Neural methods for spoken dialogue understanding

Emile Chapuis

INSTITUT POLYTECHNIQUE DE PARIS

NNT : 2021IPPAT045

Thèse de doctorat

Neural Methods For Spoken Dialogue Understanding

Thèse de doctorat de l'Institut Polytechnique de Paris
préparée à Telecom Paris

École doctorale n°626 École doctorale de l'Institut Polytechnique de Paris (EDIPP)
Spécialité de doctorat : Informatique

Thèse présentée et soutenue à Palaiseau, le 15 Décembre 2021, par

EMILE CHAPUIS

Composition du Jury :

François Yvon LISN/CNRS, Paris, France	Président/Rapporteur
Benoit Favre LIS/CNRS, Aix-Marseille, France	Rapporteur
Emmanuel Morin LS2N/CNRS, Nantes, France	Examinateur
Verena Rieser Heriot Watt University, Edinburgh, Scotland	Examinateur
Christophe Cerisara LORIA/CNRS, Nancy, France	Examinateur
Chloe Clavel Telecom Paris, Palaiseau, France	Directeur de thèse
Matthieu Labeau Telecom Paris, Palaiseau, France	Co-directeur de thèse

Abstract

Conversational AI has received a growing interest in recent years from both the research community and the industry. Products have started to emerge (*e.g.* Amazon's Alexa, Google's Home, Apple's Siri) but performances of such systems are still far from human-likeness communication. As an example, conversation with the aforementioned systems is often limited to basic question-response interactions. Among all the reasons why people communicate, the exchange of information and the strengthening of social bound appeared to be the main ones. In dialogue research, the two aforementioned problems are well known and addressed using dialogue act classification and emotion/sentiment recognition. Those problems are made even more challenging as they involve spoken dialogues in contrast to written text. A spoken conversation is a complex and collective activity that has a specific dynamic and structure. Thus, there is a need to adapt both natural language processing and natural language understanding techniques which have been tailored for written texts as it does not share the same characteristics.

This thesis focuses on methods for spoken dialogue understanding and specifically tackles the problem of spoken dialogues classification with a particular focus on dialogue act and emotion/sentiment labels. Our contributions can be divided into two parts: in the first part, we address the problem of automatically labelling English spoken dialogues. In this part, we start by formulating this problem as a translation problem which leads us to propose a seq2seq model for dialogue act classification. Then, our second contribution focuses on a scenario relying on small annotated datasets and involves both pre-training a hierarchical transformer encoder and proposing a new benchmark for evaluation. This first part addresses the problem of spoken language classification in monolingual (*i.e.* English) and monomodal (*i.e.* text) settings. However, spoken dialogue involves phenomena such as code-switching (when a speaker switch languages within a conversation) and relies on multiple channels to communicate (`e.g.` audio or visual). Hence, the second part is dedicated to two extensions of the previous contributions in two settings: multilingual and multimodal. We first address the problem of dialogue act classification when multiple languages are involved and thus, we extend the two previous contributions to a multilingual scenario. In our last contribution, we explore a multimodal scenario and focus on the representation and fusion of modalities in the scope of emotion prediction.

Résumé

L'intelligence artificielle conversationnelle a suscité un intérêt croissant ces dernières années, tant dans la communauté des chercheurs que dans l'industrie. Des applications grand publique ont commencé à voir le jour (par exemple, Alexa d'Amazon, Home de Google, Siri d'Apple), mais les performances de ces systèmes sont encore loin d'une communication semblable à celle des humains. Par exemple, la conversation avec les systèmes susmentionnés se limite souvent à des interactions de base de type question-réponse. Parmi toutes les raisons pour lesquelles les gens communiquent, l'échange d'informations et le renforcement des liens sociaux semblent être les principales. Dans la recherche sur le dialogue, ces deux problèmes sont bien connus et abordés à l'aide de la classification des actes de dialogue et de la reconnaissance des émotions/sentiments. Ces problèmes sont d'autant plus difficiles à résoudre qu'ils concernent des dialogues parlés, contrairement aux textes écrits. Une conversation parlée est une activité complexe et collective qui possède une dynamique et une structure spécifiques. Il est donc nécessaire d'adapter les techniques de traitement et de compréhension du langage naturel qui ont été conçues pour les textes écrits car elles ne partagent pas les mêmes caractéristiques.
Cette thèse se concentre sur les méthodes de compréhension des dialogues parlés et aborde spécifiquement le problème de la classification des dialogues parlés avec un accent particulier sur les étiquettes des actes de dialogue et des émotions/sentiments. Nos contributions peuvent être divisées en deux parties : dans la première partie, nous abordons le problème de l'étiquetage automatique des dialogues parlés en anglais. Dans cette partie, nous commençons par formuler ce problème comme un problème de traduction, ce qui nous amène à proposer un modèle seq2seq pour la classification des actes de dialogue. Ensuite, notre deuxième contribution se concentre sur un scénario reposant sur de petits ensembles de données annotées et implique à la fois le pré-entraînement d'un encodeur transformateur hiérarchique et la proposition d'un nouveau benchmark pour l'évaluation. Cette première partie aborde le problème de la classification du langage parlé dans des contextes monolingues (*i.e.* anglais) et monomodaux (*i.e.* texte). Cependant, les dialogues parlés impliquent des phénomènes tels que le code-switching (lorsqu'un locuteur change de langue au cours d'une conversation) et s'appuient sur plusieurs canaux pour communiquer (par exemple, audio ou visuel).
La deuxième partie est donc consacrée à deux extensions des contributions précédentes dans deux contextes: multilingue et multimodal. Nous abordons d'abord le problème de la classification des actes de dialogue lorsque plusieurs langues sont impliquées et nous étendons donc les deux contributions précédentes à un scénario multilingue. Dans notre dernière contribution, nous explorons un scénario multimodal et nous nous concentrons sur la représentation et la fusion des modalités dans le cadre de la prédiction des émotions.

Acknowledgment

En premier lieu, j'aimerais remercier mes directeurs de thèse Chloé Clavel et Matthieu Labeau. D'abord Chloé, je te suis reconnaissant de m'avoir donné l'opportunité de faire cette thèse ainsi que d'avoir veillé à ce qu'elle se déroule dans les meilleures conditions possibles du début à la fin. Matthieu, merci à toi pour tous nos échanges, pour les retours que tu as pris le temps de me donner dès que j'avais besoin d'un avis.

Merci à mes rapporteurs de thèse, François Yvon et Benoit Favre, d'avoir pris le temps de relire attentivement mon manuscrit. I sincerely thank Verena Rieser, Emmanuel Morin and Christophe Cerisara, for accepting to be part of the jury.

Je tiens à exprimer ma gratitude envers Caio Corro et Nadi Tomeh qui m'ont offert l'opportunité de travailler avec eux après mon doctorat.

Enfin je tiens à adresser mes sincères remerciements à Sandra Plancade qui a su me transmettre sa passion pour la recherche et m'auras donné l'envie de poursuivre dans cette voix.

Merci à tous ceux que j'ai pu côtoyer durant ces 3 années de doctorats au laboratoire LTCI de Télécom Paris et qui ont participé à créer une ambiance de travail conviviale. Merci à Mr Garcia qui n'aura eu de cesse de réduire à néant nos petits rêves de thésards bizuts, Pierre ne s'en est jamais complètement remis. Merci à Kimia pour son humour pas piqué des hannetons, à Amaury pour sa bonne humeur et sa bienveillance, à Anas et Guillaume pour leur amour du débat et leurs avis tranchés. Merci à Dimitri en qui j'ai pu trouver une oreille plus qu'attentive à mes litanies surfistiques. J'espère qu'on pourra se faire une session dans ta Vendée un de ces quatre. Bonne chance à la nouvelle génération pour les années de thèse qu'il vous reste.

Merci Lucien et Pierre pour toutes nos discussions, fussent-elles autour d'un café ou d'une bière. Je suis heureux de vous avoir rencontré et de vous compter parmi mes amis. Je tiens à te remercier doublement Pierre car nous avons beaucoup collaboré et tout ce travail n'aurait pas été possible sans ton énergie et ton soutien sans failles. J'ai hâte que nous repartions en conf' et pourquoi pas se refaire le Pic du midi. Mais pas de triche cette fois !

Merci à mes amis Ajmal et Léo avec qui les débats endiablés mêlent analyses géopolitiques de haut vol et invectives des plus frontales ! Merci pour votre grinta quand il s'agit d'aller affronter les houles landaises. Léo, toi le davonote, l'amateur de gibolin, j'espère que tes rêves deviendront réalités. Paris est à tes pieds ! Même si tu ne parles pas le latin et que tu n'arranges pas les foules comme un pape. Ajmal, merci encore pour toutes ces folles soirées massicoises, un jour nous aussi on l'aura notre radio libre !

J'adresse mes derniers remerciements à ma famille. D'abord à mon oncle Alain, à ma tante Marie-Paul et à mes cousins Marine, Pauline et Ronan que j'ai eu tant de plaisir à retrouver lors d'un nième confinement. Le covid n'aura pas eu que des

aspects négatifs, après tout. Merci pour toutes ces soirées à refaire le monde. Merci Marine, ainsi que toutes les petites mains qui s'affairent autour d'elles, pour tous ces cageots de Rubis des jardins qui nous ont régalé tout le printemps. Merci également à mon taquin d'oncle Jean-Yves avec qui j'ai tant de plaisir à me chamailler !

Je remercie ma grand-mère, André, pour sa joie de vivre, son amour du jeu et de la poésie. Je tiens à remercier ma grand-mère, Mamick, qui encore à l'âge de 89 ans commence ses phrases par un "quand je serai vieille...". Merci pour tout ce que tu nous as donné, pour toute ton énergie que j'admire tant.

Je remercie mes parents pour tout. Pour m'avoir donné une si belle enfance. Pour m'avoir permis de faire des études longues. Pour m'avoir soutenu à chaque instant. J'ai beaucoup de chance de vous avoir. Je remercie mes frères Yves et Pierre ainsi que ma soeur Anne, pour tous ces moments passés ensemble, je lève mon verre à nos gamineries et nos blagues atroces ! Anne n'oublie pas que ta connaissance de l'art ancestrale de la Bavaroise te confère désormais un statut que tu te dois d'honorer !

Contents

List of Figures

List of Tables

Chapter 1

Introduction and Overview

1.1 Introduction

Spoken language is the most natural way of communication by which we transmit information or orders but also create social interactions. Its complexity is now a challenge for science and industry, leading to the rise of conversational AI with popular applications including Amazon's Alexa, Apple's Siri and Google's Home. Even if progress has been made, it has often been noted that such virtual agents are limited to a simple question-answer paradigm. Hence, we wonder whether it is possible to go beyond and develop an AI that will take into account the flow of the conversation, creating a more engaging experience for the user. Conversational AI relies on dialogue systems or more precisely, in spoken dialogue systems to manage the agent interaction. Such a system gathered two main components: a Natural Language Understanding (NLU) block which understands the intent of the user and a Natural Language Generation (NLG) whose aim is to deliver the agent's responses. In this thesis, we focus on the NLU part as it will condition the quality of the agent's future response.
NLU is a sub-domain of Natural Language Processing (NLP) and deals with machine comprehension. NLU covers a wide range of applications such as question-answering, text categorisation, machine translation, and slot filling. We focus on the applications that we believe to have the most impact on the agent's response which is the same as asking *what are the main reasons humans communicate ?*. We have identified two main reasons: exchanging information and strengthening social bonds. To exchange and share information (*i.e* stories, thoughts, ideas), we often communicate with others following a certain dialogue flow. Generally, human dialogues are not limited to a basic question/answer sequence which is, as previously said, one of the most common patterns in current dialogue systems. Instead, humans often first respond to previous context and then suggest their own answer (*e.g* questions, assertions). In this way, people display their interest and involvement in the conversation. The second reason for human communication is to strengthen their social bonding with others. For this reason, daily conversations are full of emotions. By expressing emotions, people communicate information to others about their feelings, intentions, relationship with the target of the emotions, and the environment. Understanding and empathising with this emotional state will influence the speech of those involved in the conversation and thus improve their relationship. The example of conversation in Table 1.1 provides a good illustration of the two phenomena. In the second speaker turn, speaker B understands the statement expressed by A and then asks

to elaborate. In the fourth speaker's turn, speaker B (1) shows empathy and understanding by expressing his/her emotion (2) gives an advice that could help speaker A. This suggestion is original and still related to the context. This demonstrate how speaker B creates a social bound with speaker A by addressing the emotional and conversational context.

A: I'm worried about something.
B: What's that?
A: Well, I have to drive to school for a meeting this morning, and I'm going to end up getting stuck in rush-hour traffic.
B: That's annoying, but nothing to worry about. *Just breathe deeply when you feel yourself getting upset.*
A: Ok, I'll try that.
B: Is there anything else bothering you?
A: Just one more thing. A school called me this morning to see if I could teach a few classes this weekend and I don't know what to do.
B: Do you have any other plans this weekend?
A: I'm supposed to work on a paper that'd due on Monday.
B: *Try not to take on more than you can handle.*
A: You're right. I probably should just work on my paper. Thanks!

Table 1.1 – Example of conversation from DailyDialog corpus between two speakers A and B. The underlined words in purple explicitly indicate the emotions. The words in italic are ideas expressed by speaker B that are new for the other speaker A.

The communication purposes mentioned above are two well known problems addressed in NLU as dialogue act (DA) classification and emotion/sentiment (E/S) recognition. Formally, DAs are semantic labels associated with each utterance in a conversational dialogue that indicate the speaker's intention, e.g., question, backchannel, statement-non-opinion, statement opinion. A key step to model dialogue is to detect the intent of the speaker. Indeed, correctly identifying a question gives an important clue to produce an appropriate response. Thus, the identification of both Dialogue Acts (DA) and Emotion/Sentiment (E/S) in spoken language is an important step toward improving models performances on spontaneous dialogue tasks and it is essential to avoid the generic response problem, i.e., having an automatic dialogue system that generates an unspecific response — that can be an answer to a very large number of user utterances.
Hence, this thesis focuses on methods for spoken dialogue understanding and specifically tackles the problem of spoken dialogue utterances classification, DA and E/S labels in particular. Formally, we introduce new methods relying on deep neural architectures to address the problem of dialogue utterances classification. This problem is first addressed in a monolingual and mono-modal setting (*i.e.* when the input conversations are composed of text in English). In this first setting, we propose new methods for two different situations: (1) when a high amount of English-only labelled dialogues are available, (2) when the amount of labelled data is limited but a large amount of unlabelled conversations are available. We believe this second situation is of great interest due to the high annotation costs and time of the first

situation. Then, we aim at generalizing our approach and work on extended settings of dialogues classification. We first propose an extension of our systems to handle several languages even in the presence of code-switched dialogue. Last, we study the addition of multimodal signals when dealing on E/S classification and focus on their representation and fusion. This final chapter is an introductory study of a broader project where we investigate the representation of visual and acoustic signals into a textual modality which is well leveraged by current neural architectures. Notably, punctuation marks have been studied as a textual representation of prosodical cues, hence we propose to quantitatively study their impact on current neural network models in the scope of E/S prediction.
These different settings are studied through a series of research questions that we precise below.

1.2 Research questions

Over the last decade, a lot of progress has been made in neural network architectures which has had a tremendous impact on multiple fields including NLP (*e.g* text classification, dialogue modelling, text summarizing). One of the key features of deep neural networks is their ability to handle different types of data (*e.g.* written texts, dialogues, news) coming from various data sources. Handcrafted feature systems have been replaced by neural architectures allowing to automatically learn rich representations. The present work leverages deep neural architectures to handle spoken dialogue data for classifications tasks. Therefore, we aim at proposing and studying new methods for spoken dialogue understanding. This thesis is organised along two groups of research questions. The first group of questions are dedicated to English data and aims at answering the following general research question: RQ1**: How to leverage recent advances in deep learning to build a system that can automatically label English spoken dialogue utterances?** We start by addressing this question on a direct supervised approach **when large labelled datasets are available**. Thus RQ1.1 boils down to **how to leverage recent advances in deep learning to build a system that can automatically label English spoken dialogue utterances when a large corpus of labelled data is available?** As previously mentioned, there is a multilevel dependency: between emitted utterances within the conversation, utterances and their corresponding labels and between labels. Hence, RQ1.1 can be divided into the following sub-questions :

1. *How to adapt the encoder architecture to the hierarchical aspect of dialogue?*
2. *How to better model long range dependency between labels in the decoder?*
3. *How to leverage the utterance/label dependency?*

However, the proposed solution requires large annotated corpora which are not always available for both DA and E/S. Thus, to alleviate this problem we put ourselves in a situation where data is scarce and when large annotated datasets are not available. One of the current solutions to the problem of lack of annotated data is to pre-train neural architectures in a self-supervised manner. Pre-trained language models such as BERT or RoBERTa have drastically changed the shape of modern NLP systems and allowed to make an efficient transfer learning. While it has been shown that the aforementioned models achieved competitive results in a variety of NLU tasks, they may

not be optimal when dealing with spoken dialogue data due to (1) the structure of the dialogue which is not a flattened text and (2) the discrepancy that exists between the written text these models are trained on and spoken dialogue data. Hence, within the context of data scarcity RQ1.2 boils down to **how to leverage recent advances in neural networks to build a system that can automatically label English spoken dialogue utterances when there is a lack of labelled data, especially in the case of small corpora or in low-resource setting?** This question can then be decomposed into the following sub-questions:

1. *Does an evaluation benchmark suited to low-resource sequence labelling for DA and E/S exist? If no, how to gather such an evaluation benchmark?*

2. *How to gather a large pre-training data that is suited for dialogue? How does the nature of the pre-training corpus influence the final performance of the whole system?*

3. *How to adapt transformer based architecture to the hierarchical aspect of dialogue?*

4. Spoken dialogues have a hierarchical structure. *How to take into account the hierarchy during pre-training ? How does the hierarchy impact pre-training objectives?*

The second group of questions is dedicated to the study spoken dialogue understanding in two different settings: multilingual and multimodal. Thus, RQ2 can be phrased as **How to modify spoken dialogue understanding systems to handle multilingual and multimodal data? What are the specificities of multilingual and multimodal data that the system needs to handle?** Henceforth, we further precise RQ2 in the two aforementioned settings.
In the first part of this thesis we focused on English data as most of the available resources are English-centric. However, as AI increasingly blends into everyday life across the globe, researchers are investigating the development of models that can handle multiple languages. Similarly, spoken dialogue systems must be able to handle languages other than English and as well as multilinguality within the same conversations. Hence, we formulate the following research question: RQ2.1 **How to build a system that can label multilingual conversational data and thus handle code-switching ?** This research question can be further split into the following subquestions:

1. *Does an evaluation benchmark for spoken dialogue classification in a multilingual setting exist? If no, how to gather such an evaluation benchmark? How to ensure that our representations are robust to different languages as well as code switching?*

2. *How to gather a large pre-training data that is suited for multilingual dialog? What are the specifics of the multilinguality that can influence the gathering of the pre-training corpus?*

3. *How to adapt the pre-training procedure of the previously introduced neural architectures to handle multilingual conversations?*

Another possible extension of monolingual spoken dialogue classifiers is to include multimodality. Multimodality is a setting of high interest when dealing with emotion recognition as an emitted sentence can be perceived in a different manner according to the tone of the speaker or his/her visual expression and thus multi-modal cues can be used to better classify the input utterance. Current multimodal architectures join information from two or more modalities into one synthetic representation. They often rely on a fusion mechanism performed on monomodal data coming from different sources: language is a symbolic and discrete signal while audio and visual modalities are represented by continuous signals. Hence, building a common representation is a challenging task due to the different nature of modalities. It has also been shown that the textual modality is well leveraged by current architectures and carries most of the useful information. From this point of view, we investigate the possibility of translating other modalities, *i.e* visual and audio, into the textual modality. As a preliminary study, we are interested in the use of punctuation marks to indicate prosodical information, such as intonations or pauses, and thus convey emotions in textual communication as in chats or in theatre plays. Linguists have showcased the prosodic aspect of punctuation marks thereby acknowledging their importance in human communication. However, punctuation marks are often overlooked, considered as noise when working with spoken language. Thus, it is interesting, in the case of emotion recognition for spoken dialogues to study the link between punctuation and multimodality. Hence, we aim at answering the following research questions: RQ2.2 **What is the link between punctuation and multi-modality?** which we decompose into subquestions:

1. *What multimodal information is stored into the punctuation?*
2. *How much the use of punctuation marks influence the prediction quality?*
3. *What is the link between punctuation marks and emotions? Which punctuation marks convey most of the information for emotions prediction?*

1.3 Contributions and detailed thesis organisation

In this thesis, we explore and propose new methods for spoken dialogue understanding following the latest trends in NLU with a particular focus on deep neural networks. In Part I, we start by introducing characteristics of spoken dialogue and the relevant related work for DA and E/S classification. We also present deep learning methods and models used in order to understand the contributions of this thesis. In Part II, we tackle the problem of monolingual spoken dialogue utterances classification in two scenarios: (1) when large annotated corpora are available (RQ1.1) (2) in low-resource tasks (RQ1.2). Finally in Part III, we extend the work presented in Part II to two settings: multilingual (RQ2.1) and multimodal (RQ2.2). More precisely, this dissertation is organized as follows:

[**Part I**]. In this first part, we introduce the background of spoken dialogue understanding and recall deep learning methods and architectures related to NLP which will be used in Part II and Part III.

[Chapter 2] In this chapter, we briefly introduce the spoken dialogue system and its component. Then we recall the particular characteristics of spoken dialogue such

as turns, utterances, disfluencies, dialogue act. We finally present the related work relevant for sequence labelling with a particular focus on DA and E/S classification in dialogue which are the main problems we address in this work.

[Chapter 3]: Provide an overview of deep learning concepts useful for the present work. We start by introducing the supervised approach highlighting its defects. Then, we introduce the self-supervised learning that allows to overcome the shortcomings of supervised learning. This approach is the first step of a broader framework called transfer learning for which we showcase the characteristics. Finally, we recall the neural architectures used throughout this thesis.

[Part II] The second part gathers the contributions related to the RQ1. We present two approaches to label English spoken dialogue (1) in large annotated corpora and (2) in a data scarcity setting.

[Chapter 4]: This first model targets only DA classification problems as it requires large corpora (e.g. SwDA and MRDA) that are only labelled in dialogue acts. In this approach, we cast the DA classification problem in spoken dialogue as a Natural Machine Translation NMT problem. We successfully leveraged the well established Seq2Seq framework to model global dependency whereas models relying on the CRF based decoder are better suited for local dependencies. We introduced a hierarchical encoder tailored for the dialogue structure, a novel guided attention mechanism and beam search applied at both training and inference time. Compared to former models, our approach does not require any handcrafted features and is trained end-to-end. We show such an approach achieves competitive results on MRDA and SwDA.

[Chapter 5]: The methodology in Chapter 4 presents several shortcomings: 1) the architecture is based on non-contextual word embeddings 2) encoder and decoder are based on GRU cells making them hard to train 3) the model requires a huge amount of labelled data that are not always available. Hence, in this chapter, we propose a pretrained hierarchical encoder based on a transformer architecture. We adapt existing pretraining objectives, such as Masked Language Model MLM or to this new architecture. Acknowledging the discrepancy between written and spoken dialogue data, we also chose a pretraining dataset better suited for the latter. Our representations are tested on a new benchmark called Sequence labelling evaLuatIon benChmark fOr spoken laNguagE benchmark (SILICONE) which gathered 10 broadly used datasets labelled in E/S or DA. We demonstrate that such an approach allows achieving competitive results while reducing the number of parameters.

[Part III] In this third part, we gather the contributions related to the RQ2. We address problems of spoken dialogue classification exposed in Part II within two extended settings: multilingual and multimodal.

[Chapter 6]: So far, only english data have been considered. However, spoken dialogue systems should be able to handle multiple languages and multilinguality within a conversation. In this chapter, based on the architecture presented in Chapter 5, we present new pretraining losses tailored to learn multilingual spoken dialogue representations. These losses aim at exposing the model to code switched language. We automatically build a multilingual pretraining corpus composed of multilingual conversations in 5 European languages (English, French, Spanish, Ital-

ian and German). We then test our representations on a new benchmark called MIAM for **M**ultilingual d**I**alog**A**ct bench**M**ark which gathered one dataset labelled in DA for each aforementioned language. Additionally, we provide two new evaluation tasks: contextual inconsistency detection and next utterance retrieval with both monolingual and multilingual input context.

[Chapter 7]: In the final chapter, we focus on emotion recognition in spoken data. So far our work has relied mostly on the textual modality. However several works have acknowledged the multimodal aspect of communication building complex fusion mechanisms to leverage these different canals of information. Yet, latest mutlimodal SOTA architectures have showcased the dominant aspect of textual modality in performance results. We investigate the possibility to translate non textual modalities, *i.e.* audio and visual, into the textual modality. As a starting point, we focus on punctuation marks as it is a simple and easily accessible textual representation of prosodical cues. Indeed, when the communication modality is the text, such as in chats or in theatre plays, the writers use punctuation marks to indicate intonations or pauses and therefore to convey emotions. However, punctuation marks are often reduced to their semantic aspect and removed in most NLP tasks. In this chapter, we investigate the role played by punctuation marks in an emotion recognition task with current neural network architectures.

The following references summarize the published contributions of this thesis:

Conferences:

- **E. Chapuis***, P. Colombo*, M. Labeau, and C. Clavel. Code-switched inspired losses for generic spoken dialog representations. **EMNLP 2021.**
- P. Colombo, **E. Chapuis**, M. Labeau, and C. Clavel. Improving multimodal fusion via mutual dependency maximisation. Submitted at **EMNLP 2021.**
- **E. Chapuis***, P. Colombo*, M. Manica, M. Labeau, and C. Clavel. Hierarchical pre-training for sequence labelling in spoken dialog. **Findings of EMNLP 2020.**
- P. Colombo*, **E. Chapuis***, M. Manica, E. Vignon, G. Varni, and C. Clavel. Guiding attention in sequence-to-sequence models for dialogue act prediction. (oral) **AAAI 2020.**

Preprints

- **E. Chapuis**, M. Labeau, and C. Clavel. Multimodal Nature of Punctuation: Application To Emotion Recognition.

Part I

Background

Part I Abstract

This part gathers the related work useful to understand our contributions related to RQ1 and RQ2. The background is divided into two chapters and is structured as follows:

- in Chapter 2, we provide an informal introduction to fundamental concepts linked to spoken dialogue. This chapter presents the particular characteristics of spoken dialogue (*e.g* disfluencies, dialogue act, grounding) with a particular focus on the aspect that we will leverage in the next part of the thesis. The last part of this chapter is dedicated to the related work relevant to the sequence labelling problems which is one of the core problems we address in this thesis.
- in Chapter 3, we offer an introduction to deep learning concepts useful for this thesis. Although described pre-training methods have been tested in other settings (*e.g* written text, chat), we will adapt them in Part II so they can be used to better handle spoken dialogue. This chapter also includes related work dedicated to multilingual and multimodal learning that will be useful to understand Part III.

Chapter 2

Spoken Dialogue Understanding

Chapter 2 Abstract

This thesis focuses on spoken dialogue understanding. Specifically we tackle the problem of spoken dialogues classification, DA and E/S labels in particular. Hence, this chapter aims at presenting the most important concept and methods employed in the field of spoken dialogue understanding. We start by giving an overview of spoken dialogue systems. We review some of the core linguistics concepts relevant to understand the specificity of a conversation compared to an ordinary text: turn taking, grounding, dialogue act, disfluency, code-switching. Lastly we recall the related works needed to address DA and E/S classification problems presented in RQ1 as well as background related to the two extension scenarios presented in RQ2.

2.1 Dialogue Systems Overview

Conversational AI aims at developing dialogue systems, that are not only capable of conversing with human but also answering a diverse range of questions or realize complex tasks such as travel planning. Conversational AI is at the junction of multiple areas of research *e.g.* Natural language processing (NLP), Automatic speech recognition (ASR), Machine Learning ML, reinforcement learning (RL), Linguistic, Psychology [ARORA and collab., 2013; GAO and collab., 2019]. Due to its promising potential commercial values and the great advances in deep learning and reinforcement learning, it has received a lot of attention in recent years. Dialogue systems are often divided into two categories [CHEN and collab., 2017; GAO and collab., 2019]: (1) task-oriented systems and (2) non-task-oriented systems (also known as chatbots). The first helps the user to complete a certain task (*i.e.* purchasing products, booking a reservation) hence the interaction is limited by the task itself, whereas non-task oriented systems communicate with human users in order to provide an engaging conversation and entertainment. This work is in line with the latter category. Indeed, we aim at providing neural systems that are able to label emitted utterances during the conversation in DA and/or E/S. These are key information that will allow the dialogue system to generate adequate responses that retain the user's interest.
Typically, dialogue systems follow the structure presented in Figure 2.1 [ARORA and collab., 2013].

- **Input Encoder** This block converts the input into textual data. With a textual dialogue system, the textual content is directly provided by the user and this component is therefore not needed. In the case of spoken dialogue systems, the input encoder turns a speech signal into its corresponding transcript.
- **Natural Language Understanding** It maps emitted utterances with semantic labels which are then used by the dialogue manager. In our case, the NLU block provides dialogue act and emotion/sentiment labels associated with each emitted utterance.
- **Dialogue Manager** This component tracks the state and flow of the conversation and decides the next action to be taken by the system.
- **Natural Language Generation** This component generates the best response according to information provided by the dialogue manager.
- **Output Renderer** This final component renders the generated utterance.

In this thesis, we focus in particular on the NLU component. We also notice that in recent years fully data-driven and end-to-end approaches have been studied to conversational response generation [OLABIYI and MUELLER, 2019; SHANG and collab., 2015; SORDONI and collab., 2015; VINYALS and LE, 2015; ZHANG and collab., 2019b]. Such systems follow the seq2seq framework [SUTSKEVER and collab., 2014] and directly provide a response without relying on the aforementioned components of the traditional dialogue systems.

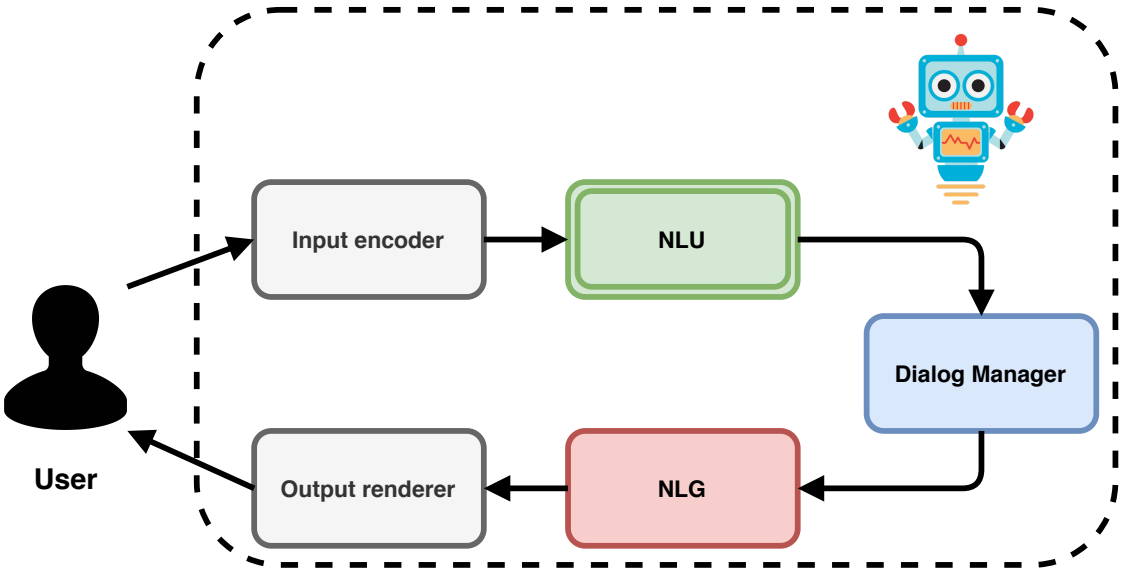


Figure 2.1 – Components of traditional dialogue systems.

2.2 Characteristics of Spoken Dialogues

As we aim at developing dialogue systems that emulate the human conversation, it is essential to understand how humans dialogue with each other. In this section, we introduce the key concepts employed in the field of spoken dialogue analysis that showcase the dynamics and structure of a dialogue.

Dialogue

Overall, a dialogue is a complex and collective activity and is often defined as follows:

Definition 2.2.1 (Dialogue)**.** talk between two (*dyadic*) or more people (*multiparty*) in which thoughts, feelings, and ideas are expressed, questions are asked and answered, or news and information are exchanged.

A dialogue can also be presented as a restricted form of conversation limited to two participants with an aim and a central topic. Alternatively, a conversation gathers a group of people that communicate without any specific goal and are somewhat more tentative. Moreover, a conversation may include change of topics. Still, since there is no consensus on the distinction between dialogue and conversation, we will use them indifferently.

Turn

A dialogue is structured by turns which are single contributions by one speaker to the dialogue. A change of turn occurs when the roles of the speaker and listener are exchanged. A turn can be either a whole sentence, several sentences or simply a word. For instance, Table 2.1 presents a dialogue between the two characters of the Samuel Becket piece *Waiting for Godo*. Since the example comes from a written play, the dialogue is segmented by turns that gather one speaker's whole speech. However, in spoken dialogue analysis, a turn is often divided into smaller units of speech that we present in the following.

Utterances

Utterances are often referred as *"the smallest unit of speech. It is a continuous span of speech beginning and ending with a clear pause"*. Over the years, many different definitions of utterances have been proposed and make use of one or more of the following factors [Traum and Heeman, 1997]:

- Speech by a single speaker, speaking without interruption by speech of the other, constituting a single Turn.
- Defines a single dialogue act.
- Has syntactic and/or semantic completion.
- Is an intonational phrase.
- Separated by a clear pause.

In spoken dialogue systems, utterances are not given and come from ASR system. Finding the beginning and the end of an utterance is a challenging problem [ANG and collab., 2005; ZIMMERMANN and collab., 2006]. Efficient and effective segmentation of spoken language remains an open problem in spoken dialogue. As an example, in Table 2.2 we can see (1) several consecutive utterances from the same speaker (2) the variety of utterances (*e.g.* "so", ".", "Uh-huh", "I'm not real").
Dialogue Act: dialogue acts (DA) or speech acts are a key component to analyse dialogue. [AUSTIN, 1962; SEARLE, 1969] explained that utterances emitted in a dialogue are not purposeless and rather described them as actions perform by the speakers. These actions are called dialogue acts and it becomes possible to characterise and classify these actions. [SEARLE, 1975] established a taxonomy of DA divided into five categories :

1. **Assertives**: Committing the speaker to the truth of a proposition.
2. **Directives**: Attempts by the speaker to get the addressee to do something.
3. **Commissives**: Committing the speaker to some future course of action.
4. **Expressives**: Expressing the psychological state of the speaker about a state of affairs.
5. **Declaratives**: Bringing about a different state of the world by the utterance.

Other annotation schemes are available with more detailed categories such as DAMSL (Dialogue Act Markup in Several Layers) introduced by [ALLEN and CORE, 1997]. Moreover there are numerous datasets annotated in dialogue acts such as Switchboard [GODFREY and collab., 1992], MRDA [SHRIBERG and collab., 2004], Dailydialog [LI and collab., 2017], each one with their own annotation scheme but, nevertheless, based on the DAMSL scheme.
Disfluency: as previously mentioned, in spoken dialogue systems the inputs comes from the user's speech. Spoken language is rarely fluent; the speaker's flow is often interrupted with disfluencies such as pause silence, self-correcting, repeating words or fillers (*e.g.* "um" or "uh") that fill pauses within an utterance during a conversation [SHRIBERG, 1994]. [DINKAR and collab., 2020] shows that representing fillers words helps to predict a speaker's stance and expressed confidence.
Code switching (CS): is a linguistic phenomena that has been observed with bilingual speakers, it is defined as the capacity of a speaker to alternate languages, effortlessly, within a conversation or speech. For instance, a well-known example from [POPLACK, 1980] of CS between English and Spanish:

> "Sometimes I'll start a sentence in Spanish **y termino en espanol**[sic] (and finish in Spanish)"

Utterance	Speaker
I'm curious to hear what he has to offer. Then we'll take it or leave it.	Vladimir
What exactly did we ask him for?	Estragon
Were you not there?	Vladimir
I can't have been listening.	Estragon
Oh . . . Nothing very definite.	Vladimir
A kind of prayer.	Estragon
Precisely.	Vladimir
A vague supplication.	Estragon
Exactly.	Vladimir

Table 2.1 – Extract of dialogue from *Waiting for Godot*

Utterance	Speaker
So I 've been concerned about crime lately .	A
Uh-huh.	B
Uh , it 's really scary to listen to the news every night and –	A
Uh-huh .	B
to hear about all the problems .	A
I wondered if you were taking any special precautions in your neighborhood ?	A
Well , I , I think we have a neighborhood watch .	B
Uh-huh .	A
I think .	B
.	A
I 'm not real,	B
we do n't get real involved.	B
We 're never home,	B
so	B
Uh-huh	A

Table 2.2 – Example of dialogue taken from the Switchboard Dialog Act Corpus.

CS is considered as an important communication strategy and has been studied extensively in linguistics, especially as a speech phenomenon [AUER, 2013; GARDNER-CHLOROS and collab., 2009; MYERS-SCOTTON and COULMAS, 1997; POPLACK, 1980]. [POPLACK, 1980] has shown that CS may occur in different ways which are divided into three categories :

- **Extra-sentencial:** also called *tag-switching* is the inserting of tag elements from one language into a monolingual discourse, *e.g.* "you don't know how to speak Spanish, **¿verdad?** (right ?)"
- **Intra-sentencial:** the switch from one language to the other occurs outside a sentence, *e.g* : "Le dije que no queria comprar el carro (I told him I didn't want to buy the car). **He got really mad**"
- **Inter-sentencial:** the switch from one language to the other take place within a single utterance *e.g.* "¡ Ay, qué **cute** se ve !".

In recent years CS has also drawn the attention of the NLP community [ÇETINOGLU and collab., 2016]. For instance [AGUILAR and SOLORIO, 2020] provides an extension

of the ELMo [PETERS and collab., 2018] language model that leverages CS phenomena, in [PAREKH and collab., 2020] authors propose a bilingual collaborative dialogue system that outputs code-switching conversations in a controlled manner. Benchmarks have also been built in order to provide a unified platform to evaluate CS data, such as Linguistic Codeswitching Evaluation (LinCE)[AGUILAR and collab., 2020] and GLUECoS [KHANUJA and collab., 2020]. In Chapter 6 we present a multilingual system where pre-training losses and evaluation tasks are inspired from the CS phenomena. In addition, spoken language presents a variety of specific phenomena that differentiate it from written communication such as surface formality change, lexical diversity and grammatical complexity and accuracy [CHAFE and TANNEN, 1987; REDEKER, 1984]. This discrepancy between written and spoken language is a core motivation of our work presented Chapter 5. Furthermore, we explore this difference in Chapter 7 through the lens of punctuation marks.

2.3 Related Work on Spoken Dialogue Understanding

In this section, we gather the related work related to spoken dialogue understanding. This section is organized as follows: we start by describing previous systems used for DA classification, then we gather the relevant related work for E/S classification.

2.3.1 DA classification

Several approaches have been proposed to tackle the DA classification problem. These methods can be divided into two different categories. The first class of methods relies on the independent classification of each utterance using various techniques, such as HMM [STOLCKE and collab., 2000], SVM [SURENDRAN and LEVOW, 2006] and Bayesian Network [KEIZER and collab., 2002]. The second class, which achieves better performance, leverages the context, that relies on the sequence of previously emitted utterances, to improve the classifier performance by using deep learning approaches to capture contextual dependencies between input sentences [BOTHE and collab., 2018; KHANPOUR and collab., 2016]. Another refinement of input context-based classification is the modelling of inter-tag dependencies. This task is tackled as sequence-based classification where output tags are considered as a DA sequence [CHEN and collab., 2018; KUMAR and collab., 2018; LI and collab., 2018; RAHEJA and TETREAULT, 2019; STOLCKE and collab., 2000]. In the latter works the tags sequence is decoded using Conditional Random Features we present in the Chapter 3. In Chapter 4 we showcase a Seq2Seq based model...

2.3.2 Emotion/Sentiment recognition

The study of emotion and sentiment in human communication by intelligent systems is a whole cross-disciplinary research field called *affective computing*. It aims machines to recognize, predict, and interpret human emotions in order to make emotional responses. It covers a wide range of areas such as computer science, AI, cognitive science, neuroscience, neuropsychology and social science. In this chapter, we recall deep learning approaches that have been developed to tackle E/S prediction in spoken dialogue. Currently, these approaches can be divided into two main lines of research:

Emotion Recogniton in Conversation ERC: with recent public available datasets (*e.g.*

MELD [PORIA and collab., 2018], DailyDialog [LI and collab., 2017], SEMAINE[MCKEOWN and collab., 2012]) ERC has gained a lot of attention from the NLP community [MAJUMDER and collab., 2018; ZHOU and collab., 2017]. As in current DA tagging systems, recent ERC models benefit from the advances in deep learning and aim at leveraging the contextual dependencies of the target utterance [HAZARIKA and collab., 2018; JIAO and collab., 2019; MAJUMDER and collab., 2018]. Efforts have been made to carefully model the dependency graph of utterances [GHOSAL and collab., 2019; SHEN and collab., 2020, 2021]. Other works also benefit from finer modelling of the speaker personality [LI and collab., 2020; MAJUMDER and collab., 2018; SHEN and collab., 2020; ZHANG and collab., 2019a].
However, these works often rely on written based pre-training models (*e.g.* BERT, RoBERTa) as a feature extractor at the utterance level which is a defect we address in Chapter 5.

Multimodal Emotion Recognition Humans employ three different modalities to communicate in a coordinated manner: the language modality with the use of words and sentences, the visual modality with gestures, poses and facial expressions and the acoustic modality through changes in vocal tones. Multimodal representation learning has shown great progress in a large variety of tasks including emotion recognition, sentiment analysis [SOLEYMANI and collab., 2017], speaker trait analysis [PARK and collab., 2014] and fine-grained opinion mining [GARCIA and collab., 2019a].
Learning from different modalities is an efficient way to improve performance on the target tasks [XU and collab., 2013]. Nevertheless, heterogeneities across modalities increase the difficulty of learning multimodal representations and raise specific challenges. BALTRUŠAITIS and collab. [2018] identifies five core challenges for multimodal learning:

- **Fusion:** join information from two or more modalities to perform a prediction.
- **Representation:** learn how to represent and summarize multimodal data.
- **Modality alignment:** identify the direct relations between (sub)elements from two or more different modalities.
- **Translation:** map data from one modality to another.
- **Co-learning:** transfer knowledge between modalities, their representation, and their predictive models.

In Chapter 7 we will have a particular look at the fusion challenge. Multimodal fusion can be divided into early and late fusion techniques: early fusion takes place at the feature level [YE and collab., 2017], while late fusion takes place at the decision or scoring level [KHAN and collab., 2012]. A plethora of architecture with new fusion mechanisms relying on deep architectures have been proposed to tackle the problem of multimodal learning. Perhaps the most known includes TFN [ZADEH and collab., 2017], LFN [LIU and collab., 2018], MARN [ZADEH and collab., 2018a], MISA [HAZARIKA and collab., 2020], MCTN [PHAM and collab., 2019], HFNN [MAI and collab., 2019], ICCN [SUN and collab., 2020]). These models are evaluated on several multimodal sentiment analysis benchmarks such as IEMOCAP [BUSSO and collab., 2008], MOSI [WÖLLMER and collab., 2013], MOSEI [ZADEH and collab., 2018b] and POM [GARCIA and collab., 2019b; PARK and collab., 2014]. The current state-of-the-art on these datasets uses architectures based on pre-trained transformers [SIRIWARDHANA and collab., 2020; TSAI and collab., 2019] such as MultiModal Bert (MAGBERT) or

MultiModal XLNET (MAGXLNET) [RAHMAN and collab., 2020]. DAI and collab. [2021] pointed out that multimodal systems rely on handcrafted features for visual and audio and propose a end-to-end architecture to process all the modalities. Lastly, it has been showcased that the textual modality carry most of the performances [DAI and collab., 2021; RAHMAN and collab., 2019]. The symbolic nature of language is well processed by neural model especially with the rise of distributed representation while audio and visual modalities, represented as signals, seems less well handle (especially with handcrafted features) in the scope of emotion recognition. In Chapter 7, we examine the possibility to translate the audio signal into a symbolic representation via punctuation marks and provide a quantitative study.

Chapter 2 Conclusion

In this chapter, we introduced the characteristics of spoken dialogue and presented the two problems of NLU we focus on this thesis: DA and E/S classification. We presented the related work associated with RQ1 and RQ2. In the next chapter, we will recall the deep learning concepts and architectures useful for this thesis.

2.4 References

AGUILAR, G., S. KAR and T. SOLORIO. 2020, "Lince: A centralized benchmark for linguistic code-switching evaluation", *arXiv preprint arXiv:2005.04322*. 32

AGUILAR, G. and T. SOLORIO. 2020, "From English to code-switching: Transfer learning with strong morphological clues", in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Online, p. 8033–8044, doi: 10.18653/v1/2020.acl-main.716. URL https://aclanthology.org/2020.acl-main.716. 31

ALLEN, J. and M. CORE. 1997, "Draft of DAMSL: Dialog act markup in several layers", Unpublished manuscript. 30

ANG, J., Y. LIU and E. SHRIBERG. 2005, "Automatic dialog act segmentation and classification in multiparty meetings", in *Proceedings. (ICASSP '05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, vol. 1, p. I/1061–I/1064 Vol. 1, doi: 10.1109/ICASSP.2005.1415300. 30

ARORA, S., K. BATRA and S. SINGH. 2013, "Dialogue system: A brief review", *CoRR*, vol. abs/1306.4134. URL http://arxiv.org/abs/1306.4134. 28

AUER, P. 2013, *Code-switching in conversation: Language, interaction and identity*, Routledge. 31

AUSTIN, J. L. 1962, *How to do things with words*, William James Lectures, Oxford University Press. URL http://scholar.google.de/scholar.bib?q=info:xI2JvixH8_QJ:scholar.google.com/&output=citation&hl=de&as_sdt=0,5&ct=citation&cd=1. 30

BALTRUŠAITIS, T., C. AHUJA and L.-P. MORENCY. 2018, "Multimodal machine learning: A survey and taxonomy", *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, nº 2, p. 423–443. 33

BOTHE, C., C. WEBER, S. MAGG and S. WERMTER. 2018, "A context-based approach for dialogue act recognition using simple recurrent neural networks", *CoRR*, vol. abs/1805.06280. 32

BUSSO, C., M. BULUT, C.-C. LEE, A. KAZEMZADEH, E. MOWER, S. KIM, J. N. CHANG, S. LEE and S. S. NARAYANAN. 2008, "Iemocap: Interactive emotional dyadic motion capture database", *Language resources and evaluation*, vol. 42, nº 4, p. 335. 33

ÇETINOGLU, Ö., S. SCHULZ and N. T. VU. 2016, "Challenges of computational processing of code-switching", *CoRR*, vol. abs/1610.02213. URL http://arxiv.org/abs/1610.02213. 31

CHAFE, W. and D. TANNEN. 1987, "The relation between written and spoken language", *Annual review of anthropology*, vol. 16, nº 1, p. 383–407. 32

CHEN, H., X. LIU, D. YIN and J. TANG. 2017, "A survey on dialogue systems: Recent advances and new frontiers", *CoRR*, vol. abs/1711.01731. URL http://arxiv.org/abs/1711.01731. 28

CHEN, Z., R. YANG, Z. ZHAO, D. CAI and X. HE. 2018, "Dialogue act recognition via crf-attentive structured network", in *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, ACM, p. 225–234. 32

DAI, W., S. CAHYAWIJAYA, Z. LIU and P. FUNG. 2021, "Multimodal end-to-end sparse model for emotion recognition", in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Association for Computational Linguistics, Online, p. 5305–5316, doi: 10.18653/v1/2021.naacl-main.417. URL https://aclanthology.org/2021.naacl-main.417. 34

DINKAR, T., P. COLOMBO, M. LABEAU and C. CLAVEL. 2020, "The importance of fillers for text representations of speech transcripts", *CoRR*, vol. abs/2009.11340. URL https://arxiv.org/abs/2009.11340. 30

GAO, J., M. GALLEY and L. LI. 2019, "Neural approaches to conversational ai", . 28

GARCIA, A., P. COLOMBO, S. ESSID, F. D'ALCHÉ BUC and C. CLAVEL. 2019a, "From the token to the review: A hierarchical multimodal approach to opinion mining", *arXiv preprint arXiv:1908.11216*. 33

GARCIA, A., S. ESSID, F. D'ALCHÉ BUC and C. CLAVEL. 2019b, "A multimodal movie review corpus for fine-grained opinion mining", *arXiv preprint arXiv:1902.10102*. 33

GARDNER-CHLOROS, P. and collab.. 2009, *Code-switching*, Cambridge university press. 31

GHOSAL, D., N. MAJUMDER, S. PORIA, N. CHHAYA and A. F. GELBUKH. 2019, "Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation", *CoRR*, vol. abs/1908.11540. URL http://arxiv.org/abs/1908.11540. 33

GODFREY, J., E. HOLLIMAN and J. MCDANIEL. 1992, "Switchboard: telephone speech corpus for research and development . acoustics,", in *IEEE International Conference on Speech, and Signal Processing, ICASSP-92*, vol. 1, p. 517–520. 30

HAZARIKA, D., S. PORIA, A. ZADEH, E. CAMBRIA, L.-P. MORENCY and R. ZIMMERMANN. 2018, "Conversational memory network for emotion recognition in dyadic dialogue videos", in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, Association for Computational Linguistics, New Orleans, Louisiana, p. 2122–2132, doi: 10.18653/v1/N18-1193. URL https://aclanthology.org/N18-1193. 33

HAZARIKA, D., R. ZIMMERMANN and S. PORIA. 2020, "Misa: Modality-invariant and-specific representations for multimodal sentiment analysis", *arXiv preprint arXiv:2005.03545*. 33

JIAO, W., H. YANG, I. KING and M. R. LYU. 2019, "Higru: Hierarchical gated recurrent units for utterance-level emotion recognition", *CoRR*, vol. abs/1904.04446. URL http://arxiv.org/abs/1904.04446. 33

KEIZER, S., R. OP DEN AKKER and A. NIJHOLT. 2002, "Dialogue act recognition with bayesian networks for dutch dialogues", in *Proceedings of the Third SIGdial Workshop on Discourse and Dialogue*. 32

KHAN, F. S., R. M. ANWER, J. VAN DE WEIJER, A. D. BAGDANOV, M. VANRELL and A. M. LOPEZ. 2012, "Color attributes for object detection", in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, p. 3306–3313. 33

KHANPOUR, H., N. GUNTAKANDLA and R. NIELSEN. 2016, "Dialogue act classification in domain-independent conversations using a deep recurrent neural network", in *COLING*. 32

KHANUJA, S., S. DANDAPAT, A. SRINIVASAN, S. SITARAM and M. CHOUDHURY. 2020, "GLUECoS: An evaluation benchmark for code-switched NLP", in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Online, p. 3575–3585, doi: 10.18653/v1/2020.acl-main.329. URL https://aclanthology.org/2020.acl-main.329. 32

KUMAR, H., A. AGARWAL, R. DASGUPTA and S. JOSHI. 2018, "Dialogue act sequence labeling using hierarchical encoder with crf", in *Thirty-Second AAAI Conference on Artificial Intelligence*. 32

LI, J., D. JI, F. LI, M. ZHANG and Y. LIU. 2020, "HiTrans: A transformer-based context- and speaker-sensitive model for emotion detection in conversations", in *Proceedings of the 28th International Conference on Computational Linguistics*, International Committee on Computational Linguistics, Barcelona, Spain (Online), p. 4190–4200, doi: 10.18653/v1/2020.coling-main.370. URL https://aclanthology.org/2020.coling-main.370. 33

LI, R., C. LIN, M. COLLINSON, X. LI and G. CHEN. 2018, "A dual-attention hierarchical recurrent neural network for dialogue act classification", *CoRR*. 32

LI, Y., H. SU, X. SHEN, W. LI, Z. CAO and S. NIU. 2017, "Dailydialog: A manually labelled multi-turn dialogue dataset", *CoRR*, vol. abs/1710.03957. URL http://arxiv.org/abs/1710.03957. 30, 33

LIU, Z., Y. SHEN, V. B. LAKSHMINARASIMHAN, P. P. LIANG, A. ZADEH and L.-P. MORENCY. 2018, "Efficient low-rank multimodal fusion with modality-specific factors", *arXiv preprint arXiv:1806.00064*. 33

MAI, S., H. HU and S. XING. 2019, "Divide, conquer and combine: Hierarchical feature fusion network with local and global perspectives for multimodal affective computing", in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, p. 481–492. 33

MAJUMDER, N., S. PORIA, D. HAZARIKA, R. MIHALCEA, A. F. GELBUKH and E. CAMBRIA. 2018, "Dialoguernn: An attentive RNN for emotion detection in conversations", *CoRR*, vol. abs/1811.00405. URL http://arxiv.org/abs/1811.00405. 33

MCKEOWN, G., M. VALSTAR, R. COWIE, M. PANTIC and M. SCHRODER. 2012, "The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent", *IEEE Transactions on Affective Computing*, vol. 3, nº 1, doi: 10.1109/T-AFFC.2011.20, p. 5–17. 33

MYERS-SCOTTON, C. and D. COULMAS. 1997, "Code-switching", . 31

OLABIYI, O. and E. T. MUELLER. 2019, "Multi-turn dialogue response generation with autoregressive transformer models", *CoRR*, vol. abs/1908.01841. URL http://arxiv.org/abs/1908.01841. 28

PAREKH, T., E. AHN, Y. TSVETKOV and A. W. BLACK. 2020, "Understanding linguistic accommodation in code-switched human-machine dialogues", in *Proceedings of the 24th Conference on Computational Natural Language Learning*, Association for Computational Linguistics, Online, p. 565–577, doi: 10.18653/v1/2020.conll-1.46. URL https://aclanthology.org/2020.conll-1.46. 32

PARK, S., H. S. SHIM, M. CHATTERJEE, K. SAGAE and L.-P. MORENCY. 2014, "Computational analysis of persuasiveness in social multimedia: A novel dataset and multimodal prediction approach", in *Proceedings of the 16th International Conference on Multimodal Interaction*, p. 50–57. 33

PETERS, M. E., M. NEUMANN, M. IYYER, M. GARDNER, C. CLARK, K. LEE and L. ZETTLEMOYER. 2018, "Deep contextualized word representations", *CoRR*, vol. abs/1802.05365. URL http://arxiv.org/abs/1802.05365. 32

PHAM, H., P. P. LIANG, T. MANZINI, L.-P. MORENCY and B. PÓCZOS. 2019, "Found in translation: Learning robust joint representations by cyclic translations between modalities", in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, p. 6892–6899. 33

POPLACK, S. 1980, "Sometimes i'll start a sentence in spanish y termino en espaÑol: toward a typology of code-switching 1", *Linguistics*, vol. 18, doi: 10.1515/ling.1980.18.7-8.581, p. 581–618. 30, 31

PORIA, S., D. HAZARIKA, N. MAJUMDER, G. NAIK, E. CAMBRIA and R. MIHALCEA. 2018, "MELD: A multimodal multi-party dataset for emotion recognition in conversations", *CoRR*, vol. abs/1810.02508. URL http://arxiv.org/abs/1810.02508. 33

RAHEJA and TETREAULT. 2019, "Dialogue act classification with context-aware self-attention", *CoRR*, vol. abs/1904.02594. URL http://arxiv.org/abs/1904.02594. 32

RAHMAN, W., M. K. HASAN, S. LEE, A. B. ZADEH, C. MAO, L.-P. MORENCY and E. HOQUE. 2020, "Integrating multimodal information in large pretrained transformers", in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 2359–2369. 34

RAHMAN, W., M. K. HASAN, A. ZADEH, L. MORENCY and M. E. HOQUE. 2019, "M-BERT: injecting multimodal information in the BERT structure", *CoRR*, vol. abs/1908.05787. URL http://arxiv.org/abs/1908.05787. 34

REDEKER, G. 1984, "On differences between spoken and written language", *Discourse processes*, vol. 7, nº 1, p. 43–55. 32

SEARLE, J. R. 1969, *Speech Acts: An Essay in the Philosophy of Language*, Cambridge University Press, Cambridge. 30

SEARLE, J. R. 1975, "A taxonomy of illocutionary acts", in *Language, Mind and Knowledge*, édité par K. Gunderson, University of Minnesota Press, p. 344–369. 30

SHANG, L., Z. LU and H. LI. 2015, "Neural responding machine for short-text conversation", *CoRR*, vol. abs/1503.02364. URL http://arxiv.org/abs/1503.02364. 28

SHEN, W., J. CHEN, X. QUAN and Z. XIE. 2020, "Dialogxl: All-in-one xlnet for multi-party conversation emotion recognition", *CoRR*, vol. abs/2012.08695. URL https://arxiv.org/abs/2012.08695. 33

SHEN, W., S. WU, Y. YANG and X. QUAN. 2021, "Directed acyclic graph network for conversational emotion recognition", *CoRR*, vol. abs/2105.12907. URL https://arxiv.org/abs/2105.12907. 33

SHRIBERG, E., R. DHILLON, S. BHAGAT, J. ANG and H. CARVEY. 2004, "The icsi meeting recorder dialog act (mrda) corpus", in *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue at HLT-NAACL 2004*. 30

SHRIBERG, E. E. 1994, *Preliminaries to a theory of speech disfluencies*, thèse de doctorat, Citeseer. 30

SIRIWARDHANA, S., A. REIS, R. WEERASEKERA and S. NANAYAKKARA. 2020, "Jointly fine-tuning" bert-like" self supervised models to improve multimodal speech emotion recognition", *arXiv preprint arXiv:2008.06682*. 33

SOLEYMANI, M., D. GARCIA, B. JOU, B. SCHULLER, S.-F. CHANG and M. PANTIC. 2017, "A survey of multimodal sentiment analysis", *Image and Vision Computing*, vol. 65, p. 3–14. 33

SORDONI, A., M. GALLEY, M. AULI, C. BROCKETT, Y. JI, M. MITCHELL, J. NIE, J. GAO and B. DOLAN. 2015, "A neural network approach to context-sensitive generation of conversational responses", *CoRR*, vol. abs/1506.06714. URL http://arxiv.org/abs/1506.06714. 28

STOLCKE, A., K. RIES, N. COCCARO, E. SHRIBERG, R. BATES, D. JURAFSKY, P. TAYLOR, R. MARTIN, C. V. ESS-DYKEMA and M. METEER. 2000, "Dialogue act modeling for automatic tagging and recognition of conversational speech", *Computational linguistics*, vol. 26, nº 3, p. 339–373. 32

SUN, Z., P. SARMA, W. SETHARES and Y. LIANG. 2020, "Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis", in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, p. 8992–8999. 33

SURENDRAN, D. and G.-A. LEVOW. 2006, "Dialog act tagging with support vector machines and hidden markov models", in *Ninth International Conference on Spoken Language Processing*. 32

SUTSKEVER, I., O. VINYALS and Q. V. LE. 2014, "Sequence to sequence learning with neural networks", *CoRR*, vol. abs/1409.3215. URL http://arxiv.org/abs/1409.3215. 28

TRAUM, D. R. and P. A. HEEMAN. 1997, "Utterance units in spoken dialogue", in *Dialogue Processing in Spoken Language Systems*, édité par E. Maier, M. Mast and S. LuperFoy, Springer Berlin Heidelberg, Berlin, Heidelberg, p. 125–140. 29

TSAI, Y.-H. H., S. BAI, P. P. LIANG, J. Z. KOLTER, L.-P. MORENCY and R. SALAKHUTDINOV. 2019, "Multimodal transformer for unaligned multimodal language sequences", in *Proceedings of the conference. Association for Computational Linguistics. Meeting*, vol. 2019, NIH Public Access, p. 6558. 33

VINYALS, O. and Q. V. LE. 2015, "A neural conversational model", *CoRR*, vol. abs/1506.05869. URL http://arxiv.org/abs/1506.05869. 28

WÖLLMER, M., F. WENINGER, T. KNAUP, B. SCHULLER, C. SUN, K. SAGAE and L.-P. MORENCY. 2013, "Youtube movie reviews: Sentiment analysis in an audio-visual context", *IEEE Intelligent Systems*, vol. 28, nº 3, p. 46–53. 33

XU, C., D. TAO and C. XU. 2013, "A survey on multi-view learning", *arXiv preprint arXiv:1304.5634*. 33

YE, J., H. HU, G.-J. QI and K. A. HUA. 2017, "A temporal order modeling approach to human action recognition from multimodal sensor data", *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 13, nº 2, p. 1–22. 33

ZADEH, A., M. CHEN, S. PORIA, E. CAMBRIA and L.-P. MORENCY. 2017, "Tensor fusion network for multimodal sentiment analysis", *arXiv preprint arXiv:1707.07250*. 33

ZADEH, A., P. P. LIANG, S. PORIA, P. VIJ, E. CAMBRIA and L.-P. MORENCY. 2018a, "Multi-attention recurrent network for human communication comprehension", in *Proceedings of the... AAAI Conference on Artificial Intelligence. AAAI Conference on Artificial Intelligence*, vol. 2018, NIH Public Access, p. 5642. 33

ZADEH, A. B., P. P. LIANG, S. PORIA, E. CAMBRIA and L.-P. MORENCY. 2018b, "Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph", in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, p. 2236–2246. 33

ZHANG, D., L. WU, C. SUN, S. LI, Q. ZHU and G. ZHOU. 2019a, "Modeling both context- and speaker-sensitive dependence for emotion detection in multi-speaker conversations", p. 5415–5421, doi: 10.24963/ijcai.2019/752. 33

ZHANG, Y., S. SUN, M. GALLEY, Y. CHEN, C. BROCKETT, X. GAO, J. GAO, J. LIU and B. DOLAN. 2019b, "Dialogpt: Large-scale generative pre-training for conversational response generation", *CoRR*, vol. abs/1911.00536. URL http://arxiv.org/abs/1911.00536. 28

ZHOU, H., M. HUANG, T. ZHANG, X. ZHU and B. LIU. 2017, "Emotional chatting machine: Emotional conversation generation with internal and external memory", *CoRR*, vol. abs/1704.01074. URL http://arxiv.org/abs/1704.01074. 33

ZIMMERMANN, M., Y. LIU, E. SHRIBERG and A. STOLCKE. 2006, "Toward joint segmentation and classification of dialog acts in multiparty meetings", in *Machine Learning for Multimodal Interaction*, édité par S. Renals and S. Bengio, Springer Berlin Heidelberg, Berlin, Heidelberg, p. 187–193. 30

Chapter 3

Deep Learning for NLU

Chapter 3 Abstract

This thesis focuses on distributed representation learning of spoken dialogue transcripts. Modern NLP models rely on deep neural architectures pre-trained in an unsupervised manner then fine-tuned on final tasks also called downstream tasks. In this chapter, we present the different components of the aforementioned framework. We start by defining the supervised approach, followed by a description of the classification task in spoken dialogue. We recall definitions of all neural architectures used throughout this thesis. Finally, we present the pretraning stage focusing on the self-supervised approach and its use in NLP.

3.1 Supervised Learning

In the supervised learning setting the goal is to build a prediction function f that maps input and output pairs (x, y) by optimizing a well chosen criterion. In our case, an example of supervised learning problem is the emotion classification as shown in Figure 3.1: from a dataset composed of utterances with their respective labels, for instance, $(Joy, Surprise, Neutral, Fear, Anger)$ the task is to build a system that labels such utterances with minimum error.

More formally, let $\mathcal{X}$ be the space of input features and $\mathcal{Y}$ the space of outputs. Depending on the nature of the output space $\mathcal{Y}$ the prediction problem be either a classification task ($\forall y \in \mathcal{Y}, y \in \{1, ..., K\}$ where $K \in \mathbb{N}$) or a regression task ($\forall y \in \mathcal{Y}, y \in \mathbb{R}$). Each sample $(x, y) \in \mathcal{X} \times \mathcal{Y}$ comes from an unknown distribution $\mathcal{P}$. We denote $\mathcal{H}$ as the family of mapping functions from $\mathcal{X}$ to $\mathcal{Y}$. Since this thesis makes use of neural networks we can limit $\mathcal{H}$ to the set of functions parameterised by a vector $\theta \in \mathbb{R}^d$ where $d \in \mathbb{N}$.
Then the risk of the predictor is defined as follows:

$$R(f) = \mathbb{E}_{(x,y)\sim\mathcal{P}}\, \mathcal{L}(f(x), y) \tag{3.1}$$

Where $\mathcal{L}$ is a non-negative loss function that measures the error made predicting $\hat{y} = f(x)$ when the correct label is y. Hence the goal is to find f^* that minimizes R. However this function cannot be computed as the input/output pairs distribution is unknown. In practice we have only access to a dataset $D = \{(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}\}_{i=1}^{N}$ called training set where each $(x, y) \sim \mathcal{P}$. Then we define the empirical risk:

$$\hat{R}(f) = \frac{1}{N} \sum_{i=0}^{N} \ell(f(x_i), y_i) \tag{3.2}$$

For classification tasks, we choose the cross-entropy loss as a function ℓ which is a combination of negative log-likelihood and softmax. We denote $\hat{y}_i \in \mathcal{R}^K$ the predicted probabilities for the K output classes for the input x_i. The f function maps the input x_i to a vector where each element is a score associated to one of the K classes, *i.e* $f : \mathcal{X} \mapsto \mathcal{R}^K$. Hence the probability associated with the class j is calculated as follows:

$$\hat{y}_{i,j} = Softmax(f(x_i))_j = \frac{e^{f(x_i)_j}}{\sum_{k=1}^{K} e^{f(x_i)_k}} \tag{3.3}$$

hence the log-likelihood L between $\hat{y}_i$ and y_i (being the one-hot encoding vector of the output label associated to x_i) is defined as follow:

$$L(\hat{y}_i, y_i) = \sum_{i=1}^{K} y_i log(\hat{y}_i) \tag{3.4}$$

It often happens that f will correspond closely to a particular set of data D meaning that f will be low for $\hat{R}$ but high for R. This situation is called overfitting. Regularisation techniques are used to prevent this phenomenon such as adding a non-negative function (e.g L_1 or L_2 norm) called penalty term to $\hat{R}$ enforcing smooth predictors. In addition, other techniques are employed with neural networks to prevent overfitting such as adding noise, dropout [Srivastava and collab., 2014], early stopping [Caruana and collab., 2000].

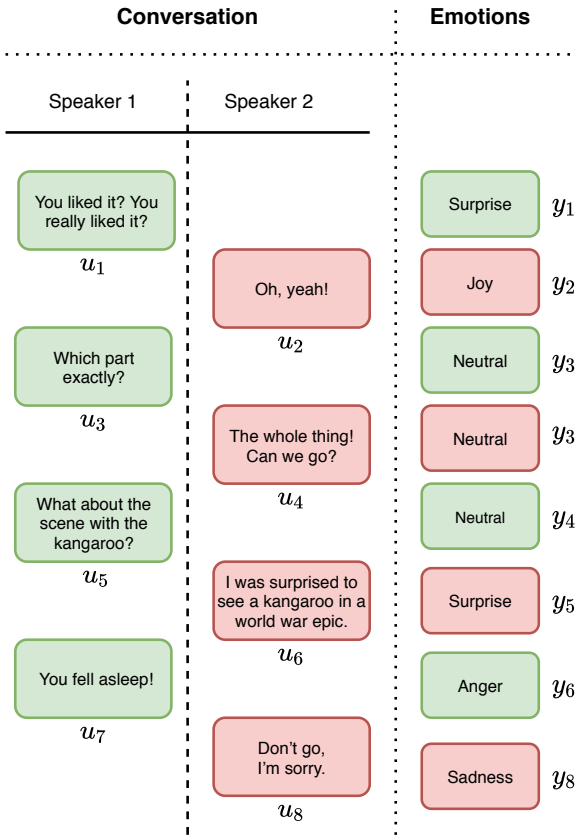


Figure 3.1 – Example of dialogue between two speakers from the MELD corpus. u_i denotes the i-th utterances and y_i the corresponding emotion.

3.1.1 Sequence Labelling Formalisation

We start by formally defining the Sequence Labelling Problem. At the highest level, we have a set D of conversations composed of utterances, i.e., $\mathrm{D} = (\mathrm{C}_1, \mathrm{C}_2, \ldots, \mathrm{C}_{|\mathrm{D}|})$ with $\mathrm{Y} = (\mathrm{Y}_1, \mathrm{Y}_2, \ldots, \mathrm{Y}_{|\mathrm{D}|})$ being the corresponding set of labels (e.g., DA, E/S). At a lower level each conversation C_i is composed of utterances u, i.e $\mathrm{C}_i = (u_1, u_2, \ldots, u_{|\mathrm{C}_i|})$ with $\mathrm{Y}_i = (y_1, y_2, \ldots, y_{|\mathrm{C}_i|})$ being the corresponding sequence of labels: each u_i is associated with a unique label y_i. At the lowest level, each utterance u_i can be seen as a sequence of words, i.e $u_i = (\omega_1^i, \omega_2^i, \ldots, \omega_{|u_i|}^i)$. Concrete examples of labelled utterances in a conversation can be found in Figure 3.1.

In this thesis, we aim at building functions f_θ that map consecutive utterances $(u_1, \ldots, u_{|\mathrm{C}_i|})$ with their corresponding labels $(y_1, \ldots, y_{|\mathrm{C}_i|})$, namely:

$$f_\theta(u_1, \ldots, u_{|\mathrm{C}_i|}) = y_1, \ldots, y_{|\mathrm{C}_i|} \tag{3.5}$$

The sequence $u_1, \ldots, u_{|\mathrm{C}_i|}$ is encoded using neural networks presented in Chapter 4 and Chapter 5. To decode the sequence of labels, we rely either on Conditional Random Fields method CRF or on a neural based decoder, both are presented and compared in Chapter 4.
Throughout this thesis, we use a simpler approach where we perform one prediction of the current utterances using the history of the conversation, namely:

$$f_\theta(u_1, \ldots, u_{|\mathrm{C}_i|}) = y_{|\mathrm{C}_i|} \tag{3.6}$$

The full sequence is then decoded in a sliding window fashion.
In the previous section, we presented the SL framework that aims at minimising

an objective over a family of function parameterised by a vector θ. In the following section, we review neural network architectures which are a specific class of function used broadly in machine learning areas and specifically in this thesis.

3.1.2 Neural Network for text classification

In this section, we start by presenting core neural architectures that are part of larger models, *i.e.* multi-layer perceptron, convolutional neural network and recurrent neural network. Then, we review methods and neural architectures that combine the aforementioned blocks in order to handle textual data.

3.1.3 Multi-Layer Perceptron

The multi-Layer Perceptron `MLP` or a fully connected Feed-Forward Neural Network `FFNN` is one of the simplest example of neural network. It consists of an input layer, one or several hidden layers and an output layer. Each layer l is composed of neurons or weights ω_{ij}^l and bias b_j^l and performs an affine transformation followed by an activation function f, namely:

$$h^{l+1} = f(\sum_i \omega_{ij}^l h_i^l + b_j^1) \tag{3.7}$$

And $h^0 = x$, *i.e.* the input of the network. As activation functions, we commonly choose among:

- `ReLU`: $f(x) = \max(0, x)$ [AGARAP, 2018].
- `GeLU`: $f(x) = \frac{x}{2}(1 + \text{erf}(\frac{x}{\sqrt{2}}))$ [HENDRYCKS and GIMPEL, 2016].
- `Sigmoid`: $f(x) = \sigma(x) = \frac{1}{1+e^{-x}}$.
- `Hyperbolic tangent`.

The final output vector is obtained via a different activation function. For classification problems, we typically use a softmax operation as presented earlier. Throughout this thesis, we use the `MLP` architecture as a decoder to perform the final decision in classification problems.

3.1.4 Convolutional Neural Network

The convolutional Neural Network [LECUN and collab., 1990] is an important neural architecture that has been a turning point in computer vision research [KRIZHEVSKY and collab., 2012]. In this architecture, each layer output unit is computed as a combination of local input units. We define the convolutional operator $*$ between a 3D tensor I (*e.g.* an image as it is organised in a 2D grid with an additional channel dimension for the different colours) and a 3D tensor K called kernel or convolutional filter, by the following equation:

$$(\mathrm{I} * \mathrm{K})_{x,y} = \sum_{i=1}^{n_\mathrm{H}} \sum_{j=1}^{n_\mathrm{W}} \sum_{k=1}^{n_\mathrm{C}} \mathrm{I}_{x+i-1,y+j-1,k} \mathrm{K}_{i,j,k} \tag{3.8}$$

In `CNN`, a layer l is composed of weights W gathering D_l kernel, namely $\mathrm{W} = (\mathrm{K}_i)_{i=1}^{\mathrm{D}_l}$. Each layer output is computed by a local convolutional operator between the input

and the layer weights, followed by a pooling operation f in order to perform a dimension reduction. Hence, units of the d^{thm} output of the layer l, h^l are calculated with the following equation:

$$h^l_{x,y,d} = f((h^{l-1} * \mathrm{K}_d)_{x,y}) \tag{3.9}$$

CNN are mostly employed in computer vision but have been also used in other areas such as speech recognition [ABDEL-HAMID and collab., 2013], NLP [COLLOBERT and collab., 2011b; GEHRING and collab., 2017]. In Chapter 7 we use CNN as sentence embedding following [KIM, 2014].

3.1.5 Recurrent Neural Network

Previous architectures, *i.e* MLP and CNN make the hypothesis of the independence of the data. After each processed input, the network's state is lost which is not desirable when the data are related in space or time as it is the case with words in a sentence or sentences in a conversation. Furthermore, the aforementioned architectures are limited to fixed-size vector examples, whereas there are a variety of applications with variable-length inputs and outputs. Recurrent Neural Network RNN [RUMELHART and collab., 1988] is a set of neural architectures that extend previous models and allow to process such sequential data $x_1, \ldots, x_n$. RNN can be viewed as a dynamical system where at each time t states are represented by a hidden state h^t. At time step t the computation of a new state depends on the previous state h^{t-1}, the input x^t and sharing parameters θ following the equation:

$$h^t = f(h^{t-1}, x^t; \theta) \tag{3.10}$$

The simplest example of this equation would be:

$$h^t = tanh(\mathrm{W}h^{t-1} + \mathrm{V}x^t + b) \tag{3.11}$$

where $\theta = (\mathrm{W}, \mathrm{V}, b)$ with W and V weights of the RNN and b the bias. For each time step t $x^t \in \mathscr{R}^{\mathrm{D}_i}$ and $h^t \in \mathscr{R}^{\mathrm{D}_h}$. Finally, a RNN is defined by a initial state h^0 which may be set to a zero vector, a random vector or be trained as a parameter.
RNN aims at providing states h^t that act as a memory and is trained to contain information of the current sequence input $x^1, \ldots, x^t$. As a consequence, the last state may represent the whole sequence and can be used further, for example through a MLP classifier. This version describes the forward RNN where each current state is computed with the knowledge of the past. However, there exists cases where future states are accessible and would be helpful to process: for instance, when processing the utterance's tokens. Similarly, a backward RNN can be then defined reading sequential data starting from the end. The combination of forward and backward RNN is called BiRNN [SCHUSTER and PALIWAL, 1997]. In that configuration each hidden state H_t is the concatenation of the forward and backward RNN hidden states, namely: $\mathrm{H}^t = [\overrightarrow{h}^t; \overleftarrow{h}^t]$. It appears that such models are hard to train as they suffer from gradient vanishing as shown in [HOCHREITER, 1998] or gradient blow up. To overcome these gradients shortcomings [HOCHREITER and SCHMIDHUBER, 1997] introduced a more complex neural architecture based on multiplicative gate unit called Long Short Term Memory LSTM. CHUNG and collab. [2014] reduces the number

of parameters of the LSTM with a new recurrent unit GRU:

$$r^t = \sigma(W_r h^{t-1} + V_r x^t + b_r), \tag{3.12}$$
$$z^t = \sigma(W_z h^{t-1} + V_z x^t + b_z), \tag{3.13}$$
$$h^t = z^t \odot h^{t-1} + (1 - z^t) tanh(W(r^t \odot h^{t-1}) + V v^t + b) \tag{3.14}$$

LSTM and GRU have intensively been used in NLP. Because of their recursive nature, states are processed one after the other making these architectures time consuming. GEHRING and collab. [2017] has provided a seq2seq architecture only based on CNN allowing the computation of the whole sequence to be done at once and be fully parallelized during training. Then VASWANI and collab. [2017b] has introduced the Transformer architecture pushing a step forward computational efficiency and performances on numerous NLP tasks. We have used Transformer architectures in Chapter 5, Chapter 6 and we present the main components of the following in this section.

3.1.6 Neural Network for language computation

In this section we have a specific focus on neural architectures that have been specifically designed for natural language computation.

Word embeddings

By representing words with continuous vectors, word embeddings allow breaking the local barrier inherent of the discrete aspect of language [HINTON and collab., 1986] and becomes the corner stone of all neural architectures in NLP. Formally each token w in a vocabulary V is represented by a one-hot encoding vector, this vector is then multiplied by a continuous matrix $E \in \mathscr{R}^{D \times |V|}$ where D is the dimension of the word embedding space. Thus, each column of E represents a token.

Encoder-Decoder

The encoder-decoder is a general paradigm that covers a wide range of neural models CHO and collab. [2014]; KALCHBRENNER and BLUNSOM [2013]; SUTSKEVER and collab. [2014]. The encoder embeds the input sequence into a hidden state vector which is then decoded to obtain the target outputs. A well-known application of this framework is the sequence-to-sequence model or Seq2Seq[SUTSKEVER and collab., 2014] originally used in machine translation systems. The encoder extracts a fixed-length vector from an input sequence of variable length (sentence in a source language A) and then is fed to a decoder to generate a output of variable length (sentence translation in a target language B). In that model, a sequence of tokens $w_1, \ldots, w_{L_i}$ is fed to a LSTM then the last hidden state h^{L_i} is used as the initial state of the LSTM decoder to produce the corresponding outputs $y_1, \ldots, y_{L_o}$, as follows:

$$p(y_1, \ldots, y_{L_o} | x_1, \ldots, x_{L_i}) = \prod_{i=1}^{L_o} p(y_i | h^{L_i}, y_1, \ldots, y_{i-1}) \tag{3.15}$$

One shortcoming of the aforementioned model is that it compresses all the length variable input information into one fixed sized hidden state which makes it difficult when coping with long range sequences. [BAHDANAU and collab., 2014] addresses

this issue introducing attention mechanism allowing the decoder to softly search among the source position containing the most relevant information. Formally, s^i denotes the i^{th} state of the decoder LSTM following $s^i = f(s^{i-1}, y_{i-1}, c_i)$ where c_i is a context vector computed as a weighted sum of the encoder hidden state :

$$c_i = \sum_{i=1}^{L_i} \alpha_{ij} h^i \tag{3.16}$$

where α_{ij} denotes the attention weights calculated as follows:

$$\alpha_{ij} = \frac{e^{score(h^i, s^{i-1})}}{\sum_{k=1}^{L_i} e^{score(h^k, s^{i-1})}} \tag{3.17}$$

In this first form, the score function was a bi-linear transformation namely:

$$score(h^k, s^i) = {h^k}^{\mathrm{T}} \mathrm{W}_a s^i \tag{3.18}$$

Due to its success, other attention mechanisms have been proposed [CHENG and collab., 2016] as well as in Chapter 4. As mentioned previously, the computational inefficiency of these neural models have led to the introduction of the Transformer. This architecture avoids recurrence by relying only on Feed Forward Network and scaled dot product attention. This attention mechanism connects all positions at constant time complexity, unlike former attention mechanisms, allowing to parallelize the computation, speeding up the training and inference times. This attention mechanism is decomposed between a set of key-value pairs (K, V) and a set of queries Q which are matrices of size d_k, d_v and d_k respectively, defined as follows:

$$\mathrm{Attention}(\mathrm{Q}, \mathrm{K}, \mathrm{V}) = softmax(\frac{\mathrm{Q}{\mathrm{K}}^{\mathrm{T}}}{\sqrt{d_k}})\mathrm{V} \tag{3.19}$$

In the encoder, all the keys, values and queries come from the same place *i.e.* outputs of the previous layer of the encoder. In this part of the network, the attention is in fact a self-attention mechanism [CHENG and collab., 2016]. The decoder makes use of a similar self-attention mechanism and a second one also based on 3.1.6 where the queries come from the previous decoder layer and the keys and values come from the outputs of the encoder. Doing so allows the decoder to attend over every position of the input sequence.
In recent years, we have seen the number of parameters of these architectures rapidly skyrocket which strongly prevents the use of a direct SL approach. Indeed, for instance, in classification tasks the number of classes can go from 2 to several thousands meaning that it will not exceed 10 bits of information. Therefore, at each step of the training phase, the classifier only predicts a small amount of information. As a result, a huge amount of training samples is needed. For example, ResNet [HE and collab., 2015], a 60 million parameters deep convolutional network, has been trained on ImageNet a dataset gathering 15 million manually labelled images. Such a dataset is hard to find in the wild as high-quality human annotation campaigns are time consuming and costly, especially for text based datasets.
In this thesis, we directly train a seq2seq model in Chapter 4 on two large corpus annotated in DA. This has prevented us from using E/S labelled corpora that are much smaller and would have led to sub-optimal results. Meanwhile, learning strategies have been developed to overcome the direct and costly SL approach. In the next section, we recall these techniques that we adapt in Chapter 5 and Chapter 6.

3.2 Pre-training for Textual Data

Pre-training is now the first step in the modern deep learning pipeline. It is an effective strategy to learn the parameters of deep neural networks which are then used/fine-tuned in further tasks (aka *downstream task, probing task*). This approach is much efficient than random initialization since it aims to learn generic representation of the data.
In NLP, the pertaining stage has been firstly used to provide distributed word embeddings. COLLOBERT and collab. [2011a] first demonstrated that learning word embeddings on large unlabelled text data could improve the generalisation on a variety of tasks. Then, MIKOLOV and collab. [2013] proposed a shallow architectures based on RNN and two novel objectives *i.e,* continuous bag-of-words CBOW and Skip-gram to learnt word embeddings. Although effective in many NLP tasks, word embeddings suffer from two main shortcomings: (1) they rely on shallow neural architectures (2) they are context-independent. Indeed, since NLP tasks are beyond word-level, the rest of the neural network needs to be trained on the downstream tasks. Hence, it appears natural to pre-train the encoder from top to bottom on sentence-level as proposed in PETERS and collab. [2018]. With the use of such deeper neural networks, in particular with the introduction of the aforementioned Transformers architecture, the number of parameters has quickly skyrocketed. Meanwhile, as previously said, a variety of supervised tasks do not gather enough data to completely exploit the power of such neural networks preventing a directly supervised training. Still, it is worth noticing that pre-training is not necessarily unsupervised: in Computer Vision (CV) models have been trained on ImageNet [DENG and collab., 2009] a dataset gathering 14 millions of pictures hand annotated over 20.000 categories. As previously mentioned there is no such dataset in NLP, for instance [CONNEAU and collab., 2017] has trained a sentence encoder on the SNLI [BOWMAN and collab., 2015] corpus, a collections of 570k human-written English sentence pairs. However, the current generation of neural models have widely adopted the unsupervised pre-training and especially the self-supervised learning approach. To summarise, the pre-training tasks can be grouped into three categories:

- **Supervised learning** (SL): aims at learning a function that maps an input to an output based on training data consisting of input-output pairs.
- **Unsupervised learning** (UL) is to find some intrinsic knowledge from unlabeled data, such as clusters, densities, latent representations.
- **Self-Supervised learning** (SSL) is a mix of supervised learning and unsupervised learning. SSL paradigm is the same as SL, but the labels of training data are generated.

In NLP, the SSL framework has attracted the most attention and has been successfully used in the learning process of a variety of well-known architectures: BERT [DEVLIN and collab., 2018], RoBERTa [LIU and collab., 2019], SpanBERT [JOSHI and collab., 2019], BART [LEWIS and collab., 2019], GPT1 [RADFORD and NARASIMHAN, 2018], GP2 [RADFORD and collab., 2019], T5 [RAFFEL and collab., 2019], ELECTRA [CLARK and collab., 2020], DeCLUTR [GIORGI and collab., 2020]. These models have achieved unprecedented results on a variety of tasks and benchmarks *e.g.* GLUE [WANG and collab., 2018], MNLI [WILLIAMS and collab., 2018], SQuAD [RAJPURKAR and collab., 2016]. In summary, SSL pre-training gives access to very large datasets allowing to increase

the numbers of parameters and capacity of models. In addition, the pre-training shows some advantages over a direct approach explaining the performances boost:

- By alleviating the inputs data manifold, SSL allows to provide universal representations that help with a wide variety of downstream tasks, especially with small datasets WANG and collab. [2020].
- Provide a better initialisation point for the neural network parameters. This allows a better generalisation and speed up the fine-tuning on downstream tasks DEVLIN and collab. [2018].
- ERHAN and collab. [2010] shows that pre-training acts as a regularisation avoiding overfitting on small datasets.

The aforementioned models were designed to encode unstructured text whereas, as we have shown previously, a dialogue has a specific structure that make it unique. Moreover, it has been shown that transfer learning using sentence embeddings tends to outperform word-level transfer [CER and collab., 2018]. This observation can be applied once again at the conversational level. In Chapter 5 we propose to pre-train an encoder tailored for conversations and adapt standard SSL objectives we recall in the following section.

3.2.1 Self Supervised Learning

Self Supervised Learning (SLL) leverages the label dependency as the labels are obtained from the data itself by using a *"semi-automatic"* process. Then, part of the data is predicted using the other parts. Specifically, the "other part" could be incomplete, transformed, distorted, or corrupted (*i.e.*, data augmentation technique). SLL has been successfully applied in a variety of machine learning domains such as graph Learning [HU and collab., 2020], speech recognition [RAVANELLI and collab., 2020], computer vision [DOERSCH and collab., 2015], NLP [DEVLIN and collab., 2018; RADFORD and NARASIMHAN, 2018; RADFORD and collab., 2019]. This section covers general concepts of SSL with a particular focus on methods applied to NLP used and adapted in this manuscript. The challenge of SLL is to define proper objectives for unlabelled data (aka *pretext task* or pre-training objectives). These tasks can be divided into three categories :

- **Generative:** train an encoder to encode input x in to an explicit vector z and a decoder to reconstruct x from z (e.g. the cloze test[TAYLOR, 1953]).
- **Contrastive:** train an encoder to encode input x into an explicit vector z to measure similarity (e.g., mutual information maximization, instance discrimination).
- **Generative/Contrastive (Adversarial):** train an encoder-decoder to generate fake samples and a discriminator to distinguish them from real samples.

In this thesis, we focused on the generative approach. Contrastive and Generative/Contrastive are also two interesting frameworks, especially with NLU. They have been successfully applied to provide sentence embeddings [CLARK and collab., 2020; GIORGI and collab., 2020; LOGESWARAN and LEE, 2018]. We leave their adaptation for the dialogue structure as future work.

3.2.2 Generative SSL

Auto-Regressive Model

In the auto-regressif AR model, giving a sequence of input $S = (w_1, \ldots, w_N)$, the prediction of each element w_i dependent on its predecessors. Hence the goal is to maximize the joint distribution with the following factorization:

$$p(S) = \prod_{i=1}^{N} p(w_i | w_{1:i-1}). \tag{3.20}$$

This is a general framework that has been successfully applied for learning continuous word representation [BENGIO and collab., 2003] or in text generation with RNN [GRAVES, 2013]. In NLP, a Language Model LM plays a key role in a variety of tasks. Their function is to assign probabilities to a sentence, *i.e* to determine how likely a sentence is in a given language. As a sentence can be naturally described as a sequence of tokens language model can be leveraged maximising the likelihood under the AR factorisation. This quantity is effectively computed in [VASWANI and collab., 2017a] with a self-attention based architecture called `Transformers`. GPT1 [RADFORD and NARASIMHAN, 2018] boards a more flexible approach with the use of a k length sliding window, namely:

$$p(S) = \prod_{i=1}^{N} p(w_i | w_{i-k:i-1}). \tag{3.21}$$

The AR formulation 3.2.2 is often referred as *forward* LM in contrast to the *backward* LM where the i th element is predicted using elements ahead, namely:

$$p(S) = \prod_{i=0}^{N-1} p(w_i | w_{i+1:N}). \tag{3.22}$$

ELMo word embedding [PETERS and collab., 2018] are trained using both *forward* and *backward* AR objectives. As these objectives are unidirectional it prevents from learning deep bidirectional context. Hence there is a discrepancy between AR language model and most of the downstream tasks in NLU since they require bidirectional mobilisation.

Auto-Encoding Model

In the auto-encoding (AE) model, the goal is to reconstruct an input S from a corrupted input $\tilde{S}$. Doing so, the model does not try to estimate the joint distribution as in AR model and allows bidirectional context for reconstruction. Moreover, AE is an intuitive and flexible framework, broadly used with numerous variants.
In its simplest AE form, the input is not corrupted: S is projected in a latent space with an encoder function f_e then reconstructed with a decoder function f_d into an output $\tilde{S}$, namely:

$$f_e(S) = z. \tag{3.23}$$

$$f_d(z) = \tilde{S}. \tag{3.24}$$

The goal is to minimize the error of reconstruction of S, *i.e* having $\tilde{S}$ close to the original input S. For instance, if $S \in R^p$ the objective would be to minimize the square distance between S and ~S. However, as such model learns the data with the output being equal to the output, the extreme form would be the identity function, thus no useful information would be learnt.

Denoising Autoencoder Model

Denoising Autoencoder (DAE) models solve this problem by corrupting the inputs adding noise or masking some parts. In recent years one of the most spectacular and broadly use of this paradigm is the Masked Language Model MLM introduced by [DEVLIN and collab., 2018] for training BERT. Inspired by the well-known *cloze task* [TAYLOR, 1953], they propose to replace part of the inputs tokens with a special token: [MASK]. As this token is only present during the pre-training stage, authors do not systematically replace the chosen tokens with [MASK] but also in a small proportion with random words or without any modification. The MLM loss is defined as follow:

$$\mathscr{L}_{\text{MLM}} = -\sum_{i \in \mathscr{M}_{\text{S}}} log(p(w_i|\hat{\text{S}}). \tag{3.25}$$

where $\mathscr{M}_{\text{S}}$ denotes the set of masked indices for the sentences S and $\hat{\text{S}}$ its corrupted form. Since then, several models have been derived from the classic MLM such as SpanBERT [JOSHI and collab., 2019] where continuous spans of tokens *i.e* consecutive tokens are masked. The masked span's start indices and length are chosen randomly. Authors show improvement over BERT in coreference resolution.
Usually, MLM is solved as a classification task: an encoder embeds the corrupted sentence into a vector that is then fed to a softmax classifier to predict the masked tokens. Another option is to use a seq2seq approach [LEWIS and collab., 2019; RAFFEL and collab., 2019; SONG and collab., 2019]. In this setting, the encoder output is fed to a retrogressive decoder that aims at reconstructing the whole sentence. These seq2seq MLM achieved better results on generative downstream tasks such as question-answering, summarization or machine translation.

Finally there exists models that combine both AR and AE. As an example, BERT also uses another objective called the Next Sentence Prediction NSP where two consecutive sentences S_1 and S_2 are concatenate. All S_2 tokens are masked and should be predicted with the knowledge of S_1. Similarly, XLNet [YANG and collab., 2019] brings the best of the two worlds introducing the Permutation Language Model PLM a generalization of the AR objective.

$$\mathscr{L}_{\text{PLM}} = \mathbb{E}_{\mathbf{z} \sim \mathscr{Z}_{\mathscr{T}}} \sum_{t=1}^{\text{T}} log\, p_\theta(w_{z_t}|w_{\mathbf{z_{<t}}}). \tag{3.26}$$

where $\mathbf{z}$ is a random permutation of the sentences tokens.

In this section, we have seen how SSL leverages the gigantic need for labelled data from raw data to build representation at the utterance level. In Chapter 5 we demonstrate how to adapt these objectives at the conversational level. In the following, we recall elements from the transfer learning paradigm relevant to this work.

3.3 Fine-tuning on Downstream Tasks

As previously mentioned the classical NLP pipeline can be divided into two stages as shown on Figure 3.2: 1) self-supervised pre-training stage on a large dataset of a general domain D 2) a fine-tuning stage on a specific task $\text{T} = \{(\text{X}_i, \text{Y}_i) \in \text{D}' \times \mathscr{Y}\}$, where D' is the domain of the target task and $\mathscr{Y}$ a specific annotation scheme. Hence several parameters need to be considered when transferring a source model to a

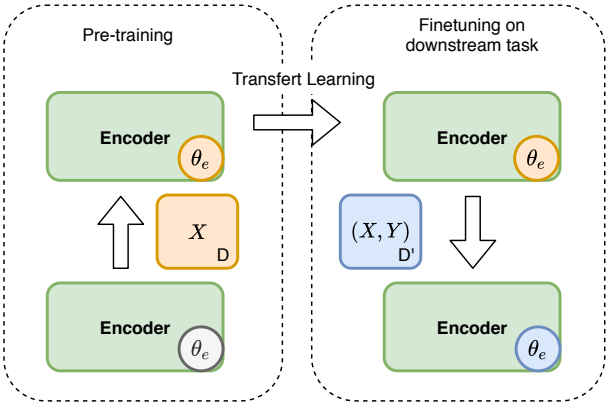


Figure 3.2 – Transfer Learning

target model.

Pre-training task vs fine-tuning task

Currently the MLM pre-training task is ubiquitous and has been useful in a wide range of NLP tasks [RADFORD and collab., 2019]. However, some pretaining tasks can be more suited than others toward specific downstream tasks, for instance, Next Sentence Prediction [DEVLIN and collab., 2018] forces the model to catch the relationship between two sentences. Such a task is useful in a context of Question-Anwsering QA, Natural Language Inferance NLI. Conversely, the MLM of BERT focused mostly on the encoder, making it more suited for NLU tasks than generation.

Source Domain vs Target Domain

The discrepancy between the source domain D, *i.e.* domain used during pre-training and the target domain D′, *i.e.* the domain used during fine-tuning, is another thing to consider. A domain is characterised by several aspects such as style, topic and genre [GURURANGAN and collab., 2020]. Hence, two texts can deal with the same subject but with two different styles. These aspects will have an impact on the vocabulary used and in fine on the data distribution. Intuitively, one could consider that the closer the data distributions of the source and target domain, the easier the transfer learning, which leads to better results for the considered tasks. Current SOTA models are pre-trained on very large corpus such as Wikipedia or book corpus. While achieving good performance on most on NLP tasks, they do not seem to be optimal for tasks involving spoken dialogue. As we showcased in Chapter 2, there is a discrepancy between written text and spoken dialogue. This motivates the choice of OpenSubtitles [LISON and TIEDEMANN, 2016] as pre-training corpus in Chapter 5. Domain adaptation is a whole area of ML and NLP in particular, while considered in this thesis, a deeper study is out of the scope of this work. It is worth noticing GURURANGAN and collab. [2020] showcases that domain shift can be leveraged by continuing the pre-training on large corpus which domain is closest to the target domain.

Layer choice

For multi-layers RNN-LSTM encoders, it has been shown [BELINKOV and collab., 2017] that layers capture different information that can benefit different tasks (POS tagging, parsing, semantic role, co-reference). The same analysis has also been made for pre-trained models such as BERT [TENNEY and collab., 2019]: basic syntactic information is captured in early layers while high-level semantic information appears at the top layers. Hence there are several ways to select the representation:

- *Word embedding*: This approach was the first generation of pre-train models and used to be very popular in the NLP community. However, they are context-independent and are used by shallow models that still need to be trained from scratch.
- *All layers*: this strategy is adopted in ELMo [PETERS and collab., 2018] where the final representation is computed as the weighted sum of all layers.
- *Last layer*: This is a classical approach adopted by a lot of current systems, especially those based on transformers. For example, in Chapter 5 and Chapter 6, an utterance representations is achieved by taking the representation of the [CLS] token of the top layer.

Chapter 3 Conclusion

In this chapter, we recalled the supervised learning framework and formally introduced both DA and E/S classification tasks in spoken dialogue. As we aim at learning fully neural based classifiers, we presented all neural architectures used throughout this thesis. We showcased the transfer learning strategy as well as pre-training objectives used in NLP. We presented the limitations of existing approaches that we will address in Chapter 5 and Chapter 6. In the next chapter, we will present our contributions related to RQ1.

3.4 References

ABDEL-HAMID, O., L. DENG and D. YU. 2013, "Exploring convolutional neural network structures and optimization techniques for speech recognition.", in *Interspeech*, vol. 11, Citeseer, p. 73–5. 45

AGARAP, A. F. 2018, "Deep learning using rectified linear units (relu)", *CoRR*, vol. abs/1803.08375. URL http://arxiv.org/abs/1803.08375. 44

BAHDANAU, D., K. CHO and Y. BENGIO. 2014, "Neural machine translation by jointly learning to align and translate", URL http://arxiv.org/abs/1409.0473, cite arxiv:1409.0473Comment: Accepted at ICLR 2015 as oral presentation. 46

BELINKOV, Y., N. DURRANI, F. DALVI, H. SAJJAD and J. R. GLASS. 2017, "What do neural machine translation models learn about morphology?", *CoRR*, vol. abs/1704.03471. URL http://arxiv.org/abs/1704.03471. 53

BENGIO, Y., R. DUCHARME, P. VINCENT and C. JANVIN. 2003, "A neural probabilistic language model", *J. Mach. Learn. Res.*, vol. 3, nº null, p. 1137–1155, ISSN 1532-4435. 50

BOWMAN, S. R., G. ANGELI, C. POTTS and C. D. MANNING. 2015, "A large annotated corpus for learning natural language inference", *CoRR*, vol. abs/1508.05326. URL http://arxiv.org/abs/1508.05326. 48

CARUANA, R., S. LAWRENCE and L. GILES. 2000, "Overfitting in neural nets: Backpropagation, conjugate gradient, and early stopping", in *Proceedings of the 13th International Conference on Neural Information Processing Systems*, NIPS'00, MIT Press, Cambridge, MA, USA, p. 381–387. 42

CER, D., Y. YANG, S. KONG, N. HUA, N. LIMTIACO, R. S. JOHN, N. CONSTANT, M. GUAJARDO-CESPEDES, S. YUAN, C. TAR, Y. SUNG, B. STROPE and R. KURZWEIL. 2018, "Universal sentence encoder", *CoRR*, vol. abs/1803.11175. URL http://arxiv.org/abs/1803.11175. 49

CHENG, J., L. DONG and M. LAPATA. 2016, "Long short-term memory-networks for machine reading", *CoRR*, vol. abs/1601.06733. URL http://arxiv.org/abs/1601.06733. 47

CHO, K., B. VAN MERRIENBOER, Ç. GÜLÇEHRE, F. BOUGARES, H. SCHWENK and Y. BENGIO. 2014, "Learning phrase representations using RNN encoder-decoder for statistical machine translation", *CoRR*, vol. abs/1406.1078. URL http://arxiv.org/abs/1406.1078. 46

CHUNG, J., Ç. GÜLÇEHRE, K. CHO and Y. BENGIO. 2014, "Empirical evaluation of gated recurrent neural networks on sequence modeling", *CoRR*, vol. abs/1412.3555. URL http://arxiv.org/abs/1412.3555. 45

CLARK, K., M. LUONG, Q. V. LE and C. D. MANNING. 2020, "ELECTRA: pre-training text encoders as discriminators rather than generators", *CoRR*, vol. abs/2003.10555. URL https://arxiv.org/abs/2003.10555. 48, 49

COLLOBERT, R., J. WESTON, L. BOTTOU, M. KARLEN, K. KAVUKCUOGLU and P. KUKSA. 2011a, "Natural language processing (almost) from scratch", *J. Mach. Learn. Res.*, vol. 12, nº null, p. 2493–2537, ISSN 1532-4435. 48

COLLOBERT, R., J. WESTON, L. BOTTOU, M. KARLEN, K. KAVUKCUOGLU and P. P. KUKSA. 2011b, "Natural language processing (almost) from scratch", *CoRR*, vol. abs/1103.0398. URL http://arxiv.org/abs/1103.0398. 45

CONNEAU, A., D. KIELA, H. SCHWENK, L. BARRAULT and A. BORDES. 2017, "Supervised learning of universal sentence representations from natural language inference data", *CoRR*, vol. abs/1705.02364. URL http://arxiv.org/abs/1705.02364. 48

DENG, J., W. DONG, R. SOCHER, L.-J. LI, K. LI and L. FEI-FEI. 2009, "Imagenet: A large-scale hierarchical image database", in *2009 IEEE conference on computer vision and pattern recognition*, Ieee, p. 248–255. 48

DEVLIN, J., M. CHANG, K. LEE and K. TOUTANOVA. 2018, "BERT: pre-training of deep bidirectional transformers for language understanding", *CoRR*, vol. abs/1810.04805. URL http://arxiv.org/abs/1810.04805. 48, 49, 51, 52

DOERSCH, C., A. GUPTA and A. A. EFROS. 2015, "Unsupervised visual representation learning by context prediction", *CoRR*, vol. abs/1505.05192. URL http://arxiv.org/abs/1505.05192. 49

ERHAN, D., Y. BENGIO, A. COURVILLE, P.-A. MANZAGOL, P. VINCENT and S. BENGIO. 2010, "Why does unsupervised pre-training help deep learning?", *Journal of Machine Learning Research*, vol. 11, nº 19, p. 625–660. URL http://jmlr.org/papers/v11/erhan10a.html. 49

GEHRING, J., M. AULI, D. GRANGIER, D. YARATS and Y. N. DAUPHIN. 2017, "Convolutional sequence to sequence learning", *CoRR*, vol. abs/1705.03122. URL http://arxiv.org/abs/1705.03122. 45, 46

GIORGI, J. M., O. NITSKI, G. D. BADER and B. WANG. 2020, "Declutr: Deep contrastive learning for unsupervised textual representations", *CoRR*, vol. abs/2006.03659. URL https://arxiv.org/abs/2006.03659. 48, 49

GRAVES, A. 2013, "Generating sequences with recurrent neural networks", *CoRR*, vol. abs/1308.0850. URL http://arxiv.org/abs/1308.0850. 50

GURURANGAN, S., A. MARASOVIC, S. SWAYAMDIPTA, K. LO, I. BELTAGY, D. DOWNEY and N. A. SMITH. 2020, "Don't stop pretraining: Adapt language models to domains and tasks", *CoRR*, vol. abs/2004.10964. URL https://arxiv.org/abs/2004.10964. 52

HE, K., X. ZHANG, S. REN and J. SUN. 2015, "Deep residual learning for image recognition", *CoRR*, vol. abs/1512.03385. URL http://arxiv.org/abs/1512.03385. 47

HENDRYCKS, D. and K. GIMPEL. 2016, "Bridging nonlinearities and stochastic regularizers with gaussian error linear units", *CoRR*, vol. abs/1606.08415. URL http://arxiv.org/abs/1606.08415. 44

HINTON, G. E., J. L. MCCLELLAND and D. E. RUMELHART. 1986, *Distributed Representations*, MIT Press, Cambridge, MA, USA, ISBN 026268053X, p. 77–109. 46

HOCHREITER, S. 1998, "The vanishing gradient problem during learning recurrent neural nets and problem solutions", *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 6, doi: 10.1142/S0218488598000094, p. 107–116. 45

HOCHREITER, S. and J. SCHMIDHUBER. 1997, "Long short-term memory", *Neural Computation*, vol. 9, nº 8, p. 1735–1780. 45

HU, Z., Y. DONG, K. WANG and Y. SUN. 2020, "Heterogeneous graph transformer", *CoRR*, vol. abs/2003.01332. URL https://arxiv.org/abs/2003.01332. 49

JOSHI, M., D. CHEN, Y. LIU, D. S. WELD, L. ZETTLEMOYER and O. LEVY. 2019, "Spanbert: Improving pre-training by representing and predicting spans", *CoRR*, vol. abs/1907.10529. URL http://arxiv.org/abs/1907.10529. 48, 51

KALCHBRENNER, N. and P. BLUNSOM. 2013, "Recurrent continuous translation models", in *EMNLP*. 46

KIM, Y. 2014, "Convolutional neural networks for sentence classification", *CoRR*, vol. abs/1408.5882. URL http://arxiv.org/abs/1408.5882. 45

KRIZHEVSKY, A., I. SUTSKEVER and G. E. HINTON. 2012, "Imagenet classification with deep convolutional neural networks", in *Advances in Neural Information Processing Systems 25*, édité par F. Pereira, C. J. C. Burges, L. Bottou and K. Q. Weinberger, Curran Associates, Inc., p. 1097–1105. URL http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf. 44

LECUN, Y., B. BOSER, J. DENKER, D. HENDERSON, R. HOWARD, W. HUBBARD and L. JACKEL. 1990, "Handwritten digit recognition with a back-propagation network", in *Advances in Neural Information Processing Systems*, vol. 2, édité par D. Touretzky, Morgan-Kaufmann. URL https://proceedings.neurips.cc/paper/1989/file/53c3bce66e43be4f209556518c2fcb54-Paper.pdf. 44

LEWIS, M., Y. LIU, N. GOYAL, M. GHAZVININEJAD, A. MOHAMED, O. LEVY, V. STOYANOV and L. ZETTLEMOYER. 2019, "BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension", *CoRR*, vol. abs/1910.13461. URL http://arxiv.org/abs/1910.13461. 48, 51

LISON, P. and J. TIEDEMANN. 2016, "OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles", in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, European Language Resources Association (ELRA), Portorož, Slovenia, p. 923–929. URL https://aclanthology.org/L16-1147. 52

LIU, Y., M. OTT, N. GOYAL, J. DU, M. JOSHI, D. CHEN, O. LEVY, M. LEWIS, L. ZETTLEMOYER and V. STOYANOV. 2019, "Roberta: A robustly optimized BERT pretraining approach", *CoRR*, vol. abs/1907.11692. URL http://arxiv.org/abs/1907.11692. 48

LOGESWARAN, L. and H. LEE. 2018, "An efficient framework for learning sentence representations", *CoRR*, vol. abs/1803.02893. URL http://arxiv.org/abs/1803.02893. 49

MIKOLOV, T., K. CHEN, G. CORRADO and J. DEAN. 2013, "Efficient estimation of word representations in vector space", . 48

PETERS, M. E., M. NEUMANN, M. IYYER, M. GARDNER, C. CLARK, K. LEE and L. ZETTLEMOYER. 2018, "Deep contextualized word representations", . 48, 50, 53

RADFORD, A. and K. NARASIMHAN. 2018, "Improving language understanding by generative pre-training", . 48, 49, 50

RADFORD, A., J. WU, R. CHILD, D. LUAN, D. AMODEI and I. SUTSKEVER. 2019, "Language models are unsupervised multitask learners", . 48, 49, 52

RAFFEL, C., N. SHAZEER, A. ROBERTS, K. LEE, S. NARANG, M. MATENA, Y. ZHOU, W. LI and P. J. LIU. 2019, "Exploring the limits of transfer learning with a unified text-to-text transformer", *CoRR*, vol. abs/1910.10683. URL http://arxiv.org/abs/1910.10683. 48, 51

RAJPURKAR, P., J. ZHANG, K. LOPYREV and P. LIANG. 2016, "Squad: 100, 000+ questions for machine comprehension of text", *CoRR*, vol. abs/1606.05250. URL http://arxiv.org/abs/1606.05250. 48

RAVANELLI, M., J. ZHONG, S. PASCUAL, P. SWIETOJANSKI, J. MONTEIRO, J. TRMAL and Y. BENGIO. 2020, "Multi-task self-supervised learning for robust speech recognition", in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, p. 6989–6993. 49

RUMELHART, D. E., G. E. HINTON and R. J. WILLIAMS. 1988, *Learning Representations by Back-Propagating Errors*, MIT Press, Cambridge, MA, USA, ISBN 0262010976, p. 696–699. 45

SCHUSTER, M. and K. PALIWAL. 1997, "Bidirectional recurrent neural networks", *IEEE Trans. Signal Process.*, vol. 45, p. 2673–2681. 45

SONG, K., X. TAN, T. QIN, J. LU and T. LIU. 2019, "MASS: masked sequence to sequence pre-training for language generation", *CoRR*, vol. abs/1905.02450. URL http://arxiv.org/abs/1905.02450. 51

SRIVASTAVA, N., G. HINTON, A. KRIZHEVSKY, I. SUTSKEVER and R. SALAKHUTDINOV. 2014, "Dropout: A simple way to prevent neural networks from overfitting", *Journal of Machine Learning Research*, vol. 15, nº 56, p. 1929–1958. URL http://jmlr.org/papers/v15/srivastava14a.html. 42

SUTSKEVER, I., O. VINYALS and Q. V. LE. 2014, "Sequence to sequence learning with neural networks", *CoRR*, vol. abs/1409.3215. URL http://arxiv.org/abs/1409.3215. 46

TAYLOR, W. L. 1953, ""cloze procedure": A new tool for measuring readability", *Journalism Quarterly*, vol. 30, nº 4, doi: 10.1177/107769905303000401, p. 415–433. URL https://doi.org/10.1177/107769905303000401. 49, 51

TENNEY, I., D. DAS and E. PAVLICK. 2019, "BERT rediscovers the classical NLP pipeline", *CoRR*, vol. abs/1905.05950. URL http://arxiv.org/abs/1905.05950. 53

VASWANI, A., N. SHAZEER, N. PARMAR, J. USZKOREIT, L. JONES, A. N. GOMEZ, L. KAISER and I. POLOSUKHIN. 2017a, "Attention is all you need", . 50

VASWANI, A., N. SHAZEER, N. PARMAR, J. USZKOREIT, L. JONES, A. N. GOMEZ, L. U. KAISER and I. POLOSUKHIN. 2017b, "Attention is all you need", in *Advances in Neural Information Processing Systems*, vol. 30, édité par I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan and R. Garnett, Curran Associates, Inc. URL https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf. 46

WANG, A., A. SINGH, J. MICHAEL, F. HILL, O. LEVY and S. R. BOWMAN. 2018, "GLUE: A multi-task benchmark and analysis platform for natural language understanding", *CoRR*, vol. abs/1804.07461. URL http://arxiv.org/abs/1804.07461. 48

WANG, S., M. KHABSA and H. MA. 2020, "To pretrain or not to pretrain: Examining the benefits of pretraining on resource rich tasks", *CoRR*, vol. abs/2006.08671. URL https://arxiv.org/abs/2006.08671. 49

WILLIAMS, A., N. NANGIA and S. BOWMAN. 2018, "A broad-coverage challenge corpus for sentence understanding through inference", in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, Association for Computational Linguistics, p. 1112–1122. URL http://aclweb.org/anthology/N18-1101. 48

YANG, Z., Z. DAI, Y. YANG, J. G. CARBONELL, R. SALAKHUTDINOV and Q. V. LE. 2019, "Xlnet: Generalized autoregressive pretraining for language understanding", *CoRR*, vol. abs/1906.08237. URL http://arxiv.org/abs/1906.08237. 51

Part II

Monolingual Spoken Dialogue Understanding

Part II Abstract

This part gathers our contributions related to RQ1 "How to leverage recent advances in neural networks to build a system that can automatically label English spoken dialogue utterances ?". This part is further divided into two different chapters:

- Our formalization of sequence labelling shares common points with Neural Machine Translation (NMT) problems. Inspired by the recent successes of seq2seq approaches widely adopted NMT, Chapter 4 proposes a seq2seq tailored towards DA classification paving the way for further innovations inspired by advances in NMT research. This seq2seq aims at improving the modelling of tags sequentially.
- The proposed seq2seq model in Chapter 4 is data-hungry and requires large corpora of labelled utterance to be trained. As an example Switchboard has over $100k$ labelled utterances, whereas the SEMAINE corpus has about $5k$ labelled utterances. Additionally, those models are very specific to the labelling scheme employed. Adapting them to different sets of emotions or dialogue acts would require more annotated data. In Chapter 5, we address the problem of E/S and DA labelling in a data scarcity setting by proposing new pre-trained objectives. In this chapter, we learn generic representations that can quickly adapt to small labelled datasets.

Chapter 4

A seq2seq Model for Sequence Labelling

Chapter 4 abstract

This chapter present our first contribution where we propose to tackle the problem of sequence labelling in spoken dialogue with a focus on DA classification. The task of predicting DA based on conversational dialogue is a key component in the development of conversational agents. Accurately predicting DAs requires a precise modeling of both the conversation and the global tag dependencies. We leverage seq2seq approaches widely adopted in Neural Machine Translation (NMT) to improve the modelling of tag sequences. Seq2seq models are known to learn complex global dependencies while currently proposed approaches using linear conditional random fields (CRF) only model local tag dependencies. In this work, we introduce a seq2seq model tailored for DA classification using: a hierarchical encoder, a novel *guided attention* mechanism and beam search applied to both training and inference. Compared to the state of the art, our model does not require handcrafted features and is trained end-to-end. Furthermore, the proposed approach achieves an unmatched accuracy score of 85% on SwDA, and state-of-the-art accuracy score of 91.6% on MRDA. The content of this chapter has been publicly released and can be found on an open preprint server: P. Colombo & E. Chapuis and collab. [2020].

4.1 Introduction

In natural language processing research, the dialogue act (DA) concept plays an important role. DAs are semantic labels associated with each utterance in a conversational dialogue that indicate the speaker's intention, e.g., question, backchannel, statement-non-opinion, statement opinion. A key to model dialogue is to detect the intent of the speaker: correctly identifying a question gives an important clue to produce an appropriate response. As can be observed in Table 4.1, DA classification relies on its conversational aspect, i.e., predicting an utterance's DA requires the knowledge of previous sentences and their associated act labels. For example, if a speaker asks a question, the interlocutor will answer with a response, analogously, a "Greeting" or a "Farwell" will be followed by a similar dialogue act. This means that in a conversation there is a sequential structure in the emitted dialogue acts. This poses the basis for the adoption of a novel perspective on the DA classification problem, i.e., from a multi-classification task to a sequence labeling one.

Speaker	Utterance
A	Is there anyone who doesn't know Nancy?
A	Do you - Do you know Nancy ?
B	Me ?
B	Mm-hmm
B	I know Nancy

Table 4.1 – Example of conversation from Switchboard Dialogue Act Corpus. A is speaking with B.

Limitations of current models: Current state-of-the-art models rely on the use of linear Conditional Random Field (CRF) combined with a recurrent neural network based encoder [CHEN and collab., 2018; LI and collab., 2018a; RAHEJA and TETREAULT, 2019] to model DA sequential dependencies. Unfortunately such approaches only capture local dependencies between two adjacent dialogue acts. For instance, if we consider the example in Table 4.1 we can see that the last statement "I know Nancy" is a response to the first question "Is there anyone who doesn't know Nancy" and the knowledge of the previous backchannel does not help the prediction of the last dialogue act. Therefore, we must consider dependencies between labels with a scope that is wider than two successive utterances. In Neural Machine Translation (NMT), the problem of global dependencies has been addressed using seq2seq models [SUTSKEVER and collab., 2014] that follow the encoder-decoder framework. The encoder embeds an input sentence into a single hidden vector which contains both global and local dependencies, and the hidden vector is then decoded to produce an output sequence. In this work, we propose a seq2seq architecture tailored towards DA classification paving the way for further innovations inspired by advances in NMT research.

Contributions: In this work

1. We formalise the Dialogue Act Prediction problem in a way that emphasises the relations between DA classification and NMT.

2. We demonstrate that the seq2seq architecture suits better to the DA classification task.

3. We present a seq2seq model leveraging NMT techniques that reaches an accuracy of 85%, outperforming the state of the art by a margin of around 2%, on the Switchboard Dialogue Act Corpus (SwDA) [STOLCKE and collab., 1998] and a state-of-the-art accuracy score of 91,6% on the Meeting Recorder Dialogue Act (MRDA). This seq2seq model exploits a hierarchical encoder with a novel *guided attention* mechanism that fits with our setting without any handcrafted features. We finetune our seq2seq using a sequence level training objective making use of the beam search algorithm. To our knowledge, this is among the first seq2seq model proposed for DA classification.

4.2 Background

4.2.1 DA classification

Several approaches have been proposed to tackle the DA classification problem. These methods can be divided into two different categories. The first class of methods relies on the independent classification of each utterance using various techniques, such as HMM [STOLCKE and collab., 2000], SVM [SURENDRAN and LEVOW, 2006] and Bayesian Network [KEIZER and collab., 2002]. The second class, which achieves better performance, leverages the context, to improve the classifier performance by using deep learning approaches to capture contextual dependencies between input sentences [BOTHE and collab., 2018; KHANPOUR and collab., 2016]. Another refinement of input context-based classification is the modelling of inter-tag dependencies. This task is tackled as sequence-based classification where output tags are considered as a DA sequence [CHEN and collab., 2018; KUMAR and collab., 2018; LI and collab., 2018a; RAHEJA and TETREAULT, 2019; STOLCKE and collab., 2000].

Two classical benchmarks are adopted to evaluate DA classification systems: the Switchboard Dialogue Act Corpus (SwDA)[STOLCKE and collab., 1998] and the Meeting Recorder Dialogue Act (MRDA) [JANIN and collab., 2003]. State-of-the-art techniques achieve an accuracy of 82.9% [LI and collab., 2018a; RAHEJA and TETREAULT, 2019]. To capture input contextual dependencies they adopt a hierarchical encoder and a CRF to model inter-tag dependencies. The main limitation of the aforementioned architecture is that a linear-CRF model is able to only capture dependencies at a local level and fails to capture non local dependencies. In this chapter, we tackle this issue with a sequence-to-sequence using a guided attention mechanism.

4.2.2 Seq2seq models

Seq2seq models have been successfully applied to NMT, where modeling non local dependencies is a crucial challenge. DA classification can be seen as a problem where the goal is to map a sequence of utterances to a sequence of DA. Thus, it can be formulated as sequence to sequence problem very similar to NMT.
The general architecture of our seq2seq models [SUTSKEVER and collab., 2014] follows a classical encoder-decoder approach with attention [LUONG and collab., 2015]. We use GRU cells [CHO and collab., 2014], since they are faster to train than LSTM ones [JOZEFOWICZ and collab., 2015]. Recent advances have improved both the learning and the inference process, producing sequences that are more coherent by means of sequence level losses [WISEMAN and RUSH, 2016] and various beam search settings [VIJAYAKUMAR and collab., 2016; WU and collab., 2016]. The closest

setting where seq2seq model have been successfully used is dependency parsing [Li and collab., 2018b], where output dependencies are crucial to achieve state-of-the-art performance. In our work we adjust NMT techniques to the specifics of DA classification.

4.3 Problem Statement

4.3.1 DA classification as an NMT problem

For the formalisation of the DA classification problem we rely on the mathematical notations introduced in subsection 3.1.1. In NMT, the goal is to associate for any sentence $X^{l_1} = (x_1^{l_1}, ..., x_{|X^{l_1}|}^{l_1})$ in language l_1 a sentence $X^{l_2} = (x_1^{l_2}, ..., x_{|X^{l_2}|}^{l_2})$ in language l_2 where $x_i^{l_k}$ is the i word in the sentence in language l_k. Using this formalism, it is straightforward to notice two main similarities ($\mathbb{S}_1$, $\mathbb{S}_2$) between DA classification and NMT. ($\mathbb{S}_1$) In NMT and DA classification, the goal is to maximise the likelihood of the output sequence given the input sequence ($P(X^{l_2}|X^{l_1})$ versus $P(Y_i|C_i)$). ($\mathbb{S}_2$) For the two tasks, there are strong dependencies between units composing both the input and output sequences. In NMT, those units are words (x_i and y_i), inDA classification those units are utterances and DA labels (u_i and y_i).

4.3.2 Specifics of DA classification

While NMT and DA classification are similar under some point of views, three differences are immediately apparent ($\mathbb{D}_i$). ($\mathbb{D}_1$) In NMT, the input units x_i represent words, in DA classification u_i are input sequences composed with words. Considering the set of all possible sequences as input (context consideration leads to superior performance) implies that the dimension of the input space several order of magnitude larger than compared to a standard NMT. ($\mathbb{D}_2$) In DA, we have a perfect alignment between input and output sequences (hence $T = T\prime$). Some languages, e.g., French, English, Italian share a partial alignment, but in DA classification we have a strong mapping between y_i and x_i. ($\mathbb{D}_3$) In NMT, the input space (number of words in l_1) is approximately the same size of the output space (number of words in l_2). In our case the output space (number of DA tags $|\mathscr{Y}| < 100$ has a limited size, with a dimension that is many order of magnitude smaller than the input space one.

In the following, we propose an end-to-end seq2seq architecture for DA classification that leverages ($\mathbb{D}_1$) using a hierarchical encoder, ($\mathbb{D}_2$) through a guided attention mechanism and ($\mathbb{D}_3$) using beam search during both training and inference, taking advantage of the limited dimension of the output space.

4.4 Models

In Seq2seq, the encoder takes a sequence of sentences and represents it as a single vector $H_i \in \mathscr{R}^d$ and then pass it to the decoder for tag generations.

4.4.1 Encoders

In this section we introduce the different encoders we consider in our experiments. We exploit the hierarchical structure of the dialogue to reduce the input space size

($\mathbb{D}_1$) and to preserve word/sentence structure. During both training and inference, the context size is fixed to T. Formally, an encoder takes as input a fixed number of utterances $(u_{i-\mathrm{T}},..,u_i)$ and outputs a vector $\mathrm{H}_i \in \mathscr{R}^d$ which will serve to initialize the hidden state of the decoder. The first level of the encoder computes $\mathscr{E}_{u_t}$, an embedding of u_t based on the words composing the utterance, and the next levels compute H_i based on $\mathscr{E}_{u_t}$.

Vanilla RNN encoder: The vanilla RNN encoder ($\mathrm{VGRU_E}$), introduced by SUTSKEVER and collab. [2014], is considered as a baseline encoder. In the vanilla encoder $\mathscr{E}_{u_i} = \frac{1}{|u_i|}\sum_{k=1}^{|u_i|}\mathscr{E}_{w_k^i}$ where $\mathscr{E}_{w_k^i}$ is an embedding of w_k^i. To better model dependencies between consecutive utterances, we use a bidirectional GRU [CHO and collab., 2014]:

$$\begin{aligned}
&\overrightarrow{h^s_{i-\mathrm{T}}} = \overleftarrow{h^s_{i-\mathrm{T}}} = \overrightarrow{0} \\
&\overrightarrow{h^s_t} = \overrightarrow{\mathrm{GRU}}(\mathscr{E}_{u_t}), t \in [i-\mathrm{T}, i] \\
&\overleftarrow{h^s_t} = \overleftarrow{\mathrm{GRU}}(\mathscr{E}_{u_t}), t \in [i, i-\mathrm{T}] \\
&\mathrm{H}_i = [\overleftarrow{h^s_i}, \overrightarrow{h^s_i}]
\end{aligned} \tag{4.1}$$

Hierarchical encoders: The vanilla encoder can be improved by computing $\mathscr{E}_{u_i}$ using bi-GRU. This hierarchical encoder (HGRU) is in line with the one introduced by SORDONI and collab. [2015]. Formally $\mathscr{E}_{u_i}$ is defined as it follows:

$$\begin{aligned}
&\overrightarrow{h^w_0} = \overleftarrow{h^w_0} = \overrightarrow{0} \\
&\overrightarrow{h^w_t} = \overrightarrow{\mathrm{GRU}}(\mathscr{E}_{w_t^i}), t \in [1, |u_i|] \\
&\overleftarrow{h^w_t} = \overleftarrow{\mathrm{GRU}}(\mathscr{E}_{w_t^i}), t \in [|u_i|, 1] \\
&\mathscr{E}_{u_i} = [\overleftarrow{h^w_{|u_i|}}, \overrightarrow{h^w_{|u_i|}}]
\end{aligned} \tag{4.2}$$

H_i is then computed using Equation 4.1. Intuitively, the first GRU layer (Equation 4.2) models dependencies between words (the hidden state of the word-level GRU is reset at each new utterance), and the second layer models dependencies between utterances.

Persona hierarchical encoders: In SwDA, a speaker turn can be splitted in several utterances. For example, if speaker A is interacting with speaker B we might encounter the sequence (AAABBBAA)[1]. We propose a novel Persona Hierarchical encoder (PersoHGRU) to better model speaker-utterance dependencies. We introduce a persona layer between the word and the sentence levels, see Figure 4.1:

$$\begin{aligned}
&\overrightarrow{h^p_t} = \begin{cases} \overrightarrow{0} \text{ if } t \text{ and } t-1 \text{ have different speakers} \\ \overrightarrow{\mathrm{GRU}}(\mathscr{E}_{u_{t-1}}) \end{cases} \\
&\overleftarrow{h^p_t} = \begin{cases} \overrightarrow{0} \text{ if } t \text{ and } t+1 \text{ have different speakers} \\ \overleftarrow{\mathrm{GRU}}(\mathscr{E}_{u_{t+1}}) \end{cases} \\
&\mathscr{E}^p_{u_k} = [\overrightarrow{h^p_k}, \overleftarrow{h^p_k}] \quad \forall k \in [i-\mathrm{T}, i]
\end{aligned} \tag{4.3}$$

H_i is then obtained following Equation 4.1 where $\mathscr{E}_{u_i}$ is replaced by $\mathscr{E}^p_{u_i}$.

4.4.2 Decoders

In this section, we introduce the different decoders we compare in our experiments. We introduce a novel form of attention that we name *guided attention*. *Guided*

[1] In SwDA arround two third of the sentence have at least a AA or BB

attention leverages the perfect alignment between input and output sequences ($\mathbb{D}_2$). The decoder computes the probability of the sequence of output tags based on:

$$p(y_{i-\mathrm{T}},\ldots,y_i|u_{i-\mathrm{T}},\ldots,u_i) = \prod_{k=i-\mathrm{T}}^{i} p(y_k|\mathrm{H}_i, y_{k-1},\ldots,y_{i-\mathrm{T}}) \tag{4.4}$$

Vanilla decoder: The vanilla decoder ($\mathrm{VGRU_D}$) is similar to the one introduced by SUTSKEVER and collab. [2014].

Decoders with attention: In NMT, the attention mechanism forces the seq2seq model to learn to focus on specific parts of the sequence each time a new word is generated and let the decoder correctly align the input sequence with output sequence. In our case, we follow the approach described by BAHDANAU and collab. [2014] and we define the context vector as:

$$c_k = \sum_{j=i-\mathrm{T}}^{i} \alpha_{j,k} h_j^s \tag{4.5}$$

where $\alpha_{j,k}$ scores how well the inputs around position k and the output at position j match. Since we have a perfect alignment ($\mathbb{D}_2$), we know a priori on which sequence the decoder needs to focus more at each time step. Taking into account this aspect of the problem, we propose three different attention mechanisms.

Vanilla attention: This attention represents our baseline attention mechanism and it is the one proposed by BAHDANAU and collab. [2014], where:

$$\alpha_{j,k} = softmax(a(h_{k-1}^{Dec}, h_j^s)) \tag{4.6}$$

and a is parametrized as a feedforward neural network.

Hard guided attention: The *hard guided attention* forces the decoder to focus only on the u_i while predicting y_i:

$$\alpha_{j,k} = \begin{cases} 0, & \text{if } k \neq j \\ 1, & \text{otherwise} \end{cases} \tag{4.7}$$

Soft guided attention: The *soft guided attention* guides the decoder to mainly focus on the u_i while predicting y_i, but allows it to have a limited focus on other parts of the input sequence.

$$\widetilde{\alpha}_{j,k} = \begin{cases} a(h_{k-1}^{Dec}, h_j^s), & \text{if } k \neq j \\ 1 + a(h_{k-1}^{Dec}, h_j^s), & \text{otherwise} \end{cases} \tag{4.8}$$

$$\alpha_{j,k} = softmax(\widetilde{\alpha}_{j,k}) \tag{4.9}$$

where a is parametrised as a feedforward neural network.

4.4.3 Training and inference

In this section, we describe the training and the inference strategies used for our models. A seq2seq model aims to find the best sentence for a given source sentence. This poses a computational challenge when the output vocabulary size is large, since even by using beam search it's expensive to explore multiple paths. Since our output vocabulary size is limited ($\mathbb{D}_3$), we do not incur in this problem and we can use beam search during both training and inference.

Beam search: In our work we measure the sequence likelihood based on the following formula:

$$s(\tilde{\boldsymbol{y}}^k, \boldsymbol{u}_i) = \frac{\log \mathrm{P}(\tilde{\boldsymbol{y}}^k | \boldsymbol{u}_i)}{lp(\tilde{\boldsymbol{y}}^k)} \tag{4.10}$$

where $\boldsymbol{u}_i = (u_{i-\mathrm{T}}, \ldots, u_i)$ and $\tilde{\boldsymbol{y}}^k = (\tilde{y}_{i-\mathrm{T}}, \ldots, \tilde{y}_{i-\mathrm{T}+k})$ is the current target, and $lp(\tilde{\boldsymbol{y}}) = \frac{(5+|\tilde{\boldsymbol{y}}|)^\alpha}{(5+1)^\alpha}$ is the length normalisation coefficient [Wu and collab., 2016]. At each time step the B most likely sequences are kept (B corresponding to the beam size).
Training objective: For training we follow [Edunov and collab., 2017] and train our model until convergence with a token level loss and fine tune it by minimising the expected risk $\mathscr{L}_{\mathrm{RISK}}$ defined as:

$$\mathscr{L}_{\mathrm{RISK}} = \sum_{\tilde{\boldsymbol{y}} \in \mathrm{U}(\mathrm{C}_i)} \frac{cost(\tilde{\boldsymbol{y}}, \boldsymbol{y}_i) p(\tilde{\boldsymbol{y}} | \boldsymbol{u}_i)}{\sum_{\tilde{\boldsymbol{y}}' \in \mathrm{U}(\boldsymbol{u}_i)} p(\tilde{\boldsymbol{y}} | \boldsymbol{u}_i)} \tag{4.11}$$

where $\mathrm{U}(\boldsymbol{u}_i)$ is the set of the sequences generated by the model using a beam search algorithm for the input $\boldsymbol{u}_i$, and $cost(\tilde{\boldsymbol{y}}, \boldsymbol{y}_i)$ is defined, for a given a candidate sequence $\tilde{\boldsymbol{y}}$ and a target $\boldsymbol{y}_i$, as:

$$cost(\tilde{\boldsymbol{y}}, \boldsymbol{y}_i) = \begin{cases} 1 & \text{if } \tilde{\boldsymbol{y}}_i = \boldsymbol{y}_i \\ 0 & \text{otherwise} \end{cases} \tag{4.12}$$

4.4.4 GRU/HGRU CRF baseline

State-of-the-art models [Kumar and collab., 2018; Li and collab., 2018a] use conditional random fields which model dependencies between tags on top of an GRU or a HGRU encoder which computed an embedding of the a variable number of utterances sentences. We have implemented our own CRF ($\mathrm{Baseline}_{\mathrm{CRF}}$) following the work of Li and collab. [2018a]:

$$p(\boldsymbol{y}|\boldsymbol{u}) = \frac{1}{\mathrm{Z}_\theta(\boldsymbol{u})} exp \left\{ \sum_{t=1}^{\mathrm{T}} \psi(y_t, y_{t-1}, \mathscr{E}_{u_t}) \right\} \tag{4.13}$$

Where $\boldsymbol{y} = y_1, \ldots, y_\mathrm{T}$, $\boldsymbol{u} = u_1, \ldots, u_\mathrm{T}$ and Z_θ is a normalisation function defined as follow:

$$\mathrm{Z}_\theta(\boldsymbol{x}) = \sum_{\boldsymbol{y} \in \mathscr{Y}} exp \left\{ \sum_{t=1}^{\mathrm{T}} \psi(y_t, y_{t-1}, \mathscr{E}_{u_t}) \right\} \tag{4.14}$$

Finally we define the feature function ψ as the sum of two scores potentials, namely:

$$\psi(y_t, y_{t-1}, \mathscr{E}_{u_t}) = \phi(\mathscr{E}_{u_t})[y_t] + \mathbf{T}_{y_t, y_{t-1}} \tag{4.15}$$

Where $\phi \colon \mathbb{R}^\mathrm{H} \rightarrow \mathbb{R}^\mathrm{K}$ is a linear transformation from the embedding space of utterances to an output space of dimension $\mathrm{K} = |\mathscr{Y}|$ and $\mathbf{T} \in \mathbb{R}^{\mathrm{K} \times \mathrm{K}}$ is the transition score matrix between labels.

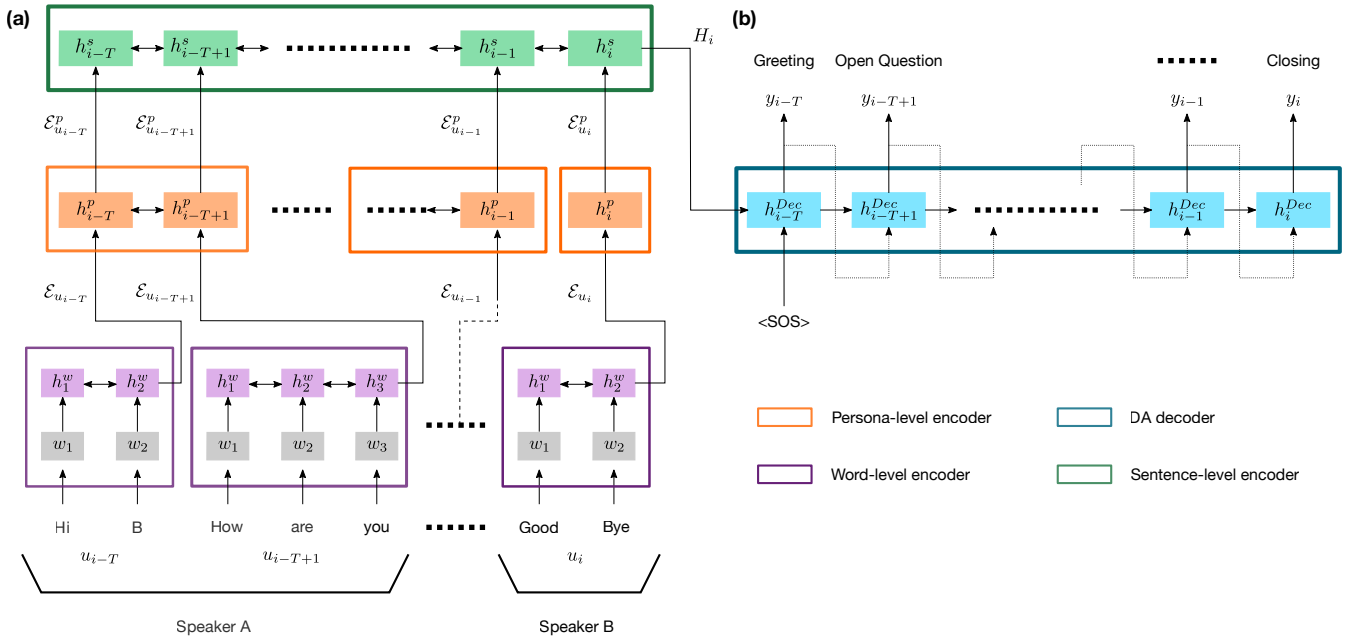


Figure 4.1 – Seq2seq model architecture forDA classification. (a) The encoder is composed with three different levels representing a different hierarchical level in the dialogue. The utterances are encoded at: word level (purple), persona level (orange) and sentence level (green). (b) The decoder (blue) is responsible to generate for each utterance aDA exploiting the last state of the encoder as initial hidden state.

4.5 Experiment Protocol

In this section we describe the experimental protocols adopted for the evaluation of our approach.

4.5.1 Datasets

We consider two classical datasets for Dialogue Act Classification: The Switchboard Dialogue Act Corpus and the MRDA. Since our models explicitly generate a sequence of tags we compute the accuracy on the last generated tag. Both datasets are already segmented in utterances and each utterance is segmented in words. For each dataset, we split each conversation C_i in sequence of utterances of length T = 5[2].

SwDA**:** The Switchboard-1 corpus is a telephone speech corpus [STOLCKE and collab., 1998], consisting of about 2.400 two-sided telephone conversation among 543 speakers with about 70 provided conversation topics. The dataset includes information about the speakers and the topics and has 42 different tags. In this dataset global dependency plays a key role due to the large amount of backchannel (19%), abandoned or turn-exit (5%), uninterpretable acts (1%). In this context, any models that only take into account local dependencies will fail at extracting information to distinguish between ambiguous tags. For the confusion matrix, we follow LI and collab. [2018a] and present it for 10 tags only: statement-non-opinion (sd), backchannel (b), statement-opinion (sv), conventional-closing (fc), wh-question (qw), response acknowledgement (bk), hedge (h), open-question (qo), other answers (no), thanking (ft).

MRDA**:** The ICSI Meeting Recorder Dialogue Act corpus [SHRIBERG and collab., 2004] contains 72 hours of naturally occurring multi-party meetings that were first converted into 75 word level conversations, and then hand-annotated with DAs using the Meeting Recorder Dialogue Act Tagset. In this work we use 5 DAs, i.e., statements (s), questions (q), floorgrabber (f), backchannel (b), disruption (d).

Train/Dev/Test Splits: For both SwDA and RMDA we follow the official split introduced by STOLCKE and collab. [2000]. Thus, our model can directly be compared to [CHEN and collab., 2018; KUMAR and collab., 2018; LI and collab., 2018a; RAHEJA and TETREAULT, 2019].

Dataset	|C|	|V|	Train	Val	Test
MRDA	5	10K	51(76K)	11(15K)	11(15K)
SwDA	42	19K	1003(173K)	112(22K)	19(9K)

Table 4.2 – Statistics for MRDA and SwDA. |C| is the number of Dialogue Act classes, |V| is the vocabulary size. Training, Validation and Testing indicate the number of conversations (number of utterances) in the respective splits.

Tags in SwDA: SwDA extends the Switchboard-1 corpus with tags from the SWBD-DAMSL tagset. The 220 tags were reduced to 42 tags. The resulting tags include dialogue acts like statement-non-opinion, acknowledge, statement-opinion, agree/accept, etc. The average speaker turns per conversation, tokens per conversation, and tokens per utterance are 195.2, 1,237.8, and 7.0, respectively.

[2] T is an hyperparameter, experiments have shown that 5 leads to the best results.

4.5.2 Training details

All the hyper-parameters have been optimised on the validation set using accuracy computed on the last tag of the sequence. The embedding layer is initialised with pre-trained fastText word vectors of size 300 [BOJANOWSKI and collab., 2017][3] , trained with subword information (on Wikipedia 2017, UMBC webbase corpus and statmt.org news dataset), and updated during training. Hyperparameter selection has been done using a random search on a fixed grid. Models have been implemented in PyTorch and trained on a single NVIDIA P100.
Parameters for SwDA**:** We used Adam optimizer [KINGMA and BA, 2014] with a learning rate of 0.01, which is updated using a scheduler with a patience of 20 epochs and a decrease rate of 0.5. The gradient norm is clipped to 5.0, weight decay is set to 1e−5, and dropout [LECUN and collab., 2015] is set to 0.2. The maximum sequence length is set to 20. Best performing model is an encoder with size of 128 and a decoder of size 48. For VGRU_E, we use two layers for the BiGRU layer. For hierarchical models, we use BiGRU with a single layer.
Parameters for MRDA**:** We used AdamW optimizer [LOSHCHILOV and HUTTER, 2017] with a learning rate of 0.001, which is updated using a scheduler with a patience of 15 epochs and a decrease rate of 0.5. The gradient norm is clipped to 5.0, weight decay is set to 5e−5, and dropout [LECUN and collab., 2015] is set to 0.3. The maximum sequence length is set to 30. Best performing model is an encoder with size of 40 and a decoder with size 400. For VGRU_E we use two layers for the BiGRU layer, for hierarchical models we use BiGRU with a single layer.

4.6 Experiments

In this section we propose a set of experiments in order to investigate the performance of our model compared to existing approaches with respect to the difficulties highlighted in the introduction.

4.6.1 Experiment 1: Are Seq2seq better suited to DA prediction than CRF ?

Current state of the art are built on CRF models. In this first section, we aim at comparing a seq2seq with a CRF based model. To provide a fair comparison we perform the same grid search for all models on a fixed grid. At this step, we do not use attention neither use beam search during training or inference. As shown in Table 4.3, with a vanilla RNN encoder the seq2seq significantly outperforms the CRF on SwDa and MRDA. With an HGRU the seq2seq exhibit significantly higher results on SwDA and reaches comparable performances on MRDA. This behaviour suggests that a model based on a seq2seq architecture tends to be achieve higher score on DA classification than a CRF based model.
Global dependencies analysis: In Table 6.4 we present two examples where our seq2seq use contextual information to disambiguate the tag and to predict the correct label. In the first example, “It can be a pain" without context can be interpreted both as statement non-opinion (sd) or statement opinion (sv). Our seq2seq uses

[3] In our work we rely on same pre-trained embedding word2vect [MIKOLOV and collab., 2013] instead of GloVe [PENNINGTON and collab., 2014].

Models	SwDa	MRDA
Baseline$_{CRF}$ (+GRU)	77.7	88.3
seq2seq (+GRU)	**81.9**	**88.5**
Baseline$_{CRF}$ (+HGRU)	81.6	90.0
seq2seq (+HGRU)	**82.4**	90.0

Table 4.3 – Accuracy of a seq2seq on dev test and Baseline$_{CRF}$ on SwDA and MRDA. Bold results exhibit significant differences (p-value < 0.01) according to the Wilcoxon Mann Whitney test performed on 10 runs using different seeds.

the surrounding context (two sentences before) to disambiguate and assign the sv label . In the second example, the correct tag assigned to "Oh, okay" is a response acknowledgement (bk) and not backchannel (b). The key difference between bk and b is that an utterance labelled with bk has to be produced within a question-answer context, whereas b is a *continuer* [4]. In our example, the global context this is a question/reply situation: the first speaker asks a question ("What school is it"), the second replies then, the first speaker answers to the reply. This observation reflects the fact CRF models only handle local dependencies where seq2seq models consider global ones as well.

Utterances	G.	seq2seq	CRF
How long does that take you to get to work?	qw	qw	qw
Uh, about forty-five, fifty minutes.	sd	sd	sd
How does that work, work out with, uh, storing your bike and showering and all that?	qw	qw	qw
Yeah ,	b	b	b
It can be a pain .	sd	sd	sv
It's, it's nice riding to school because it's all along a canal path, uh,	sd	sd	sd
Because it's just, it's along the Erie Canal up here.	sd	sd	sd
So, what school is it?	qw	qw	qw
Uh, University of Rochester.	sd	sd	sd
Oh, okay.	bk	bk	b

Table 4.4 – Example of predicted sequence of tags taken from SwDA. seq2seq is our best performing model, CRF stands for Baseline$_{CRF}$, G. is the groundtruth label.

4.6.2 Experiment 2: What is the best encoder?

In Table 4.5, we present the results of the three encoders presented in Section 4 on both datasets. ForSwDA and MRDA, we observe that a seq2seq equipped with a hierarchical encoder outperforms models with Vanilla RNN encoder, while reducing the number of learned parameters.

The VGRU$_D$ does not play well with the PersoHGRU encoder. When combined with a *guided attention mechanism*, the PersoHGRU exhibits competitive accuracy

[4]This analysis can be supported by 5.1.1 in SwDA coder manual https://web.stanford.edu/~jurafsky/ws97/manual.august1.html

	SwDA			
Enc. \ Dec.	VGRU_D	att.	soft guid.	hard guid.
Beam Size	1	1	1	1
VGRU_E	81.6	82.1	82.8	82.9
HGRU	82.4	82.3	83.1	**84.0**
PersoHGRU	49.8	79.4	84.0	83.5
	MRDA			
Enc. \ Dec.	VGRU_D	att.	soft guid.	hard guid.
Beam Size	1	1	1	1
VGRU_E	88.5	88.5	88.5	88.5
HGRU	90.0	89.9	90.0	**90.2**
PersoHGRU	66.2	87.7	88.2	86.9

Table 4.5 – Accuracy on the dev set of the different encoder/decoder combination MRDA andSwDA. ForSwDA, Wilcoxon test (10 runs with different seeds) has been performed for an HGRU encoder with a decoder with *hard guided attention* against an HGRU encoder with *soft guided attention, soft guided attention*, with attention, without attention pairwise tests exhibit p-value < 0.01.

on SwDA. However on MRDA, adding a personna layer harms the accuracy. This suggests either that the information related to the speaker is irrelevant for our task (no improvement observed while adding persona information) [5], or that the considered hierarchy is not the optimal structure to leverage this information.

Our final model makes use of the HGRU encoder since in most of the settings it exhibits superior performance.

4.6.3 Experiment 3: Which attention mechanism to use?

The seq2seq encodes a source sentence into a fixed-length vector from which a decoder generates a sequence of tags. Attention forces the decoder to strengthen its focus on the source sentences that are relevant to predicting a label.

In NMT [LUONG and collab., 2015], complementing a seq2seq with attention contributes to generate better sentences. In Table 4.5 we see that in most the case, the use of a simple attention mechanism provides a rather small improvement with VGRU and harms a bit the performances with a HGRU encoder. In case of a seq2seq composed with a PersoHGRU and a decoder without attention the learning fails: the decrease of the training loss is relatively small and seq2seq fails to generalise. It appears that inDA classification where sequences are short (5 tags), Vanilla attention does not have as much as impact as in NMT (that have longer sequences with more complex global dependencies).

If we consider an HGRU encoder, we observe that our proposed *guided attention* mechanisms improves dev accuracy which demonstrates the importance of easing the task by using prior knowledge on the alignment between the utterances and the tags. Indeed, while decoding there is a direct correspondence between labels and utterances meaning that y_i is associated with u_i. The soft guided attention will mainly focus on the current utterance with a small additional focus on the context

[5] Further investigations with several persona based model inspired from the work of LI and collab. [2016b] shows the same poor improvement (in terms of accuracy)

where hard guided attention will only consider the current utterance. Improvement due to *guided attention* demonstrates that the alignment between input/output is a key prior to include in our model.

Attention analysis: Figure 4.2 shows a representative example of the attention weights of the three different mechanisms. The seq2seq with a normal attention mechanism is characterised by a weight matrix far from the identity (especially the lower right part). While decoding the last tags, this lack of focus leads to a wrongly predicted label for a simple utterance: "Uh-Huh" (backchannel). Both *guided attention* mechanisms focus more on the sentence associated with the tag, at each time step, and predict successfully the last`DA`.

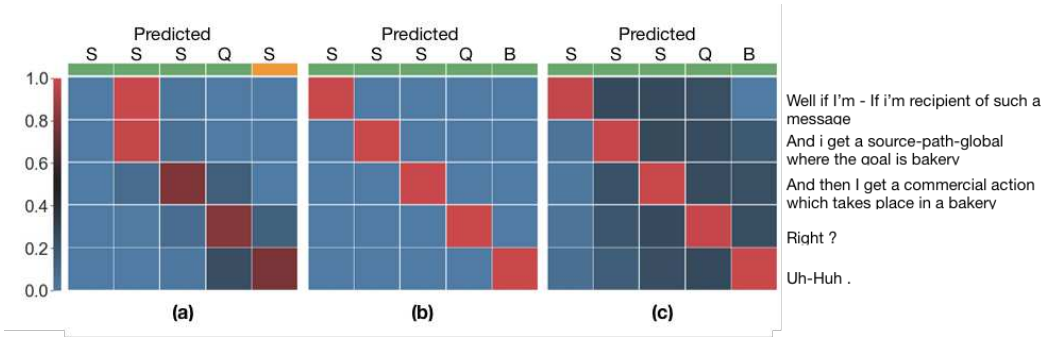


Figure 4.2 – Attention matrix visualisation on `MRDA` for the fixed context of 5 utterances. Green color for predicted label indicates a correct label, orange color indicates a mistake. (a) stands for the HGRU with attention, (b) stands for the HGRU with hard guided attention, (c) is HGRU with soft guided attention.

Since the *hard guided attention* decoder exhibit overall the best results (on both`SwDA` and `MRDA`) and does not require any additional parameter we will use it for our final model.

4.6.4 Experiment 4: How to leverage beam search to improve the performance?

Beam Search allows the seq2seq model to consider alternative paths in the decoding phase.

Beam Search during inference: Using beam search provides a low improvement (maximum absolute improvement of 0.2%) [6].

Compared to NMT, output size is drastically smaller ($\mathscr{Y}_{\texttt{SwDA}} = 42$ while $\mathscr{Y}_{\texttt{MRDA}} = 5$) for`DA` classification. When considering alternative paths with small output space in imbalanced datasets the beam search is more likely to consider very unlikely sequences as alternatives (eg. "`s s s s s`").

Fine tuning with a sequence loss: As previously mentioned, using beam search during inference only leads to a limited improvement in accuracy. We finetune a seq2seq composed with a HGRU encoder and a decoder with *hard guided attention* (this model has been selected in the previous steps) with the introduced sequence level loss describes in Section 4.4.3. Table 4.6 shows that this fine tuning steps

[6] The considered beam size are small compared to other applications [Li and collab., 2016a]. While increasing the beam size, we see that the beam search become very conservative [Gimpel and collab., 2013] and tends to output labels highly represented in the training set (e.g., `sd` for`SwDA`).

improves the performances of 1% on SwDA (84% vs 85%) and 1.2% on MRDA (90.4% vs 91.6%).

	SwDA		RMDA	
B_{inf} \ B_{train}	2	5	2	5
1	84.8	84.7	91.3	**91.6**
2	84.9	84.8	91.3	91.6
5	**85.0**	84.9	91.5	91.6

Table 4.6 – Accuracy on the dev set of seq2seq model trained with sequence level loss. B_{train} stands for the beam size during training, B_{inf} for the one during inference[7].For SwDA, Wilcoxon test (10 runs with different seeds) has been performed for $B_{train} = 2$ and $B_{inf} = 2$ against all other models. For MRDA, Wilcoxon test has been performed (10 runs with different seeds) for $B_{train} = 5$ and $B_{inf} = 1$ against all model with $B_{train} = 2$.

seq2seq$_{BEST}$: Our *seq2seq*$_{BEST}$ model is composed of a HGRU encoder and a decoder with *hard guided attention* finetuned with $B_{train} = 2$ and $B_{inf} = 5$ for SwDA and $B_{train} = 5$ and $B_{inf} = 1$ for SwDA.

4.6.5 Experiment 5: Comparison with state-of-the-art models

In this section, we compare the performances of *seq2seq*$_{BEST}$ with other state of the art models and analyse the performances of the models. Table 4.7 shows the performances of best performing model seq2seq$_{BEST}$ on the test set. seq2seq$_{BEST}$ achieves an accuracy of 85% on the SwDA corpora. This model outperforms [Chen and collab., 2018] and [Raheja and Tetreault, 2019] which achieve an accuracy of 82.9%. On MRDA, our best performing model reaches an accuracy of 91.6% where current state-of-the-art systems, [Chen and collab., 2018; Kumar and collab., 2018] achieve respectively 92.2% and 91.7%.

Models	SwDA	MRDA
[Li and collab., 2018a]	82.9	92.2
[Chen and collab., 2018]	81.3	91.7
[Kumar and collab., 2018]	79.2	90.9
[Raheja and Tetreault, 2019]	82.9	91.1
seq2seq$_{BEST}$	*85.0*	*91.6*

Table 4.7 – Accuracy of our best models (seq2seq) and Baseline$_{CRF}$ on SwDA and MRDA test sets.

4.6.6 Error analysis

The confusion matrix on SwDA (see Figure 4.3) illustrates that our model faces same difficulties as human annotator: sd is often confused with sv, bk with b, qo with qw. Due to high imbalance of SwDA, our system fails to recognise underrepresented labels (e.g. no and ft).

The confusion Matrix on MRDA shows that, here, the DA classification is easier compared to SwDA with fewer tags and classes that are more easily distinguished. *seq2seq*$_{BEST}$ reaches a perfect score at recognising questions. One of the reasons for

the mislabelling between backchannel (b) and statement (s) is that the MRDA dataset is highly imbalanced, with more than 50% of the utterances labelled as class s.

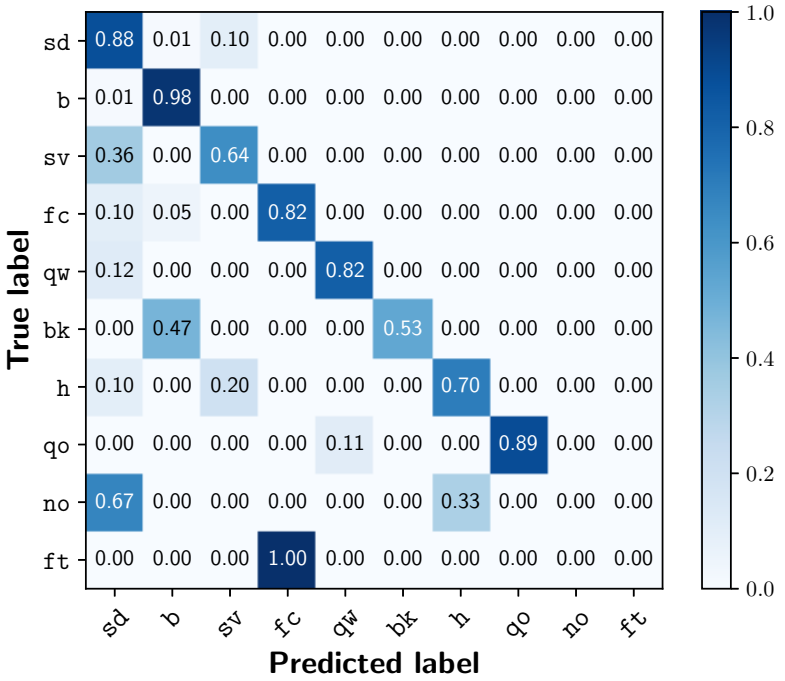


Figure 4.3 – Confusion Matrix for our best performing seq2seq model on SwDA for 10 out of 42 tags. For label designation see Section 4.5.1.

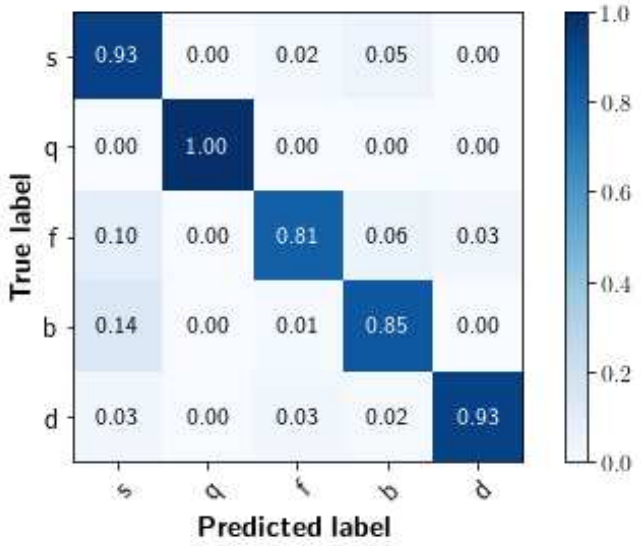


Figure 4.4 – Confusion Matrix for our best performing seq2seq model on MRDA. For label designation see Section 4.5.1.

Chapter 4 Conclusion

In this work, we have addressed the RQ1.1 and proposed a novel approach to the sequence labelling task in spoken dialogue. We have shown that our seq2seq model, using a newly devised *guided attention* mechanisms, achieves state-of-the-art results thanking its ability to better model global dependencies. However the proposed architecture need to be trained in a fully supervised manner and required large annotated corpora only available for DA. Theses shortcomings have led to RQ1.2 and are addressed in the next chapter.

4.7 References

BAHDANAU, D., K. CHO and Y. BENGIO. 2014, "Neural machine translation by jointly learning to align and translate", *CoRR*, vol. abs/1409.0473. URL http://arxiv.org/abs/1409.0473. 68

BOJANOWSKI, P., E. GRAVE, A. JOULIN and T. MIKOLOV. 2017, "Enriching word vectors with subword information", *Transactions of ACL*, vol. 5, doi: 10.1162/tacl_a_00051, p. 135–146. URL https://www.aclweb.org/anthology/Q17-1010. 72

BOTHE, C., C. WEBER, S. MAGG and S. WERMTER. 2018, "A context-based approach for dialogue act recognition using simple recurrent neural networks", *CoRR*, vol. abs/1805.06280. 65

CHEN, Z., R. YANG, Z. ZHAO, D. CAI and X. HE. 2018, "Dialogue act recognition via crf-attentive structured network", in *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, ACM, p. 225–234. 64, 65, 71, 76

CHO, K., B. VAN MERRIENBOER, Ç. GÜLÇEHRE, F. BOUGARES, H. SCHWENK and Y. BENGIO. 2014, "Learning phrase representations using RNN encoder-decoder for statistical machine translation", *CoRR*, vol. abs/1406.1078. 65, 67

EDUNOV, S., M. OTT, M. AULI, D. GRANGIER and M. RANZATO. 2017, "Classical structured prediction losses for sequence to sequence learning", *CoRR*, vol. abs/1711.04956. URL http://arxiv.org/abs/1711.04956. 69

GIMPEL, K., D. BATRA, C. DYER and G. SHAKHNAROVICH. 2013, "A systematic exploration of diversity in machine translation", in *Proceedings of EMNLP 2013*, p. 1100–1111. 75

JANIN, A., D. BARON, J. EDWARDS, D. ELLIS, D. GELBART, N. MORGAN, B. PESKIN, T. PFAU, E. SHRIBERG, A. STOLCKE and collab.. 2003, "The icsi meeting corpus", in *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).*, vol. 1, IEEE, p. I–I. 65

JOZEFOWICZ, R., W. ZAREMBA and I. SUTSKEVER. 2015, "An empirical exploration of recurrent network architectures", in *International Conference on Machine Learning*, p. 2342–2350. 65

KEIZER, S., R. OP DEN AKKER and A. NIJHOLT. 2002, "Dialogue act recognition with bayesian networks for dutch dialogues", in *Proceedings of the Third SIGdial Workshop on Discourse and Dialogue*. 65

KHANPOUR, H., N. GUNTAKANDLA and R. NIELSEN. 2016, "Dialogue act classification in domain-independent conversations using a deep recurrent neural network", in *COLING*. 65

KINGMA, D. P. and J. BA. 2014, "Adam: A method for stochastic optimization", *arXiv preprint arXiv:1412.6980*. 72

KUMAR, H., A. AGARWAL, R. DASGUPTA and S. JOSHI. 2018, "Dialogue act sequence labeling using hierarchical encoder with crf", in *Thirty-Second AAAI Conference on Artificial Intelligence*. 65, 69, 71, 76

LECUN, Y., Y. BENGIO and G. HINTON. 2015, "Deep learning", *nature*, vol. 521, nº 7553, p. 436. 72

LI, J., M. GALLEY, C. BROCKETT, J. GAO and W. B. DOLAN. 2016a, "A diversity-promoting objective function for neural conversation models", in *HLT-NAACL*. 75

LI, J., M. GALLEY, C. BROCKETT, G. SPITHOURAKIS, J. GAO and B. DOLAN. 2016b, "A persona-based neural conversation model", in *Proceedings of the 54th Annual Meeting of ACL (Volume 1: Long Papers)*, Association for Computational Linguistics, p. 994–1003. 74

LI, R., C. LIN, M. COLLINSON, X. LI and G. CHEN. 2018a, "A dual-attention hierarchical recurrent neural network for dialogue act classification", *CoRR*. 64, 65, 69, 71, 76

LI, Z., J. CAI, S. HE and H. ZHAO. 2018b, "Seq2seq dependency parsing", in *Proceedings of the 27th International Conference on Computational Linguistics*, p. 3203–3214. 66

LOSHCHILOV, I. and F. HUTTER. 2017, "Fixing weight decay regularization in adam", *arXiv preprint arXiv:1711.05101*. 72

LUONG, T., H. PHAM and C. D. MANNING. 2015, "Effective approaches to attention-based neural machine translation", in *Proceedings of EMNLP 2015*, Association for Computational Linguistics, Lisbon, Portugal, p. 1412–1421, doi: 10.18653/v1/D15-1166. URL https://www.aclweb.org/anthology/D15-1166. 65, 74

MIKOLOV, T., I. SUTSKEVER, K. CHEN, G. S. CORRADO and J. DEAN. 2013, "Distributed representations of words and phrases and their compositionality", in *NeurIPS*, p. 3111–3119. 72

P. COLOMBO & E. CHAPUIS, M. MANICA, E. VIGNON, G. VARNI and C. CLAVEL. 2020, "Guiding attention in sequence-to-sequence models for dialogue act prediction", *CoRR*, vol. abs/2002.08801. URL https://arxiv.org/abs/2002.08801. 63

PENNINGTON, J., R. SOCHER and C. MANNING. 2014, "Glove: Global vectors for word representation", in *Proceedings of EMNLP 2014*, p. 1532–1543. 72

RAHEJA and TETREAULT. 2019, "Dialogue act classification with context-aware self-attention", *CoRR*, vol. abs/1904.02594. URL http://arxiv.org/abs/1904.02594. 64, 65, 71, 76

SHRIBERG, E., R. DHILLON, S. BHAGAT, J. ANG and H. CARVEY. 2004, "The icsi meeting recorder dialog act (mrda) corpus", in *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue at HLT-NAACL 2004*. 71

SORDONI, A., Y. BENGIO, H. VAHABI, C. LIOMA, J. GRUE SIMONSEN and J.-Y. NIE. 2015, "A hierarchical recurrent encoder-decoder for generative context-aware query suggestion", in *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, ACM, p. 553–562. 67

STOLCKE, A., K. RIES, N. COCCARO, E. SHRIBERG, R. BATES, D. JURAFSKY, P. TAYLOR, R. MARTIN, C. V. ESS-DYKEMA and M. METEER. 2000, "Dialogue act modeling for automatic tagging and recognition of conversational speech", *Computational linguistics*, vol. 26, nº 3, p. 339–373. 65, 71

STOLCKE, A., E. SHRIBERG, R. BATES, N. COCCARO, D. JURAFSKY, R. MARTIN, M. METEER, K. RIES, P. TAYLOR, C. VAN ESS-DYKEMA and collab.. 1998, "Dialog act modeling for conversational speech", in *AAAI Spring Symposium on Applying Machine Learning to Discourse Processing*, p. 98–105. 65, 71

SURENDRAN, D. and G.-A. LEVOW. 2006, "Dialog act tagging with support vector machines and hidden markov models", in *Ninth International Conference on Spoken Language Processing*. 65

SUTSKEVER, I., O. VINYALS and Q. V. LE. 2014, "Sequence to sequence learning with neural networks", in *NeurIPS*, p. 3104–3112. 64, 65, 67, 68

VIJAYAKUMAR, A. K., M. COGSWELL, R. R. SELVARAJU, Q. SUN, S. LEE, D. CRANDALL and D. BATRA. 2016, "Diverse beam search: Decoding diverse solutions from neural sequence models", *arXiv preprint arXiv:1610.02424*. 65

WISEMAN, S. and A. M. RUSH. 2016, "Sequence-to-sequence learning as beam-search optimization", in *EMNLP*. 65

WU, Y., M. SCHUSTER, Z. CHEN, Q. V. LE, M. NOROUZI, W. MACHEREY, M. KRIKUN, Y. CAO, Q. GAO, K. MACHEREY and collab.. 2016, "Google's neural machine translation system: Bridging the gap between human and machine translation", *arXiv preprint arXiv:1609.08144*. 65, 69

Chapter 5

A Pre-trained Model for Low-resource DA and E/S Classification Tasks

Chapter 5 Abstract

In the previous chapter, we saw that a direct supervised approach requires a huge amount of labelled data which are not always available. In this chapter, we address the data scarcity problem through the light of pre-trained models. We propose a new approach to learn generic representations adapted to spoken dialogue, which we evaluate on a new benchmark we call Sequence labellIng evaLuatIon benChmark fOr spoken laNguagE benchmark (SILICONE). SILICONE[a] is model-agnostic and contains 10 different datasets of various sizes annotated in DA or E/S. We obtain our representations with a hierarchical encoder based on transformer architectures, for which we extend two well-known pre-training objectives. Pre-training is performed on OpenSubtitles: a large corpus of spoken dialogue containing over 2.3 billion of tokens. We demonstrate how hierarchical encoders achieve competitive results with consistently fewer parameters compared to state-of-the-art models and we show their importance for both pre-training and fine-tuning. This work has been presented in E. Chapuis & P. Colombo and collab. [2020].

[a] Benchmark can be found in the dataset library from HuggingFace [Wolf and collab., 2020] at https://huggingface.co/datasets/silicone

5.1 Introduction

As showcased in Chapter 1, it is beneficial for dialogue systems to analyse user's utterances in terms of DA and E/S. In Chapter 4, we saw that DA classification is done through sequence labelling systems that are usually trained on large corpora (with over $100k$ labelled utterances) such as Switchboard [GODFREY and collab., 1992] or MRDA [SHRIBERG and collab., 2004] which may not be always available. Moreover, even though large corpora enable learning complex models from scratch (e.g., seq2seq Chapter 4), those models are very specific to the labelling scheme employed. Adapting them to different sets of emotions or dialogue acts would require more annotated data. Meanwhile, generic representations [DEVLIN and collab., 2018; LIU and collab., 2019; MIKOLOV and collab., 2013; PENNINGTON and collab., 2014; PETERS and collab., 2018; YANG and collab., 2019] have been shown to be an effective way to adapt models across different sets of labels. Those representations are usually trained on large written corpora such as OSCAR [SUÁREZ and collab., 2019], Book Corpus [ZHU and collab., 2015] or Wikipedia [DENOYER and GALLINARI, 2006]. Although achieving state-of-the-art (SOTA) results on written benchmarks [WANG and collab., 2018], they are not tailored to spoken dialogue (SD). Indeed, [TRAN and collab., 2019] have suggested that training a parser on conversational speech data can improve results, due to the discrepancy between spoken and written language (e.g., disfluencies [STOLCKE and SHRIBERG, 1996], fillers [DINKAR and collab., 2020; SHRIBERG, 1999], different data distribution). Furthermore, capturing discourse-level features, which distinguish dialogue from other types of text [THORNBURY and SLADE, 2006], e.g., capturing multi-utterance dependencies, is key to embed dialogue that is not explicitly present in pre-training objectives [DEVLIN and collab., 2018; LIU and collab., 2019; YANG and collab., 2019], as they often treat sentences as a simple stream of tokens.

The goal of this work is to train on SD data a generic dialogue encoder capturing discourse-level features that produce representations adapted to spoken dialogue. We evaluate these representations on both DA and E/S labelling through a new benchmark SILICONE (Sequence labellIng evaLuatIon benChmark fOr spoken laNguagE) composed of datasets of varying sizes using different sets of labels. We place ourselves in the general trend of using smaller models to obtain lightweight representations [JIAO and collab., 2019; LAN and collab., 2019] that can be trained without a costly computation infrastructure while achieving good performance on several downstream tasks [HENDERSON and collab., 2020]. Concretely, since hierarchy is an inherent characteristic of dialogue [THORNBURY and SLADE, 2006], we propose the first hierarchical generic multi-utterance encoder based on a hierarchy of transformers. This allows us to factorise the model parameters, getting rid of long term dependencies and enabling training on a reduced number of GPUs. Based on this hierarchical structure, we generalise two existing pre-training objectives. As embeddings highly depend on data quality [LE and collab., 2019] and volume [LIU and collab., 2019], we preprocess OpenSubtitles [LISON and collab., 2019]: a large corpus of spoken dialogue from movies. This corpora is an order of magnitude bigger than corpora [BUDZIANOWSKI and collab., 2018b; DANESCU-NICULESCU-MIZIL and LEE, 2011; LOWE and collab., 2015] used in previous works [HAZARIKA and collab., 2019; MEHRI and collab., 2019]. Lastly, we evaluate our encoder along with other baselines on SILICONE, which lets us draw finer conclusions of the generalisation

capability of our models[1].

5.2 Method

5.2.1 Pre-training Objectives

In the following, we rely on the notations defined in subsection 3.1.1. Our work builds upon existing objectives designed to pre-train encoders: the Masked Language Model (MLM) from [DEVLIN and collab., 2018; LAN and collab., 2019; LIU and collab., 2019; ZHANG and collab., 2019a] and the Generalized Autoregressive Pre-training (GAP) from [YANG and collab., 2019].

MLM **Loss**: The MLM loss corrupts sequences (or in our case, utterances) by masking a proportion p_ω of tokens. The model learns bidirectional representations by predicting the original identities of the masked-out tokens. Formally, for an utterance u_i, a random set of indexed positions m^{u_i} is selected and the associated tokens are replaced by a masked token [MASK] to obtain a corrupted utterance u_i^{masked}. The set of parameters θ is learnt by maximizing :

$$\mathscr{L}^u_{\text{MLM}}(\theta, u_i) = \mathbb{E}\left[\sum_{t \in m^{u_i}} \log(p_\theta(\omega^i_t | \tilde{u}_i))\right] \tag{5.1}$$

where $\tilde{u}_i$ is the corrupted utterance, $m^{u_i}_j \sim unif\{1, |u_i|\}\ \forall\ j \in [1, p_\omega]$ and p_ω is the proportion of masked tokens.

GAP **Loss**: the GAP loss consists in computing a classic language modelling loss across different factorisation orders of the tokens. In this way, the model will learn to gather information across all possible positions from both directions. The set of parameters θ is learnt by maximising:

$$\mathscr{L}^u_{\text{GAP}}(\theta, u_i) = \mathbb{E}\left[\mathbb{E}_{\mathbf{z} \sim \mathbb{Z}_{|u_i|}}\left[\sum_t \log p_\theta(\omega^i_{z_t} | u_i^{\mathbf{z}<t})\right]\right] \tag{5.2}$$

where $\mathbb{Z}_{|u_i|}$ is the set of permutations of length $|u_i|$ and $u_i^{\mathbf{z}<t}$ represent the first t tokens of u_i when permuting the sequence according to $\mathbf{z} \in \mathbb{Z}_{|u_i|}$.

5.2.2 Hierarchical Encoding

Capturing dependencies at different granularity levels is key for dialogue embedding. Thus, we choose a hierarchical encoder [CHEN and collab., 2018b; LI and collab., 2018a]. It is composed of two functions f^u and f^c, satisfying:

$$\mathscr{E}_{u_i} = f^u_\theta(\omega_1, \ldots, \omega_{|u_i|}) \tag{5.3}$$

$$\mathscr{E}_{C_j} = f^d_\theta(\mathscr{E}_{u_1}, \ldots, \mathscr{E}_{C_j}) \tag{5.4}$$

where $\mathscr{E}_{u_i} \in \mathbb{R}^{d_u}$ is the embedding of u_i and $\mathscr{E}_{C_j} \in \mathbb{R}^{d_d}$ the embedding of C_j. The structure of the hierarchical encoder is depicted in Figure 5.1.

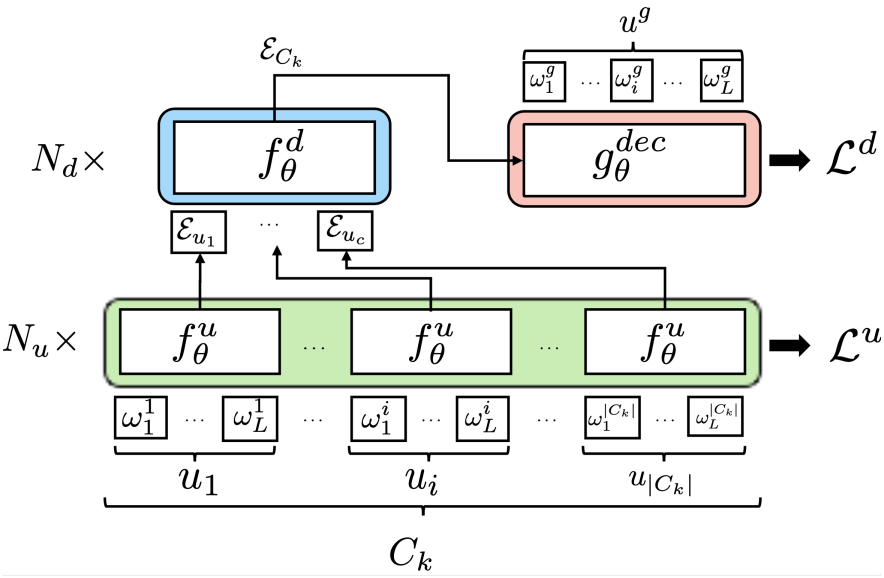


Figure 5.1 – General structure of our proposed hierarchical dialogue encoder, with a decoder: f_θ^u, f_θ^d and the sequence label decoder (g_θ^{dec}) are colored respectively in green, blue and red.

5.2.3 Hierarchical Pre-training

General Motivation

Current self-supervised pre-training objectives such as MLM and GAP are trained at the sequence level, which for us translates to only learning f_θ^u. In this section, we extend both the MLM and GAP losses at the dialogue level in order to pre-train f_θ^d. Following previous work on both multi-task learning [ARGYRIOU and collab., 2007; RUDER, 2017] and hierarchical supervision [GARCIA and collab., 2019; SANH and collab., 2019], we argue that optimising simultaneously at both levels rather than separately improves the quality of the resulting embeddings. Thus, we write our global hierarchical loss as:

$$\mathscr{L}(\theta) = \lambda_u * \mathscr{L}^u(\theta) + \lambda_d * \mathscr{L}^d(\theta) \tag{5.5}$$

where $\mathscr{L}^u(\theta)$ is either the MLM or GAP loss at the utterance level and $\mathscr{L}^d(\theta)$ is its generalisation at the dialogue level.

MLM Loss

The MLM loss at the utterance level is defined in Equation 6.5. Our generalisation at the dialogue level masks a proportion $p_{\mathscr{C}}$ of utterances and generates the sequences of masked tokens. Thus, at the dialogue level the MLM loss is defined as:

$$\mathscr{L}^d_{\text{MLM}}(\theta, \mathrm{C}_k) = \mathbb{E}\left[\sum_{j \in m^{\mathrm{C}_k}} \sum_{i=1}^{|u_j|} \log(p_\theta(\omega_i^j | \tilde{\mathrm{C}}_k))\right] \tag{5.6}$$

where $m_j^{\mathrm{C}_k} \sim unif\{1, |\mathrm{C}_k|\}\ \forall\ j \in [1, p_{\mathscr{C}}]$ is the set of positions of masked utterances in the context C_k, $\tilde{\mathrm{C}}_k$ is the corrupted context, and $p_{\mathscr{C}}$ is the proportion of masked utterances. We propose a visual illustration of the corrupted context Figure 5.2 by the MLM Loss.

[1] Upon publication, we will release the code, models and especially the preprocessing scripts to replicate our results.

(a) Initial context composed by 5 utterances.

(b) u_1 is chosen to be masked.

(c) Corrupted context with utterance u_1 masked.

(d) u_4 is chosen to be masked.

(e) Corrupted context with utterance u_4 masked.

Figure 5.2 – This figure shows an example of corrupted context. Here p_C is randomly set to 2 meaning that two utterances will be corrupted. u_1 and u_4 are randomly picked in 5.2b, 5.2d and then masked in 5.2c, 5.2e.

GAP Loss

The GAP loss at the utterance level is defined in Equation 5.2. A possible generalisation of the GAP at the dialogue level is to compute the loss of the generated utterance across all factorization orders of the context utterances. Formally, the GAP loss is defined at the dialogue level as:

$$\mathscr{L}^{d}_{\texttt{GAP}}(\theta, \mathrm{C}_k) = \mathbb{E}\left[\mathbb{E}_{\mathbf{z}\sim\mathbb{Z}_\mathrm{T}}\left[\sum_{t=1}^{|\mathrm{C}_k|}\sum_{i=1}^{|u_{z_t}|}\log p_\theta(\omega_i^{z_t}|\mathrm{C}_k^{\mathbf{z}<t})\right]\right] \tag{5.7}$$

where $\omega_i^{z_t}$ denotes the first i-th tokens of the permuted t-th utterance when permuting the context according to $\mathbf{z} \in \mathbb{Z}_\mathrm{T}$ and $\mathrm{C}_k^{\mathbf{z}<t}$ the first t utterances of C_k when permuting the context according to $\mathbf{z}$.

5.2.4 Architecture

Commonly, the functions f^u_θ and f^d_θ are either modelled with recurrent cells (Chapter 4) and Chapter 4 or Transformer blocks [VASWANI and collab., 2017]. Transformer blocks are more parallelizable, offering shorter paths for the forward and backward signals and requiring significantly less time to train compared to recurrent layers. To

the best of our knowledge this is the first attempt to pre-train a hierarchical encoder based only on transformers[2].
The structure of the model can be found in Figure 5.1. In order to optimize dialogue level losses as described in Equation 6.1, we generate (through g_θ^{dec}) the sequence with a Transformer Decoder ($\mathscr{T}_{dec}$). For downstream tasks, the context embedding $\mathscr{E}_{C_k}$ is fed to a simple MLP (simple classification), or to a CRF/GRU/LSTM (sequential prediction). In the remainder, we will name our hierarchical transformer-based encoder $\mathscr{HT}$ and the hierarchical RNN-based encoder $\mathscr{HR}$. We use θ_y^x to refer to the set of model parameters learnt using the pre-training objective y (either `MLM` or `GAP`) at the level x[3].

5.2.5 Pre-training Datasets

Datasets used to pre-train dialogue encoders [HAZARIKA and collab., 2019; MEHRI and collab., 2019] are often medium-sized (e.g. Cornell Movie Corpus [DANESCU-NICULESCU-MIZIL and LEE, 2011], Ubuntu [LOWE and collab., 2015], MultiWOz [BUDZIANOWSKI and collab., 2018a]). In our work, we focus on OpenSubtitles [LISON and TIEDEMANN, 2016][4] because (1) it contains spoken language, contrarily to the Ubuntu corpus [LOWE and collab., 2015] based on logs; (2) as Wizard of Oz [BUDZIANOWSKI and collab., 2018a] and Cornell Movie Dialog Corpus [DANESCU-NICULESCU-MIZIL and LEE, 2011], it is a multi-party dataset; and (3) OpenSubtitles is an order of magnitude larger than any other spoken language dataset used in previous work. We segment OpenSubtitles by considering the duration of the silence between two consecutive utterances. Two consecutive utterances belong to the same conversation if the silence is shorter than δ_T[5]. Conversations shorter than the context size T are dropped[6]. After preprocessing, Opensubtitles contains subtitles from 446520 movies or series which represent 54642424 conversations and over 2.3 billion of words.

5.2.6 Baseline Encoder

We compare the different methods we presented with two different types of baseline encoders: pre-trained encoders, and hierarchical encoders based on recurrent cells. The latter, achieve current SOTA performance in many sequence labelling tasks [COLOMBO and collab., 2020; LI and collab., 2018a; LIN and collab., 2017].
Pre-trained Encoder Models. We use BERT [DEVLIN and collab., 2018] through the pytorch implementation provided by the Hugging Face transformers library [WOLF and collab., 2019]. The pre-trained model is fed with a concatenation of the utterances. Formally given an input context $C_k = (u_1, \dots u_T)$ the concatenation $[u_1, \dots, u_T]$ is fed to BERT.

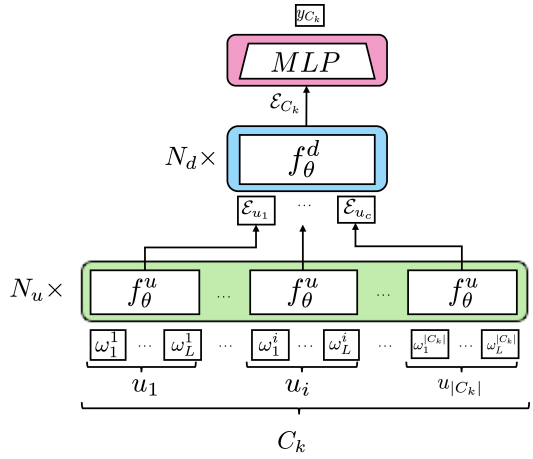
[2] Although it is possible to relax the fixed size imposed by transformers [DAI and collab., 2019] in this work we follow COLOMBO and collab. [2020] and fix the context size to 5 and the max utterance length to 50 — these choices are made to work with OpenSubtitles, since the number of available dialogues drops when considering a number of utterances greater than 5.

[3] if $x = u$ solely utterance level training is used, if $x = d$ solely dialogue level is used and if $x = u, d$ multi level supervision is used ($\lambda_u, \lambda_d \in \{0,1\}^2$ according to the case.)

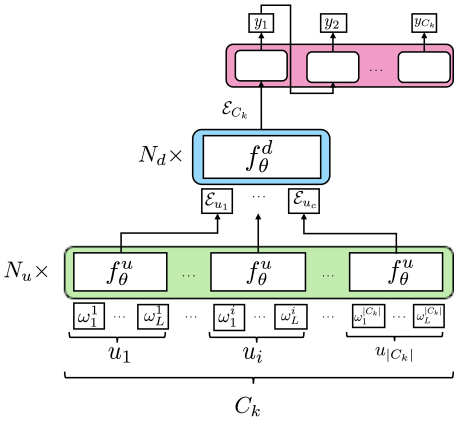
[4] http://opus.nlpl.eu/OpenSubtitles-alt-v2018.php

[5] We choose $\delta_T = 6s$

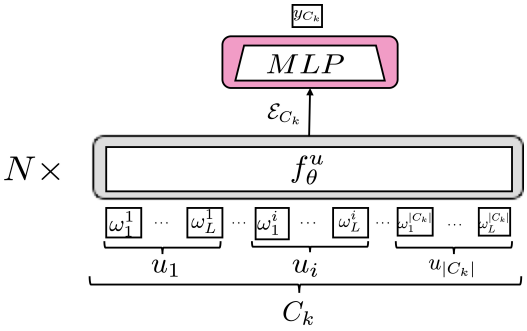
[6] Using pre-training method based on the next utterance proposed by [MEHRI and collab., 2019] requires dropping conversation shorter than T + 1 leading to a non-negligible loss in the preprocessing stage.



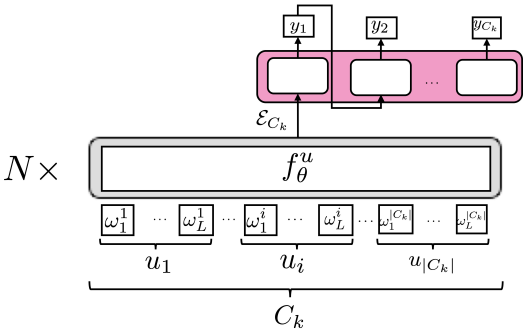
(a) Hierarchical encoder with MLP decoder performing single label prediction.



(b) Hierarchical encoder with sequential decoder (either GRU or CRF).



(c) BERT encoder with MLP decoder performing single label prediction.



(d) BERT encoder with sequential decoder (either GRU or CRF)

Figure 5.3 – Schema of the different models evaluated on SILICONE. In this figure f_θ^u, f_θ^d and the sequence label decoder (g_θ^{dec}) are respectively colored in green, blue and red for the hierarchical encoder (see Figure 5.3a and Figure 5.3d). For BERT there is no hierarchy and embedding is performed through f_θ^u colored in grey (see Figure 5.3c, Figure 5.3d)

Hierarchical Recurrent Encoders. In this work, we rely on our own implementation of the model based on $\mathscr{HR}$.
A representation for all the baselines can be found in Figure 5.3. For all models, both hidden dimension and embedding dimension is set to 768 to ensure fair comparison with the proposed model. The MLP used for decoding contains 3 layers of sizes $(768, 348, 192)$. We use RELU [AGARAP, 2018] to introduce non linearity inside our architecture.

5.3 Evaluation of Sequence Labelling

5.3.1 Related Work

As stated previously, sequence labelling tasks for spoken dialog mainly involve two different types of labels: DA and E/S. Early work has tackled the sequence labelling problem as an independent classification of each utterance. Deep neural network models that currently achieve the best results [KEIZER and collab., 2002; STOLCKE and collab., 2000; SURENDRAN and LEVOW, 2006] model both contextual dependencies between utterances [COLOMBO and collab., 2020; LI and collab., 2018b] and

labels [CHEN and collab., 2018b; KUMAR and collab., 2018; LI and collab., 2018c].

The aforementioned methods require large corpora to train models from scratch, such as: Switchboard Dialog Act (SwDA) [GODFREY and collab., 1992], Meeting Recorder Dialog Act (MRDA) [SHRIBERG and collab., 2004], Daily Dialog Act [LI and collab., 2017], HCRC Map Task Corpus (MT) [THOMPSON and collab., 1993]. This makes harder their adoption to smaller datasets, such as: Loqui human-human dialogue corpus (Loqui) [PASSONNEAU and SACHAR., 2014], BT Oasis Corpus (Oasis) [LEECH and WEISSER, 2003], Multimodal Multi-Party Dataset (MELD) [PORIA and collab., 2018a], Interactive emotional dyadic motion capture database (IEMO), SEMAINE database (SEM) [MCKEOWN and collab., 2013].

5.3.2 Presentation of SILICONE

Despite the similarity between methods usually employed to tackle DA and E/S sequential classification, studies usually rely on a single type of label. Moreover, despite the variety of small or medium-sized labelled datasets, evaluation is usually done on the largest available corpora (e.g., SwDA, MRDA). We introduce SILICONE, a collection of sequence labelling tasks, gathering both DA and E/S annotated datasets. SILICONE is built upon preexisting datasets which have been considered by the community as challenging and interesting. Any model that is able to process multiple sequences as inputs and predict the corresponding labels can be evaluated on SILICONE. We especially include small-sized datasets, as we believe it will ensure that well-performing models are able to both distil substantial knowledge and adapt to different sets of labels without relying on a large number of examples. The description of the datasets composing the benchmark can be found in the following sections, while corpora statistics are gathered in Table 5.1.

DA Datasets

In addition to SwDA and MRDA presented in subsection 4.5.1 we have collected 3 datasets annotated in DA:

DailyDialog Act Corpus (DyDA_a) has been produced by [LI and collab., 2017]. It contains multi-turn dialogues, supposed to reflect daily communication by covering topics about daily life. The dataset is manually labelled with dialog act and emotions. It is the third biggest corpus of SILICONE with $102k$ utterances. The SOTA model reports an accuracy of 88.1% [LI and collab., 2018a], using Bi-LSTMs with attention as well as additional features. We follow the official split introduced by the authors.

HCRC MapTask Corpus (MT) has been introduced by THOMPSON and collab. [1993]. To build this corpus, participants were asked to collaborate verbally by describing a route from a first participant's map by using the map of another participant. This corpus is small ($36k$ utterances). As there is no standard train/dev/test split[7] performances depends on the split. [TRAN and collab., 2017] make use of a Hierarchical LSTM encoder with a GRU decoder layer and achieves an accuracy of 65.9%.

Bt Oasis Corpus (Oasis) contains the transcripts of live calls made to the BT and operator services. This corpus has been introduced by LEECH and WEISSER [2003] and is rather small ($15k$ utterances). There is no standard train/dev/test split [8] and few studies use this dataset.

[7] We split according to the code in https://github.com/NathanDuran/Maptask-Corpus.

[8] We use a random split from https://github.com/NathanDuran/BT-Oasis-Corpus.

S/E Datasets

In S/E recognition for spoken language, there is no consensus on the choice the evaluation metric (e.g., [GHOSAL and collab., 2019; PORIA and collab., 2018b] use a weighted F-score while [ZHANG and collab., 2019b] report accuracy). For SILICONE, we choose to stay consistent with the DA research and thus follow [ZHANG and collab., 2019b] by reporting the accuracy. Additionally, emotion/sentiment labels are neither merged nor prepossessed[9].

DailyDialog Emotion Corpus (DyDA_e) has been previously introduced and contains eleven emotional labels. The SOTA model [DE BRUYNE and collab., 2019] is based on BERT with additional Valence Arousal and Dominance features and reaches an accuracy of 85% on the official split.

Multimodal EmotionLines Dataset (MELD) has been created by enhancing and extending EmotionLines dataset [CHEN and collab., 2018a] where multiple speakers participated in the dialogues. There are two types of annotations MELD_s and MELD_e: three sentiments (positive, negative and neutral) and seven emotions (anger, disgust, fear, joy,neutral, sadness and surprise). The SOTA model with text only is proposed by [ZHANG and collab., 2019b] and is inspired by quantum physics. On the official split, it is compared with a hierarchical bi-LSTM, which it beats with an accuracy of 61.9% (MELD_s) and 67.9% (MELD_e) against 60.8% and 65.2.

IEMOCAP database (IEMO) is a multimodal database of ten speakers. It consists of dyadic sessions where actors perform improvisations or scripted scenarios. Emotion categories are: anger, happiness, sadness, neutral, excitement, frustration, fear, surprise, and other. There is no official split on this dataset. One proposed model is built with bi-LSTMs and achieves 35.1%, with text only [ZHANG and collab., 2019b].

SEMAINE database (SEM) comes from the Sustained Emotionally coloured Machine human Interaction using Nonverbal Expression project [MCKEOWN and collab., 2013]. This dataset has been annotated on three sentiments labels: positive, negative and neutral by [BARRIERE and collab., 2018]. It is built on Multimodal Wizard of Oz experiment where participants held conversations with an operator who adopted various roles designed to evoke emotional reactions. There is no official split on this dataset.

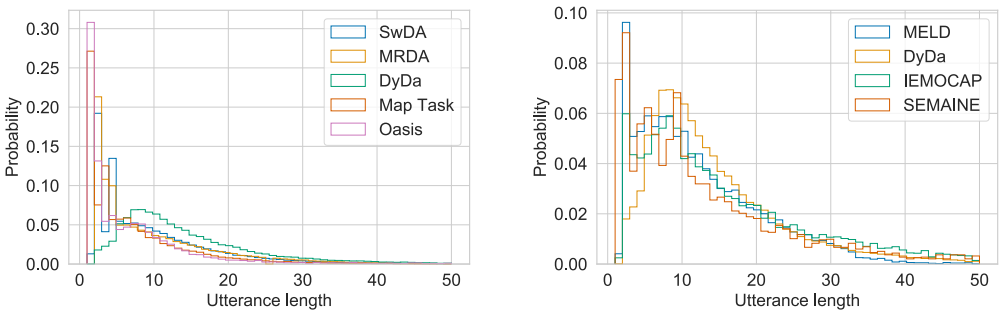
Diversity of SILICONE

We illustrate the diversity of the dataset composing SILICONE. In Figure 5.4, we plot two histograms representing the different utterance lengths for DA and E/S. As expected, for spoken dialog, lengths are shorter than for written benchmarks (e.g., GLUE).

[9] Comparison with concurrent work is more difficult as system performance heavily depends on the number of classes and label processing varies across studies [CLAVEL and CALLEJAS, 2015].

Corpus	$\|Train\|$	$\|Val\|$	$\|Test\|$	Utt.	$\|Labels\|$	Task	Utt./$\|Labels\|$
SwDA*	1k	100	11	200k	42	DA	4.8k
MRDA*	56	6	12	110k	5	DA	2.6k
DyDA$_a$	11k	1k	1k	102k	4	DA	25.5k
MT*	121	22	25	36k	12	DA	3k
Oasis*	508	64	64	15k	42	DA	357
DyDA$_e$	11k	1k	1k	102k	7	E	2.2k
MELD$^\star_s$	934	104	280	13k	3	S	4.3k
MELD$^\star_e$	934	104	280	13k	7	S	1.8k
IEMO	108	12	31	10k	6	E	1.7k
SEM	62	7	10	5,6k	3	S	1.9k

Table 5.1 – Statistics of datasets composing SILICONE. E stands for emotion label and S for sentiment label; * stands for datasets with available official split. Sizes of Train, Val and Test are given in number of conversations.



(a) SILICONE DA (b) SILICONE S/E

Figure 5.4 – Histograms showing the utterance length for each dataset of SILICONE.

5.4 Results on SILICONE

This section gathers experiments performed on the SILICONE benchmark. We first analyse an appropriate choice for the decoder, which is selected over a set of experiments on our baseline encoders: a pre-trained BERT model and a hierarchical RNN-based encoder ($\mathscr{HR}$). Since we focus on small-sized pre-trained representations, we limit the sizes of our pre-trained models to TINY and SMALL (see Table 5.7). We then study the results of the baselines and our hierarchical transformer encoders ($\mathscr{HT}$) on SILICONE along three axes: the accuracy of the models, the difference in performance between the E/S and the DA corpora, and the importance of pre-training. As we aim to obtain robust representations, we do not perform an exhaustive grid search on the downstream tasks.

5.4.1 Experimental Hyper-parameters for SILICONE

For all models, we use a batch size of 64 and automatically select the best model on the validation set according to its loss. We do not perform exhaustive grid search either on the learning rate (that is set to 10^{-4}), nor on other hyper-parameters to perform a fair comparison between all the models. We use ADAMW [KINGMA and BA, 2014; LOSHCHILOV and HUTTER, 2017] with a linear scheduler on the learning rate and the number of warm-up steps is set to 100. For all model we used a tokenizer based on WordPiece [WU and collab., 2016]. We used GELU [HENDRYCKS and GIMPEL, 2016] activations and the dropout rate [SRIVASTAVA and collab., 2014] is set to 0.1. We report in Table 6.5 the main hyper-parameters used for our model pre-training.

5.4.2 Decoder Choice

Current research efforts focus on single label prediction, as it seems to be a natural choice for sequence labelling problems (subsection 5.2.1). Sequence labelling is usually performed with CRFs [CHEN and collab., 2018b; KUMAR and collab., 2018] and GRU decoding [COLOMBO and collab., 2020], however, it is not clear to what extent inter-label dependencies are already captured by the contextualised encoders, and whether a plain MLP decoder could achieve competitive results. As can be seen

	TINY	SMALL
Nbs of heads	1	6
N_d	2	4
N_u	2	4
T	50	50
C	5	5
$\mathscr{T}_d$ nbs of heads	6	6
Inner dimension	768	768
Model Dimension	768	768
Vocab length	32000	32000
$\mathscr{T}_d$: Emb. size	768	768
d_k:	64	64
d_v:	64	64

Table 5.2 – Architecture hyperparameters used for the hierarchical pre-training.

	Avg	Avg DA	Avg E/S
BERT (+MLP)	72,8	81.5	64.0
BERT (+GRU)	69.9	80.4	59.3
BERT (+CRF)	72.8	81.5	64.1
$\mathcal{HR}$ (+MLP)	69.8	79.1	60.4
$\mathcal{HR}$ (+GRU)	67.6	79.4	55.7
$\mathcal{HR}$ (+CRF)	70.5	80.3	60.7

Table 5.3 – Experiments comparing decoder performances. Results are given on SILICONE for two types of baseline encoders (pre-trained BERT models and hierarchical recurrent encoders $\mathcal{HR}$).

	Avg	SwDA	MRDA	DyDA_{DA}	MT	Oasis	DyDA_{e}	MELD_{s}	MELD_{e}	IEMO	SEM
BERT-4layers	70.4	77.8	90.7	79.0	88.4	66.8	90.3	55.3	53.4	43.0	58.8
BERT	72.8	79.2	90.7	**82.6**	88.2	66.9	91.9	59.3	**61.4**	**45.0**	62.7
$\mathcal{HR}$	69.8	77,5	90,9	80,1	82,8	64,3	91.5	59,3	59.9	40.3	51.1
$\mathcal{HT}$ $(\theta^{u,d}_{\text{MLM}})_{\text{(TINY)}}$	73.3	**79.3**	92.0	80.1	90.0	68,3	92.5	62.6	59.9	42.0	66.6
$\mathcal{HT}$ $(\theta^{d}_{\text{GAP}})_{\text{(TINY)}}$	71.6	78.6	91.8	78.1	89.3	64.1	91.6	60.5	55.7	42.2	63.9
$\mathcal{HT}$ $(\theta^{u,d}_{\text{MLM}})_{\text{(SMALL)}}$	**74.3**	79.2	**92.4**	81.5	**90.6**	**69.4**	**92.7**	**64.1**	60.1	**45.0**	**68.2**

Table 5.4 – Performances of different encoders when decoding using a MLP on SILICONE. The datasets are grouped by label type (DA vs E/S) and ordered by decreasing size. MT stands for Map Task, IEM for IEMOCAP and Sem for Semaine.

in Table 5.3, we found that in the case of E/S prediction there is no clear difference between CRFs and MLPs, while GRU decoders exhibit poor performance, probably due to a lack of training data. It is also important to notice, that training a sequential decoder usually requires thorough hyper-parameter fine-tuning. As our goal is to learn and evaluate general representations that are decoder agnostic, in the following, we will use a plain MLP decoder for all the models compared.

5.4.3 General Performance Analysis

Table 5.4 provides an exhaustive comparison of the different encoders over the SILICONE benchmark. As previously discussed, we adopt a plain MLP as a decoder to compare the different encoders. We show that SILICONE covers a set of challenging tasks as the best performing model achieves an average accuracy of 74.3. Moreover, we observe that despite having half the parameters of a BERT model, our proposed model achieves an average result that is 2% higher on the benchmark. SILICONE covers two different sequence labelling tasks: DA and E/S. In Table 5.4 and Table 5.3, we can see that all models exhibit a consistently higher average accuracy (up to 14%) on DA tagging compared to E/S prediction. This performance drop could be explained by the different sizes of the corpora (see Table 5.1). Despite having a larger number of utterances per label (u/l), E/S tasks seem generally harder to tackle for the models. For example, on Oasis, where the u/l is inferior than those of most E/S datasets (MELD_{s}, MELD_{e}, IEMO and SEM), models consistently achieve better results.

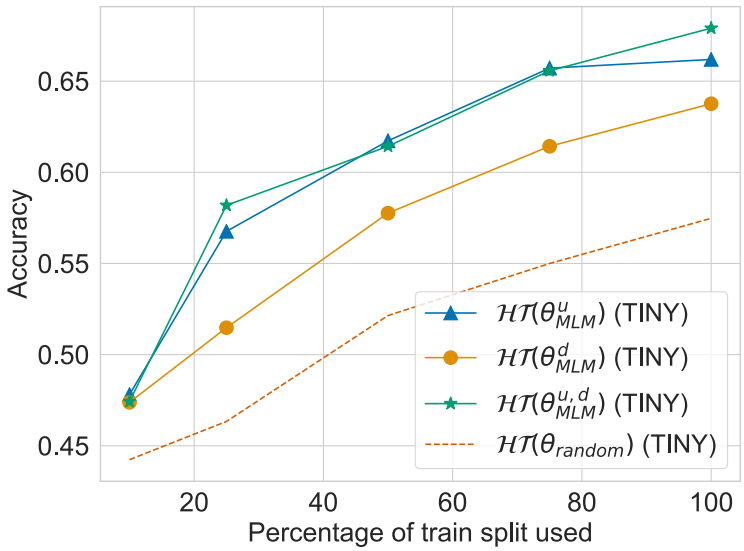


Figure 5.5 – A comparison of pre-trained encoders being fine-tuned on different percentage the training set of SEM. Validation and test set are fixed over all experiments, reported scores are averaged over 10 different random split.

5.4.4 Importance of Pre-training for SILICONE

Results reported in Table 5.4 and Table 5.3 show that pre-trained transformer-based encoders achieve consistently higher accuracy on SILICONE, even when they are not explicitly considering the hierarchical structure. This difference can be observed both in small-sized datasets (e.g. MELD and SEM) and in medium/large size datasets (e.g SwDA and MRDA). To validate the importance of pre-training in a regime of low data, we train different $\mathcal{HT}$ (with random initialisation) on different portions of SEM and MELD_s. Results shown in Figure 5.5 illustrate the importance of pre-trained representations.

Negative Results on GAP

We briefly describe few ideas we tried to make GAP works at both the utterance and dialogue level. We hypothesise that:

- giving the same weight to the utterance level and the dialogue level (see Equation 6.2) was responsible of the observed plateau. Different combinations lead to fairly poor improvements.

- the limited model capacity was part of the issue. Larger models does not give the expected results.

5.5 Model Analysis

In this section, we dissect our hierarchical pre-trained models in order to better understand the relative importance of each component. We show how a hierarchical encoder allows us to obtain a light and efficient model.

	Avg DA	Avg E/S
BERT (4 layers)	80.5	60.2
$\mathcal{HT}(\theta_{\text{BERT}-2layers})$	80.5	61.1
$\mathcal{HT}(\theta^{u}_{\text{MLM}})$	**80.8**	**64.0**

Table 5.5 – Results of ablation studies on SILICONE

5.5.1 Pre-training on Spoken vs Written Data

First, we explore the differences in training representations on spoken and written corpora. Experimentally, we compare the predictions on SILICONE made by $\mathcal{HT}(\theta^{u}_{\text{MLM}})$ and the one made by $\mathcal{HT}(\theta_{\text{BERT}-2layers})$. The latter is a hierarchical encoder where utterance embeddings are obtained with the hidden vector representing the first token [CLS] (see [DEVLIN and collab., 2018]) of the second layer of BERT. In both cases, predictions are performed using an MLP[10]. Results in Table 5.5 show higher accuracy when the pre-training is performed on spoken data. Since SILICONE is a spoken language benchmark, this result might be due to the specific features of colloquial speech (e.g. disfluencies, sentence length, vocabulary, word frequencies).

5.5.2 Hierarchy and Multi-Level Supervision

We study the relative importance of three aspects of our hierarchical pre-training with multi-level supervision. We first show that accounting for the hierarchy increases the performance of fine-tuned encoders, even without our specific pre-training procedure. We then compare our two proposed hierarchical pre-training procedures based on the GAP or MLM loss. Lastly, we look at the contribution of the possible levels of supervision on reduced training data from SEM.

Importance of hierarchical fine-tuning

We compare the performance of BERT-4layers with the $\mathcal{HT}(\theta_{\text{BERT}-2layer})$ previously described. Results reported in Table 5.5 demonstrate that fine-tuning on downstream tasks with a hierarchical encoder yields to higher accuracy, with fewer parameters, even when using already pre-trained representations.

MLM vs GAP

In this experiment, we compare the different pre-training objectives at utterance and dialogue level. As a reminder $\mathcal{HT}(\theta^{u}_{\text{MLM}})$ and $\mathcal{HT}(\theta^{u}_{\text{GAP}})$ are respectively trained using the standard MLM loss [DEVLIN and collab., 2018] and the standard GAP loss [YANG and collab., 2019]. In Table 5.6 we report the different pre-training objective results. We observe that pre-training at the dialogue level achieves comparable results to the utterance level pre-training for MLM and slightly worse for GAP. Interestingly, we observe that $\mathcal{HT}(\theta^{u}_{\text{GAP}})$ compared to $\mathcal{HT}(\theta^{u}_{\text{MLM}})$ achieves worse results, which is not consistent with the performance observed on other benchmarks, such as GLUE [WANG and collab., 2018]. The lower accuracy of the models trained using a GAP-based loss could be due to several factors (e.g., model size, pre-training using

[10] We consider the two first layer for a fair comparison based on the number of model parameters.

	Avg DA	Avg E/S
$\mathcal{HT}(\theta^{u}_{\text{MLM}})$	80.8	64.0
$\mathcal{HT}(\theta^{d}_{\text{MLM}})$	80.8	64.0
$\mathcal{HT}(\theta^{u}_{\text{GAP}})$	80.7	62.0
$\mathcal{HT}(\theta^{d}_{\text{GAP}})$	80.4	62.8
$\mathcal{HT}(\theta^{u,d}_{\text{MLM}})$	**81.9**	**64.7**

Table 5.6 – Comparison of GAP and MLM with a comparable number of parameters. For all models a MLP decoder is used on top of a TINY pre-trained encoder.

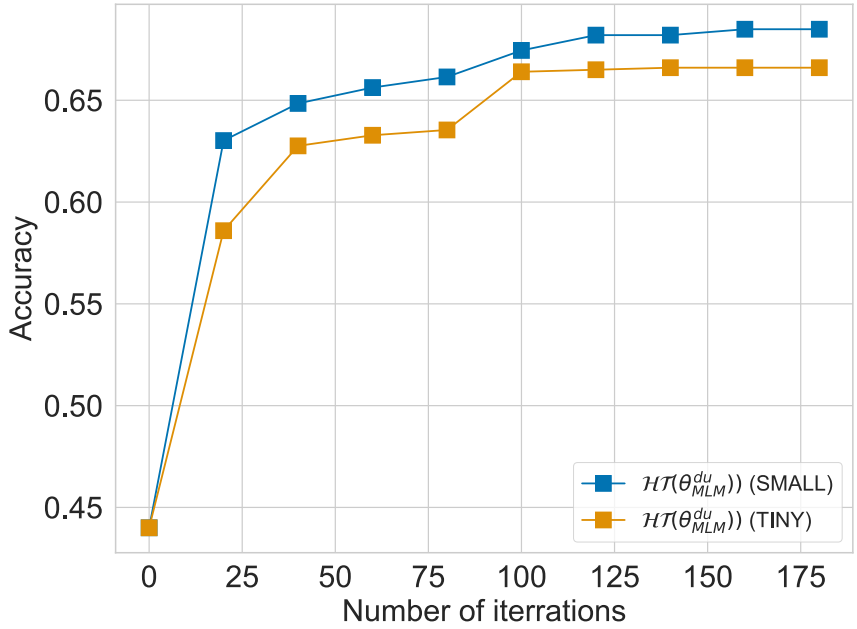


Figure 5.6 – Illustration of improvement of accuracy during pre-training stage on SEM for both a TINY and SMALL model.

the GAP loss could require a finer choice of hyper-parameters). Finally, we see that supervising at both dialogue and utterance level helps for MLM[11].

Multi level Supervision for pre-training

In this section, we illustrate the advantages of learning using several levels of supervision on small datasets. We fine-tune different model on SEM using different size of the training set. Results are shown in Figure 5.5. Overall we see that introducing sequence level supervision induces a consistent improvement on SEM.

5.5.3 Improvement over pre-training

In this experiment we illustrate how pre-training improves performance on SEM (see Figure 5.6). As expected accuracy improves when pre-training.

[11] We investigate a similar setting for GAP which lead to poor results, the loss hit a plateau suggesting that objectives are competing against each other. More advanced optimisations techniques [SENER and KOLTUN, 2018] are left for future work.

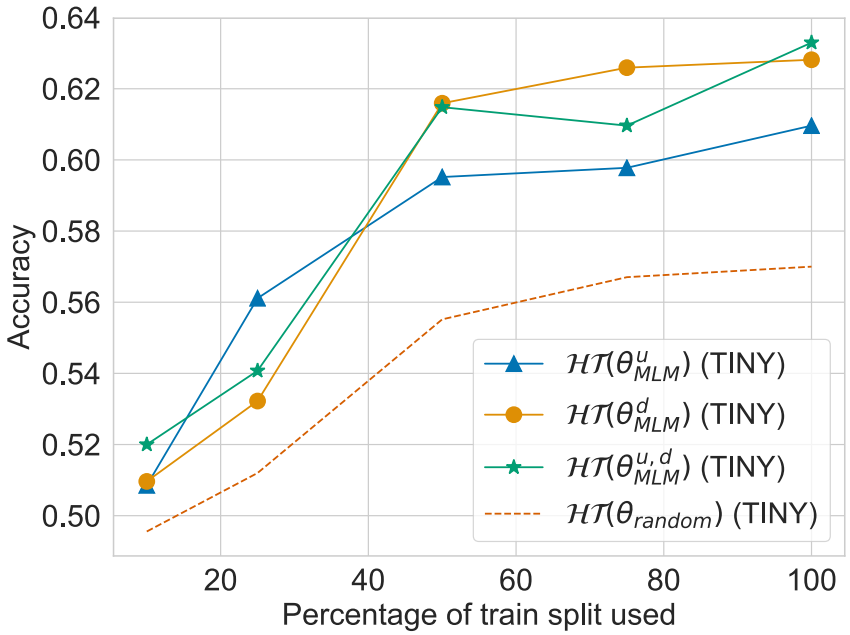


Figure 5.7 – A comparison of different parameters initialisation on $\texttt{MELD}_s$. Training is performed using a different percentage of complete training set. Validation and test set are fixed over all experimentation. Each score is the averaged accuracy over 10 random runs.

	Emb.	Word	Seq	Total
BERT		87		110
BERT (4-layer)		43		66
HMLP	23	8.6	7.8	40
(TINY)		2.9	2.8	28.7
(SMALL)		10.6	10.6	45

Table 5.7 – Number of parameters for the encoders. Sizes are given in million of parameters.

5.5.4 Multi level Supervision for pre-training MELD

In this experiment we report results of the experiment mentioned in section 5.5.2. In this experiment we see that the training process seems to be noisier for fractions lower than 40%. For larger percentages, we observe that including higher supervision (at the dialogue level) during pre-training leads to a consistent improvement.

5.5.5 Other advantages of hierarchy

Introducing a hierarchical design in the encoder allows to break dialogue into utterances and to consider inputs of size T instead of size 512. First, it allows parameters sharing, reducing the number of model parameters. The different model sizes are reported in Table 5.7. Our TINY model contains half the parameters of BERT (4-layers). Furthermore, modelling long-range dependencies hierarchically makes learning faster and allows to get rid of learning tricks (e.g., partial order prediction [YANG and collab., 2019], two-stage pre-training based on sequence length [DEVLIN and collab., 2018]) required for non-hierarchical encoders. Lastly, original BERT and XLNET are pre-trained using respectively 16 and 512 TPUs. Pre-training lasts several days

with over 500K iterations. Our TINY hierarchical models are pre-trained during 180K iterations (1.5 days) on 4 NVIDIA V100.

Chapter 5 conclusion

In this chapter, we propose a hierarchical transformer-based encoder tailored for spoken dialog. We extend two well-known pre-training objectives to adapt them to a hierarchical setting and use OpenSubtitles, the largest spoken language dataset available, for encoder pre-training. Additionally, we provide an evaluation benchmark dedicated to comparing sequence labelling systems for the NLP community, SILICONE, on which we compare our models and pre-training procedures with previous approaches. By conducting ablation studies, we demonstrate the importance of using a hierarchical structure for the encoder, both for pre-training and fine-tuning. Finally, we find that our approach is a powerful method to learn generic representations on spoken dialog, with less parameters than state-of-the-art transformer models. We hope that the SILICONE benchmark, will encourage further research to build stronger sequence labelling systems for NLP. This work could be then extended to a multilingual setting as we will show in Chapter 6.

5.6 References

AGARAP, A. F. 2018, "Deep learning using rectified linear units (relu)", *arXiv preprint arXiv:1803.08375*. 87

ARGYRIOU, A., T. EVGENIOU and M. PONTIL. 2007, "Multi-task feature learning", in *Advances in neural information processing systems*, p. 41–48. 84

BARRIERE, V., C. CLAVEL and S. ESSID. 2018, "Attitude classification in adjacency pairs of a human-agent interaction with hidden conditional random fields", in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, p. 4949–4953. 89

BUDZIANOWSKI, P., T.-H. WEN, B.-H. TSENG, I. CASANUEVA, U. STEFAN, R. OSMAN and M. GAŠIĆ. 2018a, "Multiwoz - a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling", in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 86

BUDZIANOWSKI, P., T.-H. WEN, B.-H. TSENG, I. CASANUEVA, S. ULTES, O. RAMADAN and M. GAŠIĆ. 2018b, "Multiwoz-a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling", *arXiv preprint arXiv:1810.00278*. 82

CHEN, S.-Y., C.-C. HSU, C.-C. KUO, L.-W. KU and collab.. 2018a, "Emotionlines: An emotion corpus of multi-party conversations", *arXiv preprint arXiv:1802.08379*. 89

CHEN, Z., R. YANG, Z. ZHAO, D. CAI and X. HE. 2018b, "Dialogue act recognition via crf-attentive structured network", in *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, p. 225–234. 83, 88, 91

CLAVEL, C. and Z. CALLEJAS. 2015, "Sentiment analysis: from opinion mining to human-agent interaction", *IEEE Transactions on affective computing*, vol. 7, nº 1, p. 74–93. 89

COLOMBO, P., E. CHAPUIS, M. MANICA, E. VIGNON, G. VARNI and C. CLAVEL. 2020, "Guiding attention in sequence-to-sequence models for dialogue act prediction", *arXiv preprint arXiv:2002.08801*. 86, 87, 91

DAI, Z., Z. YANG, Y. YANG, J. CARBONELL, Q. V. LE and R. SALAKHUTDINOV. 2019, "Transformer-xl: Attentive language models beyond a fixed-length context", *arXiv preprint arXiv:1901.02860*. 86

DANESCU-NICULESCU-MIZIL, C. and L. LEE. 2011, "Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs.", in *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics, ACL 2011*. 82, 86

DE BRUYNE, L., P. ATANASOVA and I. AUGENSTEIN. 2019, "Joint emotion label space modelling for affect lexica", *arXiv preprint arXiv:1911.08782*. 89

DENOYER, L. and P. GALLINARI. 2006, "The wikipedia xml corpus", in *International Workshop of the Initiative for the Evaluation of XML Retrieval*, Springer, p. 12–19. 82

DEVLIN, J., M.-W. CHANG, K. LEE and K. TOUTANOVA. 2018, "Bert: Pre-training of deep bidirectional transformers for language understanding", *arXiv preprint arXiv:1810.04805*. 82, 83, 86, 94, 96

DINKAR, T., P. COLOMBO, M. LABEAU and C. CLAVEL. 2020, "The importance of fillers for text representations of speech transcripts", *arXiv preprint arXiv:2009.11340*. 82

E. CHAPUIS & P. COLOMBO, M. MANICA, M. LABEAU and C. CLAVEL. 2020, "Hierarchical pre-training for sequence labelling in spoken dialog", *CoRR*, vol. abs/2009.11152. URL https://arxiv.org/abs/2009.11152. 81

GARCIA, A., P. COLOMBO, S. ESSID, F. D'ALCHÉ BUC and C. CLAVEL. 2019, "From the token to the review: A hierarchical multimodal approach to opinion mining", *arXiv preprint arXiv:1908.11216*. 84

GHOSAL, D., N. MAJUMDER, S. PORIA, N. CHHAYA and A. GELBUKH. 2019, "Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation", *arXiv preprint arXiv:1908.11540*. 89

GODFREY, J. J., E. C. HOLLIMAN and J. MCDANIEL. 1992, "Switchboard: Telephone speech corpus for research and development", in *Proceedings of the 1992 IEEE International Conference on Acoustics, Speech and Signal Processing - Volume 1*, ICASSP'92, IEEE Computer Society, USA, ISBN 0780305329, p. 517–520. 82, 88

HAZARIKA, D., S. PORIA, R. ZIMMERMANN and R. MIHALCEA. 2019, "Emotion recognition in conversations with transfer learning from generative conversation modeling", *arXiv preprint arXiv:1910.04980*. 82, 86

HENDERSON, P., J. HU, J. ROMOFF, E. BRUNSKILL, D. JURAFSKY and J. PINEAU. 2020, "Towards the systematic reporting of the energy and carbon footprints of machine learning", *arXiv preprint arXiv:2002.05651*. 82

HENDRYCKS, D. and K. GIMPEL. 2016, "Gaussian error linear units (gelus)", *arXiv preprint arXiv:1606.08415*. 91

JIAO, X., Y. YIN, L. SHANG, X. JIANG, X. CHEN, L. LI, F. WANG and Q. LIU. 2019, "Tinybert: Distilling bert for natural language understanding", *arXiv preprint arXiv:1909.10351*. 82

KEIZER, S., R. OP DEN AKKER and A. NIJHOLT. 2002, "Dialogue act recognition with bayesian networks for dutch dialogues", in *Proceedings of the Third SIGdial Workshop on Discourse and Dialogue*. 87

KINGMA, D. P. and J. BA. 2014, "Adam: A method for stochastic optimization", *arXiv preprint arXiv:1412.6980*. 91

KUMAR, H., A. AGARWAL, R. DASGUPTA and S. JOSHI. 2018, "Dialogue act sequence labeling using hierarchical encoder with crf", in *Thirty-Second AAAI Conference on Artificial Intelligence*. 88, 91

LAN, Z., M. CHEN, S. GOODMAN, K. GIMPEL, P. SHARMA and R. SORICUT. 2019, "Albert: A lite bert for self-supervised learning of language representations", *arXiv preprint arXiv:1909.11942*. 82, 83

LE, H., L. VIAL, J. FREJ, V. SEGONNE, M. COAVOUX, B. LECOUTEUX, A. ALLAUZEN, B. CRABBÉ, L. BESACIER and D. SCHWAB. 2019, "Flaubert: Unsupervised language model pre-training for french", *arXiv preprint arXiv:1912.05372*. 82

LEECH, G. and M. WEISSER. 2003, "Generic speech act annotation for task-oriented dialogues.", . 88

LI, R., C. LIN, M. COLLINSON, X. LI and G. CHEN. 2018a, "A dual-attention hierarchical recurrent neural network for dialogue act classification", *CoRR*, vol. abs/1810.09154. URL http://arxiv.org/abs/1810.09154. 83, 86, 88

LI, R., C. LIN, M. COLLINSON, X. LI and G. CHEN. 2018b, "A dual-attention hierarchical recurrent neural network for dialogue act classification", *CoRR*. 87

LI, R., C. LIN, M. COLLINSON, X. LI and G. CHEN. 2018c, "A dual-attention hierarchical recurrent neural network for dialogue act classification", *arXiv preprint arXiv:1810.09154*. 88

LI, Y., H. SU, X. SHEN, W. LI, Z. CAO and S. NIU. 2017, "Dailydialog: A manually labelled multi-turn dialogue dataset", . 88

LIN, Z., M. FENG, C. N. D. SANTOS, M. YU, B. XIANG, B. ZHOU and Y. BENGIO. 2017, "A structured self-attentive sentence embedding", *arXiv preprint arXiv:1703.03130*. 86

LISON, P. and J. TIEDEMANN. 2016, "Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles", . 86

LISON, P., J. TIEDEMANN, M. KOUYLEKOV and collab.. 2019, "Open subtitles 2018: Statistical rescoring of sentence alignments in large, noisy parallel corpora", in *LREC 2018, Eleventh International Conference on Language Resources and Evaluation*, European Language Resources Association (ELRA). 82

LIU, Y., M. OTT, N. GOYAL, J. DU, M. JOSHI, D. CHEN, O. LEVY, M. LEWIS, L. ZETTLEMOYER and V. STOYANOV. 2019, "Roberta: A robustly optimized bert pretraining approach", *arXiv preprint arXiv:1907.11692*. 82, 83

LOSHCHILOV, I. and F. HUTTER. 2017, "Decoupled weight decay regularization", *arXiv preprint arXiv:1711.05101*. 91

LOWE, R., N. POW, I. SERBAN and J. PINEAU. 2015, "The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems", *CoRR*, vol. abs/1506.08909. URL http://arxiv.org/abs/1506.08909. 82, 86

MCKEOWN, G., M. VALSTAR, R. COWIE, M. PANTIC and M. SCHRODER. 2013, "The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent", *Affective Computing, IEEE Transactions on*, vol. 3, doi: 10.1109/T-AFFC.2011.20, p. 5–17. 88, 89

MEHRI, S., E. RAZUMOVSAKAIA, T. ZHAO and M. ESKENAZI. 2019, "Pretraining methods for dialog context representation learning", *arXiv preprint arXiv:1906.00414*. 82, 86

MIKOLOV, T., I. SUTSKEVER, K. CHEN, G. S. CORRADO and J. DEAN. 2013, "Distributed representations of words and phrases and their compositionality", in *Advances in neural information processing systems*, p. 3111–3119. 82

PASSONNEAU, R. and E. SACHAR. 2014, "Loqui human-human dialogue corpus (transcriptions and annotations)", . 88

PENNINGTON, J., R. SOCHER and C. D. MANNING. 2014, "Glove: Global vectors for word representation", in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, p. 1532–1543. 82

PETERS, M. E., M. NEUMANN, M. IYYER, M. GARDNER, C. CLARK, K. LEE and L. ZETTLEMOYER. 2018, "Deep contextualized word representations", *arXiv preprint arXiv:1802.05365*. 82

PORIA, S., D. HAZARIKA, N. MAJUMDER, G. NAIK, E. CAMBRIA and R. MIHALCEA. 2018a, "Meld: A multimodal multi-party dataset for emotion recognition in conversations", . 88

PORIA, S., D. HAZARIKA, N. MAJUMDER, G. NAIK, E. CAMBRIA and R. MIHALCEA. 2018b, "Meld: A multimodal multi-party dataset for emotion recognition in conversations", *arXiv preprint arXiv:1810.02508*. 89

RUDER, S. 2017, "An overview of multi-task learning in deep neural networks", *arXiv preprint arXiv:1706.05098*. 84

SANH, V., T. WOLF and S. RUDER. 2019, "A hierarchical multi-task approach for learning embeddings from semantic tasks", in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, p. 6949–6956. 84

SENER, O. and V. KOLTUN. 2018, "Multi-task learning as multi-objective optimization", in *Advances in Neural Information Processing Systems*, p. 527–538. 95

SHRIBERG, E., R. DHILLON, S. BHAGAT, J. ANG and H. CARVEY. 2004, "The ICSI meeting recorder dialog act (MRDA) corpus", in *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue at HLT-NAACL 2004*, Association for Computational Linguistics, Cambridge, Massachusetts, USA, p. 97–100. URL https://www.aclweb.org/anthology/W04-2319. 82, 88

SHRIBERG, E. E. 1999, "Phonetic consequences of speech disfluency", cahier de recherche, SRI INTERNATIONAL MENLO PARK CA. 82

SRIVASTAVA, N., G. HINTON, A. KRIZHEVSKY, I. SUTSKEVER and R. SALAKHUTDINOV. 2014, "Dropout: a simple way to prevent neural networks from overfitting", *The journal of machine learning research*, vol. 15, nº 1, p. 1929–1958. 91

STOLCKE, A., K. RIES, N. COCCARO, E. SHRIBERG, R. BATES, D. JURAFSKY, P. TAYLOR, R. MARTIN, C. V. ESS-DYKEMA and M. METEER. 2000, "Dialogue act modeling for automatic tagging and recognition of conversational speech", *Computational linguistics*, vol. 26, nº 3, p. 339–373. 87

STOLCKE, A. and E. SHRIBERG. 1996, "Statistical language modeling for speech disfluencies", in *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, vol. 1, IEEE, p. 405–408. 82

SUÁREZ, P. J. O., B. SAGOT and L. ROMARY. 2019, "Asynchronous pipeline for processing huge corpora on medium to low resource infrastructures", *Challenges in the Management of Large Corpora (CMLC-7) 2019*, p. 9. 82

SURENDRAN, D. and G.-A. LEVOW. 2006, "Dialog act tagging with support vector machines and hidden markov models", in *Ninth International Conference on Spoken Language Processing*. 87

THOMPSON, H., A. ANDERSON, E. BARD, G. DOHERTY-SNEDDON, A. NEWLANDS and C. SOTILLO. 1993, "The hcrc map task corpus: natural dialogue for speech recognition", doi: 10.3115/1075671.1075677. 88

THORNBURY, S. and D. SLADE. 2006, *Conversation: From description to pedagogy*, Cambridge University Press. 82

TRAN, Q. H., G. HAFFARI and I. ZUKERMAN. 2017, "A generative attentional neural network model for dialogue act classification", in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, p. 524–529. 88

TRAN, T., J. YUAN, Y. LIU and M. OSTENDORF. 2019, "On the role of style in parsing speech with neural models", *Proc. Interspeech 2019*, p. 4190–4194. 82

VASWANI, A., N. SHAZEER, N. PARMAR, J. USZKOREIT, L. JONES, A. N. GOMEZ, Ł. KAISER and I. POLOSUKHIN. 2017, "Attention is all you need", in *Advances in neural information processing systems*, p. 5998–6008. 85

WANG, A., A. SINGH, J. MICHAEL, F. HILL, O. LEVY and S. R. BOWMAN. 2018, "Glue: A multi-task benchmark and analysis platform for natural language understanding", *arXiv preprint arXiv:1804.07461*. 82, 94

WOLF, T., L. DEBUT, V. SANH, J. CHAUMOND, C. DELANGUE, A. MOI, P. CISTAC, T. RAULT, R. LOUF, M. FUNTOWICZ and J. BREW. 2019, "Huggingface's transformers: State-of-the-art natural language processing", *ArXiv*, vol. abs/1910.03771. 86

WOLF, T., Q. LHOEST, P. VON PLATEN, Y. JERNITE, M. DRAME, J. PLU, J. CHAUMOND, C. DELANGUE, C. MA, A. THAKUR, S. PATIL, J. DAVISON, T. L. SCAO, V. SANH, C. XU, N. PATRY, A. MCMILLAN-MAJOR, S. BRANDEIS, S. GUGGER, F. LAGUNAS, L. DEBUT, M. FUNTOWICZ, A. MOI, S. RUSH, P. SCHMIDD, P. CISTAC, V. MUŠTAR, J. BOUDIER and A. TORDJMANN. 2020, "Datasets", *GitHub. Note: https://github.com/huggingface/datasets*, vol. 1. 81

WU, Y., M. SCHUSTER, Z. CHEN, Q. V. LE, M. NOROUZI, W. MACHEREY, M. KRIKUN, Y. CAO, Q. GAO, K. MACHEREY and collab.. 2016, "Google's neural machine translation system: Bridging the gap between human and machine translation", *arXiv preprint arXiv:1609.08144*. 91

YANG, Z., Z. DAI, Y. YANG, J. CARBONELL, R. R. SALAKHUTDINOV and Q. V. LE. 2019, "Xlnet: Generalized autoregressive pretraining for language understanding", in *Advances in neural information processing systems*, p. 5754–5764. 82, 83, 94, 96

ZHANG, X., F. WEI and M. ZHOU. 2019a, "Hibert: Document level pre-training of hierarchical bidirectional transformers for document summarization", *arXiv preprint arXiv:1905.06566*. 83

ZHANG, Y., Q. LI, D. SONG, P. ZHANG and P. WANG. 2019b, "Quantum-inspired interactive networks for conversational sentiment analysis.", . 89

ZHU, Y., R. KIROS, R. ZEMEL, R. SALAKHUTDINOV, R. URTASUN, A. TORRALBA and S. FIDLER. 2015, "Aligning books and movies: Towards story-like visual explanations by watching movies and reading books", in *Proceedings of the IEEE international conference on computer vision*, p. 19–27. 82

Part II Conclusions

In this part, we addressed the RQ1 and presented models that can automatically label English spoken dialogue utterances. First in Chapter 4, when large annotated corpora are available, we trained directly a seq2seq architecture composed of a hierarchical encoder and relied on specifically designed guided attention mechanisms. We showcased that our method allows to reach competitive results on DA classification. However, such an approach is difficult to apply when there is a lack of annotated data as it is the case of E/S corpora for example. In Chapter 5, we address the data scarcity problematic by relying on pre-trained architectures. We devise new pre-training approaches to learn generic representations tailored for spoken dialogue. We evaluate our representations on a new benchmark SILICONE that gathers DA and E/S corpora of different sizes.
In the next part, we address RQ2 and extend the work developed in Part II with two different settings : multilingual in Chapter 6 and multimodal in Chapter 7.

Part III

Towards Multilingual and Multi-modal Spoken Dialogue Understanding

Part III Abstract

This part gathers our contributions related to RQ2 "How to modify spoken dialogue understanding systems to handle multilingual and multimodal data? What are the specifics of multilingual and multimodal data that the system needs to handle?". This part is further divided into two different chapters:

- Chapter 6 is dedicated to the multilingual scenario where several languages are present within the conversation. We propose to adapt the neural architecture presented in Chapter 5 to handle multilingual data. We introduce new pre-training losses tailored for multilingual spoken dialogue and for the code-switching phenomena occurring within a conversation. We leverage the Opensubtitles corpus to automatically build multilingual conversations corpus. In addition, we evaluate our representations on a new benchmark called MIAM and composed of datasets in five different languages (French, Italian, English, German and Spanish) annotated in DA and two novel multilingual downstream tasks (*i.e* multilingual mask utterance retrieval and multilingual inconsistency identification).

- Chapter 7 is dedicated to the multimodal scenario. In this chapter, we focus on the multimodal aspect of conversations in the scope of emotion recognition. Textual modality is particularly well exploited by current neural architectures. Therefore, we seek to study ways to translate non-textual modality into textual modality. We primarily focus on punctuation marks, as they are easily accessible. We choose to study the potential of punctuation marks of conveying prosodical cues. In this manner, punctuation marks can be viewed as the most basic fusion scheme for multimodal representations. We provide a quantitative analysis of the impact of punctuation marks on emotion prediction with current SOTA architectures.

Chapter 6

Cross-Lingual Pre-training Methods for Spoken Dialog

Chapter 6 Abstract

Spoken dialogue systems need to be able to handle both multiple languages and multilinguality inside a conversation (*e.g* in case of code-switching). In this work, we introduce new pre-training losses tailored to learn multilingual spoken dialogue representations. The goal of these losses is to expose the model to code-switched language. To scale up training, we automatically build a pre-training corpus composed of multilingual conversations in five different languages (French, Italian, English, German and Spanish) from `OpenSubtitles`, a huge multilingual corpus composed of 24.3G tokens. We test the generic representations on `MIAM`, a new benchmark composed of five dialogue act corpora on the same aforementioned languages as well as on two novel multilingual downstream tasks (*i.e* multilingual mask utterance retrieval and multilingual inconsistency identification). Our experiments show that our new code switched-inspired losses achieve a better performance in both monolingual and multilingual settings.

6.1 Introduction

An additional difficulty to the previously mentioned modeling problem (Chapter 4 and Chapter 5) is that most people in the world are bilingual [Grosjean and Li, 2013]: therefore, progress on these systems is limited by their inability to process more than one language (English being the most frequent). For example, many people use English as a “workplace" language but seamlessly switch to their native language when the conditions are favorable [Heredia and Altarriba, 2001]. Thus, there is a growing need for understanding dialogues in a multilingual fashion [Ipsic and collab., 1999; Joshi and collab., 2020; Ruder and collab., 2019]. Additionally, when speakers share more than one language, they inevitably will engage in code-switching [Auer, 2013; Gumperz, 1982; Milroy and collab., 1995; Parekh and collab., 2020; Sankoff and Poplack, 1981]: switching between two different languages. Thus, spoken dialogue systems need to be cross lingual (*i.e* able to handle different languages) but also need to model multilinguality inside a conversation [Ahn and collab., 2020].

In this chapter, we focus on building generic representations for dialogue systems that satisfy the aforementioned requirements. Generic representations have led to strong improvements on numerous natural language understanding tasks, and can be fine-tuned when only small labelled datasets are available for the desired downstream task [DEVLIN and collab., 2018; LAN and collab., 2019; LIU and collab., 2019; MIKOLOV and collab., 2013; YANG and collab., 2019]. While there has been a growing interest in pre-training for dialogue [MEHRI and collab., 2019; ZHANG and collab., 2019d], the focus has mainly been on English datasets. Thus, these works can not be directly applied to our multilingual setting. Additionally, available multilingual pre-training objectives [LAMPLE and CONNEAU, 2019; LIU and collab., 2020; QI and collab., 2021; XUE and collab., 2020] face two main limitations when applied to dialogue modeling: (1) they are a generalization of monolingual objectives that use flat input text, whereas hierarchy has been shown to be a powerful prior for dialogue modeling. This is a reflection of a dialogue itself, for example, context plays an essential role in the labeling of dialogue acts. (2) The pre-training objectives are applied separately to each language considered, which does not expose the (possible) multilinguality inside a conversation (as it is the case for code-switching) [WINATA and collab., 2021][1].

Our main contributions are as follows:

1. *We introduce a set of code-switched inspired losses as well as a new method to automatically obtain several million of conversations with multilingual input context in different languages.* There has been limited work on proposing corpora with a sufficient amount of conversations that have multilingual input context. Most of this work focuses on social media, or on corpora of limited size. Hence, to test our new losses and scale up our pre-training, we automatically build a pre-training corpus of multilingual conversations, each of which comprises several languages, by leveraging the alignments available in OpenSubtitles (`OPS`).

2. *We showcase the relevance of the aforementioned losses and demonstrate that it leads to better performances on downstream tasks, that involve both monolingual conversations and multilingual input conversations.* For monolingual evaluation, we introduce the **M**ultilingual d**I**alog**A**ct bench**M**ark (`MIAM`): composed of five datasets in five different languages annotated with dialogue acts. Following [LOWE and collab., 2016; MEHRI and collab., 2019], we complete this task with both contextual inconsistency detection and next utterance retrieval in these five languages. For multilingual evaluation, due to the lack of code-switching corpora for spoken dialogue, we create two new tasks: contextual inconsistency detection and next utterance retrieval with multilingual input context. The datasets used for these tasks are unseen during training and automatically built from `OPS`.

In this work, we follow the recent trend [JIAO and collab., 2019; LAN and collab., 2019] in the NLP community that aims at using models of limited size that can both be pre-trained with limited computational power and achieve good performance on multiple downstream tasks. The languages we choose to work on are English, Spanish, German, French and Italian.[2]. Our implementation will be available on `github.com` and data will be available on Datasets [WOLF and collab., 2020].

[1] We refer to code-switching at the utterance level, although it is more commonly studied at the word or span level [BANERJEE and collab., 2018; BAWA and collab., 2020; FAIRCHILD and VAN HELL, 2017; POPLACK, 1980]

[2] Although our pre-training can be easily generalised to 62 languages, we use a limited number of languages to avoid exposure to the so-called "curse of multilinguality" [CONNEAU and collab., 2019]

6.2 Model and Training Objectives

Notations We start by introducing the notations. We have a set D of contexts (*i.e* truncated conversations), i.e., $D = (C_1, C_2, \ldots, C_{|D|})$. Each context C_i is composed of utterances u, i.e $C_i = (u_1^{L_1}, u_2^{L_2}, \ldots, u_{|C_i|}^{L_{|C_i|}})$ where L_i is the language of utterance u_i[3]. At the lowest level, each utterance u_i can be seen as a sequence of tokens, i.e $u_i^{L_i} = (\omega_1^i, \omega_2^i, \ldots, \omega_{|u_i|}^i)$. For DA classification y_i is the unique dialogue act tag associated to u_i. In our setting, we work with a shared vocabulary $\mathcal{V}$ thus $\omega_j^i \in \mathcal{V}$ and $\mathcal{V}$ is language independent.

6.2.1 Related work

Multilingual pre-training. Over the last few years, there has been a move towards pre-training objectives, allowing models to produce general multilingual representations that are useful for many tasks. However, they focus on the word level [FARUQUI and DYER, 2014; GOUWS and collab., 2015; MIKOLOV and collab., 2013] or the utterance level [DEVLIN and collab., 2018; ERIGUCHI and collab., 2018; LAMPLE and CONNEAU, 2019]. [WINATA and collab., 2021] shows that these models obtain poor performances in presence of code-switched data.

Pre-training to learn dialogue representation. As previously mentioned in Chapter 5, current research efforts made towards learning dialogue representation are mainly limited to the English language [CHAPUIS and collab., 2020; HENDERSON and collab., 2019; MEHRI and collab., 2019] and introduce objectives at the dialogue level such as next-utterance retrieval, next-utterance generation, masked-utterance retrieval, inconsistency identification or generalisation of the cloze task [TAYLOR, 1953]. To the best of our knowledge, this is the first work to pre-train representations for spoken dialogue in a multilingual setting.

Hierarchical pre-training As we are interested in capturing information at different granularities, we follow the hierarchical approach of Chapter 5 and decompose the pre-training objective in two terms, namely:

$$\mathcal{L}(\theta) = \underbrace{\lambda_u \times \mathcal{L}^u(\theta)}_{\text{utterance level}} + \underbrace{\lambda_d \times \mathcal{L}^d(\theta)}_{\text{dialogue level}}. \tag{6.1}$$

As in Chapter 5 these losses rely on a hierarchical encoder composed of two functions f^u and f^d:

$$\mathcal{E}_{u_i^{L_i}} = f_\theta^u(\omega_1^i, \ldots, \omega_{|u_i|}^i), \tag{6.2}$$

$$\mathcal{E}_{C_j} = f_\theta^d(\mathcal{E}_{u_1^{L_1}}, \ldots, \mathcal{E}_{u_{|C_j|}^{L_{|C_j|}}}), \tag{6.3}$$

where $\mathcal{E}_{u_i^{L_i}} \in \mathbb{R}^{d_u}$ is the embedding of $u_i^{L_i}$ and $\mathcal{E}_{C_j} \in \mathbb{R}^{d_d}$ the embedding of C_j. The encoder is built on transformer layers.

6.2.2 Utterance level pre-training

To train the first level of hierarchy (*i.e* f_θ^u), we use a Masked Utterance Modelling (MUM) loss [DEVLIN and collab., 2018]. Let $u_i^{L_i}$ be an input utterance and $\tilde{u}_i^{L_i}$ its cor-

[3]In practice, we follow SANKAR and collab. [2019a] and set the context length to 5 consecutive utterances.

Index	Speaker	Monolingual Input	Multilingual Input
0	A	Good afternoon.	Good afternoon.
1	A	I'm here to see Assistant Director Harold Cooper.	Je suis ici pour voir l'assistant directeurr Harold Cooper.
2	B	Do you have an appointment?	Do you have an appointment?
3	A	I do not.	Non.
4	A	Tell him it's Raymond Reddington.	Dites lui que c'est Raymond Reddington.

Table 6.1 – Example of automatically built input context from OPS.

rupted version, obtained after masking a proportion p_ω of tokens, the set of masked indices is denoted $\mathscr{M}_\omega$. The set of masked tokens is denoted Ω. The probability of the masked token given $\tilde{u}_i^{\mathrm{L}i}$ is given by:

$$p(\Omega|\tilde{u}_i^{\mathrm{L}i}) = \prod_{t \in \mathscr{M}_\omega} p_\theta(\omega_t^i|\tilde{u}_i^{\mathrm{L}i}). \tag{6.4}$$

6.2.3 Dialogue level pre-training

The goal of the dialogue level pre-training is to ensure that the model learns dialogue level dependencies (through f_θ^d), *i.e* the ability to handle multi-lingual input context.
Generic framework Given C_k an input context, a proportion $p_{\mathscr{C}}$ of utterances is masked to obtained the corrupted version $\tilde{\mathrm{C}}_k$. The set of masked utterances is denoted $\mathscr{U}$ and the set of corresponding masked indices $\mathscr{M}_u$. The probability of $\mathscr{U}$ given $\tilde{\mathrm{C}}_k$ is:

$$p(\mathscr{U}|\tilde{\mathrm{C}}_k) = \prod_{t \in \mathscr{M}_\mathscr{U}} \prod_{j=0}^{|u_t|-1} p_\theta(\omega_j^t|\omega_{1:j-1}^t, \tilde{\mathrm{C}}_k). \tag{6.5}$$

As shown in Equation 6.5, a masked sequence is predicted one word per step. As an example, at the j-th step, the prediction of ω_j^t is made given $(\omega_{1:j-1}^t, \tilde{\mathrm{C}}_k)$ where $\omega_{1:j-1}^t = (\omega_1^t, \cdots, \omega_{1:j-1}^t)$. In the following, we describe different procedures to build $\mathscr{M}_\mathscr{U}$ and $\tilde{\mathrm{C}}_k$ used in Equation 6.5.

Masked utterance generation (MUG)

The MUG loss aims at predicting the masked utterance from a monolingual input context. As the vocabulary is shared, this loss will improve the alignment of conversations at the dialogue level. This loss ensures that the model will be able to handle monolingual conversations in different languages.
Training Loss We rely on Equation 6.5 for MUG. The input context is composed of utterances in the same language, *i.e* $\forall k, \mathrm{C}_k = (u_1^{\mathrm{L}k}, \cdots, u_{|\mathrm{C}_k|}^{\mathrm{L}k})$. The mask is randomly chosen among all the positions.
Example Given the monolingual input context given in Table 6.1, a random mask (*e.g* [0,3]) is chosen among the positions [0,1,2,3,4]. The masked utterances are replaced by [MASK] tokens to obtain $\tilde{\mathrm{C}}_k$ and a decoder attempts to generate them.

Translation masked utterance generation (TMUG)

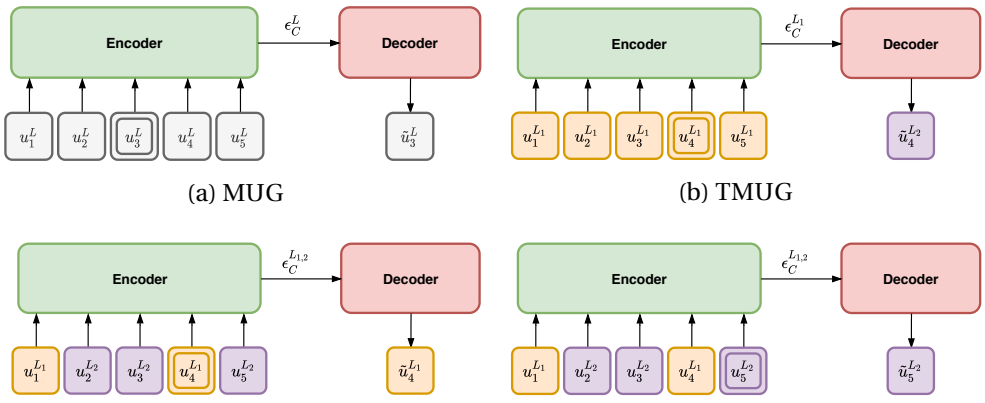
The previous objectives are self-supervised and cannot be employed with parallel data when available. In addition, these losses do not expose the model to multilinguality inside the conversation. The TMUG loss addresses this limitation using a translation mechanism: the model learns to translate the masked utterance in a new language.
Training Loss We use Equation 6.5 for TMUG with a bilingual input context C_k. C_k

contains two different languages (*i.e* L and L′) $\forall k, C_k = (u_1^{L_1}, \cdots, u_{|C_k|}^{L_k})$ with $L_i \in \{L, L'\}$. The masked positions $\mathcal{M}_u$ are all the utterances in language L′. Thus $\tilde{C}_k$ is a monolingual context.
Example Given the multilingual input context given in Table 6.1, the positions [3,4] are masked with sequences of `[MASK]` and the decoder will generate them in French.

Multilingual masked utterance generation (MMUG)

In the previous objectives, the model is exposed to monolingual input only. MMUG aims at relaxing this constraint by considering multilingual input context and generating the set of masked utterances in any possible target language.
Training Loss Given a multi-lingual input context $C_k = (u_1^{L_1}, \cdots, u_{|C_k|}^{L_{|C_k|}})$. A random set of indexes is chosen and the associated utterances are masked. The goal remains to generate the masked utterances.
Example In Table 6.1, the positions [2,3] are randomly selected from the available positions [0,1,2,3,4]. Given these masked utterances the model will generate 2 in Italien and 3 in Spanish. MMUG is closely related to code-switching as it exposes the model to multilingual context and the generation can be carried out in any language.

In Figure 6.1 we provides graphical examples for each monolingual and multilingual losses used.



(a) MUG (b) TMUG

(c) MMUG: prediction of $u_4^{L_1}$ (d) MMUG: prediction of $u_5^{L_3}$

Figure 6.1 – 6.1a and 6.1b illustrate pre-training losses using monolingual context. 6.1b and 6.1c show two scenarios for the MMUG loss using multilingual context. Double squares on the figure indicates the randomly selected utterance to predict.

Choice of scaling factor in Equation 6.1. In the case of multi-task setting, different losses may have different scales, making the optimization perform poorly. In that case, scaling factors or more advanced techniques [Sener and Koltun, 2018] can be applied. As we did not observe such phenomena, all scaling factors are set to 1.

Pre-training with generation

For both TMUG and MMUG, the model needs to be aware of the target language. Thus, the first token fed to the decoder indicates the target language (*e.g* in English the corresponding id is 99, in Spanish 98). To avoid creating a discrepancy between pre-training objectives we also add this token for MUG.

	de	en	es	fr	it
# movies	46.5K	446.5K	234.4K	127.2K	134.7K
# conversations	1.8M	18.2M	10.0M	5.2M	4.2M
# tokens	363.6M	3.7G	1.9G	1.0G	994.7M

Table 6.2 – Statistics of the processed version of OPS.

Choice of the multilingual encoder

The two dominant approaches for multilingual systems involve either using a language-specific encoder [ESCOLANO and collab., 2020] or one shared encoder across languages [ARTETXE and SCHWENK, 2019; FENG and collab., 2020]. To reduce the number of learnt parameters, we rely on the second approach.

Pre-training details

Our model is pre-trained on 4 NVIDIA V100 for 2 days (500k iterations) with a batch size of 256. We use AdamW [KINGMA and BA, 2014; LOSHCHILOV and HUTTER, 2017] with 4000 warmups steps [VASWANI and collab., 2017]. During this stage, we do not perform any grid search.

6.2.4 Pre-training corpora

There is no large corpora freely available that contains a large number of transcripts of well segmented multilingual spoken conversation with code switching phenomenon. Collecting our pre-training corpus involves two steps: the first step consists of segmenting the corpus into conversations, in the second step, we obtain aligned conversations.

Conversation segmentation Ideal pre-training corpora should contain multilingual spoken language with dialogue structure. In our work, we focus on OPS [LISON and TIEDEMANN, 2016][4] because it is the only free multilingual dialogue corpus (62 different languages). After preprocessing (see subsection 5.2.5), OPS contains around 50M of conversations and approximately 8 billion of words from the five different languages (*i.e* English, Spanish, German, French and Italian). Table 6.2 gathers statistics on the considered multilingual version of OPS. To obtain conversations from OPS, we consider that two consecutive utterances are part of the same conversation if the inter-pausal unit [KOISO and collab., 1998] (*i.e* silence between them) is shorter than $\delta_T = 6s$. If a conversation is shorter than the context size T, they are dropped and utterance are trimmed to 50 (for justification see Figure 6.2).

Obtaining aligned conversations We take advantage of the alignment files provided in OPS. They provide an alignment between utterances written in two different languages. It allows us to build aligned conversations with limited noise (solely high confidence alignments are kept). Statistics concerning the aligned conversations can be found in Table 6.3 and an example of automatically aligned context can be found in Table 6.1. The use of more advanced methods to obtain more fine-grained alignment (*e.g* word level alignment, span alignment inside an utterance) is left as future work.

[4] http://opus.nlpl.eu/OpenSubtitles-alt-v2018.php

	de-en	de-es	de-fr	de-it	en-es
# utt.	23.4M	19.9M	17.1M	14.1M	63.5M
# tokens.	217.3M	194.1M	167.0M	139.5M	590.9M
	en-fr	en-it	es-fr	es-it	fr-it
# utt.	44.2M	36.7M	37.9M	31.4M	23.8M
# tokens.	413.7M	347.1M	362.1M	304.6M	248.5M

Table 6.3 – Statistics of the processed version of the alignment files from OPS.

6.3 Evaluation Framework

This section presents our evaluation protocol. It involves two different types of evaluation depending on the input context. The first group of experiences consists in multilingual evaluations with monolingual input context and follows classical downstream tasks [DZIRI and collab., 2019; FINCH and CHOI, 2020] including sequence labeling [COLOMBO and collab., 2020], utterance retrieval [MEHRI and collab., 2019] or inconsistency detection. The second group focuses on multilingual evaluations with multilingual context.

6.3.1 Dialogue representations evaluation

Monolingual context

Sequence labeling tasks. The ability to efficiently detect and model discourse structure is an important step toward modeling spontaneous conversations. A useful first level of analysis involves the identification of dialogue act (DA) [STOLCKE and collab., 2000a] thus DA tagging is commonly used to evaluate dialogue representations. However, due to the difficulty to gather language-specific labelled datasets, multilingual sequence labeling such as DA labeling remains overlooked.
Next-utterance retrieval (NUR) The utterance retrieval task [DUPLESSIS and collab., 2017; SARACLAR and SPROAT, 2004] focuses on evaluating the ability of an encoder to model contextual dependencies. [LOWE and collab., 2016] suggests that NUR is a good indicator of how well context is modeled.
Inconsistency Identification (II) Inconsistency identification is the task of finding inconsistent utterances within a dialogue context [SANKAR and collab., 2019b]. The perturbation is as follow: one utterance is randomly replaced, the model is trained to find the inconsistent utterance.[5]

Multilingual context

To the best of our knowledge, we are the first to probe representation for multilingual spoken dialogue with multilingual input context. As there is no labeled code-switching datasets for spoken dialogue (research focuses on on synthetic data [STYMNE and collab., 2020], social media [PRATAPA and collab., 2018] or written text [KHANUJA and collab., 2020; TAN and JOTY, 2021] rather than spoken dialogue). Thus we introduce two new downstream tasks with automatically built datasets: Multilingual Next Utterance Retrieval (mNUR) and Multilingual Inconsistency Identification

[5] To ensure fair comparison, contrarily to [MEHRI and collab., 2019] the pre-training is different from the evaluation tasks.

(mII). To best assess the quality of representations, for both mII and mNUR we choose to work with train/test/validation datasets of 5k conversations. The datasets, unseen during training, are built using the procedure described in subsection 6.2.4.

Multilingual next utterance retrieval. mNUR consists of finding the most probable next utterance based on an input conversation. The evaluation dataset is built as follow: for each conversation in language L composed of T utterances, a proportion $p_{L'}$ of utterances is replaced by utterances in language L′. D utterances that we call distractors[6] in language L or L′ from the same movie. For testing, we frame the task as a ranking problem and report the recall at N (R@N) [SCHATZMANN and collab., 2005].

Multilingual inconsistency identification. The task of mII consists of identifying the index of the inconsistent sentences introduced in the conversation. Similarly to the previous task: for each conversation in language L composed of T utterances, a proportion $p_{L'}$ is replaced by utterances in language L′, a random index is sampled from [1, T] and the corresponding utterance is replaced by a negative utterance taken from the same movie.

Altering tasks difficulty

One of the interesting properties of II, mII, NUR, mNUR is the ability to alter the task difficulty in a controlled manner when sampling the negative utterances. For example, instead of randomly sampling the false utterances, the most similar to the true one as measured by a similarity metric [CELIKYILMAZ and collab., 2020; ZHANG and collab., 2019a] could be chosen. This flexibility could allow increasing the difficulty of the task as models get better.

6.3.2 Multilingual dialogue act benchmark

A plethora of freely available dialogue act dataset [GODFREY and collab., 1992; LI and collab., 2017; SHRIBERG and collab., 2004]) has been proposed to evaluate DA labeling systems in English. However, constituting a multilingual dialogue act benchmark is challenging RIBEIRO and collab. [2019b]. We introduce **M**ultilingual d**I**alogue **A**ct bench**M**ark (in short MIAM). This benchmark gathers five free corpora that have been validated by the community, in five different European languages (*i.e.* English, German, Italian, French and Spanish), examples are provided in Table 6.4. We believe that this new benchmark is challenging as it requires the model to perform well along different evaluation axis and validates the cross-lingual generalization capacity of the representations across different annotation schemes and different sizes of corpora. In Figure 6.2 we illustrate the diversity of the gathered corpora through the lens of utterance length.

DA for English For English, we choose to work on the MapTask corpus. It consists of conversations where the goal of the first speaker is to reproduce a route drawn only on the second speaker's map, with only vocal indications. We choose this corpus for its small size that will favor transfer learning approaches (36k utterances).

DA for Spanish Spanish research on DA recognition mainly focuses on three different datasets Dihana, CallHome Spanish [POST and collab., 2013] and DIME [CORIA and PINEDA, 2005; OLGUIN and CORTÉS, 2006]. Dihana is the only available corpora that contains free DA annotation [RIBEIRO and collab., 2019a]. It is a spontaneous

[6] D is set to 9 according to [LOWE and collab., 2015]

speech corpora [BENEDI and collab., 2006] composed of 900 dialogues from 225 users. Its acquisition was carried out using a Wizard of Oz setting [FRASER and GILBERT, 1991]. For this dataset, we focus on the first level of labels which is dedicated to the task-independent DA.

DA **for German** For German, we rely on the VERBMOBIL (VM2) dataset [KAY and collab., 1992]. This dataset was collected in two phases: first, multiple dialogues were recorded in an appointment scheduling scenario, then each utterance was annotated with DA using 31 domain-dependent labels. The three most common labels (*i.e.* inform, suggest and feedback) are highly related to the planning nature of the data.

DA **for French** Freely available to academic and nonprofit research datasets are limited in the French language as most available datasets are privately owned. We rely on the French dataset from the Loria Team [BARAHONA and collab., 2012] (LORIA) where the collected data consists of approximately 1250 dialogues and 10454 utterances. The tagset is composed of 31 tags.

DA **for Italian** For Italian, we rely on the Ilisten corpora [BASILE and NOVIELLI, 2018]. The corpus was collected in a Wizard of Oz setting and contains a total of 60 dialogues transcripts, 1,576 user dialogue turns and 1,611 system turns. The tag set is composed of 15 tags.

Metrics: There is no consensus on the evaluation metric for DA labelling (e.g., [GHOSAL and collab., 2019; PORIA and collab., 2018] use a weighted F-score while [ZHANG and collab., 2019c] report accuracy). As in Chapter 5 we report accuracy.

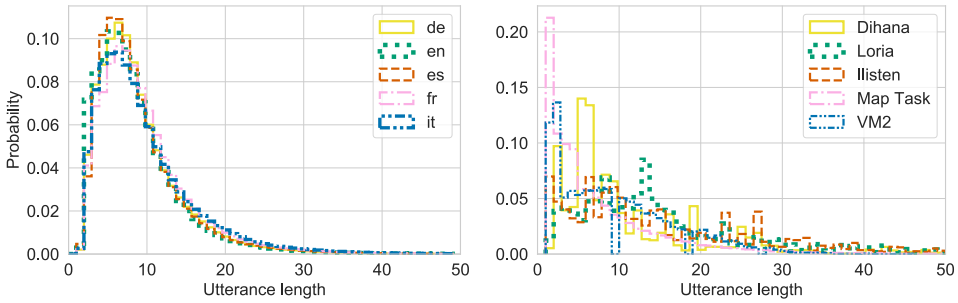


Figure 6.2 – Histograms showing the utterance length for OPS (left) and MIAM (right).

6.3.3 Baseline encoders for downstream tasks

The encoders that will serve as baselines can be divided into two different categories: hierarchical encoders based on GRU layers ($\mathscr{HR}$) and pre-trained encoders based on Transformer cells [VASWANI and collab., 2017]. The first group achieve SOTA results on several sequence labelling tasks [LI and collab., 2018; LIN and collab., 2017]. The second group can be further divided in two groups: language specific (BERT) and multilingual BERT (*m*BERT) and pre-trained hierarchical transformers from ZHANG and collab. [2019b] ($\mathscr{HT}$) are used as a common architecture to test the various pre-training losses.

Tokenizer We will work with both language specific and multlingual tokenizer. Model with multilingual tokenizer will be referred with a m (*e.g* *m*BERT as opposed to BERT).

In the following we provide details of language specific BERT and on baseline models.

Lang.	Utterances	DA
de	soll ich dann mit dem Hotel da dann die Buchung vereinbaren	OFFER
	ja das ist gut	FEED. POS.
	das wäre toll	ACCEPT
	dann kümmere ich mich um die Tickets	COMMIT
	wunderbar	ACCEPT
en	how far underneath the diamond mine	ASK
	it's about an inch or so	REPLY
	right okay five inches right along	ACK.
	up along to near a r– a ravine stuff thing	ASK
	no i don't have the ravine	REPLY
es	¿ Qué día desea salir ?	ASK
	El diez de noviembre .	REQUEST
	Quiere horarios de trenes a barcelona,	CONFIRM
	¿ desde zaragoza ?	CONFIRM
	Sí , por favor .	AFF.
fr	Bonjour	GREETINGS
	Bonjour , je suis Sophia l'opérateur (...).	GREETINGS
	Enchanté	GREETINGS
	Qu'est ce que je peux faire pour vous ?	ASK
	J'ai besoins des informations sur les composants de la manette.	INFORMER
it	mangio tre volte al giorno	STATEMENT
	Ti piace mangiare?	QUESTION
	abbastanza	ANSWER
	Che cosa hai mangiato per colazione?	QUESTION
	latte e biscotti	STATEMENT

Table 6.4 – Examples of dialogues labelled with DA taken from MapTask, Dihana, VM2, Loria and Ilisten. AFF. stands for affirmation, FEED. for feedback and ACK. for acknowledgement.

Pre-trained encoder baselines

The first group of pre-trained encoders are based on BERT. A concatenation of utterances is fed to the model to obtain a conversation embedding. For our language-specific models, we use the German BERT[7], the original BERT for English, BETO [CAÑETE and collab., 2020] for Spanish, Flaubert [LE and collab., 2019] for French and Italian BERT [SCHWETER, 2020] for Italian. We rely on the multilingual BERT (mBERT) [DEVLIN and collab., 2018][8] provided by the transformers library [WOLF and collab., 2019] implemented using the pytorch [PASZKE and collab., 2017] framework. For pre-trained hierarchical transformers, we rely on the work presented in Chapter 5 and for each considered language, we pre-train a language-specific encoder.

Decoders

Given the different nature of the proposed downstream tasks, we use various type of decoders. DA **classification**: Methods to tackle sequence labelling on monolingual

[7] https://deepset.ai/
[8] https://github.com/google-research/bert/blob/master/multilingual.md

representations can be divided into two different classes. The first one perform classification on each utterance independently using Bayesian Networks [KEIZER and collab., 2002], SVMs [SURENDRAN and LEVOW, 2006] or HMMs [STOLCKE and collab., 2000b]. The second class, which achieves stronger results, leverages the adjacency utterances by using deep representations [BOTHE and collab., 2018; KHANPOUR and collab., 2016]. Sequence labelling can be improved when sufficiently many training points are available by modelling inter-tag dependencies using RNN-based decoders [CHUNG and collab., 2014; HOCHREITER and SCHMIDHUBER, 1997], and CRFs [CHEN and collab., 2018; LAFFERTY and collab., 2001]. Thus, in this work, we choose to experiment with a MLP, a CRF and a RNN decoder based on GRU.

`II` **and** `mII` : For this task, the context embedding $\mathscr{E}_{C_k}$ is fed to a MLP. Both the encoder and the MLP are trained to predict the inconsistent utterance index by minimising a cross-entropy loss. Formally, this task is formulated as a classification problem with T classes.

`NUR` **and** `mNUR`: For this task, we first compute the context embedding $\mathscr{E}_{C_k}$ then the candidate utterance $u_c^{L_c}$ is embedded using the either f_θ^u or a chosen encoder to obtain $\mathscr{E}_{u_c^{L_c}}$. Both representations are concatenated and given to a MLP. The architecture is trained to predict if the provided candidate utterance is a suitable next utterance by minimizing a binary cross-entropy. This experiment is similar to the one in [LOWE and collab., 2015].

6.3.4 Additional details on models

In this section, we describe models used as well as details on the pre-training parameters. In Table 6.5 we report the main hyper-parameters used for our model pre-training. We used GELU [HENDRYCKS and GIMPEL, 2016] activations and the dropout rate [SRIVASTAVA and collab., 2014] is set to 0.1. Although vanilla Transformers impose a fixed context size it can be relaxed [DAI and collab., 2019]. We follow [COLOMBO and collab., 2020; SANKAR and collab., 2019a] and set T = 5. We rely on the tokenizers provided by the HuggingFace library based on the SentencePiece [KUDO and RICHARDSON, 2018] and WordPiece [WU and collab., 2016] algorithms. In all experiments, for our models relying on the $\mathscr{HT}$ we use the same architecture as the `SMALL` model developped in Chapter 5 which contains 80 millions parameters. Original BERT has 167 millions parameters and is pre-trained using 16 TPUs during several days with over 500K iterations.

For each task, model are fine-tuned and dropout [SRIVASTAVA and collab., 2014] is set to 0.1. The best learning rate is found in $\{0.01, 0.001, 0.0001\}$ and chosen based on the validation loss.

	Pre-trained Encoder
Nbs of heads	6
N_d	4
N_u	4
T	50
C	5
$\mathscr{T}_d$ nbs of heads	6
Inner dimension	768
Model Dimension	768
$\|\mathscr{V}\|$	105879
$\mathscr{T}_d$: Emb. size	768
d_k:	64
d_v:	64

Table 6.5 – Architecture hyperparameters used for the hierarchical pre-training.

6.4 Numerical Results

In this section, we empirically demonstrate the effectiveness of our code-switched inspired pre-training on downstream tasks involving both monolingual and multilingual input context.

6.4.1 Monolingual input context

`DA` labeling

Global analysis. Table 6.8 reports the results of the different models on `MIAM`. Table 6.8 is composed of two distinct groups of models: *language specifics models* (with language-specific tokenizers) and *multilingual models* (with a multilingual tokenizer denoted with a *m* before the model name). Overall, we observe that *m*MUG augmented with both TMUG and MMUG gets a boost in performance (1.8% compared to *m*MUG and 2.6% compared to a mBERT model with a similar number of parameters). This result shows that the model benefits from being exposed to aligned bilingual conversations and that our proposed losses (*i.e.* TMUG and MMUG) are useful to help the model to better catch contextual information for `DA` labeling.
Language-specific v.s. multilingual models. By comparing the performances of $\mathscr{HR}$ (with either a CRF or MLP decoder), we can notice that for these models on `DA` labelling it is better to use a multilingual tokenizer. As multilingual tokenizers are not tailored for a specific language and have roughly twice as many tokens than their language-specific counterparts, one would expect that models trained from scratch using language-specific tokenizers would achieve better results. We believe this result is related to the spoken nature of `MIAM` and further investigations are left as future work. Recent work RUST and collab. [2020] has demonstrated that pre-trained language models with language-specific tokenizers achieve better results than those using multilingual tokenizers. This result could explain the higher accuracy achieved by the language-specific versions of MUG compared to *m*MUG.
We additionally observe that some language-specific versions of BERT achieve lower results (*e.g* `Dihana`, `Loria`) than the multilingual version which could suggest that these pre-trained BERT might be less carefully trained than the multilingual one; in the next part of the analysis we will only use multilingual tokenizers.

	VM2	Map Task	Dihana	Loria	Ilisten	Total
$m\mathcal{HT}(\theta_{written})$	52.8	64.6	98.1	76.5	**74.2**	73.2
$m\mathcal{HT}_u(\theta_{spoken})$	**53.0**	**67.3**	**98.3**	**78.5**	74.0	**74.2**

Table 6.6 – Ablation studies on pre-training data. We report the accuracy on MIAM for the $m\mathcal{HT}$. $m\mathcal{HT}_u(\theta_{spoken})$ stands for the model pre-trained with the utterance level loss $m\mathcal{L}^u$ on spoken data and $m\mathcal{HT}(\theta_{written})$ stands for a hierarchical encoder where sentence embeddings is computed using a pre-trained BERT encoder.

Ablation study on pre-training data We showcase the difference between pre-training with spoken and written corpora. We compare $m\mathcal{HT}(\theta_{written})$, a hierarchical encoder where each utterance is embedded using the representation of the [CLS] token given by the second layer of BERT, and $m\mathcal{HT}_u(\theta_{spoken})$, a model pre-trained on OPS using $\mathcal{L}^u$ only. The prediction is performed by feeding the utterance embeddings to a simple MLP. In Table 6.6, we report the results on MIAM. Results demonstrate an overall higher accuracy when the pre-training is performed on spoken data. This supports the choice of OPS as pre-training corpora and demonstrates that the origin of the pre-training data matters.

Overall, pre-trained models achieve better results. Contrarily to what can be observed in some syntactic tagging tasks [ZHANG and BOWMAN, 2018], for DA tagging pre-trained models achieve consistently better results on the full benchmark. This result of multilingual models confirms what is observed with monolingual data (see [MEHRI and collab., 2019]): pre-training is an efficient method to build accurate dialogue sequence labellers.

Comparison of pre-training losses In Table 6.8 we dissect the relative improvement brought by the different parts of the code-switched inspired losses and the architecture to better understand the relative importance of each component. Similarly to Chapter 5, we see that the hierarchical pre-training on spoken data (see *m*MUG) improves over the *m*BERT model. Interestingly, we observe that the monolingual pre-training works slightly better compared to the multilingual pre-training when training using the same loss. This result surprising results might be attributed to the limited size of our models [KARTHIKEYAN and collab., 2019].

We see that in both cases, introducing a loss with aligned multilingual conversations (MMUG or TMUG) induces a performance gain (+1.5%). This suggests that our pre-training with the new losses better captures the data distribution. By comparing the results of *m*MUG + TMUG with *m*MUG, we observe that the addition of cross-lingual generation during pre-training helps. A marginal gain is induced when using MMUG over TMUG, thus we believe that the improvement of *m*MUG + MMUG over *m*MUG can mainly be attributed to the cross-lingual generation part. Interestingly, we observe that the combination of all losses out-performs the other models which suggests that different losses model different patterns present in the data.

Inconsistency Identification

In this section, we follow MEHRI and collab. [2019] and evaluate our pre-trained representations on II with a monolingual context. A random guess identifies the inconsistency by randomly selecting an index in [1, T] which corresponds to an accuracy of 20% (as we have set T = 5). Table 6.7 gathers the results. Similarly conclusion than in section 6.4.1 can be drawn: pre-trained models achieve better results and the best performing model is obtained with *m*MUG+MMUG+TMUG.

Next utterance retrieval

In this section, we evaluate our representations on NUR using a monolingual input context. As we use 9 distractors, a random classifier would achieve 0.10 for R@1, 0.20 for R@2 and 0.50 for R@5. The results are presented in Table 6.10. When comparing the accuracy obtained by the baselines models (*e.g* mBERT, mBERT (4-layers) and $\mathcal{HR}$) and our model using the contextual losses at the context level for pre-training (*i.e* MUG, TMUG and MMUG) we observe a consistent improvement.

Takeaways Across all the three considered tasks, we observe that the models pre-trained with our losses achieve better performances. We believe it is indicative of the validity of our pre-training.

6.4.2 Multilingual input context

In this section, we present the results on the downstream tasks with multilingual input context.

Multilingual inconsistency identification

Table 6.9 gathers the results for the mII with bilingual input context. As previously a random baseline would achieve an accuracy of 20%. As expected predicting inconsistency with bilingual context is more challenging than with a monolingual context: we observe a drop in performance of around 15% for all methods including the multilingual BERT. Our results confirm the observation of WINATA and collab. [2021]: multilingual pre-training does not guarantee good performance in code switched data. However, we observe that the losses, by exposing the model with bilingual context, obtain a large boost (absolute improvement of 6% which correspond to a relative boost of more than 20%). We also observe that MUG+MMUG+TMUG outperforms *m*BERT on all pairs, with fewer parameters.

Multilingual next utterance retrieval

The results on bilingual context for mNUR are presented in Table 6.11. mNUR is more challenging than NUR. Overall, we observe a strong gain in performance when exposing the model to bilingual context (gain over 9% absolute point in R@5).
Takeaways: These results show that our code-switched inspired losses help to learn better representations in a particularly effective way in the case of multilingual input context.

	de	en	es	fr	it	Avg
*m*BERT	44.6	42.9	43.7	43.5	42.3	43.4
*m*BERT (4-layers)	44.6	42.1	43.7	42.5	41.4	42.9
$m\mathcal{HR}$	44.1	42.0	40.4	41.3	41.2	41.8
*m*MUG	45.2	43.5	45.1	43.1	42.7	43.9
*m*MUG + TMUG	48.2	42.6	**47.7**	44.6	44.3	45.5
*m*MUG + MMUG	**49.6**	**43.8**	46.1	**46.2**	43.3	45.8
*m*MUG + TMUG + MMUG	49.1	43.4	46.2	45.9	**45.1**	**46.0**

Table 6.7 – Results on the II task with monolingual input context. On this task the accuracy is reported.

	Toke.	VM2	Map Task	Dihana	Loria	Ilisten	Total
BERT	lang	54.7	66.4	86.0	50.2	74.9	66.4
BERT - $4layers$	lang	52.8	66.2	85.8	55.2	76.2	67.2
$\mathcal{HR}$ + CRF	lang	49.7	63.1	85.8	73.4	75.2	69.4
$\mathcal{HR}$ + MLP	lang	51.3	63.0	85.6	58.9	75.0	66.8
MUG Chapter 5	lang	54.0	66.4	99.0	79.0	74.8	74.6
mBERT	multi	53.2	66.4	98.7	76.2	74.9	73.8
mBERT - $4layers$	multi	52.7	66.2	98.0	75.1	75.0	73.4
$m\mathcal{HR}$ + CRF	multi	49.8	65.2	97.6	75.2	76.0	72.8
$m\mathcal{HR}$ + MLP	multi	51.0	65.7	97.8	75.2	76.0	73.1
mMUG	multi	53.0	67.3	98.3	78.5	74.0	74.2
mMUG + TMUG	multi	54.8	**67.4**	99.1	**80.8**	74.9	75.4
mMUG + MMUG	multi	**56.2**	**67.4**	99.0	78.9	**77.6**	75.8
mMUG + TMUG + MMUG	multi	**56.2**	66.7	**99.3**	80.7	77.0	**76.0**

Table 6.8 – Accuracy of pre-trained and baseline encoders on MIAM. Models are divided in three groups: hierarchical transformer encoders pre-trained using our custom losses, baselines (see subsection 6.3.4) using either multilingual or language specific tokenizer. *Toke.* stands for the type of tokenizer: *multi* and *lang* denotes a pre-trained tokenizer on multilingual and language specific data respectively. When using $lang$ tokenizer, MUG pre-training and finetuning are performed on the same language.

	de-en	de-es	de-fr	de-it	en-es	en-fr	en-it	es-fr	es-it	fr-it	Avg
mBERT	31.2	28.0	28.0	27.6	28.4	33.0	32.1	35.1	31.0	28.7	30.3
mBERT (4-layers)	30.7	28.7	28.2	27.1	28.7	33.1	30.9	35.1	30.1	28.1	30.1
$m\mathcal{HR}$	28.7	27.9	26.9	27.3	25.5	25.1	30.6	34.3	30.0	26.8	28.3
mMUG	34.5	30.1	30.1	27.7	28.2	33.1	32.1	35.4	32.0	29.5	31.2
mMUG + TMUG	34.0	32.0	32.2	29.1	28.3	32.9	32.4	35.1	33.0	29.3	31.8
mMUG + MMUG	35.1	33.8	**34.0**	30.1	29.4	32.8	32.6	36.1	33.9	31.6	32.9
mMUG + TMUG + MMUG	**35.7**	**34.0**	32.5	**31.4**	**30.1**	**33.6**	**33.9**	**36.2**	**34.0**	**32.1**	**33.4**

Table 6.9 – Results on the mII task with bilingual input context.

	de			en			es			fr			it		
	R@5	R@2	R@1	R@5	R@2	R@1	R@5	R@2	R@1	R@5	R@2	R@1	R@5	R@2	R@1
mBERT	65.1	27.1	20.1	62.1	26.1	16.8	62.4	24.8	15.3	63.9	22.9	13.4	66.1	27.8	16.9
mBERT (4-layers)	65.1	27.5	20.2	61.4	25.6	15.1	62.3	24.6	15.9	63.4	22.8	12.9	65.6	27.4	15.8
$m\mathcal{HR}$	65.0	27.1	20.0	60.3	25.0	15.2	61.0	23.9	14.7	63.0	22.9	13.0	65.4	27.3	15.8
mMUG	66.9	28.0	20.0	65.9	26.4	16.3	66.7	26.4	16.4	66.2	25.2	17.2	68.9	28.9	17.2
mMUG + TMUG	67.2	**28.2**	20.1	68.3	**29.8**	17.5	69.0	26.9	17.3	67.1	**25.4**	17.3	69.9	29.4	18.6
mMUG + MMUG	66.9	28.1	20.7	68.1	26.7	18.0	68.7	26.9	17.5	67.2	25.2	**17.4**	69.7	29.4	18.6
mMUG + TMUG + MMUG	**68.3**	27.4	**21.2**	**68.9**	27.8	**18.3**	**69.3**	**27.1**	**17.9**	**67.4**	25.3	**17.4**	**70.2**	**30.0**	**18.7**

Table 6.10 – Results on the NUR task with monolingual input context. R@N stands for recall at N.

	de-en			de-es			de-fr			de-it			en-es		
	R@5	R@2	R@1	R@5	R@2	R@1	R@5	R@2	R@1	R@5	R@2	R@1	R@5	R@2	R@1
mBERT	54.4	27.0	11.6	55.9	24.8	11.9	57.9	24.2	12.9	57.5	23.9	13.0	55.4	25.6	13.0
mBERT (4-layers)	54.1	26.5	11.9	55.7	24.8	**12.4**	57.2	24.1	12.4	57.0	23.5	13.1	55.6	23.1	12.9
$m\mathcal{HR}$	52.1	25.5	12.1	54.9	14.6	10.7	56.1	22.9	11.3	56.9	24.9	13.0	53.9	23.7	12.8
mMUG	59.7	25.2	11.5	61.2	26.2	11.6	60.7	25.3	13.8	61.6	**26.4**	11.9	62.1	23.9	13.10
mMUG + TMUG	59.8	26.2	12.1	62.7	29.0	10.7	61.9	27.3	13.9	63.2	26.3	12.6	63.1	28.4	14.0
mMUG + MMUG	59.8	27.2	12.1	62.7	28.1	11.6	60.7	24.8	**14.4**	62.7	26.1	**13.8**	63.4	28.2	14.7
mMUG + TMUG + MMUG	**61.0**	**28.2**	**13.1**	**63.2**	**29.1**	11.7	**62.1**	**28.7**	14.1	**63.4**	26.3	12.9	**64.3**	**29.4**	**15.2**
	en-fr			en-it			es-fr			es-it			fr-it		
	R@5	R@2	R@1	R@5	R@2	R@1	R@5	R@2	R@1	R@5	R@2	R@1	R@5	R@2	R@1
mBERT	57.9	25.4	12.3	57.1	23.5	12.1	57.8	27.9	12.2	54.2	22.1	11.2	58.1	22.9	12.5
mBERT (4-layers)	57.8	23.2	12.1	57.1	23.4	11.9	57.1	27.6	12.1	55.1	22.0	11.1	58.9	22.6	12.7
$m\mathcal{HR}$	55.9	20.9	11.6	56.8	22.9	11.8	54.9	27.0	12.0	53.9	21.0	11.6	56.1	21.9	11.4
mMUG	61.9	24.9	12.9	61.4	27.6	11.9	64.6	29.7	13.9	59.0	24.2	13.4	59.7	23.6	12.2
mMUG + TMUG	62.9	25.2	14.3	62.7	27.8	12.9	64.9	29.9	13.8	**60.1**	25.1	13.5	61.5	25.8	13.1
mMUG + MMUG	63.9	26.3	**14.7**	61.5	27.6	13.1	65.0	30.2	13.1	**60.1**	25.3	12.9	**63.1**	**25.9**	13.6
mMUG + TMUG + MMUG	**64.0**	**26.7**	14.1	**63.5**	**28.7**	**13.7**	**66.1**	**31.4**	**14.5**	**60.1**	**25.5**	**13.6**	**63.1**	**25.9**	**14.2**

Table 6.11 – Results on the mNUR task with bilingual input context.

Chapter 6 Conclusion

In this work, we showed the strong impact of the studied set of pre-training losses, tailored for multilingual spoken dialogue data. These losses induce a significant improvement by leveraging the parallel conversations extracted from `OpenSubtitles`. They achieved solid results on the new presented benchmark for dialogue act tagging (`MIAM`), available on 5 European languages. We also observe a significant improvement on the two new introduced tasks: multilingual inconsistency identification and multilingual next utterance retrieval.

6.5 References

AHN, E., C. JIMENEZ, Y. TSVETKOV and A. W. BLACK. 2020, "What code-switching strategies are effective in dialog systems?", in *Proceedings of the Society for Computation in Linguistics 2020*, p. 213–222. 109

ARTETXE, M. and H. SCHWENK. 2019, "Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond", *Transactions of the Association for Computational Linguistics*, vol. 7, p. 597–610. 114

AUER, P. 2013, *Code-switching in conversation: Language, interaction and identity*, Routledge. 109

BANERJEE, S., N. MOGHE, S. ARORA and M. M. KHAPRA. 2018, "A dataset for building code-mixed goal oriented conversation systems", *arXiv preprint arXiv:1806.05997*. 110

BARAHONA, L. M. R., A. LORENZO and C. GARDENT. 2012, "Building and exploiting a corpus of dialog interactions between french speaking virtual and human agents", . 117

BASILE, P. and N. NOVIELLI. 2018, "Overview of the evalita 2018 italian speech act labe ling (iliste n) task", *EVALITA Evaluation of NLP and Speech Tools for Italian*, vol. 12, p. 44. 117

BAWA, A., P. KHADPE, P. JOSHI, K. BALI and M. CHOUDHURY. 2020, "Do multilingual users prefer chat-bots that code-mix? let's nudge and find out!", *Proceedings of the ACM on Human-Computer Interaction*, vol. 4, nº CSCW1, p. 1–23. 110

BENEDI, J.-M., E. LLEIDA, A. VARONA, M.-J. CASTRO, I. GALIANO, R. JUSTO, I. LÓPEZ and A. MIGUEL. 2006, "Design and acquisition of a telephone spontaneous speech dialogue corpus in spanish: Dihana", in *Fifth International Conference on Language Resources and Evaluation (LREC)*, p. 1636–1639. 117

BOTHE, C., C. WEBER, S. MAGG and S. WERMTER. 2018, "A context-based approach for dialogue act recognition using simple recurrent neural networks", *CoRR*, vol. abs/1805.06280. 119

CAÑETE, J., G. CHAPERON, R. FUENTES, J.-H. HO, H. KANG and J. PÉREZ. 2020, "Spanish pre-trained bert model and evaluation data", in *PML4DC at ICLR 2020*. 118

CELIKYILMAZ, A., E. CLARK and J. GAO. 2020, "Evaluation of text generation: A survey", *arXiv preprint arXiv:2006.14799*. 116

CHAPUIS, E., P. COLOMBO, M. MANICA, M. LABEAU and C. CLAVEL. 2020, "Hierarchical pre-training for sequence labelling in spoken dialog", in *Findings of the Association for Computational Linguistics: EMNLP 2020*, Association for Computational Linguistics, Online, p. 2636–2648, doi: 10.18653/v1/2020.findings-emnlp.239. URL https://www.aclweb.org/anthology/2020.findings-emnlp.239. 111

CHEN, Z., R. YANG, Z. ZHAO, D. CAI and X. HE. 2018, "Dialogue act recognition via crf-attentive structured network", in *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, ACM, p. 225–234. 119

CHUNG, J., C. GULCEHRE, K. CHO and Y. BENGIO. 2014, "Empirical evaluation of gated recurrent neural networks on sequence modeling", *arXiv preprint arXiv:1412.3555*. 119

COLOMBO, P., E. CHAPUIS, M. MANICA, E. VIGNON, G. VARNI and C. CLAVEL. 2020, "Guiding attention in sequence-to-sequence models for dialogue act prediction", *arXiv preprint arXiv:2002.08801*. 115, 119

CONNEAU, A., K. KHANDELWAL, N. GOYAL, V. CHAUDHARY, G. WENZEK, F. GUZMÁN, E. GRAVE, M. OTT, L. ZETTLEMOYER and V. STOYANOV. 2019, "Unsupervised cross-lingual representation learning at scale", *arXiv preprint arXiv:1911.02116*. 110

CORIA, S. and L. PINEDA. 2005, "Predicting obligation dialogue acts from prosodic and speaker infomation", *Research on Computing Science (ISSN 1665-9899), Centro de Investigacion en Computacion, Instituto Politecnico Nacional, Mexico City*. 116

DAI, Z., Z. YANG, Y. YANG, J. CARBONELL, Q. V. LE and R. SALAKHUTDINOV. 2019, "Transformer-xl: Attentive language models beyond a fixed-length context", *arXiv preprint arXiv:1901.02860*. 119

DEVLIN, J., M.-W. CHANG, K. LEE and K. TOUTANOVA. 2018, "Bert: Pre-training of deep bidirectional transformers for language understanding", *arXiv preprint arXiv:1810.04805*. 110, 111, 118

DUPLESSIS, G. D., F. CHARRAS, V. LETARD, A.-L. LIGOZAT and S. ROSSET. 2017, "Utterance retrieval based on recurrent surface text patterns", in *European Conference on Information Retrieval*, Springer, p. 199–211. 115

DZIRI, N., E. KAMALLOO, K. W. MATHEWSON and O. ZAIANE. 2019, "Evaluating coherence in dialogue systems using entailment", *arXiv preprint arXiv:1904.03371*. 115

ERIGUCHI, A., M. JOHNSON, O. FIRAT, H. KAZAWA and W. MACHEREY. 2018, "Zero-shot cross-lingual classification using multilingual neural machine translation", *arXiv preprint arXiv:1809.04686*. 111

ESCOLANO, C., M. R. COSTA-JUSSÀ, J. A. FONOLLOSA and M. ARTETXE. 2020, "Training multilingual machine translation by alternately freezing language-specific encoders-decoders", *arXiv preprint arXiv:2006.01594*. 114

FAIRCHILD, S. and J. G. VAN HELL. 2017, "Determiner-noun code-switching in spanish heritage speakers", *Bilingualism: Language and Cognition*, vol. 20, nº 1, p. 150–161. 110

FARUQUI, M. and C. DYER. 2014, "Improving vector space word representations using multilingual correlation", in *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, p. 462–471. 111

FENG, F., Y. YANG, D. CER, N. ARIVAZHAGAN and W. WANG. 2020, "Language-agnostic bert sentence embedding", *arXiv preprint arXiv:2007.01852*. 114

FINCH, S. E. and J. D. CHOI. 2020, "Towards unified dialogue system evaluation: A comprehensive analysis of current evaluation protocols", *arXiv preprint arXiv:2006.06110*. 115

FRASER, N. M. and G. N. GILBERT. 1991, "Simulating speech systems", *Computer Speech & Language*, vol. 5, nº 1, p. 81–99. 117

GHOSAL, D., N. MAJUMDER, S. PORIA, N. CHHAYA and A. GELBUKH. 2019, "Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation", *arXiv preprint arXiv:1908.11540*. 117

GODFREY, J. J., E. C. HOLLIMAN and J. MCDANIEL. 1992, "Switchboard: Telephone speech corpus for research and development", in *Proceedings of the 1992 IEEE International Conference on Acoustics, Speech and Signal Processing - Volume 1*, ICASSP'92, IEEE Computer Society, USA, ISBN 0780305329, p. 517–520. 116

GOUWS, S., Y. BENGIO and G. CORRADO. 2015, "Bilbowa: Fast bilingual distributed representations without word alignments", . 111

GROSJEAN, F. and P. LI. 2013, *The psycholinguistics of bilingualism*, John Wiley & Sons. 109

GUMPERZ, J. J. 1982, *Discourse strategies*, 1, Cambridge University Press. 109

HENDERSON, M., I. CASANUEVA, N. MRKŠIĆ, P.-H. SU, T.-H. WEN and I. VULIĆ. 2019, "Convert: Efficient and accurate conversational representations from transformers", *arXiv preprint arXiv:1911.03688*. 111

HENDRYCKS, D. and K. GIMPEL. 2016, "Gaussian error linear units (gelus)", *arXiv preprint arXiv:1606.08415*. 119

HEREDIA, R. R. and J. ALTARRIBA. 2001, "Bilingual language mixing: Why do bilinguals code-switch?", *Current Directions in Psychological Science*, vol. 10, nº 5, p. 164–168. 109

HOCHREITER, S. and J. SCHMIDHUBER. 1997, "Long short-term memory", *Neural computation*, vol. 9, nº 8, p. 1735–1780. 119

IPSIC, I., N. PAVESIC, F. MIHELIC and E. NOTH. 1999, "Multilingual spoken dialog system", in *ISIE'99. Proceedings of the IEEE International Symposium on Industrial Electronics (Cat. No. 99TH8465)*, vol. 1, IEEE, p. 183–187. 109

JIAO, X., Y. YIN, L. SHANG, X. JIANG, X. CHEN, L. LI, F. WANG and Q. LIU. 2019, "Tinybert: Distilling bert for natural language understanding", *arXiv preprint arXiv:1909.10351*. 110

JOSHI, P., S. SANTY, A. BUDHIRAJA, K. BALI and M. CHOUDHURY. 2020, "The state and fate of linguistic diversity and inclusion in the nlp world", *arXiv preprint arXiv:2004.09095*. 109

KARTHIKEYAN, K., Z. WANG, S. MAYHEW and D. ROTH. 2019, "Cross-lingual ability of multilingual bert: An empirical study", in *International Conference on Learning Representations*. 121

KAY, M., P. NORVIG and M. GAWRON. 1992, *Verbmobil: A translation system for face-to-face dialog*, University of Chicago Press. 117

KEIZER, S., R. OP DEN AKKER and A. NIJHOLT. 2002, "Dialogue act recognition with bayesian networks for dutch dialogues", in *Proceedings of the Third SIGdial Workshop on Discourse and Dialogue*. 119

KHANPOUR, H., N. GUNTAKANDLA and R. NIELSEN. 2016, "Dialogue act classification in domain-independent conversations using a deep recurrent neural network", in *COLING*. 119

KHANUJA, S., S. DANDAPAT, A. SRINIVASAN, S. SITARAM and M. CHOUDHURY. 2020, "Gluecos: An evaluation benchmark for code-switched nlp", *arXiv preprint arXiv:2004.12376*. 115

KINGMA, D. P. and J. BA. 2014, "Adam: A method for stochastic optimization", *arXiv preprint arXiv:1412.6980*. 114

KOISO, H., Y. HORIUCHI, S. TUTIYA, A. ICHIKAWA and Y. DEN. 1998, "An analysis of turn-taking and backchannels based on prosodic and syntactic features in japanese map task dialogs", *Language and speech*, vol. 41, nº 3-4, p. 295–321. 114

KUDO, T. and J. RICHARDSON. 2018, "SentencePiece: A Simple and Language Independent Subword Tokenizer and Detokenizer for Neural Text Processing", in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations (EMNLP)*, Association for Computational Linguistics, Brussels, Belgium, p. 66–71, doi: 10.18653/v1/D18-2012. URL https://www.aclweb.org/anthology/D18-2012. 119

LAFFERTY, J., A. MCCALLUM and F. C. PEREIRA. 2001, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data", . 119

LAMPLE, G. and A. CONNEAU. 2019, "Cross-lingual language model pretraining", *arXiv preprint arXiv:1901.07291*. 110, 111

LAN, Z., M. CHEN, S. GOODMAN, K. GIMPEL, P. SHARMA and R. SORICUT. 2019, "Albert: A lite bert for self-supervised learning of language representations", *arXiv preprint arXiv:1909.11942*. 110

LE, H., L. VIAL, J. FREJ, V. SEGONNE, M. COAVOUX, B. LECOUTEUX, A. ALLAUZEN, B. CRABBÉ, L. BESACIER and D. SCHWAB. 2019, "Flaubert: Unsupervised language model pre-training for french", *arXiv preprint arXiv:1912.05372*. 118

LI, R., C. LIN, M. COLLINSON, X. LI and G. CHEN. 2018, "A dual-attention hierarchical recurrent neural network for dialogue act classification", *CoRR*, vol. abs/1810.09154. URL http://arxiv.org/abs/1810.09154. 117

LI, Y., H. SU, X. SHEN, W. LI, Z. CAO and S. NIU. 2017, "Dailydialog: A manually labelled multi-turn dialogue dataset", . 116

LIN, Z., M. FENG, C. N. D. SANTOS, M. YU, B. XIANG, B. ZHOU and Y. BENGIO. 2017, "A structured self-attentive sentence embedding", *arXiv preprint arXiv:1703.03130*. 117

LISON, P. and J. TIEDEMANN. 2016, "Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles", . 114

LIU, Y., J. GU, N. GOYAL, X. LI, S. EDUNOV, M. GHAZVININEJAD, M. LEWIS and L. ZETTLEMOYER. 2020, "Multilingual denoising pre-training for neural machine translation", *Transactions of the Association for Computational Linguistics*, vol. 8, p. 726–742. 110

LIU, Y., M. OTT, N. GOYAL, J. DU, M. JOSHI, D. CHEN, O. LEVY, M. LEWIS, L. ZETTLEMOYER and V. STOYANOV. 2019, "Roberta: A robustly optimized bert pretraining approach", *arXiv preprint arXiv:1907.11692*. 110

LOSHCHILOV, I. and F. HUTTER. 2017, "Decoupled weight decay regularization", *arXiv preprint arXiv:1711.05101*. 114

LOWE, R., N. POW, I. SERBAN and J. PINEAU. 2015, "The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems", *CoRR*, vol. abs/1506.08909. URL http://arxiv.org/abs/1506.08909. 116, 119

LOWE, R., I. V. SERBAN, M. NOSEWORTHY, L. CHARLIN and J. PINEAU. 2016, "On the evaluation of dialogue systems with next utterance classification", *arXiv preprint arXiv:1605.05414*. 110, 115

MEHRI, S., E. RAZUMOVSAKAIA, T. ZHAO and M. ESKENAZI. 2019, "Pretraining methods for dialog context representation learning", *arXiv preprint arXiv:1906.00414*. 110, 111, 115, 121

MIKOLOV, T., Q. V. LE and I. SUTSKEVER. 2013, "Exploiting similarities among languages for machine translation", *arXiv preprint arXiv:1309.4168*. 110, 111

MILROY, J. and collab.. 1995, *One speaker, two languages: Cross-disciplinary perspectives on code-switching*, Cambridge University Press. 109

OLGUIN, S. R. C. and L. A. P. CORTÉS. 2006, "Predicting dialogue acts from prosodic information", in *International Conference on Intelligent Text Processing and Computational Linguistics*, Springer, p. 355–365. 116

PAREKH, T., E. AHN, Y. TSVETKOV and A. W. BLACK. 2020, "Understanding linguistic accommodation in code-switched human-machine dialogues", in *Proceedings of the 24th Conference on Computational Natural Language Learning*, p. 565–577. 109

PASZKE, A., S. GROSS, S. CHINTALA, G. CHANAN, E. YANG, Z. DEVITO, Z. LIN, A. DESMAISON, L. ANTIGA and A. LERER. 2017, "Automatic differentiation in pytorch", . 118

POPLACK, S. 1980, "Sometimes i'll start a sentence in spanish y termino en espanol: toward a typology of code-switching1", . 110

PORIA, S., D. HAZARIKA, N. MAJUMDER, G. NAIK, E. CAMBRIA and R. MIHALCEA. 2018, "Meld: A multimodal multi-party dataset for emotion recognition in conversations", *arXiv preprint arXiv:1810.02508*. 117

POST, M., G. KUMAR, A. LOPEZ, D. KARAKOS, C. CALLISON-BURCH and S. KHUDANPUR. 2013, "Improved speech-to-text translation with the Fisher and Callhome Spanish–English speech translation corpus", in *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, Heidelberg, Germany. 116

PRATAPA, A., M. CHOUDHURY and S. SITARAM. 2018, "Word embeddings for code-mixed language processing", in *Proceedings of the 2018 conference on empirical methods in natural language processing*, p. 3067–3072. 115

QI, W., Y. GONG, Y. YAN, C. XU, B. YAO, B. ZHOU, B. CHENG, D. JIANG, J. CHEN, R. ZHANG and collab.. 2021, "Prophetnet-x: Large-scale pre-training models for english, chinese, multi-lingual, dialog, and code generation", *arXiv preprint arXiv:2104.08006*. 110

RIBEIRO, E., R. RIBEIRO and D. M. DE MATOS. 2019a, "Hierarchical multi-label dialog act recognition on spanish data", *arXiv preprint arXiv:1907.12316*. 116

RIBEIRO, E., R. RIBEIRO and D. M. DE MATOS. 2019b, "A multilingual and multidomain study on dialog act recognition using character-level tokenization", *Information*, vol. 10, nº 3, p. 94. 116

RUDER, S., I. VULIĆ and A. SØGAARD. 2019, "A survey of cross-lingual word embedding models", *Journal of Artificial Intelligence Research*, vol. 65, p. 569–631. 109

RUST, P., J. PFEIFFER, I. VULIĆ, S. RUDER and I. GUREVYCH. 2020, "How good is your tokenizer? on the monolingual performance of multilingual language models", *arXiv preprint arXiv:2012.15613*. 120

SANKAR, C., S. SUBRAMANIAN, C. PAL, S. CHANDAR and Y. BENGIO. 2019a, "Do neural dialog systems use the conversation history effectively? an empirical study", *arXiv preprint arXiv:1906.01603*. 111, 119

SANKAR, C., S. SUBRAMANIAN, C. PAL, S. CHANDAR and Y. BENGIO. 2019b, "Do neural dialog systems use the conversation history effectively? an empirical study", *arXiv preprint arXiv:1906.01603*. 115

SANKOFF, D. and S. POPLACK. 1981, "A formal grammar for code-switching", *Research on Language & Social Interaction*, vol. 14, nº 1, p. 3–45. 109

SARACLAR, M. and R. SPROAT. 2004, "Lattice-based search for spoken utterance retrieval", in *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, p. 129–136. 115

SCHATZMANN, J., K. GEORGILA and S. YOUNG. 2005, "Quantitative evaluation of user simulation techniques for spoken dialogue systems", in *6th SIGdial Workshop on DISCOURSE and DIALOGUE*. 116

SCHWETER, S. 2020, "Italian bert and electra models", doi: 10.5281/zenodo.4263142. URL https://doi.org/10.5281/zenodo.4263142. 118

SENER, O. and V. KOLTUN. 2018, "Multi-task learning as multi-objective optimization", in *Advances in Neural Information Processing Systems*, p. 527–538. 113

SHRIBERG, E., R. DHILLON, S. BHAGAT, J. ANG and H. CARVEY. 2004, "The ICSI meeting recorder dialog act (MRDA) corpus", in *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue at HLT-NAACL 2004*, Association for Computational Linguistics, Cambridge, Massachusetts, USA, p. 97–100. URL https://www.aclweb.org/anthology/W04-2319. 116

SRIVASTAVA, N., G. HINTON, A. KRIZHEVSKY, I. SUTSKEVER and R. SALAKHUTDINOV. 2014, "Dropout: a simple way to prevent neural networks from overfitting", *The journal of machine learning research*, vol. 15, nº 1, p. 1929–1958. 119

STOLCKE, A., K. RIES, N. COCCARO, E. SHRIBERG, R. BATES, D. JURAFSKY, P. TAYLOR, R. MARTIN, C. V. ESS-DYKEMA and M. METEER. 2000a, "Dialogue act modeling for automatic tagging and recognition of conversational speech", *Computational linguistics*, vol. 26, nº 3, p. 339–373. 115

STOLCKE, A., K. RIES, N. COCCARO, E. SHRIBERG, R. BATES, D. JURAFSKY, P. TAYLOR, R. MARTIN, C. V. ESS-DYKEMA and M. METEER. 2000b, "Dialogue act modeling for automatic tagging and recognition of conversational speech", *Computational linguistics*, vol. 26, nº 3, p. 339–373. 119

STYMNE, S. and collab.. 2020, "Evaluating word embeddings for indonesian–english code-mixed text based on synthetic data", in *Proceedings of the The 4th Workshop on Computational Approaches to Code Switching*, p. 26–35. 115

SURENDRAN, D. and G.-A. LEVOW. 2006, "Dialog act tagging with support vector machines and hidden markov models", in *Ninth International Conference on Spoken Language Processing*. 119

TAN, S. and S. JOTY. 2021, "Code-mixing on sesame street: Dawn of the adversarial polyglots", *arXiv preprint arXiv:2103.09593*. 115

TAYLOR, W. L. 1953, ""cloze procedure": A new tool for measuring readability", *Journalism quarterly*, vol. 30, nº 4, p. 415–433. 111

VASWANI, A., N. SHAZEER, N. PARMAR, J. USZKOREIT, L. JONES, A. N. GOMEZ, Ł. KAISER and I. POLOSUKHIN. 2017, "Attention is all you need", in *Advances in neural information processing systems*, p. 5998–6008. 114, 117

WINATA, G. I., S. CAHYAWIJAYA, Z. LIU, Z. LIN, A. MADOTTO and P. FUNG. 2021, "Are multilingual models effective in code-switching?", *arXiv preprint arXiv:2103.13309*. 110, 111, 122

WOLF, T., L. DEBUT, V. SANH, J. CHAUMOND, C. DELANGUE, A. MOI, P. CISTAC, T. RAULT, R. LOUF, M. FUNTOWICZ and J. BREW. 2019, "Huggingface's transformers: State-of-the-art natural language processing", *ArXiv*, vol. abs/1910.03771. 118

WOLF, T., Q. LHOEST, P. VON PLATEN, Y. JERNITE, M. DRAME, J. PLU, J. CHAUMOND, C. DELANGUE, C. MA, A. THAKUR, S. PATIL, J. DAVISON, T. L. SCAO, V. SANH, C. XU, N. PATRY, A. MCMILLAN-MAJOR, S. BRANDEIS, S. GUGGER, F. LAGUNAS, L. DEBUT, M. FUNTOWICZ, A. MOI, S. RUSH, P. SCHMIDD, P. CISTAC, V. MUŠTAR, J. BOUDIER and A. TORDJMANN. 2020, "Datasets", *GitHub. Note: https://github.com/huggingface/datasets*, vol. 1. 110

WU, Y., M. SCHUSTER, Z. CHEN, Q. V. LE, M. NOROUZI, W. MACHEREY, M. KRIKUN, Y. CAO, Q. GAO, K. MACHEREY and collab.. 2016, "Google's neural machine translation system: Bridging the gap between human and machine translation", *arXiv preprint arXiv:1609.08144*. 119

XUE, L., N. CONSTANT, A. ROBERTS, M. KALE, R. AL-RFOU, A. SIDDHANT, A. BARUA and C. RAFFEL. 2020, "mT5: A massively multilingual pre-trained text-to-text transformer", . 110

YANG, Z., Z. DAI, Y. YANG, J. CARBONELL, R. R. SALAKHUTDINOV and Q. V. LE. 2019, "Xlnet: Generalized autoregressive pretraining for language understanding", in *Advances in neural information processing systems*, p. 5754–5764. 110

ZHANG, K. and S. BOWMAN. 2018, "Language modeling teaches you more than translation does: Lessons learned through auxiliary syntactic task analysis", in *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, p. 359–361. 121

ZHANG, T., V. KISHORE, F. WU, K. Q. WEINBERGER and Y. ARTZI. 2019a, "Bertscore: Evaluating text generation with bert", *arXiv preprint arXiv:1904.09675*. 116

ZHANG, X., F. WEI and M. ZHOU. 2019b, "Hibert: Document level pre-training of hierarchical bidirectional transformers for document summarization", *arXiv preprint arXiv:1905.06566*. 117

ZHANG, Y., Q. LI, D. SONG, P. ZHANG and P. WANG. 2019c, "Quantum-inspired interactive networks for conversational sentiment analysis.", . 117

ZHANG, Y., S. SUN, M. GALLEY, Y.-C. CHEN, C. BROCKETT, X. GAO, J. GAO, J. LIU and B. DOLAN. 2019d, "Dialogpt: Large-scale generative pre-training for conversational response generation", *arXiv preprint arXiv:1911.00536*. 110

Chapter 7

Multimodal Nature of Punctuation: Application To Emotion Recognition

Chapter 7 Abstract

In this chapter, we explore spoken dialog within a mutlimodal scenario in the scope of emotion recognition. Since humans communicate through different modalities (i.e visual, acoustic, linguistic), efforts have been made towards developing complex fusion mechanisms based on neural networks. In these works, the linguistic modality appears to be the most effective. Furthermore, when the communication modality is the text, such as in chats or in theatre plays, the writers use punctuation marks to convey emotions. Linguists have showcased the prosodic aspect of punctuation marks acknowledging their importance in human communication. However, punctuation marks are often reduced to their semantic aspect and removed in most NLP tasks. In this chapter, we investigate the role played by punctuation marks in an emotion recognition task with current neural network architectures. Results showcase their abilities to transcribe prosodic cues. We explore their role further through ablation studies and a comparison with simple multimodal models.

7.1 Introduction

Emotion recognition in spoken dialogues is an increasingly popular area of Affective Computing research. Acknowledging humans communicate through a variety of channels (i.e visual, acoustic, linguistic), multimodal systems have integrated these different unimodal representations into one synthetic representation. So far, a consequent effort has been made to develop complex architectures allowing the fusion of these modalities. However, most of the accuracy is often carried by the linguistic modality [PORIA and collab., 2018; RAHMAN and collab., 2020].
A natural way to leverage emotion in textual conversation, employed by writers in plays or by people in written chat, is the use of punctuation[1]. Linguistic studies have shown that the use of punctuation follows either a syntactic or prosodic approach [BRUTHIAUX, 1993; HALLIDAY, 1989]. MOORE [2016] distinguishes between text to be read in silence and text to be read aloud. He shows that in the latter, punctuation

[1] In this study, we focus on spoken dialogues and therefore do not consider the use of emoji

and prosody have the same function. From this perspective, when dealing with written-to-be-spoken texts or speech transcripts, punctuation can be viewed as a metalanguage that transcribes prosodic cues such as intonation and pausing into a written language.

The link between prosody and the emotion conveyed in spoken language has been well established [JUSLIN and LAUKKA, 2003; MOZZICONACCI, 2002]. Therefore, prosody has become a key component of speech emotion recognition or multimodal emotion recognition systems [BAGHER ZADEH and collab., 2018; BALTRUŠAITIS and collab., 2019; PORIA and collab., 2018; ZADEH and collab., 2016]. When considering only the textual modality, changing the punctuation sequence of an utterance can drastically alter the prosody related to it and in turn the perceived emotion. For example, the sentence "Hello, how are you today ?" can be read as neutral while adding exclamation marks as in "Hello ! how are you today ? !" is more dynamic and enthusiastic, thus conveying the utterance of a joyful tone. If humans are sensitive to such phenomena, studies on how they may affect deep learning models are still lacking. Whereas punctuation marks have been used as hand crafted features in sentiment analysis or sarcasm recognition [AGRAWAL and AN, 2014; RUBIN and collab., 2016], in most NLP tasks they are removed during pre-processing, as NLP models achieve higher accuracy on syntactically correct utterances [BAGHER ZADEH and collab., 2018; DELBROUCK and collab., 2020; GU and collab., 2018; KRATZWALD and collab., 2018; RAHMAN and collab., 2020]. Moreover a recent work [EK and collab., 2020] shows that in Natural Language Inference (NLI) neither BERT [DEVLIN and collab., 2019] nor RNN-based models take into account relevant changes in punctuation marks, but BERT seems robust to irrelevant changes. However, this study mixes data from different genres, *i.e* written and spoken English and focuses on the semantic aspect of punctuation. To the best of our knowledge, there is no quantitative study taking into account the influence of punctuation marks on emotion recognition task for spoken data. In Automatic Speech Recognition (ASR), efforts have been made toward punctuation prediction [CHO and collab., 2017; COURTLAND and collab., 2020; LU and NG, 2010]. However, these works are mostly motivated by text readability, or downstream tasks such as part-of-speech (POS) tagging or Natural Language Translation — which focus on the semantic side of punctuation. Hence, the impact of punctuation as a prosodic marker on modern deep learning architectures in emotion recognition tasks remains overlooked.

In this work, we investigate the ability of different state-of-the art neural architectures to leverage punctuation for emotion recognition tasks. We address the following research questions in two stages: (1) Are punctuation marks discriminating features for emotion recognition systems based on speech transcripts ? (2) Can punctuation marks be a substitute for acoustic modality ? In order to answer these two questions, we first present the data and the distribution of punctuation marks among emotion categories, as well as the investigated neural models section 7.2. Then, we compare models trained and tested w/o punctuation marks and w/o acoustic information for emotion recognition tasks and provide an in-depth analysis of the obtained results section 7.3.

7.2 Data and Models

7.2.1 Data Description

For this experimental study, we use the MELD[2] dataset [PORIA and collab., 2018], a multimodal dataset collected from the TV show *Friends*. The data consists in 13708 manually transcribed utterances (Table 7.1 provides statistics for each split), labelled in sentiment and emotion including : anger, sadness, disgust, surprise, fear, joy and neutral. In Table 7.2 we present several examples from the MELD corpus and Table 7.1 provides statistics for each split. Table 7.3 shows the punctuation marks distribution in MELD.

Datasets	# Dialogues			# Utterances		
	Train	Val	Test	Train	Val	Test
MELD	1039	114	280	9989	1109	2610

Table 7.1 – MELD description

Speakers	Utterances	Emotions
Joey	What is it ? Hey !	sadness
Rachel	Really it 's nothing . I 'm just	sadness
Joey	Rach come on , what ?	neutral
Rachel	I 've just been thinking about how my baby and I are gon na be all alone .	sadness
Joey	What are you talking about alone ? What about Ross ?	surprise
Monica	What is it ? !	surprise
Rachel	I do n't know ! But maybe if we keep that drawer shut , it 'll die .	disgust
Monica	I ca n't believe we 're living here !	sadness
The Teacher	Excellent ! What Rachel has shrewdly observed here	joy
Phoebe	You completely stole my answer !	anger
Rachel	Well , honey that was pretty obvious.	sadness
Phoebe	Well how would you know ? ! You did n't even read it !	anger

Table 7.2 – Some examples of dialogues from MELD dataset.

We choose this dataset for the following reasons: (1) due to its nature, it falls into the category of written-to-be-spoken text. This category of textual data gathers not only scripts for plays but also transcripts of spoken dialogues obtained either manually or to some extent, with Automatic Speech Recognition (ASR) models [BAGHER ZADEH and collab., 2018; BUSSO and collab., 2008; ZADEH and collab., 2016]. However, the latter suffers from the lack of punctuation marks. Since dialogues from MELD are manually transcribed, the transcribed punctuation marks are of good

[2]https://affective-meld.github.io/

Punctuations	Train	Val	Test
<.>	7779 **(6.48)**	832 **(6.29)**	2028 **(6.3)**
<,>	7244 **(6.04)**	790 **(5.97)**	2038 **(6.34)**
<?>	2887 **(2.41)**	312 **(2.36)**	769 **(2.39)**
<!>	4266 **(3.56)**	524 **(3.96)**	1090 **(3.39)**

Table 7.3 – Punctuation marks distribution over MELD splits. Bold number in parenthesis indicates the percentage of tokens.

quality. (2) When comparing to Ek and collab. [2020], which considers all non-alphanumeric characters, we choose to consider a limited number of punctuation marks. Indeed, not all punctuation marks are prosodic: for example, it has been demonstrated that adults and children show some pitch declination at commas, regardless of their reading ability [Kuhn and collab., 2010], while exclamation points and quotation marks appear to signal a rise in pitch [Schwanenflugel and collab., 2013]. Brackets, on the other hand, do not appear to have a clear relationship with expressive reading [Bodenbender, 1999]. Hence, we only consider the following punctuation marks : '<!> <?> <.> <,> <...>'. (3) Table 7.4 shows the proportion of utterances containing specific punctuation marks for each emotion. As expected, some emotions are more closely related to the use of certain punctuation marks. For instance '<! ! !>' is present in 36% and 33% of sentences labelled with respectively `joy` and `anger`; where '<! ?>' is present in 75% of sentences labelled with `surprise`. Alternatively the dot and comma appear in more than half the sentences labelled as `neutral`. Those observations are in line with the relationship we have previously exposed between the use of some punctuation marks and expressed emotions. (3) Finally the MELD corpus provides acoustic features allowing us to compare the use of prosodic punctuation marks with fusion between the textual and audio modalities.

Punctuations	Emotions						
	neutral	surprise	joy	sadness	anger	disgust	fear
<.>	**59.79**	5.04	14.39	8.5	6.56	3.01	2.71
<!>	10.98	20.71	**34.11**	5.39	23.02	2.94	2.84
<,>	**51.5**	7.35	16.44	8.36	10.2	3.06	3.09
<...>	**46.2**	11.71	14.56	12.97	7.91	3.16	3.48
<?>	**44.5**	26.48	8.92	4.75	10.58	2.16	2.62
<! !>	3.68	22.11	28.42	1.58	**32.63**	6.84	4.74
<! ! !>	3.03	**36.36**	15.15	0.0	33.33	6.06	6.06
<! ?>	0.0	**75.0**	0.0	0.0	25.0	0.0	0.0
<? !>	4.74	**56.2**	6.2	3.65	23.72	3.65	1.82
<? ?>	28.57	**42.86**	0.0	0.0	28.57	0.0	0.0

Table 7.4 – Distribution of sentences containing punctuation marks among emotion classes.

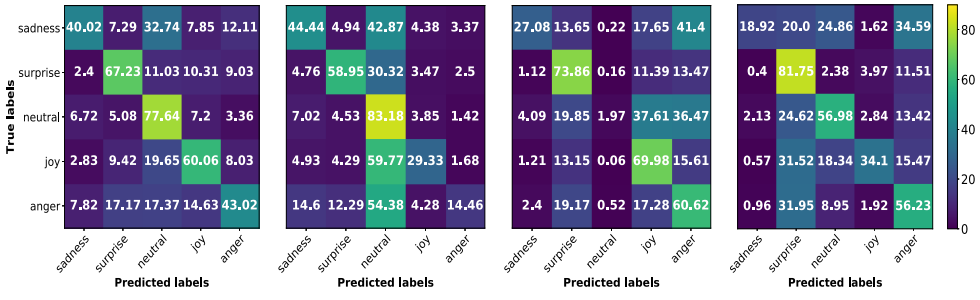
7.2.2 Baselines

We present here the baselines we selected for our experiments, as they are the most frequently used to extract context-independent utterance-level feature vectors in emotion recognition tasks.

Text-CNN In this framework introduced by Kim [2014], the embeddings of the words

composing the utterances are fed into a convolutional neural network (CNN). We denote by CNN_R and CNN_G the models corresponding to words embeddings being set randomly and with Glove embedding [PENNINGTON and collab., 2014] respectively.
Transformer-based architectures They outperform classical architectures relying on word embeddings such as glove. We use BERT and RoBERTa [LIU and collab., 2019] as they achieve competitive results and are used as sentence embedding extractors for state-of-the-art methods [GHOSAL and collab., 2020; LI and collab., 2020]. Sentence embeddings are obtained through the representation of the [CLS] token provided by the last layer.
For all our baseline classification models, the sentence embedding is fed to a multilayer perceptron, trained with dropout [SRIVASTAVA and collab., 2014]. We do not use the contextual information nor any additional information from the other modalities in order to identify the emotion or sentiment of an utterance. To obtain multimodal representations, we concatenate the sentence representations given by the aforementioned models with audio features provided with the MELD dataset.
Training Details We use dropout [SRIVASTAVA and collab., 2014] and optimise the global loss by gradient descent using AdamW [KINGMA and BA, 2015; LOSHCHILOV and HUTTER, 2017] optimiser with 4000 warmups steps for transformers based model [VASWANI and collab., 2017]. Transformers based model are trained during 7 epoch with a batch of 52 with a learning rate 1e-3 and CNN based models are trained during 100 epochs with a batch size of 256. The best learning rate is found in the grid 0.002,0.001,0.0005,0.0001. All models are trained on 1 NVIDIA V100. For Transformers based model and CNN_G we rely on weights provided by WOLF and collab. [2019] and PORIA and collab. [2018] respectively.
MELD corpus is lower cased at training and testing time. We use tokenizer introduced by DEVLIN and collab. [2019] and used the SentencePiece algorithm [KUDO and RICHARDSON, 2018].

7.3 Experiments and Analysis



(a) (b) (c) (d)

Figure 7.1 – Confusion matrix visualisation of BERT on different version of the MELD test set for the 5 most represented classes. From left to right, (a) no modification is applied, (b) <!> is removed (c) each utterance is appended with an exclamation mark <!> and (d) each utterance is appended with the punctuation mark <?>.

We introduce a variant of MELD where we remove all punctuation marks, and that we note $\neg p$. Similarly, we denote by p the original dataset (keeping all punctuation marks). We then define four settings for our experiments: $\mathcal{S}_{p\rightarrow p}$, $\mathcal{S}_{\neg p\rightarrow \neg p}$, $\mathcal{S}_{p\rightarrow \neg p}$ and $\mathcal{S}_{\neg p\rightarrow p}$. The right part of the arrow indicates the variant of MELD used as training set, while the left part indicates the variant used as test set. $\mathcal{S}_{p\rightarrow p}$ and $\mathcal{S}_{\neg p\rightarrow \neg p}$ allow

us to test the importance of punctuation as a discriminative feature in emotion recognition tasks. With the $\mathcal{S}_{p\rightarrow\neg p}$ experiment, we want to show the impact of punctuation removal at testing time, when the model has been trained with it. $\mathcal{S}_{\neg p\rightarrow p}$ acts as a sanity check experiment: if a model is not trained to associate punctuation marks to particular emotion labels, we expect it to ignore these characters at testing time. Experiments are first conducted with only textual features, then with both textual and audio features.

7.3.1 Is punctuation a discriminative feature ?

We present our results in Table 7.6. When comparing $\mathcal{S}_{p\rightarrow p}$ with $\mathcal{S}_{\neg p\rightarrow\neg p}$, we observe that every model, trained on textual modality only, achieves a better performance when punctuation marks are kept in the dataset: up to 9 points for BERT and RoBERTa. Moreover, results with the $\mathcal{S}_{p\rightarrow\neg p}$ setting show the extent to which models rely on these features to make predictions leading to a huge drop in performance when they are removed at inference time: from 10 points for transformer-based models to 15 for CNN_R model. Table 7.5 gives performance details for each emotion class. We observe an increase of the `neutral` class accuracy score for all models with the $\mathcal{S}_{p\rightarrow\neg p}$ setting: by removing punctuation marks, classifiers lose precious prosodic information and consider such altered utterances as `neutral`. Keeping punctuation marks allows the models to efficiently leverage the ambiguity of utterances labelled as `surprise`, `joy` and `anger`.
The setting $\mathcal{S}_{\neg p\rightarrow p}$ does not show major differences for models based on convolutional network. However, pre-trained transformer models show an increase of 3 points of the F1 Weighted score, stemming from better results on the `joy`, `anger` and `surprise` labels.

Models	Setting	Emotions							
		sadness	surprise	neutral	joy	anger	disgust	fear	W-Avg F1
CNN_R	$\mathcal{S}_{p\rightarrow p}$	14.11	53.26	85.35	52.02	32.66	5.00	1.23	57.04
	$\mathcal{S}_{\neg p\rightarrow\neg p}$	15.80	49.38	75.67	37.82	16.54	6.30	3.95	47.72
	$\mathcal{S}_{p\rightarrow\neg p}$	16.82	16.14	93.74	17.13	3.02	6.48	4.20	42.00
CNN_G	$\mathcal{S}_{p\rightarrow p}$	25.83	53.70	77.29	56.00	37.38	8.52	0.25	57.33
	$\mathcal{S}_{\neg p\rightarrow\neg p}$	20.48	48.19	70.61	40.46	23.54	10.19	2.72	48.39
	$\mathcal{S}_{p\rightarrow\neg p}$	31.35	15.43	86.56	24.16	6.07	18.33	2.22	44.23
BERT	$\mathcal{S}_{p\rightarrow p}$	37.43	63.59	76.58	60.17	51.12	0.00	1.67	61.70
	$\mathcal{S}_{\neg p\rightarrow\neg p}$	31.95	60.01	67.18	49.79	35.82	1.11	2.22	52.69
	$\mathcal{S}_{p\rightarrow\neg p}$	42.70	26.69	85.31	30.66	20.93	0.42	3.33	50.47
	$\mathcal{S}_{\neg p\rightarrow p}$	27.21	77.73	64.89	58.29	42.35	0.19	0.74	55.71
RoBERTa	$\mathcal{S}_{p\rightarrow p}$	38.59	66.75	76.84	59.54	41.15	11.00	10.67	61.81
	$\mathcal{S}_{\neg p\rightarrow\neg p}$	36.86	56.98	71.56	40.06	35.27	7.67	12.44	53.73
	$\mathcal{S}_{p\rightarrow\neg p}$	42.49	29.60	85.90	31.00	12.14	14.67	9.78	50.38
	$\mathcal{S}_{\neg p\rightarrow p}$	36.32	56.83	72.28	54.50	39.74	8.00	14.22	57.40

Table 7.5 – Baselines results on the test set of `MELD` using only textual modality. W-avg F1 denotes the weighted average of F1 score. Results have been averaged over 5 runs.

7.3.2 Ablation Study on Punctuation Marks

In this section, we investigate the role of each punctuation mark in emotion recognition tasks. Each model is trained with the original training set. At testing time, we remove one mark or a specific sequence of marks at a time. Results for BERT are presented in Table 7.7, we obtain similar results for other baselines. They show

Models	Setting	Modality	
		T	T + A
CNN_R	$\mathcal{S}_{p\rightarrow p}$	**57.04 ± 0.49**	**57.75 ± 0.17**
	$\mathcal{S}_{\neg p\rightarrow \neg p}$	47.72 ± 0.56	49.82 ± 0.52
	$\mathcal{S}_{p\rightarrow \neg p}$	42.00 ± 0.64	42.93 ± 0.59
CNN_G	$\mathcal{S}_{p\rightarrow p}$	**57.33 ± 0.53**	**57.88 ± 0.21**
	$\mathcal{S}_{\neg p\rightarrow \neg p}$	48.39 ± 0.29	50.74 ± 0.35
	$\mathcal{S}_{p\rightarrow \neg p}$	44.23 ± 0.82	46.93 ± 0.17
BERT	$\mathcal{S}_{p\rightarrow p}$	**61.70 ± 0.51**	**62.25 ± 0.30**
	$\mathcal{S}_{\neg p\rightarrow \neg p}$	52.69 ± 0.35	55.84 ± 0.21
	$\mathcal{S}_{p\rightarrow \neg p}$	50.47 ± 0.71	51.20 ± 0.58
RoBERTa	$\mathcal{S}_{p\rightarrow p}$	**61.81 ± 0.84**	**62.86 ± 0.31**
	$\mathcal{S}_{\neg p\rightarrow \neg p}$	53.73 ± 1.01	55.04 ± 0.17
	$\mathcal{S}_{p\rightarrow \neg p}$	50.38 ± 1.71	52.81 ± 0.27

Table 7.6 – Baseline results (weighted average of F1 score) on the test set of MELD using both textual (T) and audio (A) modalities. Results have been averaged over 5 runs.

that <!> and <?> are the most discriminating features, with <!> being more related to anger or joy and <?> to surprise. On the other hand, dot and comma seem to have no effect on the prediction, as their use follows both a syntactic and prosodic approach leading to poor improvement. This section presents the ablation study of punctuation marks for BERT. We obtain similar results for other baselines. In order to showcase the influence of each punctuation mark on the prediction we train BERT on the original MELD (*i.e* the p version) and remove one punctuation at a time at inference. Results are presented in Table 7.7.

Punctuation Marks	Emotions						
	sadness	surprise	neutral	joy	anger	disgust	fear
<.>	1.55	-2.80	-0.62	1.86	0.95	0.00	0.00
<,>	-1.77	-2.67	-3.48	4.40	-1.45	0.00	-4.17
<?>	3.23	-41.33	19.27	-2.04	-6.41	0.00	0.00
<!>	17.07	-7.48	68.42	-50.70	-43.75	0.00	0.00
<! !>	100.00	0.00	0.00	-37.50	-7.14	0.00	0.00
<? !>	0.00	-56.82	100.00	0.00	-33.33	0.00	0.00

Table 7.7 – Ablation study results for BERT on MELD test set. Each punctuation mark is removed one at a time during inference. Presented results are the difference of accuracy after removal and without modification. Results are obtained considering only utterance containing the punctuation marks to be removed. Results have been averaged over 5 runs.

To highlight the impact that some punctuation marks may have on predictions, we systematically append utterances of the original corpus with the most discriminating punctuation marks. Results for BERT are presented in Figure 7.1. Overall, this completely polarizes neutral utterances between surprise/joy/anger or surprise/anger, when <!> or <?> are used respectively. This behavior is expected, as adding punctuation marks will obviously polarize neutral sentences for human readers too: 'I don't know' is perceived as neutral but 'I don't know !' would be perceived as anger. However, we notice that punctuation marks fail at transcribing prosodic

cues related to sadness. Such a behaviour can be misleading: for example, the sentence "I am so sad" is correctly labelled by BERT as sadness however "I am so sad !" would be classified as joy.

7.3.3 Punctuation as substitute for acoustic modality ?

Table 7.6 shows that for every setting, audio information improves the performance of the models. This is especially the case for models without punctuation: by comparing $\mathcal{S}^{T+A}_{\neg p \rightarrow \neg p}$ with $\mathcal{S}^{T}_{\neg p \rightarrow \neg p}$, we notice an average increase of about 2 points. However, as previously mentioned the addition of punctuation marks alone, *i.e* $\mathcal{S}^{T}_{p \rightarrow p}$, improves the result up to 9 points. This seems to confirm the greater efficiency of punctuation marks to transmit prosodic cues. While our experiments were made with a simple fusion mechanism, similar observations have previously been made RAHMAN and collab. [2020] regarding the low performance improvement brought by other modalities. However in Table 7.8 we notice that $\mathcal{S}^{T+A}_{p \rightarrow p}$ models achieve better performances on disgust and fear than models with $\mathcal{S}^{T}_{p \rightarrow p}$ setting.

Models	Setting	Emotions							
		sadness	surprise	neutral	joy	anger	disgust	fear	W-Avg F1
CNN_R	$\mathcal{S}_{p \rightarrow p}$	16.00	56.27	82.93	50.83	41.92	0.00	0.00	57.75
	$\mathcal{S}_{\neg p \rightarrow \neg p}$	20.76	51.35	70.70	36.96	37.57	0.00	0.00	49.82
	$\mathcal{S}_{p \rightarrow \neg p}$	19.46	15.56	90.95	18.97	8.31	3.00	0.44	42.93
CNN_G	$\mathcal{S}_{p \rightarrow p}$	25.73	56.98	76.12	52.78	44.35	8.33	0.00	57.88
	$\mathcal{S}_{\neg p \rightarrow \neg p}$	23.14	48.57	68.89	43.04	40.51	1.33	0.00	50.74
	$\mathcal{S}_{p \rightarrow \neg p}$	28.32	16.43	85.35	25.44	19.11	19.67	1.33	46.93
BERT	$\mathcal{S}_{p \rightarrow p}$	35.35	65.71	74.33	58.74	52.97	9.33	12.00	62.25
	$\mathcal{S}_{\neg p \rightarrow \neg p}$	32.65	58.17	67.43	52.72	48.69	5.00	10.67	55.84
	$\mathcal{S}_{p \rightarrow \neg p}$	34.92	27.78	80.87	31.40	25.69	11.00	20.00	51.20
RoBERTa	$\mathcal{S}_{p \rightarrow p}$	32.22	62.22	77.44	62.52	49.27	15.33	9.78	62.86
	$\mathcal{S}_{\neg p \rightarrow \neg p}$	38.59	44.68	78.20	37.77	32.52	16.00	22.67	55.04
	$\mathcal{S}_{p \rightarrow \neg p}$	36.86	30.24	85.60	30.26	24.79	19.00	15.11	52.81

Table 7.8 – Baselines results on the test set of MELD using only both textual and acoustic modality. W-avg F1 denotes the weighted average of F1 score. Results have been averaged over 5 runs.

Chapter 7 Conclusion

This chapter is an preliminary study that aims to investigate ways to translate non-textual modality into textual modality. In particular, we focused on punctuations marks and we provided a quantitative analysis of their impact on emotion recognition. We showcase the ability of punctuation marks to encode some prosodic cues in textual modality, boosting classifiers performances especially with pre-trained models. Furthermore we showed how punctuation marks may be limited, as they seem to perform very well when encoding expressive tones related to anger, joy or surprise but fail at transcribing acoustic cues related to sadness or disgust.

[2]However this observation should not be over interpreted as the MELD dataset does not gather enough examples for disgust and fear

7.4 References

AGRAWAL, A. and A. AN. 2014, "Kea: Sentiment analysis of phrases within short texts", in **SEMEVAL*. 134

BAGHER ZADEH, A., P. P. LIANG, S. PORIA, E. CAMBRIA and L.-P. MORENCY. 2018, "Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph", in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, Melbourne, Australia, p. 2236–2246, doi: 10.18653/v1/P18-1208. URL https://www.aclweb.org/anthology/P18-1208. 134, 135

BALTRUŠAITIS, T., C. AHUJA and L.-P. MORENCY. 2019, "Multimodal machine learning: A survey and taxonomy", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, nº 2, doi: 10.1109/TPAMI.2018.2798607, p. 423–443. 134

BODENBENDER, C. 1999, "The punctuation and intonation of parentheticals", URL http://hdl.handle.net/1828/2768. 136

BRUTHIAUX, P. 1993, "Knowing when to stop: Investigating the nature of punctuation", *Language & Communication*, vol. 13, nº 1, doi: https://doi.org/10.1016/0271-5309(93)90019-J, p. 27–43, ISSN 0271-5309. URL https://www.sciencedirect.com/science/article/pii/027153099390019J. 133

BUSSO, C., M. BULUT, C.-C. LEE, A. KAZEMZADEH, E. MOWER, S. KIM, J. N. CHANG, S. LEE and S. S. NARAYANAN. 2008, "IEMOCAP: interactive emotional dyadic motion capture database", *Language Resources and Evaluation*, vol. 42, nº 4, p. 335–359. 135

CHO, E., J. NIEHUES and A. WAIBEL. 2017, "Nmt-based segmentation and punctuation insertion for real-time spoken language translation", in *Proc. Interspeech 2017*, p. 2645–2649, doi: 10.21437/Interspeech.2017-1320. URL http://dx.doi.org/10.21437/Interspeech.2017-1320. 134

COURTLAND, M., A. FAULKNER and G. MCELVAIN. 2020, "Efficient automatic punctuation restoration using bidirectional transformers with robust inference", in *Proceedings of the 17th International Conference on Spoken Language Translation*, Association for Computational Linguistics, Online, p. 272–279, doi: 10.18653/v1/2020.iwslt-1.33. URL https://www.aclweb.org/anthology/2020.iwslt-1.33. 134

DELBROUCK, J.-B., N. TITS, M. BROUSMICHE and S. DUPONT. 2020, "A transformer-based joint-encoding for emotion recognition and sentiment analysis", *Second Grand-Challenge and Workshop on Multimodal Language (Challenge-HML)*, doi: 10.18653/v1/2020.challengehml-1.1. URL http://dx.doi.org/10.18653/v1/2020.challengehml-1.1. 134

DEVLIN, J., M.-W. CHANG, K. LEE and K. TOUTANOVA. 2019, "BERT: Pre-training of deep bidirectional transformers for language understanding", in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Association for Computational Linguistics, Minneapolis, Minnesota, p. 4171–

4186, doi: 10.18653/v1/N19-1423. URL https://www.aclweb.org/anthology/N19-1423. 134, 137

EK, A., J.-P. BERNARDY and S. CHATZIKYRIAKIDIS. 2020, "How does punctuation affect neural models in natural language inference", in *Proceedings of the Probability and Meaning Conference (PaM 2020)*, Association for Computational Linguistics, Gothenburg, p. 109–116. URL https://www.aclweb.org/anthology/2020.pam-1.15. 134, 136

GHOSAL, D., N. MAJUMDER, A. GELBUKH, R. MIHALCEA and S. PORIA. 2020, "COSMIC: COmmonSense knowledge for eMotion identification in conversations", in *Findings of the Association for Computational Linguistics: EMNLP 2020*, Association for Computational Linguistics, Online, p. 2470–2481, doi: 10.18653/v1/2020.findings-emnlp.224. URL https://www.aclweb.org/anthology/2020.findings-emnlp.224. 137

GU, Y., K. YANG, S. FU, S. CHEN, X. LI and I. MARSIC. 2018, "Hybrid attention based multimodal network for spoken language classification", in *Proceedings of the 27th International Conference on Computational Linguistics*, Association for Computational Linguistics, Santa Fe, New Mexico, USA, p. 2379–2390. URL https://www.aclweb.org/anthology/C18-1201. 134

HALLIDAY, M. 1989, *Spoken and Written Language*, Language and Learning Series, Deakin University, ISBN 9780730003090. URL https://books.google.fr/books?id=T9RpAAAACAAJ. 133

JUSLIN, P. N. and P. LAUKKA. 2003, "Communication of emotions in vocal expression and music performance: Different channels, same code?", *Psychological Bulletin*, vol. 129, nº 5, doi: 10.1037/0033-2909.129.5.770, p. 770–814. URL https://doi.org/10.1037/0033-2909.129.5.770. 134

KIM, Y. 2014, "Convolutional neural networks for sentence classification", *CoRR*, vol. abs/1408.5882. URL http://arxiv.org/abs/1408.5882. 136

KINGMA, D. P. and J. BA. 2015, "Adam: A method for stochastic optimization", in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, édité par Y. Bengio and Y. LeCun. URL http://arxiv.org/abs/1412.6980. 137

KRATZWALD, B., S. ILIC, M. KRAUS, S. FEUERRIEGEL and H. PRENDINGER. 2018, "Decision support with text-based emotion recognition: Deep learning for affective computing", *CoRR*, vol. abs/1803.06397. URL http://arxiv.org/abs/1803.06397. 134

KUDO, T. and J. RICHARDSON. 2018, "SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing", in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Association for Computational Linguistics, Brussels, Belgium, p. 66–71, doi: 10.18653/v1/D18-2012. URL https://www.aclweb.org/anthology/D18-2012. 137

KUHN, M., P. SCHWANENFLUGEL and E. MEISINGER. 2010, "Aligning theory and assessment of reading fluency: Automaticity, prosody, and definitions of fluency", *Reading Research Quarterly*, vol. 45, p. 232–253. 136

LI, J., D. JI, F. LI, M. ZHANG and Y. LIU. 2020, "HiTrans: A transformer-based context- and speaker-sensitive model for emotion detection in conversations", in *Proceedings of the 28th International Conference on Computational Linguistics*, International Committee on Computational Linguistics, Barcelona, Spain (Online), p. 4190–4200, doi: 10.18653/v1/2020.coling-main.370. URL https://www.aclweb.org/anthology/2020.coling-main.370. 137

LIU, Y., M. OTT, N. GOYAL, J. DU, M. JOSHI, D. CHEN, O. LEVY, M. LEWIS, L. ZETTLEMOYER and V. STOYANOV. 2019, "Roberta: A robustly optimized BERT pretraining approach", *CoRR*, vol. abs/1907.11692. URL http://arxiv.org/abs/1907.11692. 137

LOSHCHILOV, I. and F. HUTTER. 2017, "Fixing weight decay regularization in adam", *CoRR*, vol. abs/1711.05101. URL http://arxiv.org/abs/1711.05101. 137

LU, W. and H. T. NG. 2010, "Better punctuation prediction with dynamic conditional random fields", in *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Cambridge, MA, p. 177–186. URL https://www.aclweb.org/anthology/D10-1018. 134

MOORE, N. 2016, "What's the point? the role of punctuation in realising information structure in written english", *Functional Linguistics*, vol. 3, doi: 10.1186/s40554-016-0029-x. 133

MOZZICONACCI, S. 2002, "Prosody and emotions", . 134

PENNINGTON, J., R. SOCHER and C. MANNING. 2014, "GloVe: Global vectors for word representation", in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Association for Computational Linguistics, Doha, Qatar, p. 1532–1543, doi: 10.3115/v1/D14-1162. URL https://www.aclweb.org/anthology/D14-1162. 137

PORIA, S., D. HAZARIKA, N. MAJUMDER, G. NAIK, E. CAMBRIA and R. MIHALCEA. 2018, "MELD: A multimodal multi-party dataset for emotion recognition in conversations", *CoRR*, vol. abs/1810.02508. URL http://arxiv.org/abs/1810.02508. 133, 134, 135, 137

RAHMAN, W., M. K. HASAN, S. LEE, A. BAGHER ZADEH, C. MAO, L.-P. MORENCY and E. HOQUE. 2020, "Integrating multimodal information in large pretrained transformers", in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Online, p. 2359–2369, doi: 10.18653/v1/2020.acl-main.214. URL https://www.aclweb.org/anthology/2020.acl-main.214. 133, 134, 140

RUBIN, V., N. CONROY, Y. CHEN and S. CORNWELL. 2016, "Fake news or truth? using satirical cues to detect potentially misleading news", in *Proceedings of the Second Workshop on Computational Approaches to Deception Detection*, Association for Computational Linguistics, San Diego, California, p. 7–17, doi:

10.18653/v1/W16-0802. URL https://www.aclweb.org/anthology/W16-0802. 134

SCHWANENFLUGEL, P. J., M. R. WESTMORELAND and R. G. BENJAMIN. 2013, "Reading fluency skill and the prosodic marking of linguistic focus", *Reading and Writing*, vol. 28, nº 1, doi: 10.1007/s11145-013-9456-1, p. 9–30. URL https://doi.org/10.1007/s11145-013-9456-1. 136

SRIVASTAVA, N., G. HINTON, A. KRIZHEVSKY, I. SUTSKEVER and R. SALAKHUTDINOV. 2014, "Dropout: A simple way to prevent neural networks from overfitting", *Journal of Machine Learning Research*, vol. 15, nº 56, p. 1929–1958. URL http://jmlr.org/papers/v15/srivastava14a.html. 137

VASWANI, A., N. SHAZEER, N. PARMAR, J. USZKOREIT, L. JONES, A. N. GOMEZ, L. KAISER and I. POLOSUKHIN. 2017, "Attention is all you need", *CoRR*, vol. abs/1706.03762. URL http://arxiv.org/abs/1706.03762. 137

WOLF, T., L. DEBUT, V. SANH, J. CHAUMOND, C. DELANGUE, A. MOI, P. CISTAC, T. RAULT, R. LOUF, M. FUNTOWICZ and J. BREW. 2019, "Huggingface's transformers: State-of-the-art natural language processing", *CoRR*, vol. abs/1910.03771. URL http://arxiv.org/abs/1910.03771. 137

ZADEH, A., R. ZELLERS, E. PINCUS and L.-P. MORENCY. 2016, "Mosi: Multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos", . 134, 135

Chapter 8

Conclusions, Limitations and Future Work

8.1 Conclusions

Throughout this thesis, we have presented different neural methods for spoken dialogue understanding. A particular emphasis has been placed on DA and E/S classification. We explored and developed new methods in various scenarios *i.e.* monolingual, lack of annotated data, multi-lingual and multimodal.
Part II is related to RQ1. In this part, we focused on monolingual conversational data (in English) and addressed the problem of sequence labelling. In its formulation (see Chapter 3), this problem is very similar to NMT problems. Thus, our first contribution presented in Chapter 4 leveraged these similarities, as we described a dialogue as a sequence of utterances to be mapped into a sequence of tags. This allowed us to apply NMT methods and architectures such as Seq2seq models widely used in this field. Compared to previous works, this approach gets rid of the currently used CRF and handcrafted features and allows for an end-to-end approach relying solely on neural networks. We proposed an encoder tailored to the hierarchical nature of dialogue and a novel attention guided mechanism. Moreover, the small size of the output space allowed us to apply a beam search during both training and inference. We showcase new SOTA results on SwDA and competitive results on MRDA.
However, in this approach, most of the network is trained from scratch which prevents us from using small annotated corpora as it is often the case with datasets labelled in E/S. Meanwhile, architectures pre-trained on large corpora (*e.g.* BERT, RoBERTa, GPT) have proved to be very effective for transferring generic representations of text to low-resource tasks. Nevertheless, these representations are pretained on written datasets and have been limited to the sentence level. In Chapter 5, we addressed these limitations and proposed a pre-trained model tailored to spoken dialogues. Our solution involves a hierarchical transformer which is pre-trained on OpenSubtitles, a large collection of spoken dialogue transcripts from movies and series. We adapted two well-known pre-training objectives to the encoder hierarchy. As the evaluation of such representations is challenging, we collected a new benchmark called SILICONE which gathers 10 datasets of different sizes that are annotated in DA and E/S. We demonstrated that our approach is suited for small and medium-size datasets and allows us for a reduction of the number of parameters while achieving better performances.
Part III is related to RQ2 and we extend the work presented in Part II in two different

scenarios: multilingual and multimodal. Research on NLU and in particular on DA or E/S classification is very English-centric, as an example Part II exclusively deals with English data, whereas, a dialogue system may be exposed to languages other than English or multilingual conversations. In Chapter 6, we proposed to extend the work presented in Chapter 5 to a multilingual setting. Based on the architectures previously introduced, we presented novel losses inspired by the CS phenomena that may occur within a speaker statement or in a conversation. We leverage the OpenSubtitles alignment files that are available in a large variety of languages to forge multilingual conversational contexts used to enrich the pre-training corpus with code-switched conversation. We evaluated our approach on a newly introduced multilingual benchmark called MIAM gathering DA annotated datasets in 5 European languages. Additionally, we test the learnt multilingual representations on two new tasks: contextual inconsistency detection and next utterance retrieval with both monolingual and multilingual input context. Finally, in Chapter 7, we explored the multimodal aspect of human communication with a focus on emotion recognition. In a spoken conversation, emotions are conveyed via a variety of channels such as visual, acoustic and linguistic. Current multimodal systems aim at integrating these different unimodal representations into one synthetic representation via complex fusion mechanisms. On the other hand, information in textual modality is well leveraged by neural architectures and has been shown as the most dominant modality regarding the system's performances. Hence, we aim at investigating the translation of non-textual modalities, i.e. visual and acoustic, into the textual modality. This last chapter is a preliminary study that focuses on punctuation marks as they are a simple and easily accessible textual representations of prosodical cues. We provided a quantitative study that showcases how effectively acoustic cues can be cram into punctuation marks boosting the performance of the predictor, especially with pre-trained models.

8.2 Limitations and future work

In this section, we put our contributions into perspective, discuss some of the limitations of our work as well as the numerous perspectives for the different parts of my work

8.2.1 Limitations and Future Directions Related to RQ1

In RQ1, we explored the problem of sequence labelling in the monolingual scenario, thus Part II opens the following research directions:

- **Including the interaction dissymmetry and speaker information in our models.** In Part II, we put a lot of effort into providing distributed representations of the dialogue context solely based on the textual content. We have neglected some aspects of the interaction such as the speaker information. Such information covers a wide range of phenomena from-turn taking [Liu and collab., 2020; Shang and collab., 2020], intra-speaker and inter-speaker dependencies [Shen and collab., 2020] to personality traits [Li and collab., 2016; Zhong and collab., 2020]. We believe they are key information to better represent the dynamic of the conversation. In Chapter 5, we have provided a hierarchical transformer to encode the conversation. In particular, the dialogue level of this

encoder is trained to learn utterances dependencies. In future work, we aim at modifying this part of the network in order to include self- and inter-speakers dependencies.

- **Explore the interplay between dialogue act and emotion** In this thesis we have addressed DA and E/S classification as two separate tasks. However, as conversations are often driven by emotion there is a mutual dependency between DA and E/S [LI and collab., 2017]. BOTHE and collab. [2020] have highlighted this relationship, for instance, the *Accept/Agree* dialogue act is more likely to occurs when the utterance is labelled with a *Joy* emotion. Future work includes the study and modelling of this dependence via structured prediction. Furthermore, recently conversational corpus annotated in both DA and E/S have been realised [BOTHE and collab., 2020], following DailyDialog [LI and collab., 2017]. We plan to include these new datasets in our SILICONE benchmark.

- **Text generation tailored for dialogue** Great advances have been made in natural language generation with large scale pre-training transformers based architectures such as [RADFORD and collab., 2019; **?**]. They have shown the capacity to produce rich, lexically diverse and fluent content similar to text written by humans. As we showed throughout this manuscript a spoken conversation has features that distinguish it from a written text. Recently, in the scope of open-dialogue systems, ZHANG and collab. [2019] have taken a step in that direction. It would be interesting to pursue the work established in Chapter 5 toward the generation of dialogues. Future work could also include style transfer following COLOMBO and collab. [2019]; YANG and collab. [2020].

- **Studying domain adaptation.** In this thesis, we have provided neural models for spoken dialogue understanding. In Chapter 5 we have justified the choice of OpenSubtitles Corpus as a pre-training dataset with the discrepancy between the spoken nature of the downstream tasks and the usual pre-training corpus from written text. We have presented experiences that showcase this domain-shit and its impact on the system's performances. However, we did not provide a quantitative analysis of this discrepancy. For example, ELSAHAR and GALLÉ [2019] studies the prediction of the performance drop when a model is evaluated on a new target domain. Three category of measures are considered to evaluate the distance between domains: H-divergence, reverse classification accuracy and confidence measures. These measures rely on parameterised classifiers that need to be learnt. Meanwhile, pre-trained language models are ubiquitous and provide an underused distributional representation of sentences. Hence, in future work, we aim at investigating untrained measures based on distribution distances.

8.2.2 Limitations and Future Directions Related to RQ2

In RQ2, we explored spoken dialogue understanding problems that involve working with multilingual and multimodal data. Therefore, Part III opens the following research directions:

- **Enrich our models developed in Chapter 6 with new languages** In the future, we plan to further work on OPS to obtain fine-grained alignments (*e.g* at the

span and word levels) and enrich the definition of code-switching (currently limited at the utterance level). Moreover, when considering interactions with voice assistants and chatbots, users may not be able to express their intent in the language in which the voice assistant is programmed. Thus, we would like to strengthen our evaluation protocol by gathering a new DA benchmark with code-switched dialogues to improve the multilingual evaluation. Lastly, the cross-lingual few shot is a setting we also consider as future work, hence it would involve re-annotating datasets with a common annotation scheme following for example [BUNT and collab., 2019]. We would extend the MIAM corpus to new language and include E/S corpus thus forming a multilingual equivalent of SILICONE.

- **Learning multimodal pre-trained representations** In this thesis, we have addressed multimodal and conversational representations separately. In Chapter 5 and Chapter 6, we have provided textual pre-trained representations tailored for spoken dialogues. In Chapter 7 we have studied multimodal systems and showcased the importance of the textual modality due in particular to its pre-trained representations. Extending [BUGLIARELLO and collab., 2020] and [DAI and collab., 2021], it would be interesting to provide pre-trained multimodal representations tailored to spoken dialogues.

- **Toward a textual representation of modality** In Chapter 7 we showed that punctuation marks convey prosodical cues that allow us to leverage the ambiguity of utterances in emotion classification. Hence, punctuation marks are discriminative features in the scope of emotion classification, especially with pre-trained models. We believe that the performances of current textual encoders may offer a better opportunity to transmit particular (social) cues, when they are available than fusion with the acoustic or visual modality — as we have shown with punctuation. In future work, we plan to translate other modalities into a textual description. As an example, the MOSEI dataset [BAGHER ZADEH and collab., 2018] provides visual features that include facial action units. Each of these units has its own label, *e.g.* the 6th facial action unit is labelled as *Cheek raiser*. Instead of relying on the one-hot encoded representation of these labels, we propose to use their textual description. This text can now be directly concatenated to the corresponding utterance transcript or embedded through a masked language model and further used in common fusion mechanisms. This approach has several advantages: (1) modalities are placed into a common space, *i.e.* language, hence the system handles data of the same nature (2) non-textual modalities can benefit from powerful representations offered by pre-trained language models.

8.3 References

BAGHER ZADEH, A., P. P. LIANG, S. PORIA, E. CAMBRIA and L.-P. MORENCY. 2018, "Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph", in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, Melbourne, Australia, doi: 10.18653/v1/P18-1208. URL https://aclanthology.org/P18-1208. 148

BOTHE, C., C. WEBER, S. MAGG and S. WERMTER. 2020, "Eda: Enriching emotional dialogue acts using an ensemble of neural annotators", . 147

BUGLIARELLO, E., R. COTTERELL, N. OKAZAKI and D. ELLIOTT. 2020, "Multimodal pretraining unmasked: Unifying the vision and language berts", *CoRR*, vol. abs/2011.15124. URL https://arxiv.org/abs/2011.15124. 148

BUNT, H., V. PETUKHOVA, A. MALCHANAU, A. FANG and K. WIJNHOVEN. 2019, "The dialogbank: dialogues with interoperable annotations", *Language Resources and Evaluation*, vol. 53, nº 2, doi: 10.1007/s10579-018-9436-9, p. 213–249. URL https://doi.org/10.1007/s10579-018-9436-9. 148

COLOMBO, P., W. WITON, A. MODI, J. KENNEDY and M. KAPADIA. 2019, "Affect-driven dialog generation", *CoRR*, vol. abs/1904.02793. URL http://arxiv.org/abs/1904.02793. 147

DAI, W., S. CAHYAWIJAYA, Z. LIU and P. FUNG. 2021, "Multimodal end-to-end sparse model for emotion recognition", *CoRR*, vol. abs/2103.09666. URL https://arxiv.org/abs/2103.09666. 148

ELSAHAR, H. and M. GALLÉ. 2019, "To annotate or not? predicting performance drop under domain shift", in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Association for Computational Linguistics, Hong Kong, China, p. 2163–2173, doi: 10.18653/v1/D19-1222. URL https://aclanthology.org/D19-1222. 147

LI, J., M. GALLEY, C. BROCKETT, J. GAO and B. DOLAN. 2016, "A persona-based neural conversation model", *CoRR*, vol. abs/1603.06155. URL http://arxiv.org/abs/1603.06155. 146

LI, Y., H. SU, X. SHEN, W. LI, Z. CAO and S. NIU. 2017, "Dailydialog: A manually labelled multi-turn dialogue dataset", *CoRR*, vol. abs/1710.03957. URL http://arxiv.org/abs/1710.03957. 147

LIU, L., Z. ZHANG, H. ZHAO, X. ZHOU and X. ZHOU. 2020, "Filling the gap of utterance-aware and speaker-aware representation for multi-turn dialogue", *CoRR*, vol. abs/2009.06504. URL https://arxiv.org/abs/2009.06504. 146

RADFORD, A., J. WU, R. CHILD, D. LUAN, D. AMODEI and I. SUTSKEVER. 2019, "Language models are unsupervised multitask learners", . 147

SHANG, G., A. J. TIXIER, M. VAZIRGIANNIS and J. LORRÉ. 2020, "Speaker-change aware CRF for dialogue act classification", *CoRR*, vol. abs/2004.02913. URL https://arxiv.org/abs/2004.02913. 146

SHEN, W., J. CHEN, X. QUAN and Z. XIE. 2020, "Dialogxl: All-in-one xlnet for multi-party conversation emotion recognition", *CoRR*, vol. abs/2012.08695. URL https://arxiv.org/abs/2012.08695. 146

YANG, Z., W. WU, C. XU, X. LIANG, J. BAI, L. WANG, W. WANG and Z. LI. 2020, "Styledgpt: Stylized response generation with pre-trained language models", *CoRR*, vol. abs/2010.02569. URL https://arxiv.org/abs/2010.02569. 147

ZHANG, Y., S. SUN, M. GALLEY, Y. CHEN, C. BROCKETT, X. GAO, J. GAO, J. LIU and B. DOLAN. 2019, "Dialogpt: Large-scale generative pre-training for conversational response generation", *CoRR*, vol. abs/1911.00536. URL http://arxiv.org/abs/1911.00536. 147

ZHONG, P., Y. SUN, Y. LIU, C. ZHANG, H. WANG, Z. NIE and C. MIAO. 2020, "Endowing empathetic dialogue systems with personas", *CoRR*, vol. abs/2004.12316. URL https://arxiv.org/abs/2004.12316. 146

DOCTORAL SCHOOL

Titre : Méthodes neuronales pour la compréhension des dialogues parlés

Mots clés : Apprentissage Profond, Traitement du Langage Naturel, Dialogues Parlés.

Résumé : L'intelligence artificielle conversationnelle a suscité un intérêt croissant ces dernières années, tant dans la communauté des chercheurs que dans l'industrie. Des applications grand publique ont commencé à voir le jour (par exemple, Alexa d'Amazon, Home de Google, Siri d'Apple), mais les performances de ces systèmes sont encore loin d'une communication semblable à celle des humains. Par exemple, la conversation avec les systèmes susmentionnés se limite souvent à des interactions de base de type question-réponse. Parmi toutes les raisons pour lesquelles les gens communiquent, l'échange d'informations et le renforcement des liens sociaux semblent être les principales. Dans la recherche sur le dialogue, ces deux problèmes sont bien connus et abordés à l'aide de la classification des actes de dialogue et de la reconnaissance des émotions/sentiments. Ces problèmes sont d'autant plus difficiles à résoudre qu'ils concernent des dialogues parlés, contrairement aux textes écrits. Une conversation parlée est une activité complexe et collective qui possède une dynamique et une structure spécifiques. Il est donc nécessaire d'adapter les techniques de traitement et de compréhension du langage naturel qui ont été conçues pour les textes écrits car elles ne partagent pas les mêmes caractéristiques.

Cette thèse se concentre sur les méthodes de compréhension des dialogues parlés et aborde spécifiquement le problème de la classification des dialogues parlés avec un accent particulier sur les étiquettes des actes de dialogue et des émotions/sentiments. Nos contributions peuvent être divisées en deux parties : dans la première partie, nous abordons le problème de l'étiquetage automatique des dialogues parlés en anglais. Dans cette partie, nous commençons par formuler ce problème comme un problème de traduction, ce qui nous amène à proposer un modèle seq2seq pour la classification des actes de dialogue. Ensuite, notre deuxième contribution se concentre sur un scénario reposant sur de petits ensembles de données annotées et implique à la fois le pré-entraînement d'un encodeur transformateur hiérarchique et la proposition d'un nouveau benchmark pour l'évaluation. Cette première partie aborde le problème de la classification du langage parlé dans des contextes monolingues (*i.e.* anglais) et monomodaux (*i.e.* texte). Cependant, les dialogues parlés impliquent des phénomènes tels que le code-switching (lorsqu'un locuteur change de langue au cours d'une conversation) et s'appuient sur plusieurs canaux pour communiquer (par exemple, audio ou visuel).

La deuxième partie est donc consacrée à deux extensions des contributions précédentes dans deux contextes : multilingue et multimodal. Nous abordons d'abord le problème de la classification des actes de dialogue lorsque plusieurs langues sont impliquées et nous étendons donc les deux contributions précédentes à un scénario multilingue. Dans notre dernière contribution, nous explorons un scénario multimodal et nous nous concentrons sur la représentation et la fusion des modalités dans le cadre de la prédiction des émotions.

Title : Neural Methods For Spoken Dialogue Understanding

Keywords : Deep Learning, Natural Language Processing, Spoken Dialogue

Abstract :
Conversational AI has received a growing interest in recent years from both the research community and the industry. Products have started to emerge (*e.g.* Amazon's Alexa, Google's Home, Apple's Siri) but performances of such systems are still far from human-likeness communication. As an example, conversation with the aforementioned systems is often limited to basic question-response interactions. Among all the reasons why people communicate, the exchange of information and the strengthening of social bound appeared to be the main ones. In dialogue research, the two aforementioned problems are well known and addressed using dialogue act classification and emotion/sentiment recognition. Those problems are made even more challenging as they involve spoken dialogues in contrast to written text. A spoken conversation is a complex and collective activity that has a specific dynamic and structure. Thus, there is a need to adapt both natural language processing and natural language understanding techniques which have been tailored for written texts as it does not share the same characteristics.
This thesis focuses on methods for spoken dialogue understanding and specifically tackles the problem of spoken dialogues classification with a particular focus on dialogue act and emotion/sentiment labels. Our contributions can be divided into two parts : in the first part, we address the problem of automatically labelling English spoken dialogues. In this part, we start by formulating this problem as a translation problem which leads us to propose a seq2seq model for dialogue act classification. Then, our second contribution focuses on a scenario relying on small annotated datasets and involves both pre-training a hierarchical transformer encoder and proposing a new benchmark for evaluation. This first part addresses the problem of spoken language classification in monolingual (*i.e.* English) and monomodal (*i.e.* text) settings. However, spoken dialogue involves phenomena such as code-switching (when a speaker switch languages within a conversation) and relies on multiple channels to communicate (e.g. audio or visual).
Hence, the second part is dedicated to two extensions of the previous contributions in two settings : multilingual and multimodal. We first address the problem of dialogue act classification when multiple languages are involved and thus, we extend the two previous contributions to a multilingual scenario. In our last contribution, we explore a multimodal scenario and focus on the representation and fusion of modalities in the scope of emotion prediction.

Institut Polytechnique de Paris
91120 Palaiseau, France