeling cross-entropy is defined as p_{LM} averaged per word:

$$H(\mathbf{e}) = -\frac{1}{M} \log p_{LM}(\mathbf{e}) \tag{3.9}$$

The cross-entropy difference is defined as follows:

$$S(\mathbf{e}) = |H(\mathbf{e})_{specific} - H(\mathbf{e})_{general}| \tag{3.10}$$

This measure compares cross-entropy values of a general-domain LM and an in-domain LM.

Cross-entropy can also be computed on both sides of a bitext to perform bilingual filtering simultaneously [Axelrod et al., 2011].

Another type of DA applied in collaborative scenarios is *Multi-Domain Adaptation* (MDA) [i.a., Sennrich et al., 2013; Huck et al., 2015; Chatterjee et al., 2017a]. MDA attempts to create a system that will be able to adapt itself to a new domain on-the-fly by choosing domain-relevant information for decoding. The rationale behind creating such models is the absence of the information on the variety of domains presented in test data during the development phase, as well as the practical issue of manipulating multiple domain-adapted models at once. For instance, Sennrich et al. [2013] implemented MDA by delaying the feature computation to the decoding phase. A TM stores a vector of several domain-specific tables, each containing sufficient statistics for feature estimation. Domain labeling of the training, development and test data is performed by means of unsupervised clustering.

3.3.4 Automatic Post-Editing

Another solution to exploit human feedback consists in online adaptation of *Automatic Post-Editing* (APE) systems instead of MT models [Simard and Foster, 2013; Lagarda et al., 2015; Chatterjee et al., 2016, 2017a]. APE seeks to automatically correct errors in MT before it is presented to the user. Motivations to use APE systems are diverse: the fact that those systems have access to new information not available to an MT system and are able to perform more direct changes [Parton et al., 2012], they are also lighter and can be more easily adapted to a style or a domain [Chatterjee et al., 2017a].

APE can be SMT-based ("translation" from MT into PE) [Simard et al., 2007a,b], rule-based [Rosa et al., 2012] or neural-based [Pal et al., 2016, 2017].

For instance, Chatterjee et al. [2017a] simultaneously learn several domain-specific APE models from user feedback. The authors use a domain-aware sampling technique to build per-sentence APE models for each new MT output. When no relevant data is available to build a model, MT is not corrected. User feedback is exploited to update the rules containing the information on MT bi-phrases and their corrections. Those rules have the following form: ($\bar{f}\#\bar{e}$, PE). A dynamic knowledge base that stores positive (correct application of a rule) and negative (application of a rule in a wrong context, the rule not used by the decoder) statistics seeks to increase the reliability of APE modifications.

In our mind, APE scenarios are more suitable for production scenarios with experienced professional post-editors, who perform reliable corrections.

3.4 Computer-Assisted Translation Systems

All the components of the human-machine collaboration described above (PE, PRE, IMT, online updates) are commonly accommodated by **Computer-Assisted Translation** (CAT, also called machine-assisted, or machine-aided translation) environments. The primary goal of those environments is to help the human during the process of translation. Thus, the key component of such systems is a user-friendly interface.

Figure 3.5 shows an example of the translation process in the MATECAT tool. The interface is plain, giving easy access to useful functionalities: e.g., change of case, search, access to external help (glossaries, translation memory suggestions), etc. [Cattelan, 2014].

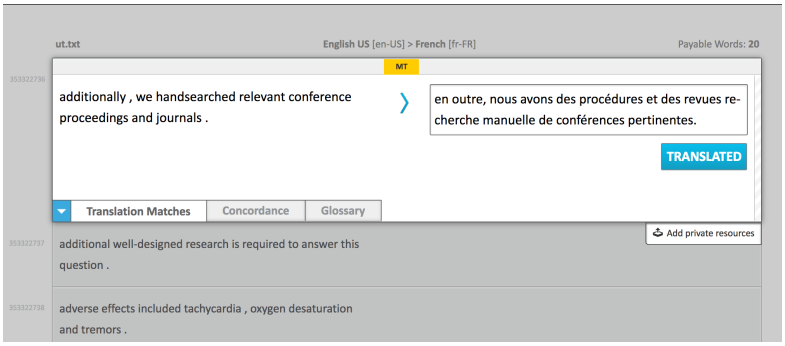


Figure 3.5 – Illustration of the translation process in the MATECAT tool

Other CAT functionalities provide:

- PRE help: spell and grammar checkers; CL suggestions, paraphrase suggestions [Seretan et al., 2014; Wu et al., 2016].
- PE help: access to terminology databases [Sheremetyeva, 2014], electronic dictionaries, Internet searches, translation memory matches,[13] concordancer searches,[14] etc.; visualization of MT-related procedures: e.g., word alignments, n-best translation hypotheses [Koehn et al., 2015], etc.
- Optimization of editing: an *e-pen* can be used to post-edit translations by means of proofreading gestures (to imitate such actions as delete, move, substitute, etc.); voice commands can be used as well [Alabau and Leiva, 2014].

[13] Translation memories offer previously translated/post-edited source-target units as suggestions. Those suggestions are based on the source similarity.
[14] Concordancers provide users with translation examples from a specified corpus.

- Extensive user action logging: it includes keeping track of all the user actions (both mouse or keyboard events) and eye-tracking. Recorded actions can be replayed [Koehn et al., 2015]. This logging is very important for the evaluation of the PE productivity.
- Reviewing aid: human reviewers are provided with automatic help that detects added or missing content, terminology inconsistency [Koehn, 2016].

3.5 Summary

This chapter has described the main aspects of human-machine collaboration within MT: we have presented the principles of human knowledge injection into MT, as well as the means of exploiting the obtained human knowledge by MT systems. In our descriptions we focused our attention on the peculiarities of this collaboration for SMT.

Human intervention into MT commonly takes place at the end of the process, and optionally at the beginning of it. The former intervention consists in correcting MT output, and is referred to as **Post-Edition** (PE); the latter consists in normalizing the input in order to prepare it for MT, and is referred to as **Pre-Edition** (PRE).

Regarding PE types, surface PE and full PE are distinguished. Hypothetically, the first type of PE is supposed to make MT understandable and the second type is supposed to bring MT to a publishable quality. In practice, the amount and quality of corrections largely depends on the experience of a post-editor and his or her level of training. Non-professional post-editors tend to make wrong lexical choices and need to be helped by professionals. Recent research tends to show that PE is less laborious for the human as compared to purely human translation, under the condition of a certain level of the corrected MT quality. However, PE still risks to become tiresome and frustrating.

PRE in SMT is meant to perform source changes aiming to make the test samples more alike to the training data, up to reducing source readability. It can consist in linguistically-motivated changes (text normalization, simplification etc.), which may require costly creation of corresponding rewriting rules, as well special training for people, who would use them.

PRE can also consist in a non-selective automatic paraphrasing or pre-reordering (modification of the source word order to make it closer to the one of the target language), but such strategies are risky and can infer both positive and negative influence on the final MT quality. In general, such monolingual help does not secure “correct” lexical choices in output, which is guaranteed by a PRE strategy involving pre-translation.

Quality Estimation (QE) predicts MT quality. Target-side QE is able to successfully indicate to post-editors which MT suggestion is "worth post-editing".

In targeted PRE scenarios, source-side QE techniques can help with detection of "badly" translated source segments, often referred to as *translation difficulties*. As soon as those difficulties are resolved by the human, the machine exploits the information to improve its second-pass MT output, thus reducing the final PE effort.

Interactive Machine Translation (IMT) also seeks to help the human during PE and anticipates the sequence that will be typed in. Predictions are often taken from available decoding search graphs or generated by *prefix-constrained decoding*. The performance of IMT depends on the experience of a post-editor, and tends to increase PE time. The PE produced with IMT is closer to MT and can be more beneficial for online updates of MT models.

As soon as human feedback becomes available, a system adapts itself to this feedback in an online fashion. Within the SMT paradigm updates are performed to model weights and to the model itself. Implementations of online adaptation depend on many factors, including the update frequency. A standard TM update solution includes two models: one dynamic, constantly updated from feedback; another one general, built from the initial corpus. **Cache-based** models provide temporary user- or text-specific models that do not bias general models. Pre-translation feedback can be integrated by restricting segmentation and translation options, or by any other means of online updates that are applied before the actual MT will take place.

Online update solutions are usually selective towards user feedback in order to reduce the risk of disturbing the initial MT quality. Strategies of selection are studied by **Active Learning** (AL). AL targets the acquisition of new information (OOV, unknown n-gram, etc.) useful to improve MT quality. In a similar way, **Domain Adaptation** (DA) selects feedback data specific to a certain domain.

Bilingual translation can be complemented by monolingual **Automatic Post-Editing** (APE). It seeks to capture PE regularities. APE systems perform as it were, a second pass "translation" of MT output, and can be also updated in an online fashion. APE solutions better fit production scenarios with experienced professional post-editors.

Modern **Computer-Assisted Translation** (CAT) environments, like CASMACAT [Koehn et al., 2013] and MATECAT [Federico et al., 2013], efficiently accommodate PE, PRE, IMT and ensure online updates. They also provide a user-friendly interface and extensive logging to better analyze user productivity. They offer various kinds of PE help and display results of internal MT procedures.

In this work, we decided to focus our attention on PRE scenarios that are able to accommodate detection and resolution of translation difficulties by pre-translation. These scenarios better fit production contexts, where both professionals and non-professionals participate in PE and the lexical consistency of final translations is the primary goal. So that, pre-translation can be ensured by few professionals, who will have control over lexical choices in MT output. Finally, the amount of PE left to non-professionals will be reduced due to improved MT output, leading to less frustration and less risk of newly-introduced errors. These improvements are reached as a result of "correct" pre-translations in output, but also as a result of potential improvements to the automatic translation of neighboring words ("free" automatic corrections). Additionally, the studied scenarios are relatively low-cost and easy to implement.

Contrarily to the target-side QE, the source-oriented QE context permits multi-target and multi-source experiments: we thus consider the latter approach as more practical, as well as a nice testbed for studying source-language and target-language effects.

In terms of the exploitation of PRE feedback, we mostly target the resolution of difficulties at the document level in a batch setting mode. The final PE stage is postponed. We thus will explore delayed update solutions for the TM.

4 | Diagnosing High-Quality Statistical Machine Translation within the Cochrane Context

Contents

MT evaluation is a necessary process as its results are important for MT system developers, users and post-editors.

MT evaluation can be performed by humans, who judge it by its fluency and adequacy (see section 4.1). This evaluation can consist in giving a general rank or score ("holistic" evaluation), or giving local judgements using an error typology. In the latter case it is usually referred to

as *error analysis*. However, human evaluations are subjective and costly, especially during MT development, when quality has to be verified systematically. The evaluation process is thus usually fully or partially automated (see section 4.2).

The second part of the chapter will present the context of our work with the MT of Cochrane systematic review abstracts (see section 4.3). Publishing and translating these abstracts is a part of the Cochrane organization mission, whose main goal is to spread medical knowledge. Cochrane implements an MT context, where PE is performed by both professional and lay post-editors with the latter being mostly professionals in the medical domain. This context is a perfect fit to implement a PRE scenario involving detection and resolution of translation difficulties.

In particular, we will describe the Cochrane English-French corpus of medical research reviews, the related translation context (see sections 4.3.1 and 4.3.2) and a high-quality English-French SMT system developed for this context (see section 4.3.3). We will finally describe our fine-grained methodology for diagnosing high-quality MT (see section 4.3.4), as well as the results of applying it to the Cochrane SMT (see section 4.3.5).

4.1 Human Evaluation and Error Analysis

Humans usually judge translations by their *fluency* and *adequacy*, defined as follows:

- *fluency* evaluation involves a judgment about how a hypothesis satisfies the grammatical and lexical norms of a language;
- *adequacy* evaluation involves a judgment about how well a hypothesis conveys the meaning of the original sentence.

Different ways to collect those evaluations were proposed over the years [White, 1994; Eck and Hori, 2005; Koehn and Monz, 2006]: for instance, ranking of sentences for both fluency and adequacy on five-point scales during the first two WMT campaigns. A recent improvement to the procedure was proposed by Graham et al. [2015] under the name *Direct Assessment*. It includes monolingual evaluation of translation fluency and adequacy as separate tasks. For instance, adequacy assessment can be designed as an assessment of similarity of meaning between reference translations and MT hypotheses on a 0-100 rating scale. No reference is displayed for fluency assessment and humans are asked to rate how much they agree that a given translation is fluent [Bojar et al., 2016d]. The design of the last type of assessment, to our mind, reduces

human cognitive effort to a minimum and can also help to mitigate subjectivity, which may introduce more significant bias in bilingual evaluation tasks.

Human MT error analysis usually consists in annotating output using a certain error typology. Usually those typologies of MT errors [Vilar et al., 2006; Bojar, 2011] are based on the main post-edit operations and comprise the following main classes with minor variations:

- *a missing word*, which corresponds to the deletion operation;
- *an extra word*, which corresponds to an insertion;
- *word order*, which corresponds to a shift, including short- and long-distance shifts;
- *an incorrect word*, which corresponds to a substitution, and includes lexical, morphological, stylistic changes, etc.

Those basic typologies are usually extended into more detailed hierarchies. For instance, the Multidimensional Quality Metrics (MQM) framework [Lommel et al., 2014][1] comprises an extended typology of errors and attributes lexical, addition and omission errors to the accuracy issues, and grammatical, style, spelling errors, etc. to the fluency issues.[2] It also includes a hierarchy of *verity* and *design* errors. Verity errors address the errors of text applicability to concrete real-world conditions (e.g., if a manual for an electrical appliance states that a ground wire will be bare copper, this may cause problems because of the difference of the wire color in different countries). Design errors address the issue of text presentation on a page, screen (e.g., highlighting text with different colors that may influence comprehension). Figure 4.1 illustrates the annotation of a hypothesis with MQM: note that in the translation into German the English abbreviation "PHE" (Public Health England) is left untranslated and hence is marked as an "accuracy:untranslated" error.

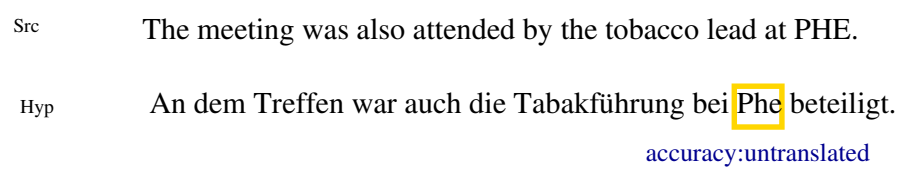


Figure 4.1 – MQM annotation of a hypothesis for the sentence "The meeting was also attended by the tobacco lead at PHE."

[1] A subset of MQM is known as the TAUS Dynamic Quality Framework (DQF, `https://qd.taus.net/`) Error Typology. It is tuned for assessing quality in localization projects.

[2] This is also one of the official metrics of the "QT21: Quality Translation 21" project .

Another way of performing MT error analysis consists in annotating which meaning components are retained in MT and how correctly they are translated (semantics-based annotation) [Lo and Wu, 2011; Birch et al., 2016]. For instance, the semantics-based Human UCCA-based MT Evaluation (HUME) [Birch et al., 2016] involves two steps:

1. Semantic annotation of sentences with the help of the Universal Conceptual Cognitive Annotation (UCCA) representation scheme [Abend and Rappoport, 2013]. Basically, this procedure includes assigning a semantic role to each source word/group of words (examples of roles in UCCA: Process, State, Linker, Participant). Figure 4.2 illustrates a UCCA-annotated hypothesis: input segments are highlighted with different colors, each corresponding to a role. For instance, the segment "the meeting" is highlighted with red, which corresponds to the Participant role.

2. HUME-annotation: human judgements of translation quality for each source semantic unit relative to its aligned MT, where units are defined according to the UCCA annotation. Figure 4.2 shows that the unit "at PHE" was generally adequately translated (marked with `A`), but the untranslated word "PHE" is marked with red as it is incorrectly translated.

Semantics-based annotations are particularly relevant for the types of texts providing precise information, e.g., medical or legal texts. They enable to create a hierarchy of errors according to their influence on text understandability. For instance, for a medical text, translations of source units bearing such roles as Participant or State (in the UCCA scheme) are more important for text comprehension than translation of source units having the role of a link (usually conjunctions).

4.2 Automatic Evaluation and Error Analysis

When quality has to be evaluated systematically, especially during MT development, human evaluations become too much costly, and are replaced by cheap and quick automatic and semi-automatic evaluations.

We have already presented an overview of automatic evaluation metrics, which compare MT output to human-translated (or post-edited) references in section 2.7, and of QE methods that predict results of those automatic evaluations without any references in section 3.1.3.

Monolingual hypothesis→reference alignments produced by some automatic metrics are often taken as the basis for automated error analyses [e.g., Popovic and Ney, 2011; Zeman et al., 2011; Berka et al., 2012]. For instance, Zeman et al. [2011] detect MT errors using an HMM-based

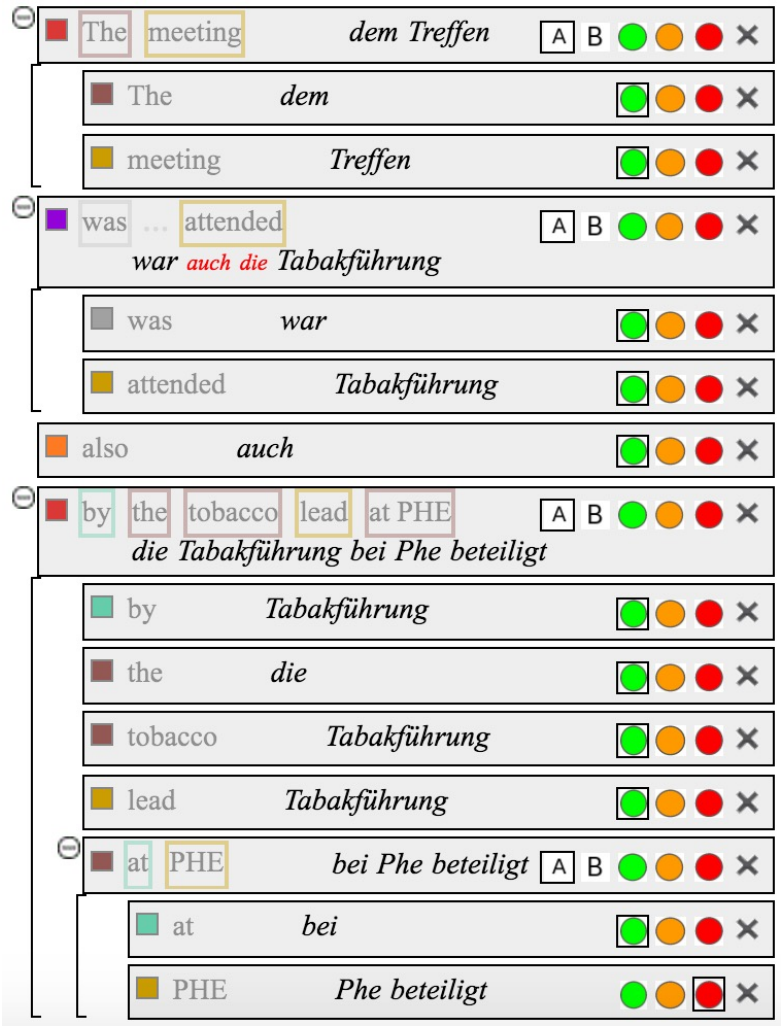


Figure 4.2 – Illustration of the HUME-Annotated sentence "The meeting was also attended by the tobacco lead at PHE." (taken from from the UCCA demo `http://vm-05.cs.huji.ac.il`)

monolingual alignment. Incorrect-word errors are classified with fine grained labels using basic automated semantic and morphological analysis techniques. Word-order errors are detected by finding corresponding paths in permutation graphs.

Other methods of automatic error analysis pay particular attention, especially in the case of high-quality MT, to the translation of certain crucial phenomena (e.g., the translation of pronouns [Guillou and Hardmeier, 2016], and other discourse-level phenomena [Beigman Klebanov and Flor, 2013], translation per part-of-speech (POS) [Max et al., 2010], word order evaluation [Stanojević and Sima'an, 2016], semantic-based evaluation [Lo et al., 2012], etc.), including the

creation of more targeted test sets and evaluation metrics.

Attempts to investigate the interconnection between source, target and system-dependent characteristics are rarer. For instance, such an attempt to find reasons for domain adaptation errors is presented in Irvine et al. [2013]. The authors of this study use the 3-way alignment source→hypothesis→reference, as well as some internal information of the system (PTs, hypothesis search spaces). Using this information, the authors attribute the reasons of MT errors to the original corpus quality (OOV, a new unseen sense); the scoring procedure (correct translation is known to the model, but was not chosen); or the search procedure (caused by pruning during the search procedure).

We will now describe the context of our work with high-quality MT within Cochrane France and our fine-grained methodology to analyze this MT with a particular focus on certain phenomena.

4.3 Automatic Translation of Cochrane Review Abstracts

Cochrane France is a part of the international non-profit Cochrane Collaboration[3], whose main mission is to globally spread high-quality evidence-based research in medicine. To this end, the Cochrane Collaboration publishes high-standard research reviews in English and selective translation of their abstracts into (as of now) 16 languages including French, Spanish, German, Japanese, and traditional Chinese. The review abstracts are publicly available online.[4]

Each Cochrane review abstract is made up of the following parts: (a) a plain language summary (PLS, 40% of the abstract, written in popular scientific style), focused on patient comprehension; (b) a scientific abstract (ABS, 60% of the open access abstract, written in scientific technical style), targeting medical experts.

The translation of English medical texts, in particular that of Cochrane systematic review abstracts, presents a series of anticipated challenges regarding:

1. the translation of the terminology and the professional jargon (e.g., abbreviations);
2. the translation of complex syntactic structures and compounds;
3. the adaptation to variations within the scientific style (this is particularly important in the Cochrane context, where different language styles are in use in the PLS and ABS sections).

[3] http://www.cochrane.org
[4] http://www.cochranelibrary.com

The last problem is more general and includes the first two. We provide here the examples of such register-dependent translation variations:

1. terminology register (e.g., "cycling", ABS: "*cyclisme*" 'cycling', PLS: "*vélo*" 'bicycle'; "surgical fixation", ABS: "*ostéosynthèse chirurgicale*" 'surgical osteosynthesis', PLS: "*fixation chirurgicale*" 'surgical fixation');

2. professional jargon (e.g., "once-daily", ABS: "*une administration quotidienne*" 'a daily administration', PLS: "*une fois par jour*" 'once a day'; "viral", ABS: "*viral*" 'viral', PLS: "*par des virus*" 'by viruses');

3. selective translation of names (e.g., "Cochrane Library", ABS: "*Cochrane Library*" (left untranslated), PLS: "*Bibliothèque Cochrane*" 'Library Cochrane'; "Cochrane Review", ABS: "*Cochrane Review*" (left untranslated), PLS: "*revue Cochrane*" 'review Cochrane');

4. general language (e.g., "to", ABS: "*afin de*" 'so that', PLS: "*pour*" 'to'; "flexible", ABS: "*flexible*" 'flexible', PLS: "*souple*" 'soft').

However, register differences between the ABS and PLS parts are not systematic. To confirm this observation, we analyzed the performance of a binary classifier trained to detect if an abstract sentence is ABS or PLS.

The classifier was trained using Random Forests [Breiman, 2001].[5] Statistics for training and test data are presented in Table 4.1 (the ABS and PLS sets were merged). We extracted 46 sentence-level features: e.g., source and target sentence token statistics, LM scores, n-gram statistics computed using a corpus of Cochrane review abstracts,[6] etc., with the `Quest++` tool [Specia et al., 2015] (for the full list of features see Appendix E.2, we used all the source features, and target features 1-5, 16-33). The classifier shows an equal performance for both classes of around $F = 0.65$ (see Table 4.2). Figure 4.3 plots a projection of the training data (represented by the two most informative features), showing the difficulty to distinguish between two classes.

4.3.1 Cochrane Production Context and Corpus

Cochrane France has been translating review abstracts into French since 2011. At first, abstracts were translated by professional human translators. Those translations were systematically veri-

[5] We used the Random Forests implementation available in `Scikit-learn` [Pedregosa et al., 2011] with the following parameters: Gini as the optimizing criterion, 700 estimators, a maximum depth of 700 and a minimum number of leaf samples of 10. All other parameters are those provided by default.

[6] Cochrane Reference Corpus was used. To extract LM-related features we built two 4-gram LMs from the corresponding monolingual parts of the corpus with modified Kneser-Ney smoothing using the `SRILM` [Stolcke, 2002] toolkit.

	set	lines	#, tok. (EN)	#, tok. (FR)
train	ABS	20K	1.4M	1.8M
	PLS	20K	1.3M	1.6M
test	ABS	500	82K	105K
	PLS	500	65K	87K

Table 4.1 – ABS/PLS training and test data statistics

set	ABS	PLS
PR	0.63	0.66
RC	0.69	0.60
F	0.66	0.63

Table 4.2 – ABS/PLS classifier performance (PR denotes precision; RC – recall; F – F-score)

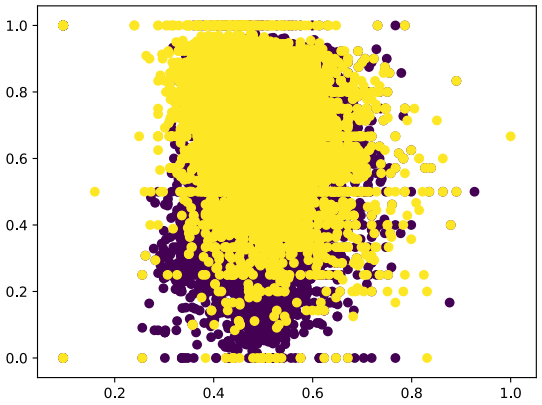


Figure 4.3 – Projection of the training data (two most informative features: ratio of number of tokens between source and target sentences, percentage of distinct bi-grams seen in the in-domain source corpus) for the binary ABS/PLS classification

fied by Cochrane experts. An SMT approach was introduced in 2013. To implement this approach we have developed an in-domain English-French PBSMT system (`Cochr-SMT`, see section 4.3.3).

Currently, around 14 abstracts are translated into French each month (13% of the total quantity of monthly abstracts written in English, on average for the period 01/03/2016-01/03/2017). The PE is performed on a regular basis by 5 volunteer domain specialists and 1 professional translator. They use the commercial Smartling[7] PE environment (in Appendix B we provide extracts of the Cochrane Application Programming Interface (API) code; the API was developed by us to integrate `Cochr-SMT` into Smartling). The abstracts post-edited into French attain around 160K views per month.[8]

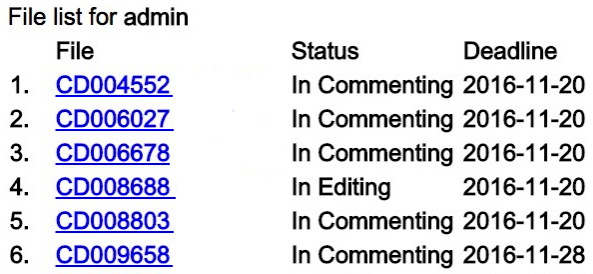
We also developed our own more intuitive PE interface, which was mostly used in PE projects with students (see extracts of Java code created using the `Google Web Toolkit` (`GWT`)[9] in Appendix C).

The interface provides access to user accounts with lists of assigned documents. The information on users and documents is managed by `Apache CouchDB`[10] (a document-oriented NoSQL database). Figure 4.4 shows a user account page with assigned documents. The interface also displays deadlines of execution and job statuses ("In Editing" – work in progress, "In Commenting" – work is being verified by a supervisor, professional translator). Once a document is opened, users can see three columns with source, MT and PE. Displayed text is divided into sections according to the original Cochrane review abstract structure (a PLS section consists of a title and a summary; an ABS section may be structured as follows: title, background, objectives, search strategy, selection criteria, data collection, results, conclusion). Users can edit MT of each sentence, submit their PE to the database, or comment on it (comments served as a means of communication with a supervisor).

Our interface also gives access to some basic CAT functionalities: translation memory matches and concordancer searches. Figure 4.5 illustrates the results of a concordancer search for the word "neuropathy": in 80% of its occurrences in a Cochrane corpus it is translated as "*neuropathie*" (left part of the bottom screen), the right part of the bottom screen provides examples of the word usage.

The English-French Cochrane parallel corpus accumulated since 2011 consists of the following parts:

[7] `https://www.smartling.com`
[8] as estimated for 2016
[9] `http://www.gwtproject.org`
[10] `http://couchdb.apache.org`



File list for admin

	File	Status	Deadline
1.	CD004552	In Commenting	2016-11-20
2.	CD006027	In Commenting	2016-11-20
3.	CD006678	In Commenting	2016-11-20
4.	CD008688	In Editing	2016-11-20
5.	CD008803	In Commenting	2016-11-20
6.	CD009658	In Commenting	2016-11-28

Figure 4.4 – User account in the in-house PE interface

- **Cochrane Reference Corpus**: a high-quality corpus consisting of human-translated review abstracts collected over a three-year period (2011-2013).
- **Cochrane Post-Editing (PE) Corpus 1**: a lower quality corpus consisting of machine-translated review abstracts post-edited mainly by volunteer domain professionals over a 6-month period (Oct. 2013-May 2014). The MT was performed by different versions of the Cochrane SMT system.
- **Cochrane Google Post-Editing (PE) Corpus**: a lower quality corpus consisting of machine-translated review abstracts produced by the Google online system.[11] They were post-edited by both professional translators and volunteer domain professionals over a 1-year period (Aug. 2014-Sep. 2015).
- **Cochrane Post-Editing (PE) Corpus 2**: a lower quality corpus consisting of machine-translated review abstracts post-edited by volunteer domain professionals over a 1-year period (Sep. 2015-Sep. 2016). The MT was performed by the latest version of the Cochrane SMT system (described in section 4.3.3).

The post-edited Cochrane data are considered to be of a lower quality since they were not subject to a quality control procedure. A procedure comprising a verification of translation by Cochrane specialists was applied to professional human translations. After the introduction of SMT the procedure was simplified. Post-editors, domain specialists are to make text "understandable" and correct it from the terminological point of view. Table 4.3 provides statistics over each part of the corpus.

We have prepared and made publicly available the above-mentioned parts of the Cochrane Corpus. We provide source texts, MT outputs (for the in-house systems) and post-edits (or

[11] `https://translate.google.com`

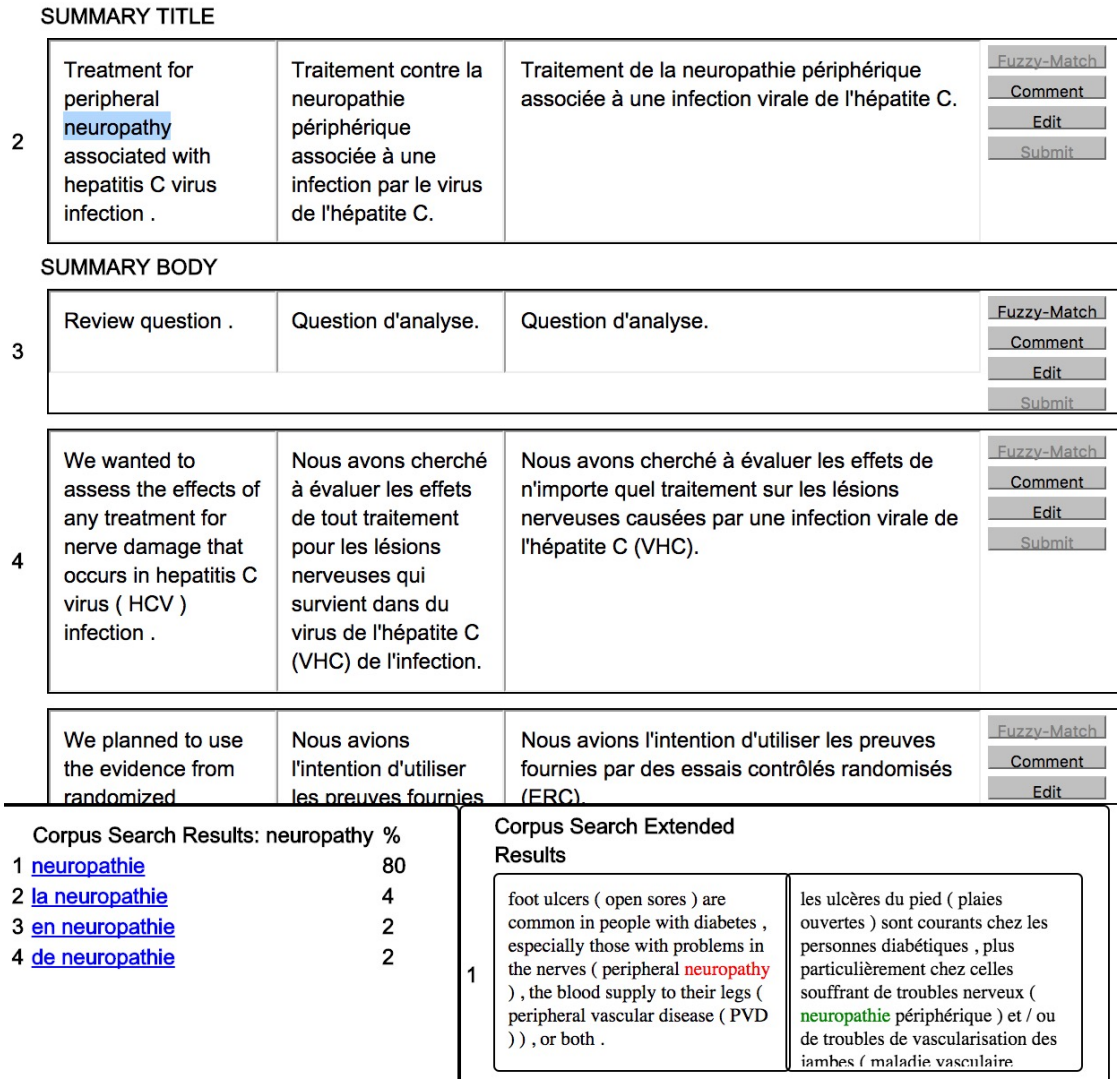


Figure 4.5 – Example of a concordancer search for the word “neuropathy” in the in-house PE interface

corpus	#, lines	# tok., EN (src.)	# tok., FR (trg.)
Cochrane Reference	130 K	2.9 M	3.6 M
Cochrane PE 1	21 K	500 K	600 K
Cochrane Google PE	31 K	740 K	890 K
Cochrane PE 2	10K	235K	288K

Table 4.3 – Cochrane corpora sizes

human translations, for Cochrane Reference Corpus) in an XML format preserving the original abstract sections. The data are available at `http://www.translatecochrane.fr/corpus` (see examples in Appendix A).

4.3.2 Manual Error Analysis of Post-Edits

For a more detailed insight of the quality of Cochrane PE we performed a manual error analysis of 50 post-edited review abstracts (25 originally translated by `Cochr-SMT` and 25 documents translated by the publicly available Google system `Google-SMT`, from the corpora Cochrane PE 2 and Cochrane Google PE respectively,[12] see Table 4.4). One person (professional translator with significant experience in post-editing Cochrane abstracts) performed annotations using the `Yawat` tool [Germann, 2008].[13]

The annotation procedure included several steps. PE words were first marked as:

- MT – words left uncorrected, a post-editor considered that MT was correct;
- essential PE (EPE) – PE that was essential for the understandability and terminological correctness of a text;
- preferential PE (PPE) – PE that was performed to match subjective stylistic preferences of a post-editor.

Then, word-level errors in the final PE were annotated according to the following typology (largely inspired by the `MeLLANGE`[14] typology):

- Adding (Add) – a word unnecessarily added in PE;
- Distortion (Dist) – a PE word that hinders the transfer of essential content and requires checking against the source (e.g., translating the word "may" as "*peut*" 'can');

[12] the version publicly available online in Sep. 2015

[13] The work was performed in collaboration with Hanna Martikainen, CLILLAC-ARP, Université Paris Diderot, Sorbonne Paris Cité, who also was the annotator.

[14] `http://mellange.eila.jussieu.fr/annot.en.shtml`

- Grammar (Gr) – a grammatical mistake in PE (word spelling, bad choice of an article etc.);
- Omission (Om) – a source word whose translation is omitted in PE;
- Phraseology (Phr) – a PE word inconsistently translating a medical expression;
- Syntax (Syn) – a syntactic error in PE (e.g., word order);
- Terminology (Tm) – an incorrect term in PE;
- Terminological Consistency (TmC) – a correct translation of a term, but not the one used in Cochrane review abstracts.

Within each error type we also distinguished major and minor errors, respectively disturbing or not disturbing text comprehension.

Annotation results show that the PE of the `Cochr-SMT` output contains 27% less errors than the PE of the `Google-SMT` output (273 errors vs. 375 errors respectively, see Figure 4.6). However, while post-editing the higher quality MT output of `Cochr-SMT`, users tend to leave more MT errors uncorrected (52% for `Cochr-SMT` vs. 43% for `Google-SMT`), and at the same time tend to make less errors in preferential corrections (18% for `Cochr-SMT` vs. 26% for `Google-SMT`). We believe that this happens because more fluent translations relax user attention, users "gain more trust" in MT and correct less.

Figure 4.7 shows statistics of PE errors per type. The most frequent errors for both systems are minor terminology consistency errors (20% for `Cochr-SMT` vs. 30% for `Google-SMT`) and grammar errors (23% for `Cochr-SMT` vs. 24% for `Google-SMT`). The first type of errors indicates the inconsistency of PE from the Cochrane corpus standpoint: for instance, the noun "outcome" translated as "*résultat/issue*" 'result/issue', instead of "*critère de jugement/d'évaluation*" 'criteria of judgement/evaluation'. Grammar errors, for instance, contain errors in the usage of articles: for instance, "Study characteristics", `Cochr-SMT`: "*Les caractéristiques de l'étude*" 'The characteristics of the study', corrected to "*Les caractéristiques des études*" 'The characteristics of studies' in PE, but should be "*Caractéristiques des études*" 'Characteristics of studies' as the usage of articles is not recommended in titles. All these observations are consistent with conclusions on the quality of PE performed by non-professionals that can be found in other works [Mitchell, 2015].

All the detected tendencies reveal the lower quality of the Cochrane PE corpora in comparison to Cochrane Reference Corpus, related both to the quality of the underlying MT, as well as to

the fact of using non-professionals as post-editors (they leave MT errors uncorrected, face more difficulties in keeping the corpus consistency, hence terminology consistency errors). Concerning the latter point, a more detailed comparison to PE performed by experienced professional post-editors would be required for a final conclusion.

MT	#, lines	#, tok. (src.)	#, tok. (trg.)	HTER	%, PPE tok.	#, post-eds.
`Cochr-SMT`	1079	158K	199K	0.34	12	4
`Google-SMT`	1250	179K	223K	0.36	11	3

Table 4.4 – Statistics for the analyzed PE (PPE denotes preferential PE; #, post-eds. – number of post-editors)

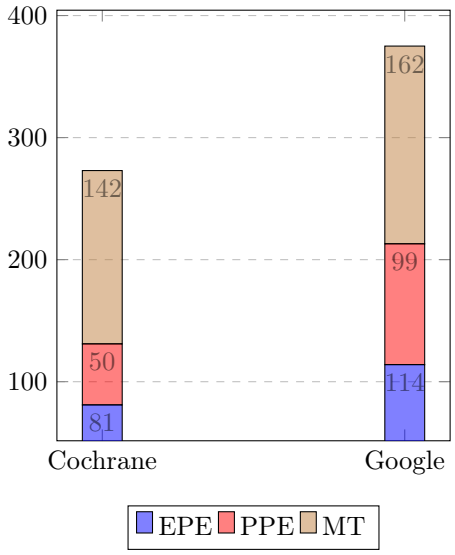
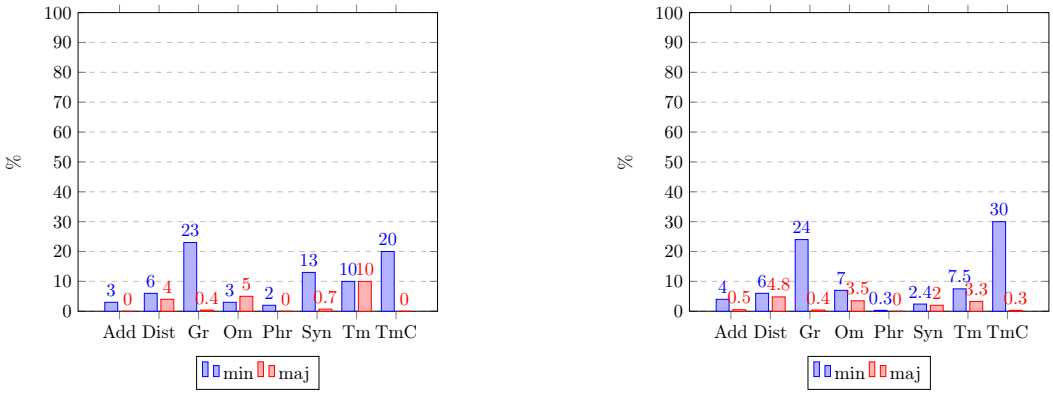


Figure 4.6 – PE error statistics per segment type

4.3.3 Cochrane High-Quality Statistical Machine Translation System

In its current form, the Cochrane SMT system (`Cochr-SMT`) uses the `Moses` toolkit [Koehn et al., 2007]. Cochrane Reference Corpus was used to train the main model (PT and RM `msd-bidirectional-fe`). The Cochrane PE 1 and additional corpora models (same components as for the main model) were used to find translations of n-grams (up to $n = 4$) absent from the first model (`Moses back-off` mode). Additional corpora (WMT'14 medical task parallel data:[15] EMEA, COPPA, PatTR, UMLS, Wikipedia) include various in-domain corpora

[15] `http://statmt.org/wmt14/medical-task`



(a) PE error statistics for `Cochr-SMT`

(b) PE error statistics for `Google-SMT`

Figure 4.7 – PE error statistics per error type

of different genres (drug instructions, medical patents, thesauri, articles intended to popularize medical knowledge, etc., see Table 4.5). We applied in-house scripts for the data cleaning and the `Stanford` tool for the tokenization of both English and French.[16] The `MGIZA` [Gao and Vogel, 2008] tool was used to compute word alignments.

The monolingual parts of the corpora mentioned above, as well as the general domain data (WMT'13 news data,[17] see Table 4.5) were used to train 4 LMs. We built two 6-gram models using the Cochrane corpora, and two 4-gram models using the WMT data. The LMs were trained with modified Kneser-Ney smoothing using the `SRILM` toolkit [Stolcke, 2002].

The system was tuned using post-edited data (820 lines, 20K English tokens, 26K French tokens) using `kb-mira` with default options on 300-best hypotheses.

corpus	**#, lines**	**#, tok., EN (src.)**	**# tok, FR (trg.)**
		Bilingual	
WMT'14 medical task	5M	89M	101M
		Monolingual	
Out-of domain WMT'13	17M	–	260M

Table 4.5 – Additional corpora sizes

An automatic evaluation of this system was performed using a test set comprising 713 sentences for the PLS part and 949 sentences for the ABS part. Those sentences were extracted from the corresponding machine-translated (by `Cochr-SMT`) and post-edited review abstracts.

Examples of the test set sentences demonstrating the translation challenge within each register

[16] `http://nlp.stanford.edu/software/tokenizer.shtml`
[17] `http://statmt.org/wmt13/translation-task.html`

are provided in Table D.1 in Appendix D.

The results, presented in Table 4.6, reveal a high level of translation performance according to the automatic metrics used, with a slightly better performance for the ABS section.

We also report a comparison with translations produced by the Google system (for the same amount of different post-edited PLS and ABS sentences from Cochrane abstracts),[18] as well as with the translations of the same test set produced by a lower performance system trained only on the WMT'14 medical task parallel data (`WMT14-SMT`). This system uses only the LMs built from the WMT data. It was tuned using the same post-edited Cochrane data as `Cochr-SMT`.

The linear lattice BLEU oracle (LB-4g) was used to estimate the potential of the systems [Sokolov et al., 2012]. The atypically low oracle improvements in terms of the automatic metrics scores (+6 BLEU, no improvement in HTER) suggest that the Cochrane system produces translations that are close to the best translations it can produce given its training data.

metric	`Cochr-SMT`			`WMT14-SMT`			`Google-SMT`		
	ALL	PLS	ABS	ALL	PLS	ABS	ALL	PLS	ABS
BLEU	**57**	55	**58**	29	30	28	49	50	48
Oracle BLEU	63	62	64	40	41	39	NA	NA	NA
(H)TER	**0.30**	0.32	0.28	**0.58**	0.54	0.62	0.36	0.37	0.35
Oracle TER	0.30	0.32	0.28	0.55	0.50	0.59	NA	NA	NA

Table 4.6 – Automatic evaluation results

Analysis of the HTERp traces confirmed the system performance differences for the PLS and ABS parts (see Table 4.7). For our experiments, we used the HTERpA configuration [Snover et al., 2009b], optimized for human adequacy judgments, with the following components for processing French: the Snowball stemmer [Porter, 2001], and a paraphrase table extracted from the concatenation of the Cochrane Reference and PE 1 corpora [Bannard and Callison-Burch, 2005; Denkowski and Lavie, 2014].

The PE operations performed to the output translation tend to be non-repetitive: only about 11% of edited tokens/pairs of tokens per operation are unique, but the most frequent PE operations (see Table 4.8) do not exceed 11% of all the changes per operation.

As shown in Table 4.9, the most common part-of-speech (POS) substitution patterns reveal frequent modifications to common nouns (NC) and to the POS that cooccur with them (DET, P, ADJ), potentially forming terms and terminological constructions, as well as grammatical changes to verbs (V gram) [Schmid, 1995; Toutanova et al., 2003].

[18] the version publicly available online in Sep. 2015

	PLS	ABS
HTERp score	25	25
#, hyp. tok.	19K	32K
#, ref. tok.	19K	32K
operation	%, hyp. tok. edited	
SH	4	5
M	74	78
T	3	3
PAR	7	6
S	8	7
D	8	6
	%, ref. tok. edited	
I	7	7

Table 4.7 – % of hyp./ref. tokens (words) aligned by an HTERp operation or a match for `Cochr-SMT` (SH denotes a shift, M – a match, T – a stem match, PAR – a paraphrased word, S – a substitution, D – a deletion, I – an insertion)

	PLS		ABS	
operation	tok.	%	tok.	%
T	de → des	11	de → des	11
PAR	les pansements → pansements à base	1	de la même fratrie → frères et sœurs	1
S	les → des	2	, → ;	8
D	de	6	les	5
I	,	4	de	4

Table 4.8 – Most frequent token changes per operation for `Cochr-SMT` (T denotes a stem match, PAR – a paraphrased word, S – a substitution, D – a deletion, I – an insertion)

PLS		ABS	
pattern	%	pattern	%
P → P	10	P → P	9
NC → NC	7	NC → NC	8
DET → DET	7	PUNC → PUNC	8
DET → P	5	DET → P	6
P → DET	4	DET → DET	4
V → V gram	3	ADJ → ADJ	4
ADJ → ADJ	3	P → DET	4
ADJ → NC	3	ADJ → NC	3
VPP → VPP	2	V → V gram	2
V → V	2	NC → P	2

Table 4.9 – Most common POS substitution patterns for `Cochr-SMT`

The presented superficial observations give us no guidance if and how the system can be improved. To this end, a fine-grained performance analysis is needed to get an insight of translation difficulties. Further, while analyzing the high-quality MT, we will discuss "residual" errors.

4.3.4 Methodology for Diagnosing High-Quality Machine Translation

Our fine-grained methodology for diagnosing MT performance investigates the interconnection between source, target and system-dependent characteristics in an attempt to study the MT of certain anticipated phenomena. In relation to the `Cochr-SMT` context, we seek to answer the following questions: Which kind of *translation difficulties does the system face*? Are those difficulties related to a greater extent to *the initial corpus quality or to the system scoring procedure*?

Taking the observations presented in Table 4.9 into account, we decided to focus on the translation quality of certain syntactic constituents and POS, in particular noun phrases (NP),[19] as potential complex terminological structures, as well as verbs and nouns [Klein and Manning, 2003].

Thus, we extract the following groups of unique source n-grams (*units*): the ones corresponding to the longest NPs, then from the rest of each sentence we extract units corresponding to the neighboring/single verbs (V) and nouns (N). The residual sentence spans of varying length, not covered so far, are put in a separate group (Rest). A sketch of our protocol is provided in Figure 4.8.

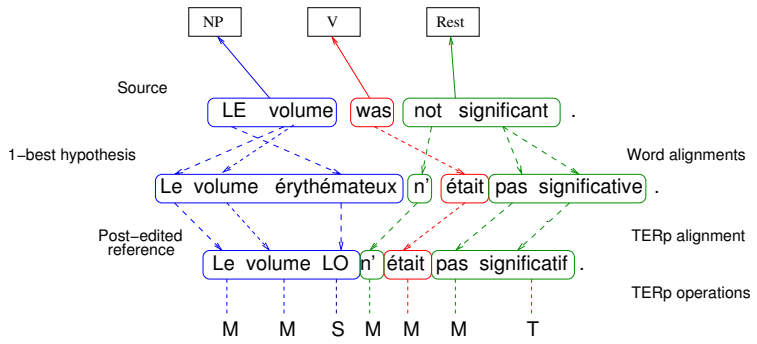


Figure 4.8 – Illustration of our analysis strategy

[19] The most frequent POS patterns within NP for our test set are: NC, DET+NC, ADJ+NC.

Further, we distinguish the following subordinate groups: the units known to a model and also present in the 1-best hypothesis segmentation in >= 80% of their occurrences (U_{1-best}); the ones known to a model but absent from the 1-best segmentation in >= 80% of their occurrences (U_{pres}); and the units unknown to a model (U_{abs}). For instance, if the unit "not significant" appears 3 times in the test data: 2 times it appears in 1-best hypothesis segmentations, and 1 time it is translated by composition (made up of several phrases $\bar{e}$), it is considered as U_{pres}. U_{pres} and U_{abs} units are naturally more prone to errors.

Knowing the alignment $\mathbf{a}^1$ between $\mathbf{f}$ and $\mathbf{e}^1$, produced by the decoder, as well as the word alignment $\mathbf{e}^1 \rightarrow \hat{\mathbf{e}}$ produced by TER, we compute the 3-way alignment $\mathbf{f} \rightarrow \mathbf{e}^1 \rightarrow \hat{\mathbf{e}}$. Then for each unit ($U$) we compute the averaged translation quality statistics for all its occurrences (u), by comparing each aligned hypothesis segment (u_{hyp}) to its aligned reference segment (u_{ref}). For instance, for the unit "not significant" with 3 occurrences we detect 3 hypothesis translations of different quality (see Table 4.10).

For each unit U we estimate the following parameters:

1. unit frequency (fr_U): we believe that more frequent units are easier to translate (3 for our running example);

2. unit length in words ($\#\, f_U$): shorter units are less likely to be translated by composition and consequently "better" translated (2 for our running example);

3. match rate (M_U): for each word within a unit, we count the percentage of times its translation is matched across occurrences, then average these values. This parameter relates the previous two parameters with an estimation of translation quality. For a single occurrence u the match rate is computed as follows:

$$M_u \;=\; \frac{\#\, M_{e^1_u}}{\#\, e^1_u} \tag{4.1}$$

Table 4.10 shows an example of the computation of the average unit match rate M_U: in its 2 occurrences the unit "not significant" is partially correctly translated, the hypothesis of 1 occurrence is totally correct.

To trace the connection between the system performance and source peculiarities (intuition suggests that units with a high concentration of rare terms are more difficult to translate), we

	u_{hyp}	u_{ref}	M_u	M_U
1	(*le volume* 'the volume') *pas significative (Fem., Sg.)*	*pas significatif (Masc., Sg.)*	0.5	
2	(*il n'a été* 'it was') *pas significative (Fem., Sg.)*	*pas significatif (Masc., Sg.)*	0.5	0.7
3	(*la laminotomie* 'the laminotomie') *pas significative (Fem., Sg.)*	*pas significative (Fem., Sg.)*	1.0	

Table 4.10 – Example of the match rate calculation for the unit "not significant" with 3 occurrences

f	mapping
treatment	Therapeutic or Preventive Procedure
for	-
IgG	-
and	-
IgA	Immunologic Factor, Pharmacologic Substance
paraproteinaemic neuropathy	Disease or Syndrome

Table 4.11 – Example of a `Metamap` mapping for the sentence "Treatment for IgG and IgA paraproteinaemic neuropathy" (finally chosen terms are in bold)

calculate the unit *term rate* (TR_U):

$$TR_U = \frac{\#\, f_U^{term}}{\#\, f_U}, \tag{4.2}$$

where f_U^{term} is a word of a unit marked as a term or a part of a complex term. For our running example this rate is equal to 0. For instance, for the unit "LE volume" (see Figure 4.8) it is equal to 0.5.

The term mapping was performed with the `Metamap` tool for medical texts [Aronson and Lang, 2010]. `Metamap` searches were parametrized to avoid mapping to general concepts. A corpus statistics filter was used to further exclude highly frequent words ($fr \geq 50K$). Table 4.11 shows a `Metamap` mapping and illustrates the filtering procedure: for instance, the frequent noun "treatment" mapped to the general concept "Preventive Procedures" was not considered.

We associate target errors with occurrences in the original training corpus, by computing the prior translation entropy ($H_{prior}(U)$) of the distribution of the phrase translation probabilities of all the possible target bi-phrases $\bar{e}$, where $\bar{f}$ is equal to U. We believe that higher $H_{prior}(U)$ values can make translation choices more difficult and disturb MT quality. We take translation probabilities from PTs with lemmatized target bi-phrases $d(\bar{e})$, since we are more interested in

$\bar{e}$	$p(\bar{e}\|U)$	$d(\bar{e})$	$p(d(\bar{e})\|U)$
non significatif	0.29		
non significatifs	0.04	*non significatif*	0.41
non significative	0.08		
pas significatif	0.08		
pas significatifs	0.17	*pas significatif*	0.46
pas significative	0.04		
pas significatives	0.17		
cependant pas significative	0.04	*cependant pas significatif*	0.04
pas révélées significatives	0.04	*pas révéler significatif*	0.04

Table 4.12 – Example of a lemmatized PT entry for the unit "not significant"

the actual variety of translation choices, rather than in morphological variety (see Table 4.12):

$$H_{prior}(U) \;=\; -\sum_{d(\bar{e})} p(d(\bar{e})|U) \log p(d(\bar{e})|U) \tag{4.3}$$

We attempt to correlate errors with the scoring procedure by measuring how well an MT system uses context to reduce the initial translation entropy. In order to do that we compute the posterior entropy ($H_{post}(\varepsilon)$) of the distribution of the path posterior probabilities $P(d(e)_u|\varepsilon)$ of $d(e)_u$ translating a word f_u across occurrences, then average these values:

$$H_{post}(\varepsilon) \;=\; -\sum_{d(e)_u} p(d(e)_u|\varepsilon) \log p(d(e)_u|\varepsilon) \tag{4.4}$$

Here, we are again interested in actual translation variants and use lemmatized translations $d(e)_u$. We compute $p(d(e)_u|\varepsilon)$ from the estimation of path posterior probabilities as defined in [de Gispert et al., 2013]:

$$p(d(e)_u|\varepsilon) \;=\; \frac{\sum_{\mathbf{e}\in\varepsilon_{d(e)_u}} \exp(\alpha S(\mathbf{f},\mathbf{e}))}{\sum_{\mathbf{e}'\in\varepsilon} \exp(\alpha S(\mathbf{f},\mathbf{e}'))}, \tag{4.5}$$

where ε is the space of translation hypotheses (a $10K$-best list was chosen), and $S(\mathbf{f},\mathbf{e})$ is the score assigned by the model to the sentence pair $(\mathbf{f},\mathbf{e})$.

Table 4.13 illustrates the computation of $H_{post}(\varepsilon)$ for our running example: in one of its occurrences the first word of the unit "not significant" is translated by 2 equally likely variants "*pas*" and "*non*" ($H_{post}(\varepsilon) = 1$), for the second word "significant" only one lemmatized translation variant is found ($H_{post}(\varepsilon) = 0$). The posterior entropy for this occurrence is thus equal to 0.5.

f_U	$d(e)_u$	$p(d(e)_u\|\varepsilon)$	$H_{post}(\varepsilon)$	$H^u_{post}(\varepsilon)$
not	*pas*	0.5	1	0.5
	non	0.5		
significant	*significatif*	1	0	

Table 4.13 – Example of the posterior entropy computation for the unit "not significant"

4.3.5 Results and Analysis

The proposed methodology was applied to the test set presented in Section 4.3.3 to analyze the performance of `Cochr-SMT`, as well as the performance of the less competitive `WMT14-SMT`.

During our analysis we attempted to answer the following questions:

1. What are the "worst" translated unit groups for the high-performance system?

We took the average percentage of matches per group M_U as an indicator of translation quality (see Figure 4.9a). We explored the group characteristics by analyzing their general statistics (see Table 4.14) and the TR_U (see Figure 4.9c).

Figure 4.9a shows that the system faces difficulties translating the units of the V group (lowest average $M_U \approx 53\%$), although the majority of those units are known to the model (97%, *1-best+Pres*, see Table 4.14). Verbs are more rare in Cochrane abstracts and we believe that all the specificities of their translations can not be reliably captured using the existing amount of data.

For the NP group, Figure 4.9a demonstrates that the most difficult-to-translate units are the units that are absent from the PT (M_U=74%, *Abs*), which have to be translated by composition.

Figure 4.9c shows the high term concentration for the N group units (average TR_U=30%). Thus, the "worst" translated units of the N group (M_U=24%, *Abs*) are mainly terms unknown to the model. The high rate of N units that are present in the 1-best segmentation (25%, *1-best*, see Table 4.14) also suggests frequent term translation inconsistency due to the lack of context.

The same difficulties are observed for the less competitive `WMT14-SMT`: the units of the V group are the "worst" translated (lowest average $M_U \approx 36\%$); the translation of the NP group units unknown to the model is of a low quality (M_U=61%, *Abs*); the translation of the term N units present in the 1-best segmentation is often inconsistent (M_U=44%, *1-best*, TR_U=34%, see Figure 4.10a, Figure 4.10c).

Thus, our high-performance system faces difficulties in translating verbs, unseen terms and terminological expressions.

NP total : 3528						
	Cochr-SMT			WMT'14 SMT		
	1-best	Pres	Abs	1-best	Pres	Abs
%	10	27	63	9	10	81
$\# f_U$	2	3	10	2	2	9
fr_U	1	1	1	1	1	1
N total : 336						
	Cochr-SMT			WMT14-SMT		
	1-best	Pres	Abs	1-best	Pres	Abs
%	**25**	71	4	**41**	47	12
$\# f_U$	1	1	1	1	1	1
fr_U	1	2	1	1	3	1
V total : 982						
	Cochr-SMT			WMT14-SMT		
	1-best	Pres	Abs	1-best	Pres	Abs
%	**18**	**79**	3	32	62	6
$\# f_U$	1	1	2	1	1	2
fr_U	1	3	1	1	4	1
Rest total : 931						
	Cochr-SMT			WMT14-SMT		
	1-best	Pres	Abs	1-best	Pres	Abs
%	13	75	12	21	57	22
$\# f_U$	2	2	2	1	1	2
fr_U	1	6	1	1	8	1

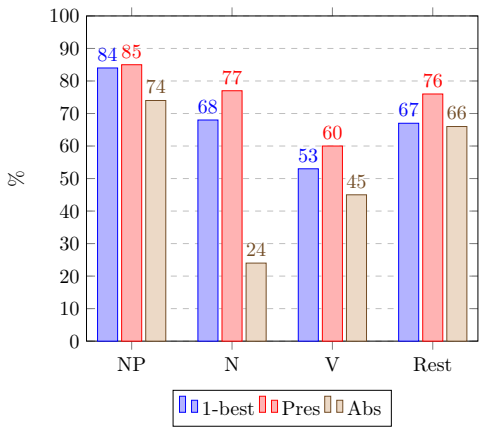
Table 4.14 – General statistics per unit group (NP denotes units containing longest NPs, V – neighboring/single verbs, N – neighboring/single nouns, Rest – residual sentence spans; U_{1-best} denotes units mostly present in the 1-best hypothesis segmentation; U_{pres} – units known to the model but mostly absent from the 1-best segmentation; U_{abs} — units unknown to the model)

2. Is the scoring procedure to blame for residual translation errors?

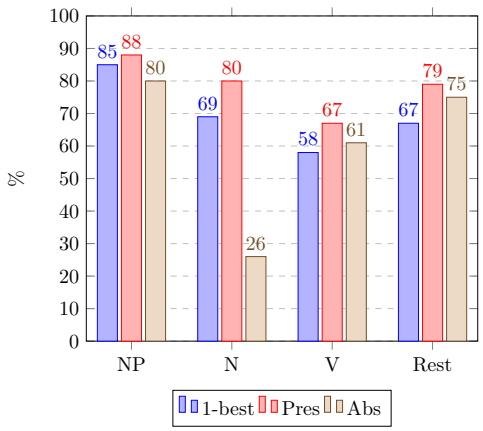
To answer this question we analyzed the per-group differences between the match percentage values ΔM_U for the system hypotheses and the oracle hypotheses (see Figures 4.10a, 4.9b).

Additionally, to evaluate the scoring procedure we studied the correlation between the low/high match percentage zones (see Figure 4.9a) and the prior/posterior entropy values (see Figures 4.11a, 4.11b). For instance, the known (*1-best+Pres*) N group units with a high match percentage (average $M_U \approx 73\%$) and the known V group units with a low match percentage (average $M_U \approx 57\%$) both correspond to the same average prior entropy value ($H_{prior}(U) \approx 2$), as well as to the absence of a significant difference between the average posterior entropy values ($H_{post}(\varepsilon) \approx 0.4$ vs. $H_{post}(\varepsilon) = 0.3$ respectively).

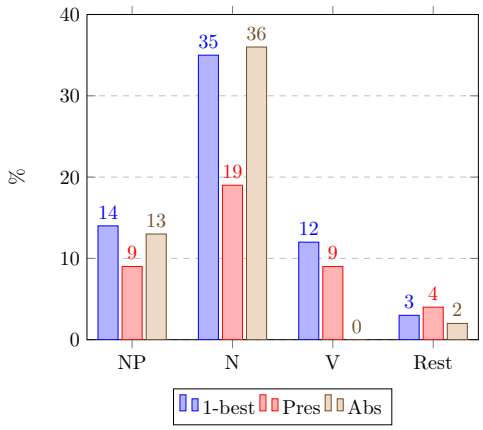
The absence of correlation between the match percentage and prior/posterior entropy values



(a) Average percentage of matches per unit group

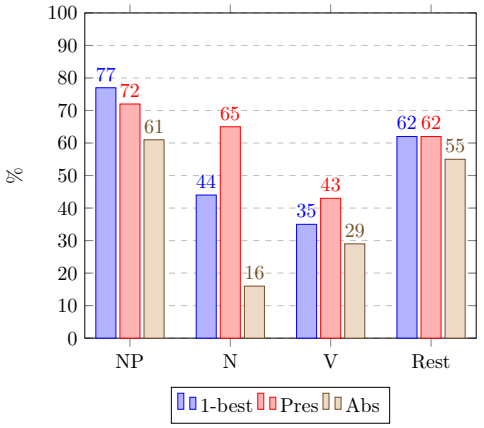


(b) Average percentage of oracle matches per unit group

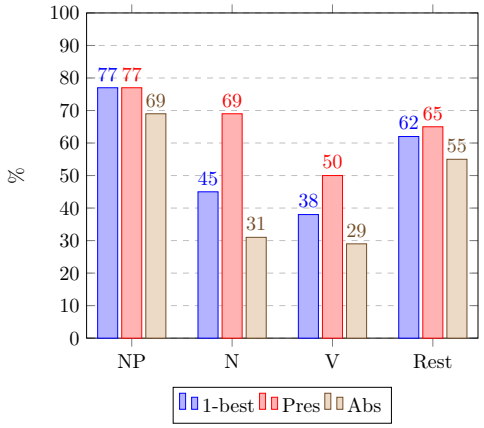


(c) Average term rate per unit group

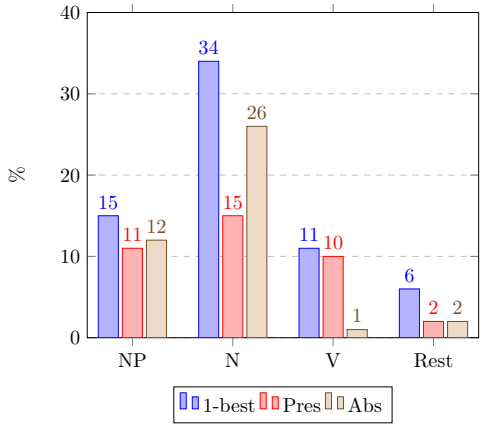
Figure 4.9 – Translation quality statistics for `Cochr-SMT` (NP denotes units containing longest NPs, V – neighboring/single verbs, N – neighboring/single nouns, Rest – residual sentence spans; U_{1-best} denotes units mostly present in the 1-best hypothesis segmentation; U_{pres} – units known to the model but mostly absent from the 1-best segmentation; U_{abs} — units unknown to the model)



(a) Average percentage of matches per unit group

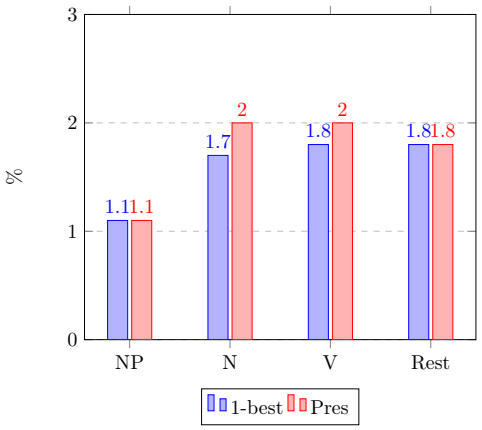


(b) Average percentage of oracle matches per unit group

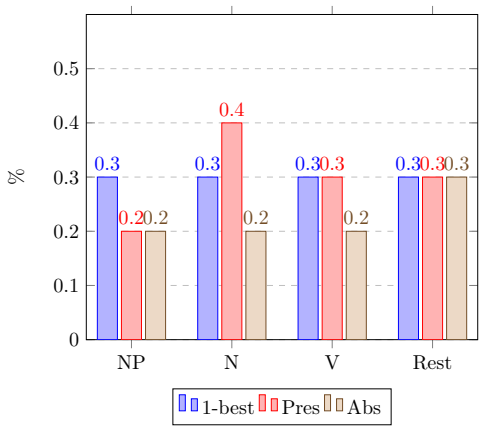


(c) Average term rate per unit group

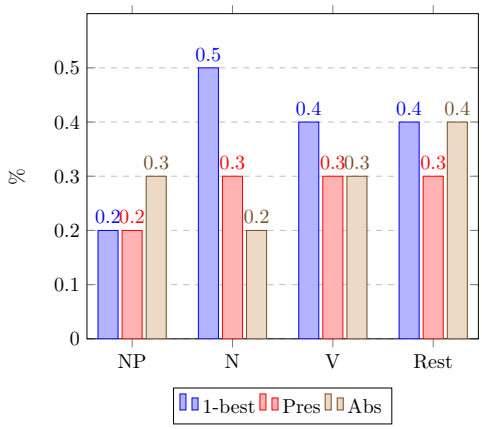
Figure 4.10 – Translation quality statistics for `WMT14-SMT` (NP denotes units containing longest NPs, V – neighboring/single verbs, N – neighboring/single nouns, Rest – residual sentence spans; U_{1-best} denotes units mostly present in the 1-best hypothesis segmentation; U_{pres} – units known to the model but mostly absent from the 1-best segmentation; U_{abs} — units unknown to the model)



(a) Average translation prior entropy per unit group (`Cochr-SMT`)



(b) Average translation posterior entropy per unit group (`Cochr-SMT`)



(c) Average translation posterior entropy per unit group (`WMT14-SMT`)

Figure 4.11 – Entropy estimations (NP denotes units containing longest NPs, V – neighboring/single verbs, N – neighboring/single nouns, Rest – residual sentence spans; U_{1-best} denotes units mostly present in the 1-best hypothesis segmentation; U_{pres} – units known to the model but mostly absent from the 1-best segmentation; U_{abs} — units unknown to the model)

confirms that the scoring procedure is not responsible for most of the errors and in the majority of cases the system is simply unable to produce "correct" translations. Our conclusion is confirmed by the fact that ΔM_U between the system hypotheses and the oracle hypotheses is quite low (about 5%).

In comparison, the scoring procedure of `WMT14-SMT` can be improved more significantly. The oracle changes to the `WMT14-SMT` output (ΔM_U of about 4%) are more significant since they are performed for more units. Table 4.14 and Figures 4.10a, 4.10b show that the translation of 41% of the *1-best* N group units is improved with ΔM_U=1% (compare to 25% of N *1-best* units with ΔM_U=1% for `Cochr-SMT`, see Figures 4.9a, 4.9b).

For `WMT14-SMT` we should also notice the presence of a more distinct correlation between the translation quality indicator and the entropy values: e.g., the high posterior entropy value ($H_{post}(\varepsilon) = 0.5$) for the *1-best* N units corresponds to the low match percentage (M_U=44%, see Figures 4.11c, 4.10a).

So improving the scoring procedure of `Cochr-SMT` can hardly help to improve its output.

3. What is the nature of the per-group residual errors?

The manual analysis of the "worst" ($M_U <= 20\%$) and "best" ($M_U >= 80\%$) translated U for `Cochr-SMT` per target group provided some insight as to the nature of residual errors.

Confirming our previous observations, the remaining errors of the N and NP groups concern mainly terms unknown to the model, as well as errors in term and professional jargon precision (e.g., "cardiotoxicity", MT: "*cardiotoxicité*" 'cardiotoxicity', Ref.: "*toxicité cardiaque*" ' toxicity cardiac', absent from the oracle hypothesis; "IDA", MT: "*une anémie ferriprive*" 'an anemia of iron deficiency', Ref.: "*l'IDA*" 'the IDA', absent from the oracle hypothesis).

In the NP group we often face complex terminological constructions translated by composition (e.g., "people with functioning kidney transplants", MT: "*les personnes atteintes de fonctionnement de greffes de rein*" 'the people suffering from functioning of kidney transplants', Ref.: "*des receveurs de greffe rénale fonctionnelle*" 'recipients of transplant functional renal', absent from the oracle hypothesis).

The residual translation errors related to the V group are mostly caused by specificities of the source language:

1. source syntactic/stylistic peculiarities (very often expletive constructions), requiring restructuring on the target language side (see Table D.2 in Appendix D);

2. tense and modality (e.g., "may reduce", MT: "*peut réduire*" 'can reduce', Or.: "*peut réduire*" 'can reduce', Ref.: "*pourrait réduire*" 'could reduce').

We also notice an increased quantity of paraphrasing corrections performed to the V group (e.g., "we **searched** all databases", MT: "*nous avons* ***effectué des recherches*** *dans toutes les bases de données*" 'we have **performed searches** in all the databases', Ref.: "*nous avons* ***interrogé*** *toutes les bases de données*" 'we have **questioned all the databases**', oracle output corresponds to MT). Those rephrasings have a negative impact on automatic evaluation. The semantic and stylistic necessity of those changes needs further investigation.

In comparison, stylistic changes within NP and N groups are quite rare (e.g., PLS: "the Canadian Institutes of Health Research", MT: "*la Canadian Institutes of Health Research*" 'the Canadian Institutes of Health Research', Or.: "*la Canadian Institutes de recherche en santé de recherche*" 'The Canadian Institutes of research in health of research', Ref.: "*les instituts de recherche en santé du Canada*" 'the institutes of research in health of Canada').

Thus, residual errors mainly concern the translation of complex terminological constructions, precision in the translation of terms and professional jargon. Errors while translating verbs are caused by structural and lexical peculiarities of the source language.

4. Which kinds of residual errors could be potentially resolved by the high-performance system given its training data?

We also performed a manual analysis of the oracle improvements to the "worst" translated unit occurrences per target group (ΔM_U of about 25%). They mostly concern:

1. grammatical errors (modifications to articles or prepositions for the N and NP groups, e.g., "with taxanes", MT: "*avec taxane*" 'with taxane', Ref.: "*avec les taxanes*" 'with the taxanes', oracle output corresponds to PE; tense changes for the V group, e.g., "were excluded", MT: "*ont été exclues*" 'have been excluded', Ref.: "*étaient exclues*" 'were excluded', oracle output corresponds to PE);

2. certain reformulations (e.g., "the trial ... **showed** a clear benefit", MT: "*l'essai ...* ***a montré*** *un bénéfice clair*" 'the trial ... **has shown** a clear evidence', Ref.: "*l'essai ...* ***a mis en évidence*** *un bénéfice clair*" 'the trial ... **has highlighted** a clear evidence', oracle output corresponds to PE);

3. some terminological precision errors, including terminological constructions translated by composition (e.g., "alternative treatments", MT: "*d'autres traitements*" 'other treatements',

MT: "*des traitements alternatifs*" 'alternative treatements', oracle output corresponds to PE; "wound management properties", MT: "*la prise en charge de la plaie propriétés*" 'the management of the wound any properties', Ref.: "*les propriétés*" 'the properties', oracle output corresponds to PE);

4. minor (rarely major) reformulations and restructurings (e.g., "a one-day training course on **how to resuscitate** newborn babies", MT: "*un schéma d'évolution de formation* ***sur la façon de réanimer*** *des nouveau-nés*" 'a scheme of development of training **on the way to resuscitate newborns**', Or.: "*un schéma d'évolution de formation* ***sur la réanimation*** *des nouveau-nés*" 'a scheme of development of training **on the resuscitation** of newborns', Ref.: "*une formation d'un jour* ***sur la réanimation*** *des nouveau-nés*" 'a training of one day **on the resuscitation** of newborns').

Finally, it seems that some improvement to `Cochr-SMT` output is possible by applying a set of source rewriting rules. But taking into account the minor improvement potential as shown by our oracle study, as well the absence of distinct PE patterns (see section 4.3.3), cost-benefit trade-offs of the development and testing of such rules are not worthwhile.

As a summary, we can enumerate the following main translation difficulties faced by `Cochr-SMT`:

1. term and professional jargon translation precision;
2. translation of complex terminological constructions;
3. translation of source-specific syntactic/stylistic constructions requiring target-side reformulation;
4. translation of verbs (grammatical/stylistic variant).

4.4 Summary

This chapter has given a brief introduction to the automatic evaluation of MT. This evaluation can be performed by standard automatic metrics that require human references and provide a general evaluation score; by automatic error analysis based on traces of those metrics; as well as by QE methods that use Machine Learning techniques to predict MT quality without any reference translation. For the high-quality translation evaluation, human intervention into the evaluation process, as well a focus on the evaluation of some specific linguistic phenomena plays an important role.

We have presented the context of our work with high-quality MT of medical research reviews at Cochrane France. The main peculiarity of this context is that PE of this MT is performed by volunteer domain specialists. We have given a brief overview of Cochrane corpora and specificities of Cochrane review abstracts with their internal stylistic variety (popular scientific vs. scientific style). A detailed description of a narrowly-specialized high-quality Cochrane SMT system was provided. We have shown that more control over the PE performed by lay post-editors is needed since non-professionals tend to introduce term inconsistency errors, which can disturb comprehension of resulting text. Moreover, further improvement of these high-quality translations is also necessary, since post-editors also tend to leave MT errors uncorrected.

Automatic evaluation techniques are good for shallow comparison, but are not informative enough to guide the improvement of systems that are already of a high quality. To tackle the issue we have developed our own fine-grained methodology to diagnose high-performance MT. The methodology searches for an interconnection between residual errors, source phenomena and system parameters, such as the original corpus quality and the scoring procedure. It uses PE traces and QE techniques, and provides some necessary hints how to better detect translation difficulties and identify their reasons.

We have found that the residual errors of the Cochrane PBSMT system most significantly concern terminology and professional jargon. Syntactic and stylistic peculiarities of the source language, often requiring reformulations on the target side, constitute the other main difficulty. We tend to relate those difficulties to the nature of the medical translation task, since they are not specific to the high-performance system. They are caused by the original corpus limitations (absence of the "correct" translation in the training data), as well as to the limitations of SMT in general. Those limitations include the inability to resolve structural differences between languages or to take the more distant context into account.

The indicated issues can be partially solved by *ad hoc* solutions (e.g., model separation to resolve stylistic differences, rule-based rewriting of source sentences, etc.). However, the most important issues related to the translation of terms, crucial for the correct understanding of Cochrane texts, can be resolved only by the introduction of external knowledge. This knowledge can be reliably provided by human experts.

5 | Detection of Translation Difficulties

Contents

In Chapter 4 we have shown that potential translation difficulties faced by a high-quality translation system are quite diverse. They range from mere unseen grammatical forms and structural source language specificities (easily resolved by post-editors) to in-domain terms and jargon peculiarities, or even genre-dependent stylistic variations (which can be resolved only by a terminology expert). The latter difficulties prevail and their resolution is crucial for spreading reliable medical information by means of Cochrane review abstracts. Given the variety of those terminological difficulties and their unsystematic nature, we will attempt to resolve the difficulty detection task as a binary Machine Learning classification task using best practices of Quality Estimation (QE).

We have also seen that most of these difficulties can be handled at the sentence level. We consequently target their detection at the subsentential level taking only the context of a sentence

into account.

In this chapter, we seek to answer the following questions: (a) Can translation difficulties be reliably identified? if yes, how and at which level: word or phrase? (b) Can translation difficulties be attributed only to the source language, or do they also depend on the target language? What are the perspectives of translation difficulty detection in a multilingual setting?

The main contributions of the chapter are twofold: (a) a system-independent methodology for translation difficulty detection; (b) a study of translation difficulties in a multilingual scenario (to our knowledge, this is the first attempt of the kind).

This chapter will detail our difficulty detection approach. To be precise, we introduce an operational notion of *ex-ante* translation difficulty and our gold label annotation procedure in section 5.1.1. We will consider several segmentations at the subsentential level and experiment with multiple sets of features (see sections 5.1.1 and 5.1.2). We then will position our approach as a derivation and an improvement of existing solutions (see sections 5.2)

We apply our methodology and search for the best way to detect translation difficulties in the `MEDICAL` domain (see section 5.3). We then turn our attention to a multi-target scenario (see section 5.4). Based on a carefully designed experimental set-up, we provide an analysis of common source-side translation difficulties for different language pairs (see section 5.4.4) and show that those difficulties are dependent on the language pair (see section 5.4.5).

5.1 Methodology

As common for subsentential QE tasks, we cast the translation difficulty detection task as a binary classification problem. At first we define the "quality" labels that are used: this is because labeling segments as "correctly" or "incorrectly" translated only works on the target side and does not apply in our setting. Our proposal, inspired by Mohit and Hwa [2007], uses a binary labeling, marking phrases as either `difficult-to-translate` (`DT`) or `easy-to-translate` (`ET`).

The translation difficulties we target are system-related. We will assume that we have access to some "draft" MT output $\mathbf{e}^1$ (1-best translation hypothesis) of a source sentence $\mathbf{f}$ with a reference sentence $\hat{\mathbf{e}}$.

We also assume an alignment $\mathbf{a}^1$ between the words f_p of the source sentence and the words e_m^1 of the 1-best hypothesis. This word alignment can be used to derive the 1-best translation $e_r^1 \cdots e_g^1 = e_{[r:g]}^1$ of arbitrary segments $f_{[k:b]}$ of the source sentence.

We next describe our preprocessing procedure (labeling and segmentation) for the dataset

used to train the difficulty detection classifier, as well as the sets of features and the training procedure.

5.1.1 Gold Annotations and Segmentations

Knowing both $\hat{\mathbf{e}}$ and the alignment $\mathbf{a}^1$ between $\mathbf{f}$ and $\mathbf{e}^1$, we label the training data using the word alignment $\mathbf{e}^1 \rightarrow \hat{\mathbf{e}}$ computed by TER. In this alignment, each word-to-word connection is labeled with a type of post-editing operation applied to e^1_m to obtain $\hat{e}_v$: match (M), shift (SH), substitution (S) or deletion (D). From the resulting 3-way alignment $\mathbf{f} \rightarrow \mathbf{e}^1 \rightarrow \hat{\mathbf{e}}$, these PE labels can be projected back onto each token f_p as follows (see Figure 5.1):

- if f_p only aligns with words labeled as M, f_p is labeled as `ET`, meaning that f_p was correctly translated;
- in all other cases, f_p is marked as `DT`.

We then turn these word-level labels into segment-level labels. Assuming a segmentation π of $\mathbf{f}$, we label each segment $f_{[k:b]}$ as `DT` if at least 50% of $f_k \cdots f_b$ are labeled as `DT`; the remaining unlabeled segments are labeled as `ET`. This approximation is based on a practical consideration: pre-translating segments with less than 50% "bad" words will be not worthwhile in terms of cost-benefit trade-offs. The human effort of providing pre-translations for those segments will be comparable to the one for "worse" segments, but the final gain in quality will be minor. Thus, we believe that the standard QE approximation, labeling a segment as `BAD` if at least one of its words is badly translated,[1] is not suitable for our case.

We will contrast 3 strategies to determine source segments $f_{[k:b]}$: the word segmentation WORD-SEG, where all the segments contain exactly one word; the segmentation induced by the MT decoder (MT-SEG); and a syntactically-motivated segmentation obtained by shallow parsing (chunking) (SYNT-SEG).

It should be mentioned that our labeling strategy is prone to errors caused by the noise in automatic alignments: those produced by the decoder, and those due to the TER computation. Thus, for non-aligned source auxiliary words, we had to make a rather arbitrary choice and label them as `DT`, where the alternative label could also have been justified. Noise in TER alignments has little effect on the quality of labeling: a word translated with an error is labeled as `DT` no matter if its alignment is correct or not. For instance, Figure 5.2 illustrates our labeling procedure for WORD-SEG: the word "eye" is marked as `DT` (as its translation is substituted by TER), even

[1] This approximation is used, for instance, in WMT shared QE tasks [Bojar et al., 2017a].

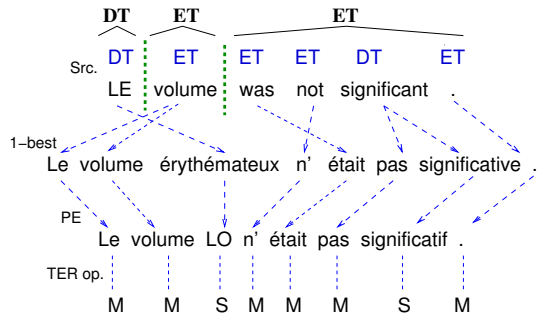


Figure 5.1 – Labeling the sentence "LE volume was not significant" segmented with MT-SEG

if the TER alignment link to the reference word "*jugement*" 'judgement' is obviously wrong and "eye" should have been aligned to "*oculaires*" 'ocular'.

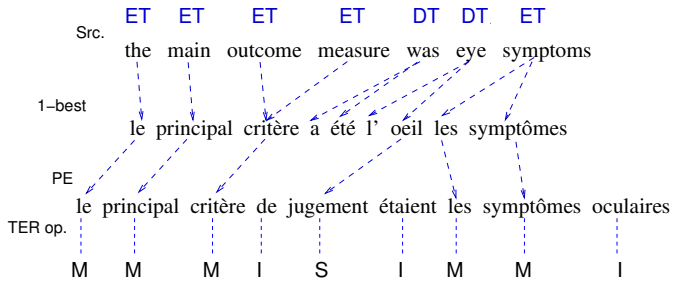


Figure 5.2 – Labeling the phrase "the main outcome measure was eye symptoms reported by participants" segmented with WORD-SEG

We admit that labeling segments according to the percentage of correctly translated words is rather naive: semantic importance estimations should be taken into account in the future. For instance, for our strategy, a segment containing two words, a determiner and a noun, is marked as `DT` independent of the fact if the determiner or the noun is `ET`. However, a correct translation of a noun is clearly more important.

Concerning the segmentation types, we believe that SYNT-SEG has more potential to "correctly" handle difficulties in labeling segments. WORD-SEG is prone to tokenization errors and does not always operate with minimal sense-bearing units. MT-SEG segments are idiosyncratic, whereas SYNT-SEG chunks are consistent. Figure 5.3 illustrates labeling a sentence with MT-SEG and SYNT-SEG. We can see that for the former case the segment "useful in the" contains two mistranslated words ("in", "the") and is consequently marked as `DT`. At the same time, the difficulty lies in translating the ambiguous English preposition "in", which represents a separate segment of one word for SYNT-SEG. For SYNT-SEG, "the" belongs to the immediately following NP

("the treatment of agitation"), which is marked as ET with the majority of correctly translated words. Moreover, using SYNT-SEG ensures that human translators will provide translations for grammatical phrases, which is probably easier than doing so for random chunks of words.

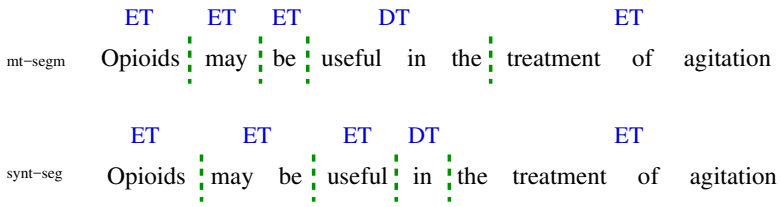


Figure 5.3 – Labeling the sentence "Opioids may be useful in the treatment of agitation" segmented with MT-SEG and SYNT-SEG

5.1.2 Main Features

In our experiments we mainly extract the standard word- and phrase-level features used in the WMT Quality Estimation (QE) tasks [Bojar et al., 2016a]. Those standard features are typically extracted from MT output. In our setting, we instead extract most features from MT *input*.

At the *word level* we distinguish 27 *black-box* and 1 *glass-box* feature. The former comprises the following groups of word-level features (we will illustrate the feature extraction procedure for the word "injections" from the sentence "5-FU injections after glaucoma surgery."):

- 3 *basic features* (`bs`): f_p, its lemma and its POS tag (for the running example: "injections", "injection" and "NNS" correspondingly);
- 20 *standard features* (`st`): this is the baseline set of the WMT'16 word-level QE task [Bojar et al., 2016a] (for the full list of features see Appendix E.1) (e.g., "5-FU"", "after" as the left and right source token contexts respectively; "*injections*" as the target token);
- 3 *syntactic features* (`snt`): the shallow parsing tag and the input dependency label of f_p, the depth of f_p in the dependency tree (distance from the root); these features are obtained from the dependency tree of $\mathbf{f}$ (for the running example: "root", "NP" and 1 correspondingly);
- 1 *term feature* (`trm`): a binary feature indicating whether a word is a term or a part of a compound term. The term mapping was performed with the `Metamap` tool for medical texts (see section 4.3.4) (1 for the running example).

We also extract the number of possible translations of f_p, as defined by the lexical translation probability model of a system as a **glass-box** feature (`trans-gl`) (117 for the running example).

At the *phrase level* for each segment $f_{[k:b]}$ we extract 72 **black-box** features and 1 **glass-box** feature (we will illustrate the feature extraction procedure for the MT-SEG segment "after glaucoma surgery" from the same sentence "5-FU injections after glaucoma surgery."):

- 8 *basic features* (`bs-phr`): the sequence of words, their hypothesis translation, lemmas and POS, plus the left (f_{k-2}, f_{k-1}) and right (f_{b+1}, f_{b+2}) contexts (for the running example: "after glaucoma surgery", "*après chirurgie du glaucome*", "after glaucoma surgery", "IN NN NN", "5-FU", "injections", "." and "</s>" respectively);
- 59 *standard phrase-level features* (`st-phr`): phrase-level features used in the WMT'16 QE task, excluding the majority of POS features except for the target POS tag sequence of the aligned $e_{[r:g]}$ (these are numerical features, see full list in Appendix E.2), for instance, the percentage of distinct trigrams seen in a source corpus (0.5 for the running example);
- 4 *syntactic features* (`snt-phr`): the constituency label covering the longest span of $f_{[k:b]}$, the percentage of words whose syntactic heads are outside the boundaries of $f_{[k:b]}$, and the maximum and minimum depth of $f_z, k \leq z \leq b$ in the dependency tree (for the running example: "NP", 0.3 (one word "surgery" has its head "injections" outside the segment boundaries), 2 and 3 respectively);
- 1 *term feature* (`trm-phr`): the percentage of $f_z, k \leq z \leq b$ marked as terms, computed as described for the word-level features (0.3 for the running example).

The only **glass-box** feature is the percentage of OOV words in $f_{[k:b]}$ (`trans-gl-phr`) (0 for the running example).

5.1.3 Classification Algorithms

In our choice of a classification algorithm we were again guided my modern QE practices. We use Support Vector Machines (SVMs) [Cortes and Vapnik, 1995] as the algorithm used in the pioneering work of Mohit and Hwa [2007] on translation difficulty detection. We also experiment with Conditional Random Fields (CRFs) [Lafferty et al., 2001] and Random Forests (RFs) [Breiman, 2001]. We also experiment with Feedforward Neural Networks (FFNNs) [Rosenblatt, 1957] as the best performing algorithm in the related work on difficulty detection of Cheng et al. [2016].

CRFs are a state-of-the-art solution in many sequence labeling tasks since they take context into account. Translations of words may be interdependent and word-level labels can influence each other, thus CRFs perfectly fit our prediction task at the word level (WORD-SEG). Phrases are supposed to capture dependencies between words their contain, their translations are less interdependent, hence other algorithms like RFs or FFNNs that handle each instance separately may be more efficient for predicting at the phrase level (in particular, for our segmentations MT-SEG and SYNT-SEG).

5.2 Detection of Difficulties as a Classification Problem

The notion of subsentential translation difficulty as a system-related measure was first introduced by Mohit and Hwa [2007]. In our attempt, we largely follow their conception and their resolution protocol.

Mohit and Hwa [2007] cast the task of detecting difficult phrases as a binary classification task (`difficult` vs. `not-difficult-to-translate`) using syntactically-motivated segmentations. They use the following notion of difficulty: a phrase is marked as `difficult-to-translate` if the removal of its translation from a hypothesis has a positive impact on the resulting BLEU score (calculated against a correspondingly modified reference). This procedure requires the knowledge of two bilingual word alignments: the source→hypothesis alignment (typically provided by the SMT decoder) and the source→reference alignment. Mohit and Hwa [2007] consider overlapping parse tree constituents whose string span is between 25% and 75% of the full sentence length. Additionally, a constituent has to be at least two levels above the tree yield (span) and two levels below the root (an average phrase length of 8.8 words). Table 5.1 illustrates the procedure for the sentence "The evidence was of moderate quality, as the visual acuity measurement was unmasked." One candidate phrase is extracted using the `Stanford Parser` [Chen and Manning, 2014] tool. The phrase is accompanied by a correspondingly modified hypothesis and a modified reference translation, which are used for the computation of BLEU.[2]

The authors use a set of only 18 features. Most of the features are system-independent. Mohit and Hwa [2007] found that the lexical ambiguity feature, reflecting the entropy of system translation choices for a phrase, as well as syntactic features were the most contributing. For

[2]For each candidate phrase Mohit and Hwa [2007] re-compute the BLEU score using the whole test set as the BLEU metric is not adapted to the evaluation of isolated sentences (it computes a geometric mean of n-gram precisions, if a higher order n-gram precision of a sentence is 0, then the BLEU score of the entire sentence is 0). Only the scores that are greater than the baseline score by a certain tunable threshold value are considered for labeling.

this definition of difficulty, SVMs performed reliably (average accuracy of 72%).

src.	The evidence was of moderate quality, as the visual acuity measurement was unmasked.
hyp.	*La preuve a été modérée de qualité, comme la mesure de l'acuité visuelle a été démasqué.*
ref.	*Les preuves étaient de qualité modérée, comme la mesure de l'acuité visuelle n'était pas masquée.*
candidate	the visual acuity measurement was unmasked
mod. hyp.	*La preuve a été modérée de qualité, comme.*
mod. ref.	*Les preuves étaient de qualité modérée, comme.*

Table 5.1 – Illustration of the procedure of translation difficulty detection proposed by Mohit and Hwa [2007]

Another closely related work is that of Cheng et al. [2016]. The authors also assume that translation difficulties are directly related to translation quality. According to their definition, a source phrase is marked as `difficult-to-translate` if, after constraining its translation, the quality of a re-translated sentence (measured in BLEU) is significantly improved as compared to results obtained for other phrases of this sentence (maximum indirect effect of re-translation). Cheng et al. [2016] explore only the phrase-level segmentations as provided by the SMT decoder (see section 2.2). They use a set of 20 features with 6 system-dependent features. They show the best prediction performance for FFNNs.

We also make an assumption that translation difficulties are directly related to translation quality, marking phrases as `difficult-to-translate` (`DT`) or `easy-to-translate` (`ET`). We use the monolingual automatic hypothesis→reference alignment as produced by TER to detect "incorrectly" translated segments (see section 5.1.1). This does not eliminate the bias introduced by the automatic metric, but allows us to use monolingual automatic alignments instead of bilingual alignments. Those alignments are more reliable and take less time to produce.

We do not use the two above-mentioned approaches as baselines in our work since their `DT` definitions are crucially different to ours. Mohit and Hwa [2007] consider rather long idiosyncratic syntactic units, which are more suitable for the local resolution at the sentence level. We target the detection of consistent units whose translations can be generalizable over their occurrences. Cheng et al. [2016] target the detection of a minimum of the most influential SMT phrases to reduce final PE effort, whereas we seek to detect a maximum of `DT` segments whose resolution will provide the human with optimal control over the MT process.

We extend both above-mentioned approaches in a multilingual setting, as Mohit and Hwa

[2007] consider only the Arabic-English translation direction and Cheng et al. [2016] consider only the Chinese-English direction.

In our work we follow modern practices in the related QE domain (features, classification algorithms, training data size), especially the ones introduced by phrase-level QE [Logacheva and Specia, 2015]. Our source-oriented context permits multi-target experiments: we thus consider our approach as a nice testbed for studying source-language and language-pairs effects.

From a more abstract viewpoint, our methodology is close to the Active Learning (AL) approach (see section 3.3.2). AL iteratively seeks to find the most informative tokens in a set of sentences and have them labeled, while we search for sentence segments whose "correct" translation will improve the translation of the whole sentence.

5.3 Intrinsic Evaluation: Experiments in the MEDICAL domain

In this section, we will validate our translation difficulty detection methodology in the medical domain for two English-French SMT systems, similar to the systems presented in section 4.3.3. We will contrast three segmentation strategies, each corresponding to the development of a dedicated classifier: 1 for the word-level prediction and 2 for the phrase-level predictions, respectively for MT-SEG and SYNT-SEG. We report the results of an intrinsic evaluation of these classifiers using standard metrics (the F-score per class ($F_{\texttt{DT}}$ and $F_{\texttt{ET}}$) and the macro-averaged F-score (F_{mcr})).

5.3.1 Data and Systems

We extract the classifier training data (*class*), as well as our development set and test set from Cochrane PE Corpus 1 (see section 4.3.1). The test set was used for both MT evaluation and classification evaluation (see Table 5.2). The classifier training data were annotated as described in section 5.1.1. Shallow parsing was performed using the `Stanford POS Tagger` [Toutanova et al., 2003] and the `OpenNLP` toolkit.[3]

For feature extraction we used the output of the `WMT14-SMT` system, described in section 4.3.3, as well as the output of the `Cochr-DT-SMT` system, built as described in section 4.3.3, using a part of Cochrane PE Corpus 2 (9K lines, 196K English tokens, 248K French tokens) instead

[3] `http://opennlp.apache.org/index.html`

of Cochrane PE Corpus 1.[4] For the computation of n-gram and corpus frequency statistics Cochrane Reference Corpus was used. For the extraction of LM-related features two 4-gram LMs were built from the corresponding monolingual parts of the corpus with modified Kneser-Ney smoothing using the SRILM toolkit [Stolcke, 2002].

For each system the quantity of the annotated difficulties is inversely proportional to its quality (see Tables 5.3 and 5.4), which is consistent with the labeling strategy deduced from TER errors. The analysis of the annotated DT chunks for both systems shows that most of the translation difficulties are in VPs and NPs (see Table 5.5). For instance, for Cochr-DT-SMT, NPs represent 42% of the detected chunks, 27% of them are DT.

set	#, lines	#, tok. (EN)	#, tok. (FR)
class	15K	344K	430K
dev	832	19K	26K
test	831	21K	26K

Table 5.2 – Basic corpus statistics for MEDICAL

	Cochr-DT-SMT	WMT14-SMT
BLEU	48.28	26.48
TER	0.34	0.59

Table 5.3 – Automatic evaluation of the MEDICAL systems

	Cochr-DT-SMT			WMT14-SMT		
	#	$\bar{l}$	% DT	#	$\bar{l}$	% DT
WORD-SEG	344K	1	25	344K	1	42
MT-SEG	198K	1.7	28	206K	1.7	50
SYNT-SEG	210K	1.6	26	210K	1.6	44

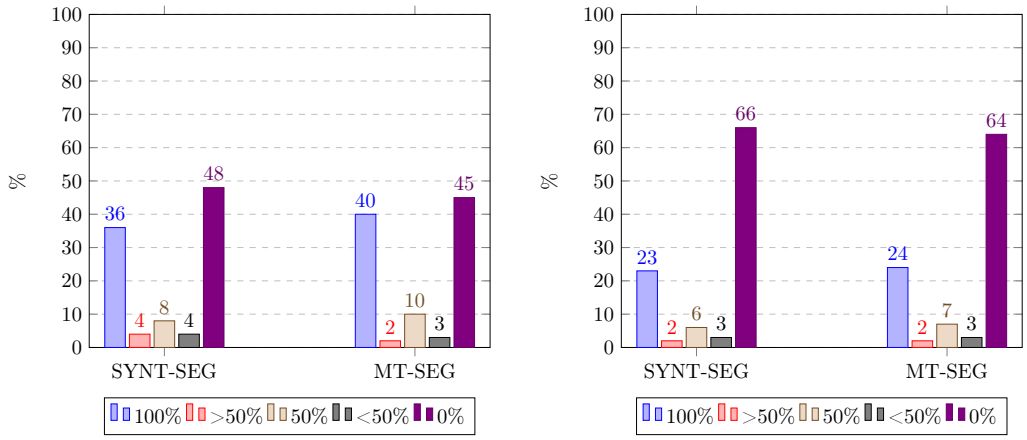
Table 5.4 – Statistics for the annotated training data (*class*) ($\bar{l}$ is the average segment length)

Figure 5.6 displays the distribution of DT and ET segments according to the percentage of DT words they contain, for both SYNT-SEG and MT-SEG, for WMT14-SMT and Cochr-DT-SMT. We can see that the vast majority of segments are unambiguously labeled as ET or DT; on average only 8% of all the SYNT-SEG and MT-SEG segments for both systems contain exactly 50% of DT words, which is the borderline case. Thus, we consider that our labeling procedure is only slightly biased by the approximations we use to define DT segments.

[4] Cochrane PE Corpus 2 is more homogeneous than Cochrane PE Corpus 1 since it was systematically translated by the latest version of Cochr-SMT. The data of Cochrane PE Corpus 1 were automatically translated by different versions of Cochr-SMT. Cochrane PE Corpus 2 was not considered in our previous experiments as its size was not sufficient by the time those experiments were designed.

tag	freq., %	Cochr-DT-SMT	WMT14-SMT
NP	42	**27**	**50**
PP	20	28	44
O	19	8	12
VP	14	**43**	**69**
ADJP	2	39	58
ADVP	2	37	53
SBAR	0.7	29	43

Table 5.5 – % of the main DT chunk tags (SYNT-SEG)



(a) WMT14-SMT (b) Cochr-DT-SMT

Table 5.6 – Distribution of source segments in the test set according to the % of DT words they contain: MEDICAL domain

5.3.2 Choice of the Classification Algorithm

To choose a proper classification algorithm we used the output of `WMT14-SMT` and experimented with the Random Forest[5] and SVM[6] implementations available in `Scikit-learn` [Pedregosa et al., 2011], as well as the CRF[7] implementation available in `Wapiti` [Lavergne et al., 2010]. We built a Feedforward Neural Network (FFNN) with 2 hidden layers using the `Keras` toolkit.[8]. For word and phrase features representing source and target sequences, as well their contexts, we used pre-trained 300-dimensional word embeddings [Mikolov et al., 2013].[9] As a baseline, we used the random classifier implementation available in `Scikit-learn`.[10] Hereinafter, each classifier experiment was run 10 times to estimate the 95% confidence interval.

The standard word-level and phrase-level sets of features were extracted using `Marmot` [Logacheva et al., 2016] and `Quest++` [Specia et al., 2015]. Lemmatization was performed using `TreeTagger` [Schmid, 1995]. Syntactic features and POS features were extracted with the `Stanford POS Tagger` and the `Stanford Parser` [Chen and Manning, 2014].

Results of our experiments are reported in Table 5.7. All the classification algorithms outperform the baseline. RFs, CRFs and FFNNs systematically outperform SVMs: for instance, for SYNT-SEG there is a $\Delta F_{mcr} = 0.16$ decrease in prediction quality for SVMs relative to the averaged $F_{mcr} = 0.75$ for CRFs, RFs and FFNNs. There are no significant differences between RFs and CRFs in terms of performance. For WORD-SEG, all the algorithms show an average performance of $F_{mcr} = 0.76$. For the phrase-level segmentations, CRFs and FFNNs are slightly outperformed by RFs: for instance, $\Delta F_{mcr} = 0.03$ for MT-SEG. Taking those minor improvements into account, as well as the fact that RFs take segments as training examples, instead of sentences, so that data can be more easily balanced when necessary, we will use RFs in all our

[5]Grid search with 5-fold cross-validation was used to tune the following parameters to optimize F-score: the optimizing criterion, the number of estimators, the maximum depth and the minimum number of leaf samples. All other parameters are those provided by default.

[6]We use a Radial Basis Function (RBF) kernel. Grid search with 5-fold cross-validation was used to tune γ and C to optimize F-score. All other parameters are those provided by default.

[7]We used the `l-bfgs` algorithm as the optimization algorithm. All other parameters are those provided by default. All the hyperparameters were tuned on the development set.

[8]`https://github.com/fchollet/keras`

[9]The following features were represented using embedding: word f_i, its left and right context f_{i-1} and f_{i+1}, its aligned word e^1_j, its left and right context e^1_{j-1} and e^1_{j+1} for WORD-SEG; sequence $f_{[k:t]}$, its aligned translation $e_{[r:g]}$, its the left (f_{k-2}, f_{k-1}) and right (f_{t+1}, f_{t+2}) contexts for phrase-level segmentations. The length of a segment sequence was limited to 10 words, masking was used for shorter segments. The first hidden layer of our network contained the quantity of units equal to the total quantity of features (the embedding dimension was taken into account, e.g., 1822 hidden units for WORD-SEG), the second layer used twice as less units. All the layers besides the output layer used the `relu` function as the activation function. The output layer uses the `sigmoid` function as the activation function. The model is trained to minimize the binary cross-entropy loss using the Adam optimizer [Kingma and Ba, 2014]. All the hyperparameters were tuned on the development set. The best performance was achieved after 10 epochs.

[10]We used the `DummyClassifier` with default parameters.

subsequent experiments and for all the segmentation types.

	Random		
	$F_{\texttt{DT}}$	$F_{\texttt{ET}}$	F_{mcr}
WORD-SEG	0.43 ± 0.0033	0.57 ± 0.0009	0.50 ± 0.0019
MT-SEG	0.51 ± 0.0031	0.49 ± 0.0046	0.50 ± 0.0036
SYNT-SEG	0.39 ± 0.0024	0.51 ± 0.0025	0.45 ± 0.0013
	SVMs		
	$F_{\texttt{DT}}$	$F_{\texttt{ET}}$	F_{mcr}
WORD-SEG	0.58 ± 0.0	0.68 ± 0.0	0.63 ± 0.0
MT-SEG	0.68 ± 0.0	$0.56 \pm 8.4e-17$	0.62 ± 0.0
SYNT-SEG	$0.53 \pm 8.4e-17$	0.65 ± 0.0	0.59 ± 0.0
	CRFs		
	$F_{\texttt{DT}}$	$F_{\texttt{ET}}$	F_{mcr}
WORD-SEG	0.73 ± 0.0	$0.79 \pm 6.9e-17$	$0.76 \pm 6.9e-17$
MT-SEG	0.75 ± 0.0	0.74 ± 0.0	0.75 ± 0.0
SYNT-SEG	$0.72 \pm 6.9e-17$	$0.76 \pm 6.9e-17$	$0.74 \pm 6.9e-17$
	RFs		
	$F_{\texttt{DT}}$	$F_{\texttt{ET}}$	F_{mcr}
WORD-SEG	0.72 ± 0.0006	0.80 ± 0.0006	0.76 ± 0.0006
MT-SEG	$\mathbf{0.78 \pm 0.0010}$	$\mathbf{0.77 \pm 0.0009}$	$\mathbf{0.78 \pm 0.0009}$
SYNT-SEG	$\mathbf{0.74 \pm 0.0011}$	$\mathbf{0.79 \pm 0.0006}$	$\mathbf{0.77 \pm 0.0008}$
	FFNNs		
	$F_{\texttt{DT}}$	$F_{\texttt{ET}}$	F_{mcr}
WORD-SEG	0.71 ± 0.0018	0.78 ± 0.0021	0.75 ± 0.0016
MT-SEG	0.76 ± 0.0026	0.75 ± 0.0025	0.75 ± 0.0017
SYNT-SEG	0.72 ± 0.0060	0.76 ± 0.0037	0.74 ± 0.0026

Table 5.7 – Performance of the classification algorithms

5.3.3 Classifier Feature Evaluation

The intrinsic evaluation results showed that for the lower quality `WMT14-SMT` translation difficulties can be reliably identified for all the segmentation strategies (average $F_{mcr} = 0.77$). For the higher quality `Cochr-DT-SMT` the prediction quality is lower: average $F_{mcr} = 0.68$ (see Tables 5.8, 5.9, 5.10). For this system with unbalanced quantity of examples per class, as common practice in Machine Learning, we artificially balanced the quantity of examples by removing the least frequent examples of `ET`. This resulted in a reduction of around 50% of the initial training data and a systematic improvement of prediction quality: for instance, for WORD-SEG $\Delta F_{mcr} = 0.10$. As a control experiment to ensure the comparability of results for the two systems, we tested how the reduction of the training data will influence the performance for `WMT14-SMT` with balanced quantity of examples per class. To do this, we randomly selected the same quantity of training

data for WMT14-SMT. This resulted in a minor decrease of prediction quality, as compared to the prediction quality, when using the full training set (e.g., $\Delta F_{mcr} = 0.02$ for MT-SEG).

The prediction performance at the word level is similar to the average prediction performance at the phrase level: an average decrease of $\Delta F_{mcr} = 0.01$ is observed when we project the predicted word-level labels to the phrase level.

According to our feature ablation experiments, the set of standard features turned out to be the most helpful for both systems and for all the segmentation strategies (e.g., $\Delta F_{mcr} = 0.08$ in the ablation experiment for WMT14-SMT MT-SEG, see Table 5.9). The other groups of features were not useful in our setting. The final set of features used in our experiments includes the basic and standard features for all the segmentation strategies. Those features are sufficient to reach an average performance of $F_{mcr} = 0.77$ for WMT14-SMT, and an average performance of $F_{mcr} = 0.68$ for Cochr-DT-SMT.

	Cochr-DT-SMT			WMT14-SMT		
set	F_{DT}	F_{ET}	F_{mcr}	F_{DT}	F_{ET}	F_{mcr}
random	0.36 ± 0.0029	0.59 ± 0.0030	0.47 ± 0.0026	0.43 ± 0.0033	0.57 ± 0.0009	0.50 ± 0.0019
all	0.57 ± 0.0009	0.75 ± 0.0006	0.66 ± 0.0007	0.72 ± 0.0006	0.80 ± 0.0006	0.76 ± 0.0006
-trans-gl	0.57 ± 0.0010	0.75 ± 0.0005		0.71 ± 0.0010	0.80 ± 0.0006	
-trm	0.57 ± 0.0008	0.75 ± 0.0006		0.72 ± 0.0007	0.80 ± 0.0005	
-snt	0.57 ± 0.0015	0.74 ± 0.0011		0.72 ± 0.0010	0.79 ± 0.0008	
-st	$\mathbf{0.52 \pm 0.0006}$	$\mathbf{0.73 \pm 0.0007}$		$\mathbf{0.61 \pm 0.0008}$	$\mathbf{0.74 \pm 0.0003}$	
-bs	0.56 ± 0.0008	0.75 ± 0.0005		0.71 ± 0.0010	0.79 ± 0.0006	

Table 5.8 – Feature ablation experiments: WORD-SEG segmentation, MEDICAL domain

	Cochr-DT-SMT			WMT14-SMT		
set	F_{DT}	F_{ET}	F_{mcr}	F_{DT}	F_{ET}	F_{mcr}
random	0.39 ± 0.0045	0.58 ± 0.0043	0.48 ± 0.0041	0.51 ± 0.0031	0.49 ± 0.0046	0.50 ± 0.0036
WORD-SEG	0.62 ± 0.0013	0.73 ± 0.0009	0.68 ± 0.0011	0.76 ± 0.0010	0.77 ± 0.0007	0.77 ± 0.0008
all	0.60 ± 0.0010	0.76 ± 0.0014	0.68 ± 0.0010	0.78 ± 0.0010	0.77 ± 0.0009	0.78 ± 0.0009
-trans-gl-phr	0.60 ± 0.0013	0.76 ± 0.0013		0.78 ± 0.0005	0.77 ± 0.0006	
-trm-phr	0.60 ± 0.0008	0.76 ± 0.0014		0.78 ± 0.0005	0.77 ± 0.0005	
-snt-phr	0.60 ± 0.0015	0.76 ± 0.0014		0.77 ± 0.0010	0.77 ± 0.0006	
-st-phr	$\mathbf{0.49 \pm 0.0019}$	$\mathbf{0.77 \pm 0.0006}$		$\mathbf{0.69 \pm 0.0010}$	$\mathbf{0.71 \pm 0.0007}$	
-bs-phr	0.61 ± 0.0007	0.77 ± 0.0005		0.77 ± 0.0007	0.77 ± 0.0007	

Table 5.9 – Feature ablation experiments: MT-SEG segmentation, MEDICAL domain (WORD-SEG denotes results for the predicted word-level labels projected to the phrase level)

5.4 Intrinsic Evaluation: Experiments in the UN domain

We now turn our attention to assessing the performance of our translation difficulty detection methodology in a multi-target scenario. This scenario perfectly fits our Cochrane setting, since

	Cochr-DT-SMT			WMT14-SMT		
set	F_{DT}	F_{ET}	F_{mcr}	F_{DT}	F_{ET}	F_{mcr}
random	0.38 ± 0.0041	0.58 ± 0.0023	0.48±0.0026	0.39 ± 0.0024	0.51 ± 0.0025	0.45±0.0013
WORD-SEG	0.60 ± 0.0007	0.73 ± 0.0006	0.67±0.0006	0.73 ± 0.0010	0.79 ± 0.0008	0.76±0.0009
all	0.60 ± 0.0008	0.78 ± 0.0006	0.69±0.0006	0.74 ± 0.0011	0.79 ± 0.0006	0.77±0.0008
-trans-gl-phr	0.60 ± 0.0010	0.78 ± 0.0009		0.74 ± 0.0007	0.79 ± 0.0005	
-trm-phr	0.60 ± 0.0006	0.78 ± 0.0006		0.74 ± 0.0005	0.79 ± 0.0006	
-snt-phr	0.60 ± 0.0007	0.78 ± 0.0004		0.75 ± 0.0009	0.79 ± 0.0008	
-st-phr	$\mathbf{0.51 \pm 0.0017}$	$\mathbf{0.75 \pm 0.0013}$		$\mathbf{0.63 \pm 0.0018}$	$\mathbf{0.74 \pm 0.0007}$	
-bs-phr	0.60 ± 0.0010	0.78 ± 0.0008		0.74 ± 0.0010	0.79 ± 0.0007	

Table 5.10 – Feature ablation experiments: SYNT-SEG segmentation, MEDICAL domain (WORD-SEG denotes results for the predicted word-level labels projected to the phrase level)

Cochrane source texts are produced in English and translated into 16 different languages, including French, German, Russian, Spanish, etc. It also opens a range of perspectives in terms of translation quality improvement: if translation difficulties largely depend on the source language (English in our case), they can be resolved once and for many languages (for instance, segments difficult-to-translate for one language can be resolved using translations into other languages, where those segments may be easy-to-translate).

In order to make our results interpretable, we needed parallel multi-target data sharing the same source. In the absence of such Cochrane data (each review abstract is translated into a different set of target languages), we used the multilingual *MultiUN* parallel data [Eisele and Chen, 2010; Tiedemann, 2012] (Arabic, English, French, Russian, Spanish), a corpus extracted from the official documents of the United Nations (UN). We carefully designed an experimental setting where each language pair was trained using *exactly* the same source data. In this multilingual setting, we will focus our attention on WORD-SEG and SYNT-SEG as language-independent and more human-friendly segmentations.

In our study we attempted to answer the following questions:

- Are translation difficulties more defined by peculiarities of only the source language or of the language pair? What is the proportion of source-side translation difficulties common for different language pairs? What is the nature of those difficulties? For the English language, these difficulties can be, for instance, the translation of polysemous prepositions (e.g., “of”, “to”, “for”), compound words (e.g., “six-year-old”, “research-based”), which require target-side reformulations, expletive constructions (e.g., “there is/are”), certain constructions (e.g., “have something done”, when we talk about someone doing something for us, which we ask them to do), etc.

- What are the differences in source-side translation difficulties for various language pairs?

Here, we are particularly interested to study those differences for high-quality MT because it is relevant for the Cochrane context.

5.4.1 Features

In the multilingual UN setting we solve the classification task using again RFs. We experiment with different feature sets with a special focus on the separation of source and target features. In this configuration we will be able to investigate how much translation difficulty can be attributed to the source language, and for how much difficulty the target language is responsible.

At the word level we extract 54 **black-box** and **glass-box** features: (a) 18 *source* features (SRC-wrd): the set of 3 basic features (f_p, its lemma and its POS tag), 12 standard source features (for the full list of features see Appendix E.1), syntactic features (the shallow parsing tag, the input dependency label of f_p, the depth of f_p in the dependency tree); (b) 9 *target* features per language pair (e.g., ES-wrd, FR-wrd etc.): 8 standard target features (for the full list of features see Appendix E.1) and 1 binary OOV `trans-gl` feature (for more details see section 5.1.2).

We extract 189 **black-box** and **glass-box** features for each segment $f_{[k:b]}$ per language pair: (a) 33 *source* features (SRC-phr): 7 basic features (the sequence of words, lemmas and POS, plus the left (f_{k-2}, f_{k-1}) and right (f_{b+1}, f_{b+2}) contexts), 22 phrase-level standard source features (for the full list of features see Appendix E.2), 4 syntactic features (the constituency label covering the longest span of $f_{[k:b]}$, the percentage of words, whose syntactic heads are outside the boundaries of $f_{[k:b]}$, and the maximum and minimum depth of $f_z, k \leq z \leq b$ in the dependency tree); (b) 39 *target* features per language pair (e.g., ES-phr, FR-phr etc.): the hypothesis translation $e_{[r:g]}$, 37 phrase-level standard target features (for the full list of features see Appendix E.2) and the percentage of OOV in $f_{[k:b]}$ (for more details see section 5.1.2).

5.4.2 Data

We applied in-house scripts for cleaning and removing duplicate lines of the *MultiUN* parallel data. We used the `Stanford Tokenizer` toolkit[11] for the majority of languages, and `TreeTagger` for Russian [Schmid, 1995; Sharoff et al., 2008]. For Arabic we used the `Stanford Word Segmenter` [Monroe et al., 2014]. The final data set included the English side intersection for the 4 language pairs (English-Arabic, English-French, English-Russian, English-Spanish).

[11] `http://nlp.stanford.edu/software/tokenizer.shtml`

Shallow parsing of the source was performed as described in section 5.3.1.

The resulting data were separated into MT and *class* training data, development and test sets. The development and test sets were used both for MT and classification evaluation (see Table 5.11).

set	#, tok.					
	#, lines	EN	AR	ES	FR	RU
MT	5.7M	164M	171M	189M	193M	149M
class	15K	430K	449K	494K	507K	391K
dev	1K	29K	30K	33K	34K	26K
test	1K	29K	30K	33K	34K	26K

Table 5.11 – Basic corpus statistics for `UN`

5.4.3 System Building

We built four PBSMT systems (English-Arabic, English-French, English-Russian and English-Spanish) using a sampling strategy to select training data. This was done in order to be able to test our difficulty detection methodology for conditions with reduced variability of translation options. To be more precise, in a first step, the suffix array structure was used to index the parallel data, as well as the corresponding word alignments (see section 3.3.1). These word alignments were computed using `fast_align` [Dyer et al., 2013]. For each source phrase, a limited number of translation examples (100 in our case), selected by *deterministic random sampling* [Callison-Burch et al., 2005], were then used to compute standard translation model scores (see section 2.3). As the inverse translation probability can not be computed exactly (because sampling is performed independently for each source phrase), the following approximation was used:

$$\phi(\bar{f}|\bar{e}) = \min\left(1.0, \frac{\phi(\bar{e}|\bar{f}) \times fr(\bar{f})}{fr(\bar{e})}\right) \tag{5.1}$$

where the $fr(\cdot)$ is the number of occurrences of the given phrase in the whole corpus, and the numerator $\phi(\bar{e}|\bar{f}) \times fr(\bar{f})$ represents the predicted joint count of $\bar{f}$ and $\bar{e}$ [Gong et al., 2014].

This provided us with a multilingual range of systems (see Table 5.12). We built 4-gram LMs trained with modified Kneser-Ney smoothing on the target parts of the MT training data using the `SRILM` toolkit. Tuning was performed using `kb-mira` with default options on 300-best hypotheses.

The *class* training data were annotated as described in section 5.1.1 (see Table 5.13). Again

for each system the quantity of the annotated difficulties is inversely proportional to its quality.

Observations regarding the percentage of predicted DT words in segments produced by SYNT-SEG and MT-SEG confirm our intuition that SYNT-SEG better models translation difficulties. For instance, for RU and FR SYNT-SEG segments are less ambiguous: we observe on average 9% more segments unambiguously labeled as DT for SYNT-SEG than for MT-SEG (see Figure 5.14).

	AR	ES	FR	RU
BLEU	38.7	50.3	45.1	36.8
TER	0.51	0.39	0.45	0.54

Table 5.12 – Automatic evaluation of UN systems

			%, DT			
	#	$\bar{l}$	AR	ES	FR	RU
WORD-SEG	420K	1	47	33	38	52
SYNT-SEG	261K	1.6	42	30	35	47

Table 5.13 – Statistics for the annotated training data (*class*) for UN ($\bar{l}$ is the average segment length)

5.4.4 Source Translation Difficulty Analysis

We will now analyze the annotated multi-target translation difficulties for the classifier training corpus both at the word level and at the phrase level.

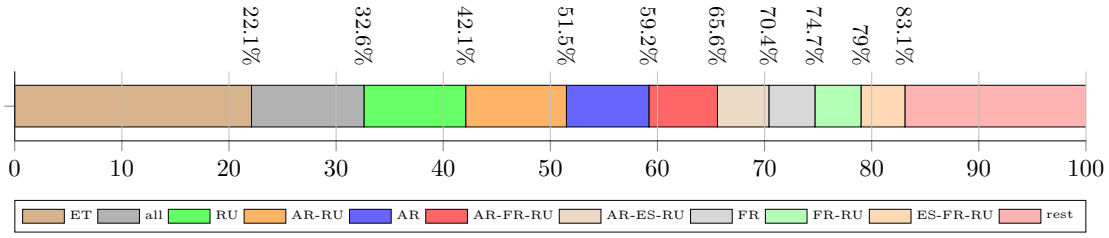
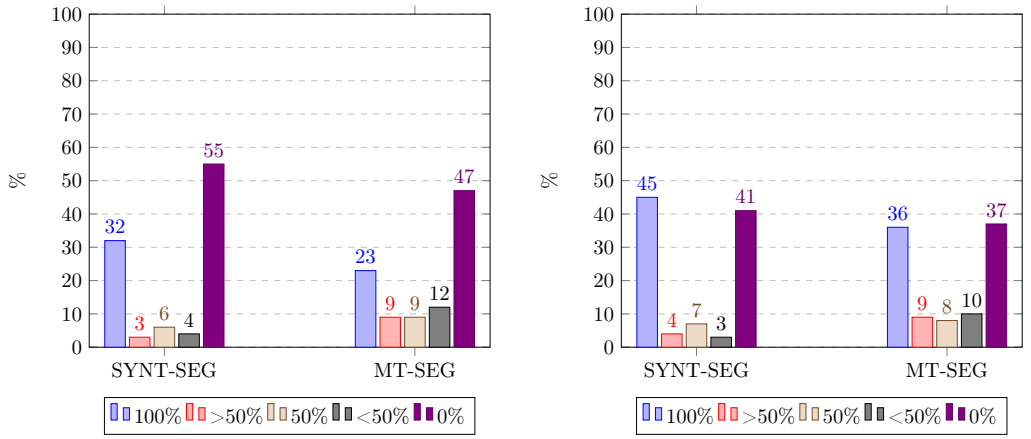


Figure 5.4 – Distribution of common DT word occurrences in the training set (ET occurrences are never difficult in any language)

Figures 5.4 and 5.5 present the distributions of common DT and ET segments for all the language pairs. The DT occurrences common to all the target languages make only around 10.5% of all the source words (8.7% of the chunks). The quantity of ET segments common for all the languages is twice as large: 22% of the words, 27% of the chunks.

Other significant groups of source difficulties are the difficulties of the systems of the "worst"



(a) FR (b) RU

Table 5.14 – Distribution of source segments in the test set according to the % of DT words they contain : UN

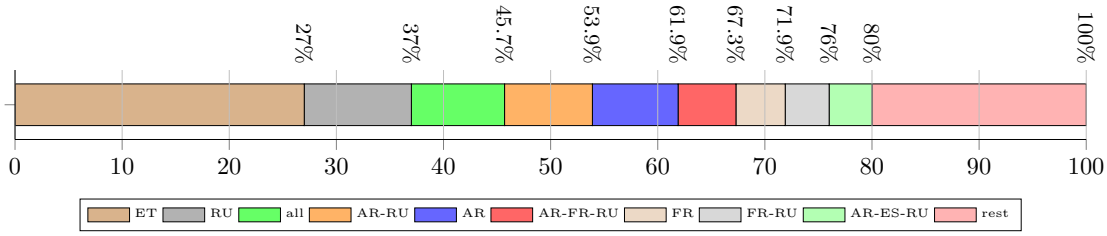


Figure 5.5 – Distribution of common DT syntactic chunk occurrences in the training set (ET occurrences are never difficult in any language)

lang	AR	ES	FR	RU
AR		21	23	31
ES	21		20	22
FR	23	20		25
RU	31	22	25	

Figure 5.6 – % of common DT words per language pair in the training set (from the total quantity of words)

lang	AR	ES	FR	RU
AR		18	20	26
ES	18		17	19
FR	20	17		22
RU	26	19	22	

Figure 5.7 – % of common DT chunks per language pair in the training set (from the total quantity of chunks)

quality (e.g., 9.5% (10%) for Russian, 9.4% (8.2%) common for Arabic and Russian, of the word and chunk occurrences respectively). The distributions do not show any significant similarities in translation difficulties for related target languages: the common Spanish and French difficulties make around 20% of all the words (17% of the chunks), which is less than the percentage of the common difficulties detected for Arabic and Russian (31% of the words, 26 % of the chunks, see Tables 5.6 and 5.7). The above observations suggest (a) that translation difficulties depend more on the language pair than solely on the source language, (b) that the "easy" languages have more easy-to-translate segments in common than difficulties.

Tables 5.15 (left, middle) show statistics over DT words and chunks common for all the language pairs. These results confirm our intuition about the nature of difficulties: the translation of highly ambiguous English prepositions (e.g., "of", "in", "to", in constructions with the case dependency label) is a recurring difficulty. Other frequent difficulties include the translation of English determiners (as it can vary greatly depending on the context), as well as the translation of the chunk "have" (a highly polysemous word, often a part of various constructions) and "there" (often a part of expletive constructions). In comparison, ET words and chunks common for all the language pairs are mainly unambiguous punctuation signs and conjunctions (see Table 5.15, right).

Tables 5.16 (left, middle) illustrate some DT POS patterns (NN, DT JJ NN) common for all the languages. Usual translation difficulties are frequent nouns or verbs (in past tense) with general meanings (e.g., "order", "development", "place", "was", "said" with the frequencies belonging to the fourth quartile as computed for the MT training corpus). For instance, the word

"place" in addition to a very general meaning (position, portion of space), is a common part of phraseological units, e.g., "in place" (established), "to put in place" (prepare), in addition, it is homonymous to the verb "to place", which contributes to the word ambiguity.

ET POS patterns common for all the languages are represented by frequent domain-specific expressions (e.g., "the report", "the Government", see Table 5.16, left). These expressions tend to have fixed translations within this text type.

We will now turn our attention to the analysis of translation difficulties that are not common for the two related languages French and Spanish. We have already mentioned that those language pairs have less common difficulties than might be expected taking their typological similarity into account (both are Romance languages). Thus, around 50% of the DT chunks for French are ET for Spanish (vice versa, around 40% of the DT chunks for Spanish are ET for French). Table 5.17, left shows statistics for the segments that are DT for French and ET for Spanish. These correspond mostly to prepositions and nouns. Table 5.18 shows an example, when the segment "of such places" is correctly translated into Spanish, but not into French ("*et tel lieu*" 'and such place').

Table 5.17 (right) shows statistics over common ET chunks for both French and Spanish. The characteristics of those segments closely resemble the ones of DT segments that are not common for both languages (they are also mostly prepositions and nouns). This makes us believe that DT differences between French and Spanish are context-dependent.

POS	dep.	f_p
IN	case	the
NN	det	to
DT	amod	of
JJ	nmod	in
VBN	root	a

POS seq.	$f_{[k:b]}$	chunk
IN	of	VP
VBD	in	NP
VBG	be	PP
NN	for	O
VBZ	to	ADVP

POS seq.	$f_{[k:b]}$	chunk
.	.	NP
IN	,	O
DT NN	and	PP
,	in	VP
CC	(	ADVP

Table 5.15 – Most frequent POS, input dependency labels, source words and syntactic chunks for the intersection AR-ES-FR-RU: DT, WORD-SEG (left); DT, SYNT-SEG (middle); ET, SYNT-SEG (right)

5.4.5 Classifier Feature Evaluation

For the intrinsic evaluation, features were extracted and classifiers were trained for each language pair as described in section 5.3.3. For feature extraction we used the MT training corpora and the target LMs trained for the corresponding SMT systems (see sections 5.4.2 and 5.4.3). For the extraction of source LM-related features, a 4-gram LM was built on the corresponding

NN	VBD	DT NN
order	was	the report
development	adopted	the Government
place	said	the implementation
action	recommended	the Conference
work	participated	the region

Table 5.16 – Examples of the most frequent POS patterns for the intersection AR-ES-FR-RU: DT, WORD-SEG (left); DT, SYNT-SEG (middle); ET, SYNT-SEG (right)

POS seq.	$f_{[k:b]}$	chunk
IN	of	NP
DT NN	,	PP
NN	in	VP
,	to	O
NNS	for	ADVP

POS seq.	$f_{[k:b]}$	chunk
IN	of	NP
,	,	PP
DT NN	in	O
NN	to	VP
CC	and	ADVP

POS seq.	$f_{[k:b]}$	chunk
IN	.	NP
.	of	O
DT NN	,	PP
,	in	VP
NN	and	SBAR

Table 5.17 – Most frequent POS sequences, segments and syntactic chunks for the following configurations: DT in FR → ET in ES (left); DT in ES → ET in FR (middle); ET in both FR and ES (right)

EN	the oversight and monitoring **of such places** is adequate
ES	*la supervisiòn y el seguimiento* ***de esos lugares*** *sea adecuada* 'the supervision and monitoring **of such sites** is appropriate'
FR	*la responsabilité du contrôle* ***et tel lieu*** *est suffisant* 'the responsibility for control **and such place** is sufficient'

Table 5.18 – Example of translation difficulties that are not common for ES and FR

monolingual part of the MT training data with modified Kneser-Ney smoothing using the SRILM toolkit.

The intrinsic evaluation results show that translation difficulties can be reliably identified for all the translation directions (average $F_{mcr} = 0.70$ for WORD-SEG and SYNT-SEG, see Table F.1 in Appendix F, Table 5.19). We artificially balanced the quantity of examples from both classes for the language pairs with a naturally unbalanced proportion of ET and DT (English-Spanish, English-French) by removing the least frequent examples of ET. This resulted in a reduction of around 34% of the initial training data and a prediction improvement of around $\Delta F_{\texttt{DT}} = 0.07$.

The prediction quality for the UN task is on average slightly lower (e.g., $\Delta F_{mcr} = 0.03$ for SYNT-SEG) than the prediction quality for MEDICAL, which we attribute to the increased syntactic complexity of the official domain texts. For UN, predicting at the word level is as difficult as predicting at the phrase level: an average $\Delta F_{mcr} = 0.015$ is observed when we project the predicted word-level labels to the phrase level.

The set of source features (SRC) is enough to reach an average performance of $F_{mcr} = 0.66$ for both WORD-SEG and SYNT-SEG. In general, adding the target features slightly improves prediction accuracy (e.g., by $\Delta F_{mcr} = 0.04$ for SYNT-SEG). At both word and phrase levels adding the features for other target languages does not improve prediction quality, which we believe can be explained by the unsystematic nature of difficulties..

5.5 Summary

In this chapter we have presented our methodology for detecting translation difficulties at the subsentential level.

Following previous approaches to difficulty detection, we cast the task as a binary classification task (difficult- (DT) vs. easy-to-translate (ET)). We have introduced our own definition of translation difficulty: difficult-to-translate segments are segments for which an MT system makes erroneous predictions. Word-level difficulty labels are projected to the phrase level. We consider words and phrases of rather short length (SMT phrases, syntactic chunks), which are consistent and can potentially influence the translation of their context. We have experimented with different classification algorithms, as well as various sets of features, including system-independent and system-dependent features.

Our approach was tested for English-French MT in the medical domain, and for multi-target MT in the official domain (translating from English into Arabic, French, Russian and Spanish).

EN-AR			
set	F_{DT}	F_{ET}	F_{mcr}
SRC-phr	0.59 ± 0.0009	0.72 ± 0.0007	0.66 ± 0.0003
AR-phr	0.64 ± 0.0015	0.77 ± 0.0007	0.71 ± 0.0007
random	0.41 ± 0.0018	0.49 ± 0.0033	0.45 ± 0.0023
WORD-SEG	0.64 ± 0.0008	0.72 ± 0.0006	0.68 ± 0.0006
all	$\mathbf{0.65 \pm 0.0016}$	$\mathbf{0.77 \pm 0.0007}$	$\mathbf{0.71 \pm 0.0008}$
-ES-phr	0.65 ± 0.0012	0.77 ± 0.0008	
-FR-phr	0.65 ± 0.0013	0.77 ± 0.0005	
-RU-phr	0.65 ± 0.0017	0.77 ± 0.0004	
-FR-phr**-**RU-phr	0.65 ± 0.0012	0.77 ± 0.0010	
-ES-phr**-**RU-phr	0.65 ± 0.0016	0.77 ± 0.0008	
-ES-phr**-**FR-phr	0.65 ± 0.0016	0.77 ± 0.0007	
EN-ES			
set	F_{DT}	F_{ET}	F_{mcr}
SRC-phr	0.57 ± 0.0008	0.75 ± 0.0010	0.66 ± 0.0007
ES-phr	0.60 ± 0.0005	0.78 ± 0.0008	0.69 ± 0.0005
random	0.36 ± 0.0026	0.54 ± 0.0029	0.45 ± 0.0025
WORD-SEG	0.60 ± 0.0009	0.74 ± 0.0006	0.67 ± 0.0007
all	$\mathbf{0.61 \pm 0.0005}$	$\mathbf{0.77 \pm 0.0007}$	$\mathbf{0.68 \pm 0.0005}$
-AR-phr	0.61 ± 0.0008	0.77 ± 0.0009	
-FR-phr	0.61 ± 0.0005	0.77 ± 0.0010	
-RU-phr	0.61 ± 0.0006	0.77 ± 0.0008	
-FR-phr**-**RU-phr	0.61 ± 0.0006	0.77 ± 0.0007	
-AR-phr**-**RU-phr	0.61 ± 0.0004	0.77 ± 0.0009	
-AR-phr**-**FR-phr	0.61 ± 0.0007	0.77 ± 0.0009	
EN-FR			
set	F_{DT}	F_{ET}	F_{mcr}
SRC-phr	0.58 ± 0.0009	0.70 ± 0.0006	0.64 ± 0.0007
FR-phr	0.61 ± 0.0008	0.74 ± 0.0008	0.68 ± 0.0007
random	0.39 ± 0.0031	0.52 ± 0.0035	0.46 ± 0.0030
WORD-SEG	0.62 ± 0.0006	0.71 ± 0.0007	0.67 ± 0.0006
all	$\mathbf{0.63 \pm 0.0005}$	$\mathbf{0.73 \pm 0.0008}$	$\mathbf{0.68 \pm 0.0006}$
-AR-phr	0.62 ± 0.0006	0.73 ± 0.0006	
-ES-phr	0.62 ± 0.0007	0.73 ± 0.0010	
-RU-phr	0.62 ± 0.0009	0.73 ± 0.0006	
-ES-phr**-**RU-phr	0.62 ± 0.0009	0.73 ± 0.0003	
-AR-phr**-**RU-phr	0.62 ± 0.0009	0.74 ± 0.0008	
-AR-phr**-**ES-phr	0.62 ± 0.0008	0.74 ± 0.0003	
EN-RU			
set	F_{DT}	F_{ET}	F_{mcr}
SRC-phr	0.69 ± 0.0005	0.66 ± 0.0020	0.67 ± 0.0012
RU-phr	0.69 ± 0.0017	0.75 ± 0.0005	0.72 ± 0.0010
random	0.46 ± 0.0041	0.46 ± 0.0031	0.46 ± 0.0031
WORD-SEG	0.72 ± 0.0006	0.70 ± 0.0006	0.71 ± 0.0005
all	$\mathbf{0.70 \pm 0.0009}$	$\mathbf{0.74 \pm 0.0008}$	$\mathbf{0.72 \pm 0.0008}$
-AR-phr	0.70 ± 0.0019	0.74 ± 0.0012	
-ES-phr	0.71 ± 0.0012	0.74 ± 0.0008	
-FR-phr	0.71 ± 0.0014	0.74 ± 0.0011	
-ES-phr**-**FR-phr	0.71 ± 0.0016	0.74 ± 0.0012	
-AR**-**FR-phr	0.70 ± 0.0010	0.75 ± 0.0007	
-AR-phr**-**ES-phr	0.71 ± 0.0006	0.75 ± 0.0006	

Table 5.19 – Feature ablation experiments: SYNT-SEG segmentation, UN domain

First, we have found that DT segments can be reliably identified, even with a relatively simple set of features, and without using any system-specific features. In general, there is no difference in prediction performance between the word and the phrase levels and predicted word-level labels can be projected to the phrase level without any major performance loss. The quality of prediction slightly varies according to the translation task (domain), but is sensitive to the translation quality of an underlying system: translation difficulties are more difficult to predict for high-quality MT (TER $>$ 0.50): on average a decrease of $\Delta F_{mcr} = 0.07$ as compared to low-quality MT (TER $<$ 0.50).

Our multi-target translation difficulty analysis has shown that difficulties seem to be more dependent on the language pair than solely on the source language, and that there is little hope to find universal source difficulties, whose resolution would help all languages across domains. However, we believe that for the medical domain common resolution of some terms can be beneficial in a multi-target scenario. For closely related languages (French and Spanish), we observed more common easy-to-translate segments than common difficulties. Difficulties that are not common to those target languages seem to be context-dependent. For instance, up to 50% of chunks that are DT in French can be corrected, knowing their translation into Spanish for which these chunks are ET. The information on differences in DT segments can be potentially used to improve translation quality in pivoting scenarios, where translation from one language into another is performed via a third language (e.g., English-Spanish-French) [Cohn and Lapata, 2007; Utiyama and Isahara, 2007; Crego et al., 2010; Durrani and Koehn, 2014].

In the next chapter, we show that given the observed prediction quality, our procedure for source-side translation difficulty detection is good enough to be integrated into a difficulty resolution protocol involving humans.

6 | Resolution of Translation Difficulties with Human Help

Contents

In the previous chapter we have shown that a simple automatic prediction methodology is sufficient to reliably detect translation difficulties at the subsentential level. The detected difficulties can be resolved by a variety of preprocessing strategies, both automatic or using human help. Automatic strategies tend to introduce noise and jeopardize the final translation quality.

The resolution of translation difficulties by humans is more certain and better fits the task of translating text types containing precise information (medical, legal, etc.). Targeted human help can consist in normalization [Seretan et al., 2014], simplification through paraphrasing [Resnik et al., 2010] or pre-translation of difficult-to-translate segments [Mohit and Hwa, 2007], etc. The first two monolingual strategies only give some hints to the machine on how to correctly translate a segment, for instance, by reformulating source segments, whereas the last solution supplies the machine with ready-to-use answers. Therefore it was chosen for the extrinsic evaluation of our difficulty detection strategy as a solution guaranteeing lexical consistency of MT output.

In this chapter, we seek to answer the following questions: (a) Is our automatic translation difficulty detection procedure efficient enough so that its results could be integrated into a PRE scenario? (b) How useful and realistic is asking the human to pre-translate difficult-to-translate segments? (c) What are the benefits of resolving difficulties before translation as opposed to resolving them during PE?

The main contributions of the chapter are as follows: (a) a proposal of a Human-Assisted MT (HAMT) protocol integrating our difficulty detection procedure and a PRE scenario for their resolution; (b) a proposal of a document-level HAMT scenario and its real-life assessment.

We will first revisit the modes of human intervention into the MT process, more specifically we will provide our motivations to prefer pre-edition (PRE, human intervention *ex-ante*) over post-edition (PE, human intervention *ex-post*) (see section 6.1). Then we introduce our Human-Assisted MT protocol (see section 6.7.1). We describe how we evaluate this protocol, as well as our way to investigate benefits of PRE (see section 6.3). As a first step, we present results of an extrinsic evaluation of our difficulty detection methodology at the sentence level for the `MEDICAL` and `UN` domains (see sections 6.5,6.6), as well as a study of PRE benefits in a simulated setting. We then focus on our target scenarios of document-level difficulty resolution for Cochrane review abstracts in a simulated setting and in a real-life setting (see section 6.7).

6.1 Pre-Edition vs. Post-Edition

We remind here that PRE is the process of supplying an MT system with information that will prevent errors in its output (human intervention before MT). PE is the process of manually correcting errors in MT output (human intervention after MT). Depending on the final translation quality requirements, PRE and PE can be combined, or applied separately (see section 3.1).

MT preceded by PRE is often referred to as Human-Assisted MT (HAMT). As opposed to

the Computer-Assisted Translation (CAT) paradigm, which holds the human responsible for producing final translation with machine help, HAMT considers that the machine uses human assistance to produce translation [Slocum, 1985].

Historically speaking, the term HAMT is often applied to the systems of the end of the previous century, which solicited input from the human by asking some specific disambiguation questions [i.a., Tomita, 1985; Brown, 1990; Blanchon, 1992]. For instance, to translate the sentence "Time flies like an arrow" the system proposed by Tomita [1985] would ask if the words "time" and "flies" are nouns or verbs. "Time" can both denote the progress of life and the act of planning. "Flies" can denote both moving through the air or insects.

As the translation process usually happens at the level of complete documents, an obvious advantage of PRE is that a text segment can be pre-edited once and for all its occurrences, whereas PE in its traditional form assumes more frustrating per occurrence corrections. Note that recent improvements to the PE task have been proposed to mitigate such problems, by immediately learning from human feedback in an online fashion (see section 3.3.1). However, the approach introduces its own risks: immediate updates, without prior feedback verification can jeopardize the resulting MT quality.

The PRE setting also provides humans with additional control, as they can actively contribute to the final MT quality by anticipating and preventing MT errors.

Another important motivation for preferring PRE over PE is the possibility of an indirect improvement in output translation: providing information for a difficult-to-translate word or phrase can have a positive impact on the automatic translation of neighboring words. Our experiments below show that these effects can be reasonably significant.

6.2 Human-Assisted Machine Translation Protocol

We propose to resolve translation difficulties (see Chapter 5) using PRE in a HAMT protocol (see Figure 6.1).

Assuming that translations are performed on a per-sentence basis, our HAMT protocol takes the following steps:

1. generate baseline translations, which are not displayed to the user;

2. detect `DT` segments in the input (see Chapter 5);

3. ask the user to provide translations for a certain amount of the DT segments displayed in their context;[1]

4. compute the actual machine translation, using the human suggestions as constraints during decoding;

5. final PE. This step is optional, depending on the final translation purpose.[2]

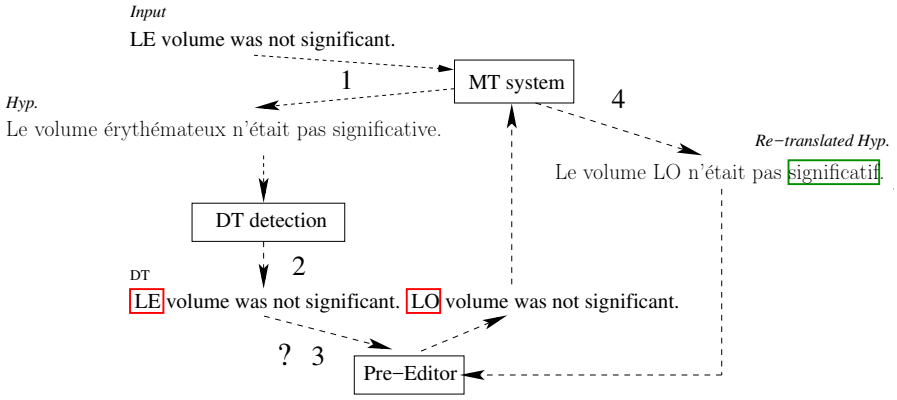


Figure 6.1 – Improving translation through HAMT. The red boxed text is pre-translated, the green boxed text is improved for free.

Mohit and Hwa [2007] also evaluate their methodology of difficulty detection at the sentence level, but do not present any protocol. Cheng et al. [2016] propose a sentence-level Pick-Revise Interactive MT protocol (PRIMT), which is supposed to break the traditional left-to-right order of interactions. In this protocol the human translator sees both the source and the first-pass MT output. He or she picks an SMT bi-phrase, which he or she considers to be the most critical translation error (optionally, also suggested by the machine), and revise its translation from the options proposed by the PT (a new translation can be input if needed as well). Then the sentence is re-translated using constrained decoding. The whole process continues until output is considered acceptable by the human. This protocol was developed as an alternative to the standard IMT procedure (see section 3.2), and involves the costly cognitive activity of repetitively analyzing MT output.

[1] In a more elaborate version, the user could select these translations from the variants proposed by an MT system, or/and from the cache of past translations of DT segments, thus potentially saving many keystrokes.

[2] E.g., for the medical domain, this step may be systematically required given the sensibility of medical information and its intended use.

6.3 Evaluation of Pre-Translation

In our experiments, the HAMT protocol was simulated and did not involve any actual human pre- or post-editor. Step 2 was performed using the methods presented in Chapter 5. Step 3 requires to automatically obtain reference translations for (some) DT segments – these were computed again using the $\mathbf{f} \rightarrow \mathbf{e}^1 \rightarrow \hat{\mathbf{e}}$ word alignments produced by TER.[3]

To measure the effectiveness of our DT detection scheme, we study the dependency between the amount of pre-translated segments (in step 3) and the overall translation quality. In each sentence, DT words or segments are ranked by their decreasing posterior class probability, meaning that segments that are most likely to be difficult are pre-translated first: we thus expect to see a sharp increase in translation quality after just a few of these difficult segments are correctly translated. Each experiment assumes that a fixed number p of DT segments is provided, where p varies between between 1 and $max_{\texttt{DT}}$, the maximum number of DT segments in a test sentence.

We also contrast the automatic DT detection with two extreme situations: in an oracle setting, we reproduce the same measurements using *reference difficulty labels* (computed as in section 5.1.1) to evaluate the difference of our method with a fully correct DT detection. Since DT segments are tagged deterministically, we use a random order for pre-translation. Our other contrast study corresponds to the case where segments are pre-translated first based on their posterior probability of being ET, thus simulating an extremely poor DT detection method.

Pre-translation is implemented in our phrase-based translation systems using the `exclusive xml-mode` of the `Moses` decoder, as the solution similar to the ones used in the closely related works of Mohit and Hwa [2007] and Cheng et al. [2016]. To ensure the comparability of the 3 segmentation strategies, for each experiment, we report in our graphs the per sentence *averages* of pre-translated words.[4]

To measure the improvement in translation quality, we mostly used TER, which is the obvious candidate to measure the residual PE effort that would be necessary to product an entirely correct translation. Note that the difference in TER between baseline outputs of step 1 and the improved outputs of a system using PRE (step 4) is due to (a) more matches, directly resulting from generating correct pre-translations; (b) indirect "contextual" improvements in the neighborhood of these good translations.

[3] For this experiment, reference words $\hat{e}$, which were mapped to non-aligned hypothesis words, as well as inserted $\hat{e}$ were aligned according to their syntactic heads. This was performed with the help of the `Stanford Parser` toolkit.

[4] We round word averages to nearest integer, which may result in several quality scores per each rounded value. We report averaged quality scores for such cases.

To evaluate these indirect effects, we compute the TER score of **pseudo PE**, where the translations of DT segments are changed *ex-post* in the baseline output. TER differences between PRE and pseudo PE precisely correspond to these indirect improvements. Figure 6.2 illustrates the procedure: for instance, the translation of the segment "overall risk" ("*risque global*" 'risk global') was replaced by its reference translation "*le risque global*" 'the risk global'. For the running example, the expected indirect improvements are a correct reordering and the correction of the grammatical error.

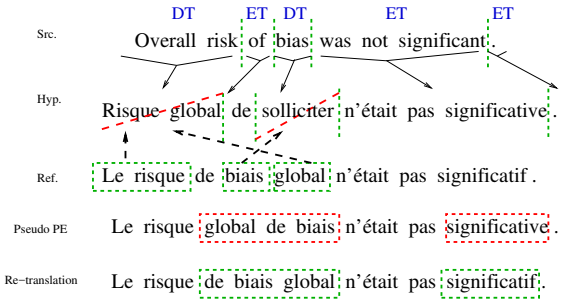


Figure 6.2 – Pseudo PE for the sentence "Overall risk of bias was low." segmented with MT-SEG

Mohit and Hwa [2007] report minor indirect effects of around $\Delta BLEU = 2$ by measuring improvements in BLEU of non-DT parts of sentences after re-translation. Cheng et al. [2016] measure indirect effects only after the re-translation of one DT SMT phrase, the most critical to improve output quality. They show that indirect improvements correspond to up to 72% of the total improvement, as measured in BLEU. Additionally, we attempt to get a closer insight into the nature of indirect effects and compute the POS statistics for the words, whose translations were positively or negatively influenced after the resolution of all the difficulties.

Measuring TER scores when a certain quantity of DT segments is pre-translated tells us only part of the story, since this might well be an unrealistic effort for the human pre-editors. To investigate how realistic this effort is, we compute the improvements in TER as a function of the number of characters a human would have to type to provide reference translations.

We compare this effort to the one involved in traditional PE, as well as in PE of artificial hypotheses generated by a system performing online updates to its models after each post-edited sentence (see section 3.3.1). To do this, we also simulated the latter procedures, using again the $\mathbf{e}^1 \rightarrow \hat{\mathbf{e}}$ alignments produced by TER. In each experiment, we "correct" a certain amount of words c in each test sentence. c is incremented from 1 to max_{TER}, the maximum quantity of words in a sentence that should be corrected to obtain a reference as prescribed by TER. "Correction"

involves the following operations: we replace substituted words with their reference translations, remove deleted words and insert the corresponding reference tokens. We again measure the number of typed-in characters as a proxy to the human effort. We use the Levenshtein edit distance [Levenshtein, 1966] for substituted words, otherwise the length of deleted or inserted words is used. More costly operations (in terms of input characters) are applied first.

6.4 HAMT: a Sentence-Level Scenario

6.5 Experiments in a Simulated Setting for MEDICAL

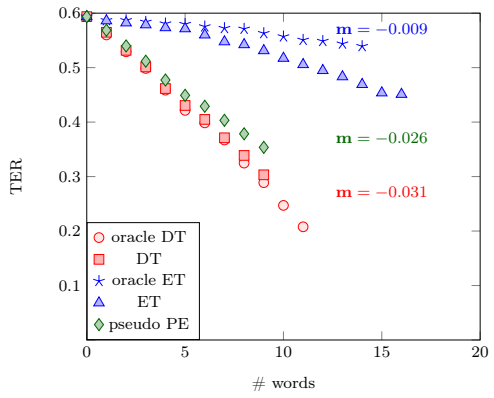
We first study the effectiveness of our DT identification strategies using Cochr-DT-SMT and WMT14-SMT in the MEDICAL domain. We saw earlier that the classification results were satisfactory (see section 5.3.3); we now evaluate whether detecting DT segments can actually help HAMT, and if so, whether one segmentation strategy is better than the others.

PE effort reduction (in TER) is plotted in Figure 6.3, for each segmentation strategy (WORD-SEG, MT-SEG and SYNT-SEG) and for each system (WMT14-SMT and Cochr-DT-SMT).

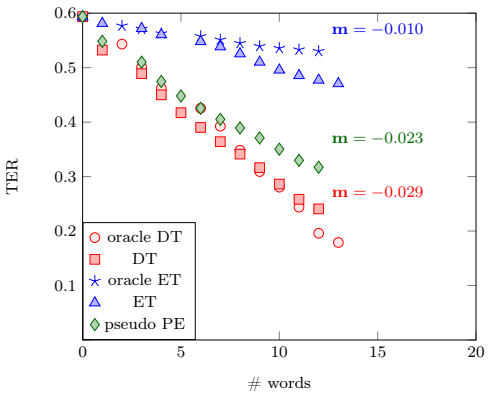
A first observation is that for the segmentation strategies, pre-translating all the DT segments results in a massive reduction of the residual PE effort. For WORD-SEG, for instance, TER improves by on average 0.24 absolute, and so does BLEU (22.35 points) for both MEDICAL systems. The results obtained in the oracle settings are not notably better ($\Delta TER = 0.31$, $\Delta BLEU = 28$ on average for both MEDICAL systems). As expected, the maximum average PE reduction achieved after constraining the translation of all the ET segments is much lower ($\Delta TER = 0.10$, $\Delta BLEU = 11.64$, oracle $\Delta TER = 0.04$, $\Delta BLEU = 6.8$ on average for both MEDICAL systems), since we mostly "correct" good translations. Those results confirm that our DT detection methodology is not only accurate, but also effective in a HAMT scenario and only mildly sensitive to erroneous labels.

The increase in performance with respect to the amount of pre-edition can be summarized by the slope of the regression function: for instance, for WORD-SEG WMT14-SMT we estimate that each pre-edited word improves the translation by about 0.031 TER point. The prediction difficulty for Cochr-DT-SMT makes this marginal improvement per additional word twice lower (0.017 TER).

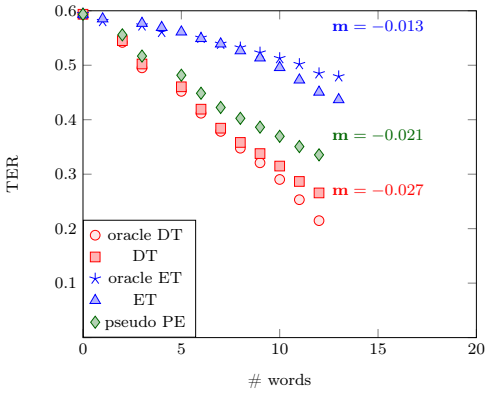
Contrasting segmentations, in the oracle setting MT-SEG strategy yields the best final improvement in translation quality, with a TER score of 0.18 ($BLEU = 63.86$) for WMT14-SMT, and a TER of 0.11 ($BLEU = 70.96$) for Cochr-DT-SMT (see Tables 6.1 and 6.2). This strategy is con-



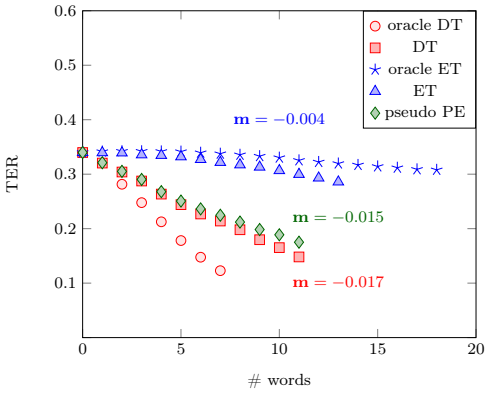
(a) PRE for WORD-SEG for `WMT14-SMT`



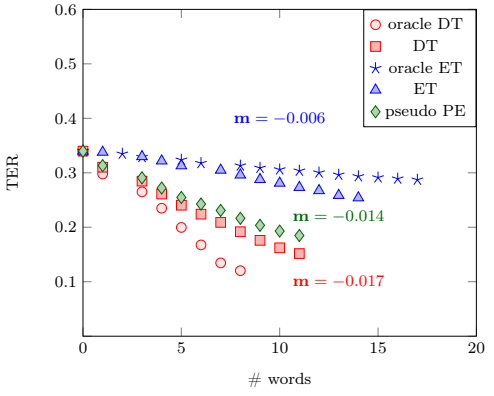
(b) PRE for MT-SEG for `WMT14-SMT`



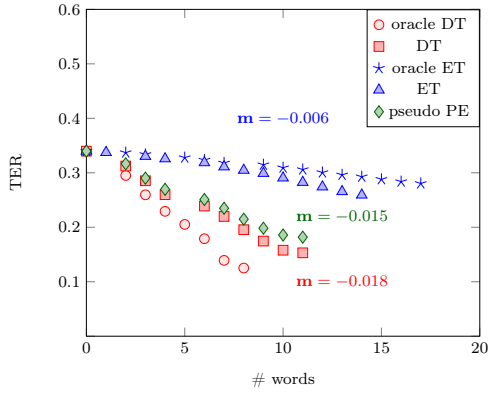
(c) PRE for SYNT-SEG for `WMT14-SMT`



(d) PRE for WORD-SEG for `Cochr-DT-SMT`



(e) PRE for MT-SEG for `Cochr-DT-SMT`



(f) PRE for SYNT-SEG for `Cochr-DT-SMT`

Figure 6.3 – PE effort reduction for `MEDICAL` (**m** denotes the slope for the non-oracle experiments, blue curves plot the results of pre-translating `ET` segments, red curves – the results of pre-translating `DT` segments, green curves – the results of pre-translating `DT` segments without re-translation)

sistent with the way MT segments input but is suboptimal (see section 5.1.1). SYNT-SEG offers a more human-friendly and almost equally effective alternative ($TER = 0.029$ vs. $TER = 0.027$ of the marginal improvement per additional word, for WMT14-SMT MT-SEG and SYNT-SEG respectively). For Cochr-DT-SMT SYNT-SEG is even slightly better than MT-SEG ($\Delta TER = 0.001$ of the marginal improvement per additional word).

Note that the residual PE effort in the simulated setting needed to go from an average $TER = 0.27$ for WMT14-SMT ($TER = 0.15$ for Cochr-DT-SMT) to $TER = 0.0$ can be explained by prediction, alignment, reordering and omission errors. Indeed, for the oracle MT-SEG, we found that 36% and 33% for WMT14-SMT (40% and 26% for Cochr-DT-SMT) of the residual edit operations were insertion and shift operations respectively (see Tables 6.1 and 6.2). Table 6.3 illustrates residual errors for MT-SEG Cochr-DT-SMT: due to a prediction error (the segment "chest" is not DT in the oracle setting) the segment "heart-related chest pain"[5] is translated with a reordering error "*douleur liée au cœur thoracique*" 'pain related to heart of chest', whereas in the oracle setting it is translated with an omission error ("*douleur liée au cœur*" 'pain related to heart'). Also note improved reordering in the re-translated version as compared to pseudo PE ("*douleur liée thoracique au cœur*" 'pain related of chest to heart' vs. "*douleur liée au cœur thoracique*" 'pain related to heart of chest').

		%					
	#, op.	D	I	S	SH	TER	BLEU
initial MT	15361	25	11	51	13	0.59	26.48
WORD-SEG	7708	15	24	35	26	**0.30**	49.52
WORD-SEG no re-translation	9070	32	13	31	24	0.35	42.61
oracle WORD-SEG	5244	12	35	17	36	0.20	59.92
MT-SEG	6162	16	25	33	27	**0.24**	57.91
MT-SEG no re-translation	8161	37	13	26	24	0.32	47.91
oracle MT-SEG	4555	11	**36**	20	**33**	**0.18**	63.86
SYNT-SEG	6811	25	23	32	19	**0.26**	59.12
SYNT-SEG no re-translation	8636	41	12	26	22	0.33	47.57
oracle SYNT-SEG	5448	23	30	23	24	0.21	64.11

Table 6.1 – Residual PE effort after resolving all the DT segments for WMT14-SMT (D denotes a deletion, I – an insertion, S – a substitution, SH – a shift)

A motivation for preferring PRE over PE are possible indirect effects induced by informing the translation of some segments. As explained above, these effects are measured by observing TER score differences between the PRE setting and pseudo post-edits. The corresponding curves, again as a function of the human effort, are in Figure 6.3. These graphs show that the indirect

[5]Actually, the segment "heart-related chest pain" is composed of three SMT phrases of one word each.

		%					
	#, op.	D	I	S	SH	TER	BLEU
initial MT	8792	18	16	53	13	0.34	48.28
WORD-SEG	3785	16	27	29	28	**0.15**	69.93
WORD-SEG no re-translation	4496	29	16	25	30	0.17	63.82
oracle WORD-SEG	3117	10	42	19	29	0.12	70.84
MT-SEG	3903	15	26	36	23	**0.15**	70.06
MT-SEG no re-translation	4762	28	16	31	25	0.18	63.65
oracle MT-SEG	3074	10	**40**	24	**26**	**0.11**	70.96
SYNT-SEG	3800	18	28	35	20	**0.15**	71.07
SYNT-SEG no re-translation	4689	29	17	29	24	0.18	64.05
oracle SYNT-SEG	3208	13	38	26	23	0.12	71.65

Table 6.2 – Residual PE effort after resolving all the DT segments for Cochr-DT-SMT (D denotes a deletion, I – an insertion, S – a substitution, SH – a shift)

effects are significant, and account for about 12-23% of the total improvement. For instance, the WORD-SEG condition for WMT14-SMT yields an indirect extra-return of about 0.005 TER by extra word, which is the half of the return observed when pre-translating ET segments.

To get a closer insight into the nature of these indirect positive changes, we computed the percentage of correctly translated words for each POS in the various settings (pseudo PE, final re-translated MT) – ignoring the words that have been input by the user. These correctly translated words were detected using the word alignments $\mathbf{e}^1 \rightarrow \hat{\mathbf{e}}$ produced by TER.

The strongest indirect positive influence is on the translation of adjectives and adverbs, for WMT14-SMT also on the translation of verbs and nouns (see Table 6.4). Adjectives are often translated "correctly" after the resolution of their head nouns. Positive changes to the translation of verbs and nouns are often lexical. Table 6.5 provides an example of positive influence on the translation of the noun "speech", which was finally correctly translated as "*parole*" after the resolution of neighboring segments (for WMT14-SMT, see Table G.1 in Appendix G for an example of positive influence on the translation of a verb; see Table G.2 in Appendix G for an example of negative influence on the translation of a preposition).

6.5.1 Comparison to Post-Edition

As a final validation of PRE, we will compare the user efforts involved in PRE and in PE using the knowledge of TER operations applied to MT as described in section 6.3. For a more thorough investigation PRE will also be compared to PE of hypotheses produced by systems adaptive to user feedback (see section 3.3.1), i.e., performing immediate updates to their models after each

src.	Percutaneous coronary intervention is regarded as a standard treatment for coronary heart disease to improve symptoms of **heart-related chest pain**.
initial MT	*L'intervention coronarienne percutanée est considéré comme un traitement standard pour la maladie coronarienne pour améliorer les symptômes de* ***la douleur thoracique cardiaques****.* 'The intervention coronary percutaneous is considered (Masc., Sg.) as a standard treatment for the disease coronary to improve the symptoms of **the pain of chest cardiac** (Masc., Pl.).'
ref.	*L'intervention coronarienne percutanée est considérée comme le traitement standard de la maladie coronarienne pour améliorer les symptômes de* ***douleur thoracique liée au cœur****.* 'The intervention coronary percutaneous is considered (Fem., Sg.) as the standard treatment for the disease coronary to improve the symptoms of **pain of chest related to heart**.'
	Predicted
src.	Percutaneous coronary intervention is regarded as a standard treatment for coronary heart disease to improve symptoms of heart-related chest pain.
pseudo PE	*L'intervention coronarienne percutanée est considérée comme le traitement standard pour la maladie coronarienne pour améliorer les symptômes de douleur liée thoracique au cœur.* 'The intervention coronary percutaneous is considered (Fem., Sg.) as the standard treatment for the disease coronary to improve the symptoms of pain related of chest to heart.'
PRE	*L'intervention coronarienne percutanée est considérée comme le traitement standard pour la maladie coronarienne pour améliorer les symptômes de douleur liée au cœur thoracique.* 'The intervention coronary percutaneous is considered (Fem., Sg.) as the standard treatment for the disease coronary to improve the symptoms of pain related to heart of chest .'
	Oracle
src.	Percutaneous coronary intervention is regarded as a standard treatment for coronary heart disease to improve symptoms of heart-related chest pain.
PRE	*L'intervention coronarienne percutanée est considérée comme le traitement standard pour la maladie coronarienne pour améliorer les symptômes de la douleur liée au cœur.* 'The intervention coronary percutaneous is considered (Fem., Sg.) as the standard treatment for the disease coronary to improve the symptoms of the pain related to heart .'

Table 6.3 – Illustration of the residual PE effort after pre-translating all the DT segments for MT-SEG Cochr-DT-SMT (underlined text denotes pre-translated segments)

	WMT14-SMT		Cochr-DT-SMT	
POS	#	Δ	#	Δ
ADJ	677	**1.5**	898	**1.8**
ADV	247	**5.7**	229	**1.7**
CONJ	1308	0.5	1531	0.4
DET	591	0.3	750	-0.7
N	2710	**1.1**	3363	-0.5
PRP	200	-1	192	-0.5
PREP	1702	-0.2	1666	-0.5
PUNC	2370	-1	2393	0.08
V	676	**3.9**	729	-0.9

Table 6.4 – Measuring the contextual influence on a per POS basis for MEDICAL SYNT-SEG: # denotes, for each POS, the number of words, which were not pre-translated; Δ denotes absolute changes in percentage of "correctly" translated POS after resolving all the DT segments.

src.	People who have this disease may develop changes in their behaviour, speech or ability to plan.
PRE	*Les personnes qui souffrent de cette maladie peuvent développer des modifications dans leur comportement,* ***la parole*** *ou la capacité de faire des projets.* 'The people who suffer from this disease can (3rd P. Pl.) develop modifications in their behavior, **the speech** or ability to make projects.'
MT	*Les personnes atteintes de cette maladie peut se développer des changements dans leur comportement,* ***du langage*** *ou de la capacité à planifier.* 'The people with this disease can (3rd P. Sg.) develop themselves changes in their behavior, **language** or the ability to plan.'
ref.	*Les personnes qui souffrent de cette maladie peuvent développer des modifications de leur comportement,* ***de leur parole*** *ou de leur capacité à faire des projets .* 'The people who suffer from this disease can develop (3rd P. Pl.) modifications in their behavior, **in their speech** or in their ability to make projects.'

Table 6.5 – Example of positive context influence for Cochr-DT-SMT (bold text denotes influenced context, underlined text – pre-translated segments)

post-edited sentence. Those systems typically store the training data in efficient data structures so that models can be estimated for each new input (e.g., by means of sampling) taking previous feedback into account. Such updates are usually accompanied by online updates of model weights, meant to increase the score of hypotheses that are close to post-edited translations.

To create such systems we took advantage of the `Moses` implementation of adaptive PBSMT [Germann, 2014, 2015]. This implementation employs two suffix array (SA) structures: one that indexes initial training data and another that indexes feedback data. To create a dynamic `Cochr-DT-SMT`, the former SA was used to index only Cochrane Reference Corpus (see section 4.3.1). We used the other standard static models described in section 5.3.1 to search for n-grams absent from the main model (`Moses back-off` mode). To create a dynamic `WMT14-SMT`, only the WMT'14 medical task parallel data were indexed by the first SA (see section 4.3.3).

For each new source phrase, this adaptive model selects a limited number of translation examples (we used the default `Moses` value) by deterministic random sampling from the indexed data. Those examples are then used to compute standard TM scores (see section 2.3). The following approximation was used to compute the inverse translation probability:[6]

$$\phi(\bar{f}|\bar{e}) = \frac{count(\bar{e},\bar{f})}{fr(\bar{e})} \cdot \frac{fr(\bar{f})}{count(\bar{f})}, \tag{6.1}$$

where $count(\bar{e},\bar{f})$ is the joint count of $\bar{e}$ and $\bar{f}$ in the sampled data, $count(\bar{f})$ is the count of $\bar{f}$ in the sampled data, and the $fr(\cdot)$ is the number of occurrences of the given phrase in the whole corpus. The direct and inverse probabilities were computed jointly for the main and the feedback corpora.

Additionally, the following scores were computed:

- phrase length ratio score: the log probability that the ratio between the length of $\bar{f}$ and the length of $\bar{e}$ is not extreme as compared to the mean ratio of the corpus;

- rarity penalty:

$$\text{rarity penalty} = \frac{\alpha}{\alpha + count(\bar{f},\bar{e})}, \tag{6.2}$$

  where α is a tunable parameter, and $count(\bar{f},\bar{e})$ is the joint count of $\bar{f}$ and $\bar{e}$ in the sampled data;

[6] The inverse translation probability can not be computed exactly because sampling is performed independently for each source phrase.

- provenance reward:

$$\text{provenance reward} = \frac{count(\bar{f}, \bar{e})}{\beta + count(\bar{f}, \bar{e})}, \tag{6.3}$$

  where β is a tunable parameter, this score is always separately for the main and the feedback corpora.

These scores are meant to penalize potentially erroneous and rare bi-phrases, as well as to reward bi-phrases coming from the main corpus or from the feedback corpus.

We used the same development and test sets as for the experiments with the static systems (see section 5.3.1). User feedback was simulated using reference translations with word alignments statically produced by `MGIZA` (the systems received a reference sentence with corresponding alignments after each automatically translated sentence). The alignments were computed for the concatenation of the MT training, development and test data. This simulated feedback was indexed by the second SA.

We re-used the LMs created for the static systems. The dynamic systems were tuned as the static systems (see section 4.3.3).

For the online adaptation of feature weights we used the solution proposed by Mathur et al. [2013], initially implemented in an older version of the `Moses` toolkit. We re-implemented this solution in a recent version of `Moses` (October 2016). The solution uses the MIRA algorithm for online weight updates (see section 3.3.1), as well as introduces a new *online feature* that seeks to increase the score of bi-phrases present in user feedback and consequently increases the chance that they will appear in 1-best hypotheses. The value of this feature is updated with the Perceptron algorithm. The initial value and the learning rate for the online feature, as well as the learning rate for other feature weights were tuned using the *Simplex Algorithm* [Nelder and Mead, 1965].

The quality of the dynamic `Cochr-DT-SMT` system is only slightly better than the quality of the corresponding static system, as measured using the same test set ($\Delta TER = 0.02$, $\Delta BLEU = 2.97$,), the improvement for the dynamic `WMT14-SMT` system is significant as compared to its static variant ($\Delta TER = 0.16$, $\Delta BLEU = 15.63$, see Table 6.6). As `Cochr-DT-SMT` is already a narrow in-domain high-quality system, regular online adaptation is only slightly efficient.

Figure 6.4 plots the human effort for PRE WORD-SEG and both kinds of PE. We reproduce the graphs of section 6.5 with a different x-axis and report per sentence averages of characters. Each point on those plots denotes an experiment, where we simulated correction or input (in the case of pre-translation) of a certain amount of characters in each test sentence. For instance, if

	#, op.	% D	% I	% S	% SH	TER	BLEU
dynamic Cochr-DT-SMT	8408	18	19	51	12	0.32	51.25
Cochr-DT-SMT	8792	18	16	53	13	0.34	48.28
dynamic WMT14-SMT	11173	20	14	51	15	0.43	42.11
WMT14-SMT	15361	25	11	51	13	0.59	26.48

Table 6.6 – Automatic evaluation of the dynamic MEDICAL systems (D denotes a deletion, I – an insertion, S – a substitution, SH – a shift)

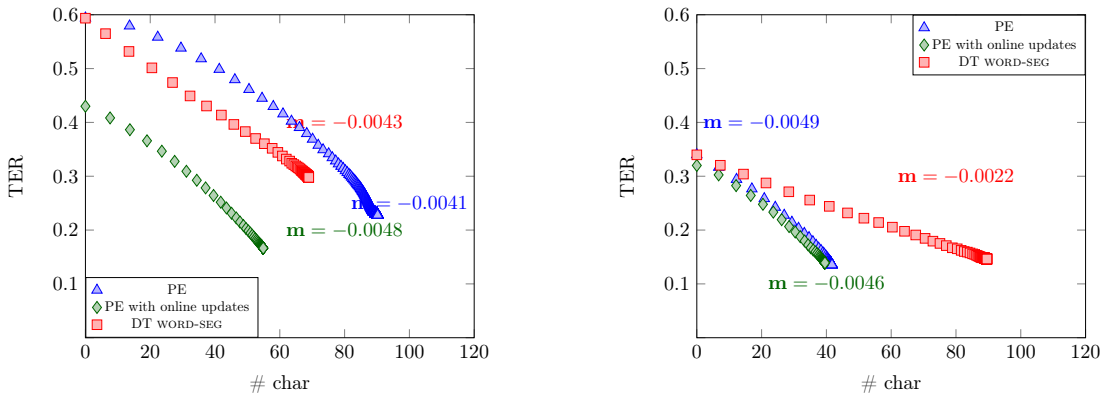


Figure 6.4 – Human effort reduction after PE of the initial or online-updated MT output (**m** denotes the slope); for WMT14-SMT (left), for Cochr-DT-SMT (right)

	ΔTER, per char.	
	WMT14-SMT	Cochr-DT-SMT
WORD-SEG	0.0043	0.0022
MT-SEG	0.0040	0.0021
SYNT-SEG	0.0035	0.0022

Table 6.7 – Marginal improvement of TER per character for PRE

a user types in around 20 characters per sentence to correct MT output the quality of the static WMT14-SMT will improve by $\Delta TER = 0.04$ (blue triangle), the quality of the adaptive WMT14-SMT output will improve by $\Delta TER = 0.06$ (green diamond). And if a user types in around 20 characters to pre-translate a certain amount of DT words (WORD-SEG) per test sentence the quality of the output will improve by $\Delta TER = 0.09$ (red square).[7]

According to those results, for WMT14-SMT the human effort reduction for the static PE is similar to the one of PRE: the marginal improvement of $TER = 0.0041$ per character vs. the marginal improvement of on average $TER = 0.0043$, for PE and PRE WORD-SEG respectively. The dynamic PE is only slightly more effective: an additional decrease of $\Delta TER = 0.0005$ per character as compared to PRE WORD-SEG. For Cochr-DT-SMT our conclusions are less optimistic: both static and dynamic PE perform better than PRE and yield an additional decrease of $\Delta TER = 0.0026$ on average. This latter result can be explained by poorer prediction quality for Cochr-DT-SMT, as well as the fact that Cochr-DT-SMT is a high-quality system that requires less major corrections to its output. Similar results are obtained for the other setups (see Table 6.7).

Note that our measurements of the human PE effort are quite optimistic, as in real-life settings we can not expect post-editors to perfectly optimize their keystrokes. Furthermore, the simulation of PE experiments does not take into account the cognitive effort needed to localize the required correction(s). We also put the dynamic PE in a very favorable condition by updating the models with the alignments statically produced by MGIZA. We therefore tend to consider that our PRE scenario remains competitive in terms of the human effort involved, and believe that the comparative merit of both approaches will only be resolved through experiments involving human pre- and post-editors.

[7] Hereinafter, in our experiments we consider the number of characters as the proxy to human effort.

6.6 Experiments in a Simulated Setting for UN

Extrinsic evaluation experiments for UN were performed in a similar way as our experiments in section 6.5.[8] The results of those experiments are plotted in Figure 6.5.

For all the language pairs, pre-translating all the DT segments results in a massive quality improvement, which is consistent with our MEDICAL experiments and confirms the efficiency of our methodology. For English-Russian, for instance, TER improves by 0.21 absolute, as does BLEU (16.78 points, see Table 6.9). Again, the maximum average PE reduction achieved after constraining the translation of all the ET segments is much lower (e.g., again for English-Russian $\Delta TER = 0.14$, $\Delta BLEU = 10.15$). The difference with the oracle condition is marked, suggesting that it would be here worthwhile to improve DT detection.

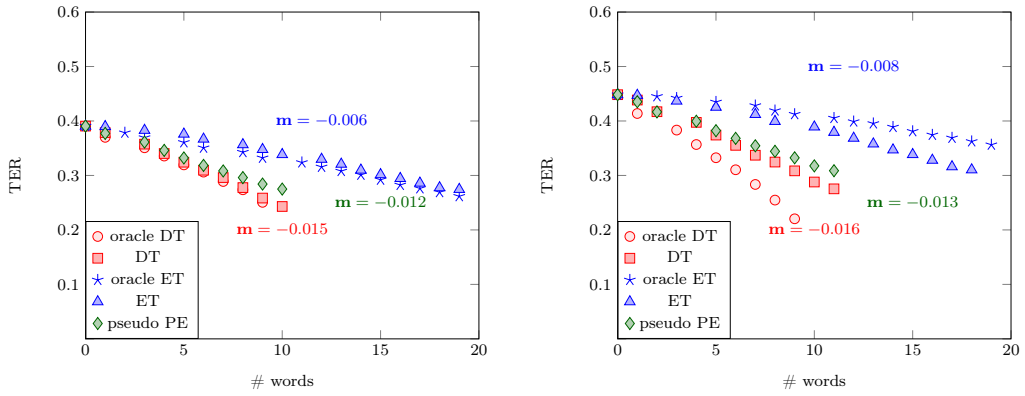
Regarding indirect improvements, they account for around 6% of the total improvement for such morphologically rich target languages as Arabic and Russian, which is less than in our MEDICAL experiments. This may be partly attributed to the reduced diversity in our translation models obtained by sampling. For English-Spanish and English-French, indirect improvements account for around 20%, which is consistent with our MEDICAL experiments.

Finally, the influence per POS (see section 6.5) for English-French is again strong for nouns. A positive influence is observed for the translation of determiners for all the language pairs, especially for English-Arabic and English-Russian (see Table 6.8, Table G.3 in Appendix G).

	AR		ES		FR		RU	
POS	#	Δ	#	Δ	#	Δ	#	Δ
ADJ	1412	1	1508	-0.3	1399	-1.5	1079	**2**
ADV	193	-0.5	135	-1.5	146	0	268	2
CONJ	1520	-0.3	1542	-0.2	1533	0.07	1381	0
DET	1192	**7**	2382	**0.5**	2134	**0.7**	784	**5.6**
N	6826	-0.9	6660	0.1	6442	**0.2**	5461	-1
PRP	348	-1.4	248	**4**	307	0.6	360	-2
PREP	2446	-1.5	2500	0.3	1993	0.7	1912	0.5
PUNC	2495	0.3	2707	0.5	2658	-0.2	2579	-0.8
V	923	-1	341	**1.5**	520	-1	571	0

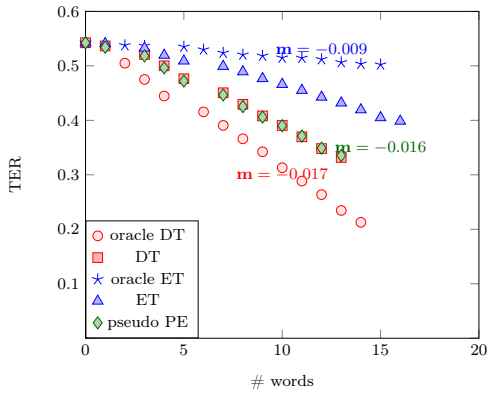
Table 6.8 – Measuring the contextual influence on a per POS basis for UN: # denotes, for each POS, the number of words, which were not pre-translated; Δ denotes absolute changes in percentage of "correctly" translated POS after resolving all the DT segments.

[8] The Stanford Arabic Parser tool [Green and Manning, 2010] and the MaltParser tool [Sharoff and Nivre, 2011] for Russian were used to obtain reference translations for segments.

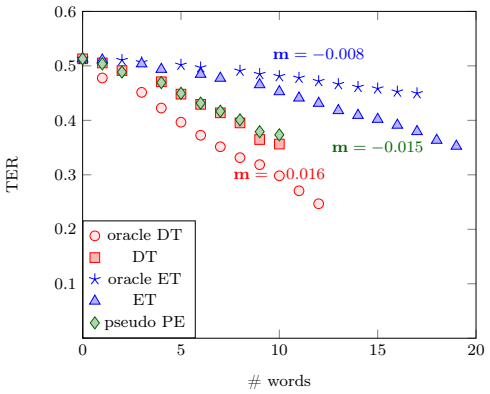


(a) EN-ES

(b) EN-FR



(c) EN-RU



(d) EN-AR

Figure 6.5 – PE effort reduction for `UN` (**m** denotes the slope for the non-oracle experiments, blue curves plot the results of pre-translating `ET` segments, red curves – the results of pre-translating `DT` segments, green curves – the results of pre-translating `DT` segments without re-translation)

		%					
	#, op	D	I	S	SH	TER	BLEU
AR							
initial MT	15297	18	17	53	12	0.51	37.12
SYNT-SEG	10623	16	27	40	17	**0.36**	48.84
SYNT-SEG no re-translation	11130	21	19	40	21	0.37	45.75
oracle SYNT-SEG	7298	10	41	26	23	0.24	58.76
ES							
initial MT	12774	18	20	49	13	0.39	49.13
SYNT-SEG	7898	16	28	36	20	**0.24**	63.02
SYNT-SEG no re-translation	8958	25	18	33	23	0.27	57.62
oracle SYNT-SEG	8108	16	29	36	19	0.25	62.11
FR							
initial MT	15132	16	20	52	12	0.45	44.38
SYNT-SEG	9262	14	32	37	17	**0.27**	59.39
SYNT-SEG no re-translation	10400	23	21	35	21	0.31	54.41
oracle SYNT-SEG	7323	11	42	28	19	0.22	63.83
RU							
initial MT	14124	17	16	59	8	0.54	35.58
SYNT-SEG	8601	17	27	40	16	**0.33**	52.36
SYNT-SEG no re-translation	8697	20	20	42	19	0.33	50.51
oracle SYNT-SEG	5485	10	36	29	25	0.21	62.11

Table 6.9 – Residual PE effort after resolving all the DT segments for UN (D denotes a deletion, I – an insertion, S – a substitution, SH – a shift)

6.7 HAMT: a Document-Level Approach

After validating our difficulty detection approach in an extrinsic evaluation at the sentence level, we now turn our attention to a document-level resolution scenario.

In this scenario a document is the translation unit and the human will have to resolve more difficulties prior to translation than at the sentence level. Hence, the scenario presents a series of new challenges. It risks to become frustrating if the human will have to provide feedback for all the DT segments of a document. Moreover, if a segment is recurrent and can not be consistently translated for all its occurrences, PRE risks to become inefficient. Thus, here we seek to answer the following main question: How to choose a set of DT segments in a document, whose translations are context-independent and whose pre-translations are crucial for the improvement of MT quality?

In this section we will describe our document-level DT resolution scenario (see section 6.7.1) and strategies for choosing DT segments for PRE (see section 6.7.2), as well our proposal to "softly" update TMs according to human feedback (see section 6.7.3). We test our approach using Cochrane review abstracts in a simulated (see section 6.7.4) and a real-life setting (see section 6.7.5).

6.7.1 Document-Level Human-Assisted Machine Translation

Our document-level HAMT protocol takes the following steps:

1. generate a baseline document translation, which is not displayed to the user;
2. detect DT and ET segments in each sentence of a document (see Chapter 5), aggregate the segments according to their sequences of source words, filter out segments classified as ET in all their occurrences;
3. ask the human pre-editor to provide translations for a **certain amount of DT segments** crucial for improving the translation of the document (see section 6.7.2). Examples of contexts are provided for each segment. The goal of the human pre-editor is to provide the most general translation that will fit into the majority of contexts;
4. compute the actual machine translation, using the obtained human suggestions as constraints during decoding;
5. produce the final PE.

Our assumption on the possibility of providing the most general translation that will fit into the majority of contexts at the document level is based on experimental validations of the "one sense per discourse" hypothesis. Originally developed for sense disambiguation algorithms, this hypothesis states that a polysemous word in a well-written discourse is extremely likely (98%) to always appear in the same sense [Gale et al., 1992]. For MT this hypothesis was formulated as "one translation per discourse". For instance, [Carpuat, 2009] shows that translating all the similar segments with a fixed translation is beneficial for translation quality and MT performance.

This document-level HAMT protocol may fit personal preferences of certain translators when they interact with the machine [Kay, 1997]. Indeed, it loosely mimicks the activity of professional translators: first, to analyze the source text, looking for parts that will be difficult for him/her to translate; then to consult any available external source of information; finally to translate, taking the obtained information into account.

As compared to the sentence-level translation difficulty resolution scenario (see section 6.4), the document-level scenario is obviously less frustrating (no repetitive resolution for frequent `DT` segments). We believe that the document-level scenario is also less demanding in terms of cognitive effort: all similar contexts of a segment can be analyzed at once, instead of analyzing one separate context at a time.

6.7.2 Selection of Crucial Difficult-to-Translate Segments

In order to choose segments, whose translations are context-independent and whose pre-translations are crucial for the improvement of MT quality, we propose the following two-step selection strategy:

1. In the first step we select segments $f_{[k:b]}$ that are more likely to be only `DT` (`only-DT`). To do that we compute the probability of a segment to be `DT` in a document as a relative frequency score:

$$p_{\texttt{only-DT}}(f_{[k:b]}) = \frac{count(f^{DT}_{[k:b]})}{fr(f_{[k:b]})}, \tag{6.4}$$

   where $count(f^{DT}_{[k:b]})$ is the count of segment occurrences as `DT`, and $fr(f_{[k:b]})$ is its frequency in a document. Here we search for segments that are never `ET` and consistently "badly" translated. If we computed, for instance, the mean of the distribution of the posterior probabilities of the `DT` class, we would also choose very frequent segments that can be `DT` or `ET` depending on the context (with high probabilities in both cases). These are the segments that we need to avoid, since their translation is not generalizable.

2. At the second step we compare 3 strategies, prioritizing:

   - `DT` segments with the highest mean $\overline{x}$ (`mean`) of the distribution of the posterior probabilities of the `DT` class ($\bar{p}$) over all their occurrences;
   - the most frequent $f_{[k:b]}$ (`freq`): we consider the frequency in a document;
   - the longest $f_{[k:b]}$ ($\#\ f_z, k \leq z \leq b$, `length`).

Other strategies of `DT` segment selection may include, for instance, prioritizing segments from certain sections of a document (e.g., conclusion, summary, etc.), selection of segments by their semantic importance, etc.

6.7.3 Update of Translation Models

Our document-level protocol targets the resolution of difficulties in a batch setting mode: we propose to first collect all the pre-translation feedback from pre-editors (expected terminology experts), postponed PE of improved MT can be performed by post-editors (expected domain professionals). In this setting there is no need for immediate online updates to the model.

We thus use document-specific TMs (extracts of the main TM), correspondingly updated using a "soft" procedure. This procedure adds only the missing information to TMs. To be more precise, if a human suggestion is not only equal, but also a part of a phrase $\bar{e}$ in a TM we inform the system about this translation variant and where it should be applied in a text. We contrast the "soft" update procedure and the "hard" update procedure (e.g., `exclusive xml-mode` in `Moses`). The latter imposes that a given segment be translated as one unit while the former gives some leeway to the decoder about how to translate the segment (see section 3.3.1).

Our "soft" update procedure takes the following steps to update TM entries:

1. alignments between human suggestions and source `DT` segments are computed in an online fashion (see section 3.3.1);

2. bi-phrases ($\bar{f}_h$,$\bar{e}_h$) are extracted for the resulting mini-corpus using standard heuristics (see section 2.3.2) ;

3. $\bar{f}_h$ are marked according to their position (@`DT` _id-word_position) to make sure we update only their translation. For instance, in the `DT` segment "venous@5-0 thromboembolism@5-1" the word "venous@5-0" is the first word (position 0) in the 5th `DT` segment in a document. Occurrences of $\bar{f}_h$ in a document are modified accordingly;

4. existing bi-phrases ($\bar{f}_{ex}$,$\bar{e}_{ex}$) that exactly match the newly extracted bi-phrases are duplicated, in each duplicated entry $\bar{f}_{ex}$ is replaced by $\bar{f}_h$. For our running example, the resulting TM would contain two entries: the original "venous thromboembolism ||| *la thromboembolie veineuse*" and the marked duplicate "venous@5-0 thromboembolism@5-1 ||| *la thromboembolie veineuse*";

5. TM entries that partially match (prefix/suffix, the middle of ($\bar{f}_{ex}$,$\bar{e}_{ex}$) exactly matches $\bar{f}_h$ or $\bar{e}_h$ correspondingly) are duplicated. Each duplicate is modified only for the matching part. For our running example, the resulting TM would contain two entries: the original "venous thromboembolism was studied ||| *la thromboembolie veineuse a été étudiée*" and the modified one "venous@5-0 thromboembolism@5-1 was studied ||| *la thromboembolie veineuse a été étudiée*";

6. bi-phrases ($\bar{f}_h$,$\bar{e}_h$) unmatched so far are added to the TM.

This new information from TMs can be later used for permanent updates of the general model (see section 3.3.2).

6.7.4 Cochrane Abstracts: Experiments in a Simulated Setting

Before proceeding with real-life experiments, we will validate our document-level scenario and our "soft" update procedure in a simulated setting using a test set of 100 Cochrane abstracts from Cochrane PE Corpus 2 (see Table 6.10). We used the same `Cochr-DT-SMT` system as in the previous experiments (see section 5.3.1). We performed our experiments for SYNT-SEG and WORD-SEG for a contrastive evaluation.

#, documents	#, lines	# tok., EN	# tok., FR
100	47	1.1K	1.4K

Table 6.10 – Averaged statistics over the document-level test set

Here, we need to note that our document-level simulated and real-life experiments were conducted in a slightly different translation difficulty detection setting (*alpha* version) than the experiments that we previously reported in this manuscript. The major difference between this first version and our final version is the definition of a `DT` segment. The former is parser-dependent: source words f aligned to reference words $\hat{e}$ corresponding to syntactic heads of deleted reference words were automatically considered as `DT`, as their translation is not complete.

The rationale behind this approach was the necessity to take into consideration omission errors in MT output. For instance, the translation "broncho-pneumopathie obstructive" 'pulmonary disease obstructive' of the segment "chronic obstructive pulmonary disease" misses the translation of the word "chronic". Hence we consider the word "disease" to be DT as it translates into the syntactic head ("*broncho-pneumopathie*") of the missing word "chronique" 'chronic'. Table 6.11 shows statistics for the MEDICAL test set (see section 5.3.1) annotated using this parser-dependent DT definition: we can see that this annotation yields an increase of around 8% in the quantity of DT segments (see Table 5.4). This preliminary DT definition was revisited to make our labeling strategy less error-prone in a multilingual setting (because of poor parsing performance for certain languages).

strategy	#	$\bar{l}$	DT
WORD-SEG	20K	1	31%
MT-SEG	13K	1.6	36%
SYNT-SEG	12K	1.68	35%

Table 6.11 – Statistics for the annotated MEDICAL test set ($\bar{l}$ is the average segment length)

In the alpha version of the difficulty detection procedure we also made an attempt to create universal MEDICAL classifiers and used only the output of WMT14-SMT for feature extraction (see section 5.3.3). Feature extraction was performed as described in section 5.3. We used CRFs as the classification algorithm.[9] The classifiers were directly applied to Cochr-DT-SMT output. Table 6.12 shows the results of the intrinsic evaluation for this configuration. The sets of standard and basic features turned out to be the most useful for WORD-SEG, only the set of standard features was useful for SYNT-SEG. All these observations are consistent with our previous conclusions on the usefulness of features for difficulty detection (see section 5.3.3). This universal approach was abandoned in favor of system-specific classifiers yielding better prediction quality: for instance, for SYNT-SEG $F_{mcr} = 0.62$ and $F_{mcr} = 0.69$, for the alpha and the final version respectively (see Table 5.10).[10]

As for the simulated sentence-level extrinsic evaluation (see section 6.4) of this preliminary approach, for Cochr-DT-SMT the 3 segmentation strategies yield a similar marginal improvement of 0.0019 TER per additional character (absolute improvement $\Delta TER = 0.19$ for MT-SEG and

[9]The CRF implementation available in Wapiti was used. We used the l-bfgs algorithm as the optimization algorithm. All other parameters are those provided by default. All the hyperparameters were tuned on the development set.

[10]These evaluation results are not directly comparable as they were estimated for the test set labeled using different definitions of DT.

	$F_{\texttt{DT}}$	$F_{\texttt{ET}}$	F_{mcr}
WORD-SEG	0.52	0.76	0.64
SYNT-SEG	0.56	0.68	0.62

Table 6.12 – CRF performance for WORD-SEG and SYNT-SEG

SYNT-SEG, $\Delta TER = 0.13$ for WORD-SEG, see Tables 6.13 and 6.7).[11]

	ΔTER, per char.	TER
initial MT		0.34
WORD-SEG	0.0019	0.21
oracle WORD-SEG	0.0039	0.11
SYNT-SEG	0.0019	0.15
oracle SYNT-SEG	0.0035	0.12

Table 6.13 – Residual PE effort after pre-translating all `DT` segments

We will now describe our experiments in the document-level difficulty resolution scenario using this alpha version of the difficulty detection procedure. Taking into account the similarity of the evaluation results for the alpha and the final version we believe that the results presented in this section would also be valid for the final difficulty detection approach.

For our "soft" TM update procedure, word alignments between human suggestions and source `DT` segments were computed using an online version of the associative sub-sentential `Anymalign` alignment method [Lardilleux et al., 2012; Gong et al., 2013]. This method was shown to be more efficient than standard forced alignment approaches (see section 3.3.1).

We replicated our extrinsic evaluation procedure on a per-document basis (see section 6.4) with the following major modification: to pre-translate each `DT` segment we chose *the most frequent reference translation* from the translations of all its occurrences. We re-translated each document using "softly" updated TMs (the `Moses exclusive xml-mode` was used for contrastive experiments).

Table 6.14 shows the results of this extrinsic evaluation.[12] The "soft" TM update method yields a performance similar to the `Moses xml-mode`. For SYNT-SEG, after pre-translating all the `DT` segments, we obtained an improvement of 0.000044 TER per additional character[13] for the `xml-mode` and of 0.000043 TER for our method (an average absolute improvement of $\Delta TER = 0.12$ for both methods). However, we believe that in a real-life setting the flexibility of our "soft"

[11] Note that for the current approach the improvement per additional character is on average 10% more efficient, even if again these evaluation results are not directly comparable (see section 6.5.1).

[12] All the presented evaluation results are averaged over the number of documents.

[13] Hereinafter, for document-level experiments we consider the number of characters normalized over the number of documents in the test set.

approach can be more beneficial.

		%					
	#, op	D	I	S	SH	TER	BLEU
initial MT	485	20	27	54	11	0.35	50.9
SYNT-SEG soft	322	14	32	38	16	0.23	62.81
SYNT-SEG `Moses xml-mode`	318	16	28	36	15	0.23	63.11

Table 6.14 – Residual PE effort after resolving all the `DT` segments in a document for SYNT-SEG (D denotes a deletion, I – an insertion, S – a substitution, SH – a shift)

As for our strategy to choose the most frequent translation across occurrences, as might be expected, the maximum absolute gain in TER after pre-translating all the `DT` segments in the document-level scenario is lower than in the sentence-level scenario, where each segment occurrence is pre-translated separately. For instance, for SYNT-SEG, an absolute improvement of $\Delta TER = 0.19$ is observed at the sentence-level (see Table 6.13), at the document level we observe an absolute improvement of $\Delta TER = 0.12$; an improvement of $\Delta TER = 0.13$ and an improvement of $\Delta TER = 0.027$ for WORD-SEG, at the sentence level and at the document level respectively.[14] These evaluations also suggest that SYNT-SEG can be considered as a more beneficial segmentation in the document-level scenario, as it opens a broader perspective for quality improvement.

Table 6.15 presents the results of our `DT` segment selection experiments. They confirm the efficiency of our `only-DT` strategy: for SYNT-SEG, after pre-translating only 50% of the `DT` segments per document we obtain on average more than half of the gain we obtain after pre-translating all the `DT` segments (54%). For WORD-SEG, the `only-DT` strategy helps to mitigate the negative effect of pre-translating `DT` words using their most frequent translation variants. Thus, after pre-translating only 50% of the `DT` words per document we obtain a gain of $\Delta TER = 0.01$ in absolute improvement as compared to the absolute improvement we obtain after pre-translating all the `DT` words.

For the second step, the `mean` strategy seems to be the most useful (the highest human effort reduction of 0.000053 TER per character for WORD-SEG, the highest human effort reduction of 0.000047 TER per character for SYNT-SEG). For WORD-SEG, the `freq` strategy performs similarly to `mean` (see Table 6.15).

Table 6.16 provides more detailed characteristics of the segments chosen by the `mean` and `freq` strategies for both WORD-SEG and SYNT-SEG. We can see that those `DT` segments are quite

[14] These evaluation results are not directly comparable as they were estimated for the different test sets.

infrequent ($fr = 1.33$ for SYNT-SEG `mean`) and short for SYNT-SEG (2.17 words on average). This suggests that the cognitive effort required for this task is quite moderate (a limited amount of contexts to look through, translation of short segments).

	SYNT-SEG			WORD-SEG		
strategy	ΔTER, per char.	#, char.	ΔTER	ΔTER, per char.	# , char.	ΔTER
`mean`	**0.000050**	1725	0.080	**0.0000527**	735	0.037
`length`	0.000036	1671	0.057	-	-	-
`freq`	0.000039	1472	0.053	**0.0000519**	735	0.037

Table 6.15 – Results of `DT` selection experiments (50% of all the detected `DT` segments are resolved)

	`mean`	`freq`
fr	1.33	1.47
l	2.14	2.72
$\bar{p}$	0.962	0.85

	`mean`	`freq`
fr	1.51	1.54
$\bar{p}$	0.84	0.84

Table 6.16 – Characteristics of the `DT` segments chosen by the `mean` and `freq` strategies ($\bar{p}$ denotes the posterior probability of a segment to be `DT`, the results for SYNT-SEG are on the left, for WORD-SEG – on the right)

Figure 6.6 shows that for the SYNT-SEG `mean` strategy almost the best attainable quality is reached after pre-translating 70% of the `DT` segments per document (on average 140 segments per document; we consider a bigger quantity of segments suboptimal for human resolution); for WORD-SEG `freq` – around 60% of the `DT` segments per document (102 words per document on average).

The final document-level PRE configuration used in real-life experiments (see section 6.7.5 below) consisted in asking the human to pre-translate 70% of `DT` chunks per document chosen by our `only-DT` + `mean` strategy. Those segments were detected using a CRF-based classifier trained on the `WMT14-SMT` data. We used the "soft" online update procedure for re-translation.

6.7.5 Cochrane Abstracts: Experiments in a Real-life Setting

The document-level resolution scenario described above was tested by 5 students of the ISIT translation school,[15] native French speakers, within the duration of an applied research project (15 November 2016 – 15 February 2017). The project was supervised by a professional translator and an experienced teacher of translation. The task was performed using the in-house PE

[15] `http://www.isitinternational.com`

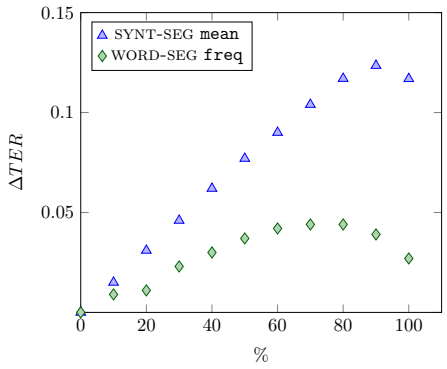


Figure 6.6 – Human effort for SYNT-SEG `mean` and WORD-SEG `freq`, results on average per document

interface described in section 4.3.1. The students worked remotely using their personal accounts, and their work was systematically verified by the supervisor.

Given this mode of work (any time remote connection to the interface, possible bad quality Internet connection), we evaluated the resulting work using only automatic metrics. Time estimations could not be reliable in this condition.

Post-Edition Task

As a preliminary task each student had to perform traditional PE of one Cochrane review abstract translated by `Cochr-DT-SMT` (1-week task). The task was given in order to test and to adjust if necessary the quantity and the quality of corrections performed by the students.

Guidelines The **main guideline** was to perform minimal PE. The final text should be understandable and grammatical, which corresponds to the main quality requirement of Cochrane France. Literal automatic translations are to be left as is unless the above-mentioned requirements are not satisfied. Before performing the task the students had access to the Cochrane Corpus and were given sufficient time to familiarize themselves with it.

#, documents	#, lines	# tok., EN	# tok., PE	HTER	BLEU
5	50	1.5K	1.7K	0.33	54.74

Table 6.17 – PE task: statistics and results (averaged per document)

Results Table 6.17 presents the results of this preliminary task. They confirm our previous observations regarding the necessary amount of PE to make `Cochr-DT-SMT` output meet the requirements (see section 4.3.3), as well as proper comprehension of the task.

Pre-Edition Task

Upon completion of the PE task, the students were given a document-level PRE task, that consisted of the following parts:

1. PRE of `DT` segments of a document (detected and selected as described in section 6.7.4) (1-week task, individual work);
2. Post-edition of a machine-translated version of the document, where the MT was produced taking the pre-edited segments into account (1-week task; note that the students expressed the wish to perform this work collectively).

In total, 9 Cochrane abstracts were pre- and post-edited in this way (see Table 6.18).

#, documents	#, lines	# tok., EN	# tok., PE
9	50	1.1K	1.5K

Table 6.18 – PRE task: statistics (averaged per document)

The students were aware of the purpose of the PRE task. They were informed that they were to provide translations for segments that the system had difficulties to translate. After the resolution of difficulties they would have to post-edit improved MT.

We will now describe the implementation of the first part of the task. The PE part was performed as described above.

Interface We have developed a special interface for the PRE task, which was integrated into the main PE interface. The PRE interface displays `DT` segments in one column. Their translations are to be provided in neighboring input boxes (see Figure 6.7).

The "See context" button opens a pop-up window displaying `DT` segments in their contexts. For instance, Figure 6.8 shows that the `DT` segment "sensory impairment" has two occurrences in a document. The names of corresponding Cochrane abstract sections are provided for each occurrence.[16] The pop-up window also displays the full-text context where `DT` segment occurrences under focus are highlighted in yellow.

[16]The students considered that this information is necessary to provide correct translation.

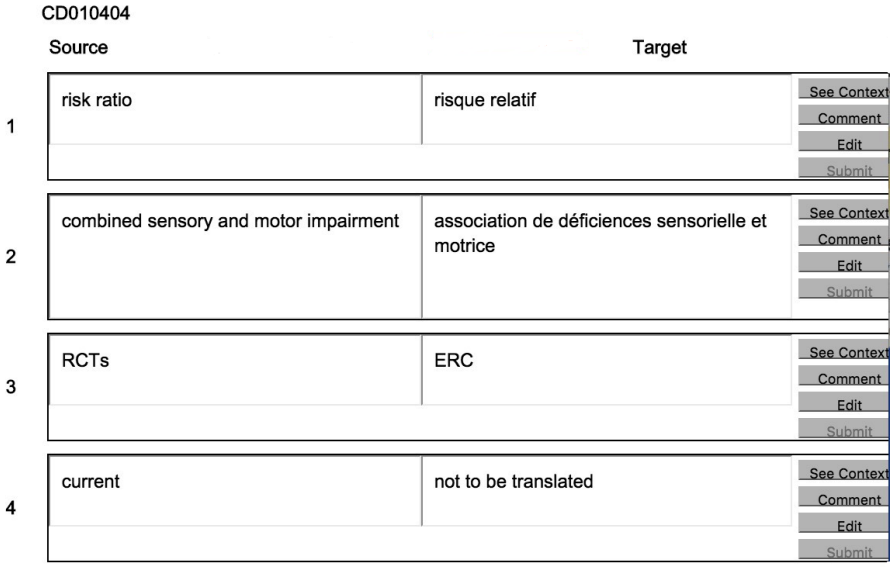


Figure 6.7 – PRE interface

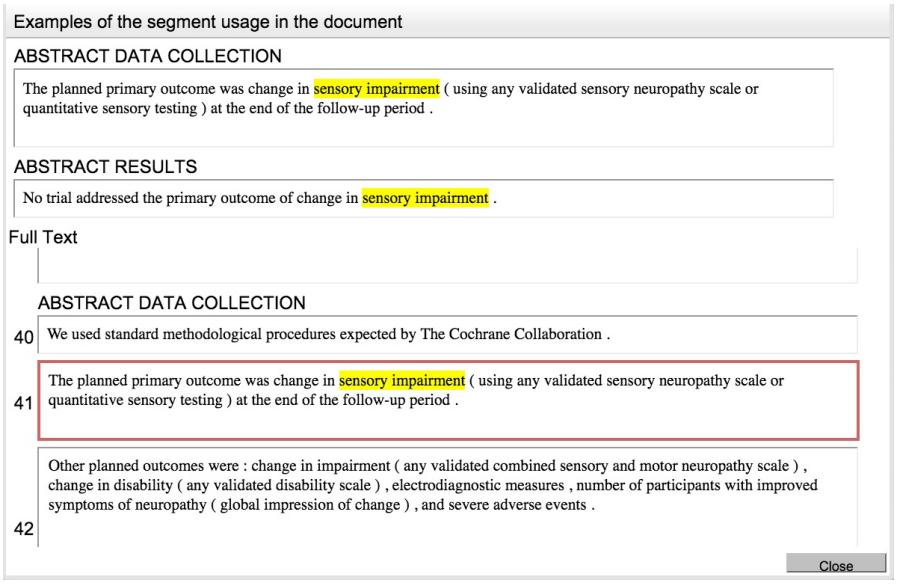


Figure 6.8 – Example of a DT segment in its context

Guidelines We asked the students to provide the most general translation for each DT segment that will fit into a maximum number of contexts. If such a translation is not possible, a translation should be provided for the first occurrence. If a segment can not be translated into French students were to input the phrase "not to be translated" into a corresponding box.

Difficulties The main difficulty faced by the students was pre-translating erroneously extracted chunks (i.e., not syntactically meaningful). For instance, the erroneous chunk "to imprecise" was extracted for the sentence "... due to imprecise estimates of effect and risk of bias". We asked the students to provide contextual translations for those segments: for the running example, "*d'imprécises*" 'imprecise (Pl.)'.

Results Table 6.18 presents the results of the task. We compute TER between the produced PE and initial MT (generated before difficulties are resolved), as well final MT (generated after difficulties are resolved). Those results are encouraging: the absolute gain in quality after pre-translating 70% DT segments in a document is on average equal to $\Delta TER = 0.09$ ($\Delta BLEU = 10.36$). Our document-level resolution strategy yields a marginal improvement of around 0.000050 TER per additional character. Those evaluations confirm our observations regarding the evolution of the MT quality in the simulated document-level setting (see section 6.7.4).

To estimate indirect effects of the difficulty resolution we compared the resulting PE to pseudo PE, where in the initial MT hypotheses we replaced translations of the DT segments (obtained from the $\mathbf{f} \rightarrow \mathbf{e}^1$ alignments, as produced by the decoder) with human suggestions.[17] The indirect effect of the human difficulty resolution is around 38% of the total improvement per additional character ($\Delta TER = 0.000019$). However, this value needs to be considered with care because of the noise added by the automatic alignments.[18] Table 6.20 gives an example of positive context influence, where with two pre-translated segments "is emerging" and "disability progression" the final MT reached almost the quality of PE (one final substitution was performed during PE: the word "*importante*" 'important' was replaced with the word "*porteuse*" 'cornerstone'). During our manual analysis we noticed that the pre-translation of in-domain terms and expressions (for instance, "disability progression") was the most rewarding: those segments are repetitive and consistently translated within a document.

[17] In a case of "not to be translated" suggestion we removed the corresponding MT.

[18] In certain cases we thus replace only partial translations of DT segments, which negatively influences automatic evaluation scores.

A full example of a document with its initial and final MT, as well as its PE is presented in Appendix H.

We should note that the translation difficulty resolution results presented here can be considered only as preliminary as the students managed to pre- and post-edit only 9 Cochrane abstracts. In addition, some of the students showed a lack of motivation to work on the project.[19] More experiments are required for a final conclusion on the efficiency of PRE.

		%							
	#, op.	D	I	S	SH	HTER	BLEU	#, char.	ΔTER, per char.
initial MT	502	14	20	54	11	**0.35**	52.49		
no-retrans MT	428	17	21	48	13	0.29	56.25	1587	0.000031
final MT	371	15	24	50	11	**0.26**	62.85	1587	0.000050

Table 6.19 – PRE task: results (on average per document; D denotes a deletion, I – an insertion, S – a substitution, SH – a shift)

src.	Neuroprotection is emerging as a potentially important strategy for preventing disability progression in people with MS.
initial MT	*La neuroprotection **est** potentiellement importante **émergentes** comme une stratégie pour prévenir **la progression du handicap** chez les personnes atteintes de SEP .* 'The neuroprotection **is** potentially important **emerging** (Fem., Pl.) as a strategy to prevent the **progression of handicap** in people with MS.'
final MT	*La neuroprotection apparaît comme une stratégie potentiellement **importante** pour prévenir la progression de l'incapacité chez les personnes atteintes de SEP .* 'The neuroprotection appears as a strategy potentially **important** to prevent the progression of disability in people with MS.'
ref.	*La neuroprotection apparaît comme une stratégie potentiellement **porteuse** pour prévenir la progression de l'incapacité chez les personnes atteintes de SEP.* 'The neuroprotection appears as a strategy potentially **cornerstone** to prevent the progression of disability in people with MS.'

Table 6.20 – Example of positive context influence in the document-level resolution scenario (underlined text denotes pre-translated segments, bold text – their human translations)

[19]The students received an inconsistent piece of information before the launch of the project. They were not informed that they would work mostly with PE, whereas they were more willing to accomplish a conventional Human Translation task.

6.8 Summary

In this chapter we have studied various ways to implement the results of our difficulty detection protocol into MT to gain more control over its output. We focused our attention on a pre-edition (PRE) scenario where the human provides translations for the detected difficulties.

Providing translations for source difficulties (and not, for instance, simplification of difficult-to-translate source segments as a hint to correct translation) ensures that the suggestion of the translator will appear as such in the automatic translation. This is very important for the translation of text types containing precise information, like medical texts.

The PRE setting, as compared to post-edition (PE), can be more beneficial in terms of the involved human effort: repetitive segments can be resolved once for many occurrences. Another important motivation for preferring PRE over PE is a possible indirect improvement in output translation: providing information for a difficult word or phrase can have a positive impact on the automatic translation of neighboring words.

For an extrinsic evaluation of our procedure for translation difficulty detection (see Chapter 5), we have introduced a sentence-level Human-Assisted MT (HAMT) protocol accommodating this procedure, as well as our PRE scenario. We have tested this protocol in a simulated setting for a set of language pairs (English-Arabic, English-French, English-Russian, English-Spanish). We have experimented with multiple segmentation strategies and system types (within the PBSMT approach).

We conclude that asking the human expert to pre-translate `DT` segments could be beneficial in a HAMT protocol. Upon reaching a certain quality of prediction (around $F_{mcr} = 0.70$), PRE can as be at least as useful as PE, and even as useful as PE performed to the output of a system regularly updating its models according to user feedback. Regarding the various segmentation strategies considered here, there are no crucial differences between their performance, and the most human-friendly strategy, like the syntactically-motivated one, can be used without any major productivity loss.

Indirect effects in PRE are genuine, accounting to up to 23% of the total effort reduction. These effects depend on the language pair: they are less important for morphologically rich languages (for instance, English-Russian).

We have then introduced a document-level HAMT protocol, where the human expert is asked to provide a translation of a `DT` segment once and for all its occurrences in a document. We based our protocol on the "one translation per discourse" hypothesis and assumed that the human pre-

editor is to provide the most general translation of a segment that will fit into the majority of contexts. We have investigated a series of approaches to choose DT segments whose translation can be generalized over contexts and also useful for MT.

We have tested our document-level protocol in a simulated setting using Cochrane review abstracts. The syntactically-motivated segmentation seemed to be the most beneficial in terms of translation quality improvement, especially combined with the strategy choosing DT segments that are consistently DT in all their contexts. A reasonable human effort is required to obtain a reduction of around one third of the initial PE effort: pre-translating around 140 segments per document each of around 2 words long. Those conclusions were confirmed in preliminary document-level PRE experiments with translation studies students. Results of those preliminary experiments suggest an encouraging beginning of the test phase, as they show a reduction in the final human PE effort and also confirm the important presence of indirect changes. Nevertheless, large-scale experiments are needed to be able to draw a more definite conclusion.

7 | Conclusion and Perspectives

Contents

In this final chapter we will briefly revisit the problem of human-machine collaboration within Machine Translation (MT), as well as choices we have made to implement this collaboration in order to improve MT of systematic medical review abstracts. We will summarize our main contributions and experimental results. We will close the chapter by discussing some perspectives of our work.

Today even high-quality MT does not produce output of a publishable quality. As some human intervention is still required, one of the potential ways to improve MT is to improve human-machine collaboration.

Traditionally, the injection of human knowledge into MT can take place at the beginning or at the end of this process. The former is usually referred to as pre-edition (PRE); the latter – as post-edition (PE).

Several questions relative to human-machine collaboration in MT need to be addressed in order to develop concrete implementations:

1. Which type of collaboration to choose? Can it be only PRE or only PE, or a combination of these interventions?

2. How to detect where the injection of human knowledge is required? Can we let humans decide, but in this case the process risks to become too costly in terms of cognitive effort. How to then automatically detect where human help is needed?

3. Which type of human knowledge will be injected? For instance, it can be monolingual knowledge in the form of source normalization or simplification; or the knowledge can be bilingual in the form of ready-to-use translations?

4. How to make the machine correctly exploit this knowledge? Should the knowledge be used in the short term (for instance, for the translation of a certain text) or in the longer term (for all future translations)?

Answers to these questions depend on many factors such as MT type, quality, translation domain, qualities and skills of human experts, etc.

7.1 Contributions

Our work has been focused on improving automatic translation of Cochrane medical review abstracts. Those abstracts are produced in English and translated into 16 different languages, including French, German, Spanish, Russian etc. Automatic translation into French is performed using a narrowly-specialized in-domain SMT system. The produced high-quality MT is post-edited by mainly domain specialists. Residual errors of this high-quality MT are still numerous. Additionally, non-professionals in PE introduce new errors in final translations, mainly due to inconsistent terminological choices they make. More control over the quality of published translations is crucial for the Cochrane production context, especially taking into account the fact that the main mission of the organization is to spread reliable medical knowledge.

Given this representative Cochrane context, we have decided to focus our work on **PRE scenarios**. As our main concern was the terminological consistency of final translations, we decided to solicit human bilingual help at the level of words and phrases, where the MT system faces difficulties. Such pre-translation help guarantees correct translations in output. Those scenarios can also provide human experts with better control over MT output and can be particularly beneficial in reducing the amount of final PE, leading to less frustration of non-professionals and less risk that they would introduce new errors. This reduction is reached as a result of "correct" pre-translations in output, but also as a result of potential improvements to the automatic translation of neighboring words ("free" automatic corrections).

Our first main contribution is thus **a system-independent methodology for translation difficulty detection**. We defined the notion of subsentential source-side translation difficulty: difficult-to-translate segments are segments for which an MT system makes erroneous predictions. We cast the problem of difficulty detection as a binary classification problem: we detect easy-to-translate (`ET`) and difficult-to-translate (`DT`) segments. We show that using this methodology translation difficulties can be reliably detected both at the word level and at the phrase level, using only a simple set of system-independent features.

Considering our goal of consistent MT of Cochrane abstracts into many languages at once, we have studied the problem of **detecting translation difficulties in a multilingual scenario**, the first attempt of the kind to our knowledge. Our study has allowed us to conclude that translation difficulties depend on the language pair, rather than solely on the source language. Moreover, they are system-dependent, and there is little hope to find a significant amount of source difficulties whose resolution would help many languages at once. Given the percentage of differences in translation difficulties for related languages (e.g., for English-French and English-Spanish, up to 50% of segments `DT` for French are `ET` for Spanish), we consider multi-source and pivoting approaches to be a promising perspective in terms of MT quality improvement.

Our second main contribution consists in a proposal of **a Human-Assisted MT (HAMT) protocol** that accommodates the results of our translation difficulty detection procedure and enables resolution of those difficulties by pre-translation.

We have tested our difficulty resolution methodology in a sentence-level scenario. Then we have presented a document-level translation difficulty resolution scenario. Simulated experiments for both scenarios show that asking the human expert to pre-translate `DT` segments is realistic and efficient, and that the indirect effects of pre-edition are genuine (up to 23% of the total quality improvement). In our document-level resolution protocol we assumed that the pre-editor

is to provide the most general translation that will fit into the majority of contexts. The results of our simulated experiments were confirmed in preliminary document-level PRE experiments with students in translation studies. Our protocol is relatively low-cost and applicable to similar production contexts.

7.2 Perspectives

However, despite our contributions, we consider that we have only started to study the complex problem of translation difficulty detection and resolution. We will now detail some possible ways for improving our methodology and continuing our work.

One of the main improvements to our methodology lies in going beyond the formal evaluation of translation quality as proposed by automatic metrics. In our mind, one of the possible ways to do this is to introduce **semantic-based labeling** of `ET` and `DT`. In a way similar to semantic-based evaluations [Lo and Wu, 2011; Birch et al., 2016], a segmentation considering semantic roles can be taken into account. `DT` segments should be defined as segments translated with a significant initial information loss.

Another potential improvement of our framework consists in **partial automatization** of translation difficulty resolution. Here, once `DT` and `ET` segments are detected for input to a main MT system, translations of `DT` segments can be searched for in outputs of various other systems of similar quality for which those segments are `ET`. Such targeted resolution could help to locally improve MT without jeopardizing the quality of already "correctly" translated segments in a system combination approach (for instance, using additional phrase-tables, word- or phrase-based lattices, etc. [i.a., Cohn and Lapata, 2007; Schroeder et al., 2009; Ma and Mckeown, 2011; Freitag et al., 2014].

In a **multi-source difficulty resolution scenario** [Och and Ney, 2001; Schwartz, 2008; Schroeder et al., 2009; Zoph and Knight, 2016], already existing translations into other languages can be exploited to search for `ET` segments. In a similar way, `ET` segments can be detected in outputs of various out-of-domain systems, pivot systems, translating from language into another via a third language, outputs for paraphrased input, etc.

An interesting perspective for document-level translation difficulty resolution is proposed by **Bandit Learning**. Bandit Learning for MT is a framework where MT improves its translation choices based on actual human activity. It has recently attracted a lot of attention [Sokolov et al., 2016; Haffari et al., 2017]. The main goal in Bandit Learning is to maximize system quality

relative to the maximal possible quality, without any prior knowledge of the optimal solution. Thus, it can used to find an optimal strategy to detect `DT` segments to be pre-translated at the document level.

Finally, translation difficulty detection and resolution need to be explored for other domains, language pairs, and more importantly for other MT types, in particular for Neural MT.

Appendices

A | Extracts from the Cochrane Corpus

A.1 Cochrane Reference Corpus

Listing A.1 – Document CD000004: Source

```
<?xml version="1.0"?>
<DOC>
  <TITLE TRANSLATED="NO">Abdominal decompression for suspected fetal
  compromise/pre-eclampsia</TITLE>
  <SUMMARY>
    <SUMMARY_TITLE>Antenatal abdominal decompression for maternal hypertension or
  impaired fetal growth</SUMMARY_TITLE>
    <SUMMARY_BODY>
<P>Abdominal decompression was first used to increase blood flow and the forward
  movement of the uterus during labour contractions as a way of relieving pain. A
  rigid covered dome is placed about the abdomen and the space around the abdomen
  is decompressed to -50 to -100 mm Hg for 15 to 30 seconds out of each minute
  for 30 minutes once to thrice daily, or continuously during labour.
  Observations that fetal wellbeing appeared to be improved led to its
  investigation for complications of pregnancy.</P>
<P>Three randomised controlled studies with a total of 356 pregnant women were
  identified from a search of the medical literature, all with the possibility of
  containing serious methodological limitations. The studies were reported on
  between 1967 and 1973. One study involved women with pre-eclampsia, essential
  hypertension, or chronic nephritis. The other two trials assigned women
  carrying babies that were small for their gestational age to abdominal
  decompression or no decompression.</P>
```

```
<P>Abdominal decompression appeared to have a beneficial effect on the
  progression of pre-eclampsia. This one trial also reported less fetal distress
  during labour and fewer low 1-minute Apgar scores in the group who received
  abdominal decompression. The apparent large improvement in birthweight and
  perinatal deaths reported in all three studies is sufficiently striking to
  warrant the further evaluation of abdominal decompression in cases of impaired
  fetal growth, and possibly for women with pre-eclampsia, by means of
  methodologically sound controlled trials. Because of the methodological
  shortcomings mentioned above, clinical use of abdominal decompression cannot be
  supported on the basis of the present trials.</P>
</SUMMARY_BODY>
  </SUMMARY>
  <ABSTRACT>
    <ABS_BACKGROUND>
<P>Abdominal decompression was developed as a means of pain relief during labour.
  It has also been used for complications of pregnancy, and in healthy pregnant
  women in an attempt to improve fetal wellbeing and intellectual development.
  </P>
</ABS_BACKGROUND>
    <ABS_OBJECTIVES>
<P>The objective of this review was to assess the effects of antenatal abdominal
  decompression for maternal hypertension or impaired fetal growth, on perinatal
  outcome.</P>
</ABS_OBJECTIVES>
    <ABS_SEARCH_STRATEGY>
<P>The Cochrane Pregnancy and Childbirth Group&amp;quot;s Trials Register (2 February
  2012).</P>
</ABS_SEARCH_STRATEGY>
    <ABS_SELECTION_CRITERIA>
<P>Randomised or quasi-randomised trials comparing abdominal decompression with
  no decompression in women with pre-eclampsia and/or fetuses thought to be
  compromised.</P>
</ABS_SELECTION_CRITERIA>
    <ABS_DATA_COLLECTION>
<P>Eligibility and trial quality were assessed by one review author.</P>
</ABS_DATA_COLLECTION>
    <ABS_RESULTS>
<P>Three studies were included, all with the possibility of containing serious
  bias. Therapeutic abdominal decompression was associated with the following
  reductions: persistent pre-eclampsia (relative risk 0.36, 95% confidence
  interval 0.18 to 0.72); fetal distress in labour (relative risk 0.37, 95%
```

```
  confidence interval 0.19 to 0.71); low birthweight (relative risk 0.50, 95%
  confidence interval 0.40 to 0.63); Apgar scores less than six at one minute
  (relative risk 0.26, 95% confidence interval 0.12 to 0.56); and perinatal
  mortality (relative risk 0.39, 95% confidence interval 0.22 to 0.71).</P>
</ABS_RESULTS>
  <ABS_CONCLUSIONS>
<P>Due to the methodological limitations of the studies, the effects of
  therapeutic abdominal decompression are not clear. The apparent improvements in
  birthweight and perinatal mortality warrant further evaluation of abdominal
  decompression where there is impaired fetal growth and possibly for women with
  pre-eclampsia.</P>
</ABS_CONCLUSIONS>
  </ABSTRACT>
</DOC>
```

Listing A.2 – Document CD000004: Human Translation into French

```
<?xml version="1.0"?>
<DOC>
  <TITLE TRANSLATED="YES">Décompression abdominale en cas de suspicion d&quot;une
  souffrance fœtale/pré-éclampsie</TITLE>
  <SUMMARY>
    <SUMMARY_TITLE>Décompression abdominale prénatale pour une hypertension
  maternelle ou un retard de croissance du fœtus</SUMMARY_TITLE>
    <SUMMARY_BODY>
<P>La décompression abdominale a d&quot;abord été utilisée pour augmenter le flux
  sanguin et le mouvement vers l&quot;avant de l&quot;utérus pendant les
  contractions au cours du travail en tant que moyen de soulager la douleur. Un
  dôme rigide couvert est placé sur l&quot;abdomen et l&quot;espace autour de
  l&quot;abdomen est décompressé à une valeur de -50 à -100 mm de Hg pendant 15 à
  30 secondes, chaque minute, pendant 30 minutes, une à trois fois par jour, ou
  en continu pendant le travail. Des observations selon lesquelles le bien-être
  fœtal semblait être amélioré ont conduit à son étude en cas de complications de
  la grossesse.</P>
<P>Trois études contrôlées randomisées incluant un total de 356 femmes enceintes
  ont été identifiées à partir d&quot;une recherche dans la littérature médicale,
  toutes avec la possibilité de contenir de graves limitations méthodologiques.
  Les études ont été rapportées entre 1967 et 1973. Une étude portait sur des
  femmes présentant une pré-éclampsie, une hypertension artérielle essentielle,
  ou une néphrite chronique. Les deux autres essais ont affecté des femmes
  portant des bébés qui étaient petits pour leur âge gestationnel à une
  décompression ou à l&quot;absence de décompression abdominale.</P>
```

```
<P>La décompression abdominale semblait avoir un effet bénéfique sur
  l&quot;évolution de la pré-éclampsie. Ce même essai a également signalé moins
  de détresse fœtale pendant le travail et un nombre moins élevé de scores Apgar
  bas à 1 minute dans le groupe ayant bénéficié de la décompression abdominale.
  L&quot;importante amélioration apparente du poids à la naissance et de la
  mortalité périnatale signalée dans les trois études est suffisamment frappante
  pour justifier une évaluation plus poussée de la décompression abdominale en
  cas de retard de croissance du fœtus et peut-être pour les femmes atteintes de
  pré-éclampsie, par le biais d&quot;essais contrôlés robustes au plan
  méthodologique. En raison des lacunes méthodologiques mentionnées ci-dessus,
  l&quot;utilisation clinique de la décompression abdominale ne peut pas être
  encouragée sur la base des essais actuels.</P>
</SUMMARY_BODY>
  </SUMMARY>
  <ABSTRACT>
    <ABS_BACKGROUND>
<P>La décompression abdominale a été développée comme un moyen de soulager la
  douleur pendant le travail. Elle a également été utilisée pour des
  complications de la grossesse, et chez des femmes enceintes en bonne santé dans
  le but d&quot;améliorer le bien-être fœtal et le développement intellectuel.
  </P>
</ABS_BACKGROUND>
    <ABS_OBJECTIVES>
<P>L&quot;objectif de cette revue était d&quot;évaluer les effets de la
  décompression abdominale prénatale pour une hypertension maternelle ou un
  retard de croissance du fœtus, sur l&quot;issue périnatale.</P>
</ABS_OBJECTIVES>
    <ABS_SEARCH_STRATEGY>
<P>Le registre des essais cliniques du groupe Cochrane sur la grossesse et la
  naissance (jeudi 2 février 2012).</P>
</ABS_SEARCH_STRATEGY>
    <ABS_SELECTION_CRITERIA>
<P>Des essais randomisés ou quasi-randomisés comparant la décompression
  abdominale à l&quot;absence de décompression chez des femmes présentant une
  pré-éclampsie et/ou des fœtus supposés être en souffrance.</P>
</ABS_SELECTION_CRITERIA>
    <ABS_DATA_COLLECTION>
<P>L&quot;éligibilité et la qualité des essais ont été évaluées par un auteur de
  la revue.</P>
</ABS_DATA_COLLECTION>
    <ABS_RESULTS>
```

```
<P>Trois études ont été incluses, toutes avec la possibilité de contenir des
  biais importants. La décompression abdominale thérapeutique a été associée aux
  réductions suivantes : la pré-éclampsie persistante (risque relatif 0,36,
  intervalle de confiance à 95 % 0,18 à 0,72) ; la détresse fœtale pendant le
  travail (risque relatif 0,37, intervalle de confiance à 95 % 0,19 à 0,71) ; le
  faible poids à la naissance (risque relatif 0,50, intervalle de confiance à 95
  % 0,40 à 0,63) ; les scores Apgar inférieurs à six à une minute (risque relatif
  0,26, intervalle de confiance à 95 % 0,12 à 0,56) ; et la mortalité périnatale
  (risque relatif 0,39, intervalle de confiance à 95 % 0,22 à 0,71).</P>
</ABS_RESULTS>
    <ABS_CONCLUSIONS>
<P>En raison des limitations méthodologiques des études, les effets de la
  décompression abdominale thérapeutique ne sont pas clairs. Les améliorations
  apparentes du poids à la naissance et de la mortalité périnatale justifient une
  évaluation plus poussée de la décompression abdominale lorsqu&quot;il existe un
  retard de croissance du fœtus et peut-être pour les femmes atteintes de
  pré-éclampsie.</P>
</ABS_CONCLUSIONS>
  </ABSTRACT>
</DOC>
```

A.2 Cochrane Post-editing Corpus 1

Listing A.3 – Document CD001099: Source

```
<?xml version="1.0"?>
<DOC>
  <TITLE TRANSLATED="NO">Fibrinolytic agents for peripheral arterial
  occlusion</TITLE>
  <SUMMARY TRANSLATED="NO">
    <SUMMARY_TITLE>Drugs to break down blood clots for people with sudden onset
  peripheral arterial occlusion</SUMMARY_TITLE>
    <SUMMARY_BODY>
<P>Acute reduction in blood flow to a limb can be caused by a blood clot blocking
  an artery or a vascular graft. If not treated promptly this condition, known as
  peripheral arterial occlusion, can result in amputation or be life threatening.
  Infusion of clot-busting drugs can restore blood flow by dissolving the clot
  (thrombolysis). This review found some evidence from five randomized controlled
  trials, involving a total of 687 patients that suggested local infusion of a
  drug into the affected artery is more effective than infusion into a vein, and
  is also associated with a lower risk of unwanted bleeding. No particular drug
```

```
  was more effective in preventing limb loss or death than another. The drugs
  investigated were streptokinase, urokinase, recombinant tissue plasminogen
  activator and pro-urokinase. More research is needed to confirm these findings.
  All of the findings of this review came from small studies that involved people
  with peripheral arterial ischaemia of differing severity.</P>
</SUMMARY_BODY>
  </SUMMARY>
  <ABSTRACT TRANSLATED="NO">
    <ABS_BACKGROUND>
<P>Peripheral arterial thrombolysis is used in the management of peripheral
  arterial ischaemia. Streptokinase was originally used but safety concerns led
  to a search for other agents. Urokinase and recombinant tissue plasminogen
  activator (rt-PA) have increasingly become established as first line agents for
  peripheral arterial thrombolysis. Potential advantages of these agents include
  improved safety, greater efficacy and a more rapid response. Recently drugs
  such as pro-urokinase, recombinant staphylokinase and alfimperase have been
  introduced. This is an update of a review first published in 2010.</P>
</ABS_BACKGROUND>
    <ABS_OBJECTIVES>
<P>To determine which fibrinolytic agents are most effective in peripheral
  arterial ischaemia.<BR/>
</P>
</ABS_OBJECTIVES>
    <ABS_SEARCH_STRATEGY>
<P>For this update the Cochrane Peripheral Vascular Diseases Group Trials Search
  Co-ordinator (TSC) searched the Specialised Register (last searched March 2013)
  and CENTRAL (2013, Issue 3) for randomised controlled trials (RCTs) comparing
  fibrinolytic agents to treat peripheral arterial ischaemia.</P>
</ABS_SEARCH_STRATEGY>
    <ABS_SELECTION_CRITERIA>
<P>RCTs comparing fibrinolytic agents to treat peripheral arterial occlusion.</P>
</ABS_SELECTION_CRITERIA>
    <ABS_DATA_COLLECTION>
<P>Data were analysed for the outcomes vessel patency, time to lysis, limb
  salvage, amputation, death, complications including major haemorrhage, stroke,
  and distal embolization.</P>
</ABS_DATA_COLLECTION>
    <ABS_RESULTS>
<P>Five RCTs involving a total of 687 participants with a range of clinical
  indications were included. No new studies were included in this update. In one
  three-pronged study, vessel patency was greater with intra-arterial recombinant
```

```
  tissue plasminogen activator (rt-PA) than with intra-arterial streptokinase (P
  &lt; 0.04) or intravenous rt-PA (P &lt; 0.01). In participants with peripheral
  arterial occlusion there was no statistically significant difference in limb
  salvage at 30 days with either urokinase or rt-PA, though this may reflect the
  small numbers in the studies. Incidences of haemorrhagic complications varied
  with fibrinolytic regime but there was no statistically significant difference
  between intra-arterial urokinase and intra-arterial rt-PA. In the three-pronged
  study intravenous rt-PA and intra-arterial streptokinase were associated with a
  significantly higher risk of haemorrhagic complications than with
  intra-arterial rt-PA (P &lt; 0.05).</P>
</ABS_RESULTS>
    <ABS_CONCLUSIONS>
<P>There is some evidence to suggest that intra-arterial rt-PA is more effective
  than intra-arterial streptokinase or intravenous rt-PA in improving vessel
  patency in people with peripheral arterial occlusion. There was no evidence
  that rt-PA was more effective than urokinase for patients with peripheral
  arterial occlusion and some evidence that initial lysis may be more rapid with
  rt-PA, depending on the regime. Incidences of haemorrhagic complications were
  not statistically significantly greater with rt-PA than with other regimes.
  However, all of the findings come from small studies and a general paucity of
  results means that it is not possible to draw clear conclusions.</P>
</ABS_CONCLUSIONS>
  </ABSTRACT>
</DOC>
```

Listing A.4 – Document CD001099: MT into French

```
<?xml version="1.0"?>
<DOC>
  <TITLE TRANSLATED="YES"> Agents fibrinolytiques contre l&quot;occlusion
  artérielle périphérique </TITLE>
  <SUMMARY TRANSLATED="YES">
    <SUMMARY_TITLE> Médicaments destinés à dissoudre les caillots de sang chez
  les personnes souffrant d&quot;une occlusion artérielle périphérique soudaine
  </SUMMARY_TITLE>
    <SUMMARY_BODY>
      <P>  Réduction du débit sanguin aiguë d&quot;un membre peut être provoquée
  par un caillot sanguin bloque une artère ou un greffon vasculaire. Si elle
  n&quot;est pas traitée rapidement, cette maladie, appelée occlusion artérielle
  périphérique, peut entraîner une amputation ou menacer le pronostic vital. La
  perfusion de médicaments thrombolytiques peut restaurer le débit sanguin en
  dissolvant le caillot (thrombolyse). Cette revue a identifié certaines preuves
```

issues de cinq essais contrôlés randomisés, portant sur un total de 687 patients, qui suggérait une perfusion locale d"un médicament dans l"artère touchée est plus efficace qu"une perfusion dans une veine, et est également associée à un moindre risque de saignements indésirables. Aucun médicament particulier n"a été plus efficace pour prévenir la perte de la jambe ou de décès qu"un autre. Les médicaments étudiés étaient la streptokinase, l"urokinase, l"activateur tissulaire recombinant du plasminogène et la pro-urokinase. Des recherches supplémentaires sont nécessaires pour confirmer ces résultats. Tous les résultats de cette revue proviennent d"études de petite taille portant sur des personnes atteintes d"une ischémie artérielle périphérique de gravité diverse. </P>
</SUMMARY_BODY>
</SUMMARY>
<ABSTRACT TRANSLATED="YES">
<ABS_BACKGROUND>
<P> La thrombolyse artérielle périphérique est utilisée dans la prise en charge de l"ischémie artérielle périphérique. La streptokinase était utilisée à l"origine, mais des préoccupations de sécurité ont conduit à une recherche pour d"autres agents. L"urokinase et l"activateur tissulaire recombinant du plasminogène recombinant (rt-PA) se sont de plus en plus établis comme agents de première intention pour la thrombolyse artérielle périphérique. Les avantages potentiels de ces agents comprennent une amélioration de l"innocuité, une plus grande efficacité et une réponse plus rapide. Récemment, des médicaments, tels que la pro-urokinase, la staphylokinase recombinante et l"alfimpérase, ont été introduits. Ceci est une mise à jour d"une revue publiée pour la première fois en 2010. </P>
</ABS_BACKGROUND>
<ABS_OBJECTIVES>
<P> Agents fibrinolytiques déterminer quels sont les plus efficaces dans l"ischémie artérielle périphérique.
 </P>
</ABS_OBJECTIVES>
<ABS_SEARCH_STRATEGY>
<P> Pour cette mise à jour, le registre du groupe Cochrane sur les maladies vasculaires périphériques coordinateur de recherche d"études (TSC) effectué des recherches dans le registre spécialisé (dernière recherche en mars 2013) et CENTRAL (2013, numéro 3) pour les essais contrôlés randomisés (ECR) comparant des agents fibrinolytiques pour traiter l"ischémie artérielle périphérique. </P>
</ABS_SEARCH_STRATEGY>
<ABS_SELECTION_CRITERIA>
<P> Les ECR comparant des agents fibrinolytiques pour traiter

l"occlusion artérielle périphérique. </P>
</ABS_SELECTION_CRITERIA>
<ABS_DATA_COLLECTION>
<P> Les données ont été analysées concernant les critères de jugement de perméabilité des vaisseaux, de temps jusqu"à la lyse, de sauvetage de jambe, d"amputation, de décès, de complications, notamment d"hémorragie massive, d"AVC et d"embolisation distale. </P>
</ABS_DATA_COLLECTION>
<ABS_RESULTS>
<P> Cinq ECR portant sur un total de 687 participants atteints de diverses indications cliniques ont été inclus. Aucune nouvelle étude n"a été inclus dans cette mise à jour. Dans une étude à trois volets, la perméabilité des vaisseaux a été supérieur avec intra-artérielle activateur tissulaire du plasminogène recombinant (rt-PA) qu"avec la streptokinase intra-artérielle (P < 0,04) ou le rt-PA intraveineux (P < 0,01). Chez des participants souffrant d"occlusion artérielle périphérique il n"y avait aucune différence statistiquement significative en termes de sauvetage de jambe à 30 jours avec soit l"urokinase ou le rt-PA, bien que cela pourrait refléter le petit nombre d"études. L"incidence des complications hémorragiques variait selon le régime fibrinolytique, mais il n"y avait aucune différence statistiquement significative entre l"urokinase intra-artérielle et le rt-PA intra-artériel. Dans l"étude à trois volets, le rt-PA intraveineux et la streptokinase intra-artérielle ont été associés à un risque significativement plus élevé de complications hémorragiques que le rt-PA intra-artériel (P < 0,05). </P>
</ABS_RESULTS>
<ABS_CONCLUSIONS>
<P> Il existe certaines preuves suggèrent que le rt-PA intra-artériel est plus efficace que la streptokinase intra-artérielle ou le rt-PA intraveineux pour améliorer la perméabilité des vaisseaux chez les personnes souffrant d"occlusion artérielle périphérique. Il n"y avait aucune preuve que le rt-PA était plus efficace que l"urokinase pour les patients souffrant d"occlusion artérielle périphérique et certaines preuves que la lyse initiale pouvait être plus rapide avec le rt-PA, selon le régime. L"incidence des complications hémorragiques n"étaient pas statistiquement significativement plus élevé avec le rt-PA par rapport à d"autres traitements. Cependant, tous les résultats proviennent d"études de petite taille et un manque général de résultats signifie qu"il n"est pas possible de tirer des conclusions claires. </P>
</ABS_CONCLUSIONS>
</ABSTRACT>

```
</DOC>
```

Listing A.5 – Document CD001099: Post-Edition

```
<?xml version="1.0"?>
<DOC>
  <TITLE TRANSLATED="YES">Agents fibrinolytiques contre l&quot;occlusion
  artérielle périphérique</TITLE>
  <SUMMARY TRANSLATED="YES">
    <SUMMARY_TITLE>Médicaments destinés à dissoudre les caillots de sang chez les
  personnes souffrant d&quot;une occlusion artérielle périphérique
  soudaine</SUMMARY_TITLE>
    <SUMMARY_BODY>
<P>La réduction brutale du débit sanguin dans un membre peut être provoquée par
  un caillot sanguin obstruant une artère ou un greffon vasculaire. Si elle
  n&quot;est pas traitée rapidement, cet état, appelé occlusion artérielle
  périphérique, peut entraîner une amputation ou menacer le pronostic vital. La
  perfusion de médicaments thrombolytiques peut restaurer le débit sanguin en
  dissolvant le caillot (thrombolyse). Cette revue a identifié certaines preuves
  issues de cinq essais contrôlés randomisés, portant sur un total de 687
  patients, qui suggéraient qu&quot;une perfusion locale d&quot;un médicament
  dans l&quot;artère touchée est plus efficace qu&quot;une perfusion dans une
  veine et est également associée à un moindre risque de saignements
  indésirables. Aucun médicament particulier n&quot;a été plus efficace
  qu&quot;un autre pour prévenir l&quot;amputation du membre ou le décès. Les
  médicaments étudiés étaient la streptokinase, l&quot;urokinase,
  l&quot;activateur tissulaire recombinant du plasminogène et la pro-urokinase.
  Des recherches supplémentaires sont nécessaires pour confirmer ces résultats.
  Tous les résultats de cette revue proviennent d&quot;études de petite taille
  portant sur des personnes atteintes d&quot;une ischémie artérielle périphérique
  de gravité diverse.</P>
</SUMMARY_BODY>
  </SUMMARY>
  <ABSTRACT TRANSLATED="YES">
    <ABS_BACKGROUND>
<P>La thrombolyse artérielle périphérique est utilisée dans la prise en charge de
  l&quot;ischémie artérielle périphérique. La streptokinase était utilisée à
  l&quot;origine, mais des préoccupations de sécurité ont conduit à une recherche
  d&quot;autres agents. L&quot;urokinase et l&quot;activateur tissulaire
  recombinant du plasminogène recombinant (rt-PA) se sont de plus en plus établis
  comme agents de première intention pour la thrombolyse artérielle périphérique.
  Les avantages potentiels de ces agents comprennent une amélioration de
```

```
  l&quot;innocuité, une plus grande efficacité et une réponse plus rapide.
  Récemment, des médicaments, tels que la pro-urokinase, la staphylokinase
  recombinante et l&quot;alfimpérase, ont été introduits. Ceci est une mise à
  jour d&quot;une revue publiée pour la première fois en 2010.</P>
</ABS_BACKGROUND>
    <ABS_OBJECTIVES>
<P>Déterminer quels sont les agents fibrinolytiques les plus efficaces dans
  l&quot;ischémie artérielle périphérique.<BR/>
</P>
</ABS_OBJECTIVES>
    <ABS_SEARCH_STRATEGY>
<P>Pour cette mise à jour, le coordinateur de recherche d&quot;essais du groupe
  Cochrane sur les maladies vasculaires périphériques a effectué des recherches
  dans le registre spécialisé (dernière recherche en mars 2013) et CENTRAL (2013,
  numéro 3) pour les essais contrôlés randomisés (ECR) comparant des agents
  fibrinolytiques pour traiter l&quot;ischémie artérielle périphérique. </P>
</ABS_SEARCH_STRATEGY>
    <ABS_SELECTION_CRITERIA>
<P>Les ECR comparant des agents fibrinolytiques pour traiter l&quot;occlusion
  artérielle périphérique. </P>
</ABS_SELECTION_CRITERIA>
    <ABS_DATA_COLLECTION>
<P>Les données ont été analysées concernant les critères de jugement de
  perméabilité des vaisseaux, de temps jusqu&quot;à la lyse, de sauvetage du
  membre, d&quot;amputation, de décès, de complications, notamment
  d&quot;hémorragie massive, d&quot;AVC et d&quot;embolisation distale.</P>
</ABS_DATA_COLLECTION>
    <ABS_RESULTS>
<P>Cinq ECR portant sur un total de 687 participants de diverses indications
  cliniques ont été inclus. Aucune nouvelle étude n&quot;a été incluse dans cette
  mise à jour. Dans une étude à trois volets, la perméabilité des vaisseaux a été
  supérieure avec l&quot;activateur tissulaire du plasminogène recombinant
  (rt-PA) intra-artériel qu&quot;avec la streptokinase intra-artérielle (P &lt;
  0,04) ou le rt-PA intraveineux (P &lt; 0,01). Chez des participants atteints
  d&quot;occlusion artérielle périphérique il n&quot;y avait aucune différence
  statistiquement significative en termes de sauvetage de membre à 30 jours que
  ce soit avec l&quot;urokinase ou avec le rt-PA, bien que cela pourrait refléter
  le petit nombre de sujets  des études. L&quot;incidence des complications
  hémorragiques variait selon le  traitement  fibrinolytique, mais il n&quot;y
  avait aucune différence statistiquement significative entre l&quot;urokinase
  intra-artérielle et le rt-PA intra-artériel. Dans l&quot;étude à trois volets,
```

```
  le rt-PA intraveineux et la streptokinase intra-artérielle ont été associés à
  un risque significativement plus élevé de complications hémorragiques que le
  rt-PA intra-artériel (P &lt; 0,05).</P>
</ABS_RESULTS>
    <ABS_CONCLUSIONS>
<P>Il existe certaines preuves suggérant que le rt-PA intra-artériel est plus
  efficace que la streptokinase intra-artérielle ou le rt-PA intraveineux pour
  améliorer la perméabilité des vaisseaux chez les personnes souffrant
  d&quot;occlusion artérielle périphérique. Il n&quot;y avait aucune preuve que
  le rt-PA était plus efficace que l&quot;urokinase pour les patients souffrant
  d&quot;occlusion artérielle périphérique et certaines preuves que la lyse
  initiale pouvait être plus rapide avec le rt-PA, dependant du traitement.
  L&quot;incidence des complications hémorragiques n&quot;était pas
  statistiquement significativement plus élevée avec le rt-PA par rapport à
  d&quot;autres traitements. Cependant, tous les résultats proviennent
  d&quot;études de petite taille et avec, en général un petit nombre de
  résultats, cela signifie qu&quot;il n&quot;est pas possible de tirer des
  conclusions claires.</P>
</ABS_CONCLUSIONS>
  </ABSTRACT>
</DOC>
```

B | Extracts of Cochrane API Code

Listing B.1 – Class CochraneTranslateService.java

```
@Path("/translate")
public class CochraneTranslateService {

  static int tagId;

  private final static Logger LOGGER = Logger
      .getLogger(CochraneTranslateService.class.getName());

  public CochraneTranslateService() {

  }

  @POST
  @Produces(MediaType.APPLICATION_JSON)
  @Consumes(MediaType.MULTIPART_FORM_DATA)
  public Response submitFile(@FormDataParam("apiKey") String key,
      @FormDataParam("user") String user, @FormDataParam("id") String id,
      @FormDataParam("file") InputStream fileInputStream,
      @FormDataParam("srcLocale") String scrL,
      @FormDataParam("targetLocale") String trgL,
      @FormDataParam("externalStepId") String stepID,
      @FormDataParam("projectId") String projectID,
      @FormDataParam("returnUrl") String returnUrl,
      @FormDataParam("stringCount") int strCount)
      throws TranslateServiceException {

    String errorMessage = "Success";

```

```
29      LOGGER.info("Receiving request from Smartling");
30      SmartlingRequest smartReq = new SmartlingRequest(key, user, id,
31          fileInputStream, scrL, trgL, stepID, projectID, returnUrl,
32          strCount);
33      boolean flag;
34      flag = smartReq.checkCompliteness();
35
36      if (!flag) {
37        LOGGER.log(Level.SEVERE, "The request failed with the message",
38            smartReq.getErrorMessage());
39        errorMessage = smartReq.getErrorMessage();
40
41      } else {
42        LOGGER.info("Message compliteness checked");
43      }
44
45      UUID requestId = UUID.randomUUID();
46
47      if (smartReq.getFileContent()!= null) {
48
49        Properties configFile = new Properties();
50        try {
51          configFile.load(CochraneTranslateService.class.getClassLoader()
52              .getResourceAsStream("config.properties"));
53        } catch (IOException e1) {
54          throw new TranslateServiceException("Config file not found", e1);
55        }
56        String path = configFile.getProperty("home");
57        String pathConfig = configFile.getProperty("homeconfig");
58        String pathDraft = configFile.getProperty("homedraft");
59
60        String fileName = requestId.toString();
61        File file = new File(pathDraft + fileName);
62
63        try {
64          Files.copy(fileInputStream, file.toPath());
65        } catch (IOException e1) {
66          throw new TranslateServiceException(
67              "Failed to write the draft file", e1);
68        }
69
```

```
70        requestParamToFile(smartReq, pathConfig, fileName);
71
72        Xliff xliff = null;
73
74        try {
75          LOGGER.info("Parsing xliff");
76          xliff = CochraneTranslateService.readFromFile(pathDraft + fileName);
77          LOGGER.info("Writing file to the disk");
78          write(xliff, fileName, path);
79
80        } catch (Exception e) {
81          LOGGER.log(Level.SEVERE, "Parsing failed", e);
82          errorMessage = errorMessage + ". " + "Wrong xliff file format";
83        }
84
85      }
86
87      SmartlingResponse smartResponse = new SmartlingResponse(requestId,
88          errorMessage);
89      LOGGER.info("Sending response: " + errorMessage + " to Smartling");
90      return Response.status(202).entity(smartResponse).build();
91
92    }
93
94    public void requestParamToFile(SmartlingRequest smartReq, String path,
95        String fileName) throws TranslateServiceException {
96
97      PrintWriter out = null;
98      try {
99        out = new PrintWriter(new OutputStreamWriter(
100           new BufferedOutputStream(new FileOutputStream(path
101               + fileName + ".config")), "UTF-8"));
102     } catch (UnsupportedEncodingException | FileNotFoundException e) {
103       throw new TranslateServiceException(
104           "File not found or wrong format", e);
105     }
106     out.println("requestId=" + fileName);
107     out.println("returnUrl=" + smartReq.getReturnUrl());
108     out.println("id=" + smartReq.getId());
109     out.println("projectId=" + smartReq.getProjectID());
110     out.println("externalStepID=" + smartReq.getStepID());
```

```
111     out.println("locale=" + smartReq.getTrgL());
112     out.close();
113
114   }
115
116   public static Xliff readFromFile(String path) throws TranslateServiceException {
117
118     File file = new File(path);
119     JAXBContext jc = null;
120     try {
121       jc = JAXBContext.newInstance(xliffsmart.Xliff.class);
122     } catch (JAXBException e) {
123       throw new TranslateServiceException("Jaxb context could not be created", e);
124     }
125     Unmarshaller u = null;
126     try {
127       u = jc.createUnmarshaller();
128     } catch (JAXBException e) {
129       throw new TranslateServiceException(
130           "Jaxb unmarshaller could not be created", e);
131     }
132     Object obj = null;
133     try {
134       obj = u.unmarshal(file);
135     } catch (JAXBException e) {
136       throw new TranslateServiceException("The file could not be unmarshalled",
    e);
137     }
138
139     if (obj instanceof Xliff)
140       return (Xliff) obj;
141     else
142       throw new TranslateServiceException("Unknown xliff object format");
143   }
144
145   public static Xliff read(InputStream st) throws TranslateServiceException {
146
147     JAXBContext jc = null;
148     try {
149       jc = JAXBContext.newInstance(xliffsmart.Xliff.class);
150     } catch (JAXBException e) {
```

```
151       throw new TranslateServiceException(
152           "Jaxb context could not be created", e);
153     }
154     Unmarshaller u = null;
155     try {
156       u = jc.createUnmarshaller();
157     } catch (JAXBException e) {
158       throw new TranslateServiceException(
159           "Jaxb unmarshaller could not be created", e);
160     }
161     Object obj = null;
162     try {
163       obj = u.unmarshal(st);
164     } catch (JAXBException e) {
165       throw new TranslateServiceException(
166           "The file could not be unmarshalled", e);
167     }
168 
169     if (obj instanceof Xliff)
170       return (Xliff) obj;
171     else
172       throw new TranslateServiceException("Unknown xliff object format");
173   }
174 
175   public static boolean write(Xliff xliff, String id, String path)
176       throws TranslateServiceException {
177 
178     File file = new File(path + id);
179     JAXBContext jc = null;
180     try {
181       jc = JAXBContext.newInstance(xliffsmart.Xliff.class);
182     } catch (JAXBException e) {
183       throw new TranslateServiceException(
184           "Jaxb context could not be created", e);
185     }
186     Marshaller m = null;
187     try {
188       m = jc.createMarshaller();
189     } catch (JAXBException e1) {
190       throw new TranslateServiceException(
191           "Jaxb marshaller could not be created", e1);
```

```
192     }
193     try {
194       m.setProperty("com.sun.xml.bind.xmlDeclaration", Boolean.FALSE);
195     } catch (PropertyException e) {
196       throw new TranslateServiceException(
197           "Marshaller property can't be set", e);
198     }
199     try {
200       m.setProperty("com.sun.xml.bind.xmlHeaders",
201           "<?xml version=\"1.0\" encoding=\"UTF-8\"?>");
202     } catch (PropertyException e) {
203       throw new TranslateServiceException(
204           "Marshaller property can't be set", e);
205     }
206     try {
207       m.setProperty(Marshaller.JAXB_SCHEMA_LOCATION,
208           "urn:oasis:names:tc:xliff:document:1.2 xliff-core-1.2-strict.xsd");
209     } catch (PropertyException e) {
210       throw new TranslateServiceException(
211           "Marshaller property can't be set", e);
212     }
213
214     try {
215       m.marshal(xliff, file);
216     } catch (JAXBException e) {
217       throw new TranslateServiceException(
218           "Te xliff object can't be marshlled", e);
219     }
220     return true;
221
222   }
223
224 }
```

C | Extracts of Cochrane UI Code

Listing C.1 – Class FuzzyWidget.java

```
1  public class FuzzyWidget extends Composite {
2
3   private static FuzzyWidgetUiBinder uiBinder = GWT
4       .create(FuzzyWidgetUiBinder.class);
5
6   interface FuzzyWidgetUiBinder extends UiBinder<Widget, FuzzyWidget> {
7   }
8
9   public FuzzyWidget() {
10    initWidget(uiBinder.createAndBindUi(this));
11  }
12
13  @UiField
14  FlexTable fuzzyList;
15  @UiField
16  Button closeButton;
17  SentenceWidget sw;
18  FuzzyDialog fd;
19
20  public FuzzyWidget(FuzzyDialog fd, final SentenceWidget sw, final
    List<FuzzyMatch> ll) {
21    initWidget(uiBinder.createAndBindUi(this));
22    closeButton.setText("Close");
23    this.sw = sw;
24    this.fd = fd;
25    String group = "FuzzyGroup";
26    HTML head1 = new HTML("Source");
27    HTML head2 = new HTML("Target");
```

```
28
29      head1.addStyleName("center");
30      head2.addStyleName("center");
31
32      Collections.sort(ll, new FuzzyMatchComparator());
33
34      fuzzyList.setWidget(0, 1, head1);
35      fuzzyList.setWidget(0, 2, head2);
36
37      int cc=1;
38      for (final FuzzyMatch fm : ll){
39
40        fuzzyList.addStyleName("flexTableForm");
41        RadioButton rb = new RadioButton(group,
    String.valueOf(fm.getPercentage())+"%");
42        if(fm.isChosen()){
43          rb.setValue(true);
44        }
45        fuzzyList.setWidget(cc, 0, rb);
46
47        rb.addValueChangeHandler(new ValueChangeHandler<Boolean>() {
48                @Override
49                public void onValueChange(ValueChangeEvent<Boolean> e) {
50                  fm.setChosen(true);
51                    sw.putFuzzyMatch(fm);
52                }});
53
54
55
56        fuzzyList.setWidget(cc, 1, new HTML(fm.getSrc()));
57        fuzzyList.setWidget(cc, 2, new HTML(fm.getMatch()));
58
59        cc++;
60      }
61
62    }
63
64    @UiHandler("closeButton")
65    void onClick2(ClickEvent e) {
66      fd.hide();
67
```

```
68   }
69
70
71
72 }
```

Listing C.2 – Class PTWidget.java

```
1  public class PTWidget extends Composite {
2
3    private static PTWidgetUiBinder uiBinder = GWT
4        .create(PTWidgetUiBinder.class);
5
6    interface PTWidgetUiBinder extends UiBinder<Widget, PTWidget> {
7    }
8
9    public PTWidget() {
10     initWidget(uiBinder.createAndBindUi(this));
11
12   }
13   @UiField
14   FlexTable ptTable;
15   TranslationViewImpl translationViewImpl;
16   List<Anchor> anchorList;
17
18   public PTWidget(Sortie sr, final TranslationViewImpl translationViewImpl) {
19     initWidget(uiBinder.createAndBindUi(this));
20     ptTable.setWidget(0, 2, new Label("Corpus Search Results: " + sr.getSrc()));
21     ptTable.setWidget(0, 4, new Label("%"));
22     anchorList = new ArrayList<Anchor>();
23     this.translationViewImpl = translationViewImpl;
24
25     Map<String, Integer> map  = sortByComparator(sr.getTrgL(), false);
26
27     int counter = 1;
28
29     for (final String str : map.keySet()){
30
31       final Anchor trg = new Anchor();
32       trg.setText(str);
33       trg.addStyleName("anchor");
34             if(counter ==1){
```

```
35
36          translationViewImpl.populateCrpView(str);
37          trg.addStyleName("red");
38        }
39
40        trg.addClickHandler(new ClickHandler() {
41              public void onClick(ClickEvent event) {
42
43                for (Anchor a : anchorList){
44                  a.removeStyleName("red");
45                }
46
47                trg.addStyleName("red");
48                translationViewImpl.populateCrpView(str);
49              }});
50
51        anchorList.add(trg);
52        ptTable.setHTML(counter, 0, String.valueOf(counter));
53        ptTable.setWidget(counter, 2, trg);
54        ptTable.setWidget(counter, 4, new HTML(String.valueOf(map.get(str))));
55        counter++;
56      }
57
58    }
59
60    private static Map<String, Integer> sortByComparator(Map<String, Integer>
    unsortMap, final boolean order)
61      {
62          List<Entry<String, Integer>> list = new LinkedList<Entry<String,
    Integer>>(unsortMap.entrySet());
63
64          // Sorting the list based on values
65          Collections.sort(list, new Comparator<Entry<String, Integer>>()
66          {
67              public int compare(Entry<String, Integer> o1,
68                      Entry<String, Integer> o2)
69              {
70                  if (order)
71                  {
72                      return o1.getValue().compareTo(o2.getValue());
73                  }
```

```
74                  else
75                  {
76                      return o2.getValue().compareTo(o1.getValue());
77
78                  }
79              }
80          });
81
82          // Maintaining insertion order with the help of LinkedList
83          Map<String, Integer> sortedMap = new LinkedHashMap<String, Integer>();
84          for (Entry<String, Integer> entry : list)
85          {
86              sortedMap.put(entry.getKey(), entry.getValue());
87          }
88
89          return sortedMap;
90      }
91  }
```

D | Examples of Medical Text Challenges

Table D.1 – Examples of PLS and ABS test set sentences

	PLS
src.	**A lack of growth** and poor nutrition are common in children with chronic diseases like cystic fibrosis and **paediatric cancer**.
Cochr-SMT	***Un manque de la croissance*** *et une mauvaise nutrition sont fréquents chez les enfants atteints de maladies chroniques comme la mucoviscidose et* ***le cancer pédiatrique****.* '**A lack of growth** and bad nutrition are common in the children suffering from chronic diseases like cystic fibrosis and **paediatric cancer**.'
oracle	***Un manque de la croissance*** *et une mauvaise nutrition sont fréquents chez les enfants atteints de maladies chroniques comme la mucoviscidose et* ***les cancers. chez les enfants*** '**A lack of growth** and bad nutrition are common in the children suffering from chronic diseases like cystic fibrosis and **cancers. in the children**'
ref.	***Une croissance réduite*** *et une mauvaise nutrition sont fréquentes chez les enfants atteints de maladies chroniques comme la mucoviscidose et* ***les cancers pédiatriques****.* '**A reduced growth** and bad nutrition are common in the children suffering from chronic diseases like cystic fibrosis and **the paediatric cancers**.'
	ABS
src.	Poor growth and nutritional status are common **in children with chronic diseases**.
Cochr-SMT	*Une mauvaise croissance et le statut nutritionnel sont fréquents* ***chez les enfants atteints de maladies chroniques****.* 'A bad growth and the nutritional status are common **in the children suffering from chronic diseases**.'

oracle	*Une mauvaise croissance et le statut nutritionnel sont fréquents* ***chez l'enfant de*** 'A bad growth and the nutritional status are common **in the child of**'
ref.	*Une croissance réduite et un mauvais statut nutritionnel sont fréquents* ***chez l'enfant atteint de maladie chronique****.* 'A reduced growth and a bad nutritional status are common **in the child suffering from a chronic disease**.'

Table D.2 – Examples of sentence restructuring

src.	Adverse events were not reported in any of the included studies.
`Cochr-SMT`	*Les événements indésirables n'étaient pas rapportés dans aucune des études incluses.* 'The adverse events were not reported in any of the included studies.'
oracle	*Les événements indésirables n'étaient rapportés dans aucune des études incluses ne.* 'The adverse events were not reported in any of the included studies not.'
PE	*Aucune des études incluses ne rapportait d'événements indésirables.* 'None of the included studies reported adverse events.'
src.	However, the evidence for survival improvement is still lacking.
`Cochr-SMT`	*Cependant, les preuves d'amélioration de la survie est encore manquantes.* 'However, the proofs of the improvement of survival is still missing.'
oracle	*Cependant, les preuves d'amélioration de la survie, il manque toujours de la.* 'However, the proofs of the improvement of survival, it misses still the.'
PE	*Cependant, il manque toujours de données probantes sur l'amélioration de la survie.* 'However, it still misses the proving data on the improvement of survival.'

E | Standard Features for Translation Difficulty Detection

E.1 List of word-level standard features

Source Features

1. source token count;
2. left context of the source token;
3. right context of the source token;
4. if source token is stopword;
5. if source token is punctuation;
6. if source token is proper noun;
7. if source token is digit;
8. source highest order n-gram left;
9. source highest order n-gram right;
10. source back-off behavior left;
11. source back-off behavior middle;
12. source back-off behavior right;

Target Features

1. target token count;
2. source target token count ratio;
3. aligned target token;
4. left context of the target token;
5. right context of the target token;
6. target highest order n-gram left;
7. target highest order n-gram right;
8. POS tag of the target token.

E.2 List of standard phrase-level features

Source Features

1. number of tokens in the source segment;
2. average source token length;
3. LM probability of source segment;
4. source segment perplexity;
5. average unigram frequency in quartile 1 of frequency (lower frequency words) in the corpus of the source language;
6. average unigram frequency in quartile 2 of frequency (lower frequency words) in the corpus of the source language;
7. average unigram frequency in quartile 3 of frequency (lower frequency words) in the corpus of the source language;
8. average unigram frequency in quartile 4 of frequency (lower frequency words) in the corpus of the source language;

9. average bigram frequency in quartile 1 of frequency (lower frequency words) in the corpus of the source language;
10. average bigram frequency in quartile 2 of frequency (lower frequency words) in the corpus of the source language;
11. average bigram frequency in quartile 3 of frequency (lower frequency words) in the corpus of the source language;
12. average bigram frequency in quartile 4 of frequency (lower frequency words) in the corpus of the source language;
13. average trigram frequency in quartile 1 of frequency (lower frequency words) in the corpus of the source language;
14. average trigram frequency in quartile 2 of frequency (lower frequency words) in the corpus of the source language;
15. average trigram frequency in quartile 3 of frequency (lower frequency words) in the corpus of the source language;
16. average trigram frequency in quartile 4 of frequency (lower frequency words) in the corpus of the source language;
17. percentage of distinct unigrams seen in the corpus of the source language (in all quartiles);
18. percentage of distinct bigrams seen in the corpus of the source language (in all quartiles);
19. percentage of distinct trigrams seen in the corpus of the source language (in all quartiles);
20. percentage of punctuation marks in source;
21. percentage of numbers in the source;
22. number source tokens that do not contain only a-z.

Target Features

1. number of tokens in the target segment;
2. ratio of number of tokens in source and target;

3. LM probability of target segment;
4. perplexity of the target;
5. number of occurrences of the target word within the target hypothesis (averaged for all words in the hypothesis - type/token ratio);
6. average number of translations per source word in the segment (threshold in giza1: prob > 0.01);
7. average number of translations per source word in the segment (threshold in giza1: prob > 0.05);
8. average number of translations per source word in the segment (threshold in giza1: prob > 0.1);
9. average number of translations per source word in the segment (threshold in giza1: prob > 0.2);
10. average number of translations per source word in the segment (threshold in giza1: prob > 0.5);
11. average number of translations per source word in the segment (threshold in giza1: prob > 0.01) weighted by the frequency of each word in the source corpus;
12. average number of translations per source word in the segment (threshold in giza1: prob > 0.5) weighted by the frequency of each word in the source corpus;
13. average number of translations per source word in the segment (threshold in giza1: prob > 0.1) weighted by the frequency of each word in the source corpus;
14. average number of translations per source word in the segment (threshold in giza1: prob > 0.2) weighted by the frequency of each word in the source corpus;
15. average number of translations per source word in the segment (threshold in giza1: prob > 0.5) weighted by the frequency of each word in the source corpus;
16. absolute difference between number of periods in source and target;
17. absolute difference between number of periods in source and target normalized by target length;

18. absolute difference between number of commas in source and target;
19. absolute difference between number of commas in source and target normalized by target length;
20. absolute difference between number of : in source and target;
21. absolute difference between number of : in source and target normalized by target length;
22. absolute difference between number of ; in source and target;
23. absolute difference between number of ; in source and target normalized by target length;
24. absolute difference between number of ? in source and target;
25. absolute difference between number of ? in source and target normalized by target length;
26. absolute difference between number of ! in source and target;
27. absolute difference between number of ! in source and target normalized by target length;
28. percentage of punctuation marks in target;
29. absolute difference between number of punctuation marks between source and target normalized by target length;
30. percentage of numbers in the target segment;
31. absolute difference between number of numbers in the source and target segment normalized by source segment length;
32. percentage of tokens in the target which do not contain only a-z;
33. ratio of percentage of tokens a-z in the source and tokens a-z in the target;
34. number of unaligned target words;
35. number of target words aligned to more than one word;
36. average number of alignments per word in the target phrase;
37. target POS sequence;

F | Feature Ablation Experiments

Table F.1 – Feature ablation experiments: WORD-SEG segmentation, UN domain

EN-AR			
set	F_{DT}	F_{ET}	F_{mcr}
SRC-wrd	0.66 ± 0.0003	0.69 ± 0.0004	0.68 ± 0.0003
AR-wrd	0.68 ± 0.0007	0.74 ± 0.0006	0.71 ± 0.0006
random	0.48 ± 0.0025	0.52 ± 0.0023	0.50 ± 0.0021
all	$\mathbf{0.68 \pm 0.0006}$	$\mathbf{0.75 \pm 0.0005}$	$\mathbf{0.71 \pm 0.0005}$
-ES-wrd	0.68 ± 0.0007	0.74 ± 0.0005	
-FR-wrd	0.68 ± 0.0005	0.74 ± 0.0003	
-RU-wrd	0.68 ± 0.0005	0.75 ± 0.0004	
-FR-wrd-RU-wrd	0.68 ± 0.0004	0.74 ± 0.0003	
-ES-wrd-RU-wrd	0.68 ± 0.0005	0.75 ± 0.0005	
-ES-wrd-FR-wrd	0.68 ± 0.0006	0.74 ± 0.0005	
EN-ES			
set	F_{DT}	F_{ET}	F_{mcr}
SRC-wrd	0.57 ± 0.0006	0.73 ± 0.0004	0.65 ± 0.0005
ES-wrd	0.60 ± 0.0005	0.77 ± 0.0003	0.69 ± 0.0004
random	0.40 ± 0.0018	0.57 ± 0.0018	0.49 ± 0.0015
all	$\mathbf{0.61 \pm 0.0005}$	$\mathbf{0.77 \pm 0.0004}$	$\mathbf{0.69 \pm 0.0004}$
-AR-wrd	0.61 ± 0.0005	0.77 ± 0.0004	
-FR-wrd	0.61 ± 0.0005	0.76 ± 0.0003	
-RU-wrd	0.61 ± 0.0007	0.77 ± 0.0004	
-FR-wrd-RU-wrd	0.61 ± 0.0005	0.76 ± 0.0003	
-AR-wrd-RU-wrd	0.61 ± 0.0004	0.77 ± 0.0003	
-AR-wrd-FR-wrd	0.61 ± 0.0005	0.76 ± 0.0004	

EN-FR			
set	F_{DT}	F_{ET}	F_{mcr}
SRC-wrd	0.60 ± 0.0006	0.69 ± 0.0005	0.64 ± 0.0005
FR-wrd	0.63 ± 0.0005	0.74 ± 0.0004	0.68 ± 0.0004
random	0.43 ± 0.0018	0.55 ± 0.0018	0.49 ± 0.0017
all	$\mathbf{0.63 \pm 0.0006}$	$\mathbf{0.74 \pm 0.0006}$	$\mathbf{0.67 \pm 0.0006}$
-AR-wrd	0.63 ± 0.0005	0.74 ± 0.0005	
-ES-wrd	0.63 ± 0.0007	0.74 ± 0.0007	
-RU-wrd	0.63 ± 0.0007	0.74 ± 0.0005	
-ES-wrd**-**RU-wrd	0.63 ± 0.0006	0.74 ± 0.0004	
-AR-wrd**-**RU-wrd	0.63 ± 0.0006	0.74 ± 0.0006	
-AR-wrd**-**ES-wrd	0.63 ± 0.0008	0.74 ± 0.0003	
EN-RU			
set	F_{DT}	F_{ET}	F_{mcr}
SRC-wrd	0.72 ± 0.0003	0.67 ± 0.0004	0.69 ± 0.0004
RU-wrd	0.74 ± 0.0004	0.72 ± 0.0006	0.73 ± 0.0004
random	0.53 ± 0.0014	0.47 ± 0.0027	0.50 ± 0.0019
all	$\mathbf{0.75 \pm 0.0007}$	$\mathbf{0.72 \pm 0.0009}$	$\mathbf{0.73 \pm 0.0008}$
-AR-wrd	0.75 ± 0.0004	0.72 ± 0.0004	
-ES-wrd	0.75 ± 0.0004	0.72 ± 0.0008	
-FR-wrd	0.75 ± 0.0005	0.72 ± 0.0007	
-ES-wrd**-**FR-wrd	0.74 ± 0.0003	0.72 ± 0.0008	
-AR-wrd**-**FR-wrd	0.75 ± 0.0004	0.72 ± 0.0006	
-AR-wrd**-**ES-wrd	0.75 ± 0.0005	0.72 ± 0.0007	

G | Examples of the Impact on the Context

Table G.1 – Example of positive context influence for the `WMT14-SMT` system (bold text denotes influenced contextual segments, underlined text – pre-translated segments)

src.	Whether social skills programmes or training can improve the social functioning of people with schizophrenia in different settings remains unclear and should be further investigated in a large multi-centre randomised controlled trial.
PRE	*Si la possibilité d'programmes de peut améliorer le fonctionnement social des personnes atteintes de schizophrénie dans différents environnements reste incertaine et doit en outre être* ***étudiée*** *dans un multicentrique à grande échelle.* 'If the possibility of programs can improve the social functioning of people with schizophrenia in different environments remains uncertain and must furthermore be **studied (Fem., Sing)** in a large multi-centric randomized controlled.'
MT	*Si les programmes habiletes sociale ou d'entraînement peut améliorer le fonctionnement social des personnes souffrant de schizophrénie dans différents réglages reste incertain et doit encore être* ***étudié*** *dans un grand multi-centriques randomisée contrôlée.* 'Whether social skills or training programs can improve the social functioning of people suffering from schizophrenia in different settings remains uncertain and must still be **studied (Masc., Sing)** in a large multi-centric randomized controlled.'
ref.	*La possibilité d'améliorer le fonctionnement social de personnes atteintes de schizophrénie par des programmes de compétences sociales dans différents environnements reste incertaine et devrait être* ***étudiée*** *plus avant par un essai contrôlé randomisé multicentrique à grande échelle.* 'The possibility of improving the social functioning of people with schizophrenia by programs of social competencies in different settings remains unclear and should be further **studied (Fem., Sing)** in a multicenter randomized controlled trial on a large scale.'

Table G.2 – Example of negative context influence for the `WMT14-SMT` system (bold text denotes influenced contextual segments, underlined text – pre-translated segments)

src.	Are treatments with eye drops of antihistamines and mast cell stabilisers, alone or in combination, effective and safe in people with seasonal and allergic conjunctivitis?
PRE	*Des traitements avec des gouttes ophtalmiques aux de stabilisateurs et des mastocytes, seuls ou en combinaison, efficace et sûre* ***dans*** *les personnes atteintes de conjonctivite saisonnière et allergique?* 'Treatments with ophthalmic drops of stabilizers and mast cells, alone (Masc. Pl.) or in combination, effective and safe **into** the people with seasonal and allergic conjunctivitis?'
MT	*Des traitements avec des gouttes ophtalmiques et d'antihistamines mastocyte thymorégulateurs, seul ou en combinaison, efficace et sûr* ***chez*** *les personnes atteintes de trouble affectif saisonnier et la conjonctivite allergique?* 'Treatments with ophthalmic drops and antihistamines mastocyte thymoregulators, alone (Masc. Sg.) or in combination, effective and safe **in** the people with seasonal affective disorder and the allergic conjunctivitis?'
ref.	*Les traitements par des gouttes ophtalmiques aux antihistaminiques et aux stabilisateurs des mastocytes, seuls ou en association, sont-ils efficaces et sûrs* ***chez*** *les personnes atteintes de conjonctivite saisonnière et allergique?* 'The treatments with ophthalmic drops with antihistamines and mast cell stabilizers, alone (Masc. Pl.) or in combination, are they effective and safe **in** the people with seasonal and allergic conjunctivitis?'

Table G.3 – Example of positive context influence for the English-Russian `UN` system (bold text denotes influenced contextual segments, underlined text – pre-translated segments)

src.	Needy countries were offered food and assistance, industries were restored, economies were rehabilitated.
PRE	***Нуждающимся странам*** *предоставлялись продовольствие и продовольственной помощи, восстанавливалась, промышленность оздоровлялась экономика* '**To needy countries (Dat. Pl.)** food was provided and of food aid, restored, industry the economy was rehabilitated.'
MT	***Нуждающихся стран*** *была оказана продовольственная помощь, восстановление и отраслях экономики, были восстановлены.* '**Needy countries (Akk. Pl.)** food aid was provided, restoration and in sectors of the economy, were restored.'
ref.	***Нуждающимся странам*** *предоставлялись продовольствие и помощь, восстанавливалась промышленность, оздоровлялась экономика.*

	'**To needy countries (Dat. Pl.)** food and aid were provided, industry was restored, and the economy was rehabilitated.'

H | Cochrane Review Abstract Pre- and Post-Edited by Humans

Table H.1 – Document CD009658

TITLE			
primary@0-0 prophylaxis@0-1 for **venous@2-0 thromboembolism@2-1** in patients undergoing cardiac or thoracic surgery .	la prophylaxie primaire de la thromboembolie veineuse chez les patients subissant une chirurgie cardiaque ou thoracique .	**prophylaxie primaire** pour la **maladie thrombo-embolique veineuse** chez les patients subissant une chirurgie cardiaque ou thoracique .	prophylaxie primaire de la maladie thrombo-embolique veineuse chez les patients subissant une chirurgie cardiaque ou thoracique .
SUMMARY TITLE			
prevention of **blood@10-0 clots@10-1** in patients undergoing cardiac or thoracic surgery .	la prévention de la formation de caillots sanguins chez les patients subissant une chirurgie cardiaque ou thoracique .	la prévention des **caillots de sang** chez les patients subissant une chirurgie cardiaque ou thoracique .	prévention des caillots de sang chez les patients subissant une chirurgie cardiaque ou thoracique .

marked src.	initial MT	final MT	PE
SUMMARY BODY			
background .	contexte .	contexte .	contexte
patients undergoing surgery have an increased probability of developing **blood@10-0 clots@10-1** in their veins (**venous@2-0 thromboembolism@2-1**) .	les patients subissant une chirurgie ont une augmentation de la probabilité de développer des caillots de sang dans les veines (thromboembolie veineuse) .	les patients subissant une intervention chirurgicale ont une probabilité accrue de développer des **caillots de sang** dans les veines (**maladie thrombo-embolique veineuse**) .	les patients subissant une intervention chirurgicale risquent davantage de développer des caillots de sang dans leurs veines (maladie thrombo-embolique veineuse) .
these clots may be in the deep veins (deep vein thrombosis) or travel to the lungs (pulmonary embolism) .	ces caillots peuvent être dans les veines profondes (thrombose veineuse profonde) ou atteignent les poumons (embolie pulmonaire) .	ces caillots peuvent être dans les veines profondes (thrombose veineuse profonde) ou atteignent les poumons (embolie pulmonaire) .	ces caillots peuvent être dans les veines profondes (thrombose veineuse profonde) ou atteindre les poumons (embolie pulmonaire) .
as in other types of surgery , effective prevention of **blood@10-0 clots@10-1** (thromboprophylaxis) after cardiac or thoracic surgery may reduce the risk of postoperative vein clots .	comme pour d' autres types de chirurgie , la prévention efficace de caillots sanguins (thromboprophylaxie) après une chirurgie cardiaque ou thoracique peut réduire le risque de caillots veineux postopératoire .	comme pour d' autres types de chirurgie , la prévention efficace de **caillots de sang** (thromboprophylaxie) après une chirurgie cardiaque ou thoracique peut réduire le risque de caillots veineux postopératoire .	comme pour d' autres types de chirurgie , la prévention efficace des caillots de sang (thromboprophylaxie) après une chirurgie cardiaque ou thoracique peut réduire le risque de caillots veineux postopératoires .
these potential benefits , however , **have@40-0 to@40-1 be@40-2 balanced@40-3** against the associated risks of bleeding .	ces bénéfices potentiels , cependant , doivent être mis en balance avec les risques associés de saignements .	ces bénéfices potentiels , cependant , **doivent être équilibrés** contre les risques associés de saignements .	cependant , ces bénéfices potentiels doivent être contrebalancés par les risques associés d' hémorragie .

marked src.	initial MT	final MT	PE
this systematic review **looked@45-0 at@46-0** the effectiveness and safety of anticoagulants (**medicines@49-0 that@49-1 reduce@50-0** the ability of the blood to clot) , mechanical interventions (such as **pneumatic@57-0 pumps@57-1** on the legs **to@60-0 promote@60-1** blood flow) , and **caval@63-0 filters@63-1** (a type of vascular filter , implanted into the main abdominal vein **to@69-0 prevent@69-1** movement of clots from the legs to the lungs) in patients undergoing cardiac or thoracic surgery .	cette revue systématique a examiné l' efficacité et l' innocuité des anticoagulants (médicaments qui réduisent la capacité du sang à coaguler) , les interventions mécaniques (tels que les pompes pneumatique sur les jambes pour promouvoir le flux sanguin) , et les filtres cave (un type de filtre vasculaire , implanter dans la principale veine abdominale pour prévenir le mouvement de caillots sanguins dans les jambes vers les poumons) chez les patients subissant une chirurgie cardiaque ou thoracique .	cette revue systématique **a examiné** l' efficacité et l' innocuité des anticoagulants (**des médicaments qui réduisant** la capacité du sang à coaguler) , les interventions mécaniques (tels que les **pompes pneumatiques** sur les jambes **favoriser** le débit sanguin) , et **filtre cave** (un type de filtre vasculaire , implanter dans la principale veine abdominale **empêcher** le mouvement de caillots sanguins dans les jambes vers les poumons) chez les patients subissant une chirurgie cardiaque ou thoracique .	cette revue systématique a examiné l' efficacité et l' innocuité des anticoagulants (médicaments qui réduisent la capacité du sang à coaguler) , les interventions mécaniques (telles que les pompes pneumatiques sur les jambes visant à favoriser le débit sanguin) , et les filtres caves (type de filtre vasculaire implanté dans la principale veine abdominale afin d' empêcher le mouvement des caillots sanguins des jambes vers les poumons) chez les patients subissant une chirurgie cardiaque ou thoracique .
study@73-0 characteristics@73-1 and **key@75-0 results@75-1** .	les caractéristiques des études et les principaux résultats .	**caractéristiques des études** et des **résultats clés** .	caractéristiques des études et résultats-clés

marked src.	initial MT	final MT	PE
we identified @num@ randomised controlled trials (@num@ participants) , six for cardiac surgery (@num@ participants) and seven for thoracic surgery (@num@ participants) .	nous avons identifié @num@ essais contrôlés randomisés (@num@ participants) , six dans la chirurgie cardiaque (@num@ participants) et sept pour la chirurgie thoracique (@num@ participants) .	nous avons identifié @num@ essais contrôlés randomisés (@num@ participants) , six dans la chirurgie cardiaque (@num@ participants) et sept pour la chirurgie thoracique (@num@ participants) .	nous avons identifié @num@ essais randomisés contrôlés (@num@ participants) , six concernant la chirurgie cardiaque (@num@ participants) et sept la chirurgie thoracique (@num@ participants) .
the@85-0 evidence@85-1 is **current@87-0** to May @num@ .	les preuves sont à jour jusqu ' à mai @num@ .	**la preuve** est **en cours** jusqu ' à mai @num@ .	la preuve date du mois de mai @num@ .
no@89-0 study@89-1 evaluated@90-0 fondaparinux , the new oral direct thrombin or direct factor Xa inhibitors , or **caval@63-0 filters@63-1** .	aucune étude n ' a évalué le fondaparinux , les nouveaux directs par voie orale de la thrombine ou des inhibiteurs directs du facteur Xa , ou les filtres cave .	**aucune étude** a **évalué** le fondaparinux , les nouveaux directs par voie orale de la thrombine ou des inhibiteurs directs du facteur Xa , ou cave filtre .	aucune étude n' a évalué le fondaparinux , le nouvel inhibiteur oral direct de la thrombine , l' inhibiteur direct du facteur Xa , ou encore les filtres caves .
data could not be combined because of the different comparisons and the lack of data .	les données n' ont pas pu être combinés en raison des différentes comparaisons et le manque de données .	les données n' ont pas pu être combinées en raison des différentes comparaisons et le manque de données .	les données n' ont pas pu être combinées en raison de la diversité des comparaisons et de leur nombre insuffisant .

marked src.	initial MT	final MT	PE
data for **clinically@101-0 relevant@101-1 outcomes@101-2** such as pulmonary embolism (**blockage@102-0** of one or more arteries of the lung) or **major@105-0 bleeding@105-1** were often lacking .	les données pour les critères de jugement cliniquement pertinents , tels que l' embolie pulmonaire (blocage d' un ou plusieurs artères du poumon) ou les saignements majeurs étaient souvent manquantes .	les données de **résultats cliniquement pertinents** tels que l' embolie pulmonaire (**obstruction** d' une ou plusieurs artères du poumon) ou d' **hémorragie grave** étaient souvent manquantes .	les données des résultats cliniquement pertinents tels que l' embolie pulmonaire (obstruction d' une ou plusieurs artères pulmonaires) ou l' hémorragie grave étaient souvent manquantes .
in cardiac surgery , **symptomatic@107-0 venous@107-1 thromboembolism@107-2 occurred@108-0** in @num@ **out@110-0 of@110-1** @num@ participants from three studies .	dans la chirurgie cardiaque , la thromboembolie veineuse symptomatique a été observée chez @num@ des @num@ participants issus de trois études .	dans la chirurgie cardiaque , de la **maladie thromboembolique veineuse symptomatique est apparu** chez @num@ **sur** @num@ participants issus de trois études .	en chirurgie cardiaque , la maladie thromboembolique veineuse symptomatique est survenue chez @num@ participants sur @num@ dans trois études différentes .

marked src.	initial MT	final MT	PE
in a study of @num@ participants , representing @num@ % of the review population in cardiac surgery , the combination of **unfractionated@117-0 heparin@117-1** with intermittent pneumatic compression was associated with an important reduction of **symptomatic@107-0 venous@107-1 thromboembolism@107-2** compared to **unfractionated@117-0 heparin@117-1** alone .	dans une étude portant sur @num@ participants , représentant @num@ % de la revue de la population dans la chirurgie cardiaque , la combinaison de l' héparine non fractionnée à la compression pneumatique intermittente a été associée à une réduction importante de la thromboembolie veineuse symptomatique par rapport à l' héparine non fractionnée seule .	dans une étude portant sur @num@ participants , représentant @num@ % de la revue de la population dans la chirurgie cardiaque , la combinaison de l' **héparine non fractionnée** à la compression pneumatique intermittente a été associée à une réduction importante de **maladie thromboembolique veineuse symptomatique** par rapport à l' **héparine non fractionnée** seule .	dans une étude portant sur @num@ participants , représentant @num@ % de la population sondée en chirurgie cardiaque , l' association de l' héparine non fractionnée à la compression pneumatique intermittente a permis de réduire considérablement la maladie thromboembolique veineuse symptomatique , contrairement à l' héparine non fractionnée seule .

marked src.	initial MT	final MT	PE
major (important) bleeding **was@126-0 reported@126-1** in one study **only@128-0** , and the best estimate was that bleedings **occurred@108-0 seven@134-0 times@134-1** more often in participants on vitamin K antagonists compared to participants on platelet inhibitors , but the true estimate **may@141-0 lay@141-1** between **one@143-0** and a half to @num@ .	majeures (saignements importants) ont été signalés dans une seule étude , et la meilleure estimation était que des hémorragies sont survenus sept fois plus souvent chez les participants sur les antagonistes de la vitamine K par rapport aux participants sur des inhibiteurs plaquettaires , mais les véritables estimation peut reposer entre un et demi à @num@ .	majeures (saignements importants) **a été rapporté** dans **une** étude **seule** , et la meilleure estimation était que des hémorragies **est apparu sept fois** plus souvent chez les participants sur les antagonistes de la vitamine K par rapport aux participants sur des inhibiteurs plaquettaires , mais les véritables estimation **peut se trouver** entre **une** et demi à @num@ .	une étude a mis en évidence une hémorragie grave et a estimé que le risque d' hémorragie était sept fois plus présent chez les participants ayant recours aux antagonistes de la vitamine K que chez les participants sous inhibiteurs plaquettaires . cependant , la véritable estimation peut se trouver entre @num@ et @num@ .
in thoracic surgery , **symptomatic@107-0 venous@107-1 thromboembolism@107-2 occurred@108-0** in @num@ **out@110-0 of@110-1** @num@ participants from six studies .	dans la chirurgie thoracique , la thromboembolie veineuse symptomatique a été observée chez @num@ des @num@ participants provenant de six études .	dans la chirurgie thoracique , la **maladie thromboembolique veineuse symptomatique est apparu** chez @num@ **sur** @num@ participants de six études .	en chirurgie thoracique , la maladie thromboembolique veineuse symptomatique est apparue chez @num@ participants sur @num@ dans six études différentes .

marked src.	initial MT	final MT	PE
combined@146-0 analysis@146-1 could not be performed , but **the@148-0 largest@148-1 study@148-2 evaluating@149-0 unfractionated@117-0 heparin@117-1** versus **an@151-0 inactive@151-1 control@151-2** did not show **a@153-0 benefit@153-1** in terms of **reduced@155-0 occurrence@155-1** of **symptomatic@107-0 venous@107-1 thromboembolism@107-2** .	une analyse combinée n' ont pas pu être réalisées , mais la plus grande étude évaluant l' héparine non fractionnée versus un contrôle inactif n ' a pas montré un bénéfice en termes de réduction de l' incidence de la thromboembolie veineuse symptomatique .	**des analyses combinées** n' ont pas pu être réalisées , mais **l' étude la plus vaste évaluant héparine non fractionnée** versus **contrôle inactif** n ' a pas montré **une amélioration** en termes de **réduction du nombre de** de **maladie thromboembolique veineuse symptomatique** .	des analyses combinées n' ont pas pu être réalisées . cependant , l' étude la plus vaste évaluant l' efficacité de l' héparine non fractionnée contre le contrôle inactif n' a pas montré une amélioration en termes de réduction de l' apparition de la maladie thromboembolique veineuse symptomatique .

marked src.	initial MT	final MT	PE
major@105-0 bleeding@105-1 was@126-0 reported@126-1 in two studies that **did@157-0 not@157-1 find@157-2 significantly@158-0 different@158-1 rates@158-2** between fixed-dose and weight-adjusted low molecular weight heparin (@num@ % versus @num@ %) and between **unfractionated@117-0 heparin@117-1** and low molecular weight heparin (@num@ % and @num@ %) .	l' hémorragie majeure a été rapporté dans deux études qui n' avaient pas trouvé de différence significative entre les taux pondéré à dose fixe et l' héparine de bas poids moléculaire (@num@ % versus @num@ %) et entre l' héparine non fractionnée et l' héparine de bas poids moléculaire (@num@ % et @num@ %) .	**hémorragie grave a été rapporté** dans deux études qui **n' ont pas trouvé des taux sensiblement différents** entre à dose fixe et pondéré de l' héparine de bas poids moléculaire (@num@ % contre @num@ %) et entre l' **héparine non fractionnée** et l' héparine de bas poids moléculaire (@num@ % et @num@ %) .	deux études ont mis en évidence une hémorragie grave mais n' ont pas trouvé de taux sensiblement différents entre l' héparine administrée à dose fixe et l' héparine de bas poids moléculaire ajustée en fonction du poids du patient (@num@ % contre @num@ %) , et entre l' héparine non fractionnée et l' héparine de bas poids moléculaire (@num@ % et @num@ %) .
quality of **the@85-0 evidence@85-1** .	la qualité des preuves .	la qualité de **la preuve** .	qualité de la preuve
overall , **the@85-0 evidence@85-1** on the use of thromboprophylaxis in cardiac and thoracic surgery **appeared@165-0 to@165-1 be@165-2** scarce , so we **are@168-0 very@169-0 uncertain@169-1 about@170-0** the benefit-to-risk balance .	dans l' ensemble , les preuves sur l' utilisation de la thromboprophylaxie dans la chirurgie cardiaque et thoracique semblait être rares , de sorte que nous avons sont très incertains quant à l' équilibre benefit-to-risk .	dans l' ensemble , **la preuve** sur l' utilisation **de** la thromboprophylaxie dans la chirurgie cardiaque et thoracique **s' est avéré être** rares , **de** sorte que nous **sommes très peu certains de** l' équilibre benefit-to-risk .	dans l' ensemble , la preuve sur l' utilisation de la thromboprophylaxie en chirurgies cardiaque et thoracique s' est avérée être rare , c' est pourquoi nous ne sommes pas sûrs du rapport bénéfices / risques .

marked src.	initial MT	final MT	PE
all studies **had@173-0** major study design flaws and most **lacked@176-0** a placebo or **no@178-0 treatment@178-1 control@178-2 group@178-3** .	toutes les études présentaient des défauts majeures , le plan d' étude et la plupart manquions à un placebo ou à l' absence de traitement , à un groupe témoin .	toutes les études **ont eu** majeures , le plan d' étude , la plupart **sont dénuées de** défauts et un placebo ou un **groupe témoin ne recevant aucun traitement** .	toutes les études ont présenté des défauts majeurs tandis que la plupart d' entre elles n' ont pas eu recours à des placebos et n' ont pas fait intervenir un groupe expérimental ainsi qu' un groupe témoin .
we **typically@179-0 graded@180-0** the quality of the overall body of **evidence@183-0** for **the@184-0 various@184-1 outcomes@184-2 and@184-3 comparisons@184-4 as@185-0 low@185-1** , due **to@187-0 imprecise@187-1** estimates of effect and **risk@190-0** of bias .	nous avons typiquement classé la qualité de l' ensemble des données disponibles pour les différentes comparaisons , et les critères de jugement comme faible en raison des estimations imprécises de l' effet et le risque de biais .	nous **avons évalué** la qualité **généralement** de l' ensemble de **preuve** pour les différents résultats et des comparaisons **aussi bas** , en raison **d' imprécises** des estimations de l' effet et le **risque** de biais .	nous avons évalué la qualité de l' ensemble des preuves pour les différents résultats et comparaisons comme étant faible , en raison du manque de précision des estimations de l' effet et du risque de biais .

marked src.	initial MT	final MT	PE
our data suggest that thromboprophylaxis can not be suggested for all patients undergoing **these@196-0 types@196-1** of surgery , but **should@197-0 rather@197-1 be@197-2 considered@197-3 case-by-case@198-0** based on the individual risk of venous thromboembolism and bleeding .	nos résultats suggèrent que la thromboprophylaxie ne peut pas être suggéré pour tous les patients subissant une de ces types de chirurgie , mais devrait être pris en compte plutôt soient basés sur le risque individuel de la thromboembolie veineuse et les saignements .	nos résultats suggèrent que la thromboprophylaxie ne peut pas être suggéré pour tous les patients subissant une **ces genres** de la chirurgie , mais devraient être étudiées **cas par cas** plutôt basés sur le risque individuel de la thromboembolie veineuse et les saignements .	nos résultats suggèrent que la thromboprophylaxie ne peut pas être recommandée à tous les patients subissant une de ces deux chirurgies . il convient de faire du cas par cas en s' appuyant sur les risques de MTEV et d' hémorragie de chaque individu .
ABSTRACT BACKGROUND			
cardiac and thoracic surgery are associated with an increased risk of **venous@2-0 thromboembolism@2-1** (**VTE@204-0**) .	la chirurgie cardiaque et thoracique sont associés à un risque accru de thromboembolie veineuse (TEV) .	la chirurgie cardiaque et thoracique sont associés à un risque accru de **maladie thrombo-embolique veineuse** (**MTE**) .	en chirurgies cardiaque et thoracique , le risque d' apparition de la maladie thrombo-embolique veineuse (MTEV) est accru .
the safety and efficacy of **primary@206-0 thromboprophylaxis@206-1** in patients undergoing **these@196-0 types@196-1** of surgery is uncertain .	l' efficacité et l' innocuité de la thromboprophylaxie primaire chez les patients subissant une de ces types de chirurgie est incertaine .	l' innocuité et l' efficacité de la **thromboprophylaxie primaire** chez les patients subissant une **ces genres** de chirurgie est incertaine .	l' innocuité et l' efficacité de la thromboprophylaxie primaire chez les patients subissant une de ces deux chirurgies sont incertaines .
ABSTRACT OBJECTIVES			

marked src.	initial MT	final MT	PE
to@208-0 assess@208-1 the effects of **primary@206-0 thromboprophylaxis@206-1** on the incidence of **symptomatic@211-0 VTE@211-1** and **major@105-0 bleeding@105-1** in patients undergoing cardiac or thoracic surgery .	évaluer les effets de la thromboprophylaxie primaire sur l' incidence de la TEV symptomatique et les saignements majeurs chez les patients subissant une chirurgie cardiaque ou thoracique .	**pour évaluer** les effets de la **thromboprophylaxie primaire** sur l' incidence de **MTE symptomatique** et d' **hémorragie grave** chez les patients subissant une chirurgie cardiaque ou thoracique .	pour évaluer les effets de la thromboprophylaxie primaire sur l' incidence de la MTEV symptomatique et de l' hémorragie grave chez les patients subissant une chirurgie cardiaque ou thoracique .
ABSTRACT SEARCH STRATEGY			
the@212-0 Cochrane@212-1 Peripheral@212-2 Vascular@212-3 Diseases@212-4 Group@212-5 Trials@212-6 Search@212-7 Co-ordinator@212-8 searched the Specialised Register (last searched May @num@) and **central@216-0 (@216-1 @num@@216-2 ,@216-3 Issue@216-4 @num@@216-5**) .	le groupe Cochrane sur les maladies vasculaires périphériques a effectué des recherches dans le coordinateur des recherches dans le registre spécialisé (dernière recherche en mai @num@) et central (@num@ , numéro @num@) .	pour Cochrane des groupes de maladie vasculaire périphérique le coordinateur de recherche des essais effectué des recherches dans le registre spécialisé (dernière recherche en mai @num@) et **le CENTRAL (@num@ , numéro @num@)** .	le Coordinateur de la recherche clinique du groupe Cochrane sur les maladies vasculaires périphériques a effectué des recherches dans son Registre spécialisé (dernière recherche en mai @num@) et dans le registre central Cochrane des essais contrôlés nommé central (@num@ , numéro @num@) .

marked src.	initial MT	final MT	PE
the@217-0 authors@217-1 searched the reference lists of **relevant@219-0 studies@219-1** , **conference@220-0 proceedings@220-1** , and clinical trial registries .	les auteurs ont effectué des recherches dans les listes bibliographiques des études pertinentes , des actes de conférence , et les registres d' essais cliniques .	**les auteurs** ont effectué des recherches dans les listes bibliographiques des **études pertinentes** , des **comptes rendus de conférences** et les registres d' essais cliniques .	les auteurs ont effectué des recherches dans les listes bibliographiques des études pertinentes , des comptes rendus de conférences et des registres d' essais cliniques .
ABSTRACT SELECTION CRITERIA			
randomised controlled trials (**RCTs@223-0**) and **quasi-RCTs@224-0** comparing any oral or parenteral anticoagulant or mechanical intervention to no intervention or placebo , or comparing two different anticoagulants .	les essais contrôlés randomisés (ECR) et quasi-ECR comparant un anticoagulant par voie orale ou parentérale intervention mécanique ou à l' absence d' intervention ou à un placebo , ou comparant deux anticoagulants différents .	les essais contrôlés randomisés (**ERC**) et **quasi-ECR** comparant un anticoagulant par voie orale ou parentérale intervention mécanique ou à l' absence d' intervention ou à un placebo , ou comparant deux anticoagulants différents .	les essais randomisés contrôlés (ERC) ainsi que les essais quasi-randomisés comparent soit un anticoagulant administré par voie orale ou parentérale ou une intervention mécanique à l' absence d' intervention ou à un placebo , soit deux anticoagulants différents .
ABSTRACT DATA COLLECTION			

marked src.	initial MT	final MT	PE
we extracted data on **methodological@231-0 quality@231-1** , **participant@232-0 characteristics@232-1** , interventions , and **outcomes@234-0 including@235-0 symptomatic@211-0 VTE@211-1** and **major@105-0 bleeding@105-1 as@236-0 the@237-0 primary@237-1 effectiveness@237-2 and@237-3 safety@237-4 outcomes@237-5** , respectively .	nous avons extrait les données sur la qualité méthodologique , les caractéristiques des participants , les interventions et les résultats , y compris la TEV symptomatique et les saignements majeurs en tant que critères principaux d' efficacité et d' innocuité , respectivement .	nous avons extrait les données sur la **qualité méthodologique** , les **critères du participant** , les interventions et les **résultats dont MTE symptomatique** et d' **hémorragie grave comme le résultat principal d' efficacité et de sécurité** , respectivement .	nous avons extrait les données sur la qualité méthodologique , les critères des participants , les interventions et les résultats tels que la MTEV symptomatique et l' hémorragie grave qui représentent les principaux résultats en termes d' efficacité et d' innocuité .
ABSTRACT RESULTS			
we identified **@num@@239-0 RCTs@239-1** and **one@240-0 quasi-RCT@240-1** (@num@ participants) , six for cardiac surgery (@num@ participants) and seven for thoracic surgery (@num@ participants) .	nous avons identifié @num@ ECR et un quasi-ECR (@num@ participants) , six dans la chirurgie cardiaque (@num@ participants) et sept pour la chirurgie thoracique (@num@ participants) .	nous avons identifié **@num@ ERC** et **un quasi-ECR** (@num@ participants) , six dans la chirurgie cardiaque (@num@ participants) et sept pour la chirurgie thoracique (@num@ participants) .	nous avons identifié @num@ ERC et un essai quasi-randomisé (@num@ participants) , six en chirurgie cardiaque (@num@ participants) et sept en chirurgie thoracique (@num@ participants) .

marked src.	initial MT	final MT	PE
no@89-0 study@89-1 evaluated@90-0 fondaparinux , the new oral direct thrombin , direct factor Xa inhibitors , or **caval@63-0 filters@63-1** .	aucune étude n ' a évalué le fondaparinux , les nouveaux directs par voie orale de la thrombine , inhibiteurs directs du facteur Xa , ou les filtres cave .	**aucune étude** a **évalué** le fondaparinux , les nouveaux directs par voie orale de la thrombine , inhibiteurs directs du facteur Xa , ou cave filtre .	aucune étude n' a évalué le fondaparinux , le nouvel inhibiteur oral direct de la thrombine , l' inhibiteur direct du facteur Xa , ou encore les filtres caves .
all studies **had@173-0** major study design flaws and most **lacked@176-0** a placebo or **no@178-0 treatment@178-1 control@178-2 group@178-3** .	toutes les études présentaient des défauts majeures , le plan d' étude et la plupart manquions à un placebo ou à l' absence de traitement , à un groupe témoin .	toutes les études **ont eu** majeures , le plan d' étude , la plupart **sont dénuées de** défauts et un placebo ou un **groupe témoin ne recevant aucun traitement** .	toutes les études ont présenté des défauts majeurs tandis que la plupart d' entre elles n' ont pas eu recours à des placebos et n' ont pas fait intervenir un groupe expérimental ainsi qu' un groupe témoin .
we **typically@179-0 graded@180-0** the quality of the overall body of **evidence@183-0** for **the@184-0 various@184-1 outcomes@184-2 and@184-3 comparisons@184-4 as@185-0 low@185-1** , due **to@187-0 imprecise@187-1** estimates of effect and **risk@190-0** of bias .	nous avons typiquement classé la qualité de l' ensemble des données disponibles pour les différentes comparaisons , et les critères de jugement comme faible en raison des estimations imprécises de l' effet et le risque de biais .	nous **avons évalué** la qualité **généralement** de l' ensemble de **preuve** pour les différents résultats et des comparaisons **aussi bas** , en raison **d' imprécises** des estimations de l' effet et le **risque** de biais .	nous avons évalué la qualité de l' ensemble des preuves pour les différents résultats et comparaisons comme étant faible , en raison du manque de précision des estimations de l' effet et du risque de biais .

marked src.	initial MT	final MT	PE
we could not pool data because of the different comparisons and the lack of data .	nous n' avons pas pu regrouper les données en raison des différentes comparaisons et le manque de données .	nous n' avons pas pu regrouper les données en raison des différentes comparaisons et le manque de données .	nous n' avons pas pu regrouper les données en raison de la diversité des comparaisons et de leur nombre insuffisant .
in cardiac surgery , **@num@@242-0 symptomatic@242-1 VTEs@242-2 occurred@108-0** in @num@ participants from four studies .	dans la chirurgie cardiaque , @num@ symptomatique ETV sont survenus chez @num@ participants provenant de quatre études .	dans la chirurgie cardiaque , **@num@ MTE symptomatiques est apparu** chez @num@ participants provenant de quatre études .	en chirurgie cardiaque , la MTEV symptomatique est apparue chez @num@ participants sur @num@ dans quatre études différentes .

marked src.	initial MT	final MT	PE
in a study of @num@ participants , representing @num@ % of the review population in cardiac surgery , the combination of **unfractionated@117-0 heparin@117-1** with pneumatic compression stockings was associated with a @num@ % reduction of **symptomatic@211-0 VTE@211-1** compared to **unfractionated@117-0 heparin@117-1** alone (@num@ % versus @num@ % ; **risk@247-0 ratio@247-1** (RR) @num@ ; @num@ % confidence interval (CI) @num@ to @num@) .	dans une étude portant sur @num@ participants , représentant @num@ % de la revue de la population dans la chirurgie cardiaque , la combinaison de l' héparine non fractionnée à bas de compression pneumatique était associée à une réduction de @num@ % des TEV symptomatiques par rapport à l' héparine non fractionnée seule (@num@ % contre @num@ % ; risque relatif (RR) @num@ ; intervalle de confiance à @num@ % (IC) @num@ à @num@) .	dans une étude portant sur @num@ participants , représentant @num@ % de la revue de la population dans la chirurgie cardiaque , la combinaison de l' **héparine non fractionnée** à bas de compression pneumatique était associée à une réduction de @num@ % de **MTE symptomatique** par rapport à l' **héparine non fractionnée** seule (@num@ % contre @num@ % ; **risque relatif** (RR) @num@ ; intervalle de confiance à @num@ % (IC) @num@ à @num@) .	dans une étude portant sur @num@ participants , représentant @num@ % de la population sondée en chirurgie cardiaque , l' association de l' héparine non fractionnée à des bas de compression pneumatique a permis de réduire de @num@ % la MTEV symptomatique , contrairement à l' héparine non fractionnée seule (@num@ % contre @num@ % ; risque relatif (RR) @num@ ; intervalle de confiance à @num@ % (IC) @num@ à @num@) .

marked src.	initial MT	final MT	PE
major@105-0 bleeding@105-1 was@251-0 only@251-1 reported@251-2 in one study , which **found@253-0 a@254-0 higher@254-1 incidence@254-2** with vitamin K antagonists compared to platelet inhibitors (@num@ % versus @num@ % , **RR@255-0 @num@@255-1** ; @num@ % CI @num@ to @num@) .	l' hémorragie majeure n ' a été rapporté que dans une étude , qui a mis en évidence une incidence plus élevée avec des antagonistes de la vitamine K par rapport aux inhibiteurs plaquettaires (@num@ % contre @num@ % , RR @num@ ; IC à @num@ % @num@ à @num@) .	**hémorragie grave n' a été rapporté** dans une étude , qui **a trouvé une incidence plus élevée** avec des antagonistes de la vitamine K par rapport aux inhibiteurs plaquettaires (@num@ % contre @num@ % , **RR @num@** ; IC à @num@ % @num@ à @num@) .	une étude a mis en évidence une hémorragie grave et a trouvé que les traitements par antagonistes de la vitamine K avaient une incidence plus élevée que les inhibiteurs plaquettaires (@num@ % contre @num@ % , RR @num@ ; IC à @num@ % @num@ à @num@) .
in thoracic surgery , **@num@@242-0 symptomatic@242-1 VTEs@242-2 occurred@108-0** in @num@ participants from six studies .	dans la chirurgie thoracique , @num@ symptomatique ETV sont survenus chez @num@ participants provenant de six études .	dans la chirurgie thoracique , **@num@ MTE symptomatiques est apparu** chez @num@ participants provenant de six études .	en chirurgie thoracique , la MTEV symptomatique est apparue chez @num@ participants sur @num@ dans six études différentes .

marked src.	initial MT	final MT	PE
in **the@148-0 largest@148-1 study@148-2 evaluating@149-0 unfractionated@117-0 heparin@117-1** versus an inactive control the rates of **symptomatic@211-0 VTE@211-1** were @num@ % versus @num@ % , respectively , giving **a@260-0 RR@260-1** of @num@ (@num@ % CI @num@ to @num@) .	dans la plus grande étude évaluant l' héparine non fractionnée versus un contrôle inactif , le taux de TEV symptomatique étaient de @num@ % versus @num@ % , respectivement , donnant un RR de @num@ (IC à @num@ % @num@ à @num@) .	dans **l' étude la plus vaste évaluant héparine non fractionnée** versus un contrôle inactif , le taux de **MTE symptomatique** étaient de @num@ % versus @num@ % , respectivement , donnant **un RR** de @num@ (IC à @num@ % de @num@ à @num@) .	dans l' étude la plus vaste évaluant une héparine non fractionnée versus un contrôle inactif , le taux de MTEV symptomatique était de @num@ % versus @num@ % , respectivement , donnant un RR de @num@ (IC à @num@ % de @num@ à @num@) .

marked src.	initial MT	final MT	PE
there@261-0 was insufficient evidence **to@263-0 determine@263-1 if@264-0 there@261-0** was a difference in the risk of **major@105-0 bleeding@105-1** from two studies **evaluating@149-0 fixed-dose@267-0 versus@267-1 weight-adjusted@267-2 low@267-3 molecular@267-4 weight@267-5 heparin@267-6** (@num@ % versus @num@ % , **RR@255-0 @num@@255-1** ; @num@ % CI @num@ to @num@) and **unfractionated@117-0 heparin@117-1** versus low molecular weight heparin (@num@ % and @num@ % , **RR@255-0 @num@@255-1** ; @num@ % CI @num@ to @num@) .	il n ' y avait pas suffisamment de preuves pour déterminer s ' il existait une différence dans le risque d' hémorragie majeure de deux études évaluant à dose fixe pondéré par rapport à l' héparine de bas poids moléculaire (@num@ % contre @num@ % , RR @num@ ; IC à @num@ % de @num@ à @num@) et de l' héparine non fractionnée versus héparine de bas poids moléculaire (@num@ % et @num@ % ; RR @num@ ; IC à @num@ % @num@ à @num@) .	**il** n ' y avait pas suffisamment de preuves **pour déterminer si il** y avait une différence dans le risque d' **hémorragie grave** à partir de deux études **évaluant** une dose fixe contre une dose adaptée au poids et faible en poids moléculaire d' héparine (@num@ % contre @num@ % , **RR @num@** ; IC à @num@ % de @num@ à @num@) et de l' **héparine non fractionnée** versus héparine de bas poids moléculaire (@num@ % et @num@ % ; **RR @num@** ; IC à @num@ % @num@ à @num@) .	il n' y avait pas suffisamment de preuves pour déterminer si le risque d' hémorragie grave était différent à partir de deux études évaluant respectivement l' héparine administrée à dose fixe contre l' héparine de bas poids moléculaire ajustée en fonction du poids du patient (@num@ % contre @num@ % , RR @num@ ; IC à @num@ % de @num@ à @num@) , et l' héparine non fractionnée contre l' héparine de bas poids moléculaire (@num@ % et @num@ % ; RR @num@ ; IC à @num@ % @num@ à @num@) .
ABSTRACT CONCLUSIONS			

marked src.	initial MT	final MT	PE
the@85-0 evidence@85-1 regarding the efficacy and safety of thromboprophylaxis in cardiac and thoracic surgery is limited .	les preuves concernant l' efficacité et l' innocuité de la thromboprophylaxie dans la chirurgie cardiaque et thoracique est limitée .	**la preuve** concernant l' efficacité et l' innocuité de la thromboprophylaxie dans la chirurgie cardiaque et thoracique est limitée .	la preuve concernant l' efficacité et l' innocuité de la thromboprophylaxie en chirurgies cardiaque et thoracique est limitée .
data for **important@271-0 outcomes@271-1** such as pulmonary embolism or **major@105-0 bleeding@105-1** were often lacking .	les données pour les critères de jugement importants tels que l' embolie pulmonaire ou les saignements majeurs étaient souvent manquantes .	les données de **résultats importants** , tels que l' embolie pulmonaire ou d' **hémorragie grave** étaient souvent manquantes .	les données de résultats importants , tels que l' embolie pulmonaire ou l' hémorragie grave , étaient souvent manquantes .
given the uncertainties around the benefit-to-risk balance , no conclusions can be drawn and a case-by-case risk evaluation of **VTE@278-0 and@278-1 bleeding@278-2** remains preferable .	étant donné les incertitudes autour de l' équilibre benefit-to-risk , aucune conclusion ne peut être tirée et une évaluation , au risque de TEV et les saignements reste préférable .	étant donné les incertitudes autour de l' équilibre benefit-to-risk , aucune conclusion ne peut être tirée et une évaluation des risques soient **MTE et de saignement** reste préférable .	étant donné les incertitudes qui persistent autour du rapport bénéfices / risques , aucune conclusion ne peut être tirée et il s' avère préférable de faire du cas par cas pour évaluer les risques engendrés par la MTEV et l' hémorragie .

I | Publications by the Author

Peer-reviewed conference proceedings

Ive, J. and Yvon, F. Parallel sentence compression. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1503–1513, 2016. URL `http://aclweb.org/anthology/C16-1142`. (Not cited)

Peer-reviewed workshop proceedings

Ive, J., Max, A., and Yvon, F. LIMSI's contribution to the WMT'16 biomedical translation task, 2016a. URL `http://www.aclweb.org/anthology/W16-2337`. (Not cited)

Ive, J., Max, A., Yvon, F., and Ravaud, P. Diagnosing high-quality statistical machine translation using traces of post-edition operations, 2016b. URL `http://www.cracking-the-language-barrier.eu/wp-content/uploads/LREC-2016-MT-Eval-Workshop-Proceedings.pdf`. (Not cited)

Marie, B., Allauzen, A., Burlot, F., Do, Q.-K., **Ive**, J., Knyazeva, E., Labeau, M., Lavergne, T., Löser, K., Pécheux, N., and Yvon, F. LIMSI@WMT'15 : Translation task, 2015. URL `http://aclweb.org/anthology/W15-3016`. (Not cited)

Pécheux, N., Gong, L., Do, Q. K., Marie, B., **Ivanishcheva (Ive)**, Y., Allauzen, A., Lavergne, T., Niehues, J., Max, A., and Yvon, F. LIMSI@WMT'14 medical translation task, 2014. URL `http://www.aclweb.org/anthology/W14-3330`. (Not cited)

Névéol, A., Max, A., **Ivanishcheva (Ive)**, Y., Ravaud, P., Zweigenbaum, P., and Yvon, F. Statistical machine translation of systematic reviews into French, 2013. URL `https://perso.limsi.fr/yvon/publications/sources/Neveol13statistical.pdf`. (Not cited)

Presentations

Martikainen, H. and **Ive**, J. Les corpus pour améliorer la qualité de la traduction automatique statistique post-éditée dans un domaine de spécialité. 2016. URL `https://traduction2016.sciencesconf.org`. (Not cited)

Bibliography

Abdelsalam, A., Bojar, O., and El-Beltagy, S. Bilingual embeddings and word alignments for translation quality estimation. In *Proceedings of the First Conference on Machine Translation*, pages 764–771, 2016. URL `http://aclweb.org/anthology/W/W16/W16-2380.pdf`. (Cited on page 38)

Abend, O. and Rappoport, A. Universal conceptual cognitive annotation (UCCA). In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 228–238, 2013. URL `http://www.aclweb.org/anthology/P13-1023`. (Cited on page 59)

Aikawa, T., Schwartz, L., King, R., Corston-Oliver, M., and Lozano, C. Impact of controlled language on translation quality and post-editing in a statistical machine translation environment. In *Proceedings of the 11th Machine Translation Summit (MT Summit XIII)*, pages 1–7, 2007. URL `http://www.mt-archive.info/05/MTS-2007-Aikawa.pdf`. (Cited on page 36)

Alabau, V. and Leiva, L. A. Proofreading human translations with an e-pen. In *Proceedings of the EACL 2014 Workshop on Humans and Computer-assisted Translation*, pages 10–15, 2014. URL `http://www.aclweb.org/anthology/W14-0302`. (Cited on page 52)

Alabau, V., González-Rubio, J., Leiva, L. A., Ortiz-Martínez, D., Sanchis-Trilles, G., Casacuberta, F., Mesa-Lao, B., Bonk, R., Carl, M., and Garcia Martinez, M. User evaluation of advanced interaction features for a computer-assisted translation workbench. In *Machine Translation Summit XIV*, pages 361–368, 2013. URL `http://www.mt-archive.info/10/MTS-2013-Alabau.pdf`. (Cited on page 38)

Arcan, M., Giuliano, C., Turchi, M., and Buitelaar, P. Identification of bilingual terms from monolingual documents for statistical machine translation. In *Proceedings of the 4th Inter-*

national Workshop on Computational Terminology (Computerm), pages 22–31, 2014. URL http://www.aclweb.org/anthology/W14-4803. (Cited on page 37)

Aronson, A. R. and Lang, F.-M. An overview of MetaMap: historical perspective and recent advances. *Journal of the American Medical Informatics Association : JAMIA*, 17(3):229–236, 2010. URL https://www.ncbi.nlm.nih.gov/pubmed/20442139. (Cited on page 75)

Axelrod, A., He, X., and Gao, J. Domain adaptation via pseudo in-domain data selection. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 355–362, 2011. URL http://www.aclweb.org/anthology/D11-1033. (Cited on page 51)

Bach, N., Huang, F., and Al-Onaizan, Y. Goodness: A method for measuring machine translation confidence. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 211–219, 2011. URL http://www.aclweb.org/anthology/P11-1022. (Cited on page 38)

Bahdanau, D., Cho, K., and Bengio, Y. Neural machine translation by jointly learning to align and translate. In *ICLR*, 2015. URL http://arxiv.org/pdf/1409.0473v6.pdf. (Cited on pages 10 and 17)

Banerjee, P., Naskar, S. K., Roturier, J., Way, A., and van Genabith, J. Domain adaptation in SMT of user-generated forum content guided by OOV word reduction: Normalization and/or supplementary data? In *Proceedings of 16th Annual Conference of the European Association for Machine Translation (EAMT)*, 2012. URL http://www.computing.dcu.ie/~away/PUBS/2012/40.pdf. (Cited on page 50)

Bannard, C. and Callison-Burch, C. Paraphrasing with bilingual parallel corpora. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 597–604, 2005. URL http://www.aclweb.org/anthology/P05-1074. (Cited on page 71)

Bar-Hillel, Y. The present state of research on mechanical translation. *American Documentation 2*, pages 229–237, 1951. URL http://www.mt-archive.info/Bar-Hillel-1951.pdf. (Cited on pages 33 and 34)

Barrachina, S., Bender, O., Casacuberta, F., Civera, J., Cubel, E., Khadivi, S., Lagarda, A., Ney, H., Tomás, J., Vidal, E., and Vilar, J.-M. Statistical approaches to computer-assisted translation. *Computational Linguistics*, 35(1), 2008. URL http://dx.doi.org/10.1162/coli.2008.07-055-R2-06-29. (Cited on page 35)

Beigman Klebanov, B. and Flor, M. Associative texture is lost in translation. In *Proceedings of the Workshop on Discourse in Machine Translation*, pages 27–32, 2013. URL `http://www.aclweb.org/anthology/W13-3304`. (Cited on page 60)

Bender, O., Hasan, S., Vilar, D., Zens, R., and Nay, H. Comparison of generation strategies for interactive machine translation. In *Proceedings of 13th Annual Conference of the European Association for Machine Translation (EAMT)*, 2009. URL `http://www.mt-archive.info/EAMT-2005-Bender.pdf`. (Cited on page 40)

Bentivogli, L., Bisazza, A., Cettolo, M., and Federico, M. Neural versus phrase-based machine translation quality: a case study. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 257–267, 2016. URL `https://aclweb.org/anthology/D16-1025`. (Cited on page 31)

Berka, J., Bojar, O., Fishel, M., Popovic, M., and Zeman, D. Automatic MT error analysis: Hjerson helping Addicter. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pages 2158–2163, 2012. URL `http://www.lrec-conf.org/proceedings/lrec2012/pdf/336_Paper.pdf`. (Cited on page 59)

Bertoldi, N., Cettolo, M., and Federico, M. Cache-based online adaptation for machine translation enhanced computer assisted translation. In *Proceedings of the XIII Machine Translation Summit*, pages 117–124, 2013a. URL `https://www.matecat.com/wp-content/uploads/2013/09/mt-summit-2013-bertoldi-et-al.pdf`. (Cited on page 47)

Bertoldi, N., Cettolo, M., Negri, M., Turchi, M., Federico, M., and Schwenk, H. Second report on lab and field tests. Technical Report 5.4, Machine Translation Enhanced Computer Assisted Translation (MateCat) project deliverable, November 2013b. URL `https://www.matecat.com/wp-content/uploads/2014/06/D5.4_Second-Report-on-Lab-and-Field-Test_v2.pdf`. (Cited on pages 38 and 48)

Birch, A., Abend, O., Bojar, O., and Haddow, B. HUME: Human UCCA-based evaluation of machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1264–1274, 2016. URL `https://aclweb.org/anthology/D16-1134`. (Cited on pages 59 and 148)

Bisazza, A., Ruiz, N., and Federico, M. Fill-up versus interpolation methods for phrase-based SMT adaptation. In *Proceedings of International Workshop on Spoken Language Translation*

(IWSLT), 2011. URL http://www.mt-archive.info/IWSLT-2011-Bisazza.pdf. (Cited on page 47)

Blain, F., Schwenk, H., and Senellart, J. Incremental adaptation using translation information and post-editing analysis. In *Proceedings of International Workshop on Spoken Language Translation (IWSLT)*, pages 229–236, 2012. URL http://www.mt-archive.info/IWSLT-2012-Blain.pdf. (Cited on pages 45 and 47)

Blain, F., Bougares, F., Hazem, A., Barrault, L., and Schwenk, H. Continuous adaptation to user feedback for statistical machine translation. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1001–1005, 2015. URL http://www.aclweb.org/anthology/N15-1103. (Cited on page 50)

Blain, F., Scarton, C., and Specia, L. Bilexical embeddings for quality estimation. In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 545–550, 2017. URL http://www.aclweb.org/anthology/W17-4760. (Cited on page 38)

Blanchon, H. A solution for the problem of interactive disambiguation. In *Proceedings of the 15th International Conference on Computational Linguistics: COLING 1992 Volume 4*, 1992. URL http://www.aclweb.org/anthology/C92-4198. (Cited on page 113)

Blatz, J., Fitzgerald, E., Foster, G., Gandrabur, S., Goutte, C., Kulesza, A., Sanchis, A., and Ueffing, N. Confidence estimation for machine translation. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING 2004)*, 2004. URL http://www.alexkulesza.com/pubs/confest_coling04.pdf. (Cited on page 37)

Bloodgood, M. and Callison-Burch, C. Bucking the trend: Large-scale cost-focused active learning for statistical machine translation. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 854–864, 2010. URL http://www.aclweb.org/anthology/P10-1088. (Cited on page 49)

Bojar, O. Analyzing error types in English-Czech machine translation. In *The Prague Bulletin of Mathematical Linguistics*, April 2011. URL https://ufal.mff.cuni.cz/pbml/95/art-bojar.pdf. (Cited on page 58)

Bojar, O., Chatterjee, R., Federmann, C., Haddow, B., Huck, M., Hokamp, C., Koehn, P., Logacheva, V., Monz, C., Negri, M., Post, M., Scarton, C., Specia, L., and Turchi, M. Findings

of the 2015 workshop on statistical machine translation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 1–46, 2015. URL http://aclweb.org/anthology/W15-3001. (Cited on page 38)

Bojar, O., Chatterjee, R., Federmann, C., Graham, Y., Haddow, B., Huck, M., Jimeno Yepes, A., Koehn, P., Logacheva, V., Monz, C., Negri, M., Neveol, A., Neves, M., Popel, M., Post, M., Rubino, R., Scarton, C., Specia, L., Turchi, M., Verspoor, K., and Zampieri, M. Findings of the 2016 conference on machine translation. In *Proceedings of the First Conference on Machine Translation*, pages 131–198, 2016a. URL http://www.aclweb.org/anthology/W/W16/W16-2301. (Cited on pages 31, 38, and 90)

Bojar, O., Děchtěrenko, F., and Zelenina, M. A pilot eye-tracking study of WMT-style ranking evaluation. In *Translation Evaluation: From Fragmented Tools and Data Sets to an Integrated Ecosystem*, pages 20–26, 2016b. URL http://www.cracking-the-language-barrier.eu/wp-content/uploads/Bojar-Dechterenko-etal.pdf. (Cited on page 35)

Bojar, O., Federmann, C., Haddow, B., Koehn, P., Post, M., and Specia, L. 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*, 2016c. URL http://www.cracking-the-language-barrier.eu/wp-content/uploads/Bojar-Federmann-etal.pdf. (Cited on pages 29 and 38)

Bojar, O., Graham, Y., Kamran, A., and Stanojević, M. Results of the WMT16 metrics shared task. In *Proceedings of the First Conference on Machine Translation*, pages 199–231, 2016d. URL http://aclweb.org/anthology/W/W16/W16-2302.pdf. (Cited on pages 29 and 57)

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. Findings of the 2017 conference on machine translation (WMT17). In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 169–214, 2017a. URL http://www.aclweb.org/anthology/W17-4717. (Cited on pages 12 and 88)

Bojar, O., Graham, Y., and Kamran, A. Results of the WMT17 metrics shared task. In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 489–513, 2017b. URL http://www.aclweb.org/anthology/W17-4755. (Cited on page 29)

Brants, T., Popat, A. C., Xu, P., Och, F. J., and Dean, J. Large language models in machine translation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 858–867, 2007. URL http://www.aclweb.org/anthology/D/D07/D07-1090. (Cited on page 25)

Breiman, L. Random forests. *Machine Learning*, 45(1):5–32, 2001. URL http://dx.doi.org/10.1023/A:1010933404324. (Cited on pages 38, 62, and 91)

Brown, P. F., Cocke, J., Pietra, S. A. D., Pietra, V. J. D., Jelinek, F., Lafferty, J. D., Mercer, R. L., and Roossin, P. S. A statistical approach to machine translation. *Computational Linguistics*, 16(2), 1990. URL http://www.aclweb.org/anthology/J90-2002. (Cited on pages 10 and 17)

Brown, P. F., Pietra, V. J. D., Pietra, S. A. D., and Mercer, R. L. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311, 1993. URL http://dl.acm.org/citation.cfm?id=972470.972474. (Cited on pages 20 and 21)

Brown, R. D. Human-computer interaction for semantic disambiguation. In *Proceedings of 13th International Conference on Computational Linguistics (COLING 1990)*, pages 42–73, 1990. URL http://www.aclweb.org/anthology/C90-3008. (Cited on page 113)

Callison-Burch, C., Bannard, C., and Schroeder, J. Scaling phrase-based statistical machine translation to larger corpora and longer phrases. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 255–262, 2005. URL http://aclweb.org/anthology/P/P05/P05-1032.pdf. (Cited on pages 46 and 102)

Callison-Burch, C., Osborne, M., and Koehn, P. Re-evaluating the role of BLEU in machine translation research. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 249–256, 2006. URL http://www.aclweb.org/anthology/E06-1032. (Cited on page 28)

Cappé, O. and Moulines, E. On-line expectation-maximization algorithm for latent data models. *Journal of the Royal Statistical Society*, 71(3):593–613, 2009. URL https://arxiv.org/pdf/0712.4273.pdf. (Cited on page 44)

Carpuat, M. One translation per discourse. In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions (SEW-2009)*, pages 19–27, 2009. URL http://www.aclweb.org/anthology/W09-2404. (Cited on page 131)

Carpuat, M., Daumé III, H., Henry, K., Irvine, A., Jagarlamudi, J., and Rudinger, R. SenseSpotting: Never let your parallel data tie you to an old domain. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1435–1445, 2013. URL `http://www.aclweb.org/anthology/P13-1141`. (Cited on page 50)

Cattelan, A. Third version of MateCat tool. Technical Report 4.3, Machine Translation Enhanced Computer Assisted Translation (MateCat) project deliverable, October 2014. `http://cordis.europa.eu/docs/projects/cnect/8/287688/080/deliverables/001-MateCatD43V12Ares20143250929.pdf`. (Cited on page 52)

Cavallanti, G., Cesa-Bianchi, N., and Gentile, C. Linear algorithms for online multitask classification. *The Journal of Machine Learning Research*, 11:2901–2934, 2010. URL `http://dl.acm.org/citation.cfm?id=1756006.1953026`. (Cited on page 49)

Ceausu, A. and Hunsicker, S. Pre-ordering of phrase-based machine translation input in translation workflow. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3589–3592, 2014. URL `http://www.lrec-conf.org/proceedings/lrec2014/pdf/1213_Paper.pdf`. (Cited on page 36)

Cettolo, M., Servan, C., Bertoldi, N., Federico, M., Barrault, L., and Schwenk, H. Issues in incremental adaptation of statistical MT from human post-edits. In *Proceedings of MT Summit XIV Workshop on Post-editing Technology and Practice*, 2013. URL `https://hal.archives-ouvertes.fr/hal-01158054/file/Servan_WPTP2013.pdf`. (Cited on page 48)

Cettolo, M., Bertoldi, N., Federico, M., Schwenk, H., Barrault, L., and Servan, C. Translation project adaptation for MT-enhanced computer assisted translation. *Machine Translation*, 28 (2):127–150, October 2014. URL `https://hal.inria.fr/hal-01157893/document`. (Cited on page 50)

Chatterjee, R., Arcan, M., Negri, M., and Turchi, M. Instance selection for online automatic post-editing in a multi-domain scenario. In *Proceedings of Association for Machine Translation in the Americas*, 2016. URL `https://amtaweb.org/wp-content/uploads/2016/10/AMTA2016_Research_Proceedings_v7.pdf`. (Cited on page 51)

Chatterjee, R., Gebremelak, G., Negri, M., and Turchi, M. Online automatic post-editing for MT in a multi-domain translation environment. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages

525–535, 2017a. URL http://www.aclweb.org/anthology/E17-1050. (Cited on pages 11 and 51)

Chatterjee, R., Negri, M., Turchi, M., Federico, M., Specia, L., and Blain, F. Guiding neural machine translation decoding with external knowledge. In *Proceedings of the Second Conference on Machine Translation, Volume 1: Research Papers*, pages 157–168, 2017b. URL http://www.aclweb.org/anthology/W17-4716. (Cited on page 48)

Chen, B. and Cherry, C. A systematic comparison of smoothing techniques for sentence-level BLEU. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 362–367, 2014. URL http://www.aclweb.org/anthology/W14-3346. (Cited on page 43)

Chen, B., Kuhn, R., Foster, G., Cherry, C., and Huang, F. Bilingual methods for adaptive training data selection for machine translation. In *Proceedings of Association for Machine Translation in the Americas*, 2016. URL http://www-labs.iro.umontreal.ca/~foster/papers/bicnn-data-sel-amta16.pdf. (Cited on page 50)

Chen, D. and Manning, C. A fast and accurate dependency parser using neural networks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 740–750, 2014. URL http://www.aclweb.org/anthology/D14-1082. (Cited on pages 92 and 97)

Chen, S. F. and Goodman, J. An empirical study of smoothing techniques for language modeling. In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*, pages 310–318, 1996. URL http://www.aclweb.org/anthology/P96-1041. (Cited on page 25)

Cheng, S., Huang, S., Chen, H., Dai, X.-Y., and Chen, J. PRIMT: A pick-revise framework for interactive machine translation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1240–1249, 2016. URL http://www.aclweb.org/anthology/N16-1148. (Cited on pages 37, 38, 39, 48, 91, 93, 94, 114, 115, and 116)

Cherry, C. and Foster, G. Batch tuning strategies for statistical machine translation. In *Proceedings of NAACL-HLT*, 2012. URL http://dl.acm.org/citation.cfm?id=2382089. (Cited on page 26)

Chiang, D. A hierarchical phrase-based model for statistical machine translation. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 263–270, 2005. URL `https://doi.org/10.3115/1219840.1219873`. (Cited on page 17)

Clarkson, P. R. and Robinson, A. J. Language model adaptation using mixtures and an exponentially decaying cache. In *Proceedings of ICASSP-97*, pages 799–802, 1997. URL `http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.28.5976`. (Cited on page 47)

Cohn, T. and Lapata, M. Machine translation by triangulation: Making effective use of multi-parallel corpora. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 728–735, 2007. URL `http://aclweb.org/anthology/P07-1092`. (Cited on pages 110 and 148)

Cortes, C. and Vapnik, V. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995. URL `http://dx.doi.org/10.1007/BF00994018`. (Cited on pages 38 and 91)

Crammer, K., Dekel, O., Keshet, J., Shalev-Shwartz, S., and Singer, Y. Online passive-aggressive algorithms. *Journal of Machine Learning Research*, 7:551–585, 2006. URL `http://dl.acm.org/citation.cfm?id=1248547.1248566`. (Cited on page 43)

Crego, J. M. and Mariño, J. B. Integration of postag-based source reordering into SMT decoding by an extended search graph. In *Proceedings of the Conference of the Association for Machine Translation in the Americas (AMTA)*, pages 29–36, 2006. URL `http://www.mt-archive.info/AMTA-2006-Crego.pdf`. (Cited on page 36)

Crego, J. M., Max, A., and Yvon, F. Local lexical adaptation in machine translation through triangulation: SMT helping SMT. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010)*, pages 232–240, 2010. URL `http://www.aclweb.org/anthology/C10-1027`. (Cited on page 110)

Cuong, H. and Sima'an, K. Latent domain translation models in mix-of-domains haystack. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1928–1939, 2014. URL `http://www.aclweb.org/anthology/C14-1182`. (Cited on page 50)

Daumé III, H. and Jagarlamudi, J. Domain adaptation for machine translation by mining unseen words. In *Proceedings of the 49th Annual Meeting of the Association for Com-*

putational Linguistics: Human Language Techologies, pages 407–412, 2011. URL http://www.aclweb.org/anthology/P11-2071. (Cited on page 50)

de Gispert, A., Blackwood, G. W., Iglesias, G., and Byrne, W. N-gram posterior probability confidence measures for statistical machine translation: an empirical study. *Machine Translation*, 27(2):85–114, 2013. URL https://link.springer.com/article/10.1007/s10590-012-9132-2. (Cited on pages 37 and 76)

Dempster, A. P., Laird, N. M., and Rubin, D. B. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society Series B (Statistical Methodology)*, 39(1):1–38, 1977. URL http://web.mit.edu/6.435/www/Dempster77.pdf. (Cited on page 21)

Deng, Y. and Byrne, W. HMM word and phrase alignment for statistical machine translation. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 169–176, 2005. URL http://www.aclweb.org/anthology/H/H05/H05-1022. (Cited on page 21)

Denkowski, M. and Lavie, A. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the EACL 2014 Workshop on Statistical Machine Translation*, 2014. URL http://www.cs.cmu.edu/~alavie/METEOR/pdf/meteor-1.5.pdf. (Cited on pages 28 and 71)

Denkowski, M., Dyer, C., and Lavie, A. Learning from post-editing: Online model adaptation for statistical machine translation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 395–404, 2014. URL http://www.aclweb.org/anthology/E14-1042. (Cited on pages 11, 12, 44, 46, and 47)

Depraetere, I. What counts as useful advice in a university post-editing training context? Report on a case study. In *Proceedings of the 14th annual conference of the European Association for Machine Translation (EAMT)*, 2010. URL http://www.mt-archive.info/EAMT-2010-Depraetere-2.pdf. (Cited on page 35)

Du, J., Jiang, J., and Way, A. Facilitating translation using source language paraphrase lattices. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 420–429, 2010. URL http://www.aclweb.org/anthology/D10-1041. (Cited on page 36)

Du, J., Srivastava, A., Way, A., Maldonado-Guerra, A., and Lewis, D. An empirical study of segment prioritization for incrementally retrained post-editing-based SMT. In *Proceedings of the XV Machine Translation Summit*, 2015. URL `http://www.computing.dcu.ie/~away/PUBS/2015/MTSummit_main.pdf`. (Cited on page 49)

Durrani, N. and Koehn, P. Improving machine translation via triangulation and transliteration. In *Proceedings of th 17th International Conference of the European Association for Machine Translation (EAMT)*, 2014. URL `http://www.mt-archive.info/10/EAMT-2014-Durrani.pdf`. (Cited on page 110)

Dyer, C., Chahuneau, V., and Smith, N. A. A simple, fast, and effective reparameterization of IBM model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648, 2013. URL `http://www.aclweb.org/anthology/N13-1073`. (Cited on page 102)

Eck, M. and Hori, C. Overview of the IWSLT 2005 evaluation campaign. In *Proceedings of International Workshop on Spoken Language Translation (IWSLT)*, pages 11–32, 2005. URL `https://pdfs.semanticscholar.org/6150/3b049b824e2dc71c03e918fc63cd6cafddf8.pdf`. (Cited on page 57)

Eck, M., Vogel, S., and Waibel, A. Low cost portability for statistical machine translation based on n-gram coverage. In *Proceedings of the X Machine Translation Summit*, 2005. URL `http://www.mt-archive.info/MTS-2005-Eck.pdf`. (Cited on page 49)

Eisele, A. and Chen, Y. MultiUN: A multilingual corpus from United Nation documents. In *Proceedings of the Seventh conference on International Language Resources and Evaluation*, pages 2868–2872, 2010. URL `https://www.dfki.de/lt/publication_show.php?id=4790`. (Cited on page 100)

Elming, L. W. B., Jakob and Carl, M. Investigating user behaviour in post-editing and translation using the CASMACAT workbench. In O'Brien, S., Balling, L. W., Carl, M., Simard, M., and Specia, L., editors, *Post-Editing of Machine Translation: Processes and Applications*, pages 147–169. Cambridge Scholars Publishing, 2014. URL `http://www.cambridgescholars.com/post-editing-of-machine-translation-6`. (Cited on page 35)

Farajian, M. A., Turchi, M., Negri, M., Bertoldi, N., and Federico, M. Neural vs. phrase-based machine translation in a multi-domain scenario. In *Proceedings of the 15th Conference of*

the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, pages 280–284, 2017. URL http://www.aclweb.org/anthology/E17-2045. (Cited on page 31)

Federico, M., Koehn, P., Schwenk, H., and Trombetti, M. Matecat: Machine translation enhanced computer assisted translation. In *Proceedings of Machine Translation Summit XIV*, 2013. URL http://www.mt-archive.info/10/MTS-2013-Matecat.pdf. (Cited on pages 11 and 54)

Fiederer, R. and O'Brien, S. Quality and machine translation: A realistic objective? *Journal of Specialised Translation*, 11, 2009. URL http://www.jostrans.org/issue11/art_fiederer_obrien.pdf. (Cited on page 35)

Freitag, M., Huck, M., and Ney, H. Jane: Open source machine translation system combination. In *Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 29–32, 2014. URL http://www.aclweb.org/anthology/E14-2008. (Cited on page 148)

Gale, W. A., Church, K. W., and Yarowsky, D. One sense per discourse. In *Proceedings of the Workshop on Speech and Natural Language*, pages 233–237, 1992. URL http://dx.doi.org/10.3115/1075527.1075579. (Cited on page 131)

Galley, M. and Manning, C. D. A simple and effective hierarchical phrase reordering model. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 848–856, 2008. URL http://www.aclweb.org/anthology/D08-1089. (Cited on page 24)

Gao, Q. and Vogel, S. Parallel implementations of word alignment tool. In *Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, pages 49–57, 2008. URL http://dl.acm.org/citation.cfm?id=1622110.1622119. (Cited on page 70)

García, I. Translating by post-editing: Is it the way forward? *Machine Translation*, 25:217–237, 2011. URL http://dx.doi.org/10.1007/s10590-011-9115-8. (Cited on page 35)

Gascó, G., Rocha, M.-A., Sanchis-Trilles, G., Andrés-Ferrer, J., and Casacuberta, F. Does more data always yield better translations? In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 152–161, 2012. URL http://www.aclweb.org/anthology/E12-1016. (Cited on page 48)

Germann, U. Yawat: Yet Another Word Alignment Tool. In *Proceedings of the ACL-08: HLT Demo Session*, pages 20–23, 2008. URL `http://www.aclweb.org/anthology/P/P08/P08-4006`. (Cited on page 67)

Germann, U. Dynamic phrase tables for machine translation in an interactive post-editing scenario. In *Proceedings of the Workshop on Interactive and Adaptive Machine Translation, AMTA Workshop*, pages 20–31, 2014. URL `http://www.accept.unige.ch/Products/Germann2014.pdf`. (Cited on page 123)

Germann, U. Sampling phrase tables for the Moses statistical machine translation system. volume 104, pages 39–50, 2015. URL `https://ufal.mff.cuni.cz/pbml/104/art-germann.pdf`. (Cited on pages 46 and 123)

Gong, L. *On-demand Development of Statistical Machine Translation Systems*. PhD thesis, Université Paris Sud - Paris XI, 2014. URL `https://tel.archives-ouvertes.fr/tel-01144656/file/VD2_GONG_LI_25112014.pdf`. (Cited on page 45)

Gong, L., Max, A., and Yvon, F. Towards contextual adaptation for any-text translation. In *Proceedings of the International Workshop on Spoken Langugage Translation*, pages 292–300, 2012. URL `http://www.mt-archive.info/10/IWSLT-2012-Gong.pdf`. (Cited on page 50)

Gong, L., Max, A., and Yvon, F. Improving bilingual sub-sentential alignment by sampling-based transpotting. In *Proceedings of International Workshop on Spoken Language Translation (IWSLT)*, 2013. URL `http://www.mt-archive.info/10/IWSLT-2013-Gong.pdf`. (Cited on page 135)

Gong, L., Max, A., and Yvon, F. Incremental development of statistical machine translation systems. In *Proceedings of the eleventh International Workshop on Spoken Language Translation (IWSLT)*, pages 214–222, 2014. URL `http://www.mt-archive.info/10/IWSLT-2014-Gong.pdf`. (Cited on pages 47 and 102)

González-Rubio, J. and Casacuberta, F. Cost-sensitive active learning for computer-assisted translation. *Pattern Recognition Letters*, 37:124 – 134, 2014. URL `http://www.casmacat.eu/uploads/Main/1prl2014.pdf`. (Cited on page 49)

González-Rubio, J., Ortiz-Martínez, D., and Casacuberta, F. Active learning for interactive machine translation. In *Proceedings of the 13th Conference of the European Chapter of the*

Association for Computational Linguistics, pages 245–254, 2012. URL http://www.aclweb.org/anthology/E12-1025. (Cited on page 49)

Graça, J. a. V., Ganchev, K., and Taskar, B. Learning tractable word alignment models with complex constraints. *Computational Linguistics*, 36(3):481–504, 2010. URL http://dl.acm.org/citation.cfm?id=1950495. (Cited on page 21)

Graham, Y., Mathur, N., and Baldwin, T. Accurate evaluation of segment-level machine translation metrics. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics Human Language Technologies*, 2015. URL http://www.aclweb.org/anthology/N15-1124. (Cited on page 57)

Green, S. and Manning, C. D. Better Arabic parsing: Baselines, evaluations, and analysis. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 394–402, 2010. URL http://www.aclweb.org/anthology/C10-1045. (Cited on page 127)

Green, S., Heer, J., and Manning, C. D. The efficacy of human post-editing for language translation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 439–448, 2013. URL http://doi.acm.org/10.1145/2470654.2470718. (Cited on page 35)

Green, S., Wang, S. I., Chuang, J., Heer, J., Schuster, S., and Manning, C. D. Human effort and machine learnability in computer aided translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1225–1236, 2014. URL http://www.aclweb.org/anthology/D14-1130. (Cited on page 42)

Guerberof Arenas, A. Correlations between productivity and quality when post-editing in a professional context. *Machine Translation*, 28(3):165–186, 2014. URL https://link.springer.com/article/10.1007/s10590-014-9155-y. (Cited on page 35)

Guillou, L. and Hardmeier, C. PROTEST: A test suite for evaluating pronouns in machine translation. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, 2016. URL http://www.lrec-conf.org/proceedings/lrec2016/pdf/327_Paper.pdf. (Cited on page 60)

Haddow, B. and Koehn, P. Analysing the effect of out-of-domain data on SMT systems. In

Proceedings of the Seventh Workshop on Statistical Machine Translation, pages 422–432, 2012. URL `http://www.aclweb.org/anthology/W12-3154`. (Cited on page 48)

Haffari, G. and Sarkar, A. Active learning for multilingual statistical machine translation. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1 - Volume 1*, pages 181–189, 2009. URL `http://dl.acm.org/citation.cfm?id=1687878.1687905`. (Cited on page 49)

Haffari, G., Roy, M., and Sarkar, A. Active learning for statistical phrase-based machine translation. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 415–423, 2009. URL `http://www.mt-archive.info/NAACL-HLT-2009-Haffari.pdf`. (Cited on page 49)

Haffari, G., Tran, T. D., and Carman, M. Efficient benchmarking of NLP APIs using multi-armed bandits. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 408–416, 2017. URL `http://www.aclweb.org/anthology/E17-1039`. (Cited on page 148)

Hardt, D. and Elming, J. Incremental re-training for post-editing SMT. In *Proceedings the 9th Conference of the Association for Machine Translation in the Americas (AMTA)*, page 10, 2010. URL `http://www.mt-archive.info/AMTA-2010-Hardt.pdf`. (Cited on page 47)

He, X. Using word dependent transition models in HMM based word alignment for statistical machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 80–87, 2007. URL `http://dl.acm.org/citation.cfm?id=1626366`. (Cited on page 21)

Hokamp, C. and Liu, Q. Lexically constrained decoding for sequence generation using grid beam search. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1535–1546, 2017. URL `http://aclweb.org/anthology/P17-1141`. (Cited on page 48)

Hu, C., Resnik, P., Kronrod, Y., Eidelman, V., Buzek, O., and Bederson, B. B. The value of monolingual crowdsourcing in a real-world translation scenario: Simulation using Haitian Creole

emergency SMS messages. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 399–404, 2011. URL http://www.aclweb.org/anthology/W11-2148. (Cited on page 34)

Huck, M., Birch, A., and Haddow, B. Mixed domain vs. multi-domain statistical machine translation. In *Proceedings of Machine Translation Summit XV: vol.1: MT Researchers' Track*, 2015. URL http://homepages.inf.ed.ac.uk/abmayne/publications/Huck-MTSummit-2015.pdf. (Cited on page 51)

Hunsicker, S. and Ceausu, A. Machine translation quality estimation adapted to the translation workflow. In *Translating and The Computer 36: Asling: International Society for Advancement in Language Technology*, pages 133–136, 2014. URL http://www.mt-archive.info/10/Asling-2014-Hunsicker.pdf. (Cited on page 38)

Hutchins, J. From first conception to first demonstration: the nascent years of machine translation, 1947–1954. A chronology. *Machine Translation*, 12(3):195–252, 1997. URL http://dx.doi.org/10.1023/A:1007969630568. (Cited on page 16)

Irvine, A., Morgan, J., Carpuat, M., Daumé III, H., and Munteanu, D. Measuring machine translation errors in new domains. *Transactions of the Association for Computational Linguistics*, 1:429–440, 2013. URL https://aclweb.org/anthology/Q/Q13/Q13-1035.pdf. (Cited on page 61)

Jelinek, F. *Statistical Methods for Speech Recognition*. MIT Press, Cambridge, MA, USA, 1997. URL http://dl.acm.org/citation.cfm?id=280484. (Cited on page 26)

Jiang, J., Way, A., and Haque, R. Translating user-generated content in the social networking space. In *Proceedings of the Conference of the Association for Machine Translation in the Americas (AMTA)*, 2012. URL http://www.mt-archive.info/10/AMTA-2012-Jiang-2.pdf. (Cited on page 36)

Joty, S., Guzmán, F., Màrquez, L., and Nakov, P. DiscoTK: Using discourse structure for machine translation evaluation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 402–408, 2014. URL http://www.aclweb.org/anthology/W14-3352. (Cited on page 29)

Kay, M. The MIND system. In Rustin, R., editor, *Natural language processing: Courant Computer Science Symposium 8: December 20-21, 1971*, pages 155–188. Algorithmics Press, 1973. (Cited on page 33)

Kay, M. The proper place of men and machines in language translation. *Machine Translation*, 12 (1):3–23, 1997. URL `http://www.mt-archive.info/70/Kay-1980.pdf`. (Cited on pages 11, 33, and 131)

Kim, H. and Lee, J.-H. A recurrent neural networks approach for estimating the quality of machine translation output. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 494–498, 2016. URL `http://www.aclweb.org/anthology/N16-1059`. (Cited on page 38)

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014. URL `http://arxiv.org/abs/1412.6980`. (Cited on page 97)

Klein, D. and Manning, C. D. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 423–430, 2003. URL `http://www.aclweb.org/anthology/P03-1054`. (Cited on page 73)

Kneser, R. and Ney, H. Improved backing-off for m-gram language modeling. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, ICASSP'95*, pages 181–184, 1995. URL `https://pdfs.semanticscholar.org/9548/ac30c113562a51e603dbbc8e9fa651cfd3ab.pdf`. (Cited on page 25)

Knight, K. Decoding complexity in word-replacement translation models. *Computational Linguistics*, 25(4), 1999. URL `http://dl.acm.org/citation.cfm?id=973226.973232`. (Cited on page 26)

Knowles, R. and Koehn, P. Neural interactive translation prediction. In *Proceedings of the Conference of the Association for Machine Translation in the Americas (AMTA)*, 2016. URL `http://www.cs.jhu.edu/~phi/publications/neural-interactive-translation.pdf`. (Cited on pages 11, 40, and 41)

Koehn, P. A web-based interactive computer aided translation tool. In *Proceedings of the ACL-IJCNLP 2009 Software Demonstrations*, pages 17–20, 2009. URL `http://www.aclweb.org/anthology/P/P09/P09-4005`. (Cited on page 39)

Koehn, P. *Statistical Machine Translation*. Cambridge University Press, New York, NY, USA, 1st edition, 2010a. URL http://dl.acm.org/citation.cfm?id=1734086. (Cited on pages 17 and 21)

Koehn, P. Enabling monolingual translators: Post-editing vs. options. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 537–545, 2010b. URL http://www.aclweb.org/anthology/N10-1078. (Cited on page 34)

Koehn, P. Computer aided translation: Advances and challenges. Tutorial Presentation, 2016. URL https://amtaweb.org/wp-content/uploads/2016/10/Computer-Aided_Tutorial_Koehn_wide-cover.pdf. (Cited on page 53)

Koehn, P. and Monz, C. Manual and automatic evaluation of machine translation between european languages. In *Proceedings of the Workshop on Statistical Machine Translation*, pages 102–121, 2006. URL https://pdfs.semanticscholar.org/18c9/bd39f6fdb20eabb0a283a2c94257016596a0.pdf. (Cited on page 57)

Koehn, P. and Schroeder, J. Experiments in domain adaptation for statistical machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 224–227, 2007. URL http://dl.acm.org/citation.cfm?id=1626355.1626388. (Cited on page 47)

Koehn, P., Och, F. J., and Marcu, D. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, pages 48–54, 2003. URL http://homepages.inf.ed.ac.uk/pkoehn/publications/phrase2003.pdf. (Cited on pages 10, 17, 22, 23, and 27)

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. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, 2007. URL http://www.aclweb.org/anthology/P07-2045. (Cited on pages 30 and 69)

Koehn, P., Carl, M., Casacuberta, F., and Marcos, E. CASMACAT: Cognitive analysis and statistical methods for advanced computer aided translation. In *Proceedings of Machine Transla-*

tion Summit XIV, 2013. URL http://www.mt-archive.info/10/MTS-2013-CASMACAT.pdf. (Cited on pages 11 and 54)

Koehn, P., Tsoukala, C., and Saint-Amand, H. Refinements to interactive translation prediction based on search graphs. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 574–578, 2014. URL http://www.aclweb.org/anthology/P14-2094. (Cited on page 40)

Koehn, P., Alabau, V., Carl, M., Casacuberta, F., García-Martínez, M., González-Rubio, J., Keller, F., Ortiz-Martínez, D., Sanchis-Trilles, G., and Germann, U. Final public report. Technical report, Cognitive Analysis and Statistical Methods for Advanced Computer Aided Translation (Casmacat) project deliverable, January 2015. URL http://www.casmacat.eu/uploads/Deliverables/final-public-report.pdf. (Cited on pages 52 and 53)

Koponen, M. How to teach machine translation post-editing? Experiences from a post-editing course. In *Proceedings of Fourth Workshop of MT Summit XV on Post-editing Technology and Practice (WPTP 4)*, 2015. URL http://www.mt-archive.info/15/MTS-2015-W1-Koponen.pdf. (Cited on pages 11 and 35)

Kuhn, T. A survey and classification of controlled natural languages. *Computational Linguistics*, 40, 2014. URL https://aclweb.org/anthology/J/J14/J14-1005.pdf. (Cited on page 36)

Lacruz, I., Denkowski, M., and Lavie, A. Cognitive demand and cognitive effort in post-editing. In *Proceedings of the eleventh conference of the Association for Machine Translation in the Americas, Workshop on Post-editing Technology and Practice (WPTP-3)*, 2014. URL https://www.cs.cmu.edu/~alavie/papers/Lacruz-et-al-AMTA-2014-PEWorkshop.pdf. (Cited on page 35)

Lafferty, J. D., McCallum, A., and Pereira, F. C. N. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, pages 282–289, 2001. URL http://dl.acm.org/citation.cfm?id=655813. (Cited on pages 38 and 91)

Lagarda, A. L., Ortiz-Martínez, D., Alabau, V., and Casacuberta, F. Translating without in-domain corpus: Machine translation post-editing with online learning techniques. *Computer Speech & Language*, 32(1):109–134, 2015. URL http://dx.doi.org/10.1016/j.csl.2014.10.004. (Cited on page 51)

Langlais, P. and Lapalme, G. TransType: Development-evaluation cycles to boost translator's productivity. *Machine Translation*, 17(2):77–98, 2002. URL http://dx.doi.org/10.1023/B:COAT.0000010117.98933.a0. (Cited on pages 39 and 40)

Lardilleux, A., Yvon, F., and Lepage, Y. Hierarchical sub-sentential alignment with Anymalign. In *Proceedings of 16th Annual Conference of the European Association for Machine Translation (EAMT)*, pages 279–286, 2012. URL http://www.mt-archive.info/EAMT-2012-Lardilleux.pdf. (Cited on pages 44 and 135)

Lavergne, T., Cappé, O., and Yvon, F. Practical very large scale CRFs. In *Proceedings the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 504–513, 2010. URL http://www.aclweb.org/anthology/P10-1052. (Cited on page 97)

Levenberg, A., Callison-Burch, C., and Osborne, M. Stream-based translation models for statistical machine translation. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 394–402, 2010. URL http://www.aclweb.org/anthology/N10-1062. (Cited on page 46)

Levenshtein, V. I. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, 10(8):707–710, 1966. (Cited on page 117)

Lin, C.-Y. and Och, F. J. 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), Main Volume*, pages 605–612, 2004. URL http://www.aclweb.org/anthology/P04-1077. (Cited on page 43)

Lison, P. and Tiedemann, J. OpenSubtitles2016: Extracting large parallel corpora from movie and tv subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, 2016. URL http://www.lrec-conf.org/proceedings/lrec2016/pdf/947_Paper.pdf. (Cited on page 18)

Lo, C.-k. and Wu, D. Structured vs. flat semantic role representations for machine translation evaluation. In *Proceedings of Fifth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 10–20, 2011. URL http://www.aclweb.org/anthology/W11-1002. (Cited on pages 59 and 148)

Lo, C.-k. and Wu, D. MEANT at WMT 2013: A tunable, accurate yet inexpensive semantic frame based MT evaluation metric. In *Proceedings of the Eighth Workshop on Statistical*

Machine Translation, pages 422–428, 2013. URL http://www.aclweb.org/anthology/W13-2254. (Cited on page 29)

Lo, C.-k., Tumuluru, A. K., and Wu, D. Fully automatic semantic MT evaluation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 243–252, 2012. URL http://www.aclweb.org/anthology/W12-3129. (Cited on page 60)

Logacheva, V. and Specia, L. Phrase-level quality estimation for machine translation. In *Proceedings of International Workshop on Spoken Language Translation (IWSLT)*, pages 143–150, 2015. URL http://workshop2015.iwslt.org/downloads/IWSLT_2015_RP_12.pdf. (Cited on pages 38 and 94)

Logacheva, V., Hokamp, C., and Specia, L. MARMOT: A toolkit for translation quality estimation at the word level. In *Tenth International Conference on Language Resources and Evaluation (LREC)*, pages 3671–3674, 2016. URL http://www.lrec-conf.org/proceedings/lrec2016/pdf/1054_Paper.pdf. (Cited on page 97)

Lommel, A. Blues for BLEU: Reconsidering the validity of reference-based MT evaluation. In *Proceedings of LREC 2016 Workshop on Translation Evaluation: From Fragmented Tools and Data Sets to an Integrated Ecosystem (MT Eval 2016)*, 2016. URL http://www.cracking-the-language-barrier.eu/wp-content/uploads/Lommel.pdf. (Cited on page 28)

Lommel, A. R., Burchardt, A., Popovic, M., Harris, K., Avramidis, E., and Uszkoreit, H. Using a new analytic measure for the annotation and analysis of MT errors on real data. In *Proceedings of the 17th Annual Conference of the European Association for Machine Translation*, pages 165–172, 2014. URL https://www.dfki.de/lt/publication_show.php?id=7426. (Cited on page 58)

Lopez, A. Statistical machine translation. *ACM Computing Surveys*, 40(3), 2008a. URL http://doi.acm.org/10.1145/1380584.1380586. (Cited on page 17)

Lopez, A. Tera-scale translation models via pattern matching. In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING 2008)*, pages 505–512, 2008b. URL http://www.aclweb.org/anthology/C08-1064. (Cited on page 46)

Ma, W.-Y. and Mckeown, K. System combination for machine translation based on text-to-text generation. In *Proceedings of the 13th Machine Translation Summit (MT Summit XIII)*,

pages 546–553, 2011. URL http://www.mt-archive.info/MTS-2011-Ma-3.pdf. (Cited on page 148)

Machacek, M. and Bojar, O. Results of the WMT14 metrics shared task. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 293–301, 2014. URL http://www.aclweb.org/anthology/W14-3336. (Cited on page 29)

Manber, U. and Myers, G. Suffix arrays: a new method for on-line string searches. In *Proceedings of the first annual ACM-SIAM symposium on Discrete algorithms*, pages 319–327, 1990. URL http://portal.acm.org/citation.cfm?id=320218. (Cited on page 45)

Marie, B. and Apidianaki, M. Alignment-based sense selection in METEOR and the RATATOUILLE recipe. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 385–391, 2015. URL http://aclweb.org/anthology/W15-3048. (Cited on page 30)

Marie, B. and Max, A. Touch-based pre-post-editing of machine translation output. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1040–1045, 2015. URL http://aclweb.org/anthology/D15-1120. (Cited on pages 11 and 37)

Martínez-Gómez, P., Sanchis-Trilles, G., and Casacuberta, F. Passive-aggressive for on-line learning in statistical machine translation. In *Proceedings of the 5th Iberian Conference on Pattern Recognition and Image Analysis*, pages 240–247, 2011. URL https://link.springer.com/chapter/10.1007/978-3-642-21257-4_30. (Cited on page 44)

Martínez-Gómez, P., Sanchis-Trilles, G., and Casacuberta, F. Online adaptation strategies for statistical machine translation in post-editing scenarios. *Pattern Recognition*, 45(9):3193 – 3203, 2012. URL http://www.sciencedirect.com/science/article/pii/S0031320312000325. (Cited on page 44)

Mathur, P., Mauro, C., and Federico, M. Online learning approaches in computer assisted translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 301–308, 2013. URL http://www.aclweb.org/anthology/W13-2237. (Cited on pages 12, 44, and 124)

Mathur, P., Cettolo, M., Federico, M., and de Souza, J. G. C. Online multi-user adaptive statistical machine translation. In *Proceedings of the Workshop on Interactive and Adaptive Machine*

Translation, the 11th Conference of the Association for Machine Translation in the Americas (AMTA), pages 152–165, 2014. URL `http://www.mt-archive.info/10/AMTA-2014-Mathur.pdf`. (Cited on page 49)

Max, A., Crego, J. M., and Yvon, F. Contrastive lexical evaluation of machine translation. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, 2010. `http://www.mt-archive.info/10/LREC-2010-Max.pdf`. (Cited on page 60)

Mikolov, T., Chen, K., Corrado, G., and Dean, J. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781, 2013. URL `http://arxiv.org/abs/1301.3781`. (Cited on page 97)

Mirkin, S., Venkatapathy, S., and Dymetman, M. Confidence-driven rewriting for improved translation. In *Proceedings of the XIV Machine Translation Summit*, 2013. URL `http://www.mt-archive.info/10/MTS-2013-Mirkin.pdf`. (Cited on page 37)

Mitchell, L. The potential and limits of lay post-editing in an online community. In *Proceedings of 18th Annual Conference of the European Association for Machine Translation (EAMT)*, pages 67–74, 2015. URL `http://www.mt-archive.info/15/EAMT-2015-Mitchell.pdf`. (Cited on pages 35 and 68)

Miyata, R., Hartley, A., Paris, C., Tatsumi, M., and Kageura, K. Japanese controlled language rules to improve machine translatability of municipal documents. In *Proceedings of Machine Translation Summit XV: vol.1: MT Researchers' Track*, 2015. URL `http://www.mt-archive.info/15/MTS-2015-Miyata.pdf`. (Cited on page 36)

Mohit, B. and Hwa, R. Localization of difficult-to-translate phrases. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 248–255, 2007. URL `http://www.aclweb.org/anthology/W/W07/W07-0737.pdf`. (Cited on pages 11, 12, 37, 38, 39, 48, 87, 91, 92, 93, 112, 114, 115, and 116)

Monroe, W., Green, S., and Manning, C. D. Word segmentation of informal Arabic with domain adaptation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 206–211, 2014. URL `http://www.aclweb.org/anthology/P14-2034`. (Cited on page 101)

Moore, R. C. and Lewis, W. Intelligent selection of language model training data. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 220–224, 2010. URL http://www.aclweb.org/anthology/P10-2041. (Cited on page 50)

Morita, D. and Ishida, T. Designing protocols for collaborative translation. In *Proceedings 12th International Conference on Principles of Practice in Multi-Agent Systems (PRIMA 2009)*, 2009. URL http://dx.doi.org/10.1007/978-3-642-11161-7_2. (Cited on page 37)

Nagao, M. A framework of a mechanical translation between Japanese and English by analogy principle. In *Proceedings of the International NATO Symposium on Artificial and Human Intelligence*, pages 173–180, 1984. URL http://dl.acm.org/citation.cfm?id=2927.2938. (Cited on page 17)

Nakov, P. Improving English-Spanish statistical machine translation: Experiments in domain adaptation, sentence paraphrasing, tokenization, and recasing. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 147–150, 2008. URL http://www.aclweb.org/anthology/W/W08/W08-0320.pdf. (Cited on page 47)

Neal, R. and Hinton, G. E. A view of the EM algorithm that justifies incremental, sparse, and other variants. In *Learning in Graphical Models*, pages 355–368, 1998. URL http://www.cs.toronto.edu/~fritz/absps/emk.pdf. (Cited on page 44)

Nelder, J. A. and Mead, R. A simplex method for function minimization. *Computer Journal*, 7:308–313, 1965. URL https://academic.oup.com/comjnl/article-abstract/7/4/308/354237/A-Simplex-Method-for-Function-Minimization?redirectedFrom=fulltext. (Cited on page 124)

Nepveu, L., Lapalme, G., Langlais, P., and Foster, G. Adaptive language and translation models for interactive machine translation. In *Proceedings of EMNLP 2004*, pages 190–197, 2004. URL http://www.mt-archive.info/EMNLP-2004-Nepveu.pdf. (Cited on page 47)

Neubig, G. and Watanabe, T. Optimization for statistical machine translation: A survey. *Computational Linguistics*, 42(1):1–54, March 2016. URL http://www.phontron.com/paper/neubig16cl.pdf. (Cited on page 26)

O'Brien, S. Towards predicting post-editing productivity. *Machine Translation*, 25(3):197–215, 2011. URL http://dx.doi.org/10.1007/s10590-011-9096-7. (Cited on page 35)

Och, F. J. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 160–167, 2003. URL `http://www.aclweb.org/anthology/P03-1021`. (Cited on page 26)

Och, F. J. and Ney, H. Statistical multi-source translation. In *Machine Translation Summit*, pages 253–258, Santiago de Compostela, Spain, September 2001. URL `http://www.eamt.org/events/summitVIII/papers/och-1.pdf`. (Cited on page 148)

Och, F. J. and Ney, H. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29, 2003. URL `https://www.cse.iitb.ac.in/~pb/cs626-2013/word-alignment/alignment-comparison-J03-1002.pdf`. (Cited on page 21)

Och, F. J., Zens, R., and Ney, H. Efficient search for interactive statistical machine translation. In *Proceedings of the Tenth Conference on European Chapter of the Association for Computational Linguistics - Volume 1(EACL '03)*, pages 387–393, 2003. URL `http://dx.doi.org/10.3115/1067807.1067858`. (Cited on pages 35 and 40)

Onishi, T., Utiyama, M., and Sumita, E. Paraphrase lattice for statistical machine translation. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 1–5, 2010. URL `http://www.aclweb.org/anthology/P10-2001`. (Cited on page 36)

Ortiz-Martínez, D. *Advances in Fully-Automatic and Interactive Phrase-Based Statistical Machine Translation*. PhD thesis, Universidad Politécnica de Valencia, 2011. URL `https://riunet.upv.es/handle/10251/12127`. (Cited on pages 46 and 47)

Ortiz-Martínez, D., García-Varea, I., and Casacuberta, F. Interactive machine translation based on partial statistical phrase-based alignments. In *Proceedings of International Conference RANLP*, 2009. URL `http://www.mt-archive.info/RANLP-2009-Ortiz-Martinez.pdf`. (Cited on page 40)

Oswald, V. A. and Fletcher, S. L. Proposals for the mechanical resolution of German syntax patterns. *Modern Language Forum*, 36(3/4):1–24, 1951. URL `http://www.mt-archive.info/Oswald-1951.pdf`. (Cited on page 16)

Pal, S., Naskar, S. K., and Bandyopadhyay, S. Word alignment-based reordering of source chunks in PB-SMT. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, 2014. URL `http://www.lrec-conf.org/proceedings/lrec2014/pdf/982_Paper.pdf`. (Cited on page 36)

Pal, S., Naskar, S. K., Vela, M., and van Genabith, J. A neural network based approach to automatic post-editing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 281–286, 2016. URL http://anthology.aclweb.org/P16-2046. (Cited on page 51)

Pal, S., Naskar, S. K., Vela, M., Liu, Q., and van Genabith, J. Neural automatic post-editing using prior alignment and reranking. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 349–355, 2017. URL http://www.aclweb.org/anthology/E17-2056. (Cited on page 51)

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, 2002. URL http://www.aclweb.org/anthology/P02-1040. (Cited on page 27)

Parton, K., Habash, N., McKeown, K., Iglesias, G., and de Gispert, A. Can automatic post-editing make MT more meaningful? In *Proceedings of th 16th International Conference of the European Association for Machine Translation (EAMT)*, pages 111–118, 2012. URL http://www.mt-archive.info/EAMT-2012-Parton.pdf. (Cited on page 51)

Pedregosa, F., Varoquaux, G., Alexandre Gramfort, V. M., Thirion, B., Grisel, O., Mathieu Blondel, P. P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Édouard Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. URL http://www.jmlr.org/papers/v12/pedregosa11a.html. (Cited on pages 62 and 97)

Plitt, M. and Masselot, F. A productivity test of statistical machine translation post-editing in a typical localisation context. volume 93, pages 7–16, 2010. URL https://ufal.mff.cuni.cz/pbml/93/art-plitt-masselot.pdf. (Cited on page 35)

Popovic, M. and Ney, H. Towards automatic error analysis of machine translation output. *Computational Linguistics*, 37:657–688, 2011. URL http://dl.acm.org/citation.cfm?id=2077694. (Cited on page 59)

Popovic, M., Lommel, A. R., Burchardt, A., Avramidis, E., and Uszkoreit, H. Relations between different types of post-editing operations, cognitive effort and temporal effort. In *The Seventeenth Annual Conference of the European Association for Machine Translation (EAMT'14)*,

pages 191–198, 2014. URL http://www.dfki.de/web/forschung/iwi/publikationen/renameFileForDownload?filename=finalVersion48.pdf&file_id=uploads_2255. (Cited on page 35)

Porter, M. F. Snowball: A language for stemming algorithms, 2001. URL http://snowball.tartarus.org/texts/introduction.html. (Cited on page 71)

Rasmussen, C. E. and Williams, C. K. I. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. MIT Press, 2005. URL http://dl.acm.org/citation.cfm?id=1162254. (Cited on page 38)

Ratnaparkhi, A. A simple introduction to maximum entropy models for natural language processing. Technical report, Institute for Research in Cognitive Science, 1997. URL http://repository.upenn.edu/cgi/viewcontent.cgi?article=1083&context=ircs_reports. (Cited on page 39)

Raybaud, S. *De l'utilisation de mesures de confiance en traduction automatique : évaluation, post-édition et application à la traduction de la parole*. PhD thesis, Université de Lorraine, 2012. URL http://docnum.univ-lorraine.fr/public/DDOC_T_2012_0260_RAYBAUD.pdf. (Cited on page 38)

Rayner, M., Bouillon, P., and Haddow, B. Using source-language transformations to address register mismatches in SMT. In *Proceedings the 10th Conference of the Association for Machine Translation in the Americas (AMTA)*, page 10, 2012. URL http://www.mt-archive.info/10/AMTA-2012-Rayner.pdf. (Cited on page 36)

Reifler, E. MT. *Studies in Mechanical Translation*, 1, 1950. URL http://www.mt-archive.info/Reifler-1950.pdf. (Cited on page 33)

Resnik, P., Buzek, O., Hu, C., Kronrod, Y., Quinn, A., and Bederson, B. B. Improving translation via targeted paraphrasing. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 127–137, 2010. URL http://aclweb.org/anthology/D10-1013. (Cited on pages 11, 37, and 112)

Rosa, R., Mareček, D., and Dušek, O. DEPFIX: A system for automatic correction of Czech MT outputs. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 362–368, 2012. URL http://www.aclweb.org/anthology/W12-3146. (Cited on page 51)

Rosenblatt, F. The perceptron, a perceiving and recognizing automaton. In *Project Para Report No. 85-460-1, Cornell Aeronautical Laboratory (CAL)*, 1957. URL http://dl.acm.org/citation.cfm?id=1074686. (Cited on pages 38, 39, 43, and 91)

Salson, M., Lecroq, T., Léonard, M., and Mouchard, L. Dynamic extended suffix arrays. *Journal of Discrete Algorithms*, 8(2):241 – 257, 2010. URL http://www-igm.univ-mlv.fr/~lecroq/articles/jda2009.pdf. (Cited on page 46)

Sanchez-Torron, M. and Koehn, P. Machine translation quality and post-editor productivity. In *Proceedings of the Conference of the Association for Machine Translation in the Americas (AMTA)*, 2016. URL http://www.cs.jhu.edu/~phi/publications/machine-translation-quality.pdf. (Cited on pages 11 and 35)

Sanchis-Trilles, G., Alabau, V., Casacuberta, F., Benedí, J. M., González-Rubio, J., Ortiz-Martínez, D., and Koehn, P. Progress report on interactive translation prediction. Technical Report D2.2, Cognitive Analysis and Statistical Methods for Advanced Computer Aided Translation (CASMACAT) project deliverable, November 2013. http://www.casmacat.eu/uploads/Deliverables/d2.2.pdf. (Cited on page 38)

Sanchis-Trilles, G., Alabau, V., Buck, C., Carl, M., Casacuberta, F., García-Martínez, M., Germann, U., González-Rubio, J., Hill, R. L., Koehn, P., Leiva, L. A., Mesa-Lao, B., Ortiz-Martínez, D., Saint-Amand, H., Tsoukala, C., and Vidal, E. Interactive translation prediction versus conventional post-editing in practice: a study with the CasMaCat workbench. *Machine Translation*, 28(3):217–235, 2014. URL https://www.cs.jhu.edu/~phi/publications/MTJ_2014_CASMACAT.pdf. (Cited on page 42)

Schmid, H. Improvements in part-of-speech tagging with an application to German. In *Proceedings of the ACL SIGDAT-Workshop*, pages 47–50, 1995. URL http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/tree-tagger2.pdf. (Cited on pages 71, 97, and 101)

Schroeder, J., Cohn, T., and Koehn, P. Word lattices for multi-source translation. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 719–727, 2009. URL http://www.aclweb.org/anthology/E09-1082. (Cited on page 148)

Schwartz, L. Multi-source translation methods. In *Proc. AMTA*, October 2008. URL http://www.mt-archive.info/05/AMTA-2008-Schwartz.pdf. (Cited on page 148)

Schwartz, L., Anderson, T., Gwinnup, J., and Young, K. Machine translation and monolingual postediting: The AFRL WMT-14 system. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 186–194, 2014. URL `http://www.aclweb.org/anthology/W14-3321`. (Cited on pages 34 and 35)

Sennrich, R. Mixture-modeling with unsupervised clusters for domain adaptation in statistical machine translation. In *Proceedings of th 16th International Conference of the European Association for Machine Translation (EAMT)*, pages 185–192, 2012. URL `http://www.mt-archive.info/EAMT-2012-Sennrich`. (Cited on page 50)

Sennrich, R., Schwenk, H., and Aransa, W. A multi-domain translation model framework for statistical machine translation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 832–840, 2013. URL `http://www.aclweb.org/anthology/P13-1082`. (Cited on page 51)

Seretan, V., Bouillon, P., and Gerlach, J. A large-scale evaluation of pre-editing strategies for improving user-generated content translation. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1793–1799, 2014. URL `http://www.lrec-conf.org/proceedings/lrec2014/pdf/676_Paper.pdf`. (Cited on pages 11, 36, 52, and 112)

Settles, B. Active learning literature survey. Computer sciences technical report, 2009. URL `http://burrsettles.com/pub/settles.activelearning.pdf`. (Cited on page 49)

Shannon, C. E. A mathematical theory of communication. *Bell system technical journal*, 27(3): 379–423, 1948. URL `http://doi.acm.org/10.1145/584091.584093`. (Cited on page 19)

Shapira, D. and Storer, J. A. Edit distance with move operations. *Journal of Discrete Algorithms*, 5(2):380 – 392, 2007. URL `http://www.sciencedirect.com/science/article/pii/S157086670600030X`. (Cited on page 28)

Sharoff, S. and Nivre, J. The proper place of men and machines in language technology: Processing Russian without any linguistic knowledge. In *Proceedings of Dialogue 2011, Russian Conference on Computational Linguistics*, 2011. URL `http://www.comp.leeds.ac.uk/ssharoff/p/b2hd-sharoff11dialog.html`. (Cited on page 127)

Sharoff, S., Kopotev, M., Erjavec, T., Feldman, A., and Divjak, D. Designing and evaluating a Russian tagset. In *Proceedings of the Seventh International Conference on Language Resources*

and Evaluation (LREC'08), 2008. URL http://lrec-conf.org/proceedings/lrec2008/pdf/78_paper.pdf. (Cited on page 101)

Sheremetyeva, S. On-the-fly translator assistant (readability and terminology handling). In *Proceedings of the EACL 2014 Workshop on Humans and Computer-assisted Translation*, pages 22–27, 2014. URL http://www.aclweb.org/anthology/W14-0304. (Cited on page 52)

Simard, M. and Foster, G. PEPr: Post-edit propagation using phrase-based statistical machine translation. In *Proceedings of the XIV Machine Translation Summit*, 2013. URL http://www.mt-archive.info/10/MTS-2013-Simard.pdf. (Cited on page 51)

Simard, M., Goutte, C., and Isabelle, P. Statistical phrase-based post-editing. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 508–515, 2007a. URL http://www.aclweb.org/anthology/N/N07/N07-1064. (Cited on page 51)

Simard, M., Ueffing, N., Isabelle, P., and Kuhn, R. Rule-based translation with statistical phrase-based post-editing. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 203–206, 2007b. URL http://dl.acm.org/citation.cfm?id=1626355.1626383. (Cited on page 51)

Slocum, J. A survey of machine translation: Its history, current status, and future prospects. *Computatonal Linguistics*, 11(1):1–17, January 1985. URL http://dl.acm.org/citation.cfm?id=5615.5616. (Cited on page 113)

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., and Makhoul, J. A study of translation edit rate with targeted human annotation. In *Proceedings of Association for Machine Translation in the Americas*, pages 223–231, 2006. URL http://www.cs.umd.edu/~snover/pub/amta06/ter_amta.pdf. (Cited on page 28)

Snover, M., Madnani, N., Dorr, B. J., and Schwartz, R. TER-plus: Paraphrase, semantic, and alignment enhancements to translation edit rate. *Machine Translation*, 23:117 – 127, 2009a. URL http://dl.acm.org/citation.cfm?id=1743646. (Cited on page 29)

Snover, M., Madnani, N., Dorr, B. J., and Schwartz, R. Fluency, adequacy, or HTER?: Exploring different human judgments with a tunable MT metric. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 259–268, 2009b. URL http://dl.acm.org/citation.cfm?id=1626431.1626480. (Cited on page 71)

Sokolov, A., Wisniewski, G., and Yvon, F. Computing lattice BLEU oracle scores for machine translation. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 120–129, 2012. URL `http://www.aclweb.org/anthology/E12-1013`. (Cited on page 71)

Sokolov, A., Kreutzer, J., Lo, C., and Riezler, S. Learning structured predictors from bandit feedback for interactive NLP. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1610–1620, 2016. URL `http://www.aclweb.org/anthology/P16-1152`. (Cited on page 148)

Specia, L. and Giménez, J. Combining confidence estimation and reference-based metrics for segment-level MT evaluation. In *Proceedings the 9th Conference of the Association for Machine Translation in the Americas (AMTA)*, page 10, 2010. URL `http://www.mt-archive.info/10/AMTA-2010-Specia.pdf`. (Cited on page 37)

Specia, L., Turchi, M., Cancedda, N., Dymetman, M., and Cristianini, N. Estimating the sentence-level quality of machine translation systems. In *Proceedings of 13th Annual Conference of the European Association for Machine Translation (EAMT)*, pages 28–37, 2009. URL `http://www.mt-archive.info/EAMT-2009-Specia.pdf`. (Cited on page 37)

Specia, L., Paetzold, G., and Scarton, C. Multi-level translation quality prediction with QuEst++. In *ACL-IJCNLP 2015 System Demonstrations*, pages 115–120, 2015. URL `http://www.aclweb.org/anthology/P15-4020`. (Cited on pages 62 and 97)

Stanojević, M. and Sima'an, K. BEER 1.1: ILLC UvA submission to metrics and tuning task. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 396–401, 2015. URL `http://aclweb.org/anthology/W15-3050`. (Cited on page 29)

Stanojević, M. and Sima'an, K. Hierarchical permutation complexity for word order evaluation. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2164–2173, 2016. URL `http://aclweb.org/anthology/C16-1204`. (Cited on page 60)

Stanojević, M., Kamran, A., Koehn, P., and Bojar, O. Results of the WMT15 metrics shared task. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 256–273, 2015. URL `http://aclweb.org/anthology/W15-3031`. (Cited on page 29)

Stolcke, A. SRILM - an extensible language modeling toolkit. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, pages 901–904, 2002. URL http://www.speech.sri.com/projects/srilm/papers/icslp2002-srilm.pdf. (Cited on pages 62, 70, and 95)

Sutskever, I., Vinyals, O., and Le, Q. V. Sequence to sequence learning with neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems*, pages 3104–3112, 2014. URL http://papers.nips.cc/paper/5346-sequence-to-sequence-learning-with-neural-networks.pdf. (Cited on pages 10, 17, and 40)

Temnikova, I. Cognitive evaluation approach for a controlled language post-editing experiment. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, 2010. URL http://www.mt-archive.info/LREC-2010-Temnikova.pdf. (Cited on pages 35 and 36)

Tiedemann, J. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, 2012. URL http://www.lrec-conf.org/proceedings/lrec2012/pdf/463_Paper.pdf. (Cited on page 100)

Tillmann, C. A unigram orientation model for statistical machine translation. In *Proceedings of HLT-NAACL 2004: Short Papers*, pages 101–104, 2004. URL http://dl.acm.org/citation.cfm?id=1613984.1614010. (Cited on page 23)

Tomita, M. Feasibility study of personal/interactive machine translation systems. In *Proceedings of the Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages*, 1985. URL http://www.mt-archive.info/TMI-1985-Tomita.pdf. (Cited on page 113)

Toutanova, K., Ilhan, H. T., and Manning, C. D. Extensions to HMM-based statistical word alignment models. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10*, pages 87–94, 2002. URL http://www.aclweb.org/anthology/W02-1012. (Cited on page 21)

Toutanova, K., Klein, D., Manning, C. D., and Singer, Y. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, 2003. URL http://www.aclweb.org/anthology/N03-1033. (Cited on pages 71 and 94)

Turchi, M., De Bie, T., and Cristianini, N. Learning performance of a machine translation system: a statistical and computational analysis. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 35–43, 2008. URL `http://www.aclweb.org/anthology/W/W08/W08-0305`. (Cited on page 48)

Turchi, M., Negri, M., and Federico, M. MT quality estimation for computer-assisted translation: Does it really help? In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 530–535, 2015. URL `http://www.aclweb.org/anthology/P15-2087`. (Cited on page 38)

Ueffing, N. and Ney, H. Word-level confidence estimation for machine translation using phrase-based translation models. In *Computational Linguistics*, pages 763–770, 2005. URL `http://www.aclweb.org/anthology/J07-1003`. (Cited on page 37)

Ueffing, N., Och, F. J., and Ney, H. Generation of word graphs in statistical machine translation. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing*, pages 156–163, 2002. URL `http://www.aclweb.org/anthology/W02-1021`. (Cited on page 40)

Underwood, N., Mesa-Lao, B., Martínez, M. G., Carl, M., Alabau, V., González-Rubio, J., Leiva, L. A., Sanchis-Trilles, G., Ortiz-Martínez, D., and Casacuberta, F. Evaluating the effects of interactivity in a post-editing workbench. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 553–559, 2014. URL `http://www.lrec-conf.org/proceedings/lrec2014/pdf/289_Paper.pdf`. (Cited on page 42)

Utiyama, M. and Isahara, H. A comparison of pivot methods for phrase-based statistical machine translation. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 484–491, 2007. URL `http://www.aclweb.org/anthology/N07-1061`. (Cited on page 110)

Vilar, D., Xu, J., D'haro, L. F., and Ney, H. Error analysis of statistical machine translation output. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC-2006)*, 2006. URL `http://www.lrec-conf.org/proceedings/lrec2006/pdf/413_pdf.pdf`. (Cited on page 58)

Vogel, S., Ney, H., and Tillmann, C. HMM-based word alignment in statistical translation. In *Proceedings of the 16th Conference on Computational Linguistics - Volume 2 (COLING '96)*, pages 836–841, 1996. URL http://dl.acm.org/citation.cfm?id=993313. (Cited on page 21)

Wang, W., Peter, J.-T., Rosendahl, H., and Ney, H. CharacTer: Translation edit rate on character level. In *Proceedings of the First Conference on Machine Translation*, pages 505–510, 2016. URL http://www.aclweb.org/anthology/W16-2342. (Cited on page 29)

Watanabe, T., Suzuki, J., Tsukada, H., and Isozaki, H. Online large-margin training for statistical machine translation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 764–773, 2007. URL http://www.aclweb.org/anthology/D/D07/D07-1080. (Cited on page 43)

Weaver, W. Translation. In Locke, W. N. and Booth, A. D., editors, *Machine translation of languages, fourteen essays*, pages 15–23. MIT Press, John Wiley & Sons, Inc., New York, 1955. URL http://www.mt-archive.info/50/Locke-1955-TOC.htm. (Cited on page 33)

White, J. S. The ARPA MT evaluation methodologies: Evolution, lessons, and further approaches. In *Proceedings of the 1994 Conference of the Association for Machine Translation in the Americas*, pages 193–205, 1994. URL https://pdfs.semanticscholar.org/bf45/f9e578cb4b43a2604d6149553ae8cfee3016.pdf. (Cited on page 57)

Wisniewski, G., Singh, A. K., Segal, N., and Yvon, F. Design and analysis of a large corpus of post-edited translations: Quality estimation, failure analysis and the variability of post-edition. In *Proceedings of Machine Translation Summit XIV*, pages 117–124, 2013. URL http://www.mt-archive.info/10/MTS-2013-Wisniewski.pdf. (Cited on page 35)

Witten, I. H. and Bell, T. C. The zero-frequency problem: Estimating the probabilities of novel events in adaptive text compression. *IEEE Transactions on Information Theory*, 37(4), 1991. URL http://dl.acm.org/citation.cfm?id=2263404.2271157. (Cited on page 25)

Wu, X., Du, J., Liu, Q., and Way, A. ProphetMT: A tree-based SMT-driven controlled language authoring/post-editing tool. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, 2016. URL http://www.lrec-conf.org/proceedings/lrec2016/pdf/749_Paper.pdf. (Cited on page 52)

Wuebker, J., Green, S., and DeNero, J. Hierarchical incremental adaptation for statistical machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1059–1065, 2015. URL `http://aclweb.org/anthology/D15-1123`. (Cited on pages 44 and 50)

Wuebker, J., Green, S., DeNero, J., Hasan, S., and Luong, M.-T. Models and inference for prefix-constrained machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75, 2016. URL `http://www.aclweb.org/anthology/P16-1007`. (Cited on pages 40 and 41)

Xia, F. and McCord, M. Improving a statistical MT system with automatically learned rewrite patterns. In *Proceedings of COLING 2004*, pages 508–514, 2004. URL `http://www.aclweb.org/anthology/C04-1073`. (Cited on page 36)

Yu, H., Ma, Q., Wu, X., and Liu, Q. CASICT-DCU participation in WMT2015 metrics task. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 417–421, 2015. URL `http://aclweb.org/anthology/W15-3053`. (Cited on pages 29 and 30)

Zeman, D., Fishel, M., Berka, J., and Bojar, O. Addicter: What is wrong with my translations? *The Prague Bulletin of Mathematical Linguistics*, 96:79–88, 2011. URL `http://ufal.ms.mff.cuni.cz/pbml/96/art-zeman-fishel-berka-bojar.pdf`. (Cited on page 59)

Zoph, B. and Knight, K. Multi-source neural translation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 30–34, 2016. URL `http://www.aclweb.org/anthology/N16-1004`. (Cited on page 148)

Synthèse

La traduction automatique (TA) a connu des progrès significatifs ces dernières années. Elle est utilisée aujourd'hui dans de nombreux contextes, y compris les environnements professionnels de traduction et les scénarios de production. Cependant, le processus de traduction requiert souvent des connaissances plus larges que celles extraites de corpus parallèles. Puisque que l'injection de connaissances humaines dans la TA s'avère aujourd'hui encore nécessaire, l'un des moyens possibles d'améliorer TA est d'assurer une collaboration optimisée entre l'humain et la machine. À cette fin, de nombreuses questions sont posées pour la recherche en TA: Comment détecter les passages où une aide humaine devrait être proposée ? Comment faire pour que les machines exploitent les connaissances humaines afin d'améliorer leurs sorties ? Enfin, et de manière importante, comment optimiser l'échange afin de minimiser l'effort humain mis en jeu tout en maximisant la qualité de TA ? Diverses solutions ont été proposés selon les scénarios de traductions considérés.

Dans cette thèse, nous avons choisi de nous concentrer sur la pré-édition, une intervention humaine en TA qui a lieu ex-ante, par opposition à la post-édition, où l'intervention humaine qui déroule ex-post. En particulier, nous étudions des scénarios de pré-édition ciblés où l'humain doit fournir des traductions pour des segments sources difficiles à traduire. Les scénarios de la pré-édition impliquant la pré-traduction restent étonnamment peu étudiés dans la communauté. Cependant, ces scénarios peuvent offrir une série d'avantages relativement, notamment, à des scénarios de post-édition non ciblés, tels que : la réduction de la charge cognitive requise pour analyser des phrases mal traduites; davantage de contrôle sur le processus; une possibilité que la machine exploite de nouvelles connaissances pour améliorer la traduction automatique au voisinage des segments pré-traduits, etc. De plus, dans un contexte multilingue, des difficultés communes peuvent être résolues simultanément pour de nombreuses langues. De tels scénarios s'adaptent donc parfaitement aux contextes de production standard, où l'un des principaux objectifs est de réduire le coût de l'intervention humaine et où les traductions sont généralement effectuées à partir d'une langue vers plusieurs langues à la fois. Dans ce contexte, nous nous concentrons sur la TA de revues systématiques en médecine. En considérant cet exemple, nous proposons une méthodologie indépendante du système pour la détection des difficultés de traduction. Nous définissons la notion de difficulté de traduction de la manière suivante : les segments difficiles à traduire sont des segments pour lesquels un système de TA fait des prédictions erronées. Nous formulons le problème comme un problème de classification binaire et montrons

que, en utilisant cette méthodologie, les difficultés peuvent être détectées de manière fiable sans avoir accès à des informations spécifiques au système. Nous montrons que dans un contexte multilingue, les difficultés communes sont rares. Une perspective plus prometteuse en vue d'améliorer la qualité réside dans des approches dans lesquelles les traductions dans les différentes langues s'aident mutuellement à résoudre leurs difficultés. Nous intégrons les résultats de notre procédure de détection des difficultés dans un protocole de pré-édition qui permet de résoudre ces difficultés par pré-traduction. Nous évaluons le protocole dans un cadre simulé et montrons que la pré-traduction peut être à la fois utile pour améliorer la qualité de la TA et réaliste en termes des efforts humains nécessaires. En outre, nous montrons que les effets indirects sont significatifs. Nous évaluons également notre protocole dans un contexte préliminaire impliquant des interventions humaines. Les résultats de ces expériences pilotes confirment les résultats obtenus dans le cadre simulé et ouvrent des perspectives pour des expériences ultérieures.

Titre : Vers une meilleure collaboration humain-machine en traduction statistique: l'exemple des revues systématiques en médecine

Mots clefs : Traduction automatique Statistique, Traduction automatique de haute qualité, Traduction automatique assistée par l'humain, Pré-édition en traduction automatique, Difficulté de traduction, Estimation de la qualité de Traduction

Résumé : La traduction automatique (TA) a connu des progrès significatifs ces dernières années et continue de s'améliorer. La TA est utilisée aujourd'hui avec succès dans de nombreux contextes, y compris les environnements professionnels de traduction et les scénarios de production. Cependant, le processus de traduction requiert souvent des connaissances plus larges que extraites de corpus parallèles. Étant donné qu'une injection de connaissances humaines dans la TA est nécessaire, l'un des moyens possibles d'améliorer TA est d'assurer une collaboration optimisée entre l'humain et la machine. À cette fin, de nombreuses questions sont posées pour la recherche en TA: Comment détecter les passages où une aide humaine devrait être proposée ? Comment faire pour que les machines exploitent les connaissances humaines obtenues afin d'améliorer leurs sorties ? Enfin, comment optimiser l'échange: minimiser l'effort humain impliqué et maximiser la qualité de TA? Diverses solutions sont possibles selon les scénarios de traductions considérés. Dans cette thèse, nous avons choisi de nous concentrer sur la pré-édition, une intervention humaine en TA qui a lieu ex-ante, par opposition à la post-édition, où l'intervention humaine qui déroule ex-post. En particulier, nous étudions des scénarios de pré-édition ciblés où l'humain doit fournir des traductions pour des segments sources difficiles à traduire et choisis avec soin. Les scénarios de la pré-édition impliquant la pré-traduction restent étonnamment peu étudiés dans la communauté. Cependant, ces scénarios peuvent offrir une série d'avantages relativement, notamment, à des scénarios de post-édition non ciblés, tels que : la réduction de la charge cognitive requise pour analyser des phrases mal traduites; davantage de contrôle sur le processus; une possibilité que la machine exploite de nouvelles connaissances pour améliorer la traduction automatique au voisinage des segments pré-traduits, etc. De plus, dans un contexte multilingue, des difficultés communes peuvent être résolues simultanément pour de nombreuses langues. De tels scénarios s'adaptent donc parfaitement aux contextes de production standard, où l'un des principaux objectifs est de réduire le coût de l'intervention humaine et où les traductions sont généralement effectuées à partir d'une langue vers plusieurs langues à la fois. Dans ce contexte, nous nous concentrons sur la TA de revues systématiques en médecine. En considérant cet exemple, nous proposons une méthodologie indépendante du système pour la détection des difficultés de traduction. Nous définissons la notion de difficulté de traduction de la manière suivante : les segments difficiles à traduire sont des segments pour lesquels un système de TA fait des prédictions erronées. Nous formulons le problème comme un problème de classification binaire et montrons que, en utilisant cette méthodologie, les difficultés peuvent être détectées de manière fiable sans avoir accès à des informations spécifiques au système. Nous montrons que dans un contexte multilingue, les difficultés communes sont rares. Une perspective plus prometteuse en vue d'améliorer la qualité réside dans des approches dans lesquelles les traductions dans les différentes langues s'aident mutuellement à résoudre leurs difficultés. Nous intégrons les résultats de notre procédure de détection des difficultés dans un protocole de pré-édition qui permet de résoudre ces difficultés par pré-traduction. Nous évaluons le protocole dans un cadre simulé et montrons que la pré-traduction peut être à la fois utile pour améliorer la qualité de la TA et réaliste en termes d'implication des efforts humains. En outre, les effets indirects sont significatifs. Nous évaluons également notre protocole dans un contexte préliminaire impliquant des interventions humaines. Les résultats de ces expériences pilotes confirment les résultats obtenus dans le cadre simulé et ouvrent des perspectives encourageantes pour des tests ultérieures.

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France

université PARIS-SACLAY
ÉCOLE DOCTORALE
Sciences et technologies de l'information et de la communication (STIC)

Title: Towards a Better Human-Machine Collaboration in Statistical Translation: Example of Systematic Medical Reviews

Key words: Statistical Machine Translation; High-Quality Machine Translation; Human-Assisted Machine Translation; Pre-Edition in translation; Translation Difficulty; Quality Estimation

Abstract: Machine Translation (MT) has made significant progress in the recent years and continues to improve. Today, MT is successfully used in many contexts, including professional translation environments and production scenarios. However, the translation process requires knowledge larger in scope than what can be captured by machines even from a large quantity of translated texts. Since injecting human knowledge into MT is required, one of the potential ways to improve MT is to ensure an optimized human-machine collaboration.

To this end, many questions are asked by modern research in MT: How to detect where human assistance should be proposed? How to make machines exploit the obtained human knowledge so that they could improve their output? And, not less importantly, how to optimize the exchange so as to minimize the human effort involved and maximize the quality of MT output? Various solutions have been proposed depending on concrete implementations of the MT process.

In this thesis we have chosen to focus on Pre-Edition (PRE), corresponding to a type of human intervention into MT that takes place *ex-ante*, as opposed to Post-Edition (PE), where human intervention takes place *ex-post*. In particular, we study targeted PRE scenarios where the human is to provide translations for carefully chosen, difficult-to-translate, source segments. Targeted PRE scenarios involving pre-translation remain surprisingly understudied in the MT community. However, such PRE scenarios can offer a series of advantages as compared, for instance, to non-targeted PE scenarios: i.a., the reduction of the cognitive load required to analyze poorly translated sentences; more control over the translation process; a possibility that the machine will exploit new knowledge to improve the automatic translation of neighboring words, etc. Moreover, in a multilingual setting common difficulties can be resolved at one time and for many languages. Such scenarios thus perfectly fit standard production contexts, where one of the main goals is to reduce the cost of PE and where translations are commonly performed simultaneously from one language into many languages. A representative production context – an automatic translation of systematic medical reviews – is the focus of this work.

Given this representative context, we propose a system-independent methodology for translation difficulty detection. We define the notion of translation difficulty as related to translation quality: difficult-to-translate segments are segments for which an MT system makes erroneous predictions. We cast the problem of difficulty detection as a binary classification problem and demonstrate that, using this methodology, difficulties can be reliably detected without access to system-specific information. We show that in a multilingual setting common difficulties are rare, and a better perspective of quality improvement lies in approaches where translations into different languages will help each other in the resolution of difficulties.

We integrate the results of our difficulty detection procedure into a PRE protocol that enables resolution of those difficulties by pre-translation. We assess the protocol in a simulated setting and show that pre-translation as a type of PRE can be both useful to improve MT quality and realistic in terms of the human effort involved. Moreover, indirect effects are found to be genuine. We also assess the protocol in a preliminary real-life setting. Results of those pilot experiments confirm the results in the simulated setting and suggest an encouraging beginning of the test phase.

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France