Semantic and Discursive Representation for Natural Language Understanding

Damien Sileo

THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par l'Université Toulouse 3 - Paul Sabatier

Présentée et soutenue par

Damien SILEO

Le 18 décembre 2019

Représentations sémantiques et discursives pour la compréhension automatique du langage naturel

---

Ecole doctorale : **EDMITT - Ecole Doctorale Mathématiques, Informatique et Télécommunications de Toulouse**

Spécialité : **Informatique et Télécommunications**

Unité de recherche :
IRIT : Institut de Recherche en Informatique de Toulouse

Thèse dirigée par
Philippe MULLER et Tim VAN DE CRUYS

Jury

Mme Sophie ROSSET, Rapporteure
M. Ludovic DENOYER, Rapporteur
Mme Nathalie AUSSENAC-GILLES, Examinatrice
Mme Marta ABRUSAN, Examinatrice
M. Philippe MULLER, Directeur de thèse
M. Tim VAN-DE-CRUYS, Co-directeur de thèse

Thèse de Doctorat en Informatique

Représentations Sémantiques et Discursives pour la Compréhension Automatique du Langage Naturel

Présentée et soutenue publiquement
par Damien Sileo, le 18 décembre 2019

Composition du jury:

Sophie Rosset, Rapportrice, Directrice de Recherche CNRS (LIMSI, Université Paris-Sud)
Ludovic Denoyer, Rapporteur, Research Scientist (Facebook AI Research)
Nathalie Aussenac-Gilles, Examinatrice, Directrice de Recherche CNRS (IRIT, Université Toulouse 3)
Márta Abrusán, Examinatrice, Chargée de Recherche CNRS (Institut Jean Nicod, ENS Paris)
Philippe Muller, Directeur de Thèse, Maître de Conférences (IRIT, Université Toulouse 3)
Tim Van de Cruys, Co-Directeur de Thèse, Chargé de Recherche CNRS (IRIT, Université Toulouse 3)
Camille Pradel, Invité, Directeur de la Recherche et du Développement, Synapse Développement

Abstract

Computational models for automatic text understanding have gained a lot of interest due to unusual performance gains over the last few years, some of them leading to super-human scores. This success reignited some grandeur claims about artificial intelligence, such as *universal sentence representation*. In this thesis, we question these claims through two complementary angles.

Firstly, are neural networks and vector representations expressive enough to process text and perform a wide array of complex tasks? In this thesis, we will present currently used computational neural models and their training techniques. We propose a criterion for expressive compositions and show that a popular evaluation suite and sentence encoders (SentEval/InferSent) have an expressivity bottleneck; minor changes can yield new compositions that are expressive and insightful, but might not be sufficient, which may justify the paradigm shift towards newer Transformers-based models.

Secondly, we will discuss the question of universality in sentence representation: what actually lies behind these universality claims? We delineate a few theories of meaning, and in a subsequent part of this thesis, we argue that semantics (unsituated, literal content) as opposed to pragmatics (meaning as use) is preponderant in the current training and evaluation data of natural language understanding models. To alleviate that problem, we show that discourse marker prediction (classification of hidden discourse markers between sentences) can be seen as a pragmatics-centered training signal for text understanding. We build a new discourse marker prediction dataset that yields significantly better results than previous work. In addition, we propose a new discourse-based evaluation suite that could incentivize researchers to take into account pragmatic considerations when evaluating text understanding models.

Résumé

Les modèles computationnels pour la compréhension automatique des textes ont suscité un vif intérêt en raison de gains de performances inhabituels au cours des dernières années, certains d'entre eux conduisant à des scores d'évaluation surhumains. Ce succès a conduit à affirmer la création de représentations *universelles* de phrases. Dans cette thèse, nous questionnons cette affirmation au travers de deux angles complémentaires.

Premièrement, les réseaux de neurones et les représentations vectorielles sont-ils suffisamment expressifs pour traiter du texte de sorte à pouvoir effectuer un large éventail de tâches complexes? Dans cette thèse, nous présenterons les modèles neuronaux actuellement utilisés et les techniques d'entraînement associées. Nous proposons des critères pour l'expressivité de composition des représentations vectorielles et montrons que la suite d'évaluations et les encodeurs de phrases très répandus (SentEval / InferSent) sont limités dans leur expressivité; des changements mineurs peuvent permettre de nouvelles compositions expressives et interprétables, mais pourraient ne pas suffire, ce qui peut justifier le changement de paradigme vers de nouveaux modèles basés sur les Transformers.

Deuxièmement, nous aborderons la question de l'universalité dans les représentation des phrases: que cachent en réalité ces prétentions à l'universalité? Nous décrivons quelques théories de ce qu'est le sens d'une expression textuelle, et dans une partie ultérieure de cette thèse, nous soutenons que la sémantique (contenu littéral, non situé) par rapport à la pragmatique (la partie du sens d'un texte définie par son rôle et son contexte) est prépondérante dans les données d'entrainement et d'évaluation actuelles des modèles de compréhension du langage naturel. Pour atténuer ce problème, nous montrons que la prédiction de marqueurs de discours (classification de marqueurs de discours initialement présents entre des phrases) peut être considérée comme un signal d'apprentissage centré sur la pragmatique pour la compréhension de textes. Nous construisons un nouvel ensemble de données de prédiction de marqueurs de discours qui donne des résultats nettement supérieurs aux travaux précédents. Nous proposons également un nouvel outil d'évaluation de la compréhension du langage naturel en se basant sur le discours et la pragmatique. Cet

outil pourrait inciter la communauté du traitement des langues à prendre en compte les considérations pragmatiques lors de l'évaluation de modèles de compréhension du langage naturel.

Acknowledgments

Je lis parfois dans les remerciements d'autres manuscrits que la thèse est une entreprise longue et difficile. Je voudrais remercier tous ceux qui ont fait que la mienne ne le soit pas tellement. J'espère que ce que je cite n'occultera pas trop ce que par concision je ne cite pas. Je remercie ceux que j'oublie.

Merci à mes trois encadrants, disponibles et complémentaires. Merci à Philippe pour sa sagesse, son intransigeance et son soutien, Merci à Tim pour son style, sa vision et sa vivacité. Merci à Camille pour son inspiration, son acidité et son efficacité.

Merci à mon jury: Sophie, Ludovic, Martà, Nathalie pour leurs lectures avisées et si personnelles, ainsi que pour la richesse de leurs questions.

Merci au personnel de l'IRIT pour mettre tant de couleur dans ce bâtiment gris.

Merci à Sonia pour avoir illuminé (et sonorisé) le bureau 274, et avoir si bien assumé la lourde tâche de remplacer Antoine. Merci à Tom pour tant de conseils que je peine à les rendre. Merci à Sebastien et Abdel pour tous leurs passages. Merci à Patricia, pour avoir été un compagnon de conférences de couloir exceptionnel.

Merci Nicholas, Laure, Nathalie et vos personnalités qui portent l'équipe MELODI.

Merci à l'équipe de Synapse qui mériterait une énumération plus complète. Merci Patrick d'avoir permis cette thèse CIFRE, merci Emilie, Charles, Anais, Natalia pour leurs précieuses expériences de doctorants ou ex-doctorants. Merci Nicolas, Clément, Thibault, Anouk, Baptiste, Guilhem, Audrey, Sophie, pour tous nos échanges.

Je remercie mes amis. Ilyes avec qui j'ai tant partagé. Nicolas pour son influence et son style prodigieux. David pour être resté mon parrain.

Je remercie mes parents pour leur soutien, leur attention indéfectible, et leurs appels réguliers me faisant oublier la distance entre Toulouse et Lusigny-sur-Barse.

Et enfin, je remercie Hélène pour me mettre tant en question et être tant une réponse.

CONTENTS

CHAPTER 1

INTRODUCTION

Natural Language Processing (NLP) is a practical field with growing applications. Understanding the needs of humans paves the way for their automatic fulfilment (as in chatbot systems, robotics or information retrieval). Automated analysis of text can allow humans (or other agents, as in algorithmic trading) to learn new insights, such as sentiment analysis, summarization, relation extraction, trends detection, from streaming and massive data. These tasks can be framed as the prediction of an output (e.g. class, text, relation between objects) given an input (e.g. text). Figure 1-1 illustrates a selection of such tasks.

Humans can be lead to perform such tasks, and automated dedicated NLP systems can be designed or trained to assist or replace them with greater speed, consistency, and lower operating costs. Some tasks can rely on other tasks; for instance, a chatbot system (1-1c) can be decomposed into modules. A similarity estimation module (1-1d) can be used to extract answers to an utterance from previous conversations (if in the past, an human answered *Did you plug it in?* to *My computer mouse doesn't work :(*, it might be relevant to reuse that answer to a similar utterance (e.g. *my pointer is stuck!*). Natural Language Inference (1-1e) and discourse relation prediction (1-1f) systems are other examples of components that could also be leveraged in that use case, or other use cases like summarization.

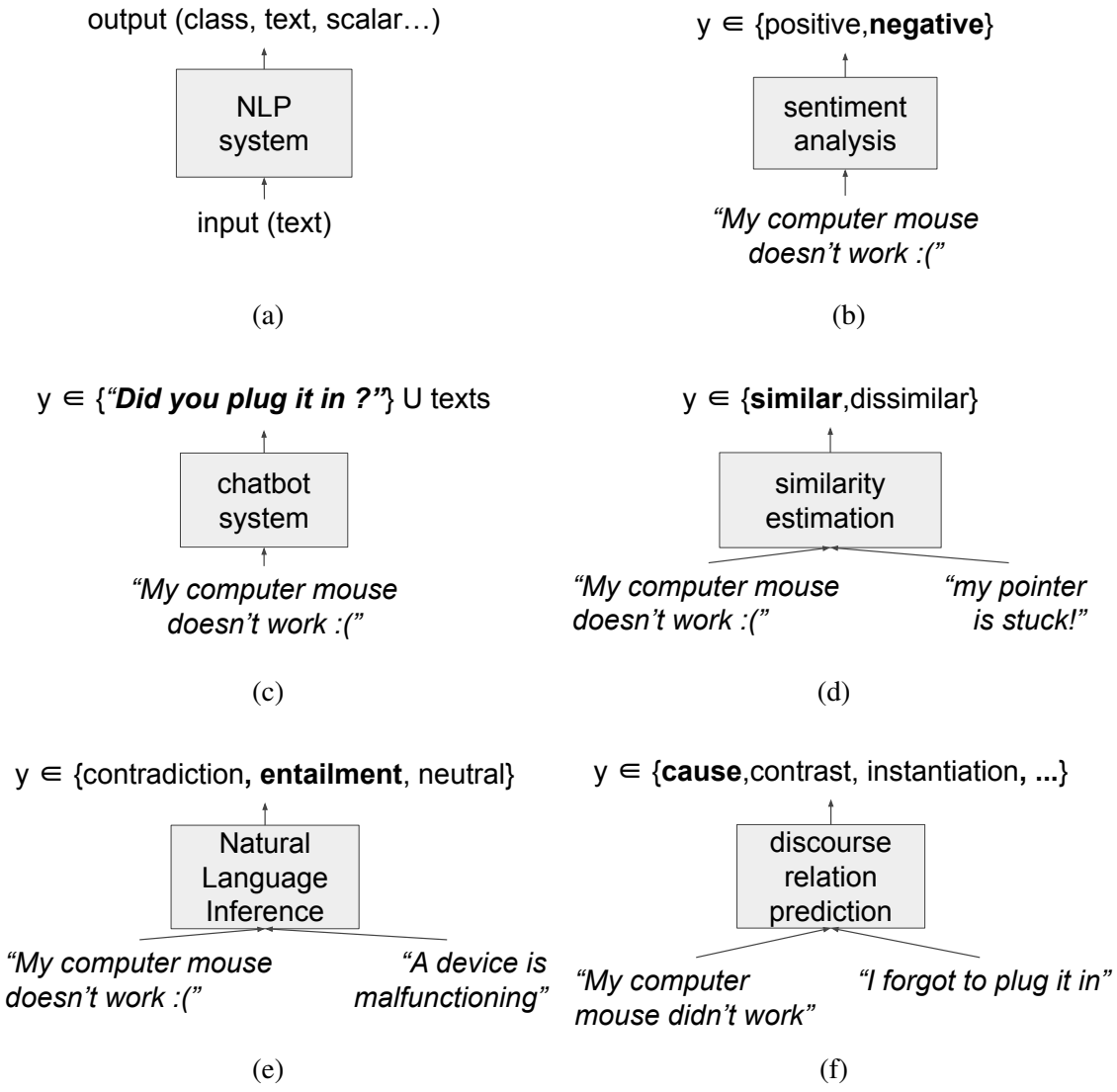


Figure 1-1 – NLP system for various tasks framed under template 1-1a

1.1 Theories behind NLP

Because of this practical orientation, NLP is driven by evaluations that are often standard automatic evaluations. Achieving state of the art results is a key factor for the popularity of new techniques. No matter how empirical NLP is, the underlying theories and models of meaning have permeated its trends across the different dominant paradigms in its history, carrying biases as assumptions about language. These biases might be wanted if they help to achieve desirable goals, but it is still worthwhile to uncover them. They expressed them-

selves throughout the evolution of artificial intelligence. This evolution has not occurred through clear-cut stages, but three paradigms can be distinguished according to (Wermter et al., 1996): symbolic, statistical, and connectionist methods (representation learning).

1.1.1 Symbolic AI

In the earliest dominant paradigm of symbolic AI, human intuitions and structural bias were transcribed programmatically into rules and ontologies. Hindsight allows us to identify those assumptions and the limitations they imply, but it was not always as obvious, as illustrated by Hubert Dreyfus in the following quote:

> “When I was teaching at MIT in the 1960s, students from the Artificial Intelligence Laboratory would come to my Heidegger course and say in effect: *”You philosophers have been reflecting in your armchairs for over 2000 years and you still don’t understand intelligence. We in the AI Lab have taken over and are succeeding where you philosophers have failed.”* But in 1963, when I was invited to evaluate the work of Alan Newell and Herbert Simon on physical symbol systems, I found to my surprise that, far from replacing philosophy, the pioneers in CS had learned a lot, directly and indirectly from the philosophers. They had taken over Hobbes’ claim that reasoning was calculating, Descartes’ mental representations, Leibniz’s idea of a “universal characteristic” – a set of primitives in which all knowledge could be expressed, – Kant’s claim that concepts were rules, Frege’s formalization of such rules, and Russell’s postulation of logical atoms as the building blocks of reality. In short, without realizing it, AI researchers were hard at work turning rationalist philosophy into a research program.” ”

Quote. 1-1 – quote from *Why Heideggerian AI Failed and how Fixing it would Require making it more Heideggerian* (Dreyfus, 2007)

The idea of rule-based systems depending on a clear cut logical structure was in itself a strong assumption, but the ontologies and rules had to be instantiated, and some natural language understanding problems require a lot of commonsense knowledge that had to be

stated by humans specification, even though we used some concepts that we aren't able to clearly define. For example, in order to know what was *too big* in the sentence *the trophy didn't fit the suitcase because it was too big*, knowledge that we put trophies in suitcases and not the converse can be necessary (Levesque et al., 2012).

1.1.2 Featured-based statistical methods

The later paradigm of statistical NLP aims to make statistical models learn to predict the correct output from previous examples rather than using enforced, hard-coded algorithmic structures. It is in itself theory laden due to the assumption of a statistical structure in language, but it marks the beginning of a shift from human asserted knowledge to data derived knowledge. However, bias is still present in inductive bias of statistical models (Baxter, 2000) and representations of data and data itself. Feature engineering was used to represent data. For instance, in classification tasks (e.g. spam detection), input text can be represented as a vector of word frequencies that serve as an input for a statistical model such as a support vector machine that uses *training data* to automatically discover patterns in the input representation to perform predictions (e.g. associations between some specific words and the spam/non-spam nature of a text message). In the past decades, the leading tendency has been to increase the expressivity of statistical model and reduce the amount of preprocessing of the input data.

1.1.3 Representation learning

The advent of neural processing systems takes this a step further by replacing handcrafted features with representation learning from ever more basic inputs. As of 2019, current state of the art models (Devlin et al., 2019) do not even tokenize text into words and use data-derived subwords instead. Since large neural networks and deep learning allow for tractable learning of complex functions (Raghu et al., 2017), they have the power to replicate some of the feature engineering humans would have done, without being limited by it.

Deep learning systems need to learn to derive features themselves. In order to do so, they rely on data whose annotations are costly to obtain. For this reason, transfer learning has become a standard solution to this problem. In transfer learning, a source task is used as a proxy to learn useful representations that are reused in the target task that needs to be solved. This source task can be unsupervised (i.e. requires no human annotation) or use a standard already annotated dataset.

Recently, the success of transfer learning has led to the spreading claim of universal natural language understanding systems, generalizing well to many tasks. (Kiros et al., 2015; Wieting et al., 2015; Conneau and Bordes, 2017; Subramanian et al., 2018; Cer et al., 2018b; Tamaazousti, 2018; Howard and Ruder, 2018).

1.1.4 Synthesis : focusing on data without neglecting theory

The notion of universality in this context is arguably a misnomer. Models are called "general purpose" or "universal" when they perform well in the few tasks they were evaluated on, and on a few domains. Strong assumptions can lie behind the choice of evaluation tasks when these tasks are not the target task. In fact, it would be impossible for a representation to be appropriate for all tasks (Watanabe, 1985), and even though it might be possible to have a form of universality that is limited to some practical purposes, it's hard not to underestimate the variety of tasks that is needed to evaluate the universality we actually want.

Current natural language systems automatically derive meaning representations from text; in order to do so, these systems rely on specific computational models, that are pretrained on large amounts of text, and automatically evaluated using standard evaluation benchmarks. In this thesis, we will firstly investigate the expressivity of these models, which we believe to have been largely ignored—leading to significant limitations within recent research. Secondly, we argue that the currently used training and evaluation datasets promote semantic aspects in representation, at the expense of pragmatics aspects, despite the wide acceptance of the importance of pragmatics in linguistics and computational linguis-

tics. Hence, in order to induce more genuinely universal meaning representations, we make the case for a stronger integration of discourse into neural models for NLP.

1.2 Thesis outline and contributions

We will introduce theories of meaning in chapter 2, neural models and distributed representation in chapter 3, and techniques to train neural models in chapter 4, and we will discuss the link between these signals and theories of meaning.

Then, we will present our contributions: we consider the expressivity of computational models, and integration of discourse through improvement of training signals and the proposal of a new evaluation benchmark.

More specifically, in chapter 5 we will analyze how different models compose representations to predict relations. We will propose desirable properties of composition through a case study of sentence embedding composition. Our analysis will highlight flaws in popular setups for evaluation and training, viz. SentEval and InferSent (Conneau and Bordes, 2017), and we propose solutions to these flaws. This chapter also proposes a view of semantic reasoning as geometrical operations in vector space. Finally, based on this study, we shed light on the success of recent methods based on transfer learning with pretrained transformers.

In chapter 6 we introduce Discovery, a dataset for unsupervised discourse based supervision of machine understanding models. We propose a method to automatically discover sentence pairs with relevant discourse markers, and apply it to massive amounts of data. Our resulting dataset contains 174 discourse markers with at least 10K examples each, even for rare markers such as *coincidentally* or *amazingly*. We use the resulting data as supervision for learning transferable sentence embeddings, and outperform models trained with Natural Language Inference (NLI).

In chapter 7 we introduce DiscEval, a new benchmark compiling 11 evaluation datasets with a focus on discourse aspects, that is designed for evaluation of English Natural Lan-

guage Understanding (NLU) methods.We make the case that discourse and pragmatics should be the center for evaluation of natural language understanding in current evaluation frameworks while current benchmarks are centered toward semantic considerations. We leverage our evaluation suite to show that the widely used NLI pretraining may not lead to the learning of really universal representations.

In chapter 8 we rely on Discovery and DiscEval to propose an automatic method in order to provide a semantics of discourse markers. This semantics explains and confirms the benefits of using discourse marker prediction as a training task. More concretely, using a model trained to predict discourse markers between sentence pairs, we predict plausible markers between sentence pairs with a known semantic relation (provided by existing classification datasets). These predictions allow us to study the link between markers and associated semantic relations. Thus, we provide a characterization of markers in use, as opposed to previous manual marker annotations.

Lastly, in chapter 9 we provide a general conclusion about our work and the interplay between the preceding chapters. We also discuss possible applications of our work and future directions.

A part of chapters 5, 6 have been published respectively in:

- Composition of Sentence Embeddings: Lessons from Statistical Relational Learning
  Damien Sileo, Tim Van-de-Cruys, Camille Pradel, Philippe Muller
  Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)

- Mining Discourse Markers for Unsupervised Sentence Representation Learning
  Damien Sileo, Tim Van-de-Cruys, Camille Pradel, Philippe Muller
  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)

. A part of chapters 7, 8 are under submission (ICLR2020 and LREC2020).

CHAPTER 2

THEORIES OF MEANING

NLP systems need a form of understanding in order to perform useful tasks. While it is not clear that human-like understanding is necessary to perform well in useful tasks (Dennett, 1989), theories of meaning can help the design or analysis of machine natural language understanding systems. Several views (and several categorizations of views) were proposed to define the meaning of texts.

2.1 Denotational view of meaning

In the denotational view, meaning is defined as a picture of the state of the world. A sentence then corresponds to a description or a picture of the world, or a set of possible descriptions. Meaning also have been argued to be a picture of mental representations instead of the world. Consider the following sentence:

The cat is either black or white. (2.1)

The denotational view can make statements truth conditional (Field, 1972), i.e. true if they correspond to the actual state of the world. In formal semantics, logical formulas have been

used as a non-ambiguous representation that can describe the parts of the world statements refer to. For instance, the formula in equation 2.2 could represent the sentence 2.1.

$$\exists x \text{is\_cat}(x) \wedge (\text{is\_white}(x) \vee \text{is\_black}(x)) \tag{2.2}$$

Assuming such a logical mapping, one can tell whether a sentence s_1 entails or contradicts a sentence s_2 based on classical logics' inference rules. This provides a test for sentence understanding: a system that captures the meaning of a sentence should be able to predict what other sentences it entails or contradicts for all plausible sentences. The logical mappings combined with inference rules provides a formalism to claim that *the cat is white* entails *the cat is either black or white*.

The denotational view also provides a way to define sentence similarity between sentences: two sentences are similar if they represent similar states of affairs (even though this can rely on human notions of significance). The field of semantics focuses on the study of meaning and truth conditions and ignores issues of context and communication considerations, being focused on literal content. However, even with a restriction under this scope, many problems arise in formal semantics, such as pronoun resolution, that require handling information across sentence boundaries. The theories proposed (Kamp et al., 2011; Groenendijk and Stokhof, 1991) to address these issues were an inspiration for some discursive approaches described later (subsection 2.2.2).

2.2 Pragmatics view of meaning

In the pragmatics view, the meaning of an utterance is associated with its use; it stems from the interaction of locutors in the world. Meaning is thus intrinsically dependent on the context. This context can be a discussion, a document, a workplace situation, or it might not be specified and just inferred. Upon reading a contextless sentence a human reader formulates presuppositions (Karttunen, 1974) about the context using common sense.

Semantics and denotational aspects are not ignored, since they can be a part of the mean-

ing, but they are seen as a means and not an end. In contrast to the denotational view, when considering pragmatics, two utterances have the same meaning if they have similar expected effects on the world.

Discourse analysis provides tools and concepts for studying pragmatics. We present some of them in this section.

2.2.1 Speech act theory

Speech act theory (Austin, 1962; Searle et al., 1980) proposes distinctions between utterances according to the kind of interaction with the world they allow. Table 2.1 shows a possible coarse classification of speech acts (Searle et al., 1980). While the denotational view of meaning provides tools to describe the content of utterances, speech acts describe planes of communication in which the content can be situated. An utterance such as *Bill was an accountant* is not a mere description of the world, but usually has communication purposes and in most cases is expected to be true, thus committing the speaker. Other kinds of speech acts are even harder to account for within the denotational view, since they can change the world upon utterance. A judge uttering *We find the defendant guilty* actually makes the defendant guilty. Such sentences are arguably neither true nor false, and entailment relations might not make sense. Instead, the notion of felicity conditions (Austin, 1962) was proposed in order to model the success of an utterance depending on a type of speech act (e.g. that an oath followed a conventional form).

Category	Description	Example
Representatives	commit a speaker to the truth of an expressed proposition.	Bill was an accountant.
Commissives	commit a speaker to some future action.	I'll call you tonight.
Directives	used by a speaker who attempts to get the addressee to carry out an action.	Sit down.
Declarations	affect an immediate change of affairs.	We find the defendant guilty.

Table 2.1 – A typology of speech acts
Adapted from http://www.ello.uos.de/field.php/Pragmatics/PragmaticsTypesofSpeechActs

2.2.2 Discourse representation theories

Since the situatedness of a sentence is key in the pragmatics view of meaning, it is useful to model the context in which an utterance occurs and how a sentence contributes to a larger ensemble. The discursive structure is a representation of a textual context. Several formalisms were proposed to characterize discourse structure, such as Rhetorical Structure Theory (RST) (Mann and Thompson, 1988) or Segmented Discourse Representation Theory (SDRT) (Lascarides and Asher, 2008). These formalisms represent a text or a conversation as graph of discourse units, either sentences, clauses, or speech acts, related by rhetorical relations of various types. They differ by their definition of an elementary discourse unit, by conditions on the structure of the graph and by the typology of the relations in the graph. The discourse relations can be seen as a situated extension of speech acts (Asher and Lascarides, 2003).

Figure 2-1 shows a RST discourse analysis of a document. This framework allows for a characterization of a sentence meaning as use. Even if we discard the problem of anaphora resolution, segment (3) is arguably more meaningful as a contrast segment (2) than in isolation, and the discourse analysis aims to uncover the structure behind documents or conversations.

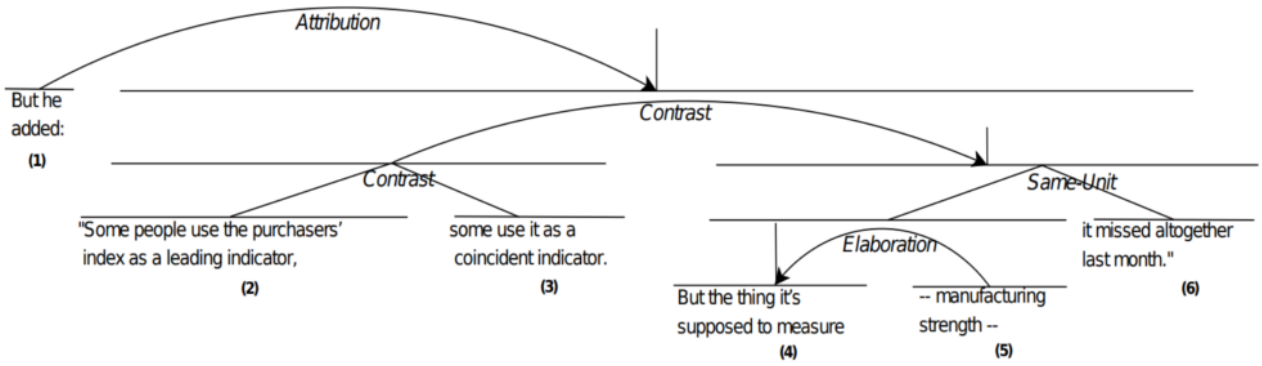


Figure 2-1 – RST-DT discourse parsing

Several resources gather annotations of discursive structures using different formalisms. The Penn Discourse TreeBank (PDTB) (Prasad et al., 2014) is one of them, and aims to

be theory neutral by annotating relation between discourse units independently without delving composite structures (e.g. trees).

2.2.3 Gricean implicatures

When pragmatics is taken into account, the meaning of a sentence isn't constrained to its literal content. The parts of the meaning that are not contained in the literal content are called implicature (Grice, 1975). For instance, an implicature of the sentence *I ate some of the cake* is that the speaker did not eat all of the cake.

Decoupling meaning from the literal content could open doors to indeterminacy (not being able to figure out the meaning among the possible interpretations). While indeterminacy might be intrinsic to textual communication (Mantzavinos, 2016), some principles can reduce this indeterminacy. Grice (1975) draws a list of maxims that the reader of a text assumes the writer to follow. The quote 2-2 displays a proposition for such a list:

> “
>
> **Maxim of quantity :** be as informative as one possibly can, and gives as much information as is needed, and no more.
>
> **Maxim of quality :** be truthful, and do not give information that is false or that is not supported by evidence.
>
> **Maxim of relation :** be relevant, and say things that are pertinent to the discussion.
>
> **Maxim of manner :** be as clear, as brief, and as orderly as one can in what one says, and avoid obscurity and ambiguity.
>
> ”

Quote. 2-2 – The four gricean maxims
adapted from https://www.sas.upenn.edu/~haroldfs/dravling/grice.html

For instance, a reader can infer that *I ate some of the cake* means that not all the cake was

eaten using the maxim of quantity.

2.3 Mentalist view of meaning

The mentalist theory identifies meaning with the ideas or concepts that correspond to an expression. The ideas encountered upon reading a sentence differ according to the reader, but there could still be a common basis to these experiences. It might be argued that ultimately, meaning is the content of communication that allow situations to evolve according to some needs. But, meaning is mental content as well, and it is not clear that we would attribute understanding to any system that performs the language tasks we want it to perform, as illustrated by the still ongoing *chinese room experiment* (Searle, 2006) argument debate. Significance and emotional content can be key to language understanding.

The mentalist side of meaning can also be seen through the lens of pragmatics and semantics. Searle et al. (1983) argues that mental representations have a psychological mode (e.g. remembering, perceiving, imagining) and a propositional content. The psychological modes can be seen as analogous to speech act categories addressed previously.

The mentalist view of meaning is mind-centered while pragmatics and semantics are world-centered, which might make them more relevant to model NLP systems that mediate between users and the world, and operate with observable, behavior data. Still, it is worthwhile to keep in mind the psychological, phenomenological sides of language understanding.

2.4 Sentence meaning and word meaning

In this chapter, we focused on the meaning of expressions like sentences. The meaning of words is also important to consider, since most language understanding system are based on word representations (or other elementary units such as subwords). In a semantic view, some words (e.g. names) can be seen as referring to specific parts of the world, predicates

or properties. Other expressions correspond to functors (quantification, negation) in logic (e.g. *not* can be mapped to a negation functor $\neg$). The meaning of words can also be defined only based on their use in sentences, marking less formal distinctions between words. In the next section, we will introduce computational word representations, and composition functions that yield sentence representations from them.

CHAPTER 3

COMPUTATIONAL MODELING OF MEANING

The common way of computationally processing text for language understanding is to split it into elementary units (characters, morphemes or other subwords, words, or word n-grams) from a fixed vocabulary, to represent those units (usually with vectors), and to compose those units in a text into another representation (e.g. a vector or a set of vectors), as illustrated in figure 3-1. Here, the system receiving a text string as input and producing a vector representation that reflects meaning is called a text encoder.

Words are a popular choice of elementary unit for text encoding. The smaller the elementary units, the smaller the vocabulary is (there can typically be millions of distinct words symbols, thousands of subwords, and hundreds of characters in a large corpus like English Wikipedia). The upside of using small units it that it allows text encoders to have a comprehensive vocabulary (so the probability of encountering an unknown unit is small). Smaller vocabularies have a smaller memory footprint. However, smaller units make the sequences longer, and increase the need for encoders to take in account long term dependencies between input units which is a hard problem (Hochreiter et al., 2001).

As seen previously in section 2.4, word representation require knowledge about the world, that can be useful for several NLP tasks. As a result, using word embeddings as a central building block has been fruitful for NLP over the last decade (Young et al., 2018). Thus,

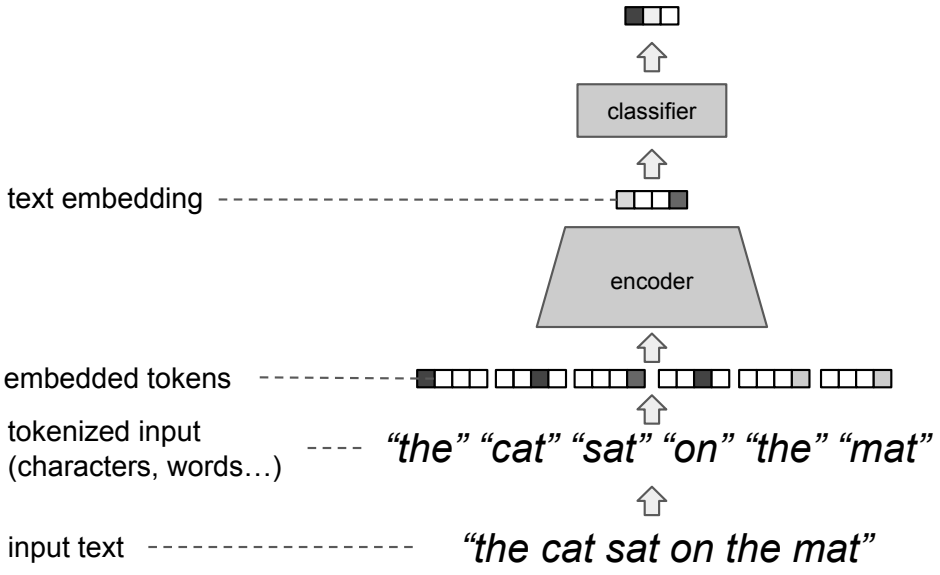


Figure 3-1 – Architecture of a text encoder

we will focus on word embeddings as an illustration of the representation of an elementary unit and then discuss ways to derive sentence embeddings from them.

3.1 Word Embeddings

3.1.1 Attribute based representation

Words can be represented with the help of various kinds of attributes. Many aspects of human representations of words can be reduced to features that can fit into a vector. Properties such as grammatical category can be mapped to one hot vectors. This also holds for surface aspects of words (e.g. size, presence of n-grams), and the features of the objects possibly denoted by a word sense (e.g. typical color, weight of an object).

Words can also be characterized by their relation to other words. In an explicit way, relations such as synonymy, antonymy or meronymy, carry substantial information about words. More implicitly, the contexts in which a word occurs, (*the company it keeps* (Firth, 1957) carries a lot of information about it. This principle is called the distributional hypothesis. There are several ways to cast that contextual information into vectors (Van de Cruys,

2010). Table 4.1 shows various context instances for the same words. Here, the context is limited to 5 words both to the left and right, and does not allow words outside sentences boundaries of contiguous sentences, but there are many ways to define contexts. Such contexts can be vectorized using a bag of word frequencies that carry substantial information about the word *science*. This process can be repeated for other words such as *technology* or *turquoise*, and standard vector similarity metrics (e.g. cosine similarity) between their vectors reflect human notions of relatedness or similarity (Torabi Asr et al., 2018) to some extent.

left context	word	right context
...functional interplay of philosophy and		should, as a minimum, guarantee..
...and among works of dystopian		fiction.
The rapid advance in		today suggests that the existing...
...calculus, which are more popular in		-oriented schools

Table 3.1 – Context of the word *science*, adapted from Piedeleu et al. (2015)

Task oriented lexicons can also represent words. Word lexicons can be annotated for sentiment analysis, emotion analysis or discourse analysis (Kaur and Gupta, 2013; Das et al., 2018).

Several initiatives, such as WordNet and ConceptNet can be seen as instantiations of that feature-based view of language. The concatenation of these representations for all considered aspects could yield a very high dimensional sparse vector containing a lot of (redundant) information on the word.

3.1.2 Dimensionality reduction of attribute vectors

While attribute based representations can yield decent results for similarity computation and token representation, their size can be problematic. The vector space dimensions aren't used very efficiently. For instance, the similarity between attributes is not used in categorical subspaces: in a subspace where *cat* is represented by $[1, 0, 0]$, *feline* is represented by $[0, 1, 0]$ and *submarine* by $[0, 0, 1]$, the three vectors will be separated by the same cosine distance, but it might be desirable to represent the words so that *cat* is closer to *feline* than

to *submarine*.

Dimensionality reduction techniques can solve this problem by leveraging redundancy in features and condensing representations in a limited number of dimensions which contain most of the information. The resulting vectors are called embeddings. Learning embeddings that perform dimensionality reduction of word contexts yielded great improvement over previous representations in the popular Word2Vec (Bordes et al., 2013b) method. The context matrix can also be computed explicitly before factorization, as in the GloVe model (Pennington et al., 2014) or implicitely as in the Word2Vec model (Levy and Goldberg, 2014b). Word2Vec and GloVe precomputed word embeddings are publicly available [1] and the knowledge they carry is a useful means to bootstrap text encoders.

Baroni et al. (2014b) showed that in the context of word counts, dimensionality reduction has a more profound effect than just the reduction of the number of parameters, since it leads to a systematic improvement in downstream evaluations.

Some directions in precomputed embedding feature spaces could be interpreted as corresponding to human concepts like gender. The widely cited example $\vec{king} - \vec{queen} \approx \vec{man} - \vec{woman}$ can hint at the existence of a gender direction on the $\vec{man} - \vec{woman}$ axis. This confirms that a form of abstraction occurs during the learning of word embeddings.

The embeddings can also be used as input features for classification. Gender, plurality, and sentiment can be fruitfully predicted using logistic regression with such features (Chen et al., 2013). To do so, these words need to be combined into sentence or text embeddings.

3.2 Sentence embeddings

Text encoders rely on a form of the compositionality principle (Frege, 1884; Szabó, 2017). It states that the meaning of word combinations is derived from the meaning of the individual words, and the manner in which those words are combined. For instance, according to this principle, the meaning of the noun phrase *carnivorous plants*, can be derived from the

[1] http://vectors.nlpl.eu/repository/

meaning of *carnivorous* and the meaning of *plant* through a process named composition. Successive composition steps can yield a sentence representation that can be used as features for other tasks such as similarity estimation or classification, using statistical models such as logistic regression.

A very simple composition consists in averaging the word embeddings of all words occuring in a sentence (Shen et al., 2018). However, the order of words isn't taken in account during averaging (*the cat sat on the mat* and *the mat sat on the cat* have the same representation with embedding averaging). But accounting for word order is necessary for deeper understanding. This is made possible with more sophisticated combinations of embeddings, as in convolutional neural networks (Collobert and Weston, 2008), recurrent neural networks or transformers (Vaswani et al., 2017).

A sentence embedding is a fixed sized vector that characterizes the meaning of a sentence. Sentence embeddings can be constructed on the basis of the previously cited word embeddings. The word embeddings can also be learnt from scratch and during that process acquire similar properties. In sentence embeddings, words can be seen merely as tools for composition, which can be why state of the art methods are departing from pretrained word embeddings (Devlin et al., 2019).

3.2.1 Composition

The compositionality principle can be operationalized with vector representations (Baroni et al., 2014a; Clark, 2015). If we want embeddings for a particular kind of phrase (e.g. adjective/nouns like *white cat*), it is possible to start by representing *white* and *cat* into two respective embeddings h_1 and h_2. The composed representation $h_{1,2}$ can be computed as a function of h_1 and h_2 as in equation 3.1.

$$h_{1,2} = f_\theta(h_1, h_2) \text{ where } f \text{ is a function parametrized by } \theta \tag{3.1}$$

When humans read a phrase, they compose their representations of the words into a phrase representation (Szabó, 2017). This task requires a lot of background knowledge. For instance, *white cat* and *white wine* actually have a different color even though the adjective specifying the color is the same, as shown in figure 3-2. This phenomenon is known as co-composition (Pustejovsky, 2012). A broader context can also be necessary for composition. For instance, the way the words in *green light* should be composed depends on the situation. A green light can denote an authorization or an actual green light. The meaning of some idiomatic expressions requires a form of memorization rather than composition per se. Thus, performing those compositions in the vector space requires powerful nonlinear functions. Deep neural network can implement such functions, and also have memorization abilities (Arpit et al., 2017) that can allow them to deal with idioms.

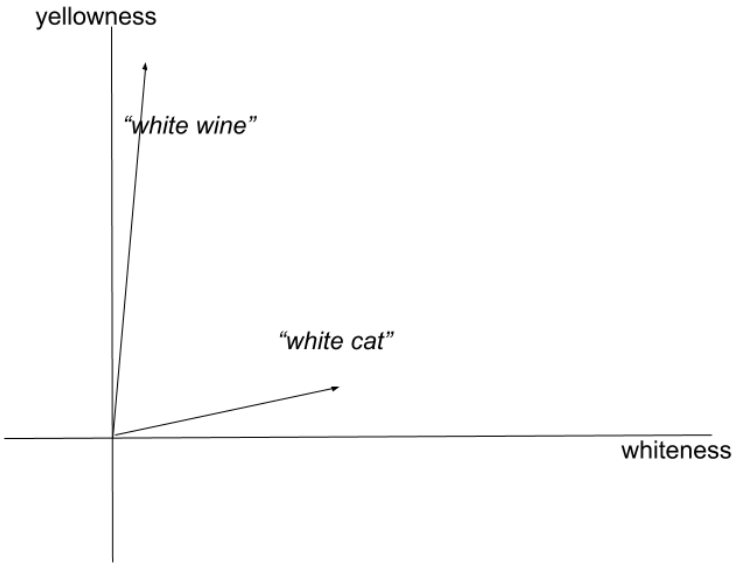


Figure 3-2 – How adjective meaning can depend on the noun

3.2.2 Guiding compositions through parsing

A form of parsing is required in order to find what words or groups of words should be composed in order to perform the composition.

Constituency based parsing is a standalone instrument that is useful for text analysis. It can

be used to determine what parts of a sentence should be composed together (Socher et al., 2013). Figure 3-3 shows an example of such a parse tree.

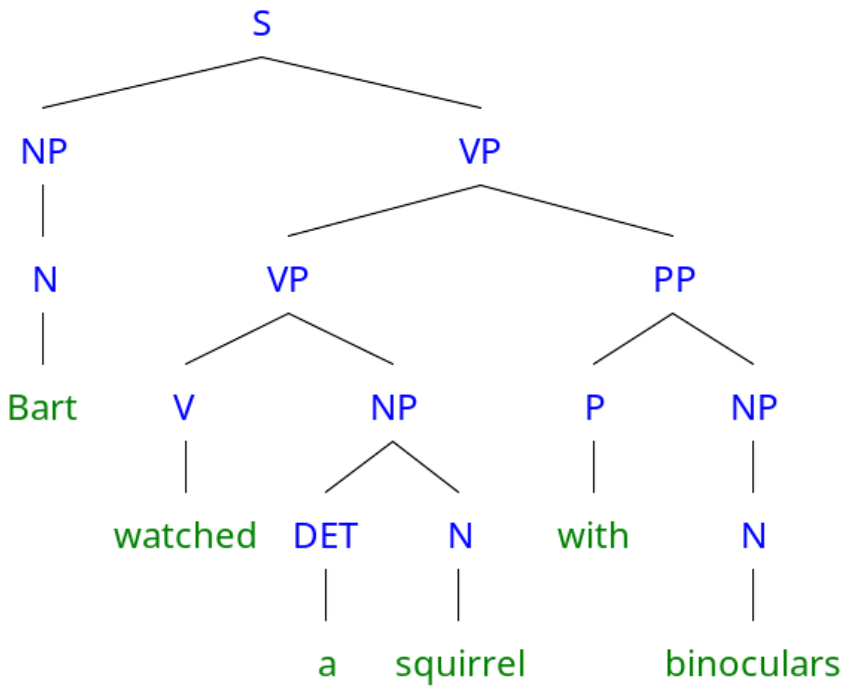


Figure 3-3 – A constituency-based parse tree of the sentence "Bart watched a squirrel with binoculars"

However, a constituency based parse tree might be a too simplistic structure. Words arguably all influence each other forming a form of rhizome (Deleuze and Guattari, 1988) rather than a tree. Besides, there can be multiple parse trees for a given sentence, as shown in figure 3-4.

The parse tree from figure 3-3 is more likely that the alternative from figure 3-4. However, already having already achieved a form of understanding is necessary in order to compare the likelihoods of those parses. This can be done through compositions in parse candidates and an iterative process, but this means that parsing and composition are intertwined.

Text encoders can perform a form of parsing as well, since they can selectively compose words. Convolutional Neural Networks (CNN) compose adjacent words. Recurrent Neural Networks (RNN) compose words with a memory of the previous words at a given position in the sentence. Transformers selectively and fuzzily associate words according to their

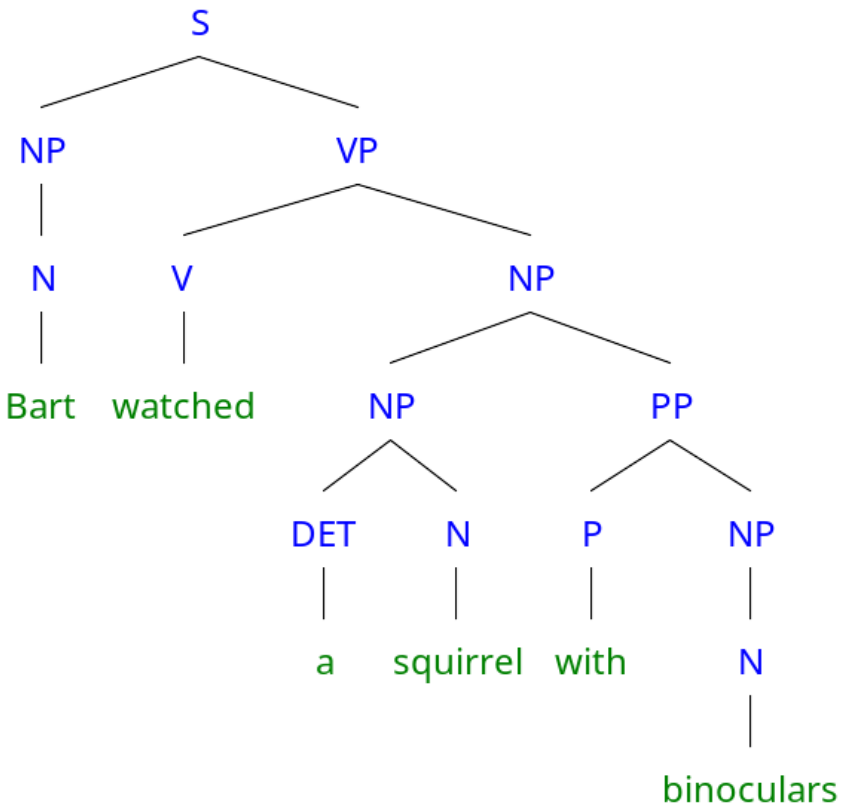


Figure 3-4 – Another constituency-based parse tree of the sentence "Bart watched a squirrel with binoculars"

nature and position through different channels and reunite these channels in token-wise representations (one token serving as a placeholder for a text representation) at each layer.

All methods for sentence embedding compose words until a single sentence embedding is derived. This can be realized with a *pooling* operation that derives a single embedding from a sequence of embeddings. For instance, the embedding corresponding to the last element of the sequence can be chosen to represent the sentence. Alternatively, element-wise max pooling can be used (representing a sentence with a vector for which each dimension is the maximum value among corresponding dimensions in the input sequence vectors). Max pooling is more common than last state pooling and alleviates the problem of forgetting the beginning of sentences.

It is worthwhile to note that recurrent neural networks and transformers are Turing complete (Pérez et al., 2019) (convolutional neural networks and average pooling are not), so they

have the ability to implement arbitrary algorithms (e.g. complex parsing schemes, complex compositions over input sequence of tokens). This property is quite desirable but it is not sufficient, since we want these architectures to *learn* those algorithms from a limited set of examples, and limited model capacity. In such conditions, it is possible that convolutional neural networks outperform RNN which are more expressive. The architecture have to allow not only possible but efficient flows of information that suits the kind of algorithms that should be learnt to perform well at a given set of tasks.

3.2.3 Recurrent Neural Networks

Humans can understand sentences presented one word at a time (Kroll, 1980) quite efficiently (twice as fast as they would read the full sentence, without enforced sequential presentation, even though it becomes more difficult at paragraph level). When we read a sentence in that way, we have a mental representation for the state of the sentence and update each time we encounter a new word. Recurrent Neural Networks (RNN) process input tokens in such a sequential way, as illustrated in figure 3-6. RNNs are based on a

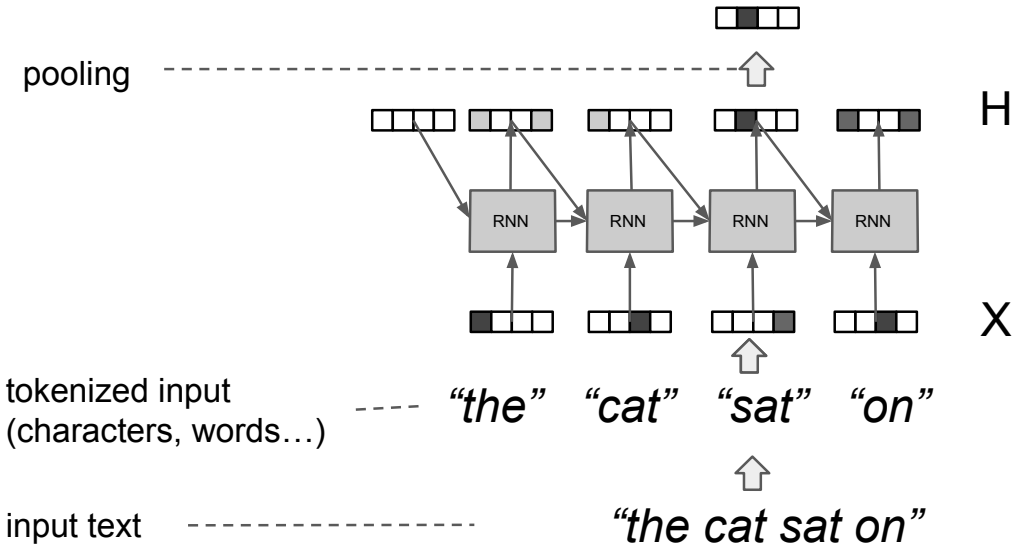


Figure 3-5 – RNN layer

parametrized function f_θ, as shown in equation 3.2.

$$h_t = f_\theta(h_{t-1}, x_t) \tag{3.2}$$

This is analogous to the composition described in equation 3.1 except that instead of composing several words, RNN compose words with a memory. RNN have to perform both composition and parsing at the same time. f_θ can select the relevant parts of that memory for each word, thus performing a form of implicit parsing (Bowman et al., 2015c). Long Short Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) and Gated Recurrent Unit (GRU) (Chung et al., 2014) are specific RNN architectures. While the most standard RNN architecture implements f in the following way:

$$h_t = \tanh(x_t U + W h_{t-1}) \text{ where } W, U \in \mathbb{R}^{d \times d}, h \in \mathbb{R}^d \tag{3.3}$$

an LSTM uses gates (i, f, o) that mediate the modification of the state, and a protected memory C, as described in equation 3.4. f controls the information that should be kept in the updated protected memory C_t, while i controls the information that should be extracted from $\tilde{C}_t$. o controls the information that should be disclosed in the visible state h.

$$\begin{aligned}
i_t &= \sigma\left(x_t U^i + h_{t-1} W^i\right) \\
f_t &= \sigma\left(x_t U^f + h_{t-1} W^f\right) \\
o_t &= \sigma\left(x_t U^o + h_{t-1} W^o\right) \\
\tilde{C}_t &= \tanh\left(x_t U^g + h_{t-1} W^g\right) \\
C_t &= \sigma\left(f_t * C_{t-1} + i_t * \tilde{C}_t\right) \\
h_t &= \tanh(C_t) * o_t
\end{aligned} \tag{3.4}$$

3.2.4 Attention and Transformers

Transformers (Vaswani et al., 2017) have gained a lot of popularity recently. In Transformers, tokens are represented with a positional encoding that contains features fully characterizing the position of tokens in the sequence, in addition to the standard trainable token

embeddings that reflect token properties.

For each token of the input sequence H, a new representation H' is computed based on the input sequence, with the help of an *attention mechanism* composed of several *attention heads* that are similarly structured modules focusing on different aspects of input sequence through different weights.

More specifically, for each token h_i of the input sequence $h_1, h_2...h_n$, an attention head k computes to what extent h_i should interact with other tokens and what information should be extracted from these tokens.

A token representation is mapped to three representations that account for different functions of the token:

- $K = W_K^k H$ represents what the token is looking for in the query representations of other tokens
- $Q = W_Q^k H$ represents what the token is presenting to other tokens
- $V = W_V^k H$ represents the information that should be passed on by a token if the information it presents has been picked by a query

$$H^k = \text{softmax}(\frac{QK^T}{d})V \tag{3.5}$$

The representations H^k of various attention heads indexed by k are merged into H' with a multilayer perceptron taking concatenation $[H^1, H^2....H^n]$ as input. That step allows a reasoning to occur when aggregating the different aspects of the relation between the words (taking the positions into account).

The transformer architecture is able to deal with long term dependencies, and might be better suited to subwords than RNNs.

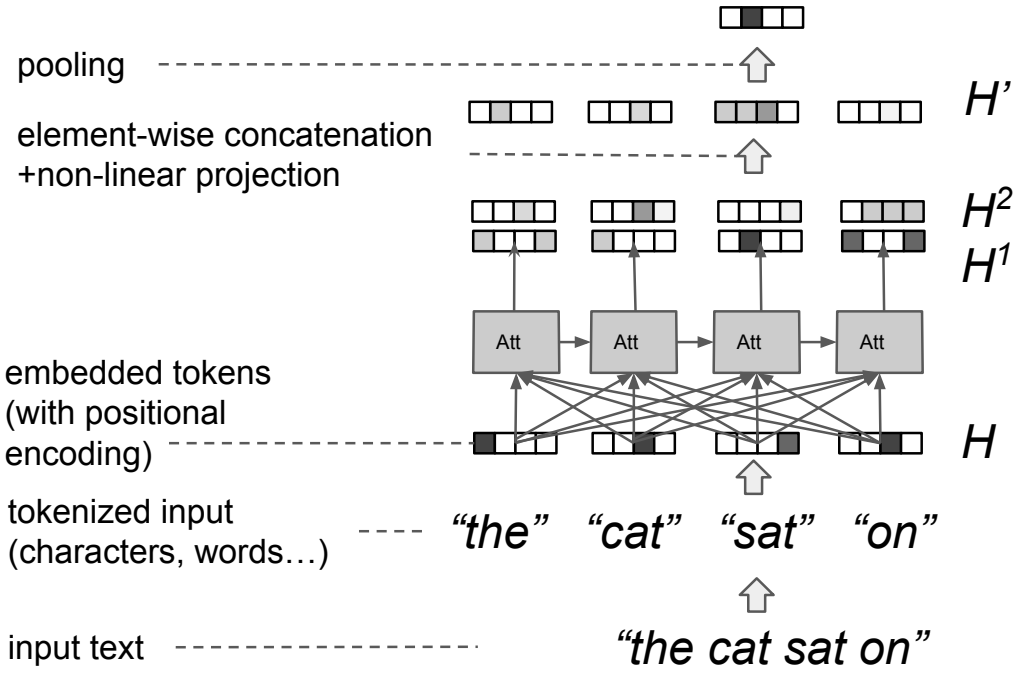


Figure 3-6 – Transformer layer

3.2.5 Depth, bidirectionality: stacking encoding layers

We just described a RNN layer and a transformer layer. These layers can be stacked (Pascanu et al., 2014) in order to allow depth distributed computations that significantly improve the capacity (Zhang et al., 2016). With depth, the lower layers can provide features that can be used by higher level layers, allowing a factorization of computations. The features can be hierarchies, and higher layers are seen as able to represent high level concepts.

The first layer, the embedding layer, represents all tokens of the input sequence H. The RNN or Transformer layer can compute contextualized representation H'. But the same layer architecture (with different weights) can perform the same operation taking H' as input instead of H, and so on. State of the art NLP models stack up to 24 layers (Devlin et al., 2019).

RNN encode tokens sequentially (from left to right or right to left) which can pose two problems: the first element of the sequence (if using left to right RNN) cannot be contextualized, and the first elements are more prone to forgetting. To mitigate those problems,

bidirectional RNN were proposed (Schuster and Paliwal, 1997), where two RNN with different weights (left to right and right to left) are used and combined (e.g. with concatenation of states).

3.2.6 Interpretation of sentence embeddings

Unlike word representations derived from Principle Component Analysis or Latent Dirichlet Allocation (Blei et al., 2003), the dimensions of standard word embeddings are not intepretable by design, even though vector space analogies or classification tasks can give an idea of what information is stored in word embeddings. Sentence embeddings might be conceptually even harder to interpret.

The dimensions of popular sentence embeddings have been shown to contain a variety of information, such as lexical, syntactic and surface information Conneau et al. (2018a). Different views of meaning can yield different interpretations of the structure of the vector space. For instance, according to correspondence theory of meaning, sentence that depict similar worlds should be close. The vector space might be structured to contain predicate, objects and subjects of a sentence. By contrast, according to the theory of *meaning as use*, the positions in the vector space of meaning should allow sentences that have the same use to be close.

Sentence embeddings have also been called ”thought vector” in SkipThought article (Kiros et al., 2015), which can remind us of the mentalist view of meaning from section 2.3, even though SkipThought vectors arguably conform to meaning as use. There have been studies correlating neural representation of sentences and brain activity measurements (Gauthier and Ivanova, 2018).

Using fixed size vectors for sentence representation is questionable itself, since text can be ambiguous and one could think that a sentence embedding corresponds to a single meaning. Alternatives to vector representations have been proposed for words, such as Gaussian representations (Vilnis and McCallum, 2015; Athiwaratkun and Wilson, 2017) where standard deviation and multi-modality can represent ambiguities. Gaussian representations can also

be used for sentences (Bowman et al., 2016) even though they were used for text generation and not evaluated for representation learning purposes. However, we make the hypothesis that single vectors are powerful enough to represent ambiguity since a projection in vector space can select the relevant content, even though this has not been explored in previous work to the best of our knowledge.

Until now, we focused on word and sentence embeddings; but sentence pairs, or larger texts can also be embedded into a single vector; this embedding can be useful in itself as a text representation, or can be used as an input for relation prediction between sentence pairs. The information that should be contained is different in each use case, even though they could be represented jointly. In the case of a relation representation, the embedding should characterize relations between the input, and encode information about various aspects, and various relations such as contrast or contradiction. In the case of text representation, the embedding should use those relations to produce a synthesis of the input sentences.

3.3 Transfer without explicit sentence embedding

Explicit sentence embeddings are not necessary to perform transfer learning, and can induce an unnecessary information bottleneck. For a target task that is a relation prediction between two sentences, starting with the sentence embeddings means that we have to train a model to compose the sentence embeddings. Two sentences $s_1 = w_1^1, w_2^1, w_3^1...$ and $s_2 = w_1^2, w_2^2, w_3^2...$ are embedded with $f_{sentence\_encoder}$ into h_1 and h_2 and a sentence composition $f_{relation}$ produces a relation representation

$$h_{relation} = f_{relation}(f_{sentence\_encoder}(w_1^1, w_2^1, w_3^1...), (f_{sentence\_encoder}(w_1^2, w_2^2, w_3^2...)) \quad (3.6)$$

Instead, it is possible to learn a text encoder that encode multiple sentences, and the relation between them at the same time.

$$h_{relation} = f_{relation}(w_1^1, w_2^1, w_3^1..., w_{SEP}, w_1^2, w_2^2, w_3^2...) \quad (3.7)$$

SEP is a special token that denotes the separation between sentences. This formulation allows a model to compare different aspects of the sentences without having to squeeze all the information into sentence embeddings. If $f_{relation}$ is a RNN/LSTM, the information will still be squeezed into a single vector (the memory), but if $f_{relation}$ is a Transformer, this strategy can be helpful. Training such a model incentivizes implicit sentence representation capabilities and the learning of composition of these implicit sentence representations jointly. Representations are captured in functions instead of being captured in vectors. This strategy has been employed in recent popular work (Rocktäschel et al., 2015), though it has been formulated explicitly more recently (Devlin et al., 2019).

CHAPTER 4

TRAINING, TRANSFER, EVALUATION

In the previous chapter, we presented parametrized computational models that can perform a form of reasoning and mapping from a string to a representations that could be used to perform various tasks. However, these models need to learn the right parameters in order to do anything useful. In this chapter, we introduce techniques and resources that are used to instantiate the previously described architectures into useful encoders.

4.1 Transfer Learning

The goal of transfer learning is to train a neural network on a set of tasks (called *source tasks*) in order to learn a form of knowledge or skill, and to *transfer* that knowledge into other tasks we actually want to perform (called the *target tasks*). The source tasks are only used as a proxy to improve the results in a target task. For instance, humans can learn driving in simulators (source task) in order to get better at driving in real cars (target task). They don't necessarily care about being able to drive well in a simulator, but both tasks require some of the same skills, and driving in a simulator is cheaper while actually improving driving skills in real situations. There are many forms of knowledge or skills that are common across NLP tasks, such as representation of words, composition capabilities,

parsing the structure of text, possession of commonsense (or specialized) knowledge and ability to use it, representation of human psychology (including Gricean maxims). Thus, if we have a reusable, generic text encoder that successfully learned these skills with a source task, a model based on that reusable text encoder will mainly have to learn *what* it has to do instead of *how* it has to do it. Figure 4-1 depicts a traditional supervised learning model without transfer, where the model learns everything from scratch in order to perform the task, and a model based on a reusable text encoder.

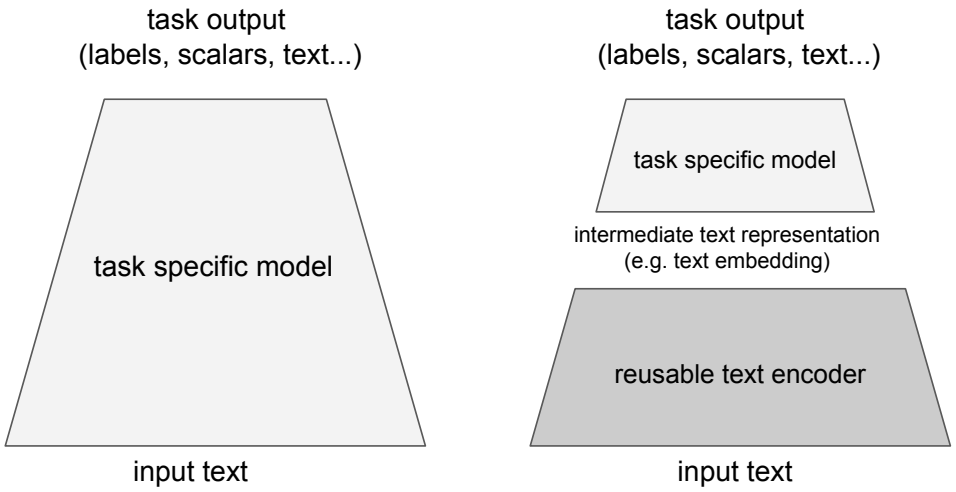


Figure 4-1 – Reusable text encoder

Instead of a sequential (source task, then target task training) training, a joint training can be used. This method is called multi-task learning (Caruana, 1997) and forces the model not to ”forget” how to solve the base task.

4.2 Training signals

Various training signals can incentivize text encoders to learn different capabilities and induce different kinds of representations. Sentence embeddings can contain lexical information, syntactic, semantic, or pragmatic information that are more or less desirable depending on the intended use, and the training signals have a key influence on the representa-

tions. For instance, we might want to incentivize encoders to discard syntactic information in the final representation if the intended use is semantic (or discursive) similarity, since two paraphrases should always be close but can have quite different syntactic structures.

4.2.1 Language modeling

A popular training signal is language modeling, i.e. prediction of masked words in a text. This task is quite general since we can find instances of text where the prediction of the right words might require the model to perform various tasks such as translation, summarizing, or question answering (Radford et al., 2019). The following text is a verbatim extract from a web text corpus[1]

s_A=*Brevet Sans Garantie Du Gouvernement", translated to English: "Patented without government warranty"* (4.1)

Masking $Patented$ in s_A and training a model to predict it would incentivize the model to learn context understanding and translation capabilities.

Similarly, masking *rejoiced* in s_B in 4.2.1 would incentivize the model to learn sentiment analysis (words like *mourned* could fit in this syntactic context, but it would make less sense because of sentiment consistency.)

[1] from https://www.sliderulemuseum.com/France.htm

s_B= *Qaro's son married Luria's daughter, and Qaro rejoiced at the connexion, for he had a high opinion of Luria's learning.* (4.2)

Even though other easier cues might help, the wide variety of sentence kinds and possible maskings make the task challenging and interesting as a training signal for semantic and discursive understanding.

Language models can learn to predict probabilities of masked positions that maximize the likelihood of a text (through the tractable log-likelihood) given a density modeled as in equation 4.3 for a sentence composed of tokens $t_1..t_n$.

$$p(s) = p(t_1...t_n) = \prod_{i=1}^{n} p(t_i|t_1...t_{i-1}) \tag{4.3}$$

This training objective operate on the token level. It does not directly involve the computation of a sentence embedding, though variations can be derived to allow it, as shown in the next subsection. The next objectives we are going to present are sentence based.

4.2.2 Sentence-level distributional hypothesis

Language models exploit the distributional hypothesis described in 3.1.1 at the word level. The distributional hypothesis can also be used at sentence level. Sentences occurring in similar contexts are similar. Thus, deriving training signals from the sentence level distributional hypothesis can possibly induce discursive knowledge, together with semantic knowledge.

Table 4.1 shows left and right contexts for a sentence disclosed in the caption.

left context	sentence	right context
The distributional hypothesis can be generalized to sentences.		So not every sentence can plausibly occur in a context.

Table 4.1 – Possible context for the sentence *Left and right context usually impose coherence constraints on a sentence.*

4.2.2.1 Prediction of sentence

Training tasks can be derived from this principle. One way is to consider naturally occurring sequence of sentences $s_0, s_1, ...s_n$ and predict whether two sentences are consecutive or not (Logeswaran et al., 2018). However, local coherence can be straightforwardly predicted with relatively shallow features (Barzilay and Lapata, 2008) such as reuse of words or syntactic consistency.

4.2.2.2 Prediction of words

Another is to consider consecutive sentences with their neighbourhood s_{i-1}, s_i, s_{i+1}, and to predict words from s_{i-1} or/and s_{i+1} given s_i. A sentence embedding of s_i can be used for that purpose, thus the training would incentivize s_i to carry useful information for prediction of words in $s_{i\pm1}$. This technique has been used in (Kiros et al., 2015).

4.2.3 Natural Language Inference

As we have seen in section 2.1, detecting entailment relations can be seen as a test for understanding according to the semantic view of meaning. Models fine-tuned on this dataset transfer well to other semantic tasks (paraphrase detection, semantic similarity) (Conneau and Bordes, 2017), and sentiment analysis. Because of this, the natural language inference task has been increasingly popular (Bowman and Zhu, 2019) as a source training

task. Many datasets exist, but two of them stand out due to their sizes: Stanford Natural Language Inference (SNLI) and Multigenre Natural Language Inference (MNLI).

Table 4.2 shows some examples of the SNLI dataset.

premise	hypothesis	label
An older and younger man smiling.	Two men are smiling and laughing at the cats playing on the floor.	neutral
A black race car starts up in front of a crowd of people.	A man is driving down a lonely road.	contradiction
A soccer game with multiple males playing.	Some men are playing a sport.	entailment

Table 4.2 – Examples from the SNLI dataset

The semantic bias in the view of meaning is clearly stated by the main contributor of those datasets, as shown in quote 4-3.

> “Pragmatic inference plays a substantial role in almost any instance of human language understanding, and it is relevant to any reasonable view of sentence meaning. In the context of applied natural language inference, it has been studied in work like that of (de Marneffe et al., 2012) but it is not a major focus of current research. It is not highlighted in any of the evaluation tasks that I use in this dissertation, and I choose [to] leave it as a background issue in much of what follows.”

Quote. 4-3 – Quote from *Modeling Natural Language Semantics In Learned Representations* (Bowman, 2016)

The SNLI dataset premises are sentences extracted from a corpus of Flickr image descriptions. This reinforces the semantic connotation of the dataset since it deals only with texts literally conforming to the picture theory of language (Wittgenstein, 2013) which is a semantic view of meaning considering that utterances can be mapped to images of the world.

4.2.4 Discourse marker prediction

A natural pragmatics-oriented counterpart to NLI could be the prediction of the semantic or rhetorical relation between two sentences, as is the goal of discourse parsing. A number of annotated corpora exist, such as RST-DT (Carlson et al., 2001) and PDTB (Prasad et al., 2008), but in general the available data is fairly limited, and the task of discourse relation prediction is rather difficult. The problem, however, is much easier when there is a marker that makes the semantic link explicit (Pitler et al., 2008a), and this observation has often been used in a semi-supervised setting to predict discourse relations in general (Rutherford and Xue, 2015). Building on this observation, one approach to learn sentence representations is to predict such markers or clusters of markers explicitly (Jernite et al., 2017; Malmi et al., 2018; Nie et al., 2019). Consider the following sentence pair:

I live in Paris. But I'm often abroad.

The discourse marker *but* highlights an opposition between the first sentence (the speaker lives in Paris) and the second sentence (the speaker is often abroad). The marker can thus be straightforwardly used as a label between sentence pairs. In this case, the task is to predict $c =$ *but* (among other markers) for the pair (*I live in Paris*, *I'm often abroad*).

Table 4.3 shows further examples from the DisSent dataset.

S1	marker	S2
Her eyes flew up to his face.	and	Suddenly she realized why he looked so different.
The concept is simple.	but	The execution will be incredibly dangerous.
You used to feel pride.	because	You defended innocent people.
I'll tell you about it.	if	You give me your number.

Table 4.3 – Examples from the Book8 dataset

Discourse marker prediction has been used to improve discourse relation prediction, where automatically extracted explicit instances feed a model targeting implicit instances (Marcu and Echihabi, 2002; Sporleder and Lascarides, 2008; Pitler and Nenkova, 2009; Rutherford

and Xue, 2015). Jernite et al. (2017) used a similar approach but add additional objectives such as order prediction (predicting whether the order of two initially consecutive sentences has been reverted) and detection of consecutive sentences as defined in 4.2.2. This supervision is arguably more pragmatics-oriented than NLI supervision, even though semantics is mentioned by Nie et al. (2019) and not pragmatics. (Nie et al., 2019) is quite popular. However, we will propose criticism and improvements over this dataset in chapter 6.

4.3 Evaluation methods

Since a goal of NLP is to improve performance in various tasks, evaluation is key to NLP practitioners as a guidance for model selection. The best evaluation is arguably feedback from real use. However, it is not always possible to evaluate systems that are not ready for production. A Twitter Bot released by Microsoft, named *Tay* (Wolf et al., 2017) that was manipulated by some users to propagate hateful speech is a revealing example of how NLP systems could go wrong in production. More generally, we need to test a system because we do not know how unsatisfactory it is, and we cannot always afford to provide an unsatisfactory system to real users. But if it is possible, A/B testing can be used.

Automatic evaluations allow a fully offline evaluation based on datasets of previously collected annotated data. The dataset might reflect the target task (e.g. intent classification) in a dialog system. Using additional datasets that are less related to the target task can also help understanding the weaknesses and strengths of a model. Using other datasets is then particularly useful when little data is available for the real target tasks. To perform well with those datasets, using supervised learning while training on similar data is a winning strategy, even though it is worthwhile to keep Goodhart's law (*When a measure becomes a target, it ceases to be a good measure*) in mind.

4.4 Evaluation benchmarks

Task specific evaluations have been widely used but different methodologies (e.g. cross validation splits, hyperparameter tuning) can make comparisons difficult. Evaluation benchmarks emerge spontaneously when authors compare themselves on the same set of tasks and try to have comparable methodologies.

These benchmarks are a convenient tool, even though they are an additional layer of abstraction that can lead to interpretation errors. A telling example is the use of the SUBJ dataset (Pang and Lee, 2004) in SentEval. In Pang and Lee (2004), the authors properly describe the dataset they use for subjectivity analysis. They extracted sentences or phrases from movie abstracts and from movie reviews; they consider text from abstract as *objective* and text from reviews as *subjective*. Thus, SUBJ is arguably an abstract/review discrimination task and not a subjectivity analysis task, but the SentEval paper's only description of the dataset is *subjectivity/objectivity (SUBJ)*. In this section, we will present popular evaluation benchmarks.

4.4.1 SentEval

Kiros et al. (2015) gathered a set of tasks and tools for evaluation of understanding. These tasks were compiled in the SentEval (Conneau and Bordes, 2017) evaluation suite designed for automatic evaluation of pre-trained sentence embeddings.

SentEval tasks are mostly based on sentiment analysis, sentence similarity and natural language inference, and the framework forces the user to provide a sentence encoder that is not finetuned during the evaluation.

Table 4.4 displays the tasks used in SentEval. The tasks selection is arguably centered towards semantics since SentEval consist mainly of similarity/relatedness estimation tasks (STS, MRPC, SICK-R) and natural language inference tasks (SICK-E, MNLI).

dataset	categories	exemple&class		N_{train}
MR	sentiment (movie)	“bland but harmless”	neg	11k
SST	sentiment (movie)	“a quiet , pure , elliptical film ”	pos	70k
CR	sentiment (products)	“the customer support is pathetic.”	neg	3k
SUBJ	subjective/objective	“it is late at night in a foreign land”	obj	10k
MPQA	opinion polarity	“would like to tell”	pos	11k
TREC	question-type	“What are the twin cities ?”	LOC:city	6k
MRPC	paraphrase	“i ’m never going to [...]”/“i am [...]”	paraphrase	4k
STS14	similarity	“a man is running.”/“a man is mooing.”	1.0	4k
SICK-E	inference relation	“a man is puking”/“a man is eating”	neutral	4k
SICK-R	relatedness score	“a man is puking”/“a man is eating”	2.30	4k
COCO	rank	(image)/“A man in a suite sits at a table”	rank 3	565k
SNLI	inference relation	“dog leaps out”/“a dog jumps”	entailment	570k

Table 4.4 – SentEval classification datasets

SentEval internally composes the sentence embeddings of two sentence when the task deals with sentence pairs (e.g. MRPC); we will discuss the implications of this strategy in the upcoming chapter 5.

4.4.2 GLUE

Wang et al. (2018) propose to evaluate language understanding with less constraints than SentEval, allowing users not to rely on explicit sentence embedding based models. GLUE’s 9 tasks are classification or regression based, and are carried out for sentences or sentence pairs. Additionally, they propose *diagnostic* NLI tasks where various annotated linguistic phenomena occur, which could be necessary to make the right predictions, as in Poliak et al. (2018b). Table 4.5 provides an overview of the GLUE benchmark tasks.

dataset	categories	exemple&class		N_{train}
MNLI	inference relation	“they renewed inquiries”/“they asked again”	entailment	391k
QNLI	inference relation	“Who took over Samoa?”/“Sykes–Picot Agreement.”	entailment	105k
MNLI	inference relation	“they renewed inquiries”/“they asked again”	entailment	391k
SST	sentiment (movie)	“a quiet , pure , elliptical film ”	pos	70k
STSB	similarity	“a man is running.”/“a man is mooing.”	1.0	1k
CoLA	linguistic acceptability	“They drank the pub.”	not-acceptable	8k
QQP	paraphrase	“Is there a soul?”/“What is a soul?”	Non-duplicate	364k
RTE	inference relation	“Oil prices fall back as Yukos oil threat lifted”/“Oil prices rise.”	not-entailment	2k
WNLI	inference relation	“The fish ate the worm. It was tasty.”/“The fish was tasty.”	entailment	643

Table 4.5 – GLUE classification datasets

4.4.3 XNLI

Another dimension of universality of sentence encoders model is the ability to process multiple languages. English is seen as a default but might not be very representative (Bender, 2011) of other languages. Tools for multilingual evaluations are more restricted, and mostly semantic: XNLI (Conneau et al., 2018b) was proposed as an evaluation suite for multilingual sentence representation learning.

4.5 Representation learning models

The several text encoders that were proposed often rely on the building blocks we cited previously: an elementary unit representation (pretrained word embeddings or subwords embeddings learned from scratch), an encoder architecture (LSTM, CNN, Transformers), and a training signal (Language modeling or Natural Language Inference). These choices, alongside experimental parameters provide an overall characterization of the following encoders.

4.5.1 FastText

A strong baseline for text encoding is FastText. FastText refers to either a word embedding method (FastText embedding), or a classification method (FastText classifier). Both are based on representations with the average of token embeddings. Thus, FastText text classifier is just an embedding layer followed by a pooling layer. In the FastText classifier, texts are represented with the average of the word/word ngram embeddings, with the optional addition of character ngram embeddings contained in the words. These embeddings can optionally have been pretrained with the FastText embedding method (which boils down to Word2Vec but with integration of ngram embeddings in word representations).

The composition is additive even though the word ngram embeddings can only crudely process compositions. Still, this model is quite fast and outperformed more complex architectures (Joulin et al., 2017).

4.5.2 SkipThought

SkipThought (Kiros et al., 2015) was the first expressive and universality claiming sentence encoder. It uses Word2Vec embeddings as input and a bidirectional GRU encoder. Building upon Le and Mikolov (2014), it leverages a sentence level distributional hypothesis described in subsection 4.2.2.2. The training examples are unsupervisedly derived from the BookCorpus dataset. A GRU encoder maps a sentence s_i into a vector $h_i = GRU_{encoder}(s_i)$. This vector is fed to two distinct GRUs that perform language modeling on respectively s_{i-1} and s_{i+1}, the sentences that are contiguous to s_i. h_i is fed as additional input to these GRUs to help the language modeling task; the joint training of the three GRUs incentivizes $GRU_{encoder}$ to encode useful information about its input sentence.

This sentence encoder has been widely use and is allegedly able to leverage the complexity of its GRU encoder, but has been outperformed by the simpler FastText model in several tasks.

4.5.3 InferSent

Conneau and Bordes (2017) introduced the idea of using supervised learning as a training signal for generalizable sentence representation. They experimented with using sentiment analysis or natural language inference as supervised pretraining and found that natural language inference transfers better even to sentiment analysis tasks. InferSent is a bidirectional LSTM model trained on the concatenation of SNLI and MNLI datasets.

4.5.4 DisSent

Nie et al. (2019) proposed to use discourse marker prediction described in section 4.2.4 instead of natural language tasks. Their results are not quite on par with InferSent on the SentEval benchmark, but required no human annotations and yield superior results in discourse relation prediction.

4.5.5 BERT

BERT makes use of the transformer architecture, as well as two different training signals. One of them is a masked language modeling: an input text extracted from a large corpus is fed to a transformer with some masked tokens, as described in section 4.2.1

The other is prediction of sentence contiguity as defined in section 4.2.2; even though upon inspection of the authors' implementation, what they call sentences are chunks of text containing multiple sentences.

A peculiarity of BERT is that it uses left and right context jointly for the prediction of masked words, as opposed to previous models (Peters et al., 2018b; Radford, 2018). This allows richer contextualized representations and incentivizes the model to compose the left and right contexts. BERT provides a structure that allows general text representation (contextualized word embedding, prediction of relation between texts). As opposed to the other cited models, BERT does not provide explicit sentence embeddings, as described in section

3.3. With the previously cited models, in order to predict a relation between two sentences, both embeddings have to be computed and fed to another composition model that composes the sentence embeddings, as described in chapter 5. BERT learns composition and text representation jointly which makes it particularly suitable to multitask finetuning (Liu et al., 2019).

CHAPTER 5

EXPRESSIVITY OF EMBEDDING COMPOSITIONS

5.1 Motivation

Predicting relations between textual units is a widespread problem, essential for discourse analysis, dialog systems, information retrieval, or paraphrase detection among others. Since relation prediction often requires a form of understanding, it can also be used as a proxy to train and evaluate transferable sentence representations.

As seen in section 4.2, several tasks that are useful to build sentence representations are derived directly from text structure, with training data that can be obtained without human annotation: sentence order prediction (Logeswaran et al., 2016; Jernite et al., 2017), the prediction of previous and subsequent sentences (Kiros et al., 2015; Jernite et al., 2017), or the prediction of explicit discourse markers between sentence pairs (Nie et al., 2019; Jernite et al., 2017). Human labeled relations between sentences can also be used for that purpose, e.g. inferential relations (Conneau and Bordes, 2017).

While most work on sentence similarity estimation, entailment detection, answer selection, or discourse relation prediction seemingly uses task-specific models, they all involve predicting whether a relation R holds between two sentences s_1 and s_2. This genericity has

been noticed in the literature before (Baudiš et al., 2016) and it has been leveraged for the evaluation of sentence embeddings within the SentEval framework (Conneau and Bordes, 2017).

A straightforward way to predict the probability of (s_1, R, s_2) being true is to represent s_1 and s_2 with d-dimensional embeddings h_1 and h_2, and to compute sentence pair features $f(h_1, h_2)$, where f is a composition function (e.g. concatenation, product, ...). A softmax classifier g_θ can learn to predict R with those features. $g_\theta \circ f$ can be seen as a reasoning based on the content of h_1 and h_2 (Socher et al., 2013).

In the SentEval evaluation suite (described in 4.4.1), users provide sentence embedding models that are composed in the framework to evaluate the quality of embeddings though performance scores in various tasks. So the compositions that are used need to be expressive enough for the evaluation to make sense. What if even the best possible sentence embeddings could not allow a model to reach human level performance ? This question has not yet been addressed in previous work to our knowledge and will be the subject of the present chapter.

Our contributions are as follows:

- we review composition functions used in textual relational learning, propose expressiveness requirements and show that existing functions are lacking in that respect (section 5.2);
- we draw analogies with existing SRL models (section 5.3) and design new compositions inspired from SRL (section 5.4) that are more expressive;
- we perform extensive experiments to test composition functions and show that some of them can improve the learning of representations and their downstream uses, and also impact evaluation (section 5.6).

In this chapter we focus on sentence embeddings as a case study, although our framework can straightforwardly be applied to other levels of language granularity (such as words, clauses, or documents).

5.2 Composition functions for relation prediction

We review here popular composition functions used for relation prediction based on sentence embeddings. Ideally, they should simultaneously fulfill the following minimal requirements:

- make use of interactions between representations of sentences to relate;
- allow for the learning of asymmetric relations (e.g. entailment, order);
- be usable with high dimensionalities; high dimensionality can improve model capacity but comes with operational constraints[1].

Additionally, if the main goal is transferable sentence representation learning, compositions should also incentivize gradually changing sentences to lie on a linear manifold, since transfer usually uses linear models. Figure 5-1 illustrates this with a manifold where sentences keep the same content while changing their style towards more formality when following the arrow direction. If sentences lie on such a non-linear manifold, a linear model that predicts formality level will be better than random but still incapable of perfect accuracy.

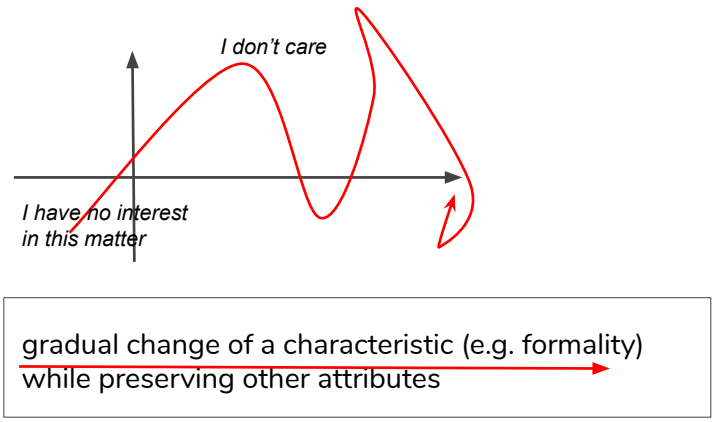


Figure 5-1 – Non linear manifold in which sentences become less and less formal

Another use case of transfer can be learning/evaluation of transferable *relation* representations. Concretely, a sentence encoder and f can be trained on a source task, and $f(h_1, h_2)$

[1] parameters of θ and f should fit in GPU memory

can be used as features for transfer in target tasks. In that case, the geometry of the sentence embedding space is less relevant, as long as the $f(h_1, h_2)$ space works well for transfer learning. Our evaluation will cover both cases.

A straightforward instantiation of f is concatenation (Hooda and Kosseim, 2017):

$$f_{[,]}(h_1, h_2) = [h_1, h_2] \tag{5.1}$$

However, strong interactions between s_1 and s_2 cannot be modeled with $f_{[,]}$ followed by a softmax regression. We define strong interaction as follows: When h_1 and h_2 strongly interact, a change in h_1 should be able to influence how a change of h_2 changes $P(s_1, R, s_2)$. That is, $\frac{\partial}{\partial h_1}\frac{\partial}{\partial h_2}P(s_1, R, s_2)$ should not be always zero.

Consider a paraphrase detection task: given a sentence s_1, the sentence s_2 maximizing the probability of s_2 being an s_1 paraphrase should depend on s_1. s_1 cannot be a paraphrase on its own. But it is not the case if the probability computed with concatenation is followed by softmax regression, since $P(s_1, R, s_2)$ depends on $h_1.W^R_{[0:n]} + h_2.W^R_{[n:2n]}$ [2], where W_R denotes softmax weights for relation R and n is the size of h. This expression is a sum over two independent terms. How h_1 should change to change $P(s_1, R, s_2)$ only depends on W. The effects of those considerations have been noticed experimentally in Levy et al. (2015) regarding lexical relations.

To promote interactions between h_1 and h_2, element-wise product has been used by Baudiš et al. (2016):

$$f_{\odot}(h_1, h_2) = h_1 \odot h_2 \tag{5.2}$$

Absolute difference is another solution for sentence similarity (Mueller and Thyagarajan, 2016), and its element-wise variation may equally be used to compute informative features:

$$f_{-}(h_1, h_2) = |h_1 - h_2| \tag{5.3}$$

The latter two were combined into a popular instantiation, sometimes referred as *heuristic*

[2] It also depends on the score for other relations than R due to the softmax normalization, but the softmax still does not induce a strong interaction.

matching (Tai et al., 2015; Kiros et al., 2015; Mou et al., 2016):

$$f_{\odot -}(h_1, h_2) = [h_1 \odot h_2, |h_2 - h_1|] \tag{5.4}$$

Although possibly effective for certain similarity tasks, $f_{\odot -}$ is symmetrical, and should be a poor choice for tasks like entailment prediction or prediction of discourse relations. For instance, if R_e denotes entailment and (s_1, s_2)= ("It just rained", "The ground is wet"), (s_1, R_e, s_2) should hold but not (s_2, R_e, s_1). The $f_{\odot -}$ composition function is nonetheless used to train models with NLI (Conneau and Bordes, 2017) or discourse relation prediction (Nie et al., 2019). This composition is also used in all works using the SentEval evaluation suite, which should be concerning since it might not allow any sentence embeddings to be used to perform some tasks correctly.

Sometimes $[h_1, h_2]$ is concatenated to $f_{\odot -}(h_1, h_2)$ (Ampomah et al., 2016; Conneau and Bordes, 2017). While the resulting composition is asymmetrical, the asymmetrical component involves no interaction as noted previously. So it can deal with cases where asymmetry is needed without needing strong interaction, but not cases needing both. We note that this composition is very commonly used. On the SNLI benchmark,[3] 12 out of the 25 listed sentence embedding based models use it, and 7 use a weaker form (e.g. omitting $f_{\odot}$).

The outer product $\otimes$ has been used instead for asymmetric multiplicative interaction (Jernite et al., 2017):

$$f_{\otimes}(h_1, h_2) = h_1 \otimes h_2 \text{ where } (h_1 \otimes h_2)_{i,j} = h_{1i} h_{2j} \tag{5.5}$$

This formulation is expressive but it forces g_θ to have d^2 parameters per relation, which is prohibitive when there are many relations and d is high.

The problems outlined above are well known in SRL. Thus, existing compositions (except $f_{\otimes}$) can only model relations superficially for tasks currently used to train state of the art sentence encoders, like NLI or discourse connectives prediction.

[3] nlp.stanford.edu/projects/snli/, as of February 2019.

Model	Scoring function	Parameters
Unstructured	$\|\|e_1 - e_2\|\|_p$	-
TransE	$\|\|e_1 + w_r - e_2\|\|_p$	$w_r \in \mathbb{R}^d$
RESCAL	$e_1^T W_r e_2$	$W_r \in \mathbb{R}^{d^2}$
DistMult	$< e_1, w_r, e_2 >$	$w_r \in \mathbb{R}^d$
ComplEx	Re $< e_1, w_r, \overline{e_2} >$	$w_r \in \mathbb{C}^d$

Table 5.1 – Selected relational learning models. Unstructured is from (Bordes et al., 2013a), TransE from (Bordes et al., 2013b), RESCAL from (Nickel et al., 2011), DistMult from (Yang et al., 2015) and (Trouillon et al., 2016). Following the latter, $< a, b, c >$ denotes $\sum_k a_k b_k c_k$. Re(x) is the real part of x, and p is commonly set to 1.

5.3 Statistical Relational Learning models

In this section we introduce the context of statistical relational learning (SRL) and relevant models. Recently, SRL has focused on efficient and expressive relation prediction based on embeddings and we believe that its techniques are overlooked in NLP. A core goal of SRL (Getoor and Taskar, 2007) is to induce whether a relation R holds between two arbitrary entities e_1, e_2. As an example, we would like to assign a score to (e_1, R, e_2) = (Paris, LOCATED_IN, France) that reflects a high probability. In embedding-based SRL models, entities e_i have vector representations in $\mathbb{R}^d$ and a scoring function reflects truth values of relations. The scoring function should allow for relation-dependent reasoning over the latent space of entities. Scoring functions can have relation-specific parameters, which can be interpreted as relation embeddings. Table 5.1 presents an overview of a number of state of the art relational models. We can distinguish two families of models: subtractive and multiplicative.

The TransE scoring function is motivated by the idea that translations in latent space can model analogical reasoning and hierarchical relationships. Dense word embeddings trained on tasks related to the distributional hypothesis naturally allow for analogical reasoning with translations without explicit supervision (Mikolov et al., 2013). TransE generalizes the older Unstructured model. We call them subtractive models.

The RESCAL, Distmult, and ComplEx scoring functions can be seen as dot product matching between e_1 and a relation-specific linear transformation of e_2 (Liu et al., 2017). This

transformation helps checking whether e_1 matches with some aspects of e_2. RESCAL allows for a full linear mapping $W_r e_2$ but has a high complexity, while Distmult is restricted to a component-wise weighting $w_r \odot e_2$. ComplEx has fewer parameters than RESCAL but still allows for the modeling of asymmetrical relations. As shown in Liu et al. (2017), ComplEx boils down to a restriction of RESCAL where W_r is a block diagonal matrix. These blocks are 2-dimensional, antisymmetric and have equal diagonal terms. Using such a form, even and odd indexes of e's dimensions play the roles of real and imaginary numbers respectively. The ComplEx model (Trouillon et al., 2016) and its variations (Lacroix et al., 2018) yield state of the art performance on knowledge base completion on numerous evaluations.

5.4 Embeddings composition as SRL models

We claim that several existing NLP models (Conneau and Bordes, 2017; Nie et al., 2019; Baudiš et al., 2016) boil down to SRL models where the sentence embeddings (h_1, h_2) act as entity embeddings (e_1, e_2). This framework is depicted in figure 5-2.

Recent popular models (Chen et al., 2017b; Seo et al., 2017; Gong et al., 2018; Radford, 2018; Devlin et al., 2019) do not rely on explicit sentence encodings to perform relation prediction. They combine information of input sentences at earlier stages, using conditional encoding or cross-attention. There is however no straightforward way to derive transferable sentence representations in this setting. Thus, we leave them out for the moment but we will discuss them in section 5.8, as they sometimes make use of composition functions, so our work could still be relevant to them in some respect.

In this section we will make a link between sentence composition functions and SRL scoring functions, and propose new scoring functions drawing inspiration from SRL.

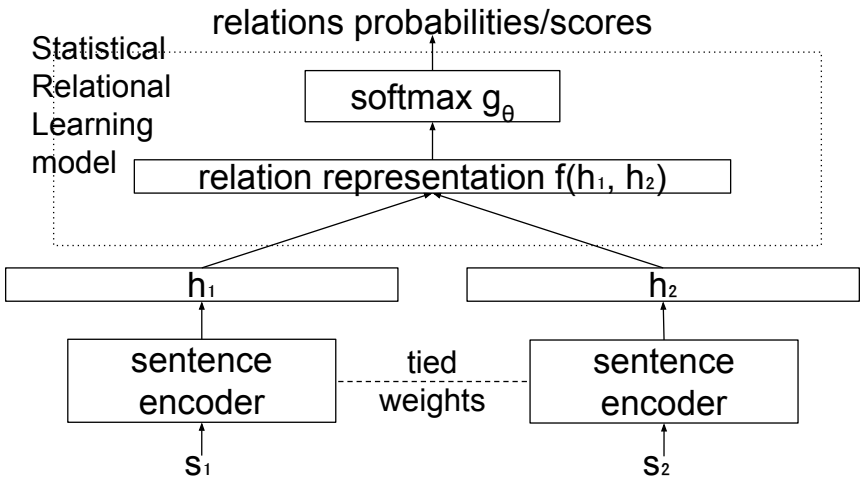


Figure 5-2 – Implicit SRL model in text relation prediction

5.4.1 Linking composition functions and SRL models

The composition function $f_\odot$ from equation 5.2 followed by a softmax regression yields a score whose analytical form is identical to the Distmult model score described in section 5.3. Let θ_R denote the softmax weights for relation R. The logit score for the truth of (s_1, R, s_2) is $f(h_1, h_2)\theta_R = (h_1 \odot h_2)\theta_R$ which is equal to the Distmult scoring function $< h_1, \theta_R, h_2 >$ if h_1, h_2 act as entities embeddings and θ_R as the relation weight w_R .

Similarly, the composition f_- from equation 5.3 followed by a softmax regression can be seen as an element-wise weighted score of Unstructured (both are equal if softmax weights are all unitary).

Thus, $f_{\odot-}$ from 5.4 (with softmax regression) can be seen as a weighted ensemble of Unstructured and Distmult. These two models are respectively outperformed by TransE and ComplEx on knowledge base link prediction by a large margin (Trouillon et al., 2016; Bordes et al., 2013a). We therefore propose to change the Unstructured and Distmult in $f_{\odot-}$ such that they match their respective state of the art variations in the following sections. We will also show the implications of these refinements.

5.4.2 Casting TransE as a composition

Simply replacing $|h_2 - h_1|$ with

$$f_t(h_1, h_2) = |h_2 - h_1 + t|, \text{ where } t \in \mathbb{R}^d \tag{5.6}$$

would make the model analogous to TransE. t is learned and is shared by all relations. A relation-specific translation t_R could be used but it would make f relation-specific. Instead, here, each dimension of $f_t(h_1, h_2)$ can be weighted according to a given relation. Non-zero t makes f_t asymmetrical and also yields features that allow for the checking of an analogy between s_1 and s_2. Sentence embeddings often rely on pre-trained word embeddings which have demonstrated strong capabilities for analogical reasoning. Some analogies, such as *part-whole*, are computable with off-the-shelf word embeddings (Chen et al., 2017a) and should be very informative for natural language inference tasks. As an illustration, let us consider an artificial semantic space (depicted in figures 5-3a and 5-3b) where we posit that there is a "to the past" translation t so that $h_1 + t$ is the embedding of a sentence s_1 changed to the past tense. Unstructured is not able to leverage this semantic space to correctly score $(s_1, R_{to\_the\_past}, s_2)$ while TransE is well tailored to provide highest scores for sentences near $h_1 + \hat{t}$ where $\hat{t}$ is an estimation of t that could be learned from examples.

5.4.3 Casting ComplEx as a composition

Let us partition h dimensions into two equally sized sets $\mathcal{R}$ and $\mathcal{I}$, e.g. even and odd dimension indices of h. We propose a new function $f_{\mathbb{C}}$ as a way to fit the ComplEx scoring function into a composition function.

$$f_{\mathbb{C}}(h_1, h_2) = [h_1^{\mathcal{R}} \odot h_2^{\mathcal{R}} + h_1^{\mathcal{I}} \odot h_2^{\mathcal{I}}, h_1^{\mathcal{R}} \odot h_2^{\mathcal{I}} - h_1^{\mathcal{I}} \odot h_2^{\mathcal{R}}] \tag{5.7}$$

$f_{\mathbb{C}}(h_1, h_2)$ multiplied by softmax weights θ_r is equivalent to the ComplEx scoring function $\text{Re} < h_1, \theta_r, \overline{h_2} >$. The first half of θ_r weights corresponds to the real part of ComplEx relation weights while the last half corresponds to the imaginary part.

$f_{\mathbb{C}}$ is to the ComplEx scoring function what $f_{\odot}$ is to the DistMult scoring function. Intuitively, ComplEx is a minimal way to model interactions between distinct latent dimensions while Distmult only allows for identical dimensions to interact.

Let us consider a new artificial semantic space (shown in figures 5-3c and 5-3d) where the first dimension is high when a sentence means that it just rained, and the second dimension is high when the ground is wet. Over this semantic space, Distmult is only able to detect entailment for paraphrases whereas ComplEx is also able to naturally model that ("it just rained", $R_{entailment}$, "the ground is wet") should be high while its converse should not.

We also propose two more general versions of $f_{\mathbb{C}}$:

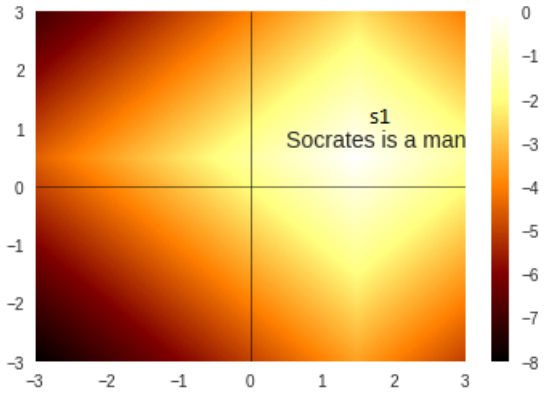
$$f_{\mathbb{C}^\alpha}(h_1, h_2) = [h_1^{\mathcal{R}} \odot h_2^{\mathcal{R}}, h_1^{\mathcal{I}} \odot h_2^{\mathcal{I}}, h_1^{\mathcal{R}} \odot h_2^{\mathcal{I}} - h_1^{\mathcal{I}} \odot h_2^{\mathcal{R}}] \tag{5.8}$$

$$f_{\mathbb{C}^\beta}(h_1, h_2) = [h_1^{\mathcal{R}} \odot h_2^{\mathcal{R}}, h_1^{\mathcal{I}} \odot h_2^{\mathcal{I}}, h_1^{\mathcal{R}} \odot h_2^{\mathcal{I}}, h_1^{\mathcal{I}} \odot h_2^{\mathcal{R}}] \tag{5.9}$$

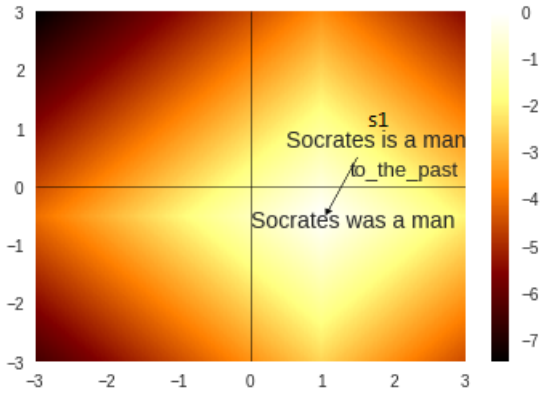
$f_{\mathbb{C}^\alpha}$ can be seen as Distmult concatenated with the asymmetrical part of ComplEx and $f_{\mathbb{C}^\beta}$ can be seen as RESCAL with unconstrained block diagonal relation matrices. These compositions have higher dimensionality but this is not as problematic as in SRL (Freebase contains $35k$ relation types which can make it hard to learn high dimensional relation embeddings). NLP problems tend to have a moderate number of relations and we can afford to use slightly more relation parameters.

5.5 On the evaluation of relational models

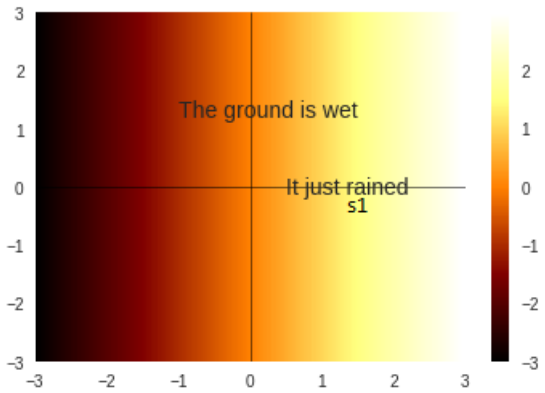
The SentEval framework (Conneau and Bordes, 2017) provides a general evaluation for transferable sentence representations, with open source evaluation code. One only needs



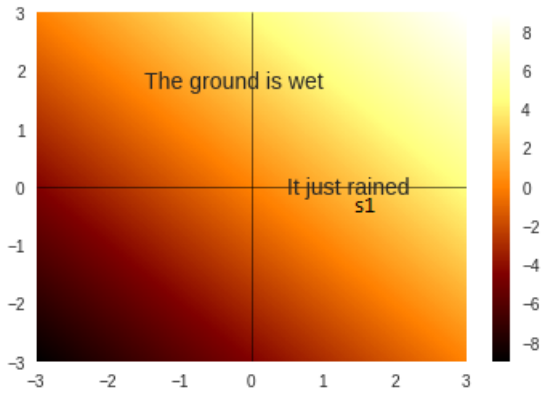
(a) Score map of $(s_1, R_{to\_the\_past}, s_2)$ over possible sentences s_2 using Unstructured composition.



(b) Score map of $(s_1, R_{to\_the\_past}, s_2)$ over possible sentences s_2 using TransE composition.



(c) Score map of $(s_1, R_{entailment}, s_2)$ over possible sentences s_2 using DistMult composition.



(d) Score map of $(s_1, R_{entailment}, s_2)$ over possible sentences s_2 using ComplEx composition.

Figure 5-3 – Possible scoring function values according to different composition functions. s_1 and R are fixed and color brightness reflects likelihood of (s_1, R, s_2) for each position of embedding s_2. (b) and (d) are respectively more expressive than (a) and (c).

to specify a sentence encoder function, and the framework performs classification tasks or relation prediction tasks using cross-validated logistic regression on embeddings or composed sentence embeddings. Tasks include sentiment analysis, entailment, textual similarity, textual relatedness, and paraphrase detection. These tasks are a rich way to train or evaluate sentence representations since in a triple (s_1, R, s_2), we can see (R, s_2) as a label for s_1 (Baudiš et al., 2016). Unfortunately, the relational tasks hard-code the composition function from equation 5.4. From our previous analysis, we believe this composition function favors the use of contextual/lexical similarity rather than high-level reasoning and can penalize representations based on more semantic aspects. This bias could harm research

since semantic representation is an important next step for sentence embedding. Training/evaluation datasets are also arguably flawed with respect to relational aspects since several recent studies (Dasgupta et al., 2018; Poliak et al., 2018b; Gururangan et al., 2018; Glockner et al., 2018) show that InferSent, despite being state of the art on SentEval evaluation tasks, has poor performance when dealing with asymmetrical tasks and non-additive composition of words. In addition to providing new ways of training sentence encoders, we will also extend the SentEval evaluation framework with a more expressive composition function when dealing with relational transfer tasks, which improves results even when the sentence encoder was not trained with it.

5.6 Experiments

Our goal is to show that transferable sentence representation learning and relation prediction tasks can be improved when our expressive compositions are used instead of the composition from equation 5.4. We train our relational model adaptations on two relation prediction source tasks ($\mathcal{T}$), one supervised ($\mathcal{T} = NLI$) and one unsupervised ($\mathcal{T} = Disc$) described below, and evaluate sentence/relation representations on source and target tasks using the SentEval framework in order to quantify the generalization capabilities of our models. Since we use minor modifications of InferSent and SentEval, our experiments are easily reproducible.

5.6.1 Training tasks

Natural language inference ($\mathcal{T}$ = *NLI*)'s goal is to predict whether the relation between two sentences (premise and hypothesis) is *Entailment*, *Contradiction* or *Neutral*. We use the combination of SNLI dataset (Bowman et al., 2015a) and MNLI dataset (Williams et al., 2018b). We call *AllNLI* the resulting dataset of $1M$ examples. Conneau and Bordes (2017) claim that NLI data allows for universal sentence representation learning. They used the $f_{\odot,-}$ composition function with concatenated sentence representations in order to train their

name	N	task	C	representation(s) used
MR	11k	sentiment (movies)	2	h_1
SUBJ	10k	subjectivity/objectivity	2	h_1
MPQA	11k	opinion polarity	2	h_1
TREC	6k	question-type	6	h_1
SICK$_s^m$	10k	NLI	3	$f_{m,s}(h_1, h_2)$
MRPC$_s^m$	4k	paraphrase detection	2	$(f_{m,s}(h_1, h_2) + (f_{m,s}(h_2, h_1))/2$
PDTB$_s^m$	17k	discursive relation	5	$f_{m,s}(h_1, h_2)$
STS14	4.5k	similarity	-	$\cos(h_1, h_2)$

Table 5.2 – Transfer evaluation tasks. N = number of training examples; C = number of classes if applicable. h_1, h_2 are sentence representations, $f_{m,s}$ a composition function from section 5.4.

Infersent model.

We also train on the prediction of discourse markers between sentences/clauses ($\mathcal{T}$ = *Disc*). Discourse connectives make discourse relations between sentences explicit. In the sentence *I live in Paris but I'm often elsewhere*, the word *but* highlights that there is a contrast between the two clauses it connects, as introduced in section 4.2.4. We use Malmi et al.'s (2018) dataset of selected $400k$ instances with 20 discourse connectives (e.g. *however*, *for example*) with the provided train/dev/test split. This dataset has no other supervision than the list of 20 connectives. Nie et al. (2019) used $f_{\odot,-}$ concatenated with the sum of sentence representations to train their model, *DisSent*, on a similar task and showed that their encoder was general enough to perform well on SentEval tasks. They use a dataset that was, at the time of these experiments, not publicly available. We will tackle this unsupervised approach in an arguably better way in chapter 6, and deepen the comparison of NLI and marker prediction in chapter 7.

5.6.2 Evaluation tasks

Table 6.8 provides an overview of different transfer tasks that will be used for evaluation. We added another relation prediction task, the PDTB coarse-grained implicit discourse relation task, to SentEval. This task involves predicting a discursive link between two sentences among {Comparison, Contingency, Entity based coherence, Expansion, Temporal}.

We emphasized the importance of discourse in evaluation in the introduction; a more refined evaluation will be proposed in chapter 7.

We followed the setup of Pitler et al. (2009), without sampling negative examples in training. MRPC, PDTB and SICK will be tested with two composition functions: besides SentEval composition $f_{\odot,-}$, we will use $f_{\mathcal{C}^\beta,-}$ for transfer learning evaluation, since it has the most general multiplicative interaction and it does not penalize models that do not learn a translation. For all tasks except STS14, a cross-validated logistic regression is used on the sentence or relation representation. The evaluation of the STS14 task relies on Pearson or Spearman correlation between cosine similarity and the target. We force the composition function to be symmetrical on the MRPC task since paraphrase detection should be invariant to permutation of input sentences.

5.6.3 Setup

We want to compare the different instances of f. We follow the setup of Infersent (Conneau and Bordes, 2017): we learn to encode sentences into h with a bi-directional LSTM using element-wise max pooling over time. The dimension size of h is 4096. Word embeddings are fixed GloVe with 300 dimensions, trained on Common Crawl 840B.[4] Optimization is done with SGD and decreasing learning rate until convergence.

The only difference with regard to Infersent is the composition. Sentences are composed with six different compositions for training according to the following template:

$$f_{m,s,1,2}(h_1, h_2) = [f_m(h_1, h_2), f_s(h_1, h_2), h_1, h_2] \tag{5.10}$$

f_s (subtractive interaction) is in $\{f_-, f_t\}$, f_m (multiplicative interaction) is in $\{f_\odot, f_{\mathbb{C}^\alpha}, f_{\mathbb{C}^\beta}\}$. We do not consider $f_\mathbb{C}$ since it yielded inferior results in our early experiments using NLI and SentEval development sets.

$f_{m,s,1,2}(h_1, h_2)$ is fed directly to a softmax regression. Note that Infersent uses a multi-

[4] https://nlp.stanford.edu/projects/glove/

layer perceptron before the softmax, but uses only linear activations, so $f_{\odot,-,1,2}(h_1, h_2)$ is analytically equivalent to Infersent when $\mathcal{T} = NLI$.

5.6.4 Results

Models trained on natural language inference ($\mathcal{T} = NLI$)										
m,s	MR	SUBJ	MPQA	TREC	MRPC$^{\odot}_{-}$	PDTB$^{\odot}_{-}$	SICK$^{\odot}_{-}$	STS14	$\mathcal{T}$	AVG
$\odot, -$	81.2	92.7	90.4	89.6	76.1	46.7	86.6	69.5	84.2	79.1
$\alpha, -$	**81.4**	**92.8**	90.5	89.6	75.4	46.6	86.7	69.5	84.3	79.1
$\beta, -$	81.2	92.6	90.5	89.6	76	46.5	86.6	69.5	84.2	79.1
$\odot, t$	81.1	92.7	90.5	**89.7**	**76.5**	46.4	86.5	**70.0**	**84.8**	**79.2**
α, t	81.3	92.6	**90.6**	89.2	76.2	47.2	86.5	**70.0**	84.6	**79.2**
β, t	81.2	92.7	90.4	88.5	75.8	**47.3**	**86.8**	69.8	84.2	79.1

Table 5.3 – SentEval and source task evaluation results for the models trained on natural language inference ($\mathcal{T} = NLI$); AllNLI is used for training. All scores are accuracy percentages, except STS14, which is Pearson correlation percentage. AVG denotes the average of the SentEval scores.

Models trained on discourse connective prediction ($\mathcal{T} = Disc$)										
m,s	MR	SUBJ	MPQA	TREC	MRPC$^{\odot}_{-}$	PDTB$^{\odot}_{-}$	SICK$^{\odot}_{-}$	STS14	$\mathcal{T}$	AVG
$\odot, -$	**80.4**	92.7	90.2	89.5	74.5	47.3	83.2	57.9	35.7	77
$\alpha, -$	**80.4**	**92.9**	90.2	90.2	75	47.9	83.3	57.8	35.9	77.2
$\beta, -$	80.2	92.8	90.2	88.4	74.9	47.5	82.9	57.7	35.9	76.8
$\odot, t$	80.2	92.8	90.2	**90.4**	74.6	**48.5**	83.4	**58.6**	**36.1**	**77.3**
α, t	80.2	**92.9**	**90.3**	90.3	**75.1**	47.8	83.2	58.3	**36.1**	**77.3**
β, t	80.2	92.8	**90.3**	89.7	74.4	47.9	**83.7**	58.2	35.7	77.2

Table 5.4 – SentEval and source task evaluation results for the models trained on discourse connective prediction ($\mathcal{T} = Disc$). All scores are accuracy percentages, except STS14, which is Pearson correlation percentage. AVG denotes the average of the SentEval scores.

Having run several experiments with different initializations, the standard deviations between them do not seem to be negligible. We decided to take these into account when reporting scores, contrary to previous work (Kiros et al., 2015; Conneau and Bordes, 2017): we average the scores of 6 distinct runs for each task and use standard deviations under normality assumption to compute significance. Table 5.3 shows model scores for $\mathcal{T} = NLI$, while Table 5.4 shows scores for $\mathcal{T} = Disc$. For comparison, Table 5.5 shows a number of

Comparison models									
model	MR	SUBJ	MPQA	TREC	MRPC$^{\odot}_{-}$	PDTB$^{\odot}_{-}$	SICK$^{\odot}_{-}$	STS14	AVG
Infersent	81.1	92.4	90.2	88.2	76.2	46.7-	86.3	70	78.9
SkipT	76.5	93.6	87.1	92.2	73	-	82.3	29	-
BoW	77.2	91.2	87.9	83	72.2	43.9	78.4	54.6	73.6

Table 5.5 – Comparison models from previous work. InferSent represents the original results from Conneau and Bordes (2017), SkipT is SkipThought from Kiros et al. (2015), and BoW is our re-evaluation of GloVe Bag of Words from Conneau and Bordes (2017). AVG denotes the average of the SentEval scores..

	$\mathcal{T} = Disc$				$\mathcal{T} = NLI$			
m,s	MRPC$^{\beta}_{-}$	PDTB$^{\beta}_{-}$	SICK$^{\beta}_{-}$	AVG	MRPC$^{\beta}_{-}$	PDTB$^{\beta}_{-}$	SICK$^{\beta}_{-}$	AVG
$\odot, -$	74.8	48.2	83.6	68.9	**76.2**	47.2	86.9	70.1
$\alpha, -$	74.9	**49.3**	83.8	**69.3**	75.9	47.1	86.9	70
$\beta, -$	75	48.8	83.4	69.1	75.8	47	87	69.9
$\odot, t$	74.9	48.7	83.6	69.1	**76.2**	**47.8**	86.8	70.3
α, t	**75.2**	48.6	83.5	69.1	**76.2**	47.6	**87.3**	**70.4**
β, t	74.6	48.9	**83.9**	69.1	**76.2**	**47.8**	87	70.3

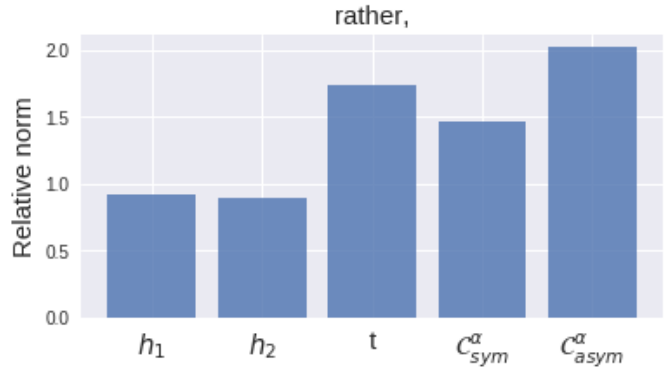
Table 5.6 – Results for sentence relation tasks using an alternative composition function ($f_{\mathbb{C}^{\beta},-}$) during evaluation. AVG denotes the average of the three tasks.

important models from previous work. Finally, in Table 5.6, we present results for sentence relation tasks that use an alternative composition function ($f_{\mathbb{C}^{\beta},-}$) instead of the standard composition function used in SentEval.

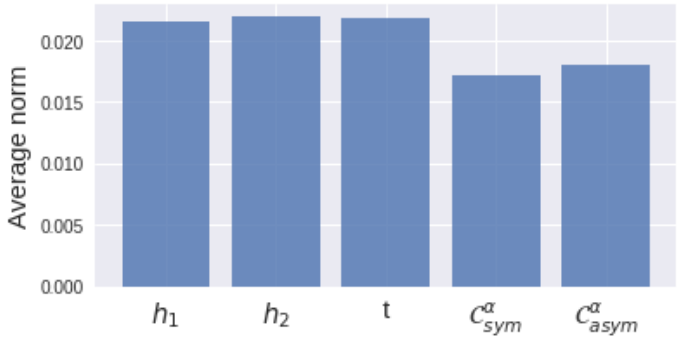
For sentence representation learning, the baseline, $f_{\odot}-$ composition already performs rather well, being on par with the InferSent scores of the original paper, as would be expected. However, macro-averaging all accuracies, it is the second worst performing model. $f_{\mathbb{C}^{\alpha},t,1,2}$ is the best performing model, and all three best models use the translation ($s = t$). On relational transfer tasks, training with $f_{\mathbb{C}^{\alpha},t,1,2}$ and using complex $\mathbb{C}^{\beta}$ for transfer (Table 5.6) always outperforms the baseline ($f_{\odot,-,1,2}$ with $\odot-$ composition in Tables 5.3 and 5.4). Averaging accuracies of those transfer tasks, this result is significant for both training tasks at level $p < 0.05$ (using Bonferroni correction accounting for the 5 comparisons). On source tasks and the average of non-relational target tasks (MR, MPQA, SUBJ, TREC), our proposed compositions are on average slightly better than $f_{\odot,-,1,2}$. Representations learned

with our proposed compositions can still be compared with simple cosine similarity: all three methods using the translational composition ($s = t$) very significantly outperform the baseline (significant at level $p < 0.01$ with Bonferroni correction) on STS14 for $\mathcal{T} = NLI$. Thus, we believe $f_{\mathcal{C}^\alpha,t,1,2}$ has more robust results and could be a better default choice than $f_{\odot,-,1,2}$ as composition for representation learning. [5]

Additionally, using $\mathbb{C}^\beta$ (Table 5.6) instead of $\odot$ (Tables 5.3 and 5.4) for transfer learning in relational transfer tasks (PDTB, MRPC, SICK) yields a significant improvement on average, even when $m = \odot$ was used for training ($p < 0.001$). Therefore, we believe $f_{\mathbb{C}^\beta,-}$ is an interesting composition for inference or evaluation of models regardless of how they were trained.



(a) Importance feature type for the *rather* relation



(b) Average importance by feature type

Figure 5-4 – A graphical representation of feature importance

[5] Note that our compositions are also beneficial with regard to convergence speed: on average, each of our proposed compositions needed less epochs to converge than the baseline $f_{\odot,-,1,2}$, for both training tasks.

5.6.5 Feature importance

As a qualitative analysis, we inspected the importance of each feature type. We used an encoder trained on $Disc$ task with $f_{\mathcal{C}^\alpha,t,1,2}$ composition and ran a logistic regression on unit scaled composed features, once again on $Disc$. $\mathcal{C}^\alpha_{sym}$ denotes the symmetrical part of $f_{\mathcal{C}^\alpha}$. We use the L_2 norm of each feature group as a measure of importance. Figure 5-4b indicates that on average, the model uses all kinds of features, especially $\mathcal{C}^\alpha_{asym}$ which we proposed. Figure 5-4a shows the norms for a particular relation class – *rather* – divided by those average norms. *rather* has the highest $\mathcal{C}^\alpha_{asym}$ norm and should be an asymmetrical connective. This makes sense, as if (s_1, R_{rather}, s_2) holds, s_2 must be a refinement of s_1 and not the converse. This example indicates that our framework can provide some insights for the description of relations.

5.7 Related work

There are numerous interactions between SRL and NLP. We believe that our framework merges two specific lines of work: relation prediction and modeling textual relational tasks.

Some previous NLP work focused on composition functions for relation prediction between text fragments, even though they ignored SRL and only dealt with word units. The use of embeddings is not necessary for composition in vector space, and previous work tackled composition in non-distributional vector spaces (Mitchell and Lapata, 2008; Bride et al., 2015). However, Word2vec (Mikolov et al., 2013) has sparked a great interest for this task with word analogies in the latent space. Levy and Goldberg (2014a) explored different scoring functions between words, notably for analogies. Hypernymy relations were also studied, by Chang et al. (2018) and Fu et al. (2014). Levy et al. (2015) proposed tailored scoring functions. Even the skipgram model (Mikolov et al., 2013) can be formulated as finding relations between context and target words. To account for asymmetry, two separate representations are used for each word (word embedding and context embedding (Torabi Asr et al., 2018). We did not empirically explore textual relational learning at the

word level, but we believe that it would fit in our framework, and could be tested in future studies. Numerous approaches (Chen et al., 2017b; Seok et al., 2016; Gong et al., 2018; Joshi et al., 2019) were proposed to predict inference relations between sentences, but don't explicitly use sentence embeddings. Instead, they encode sentences jointly, possibly with the help of previously cited word compositions, therefore it would also be interesting to try applying our techniques within their framework.

Some modeling aspects of textual relational learning have been formally investigated by Baudiš et al. (2016). They noticed the genericity of relational problems and explored multi-task and transfer learning on relational tasks. Their work is complementary to ours since their framework unifies tasks while ours unifies composition functions. Subsequent approaches use relational tasks for training and evaluation on specific datasets (Conneau and Bordes, 2017; Nie et al., 2019).

5.8 Composition in attention-based models

As seen in section 3.3, explicit sentence embeddings are not required in order to perform relation prediction between sentences. However, compositions still occur, and in many models, heuristic matching is used, in order to compute joint representations of contextualized words (Pan et al., 2018a; Chen et al., 2017b; Tay et al., 2018; Ghaeini et al., 2018).

In transformers, the different aspects of tokens, obtained by attention heads are composed through a concatenation of attention head outputs followed by a non-linear composition, as shown in figure 5-5.

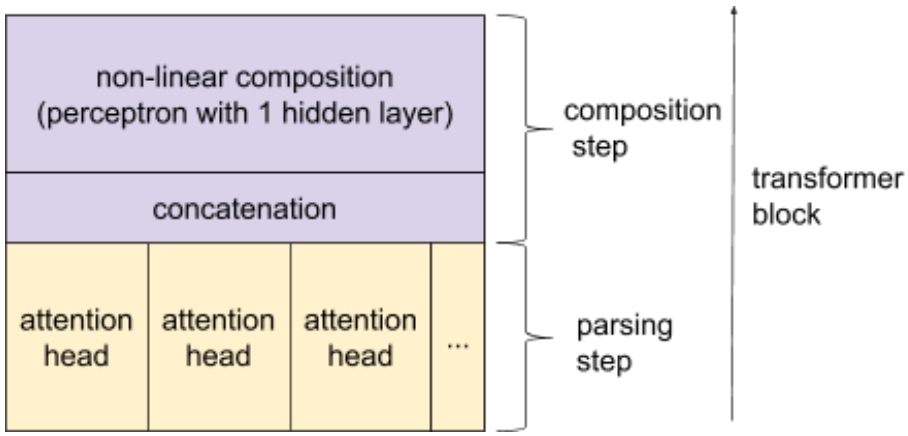


Figure 5-5 – Transformer block as successive parsing step and composition

The attention mechanism in Transformer architectures is being regarded as central, as illustrated by the *Attention is all you need* title (Vaswani et al., 2017), but we believe that what we call the composition step is equally important.

5.9 Conclusion

We have demonstrated that a number of existing models used for textual relational learning rely on composition functions that are already used in Statistical Relational Learning. By taking into account previous insights from SRL, we proposed new composition functions to fix expressivity problems and evaluated them. These composition functions are all simple to implement and we hope that it will become standard to try them on relational problems. Larger scale data might leverage these more expressive compositions, as well as more compositional, asymmetric, and arguably more realistic datasets (Dasgupta et al., 2018; Gururangan et al., 2018) could also widen the gap between symmetrical and more expressive compositions. Thus, while the improvement of more expressive compositions in our sentence embeddings setup is marginal, it is dependant on sentence encoders, training signals, and evaluation. Our compositions can also be helpful to improve interpretability of embeddings, since they can help measure relation prediction asymmetry. Analogies through translations helped interpreting word embeddings, and perhaps analyzing learned t translation depending on different relations could help interpreting sentence embeddings.

Finally, our analysis can also be transferred to Transformers and explain their expressiveness.

CHAPTER 6

MINING DISCOURSE MARKERS FOR UNSUPERVISED SENTENCE REPRESENTATION LEARNING

6.1 Motivation

Discourse markers are a common language device used to make explicit the semantic and/or pragmatic relationship between clauses or sentences.

For example, the marker *so* in sentence 6.1 indicates that the second clause is a consequence of the first.

We're standing in gasoline, so you should not smoke (6.1)

As such, discourse markers indicate how a sentence contributes to the meaning of a text. Because of this, they provide an appealing supervision signal for sentence representation learning based on language use. Fraser (1996) theorized a class of expressions called pragmatic markers that include discourse markers (see quote 6-4).

> “ [...] pragmatic markers, taken to be separate and distinct from the propositional content of the sentence, are the linguistically encoded clues which signal the speaker's potential communicative intentions ”

Quote. 6-4 – quote from *Pragmatic Markers* (Fraser, 1996)

As such, discourse markers can also be considered as noisy labels for various semantic tasks, such as entailment ($c =$ *therefore*), subjectivity analysis ($c =$ *personally*), sentiment analysis ($c =$ *sadly*), similarity ($c = similarly$), typicality ($c = curiously$), temporal order ($c = then$), frequency ($c = sometimes$), or importance ($c = mostly$). A wide variety of discourse usages would be desirable in order to learn general sentence representations. Extensive research in linguistics has resulted in elaborate discourse marker inventories for many languages.[1] These inventories were created by manual corpus exploration or annotation of small-scale corpora: the largest annotated corpus, the English PDTB consists of a few tens of thousands examples, and provides a list of about 100 discourse markers, organized in a number of categories.

However, previous work on sentence representation learning with discourse markers makes use of even more restricted sets of discourse markers, as shown in table 6.1. Jernite et al. (2017) use 9 categories as labels, accounting for 40 discourse markers in total. It should be noted that the aggregate labels do not allow for any fine-grained distinctions; for instance, the TIME label includes both *now* and *next*, which is likely to impair the supervision. Moreover, discourse markers may be ambiguous; for example, *now* can be used to express contrast. On the other hand, Nie et al. (2019) make use of 15 discourse markers. These classes do not account for the variety of phenomena outlined in the previous paragraph.

That low diversity of markers is amplified by data imbalance, since among the 15 markers used by Nie et al. (2019), 5 markers are accounting for more than 80% of their training

[1] See for instance a sample of language on the Textlink project website: http://www.textlink.ii.metu.edu.tr/dsd-view

data, as illustrated in figure 6-7.

author	discourse markers / classes	classes	markers
Jernite et al. (2017)	ADDITION, CONTRAST, TIME, RESULT, SPECIFIC, COMPARE, STRENGTH, RETURN, RECOGNIZE	9	40
Nie et al. (2019)	*and, but, because, if, when, before, though, so, as, while, after, still, also, then, although*	15	15
current work	*later, often, understandably, gradually, or, ironically, namely, . . .*	174	174

Table 6.1 – Discourse markers or classes used by previous work on unsupervised representation learning

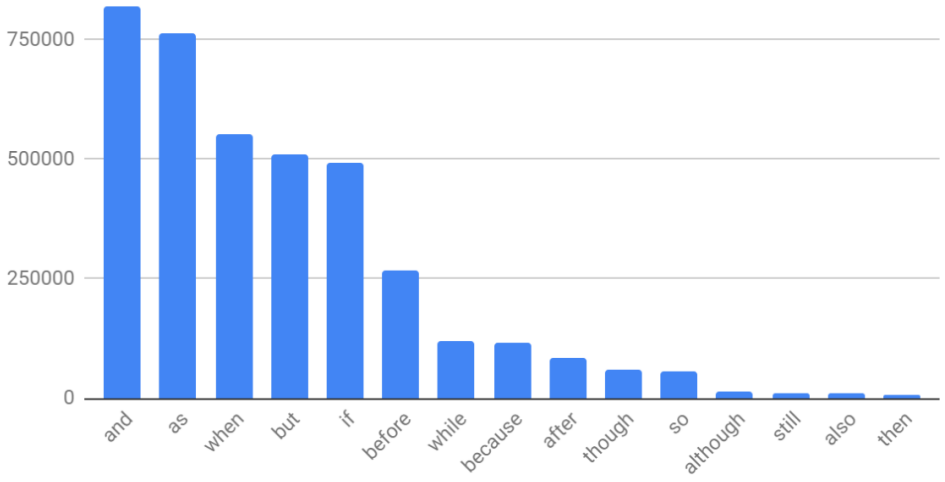


Figure 6-7 – Frequency distribution of markers of discourse markers in Nie et al. (2019)

Furthermore, they only select pairs matching a dependency pattern manually specified for each marker in order to ensure the quality of their examples, which makes their method tedious and still dependent on dependency parsers. However, *and, as, when* and *if* arguably don't appear in the same syntactic context. These examples constitute a large part of the training data, which might be a strong incentive for the model to learn syntactic cues rather than using semantic or pragmatic cues.

As such, both of these studies use a restricted or impoverished set of discourse markers; they also both use the BookCorpus dataset, whose size ($4.7M$ sentences that contain a

discourse marker, according to Nie et al., 2019) is prohibitively small for the prediction of rare discourse markers.

In the context of the present work, we use web-scale data in order to explore the prediction of a wide range of discourse markers, with totally balanced frequency distributions, along with an application to sentence representation learning. We use English data for the experiments, but the same method could be applied to any language that bears a typological resemblance with regard to discourse usage, and has sufficient amounts of textual data available (e.g. German or French). Inspired by recent work (Dasgupta et al., 2018; Poliak et al., 2018b; Glockner et al., 2018) on the unexpected properties of recent manually labelled datasets (e.g. SNLI), we will also analyze our dataset to check whether labels are easy to guess, and whether the proposed model architectures make use of high-level reasoning for their predictions. Our contributions are as follows:

- we propose a simple and efficient method to discover new discourse markers, and present a curated list of 174 markers for English;
- we provide evidence that many connectives can be predicted with only simple lexical features;
- we investigate whether relation prediction actually makes use of the relation between sentences;
- we carry out extensive experiments based on the Infersent/SentEval framework.

6.2 Discovering discourse markers

Our goal to collect unambiguous instances of potential discourse markers. To do so, previous work used heuristics based on specific constructs, especially syntactic patterns for intra-sentential relations, based on a fixed list of manually collected discourse markers. Since we focus on sentence representations, we limit ourselves to discourse arguments that are well-formed sentences, thus also avoiding clause segmentation issues.

6.2.1 Comparison to previous work

Following a heuristic from Rutherford and Xue (2015), also considered by Malmi et al. (2018) and Jernite et al. (2017), we collect pairs of sentences (s_1, s_2) where s_2 starts with marker c. We only consider the case where c is a single word, as detecting longer adverbial constructions is more difficult. We remove c from the beginning of s_2 and call the resulting sentence s_2'. Malmi et al. (2018) make use of a list of the 80 most frequent discourse markers in the PDTB in order to extract suitable sentence pairs. We stay faithful to Rutherford and Xue (2015)'s heuristic, as opposed to Malmi et al. (2018) and Jernite et al. (2017): if s_2 starts with c followed by a comma, and c is an adverbial or a conjunction, then it is a suitable candidate. By limiting ourselves to sentences that contain a comma, we are likely to ensure that s_2' is meaningful and grammatical. As opposed to all the cited work mentioned above, we do not restrict the pattern to a known list of markers, but try to collect new reliable cues.

This pattern is obviously restrictive, since discourse markers often appear at the clausal level (e.g. *I did it but now I regret it*). But clauses are not meant to be self contained, and it is not obvious that they should be included in a dataset for *sentence* representation learning. At the same time, one could easily think of cases where c is not a discourse marker, e.g. (s_1, s_2)= ("It's cold.", "Very, very cold."). However, these uses might be easily predicted with shallow language models. In the next section, we use the proposed method for the discovery of discourse markers, and we investigate whether the resulting dataset leads to improved model performance.

s1	Paul Prudhomme's Louisiana Kitchen created a sensation when it was published in 1984.
c	happily,
s2'	This family collective cookbook is just as good

Table 6.2 – Sample from our *Discovery* dataset

6.2.2 Methodology

We use sentences from the Depcc corpus (Panchenko et al., 2017), which consists of English texts harvested from commoncrawl web data. We sample 8.5 billion consecutive sentence pairs from the corpus. We keep 53% of sentence pairs that contain between 3 and 32 words, have a high probability of being English ($> 75\%$) using FastText langid from Grave et al. (2018), have balanced parentheses and quotes, and are mostly lowercase. We use NLTK (Bird et al., 2009) as sentence tokenizer and NLTK PerceptronTagger as part of speech tagger for adverb recognition. In addition to our automatically discovered candidate set, we also include all (not necessarily adverbial) PDTB discourse markers that are not induced by our method. Taking this into account, 3.77% of sentence pairs contained a discourse marker candidate, which is about 170M sentence pairs. An example from the dataset is shown in table 6.2. We only keep pairs in which the discourse marker occurs at least 10K times. We also subsample pairs so that the maximum occurrence count of a discourse marker is 200K. The resulting dataset contains 19M pairs.

We discovered 243 discourse marker candidates. Figure 6-8 shows their frequency distributions. As expected, the most frequent markers dominate the training data, but when a wide range of markers is included, the rare ones still contribute up to millions of training instances. Out of the 42 single-word PDTB markers that precede a comma, 31 were found by our procedure. Some markers are missing because of NLTK errors, which mainly result from morphological issues.[2]

6.2.3 Controlling for shallow features

As previously noted, some candidates discovered by our rule may not be actual discourse markers. In order to discard them, we make the hypothesis that actual discourse markers cannot be predicted with shallow lexical features. Inspired by Gururangan et al. (2018), we use a Fasttext classifier (Joulin et al., 2017) in order to predict c from s'_2. The Fasttext

[2]For instance, *lovely* is tagged as an adverb because of its suffix, while *besides* was never tagged as an adverb

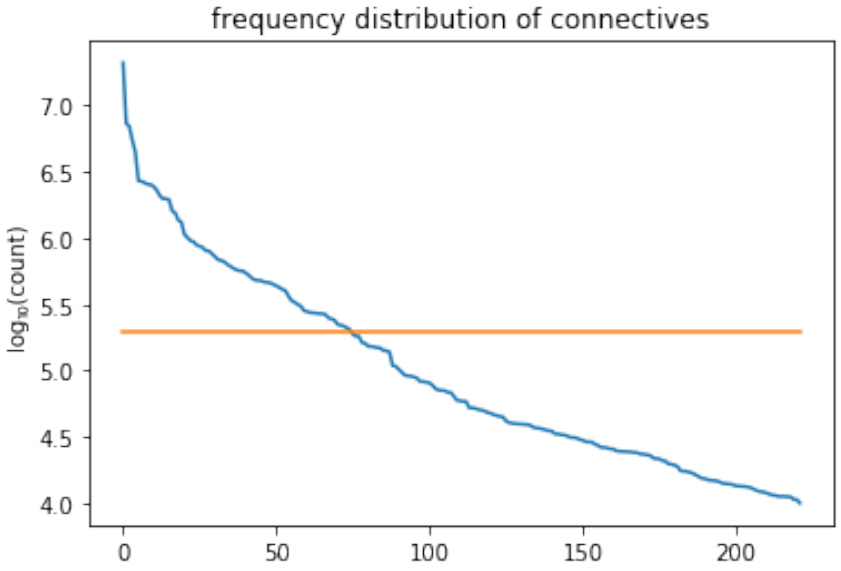


Figure 6-8 – Frequency distribution of candidate discourse markers; the horizontal line indicates the subsampling threshold.

classifier predicts labels from an average of word embeddings fed to a linear classifier. We split the dataset in 5 folds, and we predict markers for each fold, while training on the remaining folds. We use a single epoch, randomly initialized vectors of size 100 (that can be unigrams, bigrams or trigrams) and a learning rate of 0.5.

In addition, we predict c from the concatenation of s_1 and s'_2 (using separate word representations for each case). One might assume that the prediction of c in this case relies on the interaction between s_1 and s_2; however, the features of s_1 and s_2 within Fasttext's setup only interact additively, which means that the classification most likely relies on individual cues in the separate sentences, rather than on their combination. In order to test this hypothesis, we introduce a *random shuffle* operation: for each example (s_1, s'_2, c), s'_2 is replaced by a random sentence from a pair that is equally linked by c (we perform this operation separately in train and test sets).

Table 6.3 indicates that shallow lexical features indeed yield relatively high prediction rates. Moreover, the shuffle operation indeed increases accuracy, which corroborates the hypothesis that classification with shallow features relies on individual cues from separate sentences, rather than their combination.

features	accuracy (%)
majority rule	1.2
s_2	18.6
s_1-s_2	21.9
s_1-s_2 (shuffled)	24.8

Table 6.3 – Accuracy when predicting candidate discourse markers using shallow lexical features

Tables 6.4 and 6.5 show the least and most predictable discourse markers, and the corresponding recognition rate with lexical features.

candidate marker	accuracy (%)
evidently,	0.0
frequently,	0.0
meantime,	0.0
truthfully,	0.0
supposedly,	0.1

Table 6.4 – Candidate discourse markers that are the most difficult to predict from shallow features

candidate marker	accuracy (%)
defensively,	65.5
afterward	71.1
preferably,	71.9
this,	72.7
very,	90.7

Table 6.5 – Candidate discourse markers that are the easiest to predict from shallow features. This shows candidates that are unlikely to be interesting discourse cues.

Interestingly, the two most predictable candidates are not discourse markers. Upon inspection of harvested pairs, we noticed that even legitimate discourse markers can be guessed with relatively simple heuristics in numerous examples. For example, $c =$ *thirdly* is very likely to occur if s_1 contains *secondly*. Therefore, a training example where s_1 doesn't contain *secondly* could be more challenging and incentivize the encoder to represent the richer information that s_2 is pursuing an enumeration that was already started in s_1. We

will test this hypothesis with a variant of our dataset, called *Hard*. We use this information to optionally filter out such simple instances, as described in the next section.

6.2.4 Dataset variations

In the following, we call our method *Discovery*. We create several variations of the sentence pairs dataset. In *DiscoveryHard*, we remove examples where the candidate marker was among the top 5 predictions in our Fasttext shallow model and keep only the 174 candidate markers with a frequency of at least $10k$. Instances are then sampled randomly so that each marker appears exactly $10k$ times in the dataset.

Subsequently, the resulting set of discourse markers is also used in the other variations of our dataset. *DiscoveryBase* designates the dataset for which examples predicted with the Fasttext model were not removed. In order to measure the extent to which the model makes use of the relation between s_1 and s'_2, we also create a *DiscoveryShuffled* dataset, which is the *DiscoveryBase* dataset subjected to the *random shuffle* operation described previously. To isolate the contribution of our discovery method, the dataset *DiscoveryAdv* discards all discourse markers from PDTB that were not found by our method. Also, in order to measure the impact of label diversity, *Discovery10* uses $174k$ examples for each of the 10 most frequent markers,[3] thus totaling as many instances as *DiscoveryBase*. Finally, *DiscoveryBig* contains almost twice as many instances as *DiscoveryBase*, i.e. $20k$ instances for each discourse marker (although, for a limited number of markers, the number of instances is slightly lower due to data sparseness).

[3]They are: *however, hence, moreover, additionally, nevertheless, furthermore, alternatively, again, next, therefore*

6.3 Evaluation of sentence representation learning

6.3.1 Setup

Our goal is to evaluate the effect of using our various training datasets on sentence encoding, given encoders of equivalent capacity and similar setups. Thus, we follow the exact setup of Infersent (Conneau and Bordes, 2017), also used in the Dissent (Malmi et al., 2018) model: we learn to encode sentences into h with a bi-directional LSTM sentence encoder using element-wise max pooling over time. The dimension size of h is 4096. Word embeddings are fixed GloVe embeddings with 300 dimensions, trained on Common Crawl 840B.[4] A sentence pair (s_1, s_2) is represented with $[h_1, h_2, h_1 \odot h_2, |h_2 - h_1|]$,[5] which is fed to a softmax in order to predict a marker c. We have seen in chapter 5 that this composition had drawbacks, but we use it to allow for easier comparison

Our datasets are split in 90% train, 5% validation, and 5% test. Optimization is done with SGD (learning rate is initialized at 0.1, decayed by 1% at each epoch and by 80% if validation accuracy decreases; learning stops when learning rate is below 10^{-5} and the best model on training task validation loss is used for evaluation; gradient is clipped when its norm exceeds 5). Once the sentence encoder has been trained on a base task, the resulting sentence embeddings are tested with the SentEval library (Conneau and Bordes, 2017).

We evaluate the different variations of our dataset we described above in order to analyze their effect, and compare them to a number of existing models. Table 6.8 displays the tasks used for evaluation. For further analysis, table 6.7 displays the result of *Linguistic Probing* using the method by Conneau et al. (2018a). Although these tasks are primarily designed for understanding the content of embeddings, they also focus on aspects that are desirable to perform well in general semantic tasks (e.g. prediction of tense, or number of object).

[4] https://nlp.stanford.edu/projects/glove/

[5] $h_1 \odot h_2 = (h_{11}.h_{21}, .., h_{1i}.h_{2i}, ...)$

6.3.2 Results

Table 6.6 gives an overview of transfer learning evaluation, also comparing to other supervised and unsupervised approaches.

	N	MR	CR	SUBJ	MPQA	SST2	TREC	SICK-R	SICK-E	MRPC	AVG
InferSent (AllNLI)	1.0	81.1	86.3	92.4	90.2	84.6	88.2	88.4	86.1	76.2	85.9
MTL	124	82.5	**87.7**	94	90.9	83.2	93	**88.8**	**87.8**	**78.6**	**87.4**
SkipThought	74	76.5	80.1	93.6	87.1	82	92.2	85.8	82.3	73	83.6
QuickThought	174	81.3	84.5	**94.6**	89.5	-	92.4	87.1	-	75.9	-
DisSent	4.7	80.1	84.9	93.6	90.1	84.1	**93.6**	84.9	83.7	75	85.6
DiscoveryBase	1.7	82.5	86.3	94.2	90.5	85.2	91.8	85.7	84	75.8	86.2
DiscoveryHard	1.7	81.6	86.5	93.9	90.5	84.8	90	85.4	83.2	76.5	85.8
Discovery10	1.7	81.2	85.1	93.7	90.2	83	90	85.9	83.8	75.8	85.4
DiscoveryAdv	1.4	81.4	85.8	93.8	90.5	83.4	92	86	84.3	75.7	85.9
DiscoveryShuffled	1.7	81.4	86.1	94.1	90.9	**85.3**	90.4	85.6	83.6	75.4	85.9
DiscoveryBig	3.4	**82.6**	87.4	94.5	**91.0**	85.2	93.4	86.4	84.8	76.6	86.9

Table 6.6 – SentEval evaluation results with our models trained on various datasets. The first two models are supervised, the other ones unsupervised. All scores are accuracy percentages, except SICK-R, which is Pearson correlation percentage. InferSent is from Conneau and Bordes (2017), MTL is the multi-task learning based model from Subramanian et al. (2018). Evaluation tasks are described in table 6.8, and N denotes the number of examples for each dataset (in millions). Dissent is from Nie et al. (2019), QuickThought is from Logeswaran et al. (2018) with fixed embeddings configuration. The best result per task appears in bold, the best result for unsupervised setups is underlined.

Note that we outperform DisSent on all tasks except TREC[6] with less than half the amount of training examples. In addition, our approach is arguably simpler and faster.

MTL (Subramanian et al., 2018) only achieves stronger results than our method on the MRPC and SICK tasks. The MTL model uses $124M$ training examples with an elaborate multi-task setup, training on $45M$ sentences with manual translation, $1M$ pairs from SNLI/MNLI, $4M$ parse trees of sentences, and $74M$ consecutive sentence pairs. The model also fine-tunes word embeddings in order to achieve a higher capacity. It is therefore remarkable that our model outperforms it on many tasks. Besides, MTL is not a direct competitor to our approach since its main contribution is its multi-task setup, and it could benefit from using our training examples.

[6]This dataset is composed of questions only, which are underrepresented in our training data.

Our best model rivals (and indeed often outperforms) QuickThought on all tasks, except relatedness (SICK-R). QuickThought's training task is to predict whether two sentences are contiguous, which might incentivize the model to perform well on a relatedness task. We also outperform InferSent on many tasks except entailment and relatedness. Entailment prediction is the explicit training signal for Infersent.

To help the analysis of our different model variations, table 6.9 displays the test scores on each dataset for the original training task. It also shows the related PDTB implicit relation prediction scores. The PDTB is annotated with a hierarchy of relations, with 5 classes at level 1 (including the EntRel relation), and 16 at level 2 (with one relation absent from the test). It is interesting to see that this form of simple semi-supervised learning for implicit relation prediction performs quite well, especially for fine-grained relations, as the best model slightly beats the best current dedicated model, listed at 40.9% in Xue et al. (2017). This is notable since direct transfer from explicit to implicit cases was not thought to be useful (Sporleder and Lascarides, 2008) without filtering or domain adaptation (Rutherford and Xue, 2015).

	BShift	CoordInv	Depth	ObjNum	SubjNum	OddM	Tense	TC	WC	AVG
InferSent	56.5	65.9	37.5	79.9	84.3	53.2	87	78.1	95.2	70.8
SkipThought	**69.5**	69	39.6	83.2	86.2	54.5	**90.3**	82.1	79.6	72.7
QuickThought	56.8	70	40.2	79.7	83	55.3	86.2	80.7	90.3	71.4
DiscoveryBase	63.1	70.6	45.2	83.8	87.2	57.3	89.1	83.2	94.7	74.9
DiscoveryHard	62.7	70.4	44.5	83.4	**88.1**	57.3	89.5	82.8	94.1	74.8
Discovery10	61.3	69.7	42.9	81.8	86.7	55.8	87.8	81.4	**96.1**	73.7
DiscoveryAdv	61.5	70	43.9	82.6	86.2	56.2	89.1	82.8	**96.1**	74.3
DiscoveryShuffled	62.6	**71.4**	45.3	**84.3**	88	**58.3**	89.3	82.8	93.4	75
DiscoveryBig	63.3	**71.4**	**46.0**	84.1	87.8	57.1	89.4	**84.2**	96	**75.5**

Table 6.7 – Accuracy of various models on linguistic probing tasks using logistic regression on SentEval. BShift is detection of token inversion. CoordInv is detection of clause inversion. ObjNum/SubjNum is prediction of the number of object resp. subject. Tense is prediction of the main verb tense. Depth is prediction of parse tree depth. TC is detection of common sequences of constituents. WC is prediction of words contained in the sentence. OddM is detection of random replacement of verbs/nouns by other verbs/nouns. AVG is the average score of those tasks for each model. For more details see Conneau et al. (2018a). SkipThought and Infersent results come from Perone et al. (2018), QuickThought results come from Brahma (2018).

DiscoveryHard scores lower on its training task than *DiscoveryBase*, and it also performs worse on transfer learning tasks. This makes sense, since lexical features are important to solve the evaluation tasks. Our initial hypothesis was that more difficult instances might force the model to use higher-level reasoning, but this does not seem to be the case. More surprisingly, preventing the encoders to use the relationship between sentences, as in *DiscoveryShuffled*, does not substantially hurt the transfer performance, which remains on average higher than Nie et al. (2019). Additionally, our models score well on linguistic probing tasks. They outperform Infersent on all tasks, which seems to contradict the claim that SNLI data allows for learning of *universal* sentence representations (Conneau and Bordes, 2017). And a final interesting outcome is that the diversity of markers (e.g. using *DiscoveryBase* instead of *Discovery10*) seems to be important for good performance on those tasks, since *Discovery10* has the worst overall performance on average.

name	N	task	C
MR	11k	sentiment (movie reviews)	2
CR	4k	sentiment (product reviews)	2
SUBJ	10k	subjectivity/objectivity	2
MPQA	11k	opinion polarity	2
TREC	6k	question-type	6
SST	70k	sentiment (movie reviews)	2
SICK-E	10k	entailment	3
SICK-R	10k	relatedness	3
MRPC	4k	paraphrase detection	2
PDTB_5	17k	implicit discourse relation (coarse)	5
PDTB_{16}	17k	implicit discourse relation (fine)	15

Table 6.8 – Transfer evaluation tasks. N is the number of training examples and C is number of classes for each task.

6.3.3 Visualisation

The softmax weights learned during the training phase can be interpreted as embeddings for the markers themselves, and used to visualize their relationships. Figure 6-9 shows a TSNE (van der Maaten and Hinton, 2008) plot of the markers' representations. Proximity in the feature space seems to reflect semantic similarity (e.g. *usually/normally*). In addition, the

	$PDTB_5$ coarse	$PDTB_{16}$ fine	$\mathcal{T}$
InferSent	46.7	34.2	-
DisSent	48.9	36,9	-
DiscoveryBase	**52.5**	40.0	20.6
DiscoveryHard	50.7	39.8	9.3
Discovery10	48.3	37.7	51.9
DiscoveryAdv	49.7	37.6	26.1
DiscoveryShuffled	51.0	39.5	11.5
DiscoveryBig	51.3	**41.3**	22.2

Table 6.9 – Test results (accuracy) on implicit discursive relation prediction task (PDTB relations level 1 and 2, i.e coarse-grained and fine-grained) and training tasks $\mathcal{T}$. Note that scores for $\mathcal{T}$ are not comparable since the test set changes for each version of the dataset.

markers we discovered, colored in red, blend with the PDTB markers (depicted in black). It would be interesting to cluster markers in order to empirically define discourse relations, but we leave this for future work.

6.4 Conclusion

In this chapter, we introduced a novel and efficient method to automatically discover discourse markers from text, and we use the resulting set of candidate markers for the construction of an extensive dataset for semi-supervised sentence representation learning. A number of dataset variations are evaluated on a wide range of transfer learning tasks (as well as implicit discourse recognition) and a comparison with existing models indicates that our approach yields state of the art results on the bulk of these tasks. Additionally, our analysis shows that removing 'simple' examples is detrimental to transfer results, while preventing the model to exploit the relationship between sentences has a negligible effect. This leads us to believe that, even though our approach reaches state of the art results, there is still room for improvement: models that adequately exploit the relationship between sentences would be better at leveraging the supervision of our dataset, and could yield even better sentence representations. In future work, we also aim to increase the coverage of our method. For instance, we can make use of more lenient patterns that capture an even

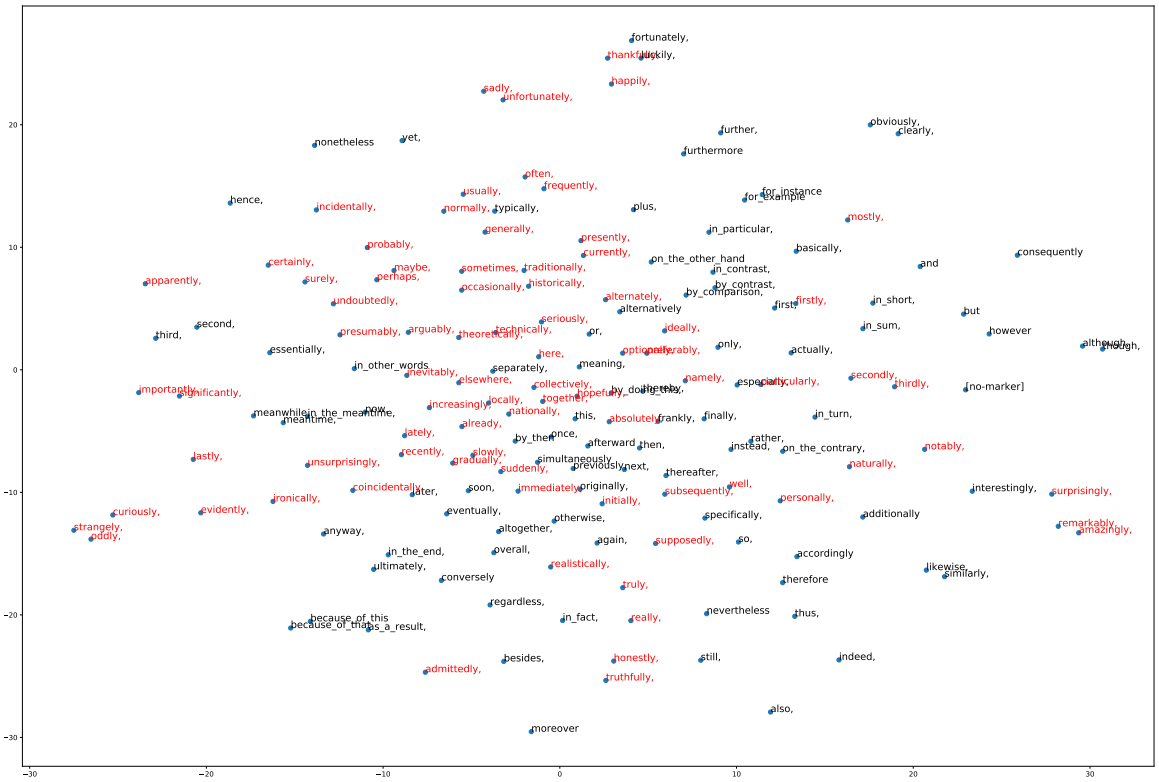


Figure 6-9 – TSNE visualization of the softmax weights from our *DiscoveryBig* model for each discourse marker, after unit norm normalization. Markers discovered by our method (e.g. absent from PDTB annotations) are colored in red.

wider range of discourse markers, such as multi-word markers. Training signals are key to general natural language understanding, but they go in pair with evaluation, and in the next chapter we will show that discourse based evaluation highlights the potential of Discovery even more.

CHAPTER 7

DISCOURSE-BASED EVALUATION OF LANGUAGE UNDERSTANDING

7.1 Motivation

Over the last year, novel models for natural language understanding (NLU) have made a remarkable amount of progress on a number of widely accepted evaluation benchmarks. The GLUE benchmark (Wang et al., 2018), for example, was designed to be a set of challenging NLU tasks, such as question answering, sentiment analysis, and textual entailment; yet, current state of the art systems surpass human performance estimates on the average score of its subtasks (Yang et al., 2019). Similarly, the NLU subtasks that are part of the SentEval framework, a widely used benchmark for the evaluation of sentence-to-vector encoders, are successfully dealt with by current neural models, with scores that exceed the 90% mark.[1]

The impressive results on these benchmarks might lead one to believe that natural language understanding is largely a solved problem. Based on the resulting performance on the above-mentioned benchmarks, a considerable number of researchers has even put for-

[1] http://nlpprogress.com/english/semantic_textual_similarity.html

ward the claim that their models induce *universal* representations (Cer et al., 2018a; Kiros and Chan, 2018; Subramanian et al., 2018; Wieting et al., 2015; Liu et al., 2019). It is important to note, however, that benchmarks like SentEval and GLUE are primarily focusing on semantic aspects, i.e. the literal and uncontextualized content of text. While the semantics of language is without doubt an important aspect of language, we believe that a single focus on semantic aspects leads to an impoverished model of language.

For a versatile model of language, other aspects of language, viz. pragmatic aspects, equally need to be taken into account. Pragmatics focuses on the larger context that surrounds a particular textual instance, and they are central to meaning representations that aspire to lay a claim to universality. Consider the following utterance:

(1) You're standing on my foot.

The utterance in (1) has a number of direct implications that are logically entailed by the utterance above, such as the implication that the hearer is standing on a body part of the speaker, and the implication that the speaker is touching the hearer. But there are also more indirect implications, that are not literally expressed, but need to be inferred from the context, such as the implication that the speaker wants the hearer to move away from them. The latter kind of implication, that is indirectly implied by the context of an utterance, is called *implicature*—a term coined by Grice (1975). In real world applications, recognizing the implicatures of a statement is arguably more important than recognizing its mere semantic content.

The implicatures that are conveyed by an utterance are highly dependent on its illocutionary force (Austin, 1975). In Austin's framework, the *locution* is the literal meaning of an utterance, while the *illocution* is the goal that the utterance tries to achieve. When we restrict the meaning of (1) to its locution, the utterance is reduced to the mere statement that the hearer is standing on the speaker's foot. However, when we also take its illocution into account, it becomes clear that the speaker actually formulates the request that the speaker step away. The utterance's illocution is clearly an important part of the entire meaning of

the utterance, that is complementary to the literal content (Green, 2000).[2]

The example above makes clear that pragmatics is a fundamental aspect of the meaning of an utterance. Semantics focuses on the literal content of utterances, but not on the kind of goal the speaker is trying to achieve. Pragmatic (i.e. discourse-based) tasks focus on the actual use of language, so a discourse-centric evaluation could *by construction* be a better fit to evaluate how NLU models perform in practical use cases, or at least should be used as a complement to semantics-focused evaluations benchmarks. Ultimately, many use cases of NLP models are related to conversation with end users or analysis of structured documents. In such cases, discourse analysis (i.e. the ability to parse high-level textual structures that take into account the global context) is a prerequisite for human level performance. Moreover, standard benchmarks often strongly influence the evolution of NLU models, which means they should be as exhaustive as possible, and closely related to the models' end use cases.

In this work, we compile a list of 11 discourse-focused tasks that are meant to complement existing benchmarks. We propose: (i) A new evaluation benchmark, named *DiscEval*, which we make publicly available.[3] (ii) Derivations of human accuracy estimates for some of the tasks. (iii) Evaluation on these tasks of state of the art generalizable NLU model, viz. BERT, alongside BERT augmented with auxiliary finetunings. (iv) New comparisons of discourse-based and Natural Language Inference based training signals showing that the most widely used auxiliary finetuning dataset, viz. MNLI, is not the best performing on DiscEval, which suggests a margin for improvements.

[2]In order to precisely determine their illocution, utterances have been categorized into classes called speech acts (Searle et al., 1980), such as ASSERTION, QUESTION or ORDER which have different kinds of effects on the world. For instance, constative speech acts (e.g. *the sky is blue*) describe a state of the world and are either true or false while performative speech acts (e.g. *I declare you husband and wife*) can change the world upon utterance (Austin, 1975).

[3]https://github.com/disceval/DiscEval

7.2 Related Work

Evaluation methods of NLU have been the object of heated debates since the proposal of the Turing Test. Automatic evaluations relying on annotated datasets are arguably limited but they became a standard. They can be based on sentence similarity (Agirre et al., 2012), leveraging human annotated scores of similarity between sentence pairs. Predicting similarity between two sentences requires some representation of their semantic content beyond their surface form, and sentence similarity estimation tasks can potentially encompass many aspects, but it is not clear how humans annotators weight semantic, stylistic, and discursive aspects while rating.

Using a set of more focused and clearly defined tasks has been a popular approach. Kiros et al. (2015) proposed a set of tasks and tools for sentence understanding evaluation. These 13 tasks were compiled in the SentEval (Conneau and Bordes, 2017) evaluation suite designed for automatic evaluation of pre-trained sentence embeddings. SentEval tasks are mostly based on sentiment analysis, semantic sentence similarity and natural language inference. Since SentEval evaluates sentence embeddings, the users have to provide a sentence encoder that is not finetuned during the evaluation.

GLUE (Wang et al., 2018) proposes to evaluate language understanding with less constraints than SentEval, allowing users not to rely on explicit sentence embedding based models. They compile 9 classification or regression tasks that are carried out for sentences or sentence pairs. 3 tasks are semantic similarity, and 4 tasks are based on NLI.

NLI can be regarded as a universal framework for evaluation. In the *Recast* framework (Poliak et al., 2018a), existing datasets (e.g. sentiment analysis) are formulated as NLI tasks. For instance, based on the sentence *don't waste your money*, annotated as a negative review, they use handcrafted rules to generate the following example: (PREMISE: *When asked about the product, liam said "don't waste your money"* , HYPOTHESIS: *Liam didn't like the product*, LABEL: entailment). However, the generated datasets prevent the evaluation to measure directly how well a model deals with the semantic phenomena present in the original dataset, since some sentences use artificially generated reported speech. Thus,

NLI data could be used to evaluate discourse analysis, but it is not clear how to generate examples that are not overly artificial. Moreover, it is unclear to what extent instances in existing NLI datasets need to deal with pragmatic aspects (Bowman, 2016).

SuperGLUE (Wang et al., 2018) updates GLUE with six novel tasks that are selected to be even more challenging. Two of them deal with contextualized lexical semantics, two tasks are a form of question answering, and two of them are NLI problems. One of those NLI tasks, CommitmentBank (de Marneffe et al., 2019), is the only explicitly discourse-related task.

Another effort towards evaluation of general purpose NLP systems is DecaNLP (McCann et al., 2018). The 10 tasks of this benchmarks are all framed as question answering. For example, a question answering task is derived from a sentiment analysis task using artificial questions such as *Is this sentence positive or negative?*. Four of these tasks deal with semantic parsing, and other tasks include NLI and sentiment analysis. Discourse phenomena can be involved in some tasks (e.g. the summarization task) although it is hard to assess to what extent.

Discourse relation prediction has punctually been used for sentence representation learning evaluation, by Nie et al. (2019) which we followed in chapter 6, but it consisted in only one dataset (viz. the PDTB (Prasad et al., 2008)), which we included in our benchmark. Discourse for evaluation has also been considered in the field of machine translation. Läubli et al. (2018) showed that neural models achieve superhuman results on sentence-level translations but that current models yield underwhelming results when considering document-level translations, also making a case for discourse-aware evaluations.

Other evaluations, such as linguistic probing or GLUE diagnostics (Conneau et al., 2018a; Belinkov and Glass, 2019; Wang et al., 2019c), focus on an internal understanding of what is captured by the models (e.g. syntax, lexical content), rather than measuring performance on external tasks, and are outside the scope of this work, while providing a complementary viewpoint.

dataset	categories	exemple	class	N_{train}
PDTB	discourse relation	“it was censorship”/“it was outrageous”	conjunction	13k
STAC	discourse relation	“what ?”/“i literally lost”	question-answer-pair	11k
GUM	discourse relation	“Do not drink”/“if underage in your country”	condition	2k
Emergent	stance	“a meteorite landed in nicaragua.”/“small meteorite hits managua”	for	2k
SwitchBoard	speech act	“well , a little different , actually ,”	hedge	19k
MRDA	speech act	“yeah that ’s that ’s that ’s what i meant .”	acknowledge-answer	14k
Persuasion	E/S/S/R	“Co-operation is essential for team work”/“lions hunt in a team”	low specificity	0.6k
SarcasmV2	sarcasm presence	“don’t quit your day job”/“[...] i was going to sell this joke. [...]”	sarcasm	9k
Squinky	I/I/F	“boo ya.”	uninformative, high implicature, unformal	4k
Verifiability	verifiability	“I’ve been a physician for 20 years.”	verifiable-experiential	6k
EmoBank	V/A/D	“I wanted to be there..”	low valence, high arousal, low dominance	5k

Table 7.1 – DiscEval classification datasets. N_{train} is the number of examples in the training set. E/S/S/R denotes Eloquence/Strength/Specificity/Relevance; I/I/F is Information/Implicature/Formality ; V/A/D denotes Valence/Arousal/Dominance

7.3 Proposed Tasks

Our goal is to compile a set of diverse discourse-related tasks. We restrict ourselves to classification either of sentences or sentence pairs and only use publicly available datasets that are absent from other benchmarks (SentEval/GLUE/SuperGLUE).

The scores in our tasks are not all meant to be compared to previous work, since we alter some datasets to yield more meaningful evaluations (we perform duplicate removal or class subsampling when mentioned). We found these operations necessary in order to leverage the rare classes and yield more meaningful scores. As an illustration, GUM initially consists of more than 99% of *unattached* labels, and SwitchBoard contains 80% of *statements*.

We first present the tasks we selected, also described in table 7.1, and then propose a rudimentary taxonomy of how they address different aspects of meaning.

PDTB The Penn Discourse Tree Bank (Prasad et al., 2014) contains a collection of fine-grained implicit (i.e. not signaled by a discourse marker) relations between sentences from the news domain in the Penn Discourse TreeBank 2.0. We select the level 2 relations as categories.

STAC (Strategic Conversation) is a corpus of strategic chat conversations manually annotated with negotiation-related information, dialogue acts and discourse structures in the framework of Segmented Discourse Representation Theory (SDRT, Asher and Lascarides, 2003). We only consider pairwise relations between all dialog acts, following Badene et al. (2019). We remove duplicate pairs and dialogues that only have non-linguistic utterances (coming from the game server). We subsample dialog act pairs with no relation so that they constitute 20% of each fold.

GUM (Zeldes, 2017) is a corpus of multilayer annotations for texts from various domains; it includes Rhetorical Structure Theory (RST, **?**) discourse structure annotations. Once again, we only consider pairwise interactions between discourse units (e.g. sentences/clauses). We subsample discourse units with no relation so that they constitute 20% of each document. We split the examples in train/test/dev sets randomly according to the document they belong to.

Emergent (Ferreira and Vlachos, 2016) is composed of pairs of assertions and titles of news articles that are *against*, *for*, or *neutral* with respect to the opinion of the assertion.

SwitchBoard (Godfrey et al., 1992) contains textual transcriptions of dialogs about various topics with annotated speech acts. We remove duplicate examples and subsample *Statements* and *Non Statements* so that they constitute 20% of the examples. We use a custom train/dev validation split (90/10 ratio) since our preprocessing leads to a drastic size reduction of the original development set. The label of a speech act can be dependent on the context (previous utterances), but we discarded it in this work for the sake of simplicity, even though integration of context could improve the scores (Ribeiro et al., 2015).

MRDA (Shriberg et al., 2004) contains textual transcription of multi-party real meetings, with speech act annotations. We remove duplicate examples. We use a custom train/dev validation split (90/10 ratio) since this deduplication leads to a drastic size reduction of

the original development set, and we subsample *Statement* examples so that they constitute 20% of the dataset. We also discarded the context.

Persuasion (Carlile et al., 2018) is a collection of arguments from student essays annotated with factors of persuasiveness with respect to a claim; considered factors are the following: Specificity, Eloquence, Relevance and Strength. For each graded target, we cast the ratings into three quantiles and discard the middle quantile.

SarcasmV2 (Oraby et al., 2016) consists of messages from online forums with responses that may or may not be sarcastic according to human annotations.

Squinky dataset (Lahiri, 2015) gather annotations in Formality and Informativeness and Implicature where sentences were graded on a scale from 1 to 7. They define the Implicature score as the amount of not explicitly stated information carried in a sentence. For each target, we cast the ratings into three quantiles and discard the middle quantile.

Verifiability (Park and Cardie, 2014) is a collection of online user comments annotated as *Verifiable-Experiential* (verifiable and about writer's experience) *Verifiable-Non-Experiential* or *Unverifiable*.

EmoBank (Buechel and Hahn, 2017) aggregates emotion annotations on texts from various domains using the VAD representation format. The authors define Valence as *corresponding to the concept of polarity*[4], Arousal as *degree of calmness or excitement* and Dominance as *perceived degree of control over a situation*. For each target, we cast the ratings into three quantiles and discard the middle quantile.

It has been argued by Halliday (1985) that linguistic phenomena fall into three metafunctions: *ideational* for semantics, *interpersonal* for appeals to the hearer/reader, and *textual* for form-related aspects. This forms the basis of discourse relation types by (Hovy and

[4]This is the dimension that is widely used in sentiment analysis.

Maier, 1992) in which they are called semantic, interpersonal and presentational. DiscEval tasks cut across these categories, because some of the tasks integrate all aspects when they characterize the speech act or discourse relation category associated to a discourse unit (mostly sentences), an utterance or a pair of these. However, most discourse relations involved focus on *ideational* aspects, which are thus complemented by tasks insisting on more interpersonal aspects (e.g. using appeal to emotions, or verifiable arguments) that help realizing speech act's intentions. Finally, intentions can achieve their goals with varying degrees of success. This leads us to a rudimentary grouping of our tasks:

– The speech act classification tasks (SwitchBoard, MRDA) deal with the detection of the intention of utterances. They use the same label set (viz. DASML, Allen and Core, 1997) but different domains and annotation guidelines. A discourse relation characterizes how an utterance contributes to the coherence of a document/conversation (e.g through *elaboration* or *contrast*), so this task requires a form of understanding of the use of a sentence, and how a sentence fits with another sentence in a broader discourse. Here, three tasks (PDTB, STAC, GUM) deal with discourse relation prediction with varying domains and formalisms[5]. The Stance detection task can be seen as a coarse-grained discourse relation classification.

– Detecting emotional content, verifiability, formality, informativeness or sarcasm is necessary in order to figure out in what realm communication is occurring. A statement can be persuasive, yet poorly informative and unverifiable. Emotions (Dolan, 2002) and power perception (Pfeffer, 1981) can have a strong influence on human behavior and text interpretation. Manipulating emotions can be the main purpose of a speech act as well. Sarcasm is another means of communication and sarcasm detection is in itself a straightforward task for evaluation of pragmatics, since sarcasm is a clear case of literal meaning being different from the intended meaning.

– Persuasiveness prediction is a useful tool to assess whether a model can measure how well a sentence can achieve its intended goal. This aspect is orthogonal to the determination of the goal itself, and is arguably equally important.

[5]These formalisms have different assumptions about the nature of discourse structure.

7.4 Evaluations

7.4.1 Models

Our goal is to assess the performance of popular NLU models and the influence of various training signals on DiscEval scores. We evaluate state of the art models and baselines on DiscEval using the Jiant (Wang et al., 2019d) framework. Our baselines include average of GloVe (Pennington et al., 2014) embeddings (CBoW) and BiLSTM with GloVe and ELMo (Peters et al., 2018a) embeddings. We also evaluate BERT (Devlin et al., 2019) base uncased models, and perform experiments with *Supplementary Training on Intermediate Labeled-data Tasks* (STILT) (Phang et al., 2018). STILT is a further pretraining step on a data-rich task before the final fine-tuning evaluation on the target task. STILTs can be combined using multitask learning. We use Jiant default parameters[6], and uniform loss weighting when multitasking (a different task is optimized at each training batch).

We finetune BERT with four of such training signals:

MNLI (Williams et al., 2018a) is a collection of 433k sentence pairs manually annotated with *contradiction*, *entailment*, or *neutral* relations. Finetuning with this dataset leads to accuracy improvement on all GLUE tasks except CoLA (Phang et al., 2018).

DisSent data is from (Nie et al., 2019) introduced in section 4.2.4, consisting of $4.7M$ sentences or clauses that were separated by a discourse marker from a list of 15 markers. Prediction of discourse markers based of the context clauses/sentences with which they occurred have been used as a training signal for sentence representation learning. Authors used handcrafted rules for each marker in order to ensure that the markers actually signal a form of relation. DisSent has underwhelming results on the GLUE tasks as a STILT (Wang et al., 2019a).

[6] https://github.com/nyu-mll/jiant/blob/706b6521c328cc3dd6d713cce2587ea2ff887a17/jiant/config/examples/stilts_example.conf

Discovery (Sileo et al., 2019) is the dataset for discourse marker prediction that we introduced in the previous chapter, composed of 174 discourse markers with $10k$ usage examples for each marker. Sentence pairs were extracted from web data, and the markers come either from the PDTB or from an heuristic automatic extraction.

DiscEval refers to all DiscEval tasks used in a multitask setup; we discard Persuasion subtasks other than Strength (since other subtasks are factors for strength) and weight tasks and subtasks identically otherwise.

7.4.2 Human accuracy estimates

For a more insightful comparison, we propose derivations of human accuracy estimates from the datasets we used.

The authors of SarcasmV2 (Oraby et al., 2016) dataset directly report 80% annotator accuracy compared to the gold standard. Prasad et al. (2014) report 84% annotators agreement for PDTB 2.0, which is a lower bound of accuracy. GUM (Zeldes, 2017) authors report *attachment accuracy of* 87.22% *and labelling accuracy of* 86.58% *as compared to the 'gold standard' after instructor adjudication*. We interleaved attachment and labelling in our task. Assuming human annotators never predict the non-attached relation, 69.3% is a lower bound for human accuracy. Authors of the Verifiability (Park and Cardie, 2014) dataset report an agreement $\kappa = 0.73$ which yields an agreement of 87% given the classes distribution which is a lowerbound of human accuracy. We estimated human accuracy on EmoBank (Buechel and Hahn, 2017) with the intermediate datasets provided by the authors. For each target (V,A,D) we compute the average standard deviation, and compute the probability (under normality assumption) of each example rating of falling under the wrong category.

Unlike the GLUE benchmark (Nangia and Bowman, 2019), we do not yet provide human accuracy estimates obtained in a standardized way. The high number of classes would make that process rather more difficult. But these estimates are still useful even though

they should be taken with a grain of salt.

7.4.3 Overall Results

	PDTB	STAC	GUM	Emergent	SwitchB.	MRDA	Persuasion	Sarcasm	Squinky	Verif.	EmoBank
CBoW	27.4	32	20.5	59.7	3.8	0.7	70.6	61.1	75.5	74	64
BiLSTM	25.9	27.7	18.5	45.6	3.7	0.7	62.6	63.1	72.1	74	63.5
BiLSTM+ELMo	27.5	33.5	18.9	55.2	3.7	0.7	67.4	68.9	82.5	74	66.9
BERT	48.8	48.2	40.9	79.2	38.8	22.3	74.8	**77.1**	87.5	86.7	76.2
BERT+MNLI	49.1	49.1	42.8	81.2	38.1	22.7	71.7	73.4	88.2	86	76.3
BERT+DiscEval	49.1	**57.1**	42.8	80.2	**40.3**	**23.1**	**76.2**	75	87.6	85.9	76
BERT+DisSent	49.4	49	43.9	79.8	39.2	22	74.7	74.9	87.5	85.9	76.2
B+DisSent+MNLI	49.6	49.2	**44.2**	80.9	39.8	22.1	74	74.1	87.6	85.6	76.4
BERT+Discovery	50.7	49.5	42.7	**81.7**	39.5	22.4	71.6	76.7	88.6	86.3	**76.6**
B+Discovery+MNLI	**51.3**	49.4	43.1	80.7	**40.3**	22.2	73.6	75.1	**88.9**	**86.8**	76
Human estimate	**84.0**	-	**69.3**	-	-	-	-	**80.0**	-	**87.0**	73.1

Table 7.2 – Transfer test scores across DiscEval tasks; We report the average when the dataset has several classification tasks (as in Squinky, EmoBank and Persuasion); B(ERT)+$\mathcal{X}$ refers to BERT pretrained classification model after auxiliary finetuning phase on task $\mathcal{X}$. All scores are accuracy scores except SwitchBoard/MRDA (which are macro-F1 scores)

Task-wise results are presented in table 7.2. We report the average scores of 6 runs of STILT and finetuning phases.

DiscEval seem to be challenging even to BERT base model, which has shown strong performance on GLUE (and vastly outperform the baselines on our tasks). For many tasks, there is a STILT that significantly improves the accuracy of BERT. The gap between human accuracy and BERT model is particularly high on implicit discourse relation prediction (PDTB and GUM). This task is known as hard, and previous work has shown that task dedicated models are not yet on par with human performance (Morey et al., 2017). Pretraining on MNLI worsens the DiscEval average score for BERT base model. A lower sarcasm detection score could indicate that BERT+MNLI has more focus on the literal content of statements, even though no STILT improves sarcasm detection. All models score below human accuracies, with the exception of emotion classification (but only for the valence classification subtask).

Table 7.3 shows aggregate results alongside comparisons with GLUE scores. The best overall unsupervised result is achieved with Discovery STILT. Combining Discovery and MNLI yields both a high DiscEval and GLUE score, and also yields a high GLUE diagnostics score. All discourse based STILT improve GLUE score, while MNLI does not improve DiscEval average score. DiscEval tasks based on sentence pairs seem to account for the variance across STILTs.

MNLI has been suggested as a good default auxiliary training task based on evaluation on GLUE (Phang et al., 2018) and SentEval (Conneau and Bordes, 2017). However, our evaluation suggests that finetuning a model with MNLI alone has significant drawbacks.

More detailed results for datasets with several subtasks are shown in table 7.4. We note that MNLI STILT significantly decreases relevance estimation performance (on BERT base and while multi-tasking with DisSent). Many models surpass the human estimate at valence prediction, a well studied task, but interestingly it's not the case for Arousal and Dominance prediction.

	DiscEval$_{AVG}$	D.E.-Pairs$_{AVG}$	D.E.-Single$_{AVG}$	GLUE$_{AVG}$	GLUE$_{diagnostics}$
BERT	61.8±.4	57.9±.5	62.3±.3	74.7±.2	31.7±.3
BERT+MNLI	61.7±.5	57.2±.5	62.2±.4	**77.0**±.2	32.5±.6
BERT+DiscEval MTL	**63.0**±.4	**60.0**±.4	62.6±.2	75.3±.2	31.6±.3
BERT+DisSent	62.0±.4	58.4±.4	62.2±.3	75.1±.2	31.5±.3
B+DisSent+MNLI	62.1±.4	58.2±.4	62.3±.2	76.6±.1	32.4±.0
BERT+Discovery	62.4±.3	58.2±.4	62.7±.3	75.0±.2	31.3±.2
B+Discovery+MNLI	62.5±.4	58.5±.5	**62.8**±.3	76.6±.2	**33.3**±.2

Table 7.3 – Aggregated transfer test accuracies across DiscEval and comparison with GLUE validation downstream and diagnostic tasks (GLUE diagnostic tasks evaluate NLI performance under presence of linguistic phenomena such as negation, quantification, use of common sense); BERT+$\mathcal{X}$ refers to BERT pretrained classification model after auxiliary finetuning phase on task $\mathcal{X}$; D.E.-Pairs$_{AVG}$ is the average of DiscEval sentence pair classification tasks.

The categories of our benchmark tasks cover a broad range of discourse aspects. The overall accuracies only show a synthetic view of the tasks evaluated in DiscEval. Some datasets contain many subcategories that allow for a fine grained analysis through a wide array of classes (viz. 51 categories for MRDA). Table 7.5 shows a fine grained evaluation which yields some insights on the capabilities of BERT. We report the 6 most frequent

	Persuasiveness				EmoBank			Squinky		
	Eloquence	Relevance	Specificity	Strength	Valence	Arousal	Dom.	Inf.	Implicature	Formality
BERT	75.6	63.5	81.6	78.3	87.1	72	69.5	92.2	72.1	98.3
BERT+MNLI	74.7	57.5	82.3	72.2	86.6	72.4	69.9	92.5	73.9	98.1
BERT+DiscEval	75.6	64	83.2	**82.0**	86.8	71.9	69.2	92.3	71.8	**98.6**
BERT+DisSent	73.8	63	82.6	79.5	87.1	71.4	70.1	92.6	72	97.7
B+DisSent+MNLI	**76.9**	61.5	**83.9**	73.9	**87.6**	72.1	69.4	91.5	73.4	97.9
BERT+Discovery	76	59.1	80.1	71.4	86.8	**72.6**	**70.5**	**93.2**	74.2	98.5
B+Discovery+MNLI	74.1	60.4	79.4	80.4	86.4	72.1	69.6	93.1	**75.3**	98.4
Human estimate	-	-	-	-	74.9	**73.8**	**70.5**	-	-	-

Table 7.4 – Transfer test accuracies across DiscEval subtasks (Persuasiveness, EmoBank, Squinky) BERT+$\mathcal{X}$ refers to BERT pretrained classification model after auxiliary finetuning phase on task $\mathcal{X}$.

classes per task for conciseness sake. It is worth noting that the BERT models do not neglect rare classes. These detailed results reveal that BERT+MNLI scores for discourse relation prediction are inflated by good scores at predicting absence of relation (possibly close to the neutral class in NLI), which is useful but not sufficient for discourse understanding. The STILTs have complementary strengths even with given tasks, which can explain why combining them is helpful. However, we used a quite simplistic multitasking setup and efficient combination of the tasks remains an open problem.

7.5 Conclusion

We proposed DiscEval, a set of discourse related evaluation tasks, and used them to evaluate BERT finetuned on various auxiliary finetuning tasks. The results lead us to rethink the efficiency of mainly using NLI as an auxiliary training task. DiscEval can be used for training or evaluation in general NLU or discourse related work. Much effort has been devoted to NLI for training and evaluation for general purpose sentence understanding, but we just scratched the surface of the use of discourse oriented tasks. In further investigations, we plan to use more general tasks than classification on sentence or pairs, such as longer and possibly structured sequences. Several of the datasets we used (MRDA, SwitchBoard, GUM, STAC) already contain such higher level structures. In addition, a more inclusive

	BERT	B+MNLI	B+DisSent	B+Discovery	B+DiscEval	Support
GUM.no_relation	48.9	**51.0**	46	45.4	43.3	48
GUM.circumstance	77.1	**80.6**	73.2	77.8	74.6	35
GUM.elaboration	41.5	38.5	40	**46.1**	42.9	32
GUM.background	22.6	25.3	34.3	**38.2**	35.8	23
GUM.evaluation	20.4	22.6	**36.8**	29.9	35.1	20
STAC.no_relation	59.9	**63.8**	55.4	61.3	46.9	117
STAC.Comment	77.8	76.1	74.9	**78.6**	54.4	115
STAC.Question_answer_pair	79.1	80.1	**83.3**	76.9	83	93
STAC.Q_Elab	32.1	34.3	32	38.1	**63.7**	86
STAC.Contrast	29.6	37.4	25.9	27.5	**49.9**	53
SwitchBoard.Uninterpretable	86	86	85.5	86.1	**86.3**	382
SwitchBoard.Statement-non-opinion	72	72.1	**72.4**	**72.4**	**72.4**	304
SwitchBoard.Yes-No-Question	**85.9**	85.2	85.5	**85.9**	85.8	303
SwitchBoard.Statement-opinion	46.3	46.3	48.6	48.8	**49.5**	113
SwitchBoard.Appreciation	**73.5**	71.1	70.2	71.7	72.9	108
PDTB.Cause	55.2	55.7	53.1	**57.2**	55.9	302
PDTB.Restatement	40.4	40	41.3	**43.9**	41	263
PDTB.Conjunction	52.8	**53.9**	52.1	53.3	52.5	262
PDTB.Contrast	45.8	**49.0**	47.2	48	46	172
PDTB.Instantiation	56.6	55.6	52.8	**58.7**	55.7	109
MRDA.Statement	51.2	51.8	48.9	**53.4**	51.4	364
MRDA.Defending/Explanation	52.8	54.1	**55.3**	52.8	52	166
MRDA.Expansions of y/n Answers	**51.7**	48.7	50.3	49.6	49.4	139
MRDA.Offer	48.6	46.9	**50.7**	49.4	49.4	102
MRDA.Rising Tone	39.3	40.1	40.3	**40.7**	38.8	98

Table 7.5 – Transfer F1 scores across the categories of DiscEval tasks; B(ERT)+$\mathcal{X}$ denotes BERT pretrained classification model after auxiliary finetuning phase on task $\mathcal{X}$.

comparison with human annotators on discourse tasks could also help to pinpoint the weaknesses of current models dealing with discourse phenomena. Yet another step would be to study the correlations between performance metrics in deployed NLU systems and scores of the automated evaluation benchmarks (GLUE/DiscEval) in order to validate our claims about the centrality of discourse.

CHAPTER 8

REPURPOSING CLASSIFICATION DATASETS FOR SEMANTIC ANALYSIS OF DISCOURSE MARKERS

8.1 Motivation

Discourse markers are a common language device used to make explicit the semantic and/or pragmatic relationship between clauses or sentences. As we have seen in the two previous chapters, pretraining with discourse marker prediction is a fruitful strategy for text representation learning. We hypothesized that discourse markers can act as noisy labels for relations or sentence types. In this chapter, we propose a method to actually verify that hypothesis.

Several resources enumerate discourse markers and their use in different languages, either in discourse marker lexicons (Knott, 1996; Stede, 2002; Roze et al., 2012; Das et al., 2018) or in corpora annotated with discourse relations, such as the well-known English Penn Discourse TreeBank (PDTB; Prasad et al., 2008), which inspired other efforts in Turkish, Chinese and French (Zeyrek and Webber, 2008; Zhou et al., 2014; Danlos et al., 2015). The PDTB identifies different types of discourse relation categories (such as *conjunction* and *contrast*) and the respective markers that frequently instantiate these categories (such

as *and* and *however*, respectively), and organizes them in a three-level hierarchy. It must be noted, however, that there is no general consensus on the typology of these markers and their rhetorical functions. As such, theoretical alternatives to the PDTB exist, such as RST (Carlson et al., 2001) or SDRT (Asher and Lascarides, 2003). Moreover, marker inventories are by no means exhaustive, and the role of markers is not purely semantic, but also depends on the grammatical, stylistic and pragmatic context of their use.

Meanwhile, there exist a number of NLP classification tasks (with associated datasets) that equally consider the relationship between sentences or clauses, but with a set of relations that is rather different in nature; these tasks focus on phenomena such as implication and contradiction (Bowman et al., 2015b), semantic similarity, or paraphrase (Dolan et al., 2004). Furthermore, a number of tasks consider single sentence phenomena, such as sentiment, subjectivity, and style. Such characteristics have been largely ignored for the linguistic analysis and categorization of discourse markers *per se*, even though discourse markers have been successfully used to improve categorization performance for these tasks (Jernite et al., 2017; Nie et al., 2019; Pan et al., 2018a; Sileo et al., 2019). Specifically, the afore-mentioned research shows that the prediction of discourse markers between pairs of sentences can be exploited as a training signal that improves performance on existing classification datasets. We build on these results, but we look at the task from a different perspective: we make use of a model trained on discourse marker prediction in order to predict plausible discourse markers between sentence pairs from existing datasets, which are annotated with the correct semantic categories. Specifically, we explore the following questions:

- which semantic categories are applicable to a particular discourse marker (e.g. is a marker like *but* associated with other semantic categories than just mere contrast)?
- which discourse markers can be associated with the semantic categories of different datasets (e.g. what are the most likely markers between two paraphrases)?
- to what extent do discourse markers differ between datasets with comparable semantic categories (e.g. for two sentiment analysis datasets, one on films and one on

product reviews, are the markers different)?

In order to answer the above-mentioned questions, we train a model for discourse marker prediction between sentence pairs, using millions of examples. We then use this model to predict markers between sentences whose semantic relationships have already been annotated—for example, pairs of sentences (s_1,s_2,y) where y is in *Paraphrase*, *Non-Paraphrase*. These predictions allow us to examine the relationship between each category y and the discourse markers that are most often predicted for that category.

Thus, we propose *DiscSense*, a mapping between markers and senses, that has several applications:

- It explains why it is useful to employ discourse marker prediction as a training signal for sentence representation learning
- The characterization of discourse markers with categories provides new knowledge about the connotation of discourse markers; Our characterization is arguably richer since it does not only use PDTB categories. For instance, our mapping shows that the use of some markers is associated with negative sentiment or sarcasm; this might be useful in writing-aid contexts, or as a resource for second language learners; it could also be used to guide linguistic analyses of markers
- The characterization of categories with discourse markers can help "diagnosing" a classification dataset; As shown in table 8.2 below, SICK/MNLI dataset categories have different associations and our method can provide a sanity check for annotations (e.g. a Contradiction class should be mapped to markers expected to denote a contradiction)

8.2 Related work

Previous work has amply explored the link between discourse markers and semantic categories. Pitler et al. (2008b), for example, use the PDTB to analyze to what extent discourse

markers *a priori* reflect relationship category. Asr and Demberg (2012) have demonstrated that particular relationship categories give rise to more or less presence of discourse markers. And a recent categorization of discourse markers for English is provided in the DimLex lexicon (Das et al., 2018).

As mentioned before, discourse markers have equally been used as a learning signal for the prediction of implicit discourse relations (Liu et al., 2016; Braud and Denis, 2016) and inference relations (Pan et al., 2018b). This work has been generalized by DiscSent (Jernite et al., 2017), DisSent (Nie et al., 2019), and Discovery (Sileo et al., 2019) which has been presented chapter 6 which use discourse markers to learn general representations of sentences, which are transferable to various NLP classification tasks. However, none of these examine the individual impact of markers on these tasks.

8.3 Experimental setup

8.3.1 Discourse marker corpus

In order to train a model to predict plausible discourse markers between sentence pairs, we use the English *Discovery* (Sileo et al., 2019) presented in chapter 6, as it has the richest set of markers. It is composed of 174 discourse markers with 20K usage examples for each (sentence pairs where the second sentence begins by a given marker). Sentence pairs were extracted from web data (Panchenko et al., 2017), and the markers come either from the PDTB or from an automatic extraction method based on heuristics. An example of the dataset is provided in (1).

(1) Which is best? Undoubtedly, that depends on the person.
s_1 c s_2

For reasons of comparison, we equally use another dataset by Malmi et al. (2018) which contains 20K usage examples for 20 markers extracted from Wikipedia articles (the 20 markers are a subset of the markers considered in the *Discovery* dataset); we call this

dataset *Wiki20*.

We plan to use marker prediction on sentence pairs from classification datasets, in which some sentence pairs cannot plausibly occur consecutively, for instance two entirely unrelated sentences. Therefore, we augment the *Discovery* dataset with non-consecutive sentence pairs from the DepCC corpus, that were separated by 2 to 100 sentences. Besides, we also want to predict markers beginning single sentences, so the first sentence of example pairs is masked in 10% of cases by replacing it with a special symbol, which will be used as a placeholder for predictions of single sentences as in the CR (Customer Reviews) dataset.

8.3.2 Classification datasets

We leverage classification datasets from DiscEval (see chapter 7), alongside GLUE classification tasks (see section 4.4.2) augmented with SUBJ, CR and SICK tasks from SentEval (see section 4.4.1) in order to have a different domains for sentiment analysis and NLI. We also map STS semantic similarity estimation task from GLUE/SentEval into a classification task by casting the ratings into three quantiles and discarding the middle quantile.

8.3.3 Model

For our experiments, we make use of a state of the art NLP model for language understanding, viz. BERT (Devlin et al., 2019),[1] which is a text encoder pre-trained using language modeling. The parameters are initialized with the pre-trained unsupervised *base-uncased* model and then fine-tuned using the Adam (Kingma and Ba, 2014) optimizer with 2 iterations on our corpus data, using default hyperparameters otherwise. We ran experiments using BERT on both *Discovery* and *Wiki20*.

[1] https://github.com/huggingface/pytorch-pretrained-BERT/

8.4 Results

8.4.1 Marker prediction accuracy

Table 8.1 shows the results of the different models on the prediction of discourse markers. The accuracy of BERT on the *Discovery* test data is quite high given the large number of classes (174, perfectly balanced) and sometimes their low semantic distinguishability. This accuracy is significantly higher than the score of the Bi-LSTM model of the chapter 6 setup. The BERT model finetuned on *Discovery* outperforms human performance reported on *Wiki20* with no other adaptation than discarding markers not in *Wiki20* during inference.[2] With a further step of fine-tuning (1 epoch on *Wiki20*), we also outperform the best model from Malmi et al. (2018). These results suggest that the BERT+Discovery model captures a significant part of the use of discourse markers; in the following section, we will apply it to the prediction of discourse markers for individual categories.

	Wiki20	Discovery
Majority Class	5.0	0.6
Human Raters	23.1	-
Decomposable Attention	31.8	-
Bi-LSTM	-	22.2
BERT+Discovery	30.6	**32.9**
BERT+Discovery+Wiki20	**47.6**	-

Table 8.1 – Discourse marker prediction accuracy percentages on *Wiki20* and *Discovery* datasets. Human Raters and Decomposable Attention are from Malmi et al. (2018). Bi-LSTM is from chapter 6 and the last two are ours as well.

8.4.2 Prediction of markers associated to semantic categories

For each semantic dataset, consisting of either annotated sentences (s_1, y) or annotated sentence pairs (s_1,s_2,y), where y is a category, we use the BERT+Discovery model to predict the most plausible marker m in each example. The classification datasets thus yield a list

[2] But note that there is some overlap between training data since BERT pretraining uses Wikipedia text.

marker	category	support	confidence (prior)
unfortunately	CR.neg	64	94.1 (36.2)
initially	CR.neg	25	61.0 (36.2)
sadly	SST2.neg	622	87.4 (44.2)
unfortunately	SST2.neg	260	85.8 (44.2)
in contrast	MNLI.contradiction	1138	73.4 (33.3)
curiously	MNLI.contradiction	2835	70.6 (33.3)
technically	SICKE.contradiction	28	39.4 (14.8)
only	SICKE.contradiction	204	35.3 (14.8)
similarly	MRPC.paraphrase	75	57.3 (67.7)
likewise	MRPC.paraphrase	92	54.8 (67.7)
clearly	PDTBf.Cause	50	56.8 (26.7)
additionally	PDTBf.Conjunction	41	59.4 (22.5)
but	PDTBf.Contrast	76	55.5 (12.4)
elsewhere	PDTBf.List	38	15.3 (02.6)
specifically	PDTBf.Restatement	78	65.0 (18.8)
seriously	SARC.sarc	173	61.1 (49.9)
surely	SARC.sarc	37	60.7 (49.9)

Table 8.2 – Categories and most associated markers. CR.neg denotes the negative class in the CR dataset. Support is the number of examples where the marker was predicted given a dataset. Confidence is the estimated probability of the class given the prediction of the marker i.e. $P(y|m)$. The prior is $P(y)$.

of (y, m) pairs. We discard examples where no marker is predicted, and we discard markers that we predicted less than 20 times for a particular dataset. Table 8.2 shows a sample of markers with the highest probability of $P(y|m)$, i.e. the probability of a marker given a class. An extended table, which includes a larger sample of significant markers for all datasets included in our experiments, is available in appendix A.

The associations for some markers are intuitively correct (*likewise* denotes a semantic similarity expected in front of a paraphrase, *sadly* denotes a negative feeling, etc.) and they display a predictive power much higher than random choices. Other associations seem more surprising at first glance, for example, *seriously* as a marker of sarcasm—although on second thought, it seems a reasonable assumption that *seriously* does not actually signal a serious message, but rather a sarcastic comment on the preceding sentence. Generally speaking, we notice the same tendency for each class: our model predicts both fairly obvious markers (*unfortunately* as a marker for negative sentiment, *in contrast* for contradiction), but equally more inconspicuous markers (e.g. *initially* and *curiously* for the same respective categories) that are perfectly acceptable, even though they might have

sentence$_1$	sentence$_2$	marker
every act of god is holy because god is holy .	*every act of god is loving because god is love .*	likewise,
it gives you a schizophrenic feeling when trying to navigate a web page .	*it 's just a bad experience .*	sadly,
the article below was published a few months back .	*there is all too much truth in this .*	sadly,
i do n't think i can stop with the exclamation marks ! ! !	*this could be a problem ! ! ! !*	seriously,
yesterday she was elevated to " super stone whisperer " when we got out of the car and i heard , " wait , found 'em " .	*5 minutes in the cemetery , it has to be a record .*	seriously,
i am glad you tried to explain your viewpoint .	*i can tell you put some effort into that .*	seriously,
ayite , think of link building as brand building .	*there are no shortcuts .*	unfortunately,
does it make sense to you ?	*i am still struggling .*	unfortunately,
you will seldom meet new people .	*in medellin you will definitely meet people .*	in_contrast,
if i burn a fingertip , i 'll moan all night .	*it did n't look too bad .*	initially,
he puncture is about the size of a large pea .	*he can see almost no blood .*	curiously,

Table 8.3 – Examples of the Discovery datasets illustrating various relation senses

been missed by (and indeed are not present in) *a priori* approaches to discourse marker categorization. The associations seem to vary across domains (e.g. between CR and SST2) but some markers (e.g. *unfortunately*) seem to have more robust associations than others. Table **??** provides some Discovery samples where the marker are used accordingly.

On a related note, it is encouraging to see that the top markers predicted on the implicit PDTB dataset are similar to those present in the more recent English-DimLex lexicon which annotates PDTB categories as senses for discourse markers (Das et al., 2018).

This indicates that our approach is able to induce genuine discourse markers for discourse categories that coincide with linguistic intuitions; however, our approach has the advantage to lay bare less obvious markers, that might easily be overlooked by an *a priori* categoriza-

tion.

8.5 Conclusion

Based on a model trained for the prediction of discourse markers, we have established links between the categories of various semantically annotated datasets and discourse markers. Compared to *a priori* approaches to discourse marker categorization, our method has the advantage to reveal more inconspicuous but perfectly sensible markers for particular categories. The resulting associations can straightforwardly be used to guide corpus analyses, for example to define an empirically grounded typology of marker use. More qualitative analyses would be needed to elucidate subtleties in the most unexpected results. In further work, we plan to use the associations we found as a heuristic to choose discourse markers whose prediction is the most helpful for transferable sentence representation learning.

CHAPTER 9

CONCLUSION

The field of general purpose text representation has gained significant traction over the last few years. In this thesis, we took a step back and critiqued this progress on two complementary angles. We argued that expressive compositions are a condition of possibility of natural language understanding, and we framed pragmatics understanding as one of its goal.

9.1 Importance of composition

We show that analysing composition more thoroughly with criteria such as asymmetry or possibility of strong interaction has been overlooked. The *heuristic matching* composition and SentEval are still routinely used as of end of 2019, so this analysis is still relevant. The trend of transformers and integrated multi-level composition seems promising, but recent work (Shwartz and Dagan, 2019; Nandakumar et al., 2019) shows that the problem of lexical composition is still far from being solved.

9.2 Integration of pragmatics

Discourse and pragmatics are well established topics in NLP, but we found this importance not to carry over into the paradigm of generalisable natural language language understanding and evaluation.

Our three contributions on discourse are closely linked and oriented towards that goal. DiscEval reveals strengths of Discovery, but by doing so Discovery reveals the importance of DiscEval, because of the gap between previous versions of BERT and BERT+Discovery, it actually discloses the room of improvement for these previous models. On the other hand, DiscSense is made possible both Discovery and DiscEval, but also draws a link between them, and yields an interpretation of the performance of the Discovery dataset.

We proposed practical and publicly available resources for the integration of pragmatics into state of the art systems. Their adoption could allow a more comprehensive view on these systems and impact real use cases. Discovery English dataset and marker extraction method has been used in state of the art NLP (Sun et al., 2019).

9.3 Future work

Whether representations should be focused towards semantics or pragmatics is an interesting question, but it is arguable that current representations are mostly reliant of low level statistical cues (McCoy et al., 2019; Niven and Kao, 2019) that constitute neither semantic nor pragmatics understanding yet.

Many open problems remain, such as the evaluation of understanding as more complex structures than just shallow parsing. This evaluation could incentivize the need of more complex training data as well. Discovery could be easily extended to broader context (e.g. prediction of marker m in sentences sequences s_1, s_2, m, s_3, s_4). In addition, it would be worthwhile to investigate whether pragmatics integration can emphasize bias in the learned representation (May et al., 2019), since it relies on presuppositions.

LIST OF FIGURES

LIST OF TABLES

CHAPTER 10

RÉSUMÉ LONG

10.1 Introduction

De nombreuses tâches de l'intelligence artificielle impliquent l'exploitation d'un texte par un modèle pour résoudre diverses taches (analyse de sentiment (figure10-1b), détection de similarité (figure10-1b), agent conversationnel (figure10-1c).) La figure 10-1 illustre certaines d'entre elles.

L'accomplissement de ces multiples tâches implique en partie la résolution des mêmes sous-problèmes inhérents au texte : représenter les mots, les composer en prenant en compte notamment la syntaxe et les idiomes, interpréter le contenu des phrases en fonction de ce que le contexte et le sens commun rendent vraisemblable. Une tendance de plus en plus répandue consiste à décomposer les modèles en deux parties :

– un modèle dédié aux problèmes récurrents de la compréhension du texte, qui renvoie des primitives réutilisables (représentations vectorielles, fonctions), et qui peut être appelé encodeur de texte

– un modèle plus spécialisé, dédié à l'apprentissage des particularités de la tâche (typiquement une régression logistique qui pondère les primitives réutilisables).

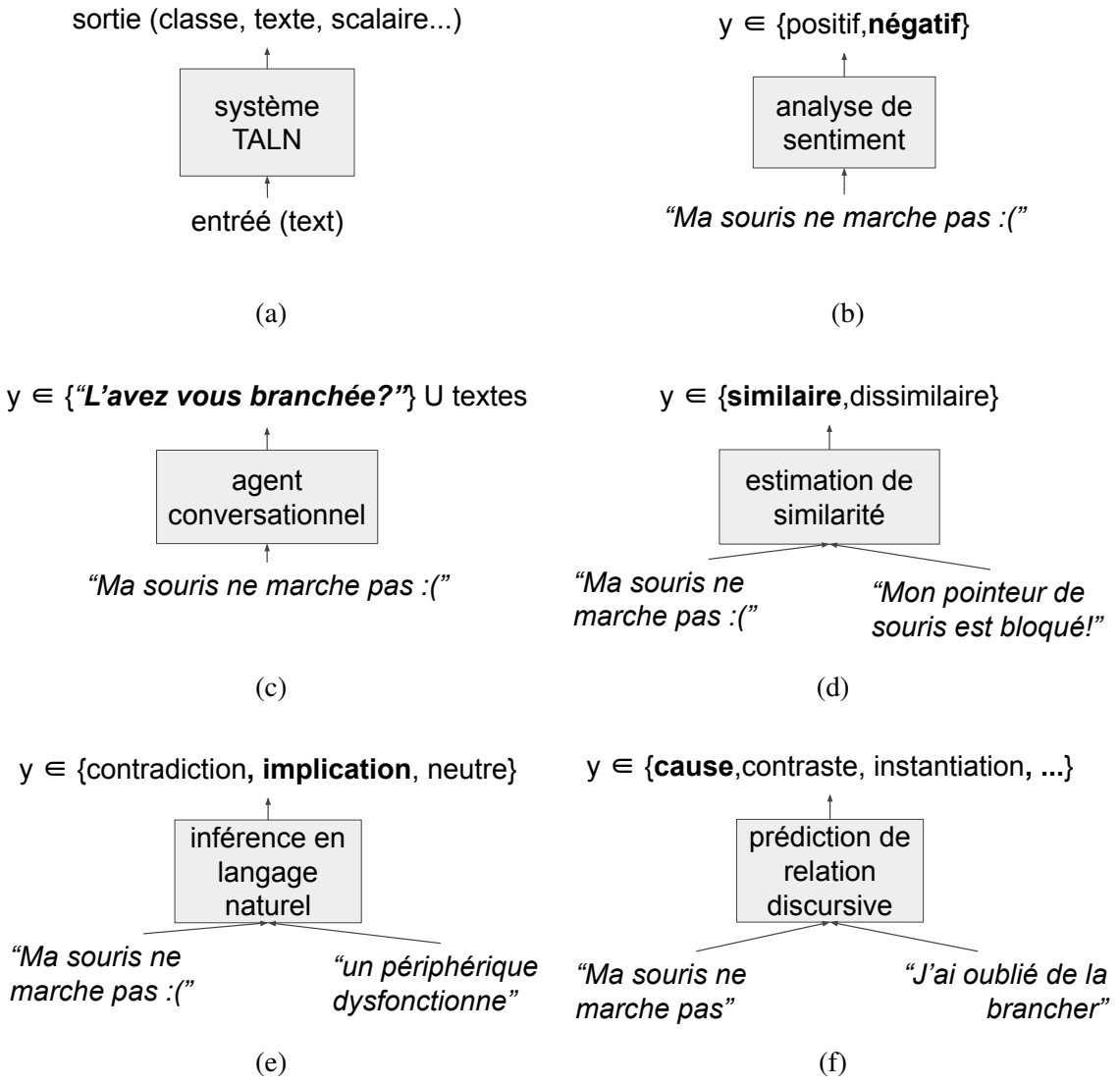


Figure 10-1 – Systèmes de traitement des langues pour plusieurs tâches

Un des avantages de cette décomposition est la possibilité d’utiliser un encodeur de texte qui a déjà été entrainé sur une certaine tâche, ce qui facilite l’apprentissage d’autres tâches, surtout si elles sont similaires. Cette pratique, appelée *apprentissage par transfert* consiste à utiliser une tâche *source* afin d’entrainer un modèle, puis de réutiliser le modèle entrainé sur une tâche dite *cible* qui est celle que l’on a réellement besoin de résoudre. La tâche source peut être vue comme un prétexte pour que le modèle acquisse des capacités intéressantes. Une analogie possible est l’utilisation de simulateur d’avions par des humains pour apprendre à piloter des avions dans le monde réel.

L'utilisation d'une tâche source permet donc d'entrainer des modèles de représentation du texte, qui peuvent être réutilisés pour mieux réaliser des taches cibles.

Des tâches dites sources se détachent par leur performance pour l'entrainement de représentations réutilisables. Certaines d'entre elles sont non-supervisées, c'est à dire ne nécessitant pas d'annotations réalisées par des humains, c'est le cas de la tâche de prédiction de mots sachant un contexte (aussi appelée modèle de langue)

De la même manière, des tâches cibles ont été retenues pour l'évaluation de représentation génériques de textes par la communauté du traitement automatique du language naturel (TALN). Ces tâches sont regroupées dans des bancs d'évaluation tels que SentEval (Conneau and Bordes, 2017) ou GLUE (Wang et al., 2019b). Ces tâches incluent la similarité sémantique, l'analyse de sentiment, l'inférence en langage naturel et la détection de paraphrases.

Cela dit, ces tâches d'évaluation ont été choisies par la dynamique de la recherche en TALN et pas directement en correspondance avec les utilisations monde réel. Pourtant, ces jeux de données, SentEval et GLUE sont utilisés pour affirmer que des encodeurs de texte produisent des représentations *universelles* du sens d'un texte.

Mais avant d'affirmer cela, il semble nécessaire de clarifier ce qu'on entend par le sens d'un texte.

En philosophie du langage, deux aspects principaux se détâchent. L'aspect sémantique du sens correspond à l'interprétation littérale d'un énoncé textuel ; des formules logiques peuvent être tirées de cette interprétation, et ces formules peuvent être confrontées à l'état du monde, ce qui peut permettre de qualifier la vérité d'un énoncé.

L'aspect pragmatique du sens est davantage lié à la finalité d'un énoncé : à quoi sert son énonciation ? La phrase suivante : ex:piedfr Tu es sur mon pied. exprime plus que son contenu littéral, par exemple une volonté d'écarter son interlocuteur de soi. Lorsqu'on raisonne sur les finalités des énoncés, le sens commun et la complexité de la psychologie humaine sont à prendre en compte. L'analyse du discours est une discipline qui fournit des outils conceptuels pour analyser les aspects pragmatiques, de la même manière que la

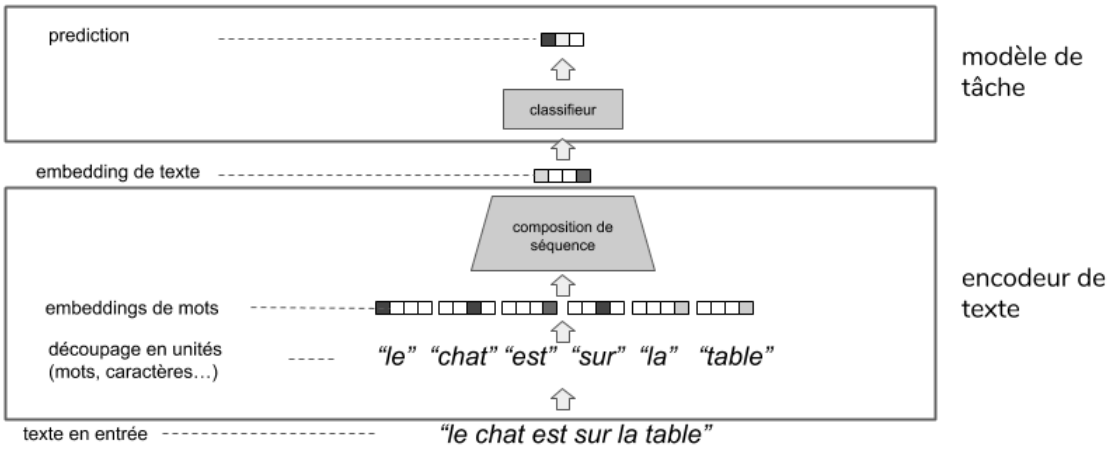


Figure 10-2 – Anatomie d'un encodeur d texte

logique permet de formaliser la sémantique.

L'examen des jeux d'évaluation existants semble montrer que les bancs d'évaluation sont focalisés principalement sur la sémantique. Cela dit, quand bien même les données d'entrainement et d'évaluation et d'entrainement seraient adaptés aux usages, cela ne suffirait pas à garantir le succès de la production de représentations universelles, puisque de nombreux verrous technologiques peuvent se situer notamment au niveau des capacités de généralisation des modèles, et de l'expressivité des modèles, c'est-à-dire la capacité à résoudre des tâches. Commençons donc par introduire brièvement les modèles sur lesquels on se base :

10.1.1 Modèles neuronaux pour le traitement des langues

Une méthode devenue la norme pour représenter des entrées textuelles consiste à diviser un texte en unités élémentaires (e.g; mots, caractères), à représenter ces unités à l'aide de représentations vectorielles, puis à composer ces représentations vectorielles, soit directement pour la résolution d'une tâche, soit pour produire une représentation de phrase qui peut servir à estimer des similarités entre pharses, ou servir comme une entrée pour finalement résoudre des tâches cibles. La figure 10-2 illustre cette chaîne de traitement.

Le modèle de composition de séquence peut notamment être un réseau de neurones récurrents,

un réseau convolutif ou un Transformer. Ce module compose les embeddings des unités découpées dans le texte afin de produire une représentation vectorielle du texte. Ce sont ces modèles qui conditionnent la capacité et l'expressivité des systèmes de TALN.

10.1.2 Contributions de la thèse

Cette thèse aborde un aspect de l'expressivité des modèles, puis s'attache à intégrer les aspects pragmatiques du langage dans l'entrainement et l'évaluation des modèles de compréhension automatique du langage naturel. Le manuscrit de cette thèse s'articule autour de 4 articles dont les contributions se focalisent sur la compréhension générale du language naturel. La section10.2 questionne les compositions vectorielles utilisées dans les travaux dans le cadre de la composition de représentation de phrases, et montre des failles dans des compositions répandues qui interviennent notamment dans l'évaluation des modèles. La section 10.3 décrit la conception d'un jeu de données basé sur la prédiction de marqueurs de discours nommé *Discovery* permettant d'améliorer l'entrainement des modèles neuronaux de représentation de texte. La section 10.4 propose d'utiliser le discours pour évaluer la compréhension automatique du texte afin de mieux évaluer les aspects pragmatiques, avec la centralisation de plusieurs jeu de données vers un bac d'évaluation nommé *DiscEval*. La section 10.5 fait le lien entre les marqueurs de discours et diverses tâches, notamment de DiscEval, et les marqueurs de discours du jeu de donnée Discovery, en proposant une manière automatique d'associer les marqueurs à des catégories de diverse tâches, ce qui permet d'obtenir une sémantique des marqueurs de discours et d'expliquer le succès de l'utilisation de Discovery.

10.2 Expressivité des compositions de représentations vectorielles

La prédiction de relation est un type de tâche répandu dans le TALN. La similarité sémantique, la détection de relations sémantiques (contradiction, implication) la cohérence de discours,

la prédiction de relations discursives, peuvent se formuler comme tel. Une manière simple et populaire de traiter ces tâches et de produire des représentations vectorielles de phrases h_1, h_2, et de composer ces vecteurs au moyen de fonctions de compostions. Par exemple, la fonction de composition $f_\odot : h_1, h_2 \rightarrow h1 \odot h_2$ associe aux deux vecteurs d'entrée une représentation jointe qui permet de représenter l'association de ses entrées h_1 et h_2. Ces fonctions conditionnent de manière critique le fonctionnement global d'un modèle. En effet, si choisit de réprésenter la relation entre deux phrases en se limitant à la fonction $f_\odot$, il devient alors impossible de modéliser des relations asymétriques correctement. En effet, $f_\odot(h_1, h_2)$ est égal à $f_\odot(h_2, h_1)$, ce qui signifie qu'un modèle basé sur cette composition aura la même représentation de relation. Il en va de même pour valeur absolue de la différence $f_- : h_1, h_2 \rightarrow |h_1 - h2|$. Une autre fonction répandue est la concaténation $f_{1,2} : h_1, h_2 \rightarrow [h_1, h_2]$. Cette fonction est asymétrique mais ne permet pas de modéliser les interactions non-additives entre les deux phrases si elle est donnée en entrée à un modèle linéaire. L'une des fonctions les plus couramment utilisée est la suivante:

$$f_{\odot,-,1,2} = [h_1 \odot h_2, |h_1, h_2|, h1, h2] \tag{10.1}$$

Cette fonction, utilisée dans la librairie d'évaluation d'embeddings de phrase SentEval, est certes asymétrique, certes capable de modéliser des interactions non-additives entre les phrases, mais pas les deux en même temps, ce qui est problématique pour l'expressivité et donc l'évaluation. En s'inspirant du champ de l'apprentissage relationnel(Getoor and Taskar, 2007), on propose les fonctions définies comme suit:

$$f_t(h_1, h_2) = |h_2 - h_1 + t|, \text{ avec } t \in \mathbb{R}^d \tag{10.2}$$

$$f_{\mathbb{C}}(h_1, h_2) = [h_1^{\mathcal{R}} \odot h_2^{\mathcal{R}} + h_1^{\mathcal{I}} \odot h_2^{\mathcal{I}}, h_1^{\mathcal{R}} \odot h_2^{\mathcal{I}} - h_1^{\mathcal{I}} \odot h_2^{\mathcal{R}}] \tag{10.3}$$

$\mathcal{R}$ et $\mathcal{C}$ pouvant correspondre aux indices pairs et impairs de h. Ces deux fonctions peuvent êtres vues comme des généralisations respectives de f_- et $f_\odot$ et sont à la fois asymmétriques et capables de gérer des interactions non-additives. Les utiliser en remplacement sur SentEval peut conduire à des gains significatifs mais assez faibles malgré le gain d'expressivité,

ce qui peut être vu comme une remise en cause du système plus global (encodeurs choisis, pertinence des jeux de données d'évaluation). Ainsi, ces résultats peuvent justifier le fait que les modèles plus récents (Devlin et al., 2019) obtiennent de meilleurs scores en s'affranchissant de l'utilisation d'un vecteur unique pour représenter les phrases.

10.3 Extraction de marqueurs de discours pour l'apprentissage non-supervisé

Les marqueurs de discours (*donc, mais, par conséquent, historiquement*) existent par centaines et peuvent être vus comme annotations bruitées pour une multitude de phénomènes linguistiques. Puisqu'ils servent souvent à annoncer l'usage d'une phrase dans son texte, ils peuvent être vus comme une supervision pragmatique, de la même manière que la tâche d'ILN peut être vue comme une supervision axée sur la sémantique. La tâche d'apprentissage est la suivante: à partir de deux phrases s_1 et s_2, prédire le marqueur de discours m qui les liait initialement.

Une telle idée a déjà été mise en application (Nie et al., 2019), cependant, l'aspect pragmatique semble ne pas avoir été exploité au mieux. En effet, la grande variété des usages signalés par les marqueurs de discours ne se retrouve pas dans les 15 marqueurs proposés dans DisSent, d'autant que la répartition est déséquilibrée illustrée dans la figure 10-3 , ensuite, la syntaxe peut jouer un rôle dans la prédiction des marqueurs : when et so n'apparaissent pas dans les mêmes contextes

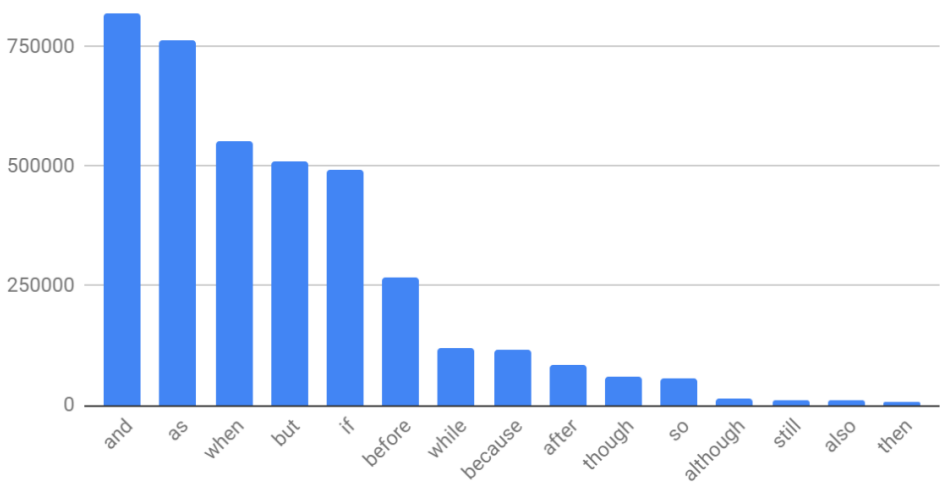


Figure 10-3 – Frequency distribution of markers of discourse markers in Nie et al. (2019)

On propose une méthode plus simple; mais permettant de traiter bien plus de marqueurs de discours: se focaliser sur les phrases consécutives s_1, s_2 où s_2 commence soit par un adverbe, soit par un marqueur de discours déjà identifié dans le PDTB. En se basant sur le corpus WebCC (Panchenko et al., 2017) dont on extrait 8.5 milliards de phrases, on sélectionne les phrases consécutives satisfaisant cette condition et on les considère comme des candidats de phrases séparées par des marqueurs.

On utilise ensuite un modèle linéaire simple, i.e. FastText, pour identifier les adverbes qui peuvent être prédits facilement, et donc, en vertu de la linéarité, sans faire intervenir la relation entre les deux phrases. Les adverbes les plus faciles à prédire, comme *very* sont souvent autre chose que des marqueurs de discours et sont éliminés du jeu de données. Cette technique permet de manière automatique de sélectionner des adverbes susceptibles d'être des marqueurs de discours, pour former le jeu de données Discovery. L'utilisation de Discovery plutôt que DisSent sur l'évaluation de SentEval conduit à un gain moyen substantiel de 1.3% .

10.4 Evaluation discursive et pragmatique pour la compréhension du langage naturel

Les jeux de données SentEval et GLUE se reposent majoritairement sur des tâches de similarité sémantique et d'inférence en langage naturel. L'inférence en langage naturel est devenue une tâche de référence pour entrainer les modèles neuronaux de compréhension du texte, mais ce choix est guidé par l'évaluation. SentEval et GLUE constituent-ils une évaluation complète de la compréhension du langage naturel ? L'inférence en langage naturel est une tâche qui sous-tend les aspects sémantiques du sens des textes. Or, les aspects pragmatiques sont d'une importance capitale puisque par définition elles traitent le sens sous l'angle de l'utilisation finale.

Afin de remédier à ce manque dans les évaluations existantes, on propose un banc d'évaluation construit sur des tâches déjà existantes mais jamais rassemblées par une évaluation unifiée:

PDTB, STAC et **GUM** contiennent des paires de phrases liées par des relations de discours parmi des catégories fixées (e.g. contaste, élaboration) **Emergent** des paires de phrase, munies d'une annotation décrivant la position de la seconde phrase par rapport à un énoncé dans la première.

SwitchBoard et **MRDA** regroupent des transcriptions de dialogues, les actes de dialogues étant annotés par diverses classes (e.g. commentaire, offre, appréciation).

Persuasion rassemble des phrases tirées de dissertation, avec des annotations sur leurs degrés de spécificité, d'éloquence ou de pertinence, qui sont autant de facteurs de persuasion.

SarcasmV2 regroupe des paires de messages issues de discussions en ligne, le 2eme message répondant de manière sarcastique ou non au premier, ce qui a fait l'objet d'une annotation

Squinky regroupe des annotations de texte sur leur degré de formalité, d'informativité et la quantité d'information présupposée

Verifiability regroupe des phrases qui ont été annotées selon leur verifiabilité.

EmoBank aggrège des annotations sur des textes concernant la valence (ce dont en parle habituellement en analyse de sentiment), le degré de dominance s'en dégageant, et le degré d'excitation s'en dégageant.

	PDTB	STAC	GUM	Emergent	SwitchB.	MRDA	Persuasion	Sarcasm	Squinky	Verif.	EmoBank
CBoW	27.4	32	20.5	59.7	3.8	0.7	70.6	61.1	75.5	74	64
BiLSTM	25.9	27.7	18.5	45.6	3.7	0.7	62.6	63.1	72.1	74	63.5
BiLSTM+ELMo	27.5	33.5	18.9	55.2	3.7	0.7	67.4	68.9	82.5	74	66.9
BERT	48.8	48.2	40.9	79.2	38.8	22.3	74.8	**77.1**	87.5	86.7	76.2
BERT+MNLI	49.1	49.1	42.8	81.2	38.1	22.7	71.7	73.4	88.2	86	76.3
BERT+DiscEval	49.1	**57.1**	42.8	80.2	**40.3**	**23.1**	**76.2**	75	87.6	85.9	76
BERT+DisSent	49.4	49	43.9	79.8	39.2	22	74.7	74.9	87.5	85.9	76.2
B+DisSent+MNLI	49.6	49.2	**44.2**	80.9	39.8	22.1	74	74.1	87.6	85.6	76.4
BERT+Discovery	50.7	49.5	42.7	**81.7**	39.5	22.4	71.6	76.7	88.6	86.3	**76.6**
B+Discovery+MNLI	**51.3**	49.4	43.1	80.7	**40.3**	22.2	73.6	75.1	**88.9**	**86.8**	76
Human estimate	**84.0**	-	**69.3**	-	-	-	-	**80.0**	-	**87.0**	73.1

Table 10.1 – Evaluations des modèles de l'état de l'art sur les tâches de DiscEval B(ERT)+$\mathcal{X}$ désigne un modèle BERT dont les poids ont été entrainés sur la tâche $\mathcal{X}$.)

La table 10.1 montre les résultats de différents modèles sur DiscEval. Il est intéressant de constater que l'utilisation d'inférence en langage naturel comme tâche d'entrainement n'est pas la source de représentations réellement universelle puisqu'elle conduit à une régression sur plusieurs tâches. Par ailleurs, le jeu de données Discovery conduit en moyenne au meilleurs résultats, en particulier lorsqu'il est combiné à MNLI dans un apprentissage multi-tâche.

10.5 Sémantique des marqueurs de discours

La section précédente donne du poids à l'idée selon laquelle les marqueurs de discours sont des annotations bruitées. Mais à quel point ces formes d'annotations sont-elles bruitées ? Peut-on aller plus loin dans la vérification des liens entre les marqueurs de discours et les Des lexiques construits manuellement permettent de répertorier les conditions d'utilisation des marqueurs de discours, par exemple que le marqueur Mais sert à exprimer une relation

de contraste entre deux unités de discours, le sens des relations (e.g. contraste) étant défini par un guide d'annotation comme le PDTB.

On propose de construire une sémantique basée sur l'utilisation réelle des marqueurs. Pour ce faire, on entraine un modèle BERT pour lui apprendre prédire des marqueurs entre des phrases. Ayant obtenu un modèle précis, (possiblement plus précis que des annotateurs humains), on utilise de modèle pour prédire des marqueurs entre des phrases liées par une relation déjà annotée dans un jeu de donné. Par exemple, le jeu de donnée MNLI qui rassemble des phrases liées par des relations de contradiction, implication ou aucun des deux (neutre). La prédiction de marqueurs entre ces paires de phrases permet alors d'analyser le lien entre catégories et marqueurs, et de montrer par exemple dans quelle proportion des exemples le marqueur *but* dénote une contradiction logique. Un échantillon de cette sémantique des marqueurs est présenté dans la table 10.2 et de manière plus complète dans l'annexe A. Cette méthode est transposable à d'autres langues, et peut permettre de sélectionner des données pour de l'apprentissage non supervisé, guider des analyses linguistiques, et qualifier les connotations des marqueurs dans une optique de rédaction ou d'apprentissage d'une langue.

10.6 Conclusion

Dans cette thèse, on a exploré plusieurs apects de la compréhension automatique du langage naturel. Les modèles neuronaux semblent aller vers plus d'expressivité grâce aux Transformers, mais notre analyse montre que cette expressivité doit être guidé par une notion de sens tournée vers l'usage concret et situé dans le monde ; les jeux de données Discovery et DiscEval sont des contributions concrètes permettant d'aller dans cette direction.

marqueur	catégorie	support	confiance (prior)
unfortunately,	CR.negative	66	100.0 (21.8)
sadly,	CR.negative	20	95.2 (21.8)
unfortunately,	SST-2.negative	240	96.0 (22.5)
as a result,	SST-2.negative	65	94.2 (22.5)
in contrast,	MNLI.contradiction	1182	74.1 (16.9)
curiously,	MNLI.contradiction	2912	70.8 (16.9)
technically,	SICKE.contradiction	29	87.9 (7.8)
rather,	SICKE.contradiction	147	69.7 (7.8)
similarly,	MRPC.paraphrase	85	87.6 (35.5)
likewise,	MRPC.paraphrase	103	84.4 (35.5)
instead,	PDTB.Alternative	27	22.5 (0.6)
then,	PDTB.Asynchronous	60	38.7 (2.4)
previously,	PDTB.Asynchronous	36	36.4 (2.4)
by doing this,	PDTB.Cause	22	61.1 (14.8)
additionally	PDTB.Conjunction	47	63.5 (12.5)
but	PDTB.Contrast	89	61.4 (7.0)
elsewhere,	PDTB.List	41	16.2 (1.3)
specifically,	PDTB.Restatement	100	67.6 (10.6)
seriously,	SarcasmV2.sarcasm	225	71.2 (26.7)
so,	SarcasmV2.sarcasm	81	65.6 (26.7)

Table 10.2 – Echantillon du jeu de donnée DiscSense. CR.negative dénote la classe du jeu de donnée Customer Reviews (CR). Le support est le nombre d'exemples des jeux de données pour lesquels le marqueur de discours donné a été prmis. La confiance est la probabilité estimée de catégorie sachant le marqueur. Version complète ici: https://github.com/synapse-developpement/DiscSense.

APPENDIX A

DISCSENSE MAPPING

antecedents	consequents	support	confidence+prior
unfortunately,	CR.neg	66.0	100.0 (36.6)
sadly,	CR.neg	20.0	95.2 (36.6)
regardless,	CR.pos	31.0	96.9 (63.4)
fortunately,	CR.pos	27.0	96.4 (63.4)
meaning,	Cola.not-well-formed	21.0	48.8 (29.7)
on the contrary,	Cola.not-well-formed	50.0	45.9 (29.7)
regardless,	Cola.well-formed	23.0	95.8 (70.3)
undoubtedly,	Cola.well-formed	25.0	92.6 (70.3)
only,	Emergent.against	24.0	88.9 (15.5)
meantime,	Emergent.against	21.0	24.1 (15.5)
normally,	Emergent.for	22.0	78.6 (47.5)
later,	Emergent.for	89.0	78.1 (47.5)
separately,	Emergent.observing	148.0	59.2 (37.0)
now,	Emergent.observing	35.0	50.0 (37.0)
anyway,	EmoBankA.high	24.0	85.7 (48.5)
suddenly,	EmoBankA.high	103.0	73.6 (48.5)
originally,	EmoBankA.low	27.0	90.0 (51.5)

antecedents	consequents	support	confidence+prior
presently,	EmoBankA.low	24.0	80.0 (51.5)
together,	EmoBankD.high	20.0	62.5 (38.7)
absolutely,	EmoBankD.high	68.0	62.4 (38.7)
inevitably,	EmoBankD.low	21.0	91.3 (61.3)
only,	EmoBankD.low	31.0	81.6 (61.3)
plus,	EmoBankV.high	45.0	90.0 (43.2)
hopefully,	EmoBankV.high	26.0	86.7 (43.2)
sadly,	EmoBankV.low	36.0	92.3 (56.8)
frankly,	EmoBankV.low	33.0	89.2 (56.8)
separately,	Formality.high	295.0	100.0 (48.2)
significantly,	Formality.high	73.0	100.0 (48.2)
well,	Formality.low	49.0	100.0 (51.8)
seriously,	Formality.low	35.0	100.0 (51.8)
this,	GUM.circumstance	24.0	35.3 (57.8)
especially,	GUM.circumstance	40.0	30.5 (57.8)
or,	GUM.condition	31.0	50.0 (42.2)
especially,	GUM.condition	41.0	31.3 (42.2)
instead,	Implicature.high	28.0	77.8 (46.7)
absolutely,	Implicature.high	36.0	70.6 (46.7)
by comparison,	Implicature.low	24.0	88.9 (53.3)
strangely,	Implicature.low	22.0	81.5 (53.3)
significantly,	Informativeness.high	57.0	100.0 (46.3)
altogether,	Informativeness.high	31.0	100.0 (46.3)
seriously,	Informativeness.low	37.0	100.0 (53.7)
really,	Informativeness.low	25.0	100.0 (53.7)
in contrast,	MNLI.contradiction	1182.0	74.1 (33.3)
curiously,	MNLI.contradiction	2912.0	70.8 (33.3)
in turn,	MNLI.entailment	7475.0	65.4 (33.4)
likewise,	MNLI.entailment	2430.0	63.0 (33.4)
for instance	MNLI.neutral	177.0	70.8 (33.3)

antecedents	consequents	support	confidence+prior
for example	MNLI.neutral	170.0	70.0 (33.3)
so,	MRDA.Accept	57.0	12.9 (3.1)
well,	MRDA.Acknowledge-answer	85.0	10.3 (3.6)
well,	MRDA.Action-directive	20.0	2.4 (1.5)
actually,	MRDA.Affirmative Non-yes Answers	37.0	12.2 (3.0)
personally,	MRDA.Assessment/Appreciation	25.0	15.9 (4.0)
especially,	MRDA.Collaborative Completion	25.0	7.4 (2.2)
really,	MRDA.Declarative-Question	48.0	11.9 (1.4)
mostly,	MRDA.Defending/Explanation	114.0	62.3 (10.9)
probably,	MRDA.Dispreferred Answers	25.0	1.5 (1.3)
namely,	MRDA.Expansions of y/n Answers	37.0	33.6 (8.5)
so,	MRDA.Floor Grabber	56.0	12.7 (4.4)
and	MRDA.Floor Holder	53.0	8.2 (4.9)
and	MRDA.Hold Before Answer/Agreement	26.0	4.0 (1.1)
absolutely,	MRDA.Interrupted/Abandoned/Uninterpretable	24.0	1.2 (1.2)
probably,	MRDA.Negative Non-no Answers	28.0	1.7 (0.9)
though,	MRDA.Offer	27.0	18.9 (7.1)
honestly,	MRDA.Other Answers	31.0	36.0 (1.2)
actually,	MRDA.Reject	34.0	11.2 (0.9)
probably,	MRDA.Reject-part	20.0	1.2 (0.3)
also,	MRDA.Rising Tone	66.0	36.7 (7.0)
originally,	MRDA.Statement	20.0	37.0 (22.1)
surely,	MRDA.Understanding Check	26.0	40.6 (5.2)
realistically,	MRDA.Wh-Question	24.0	27.6 (2.3)
or,	MRDA.Yes-No-question	61.0	16.1 (1.7)
elsewhere,	MRPC.not-paraphrase	30.0	81.1 (32.5)
meanwhile,	MRPC.not-paraphrase	21.0	61.8 (32.5)
similarly,	MRPC.paraphrase	85.0	87.6 (67.5)
likewise,	MRPC.paraphrase	103.0	84.4 (67.5)
but	PDTB.Comparison	97.0	52.4 (6.5)

antecedents	consequents	support	confidence+prior
however	PDTB.Comparison	27.0	38.6 (6.5)
by doing this,	PDTB.Contingency	22.0	57.9 (11.7)
theoretically,	PDTB.Contingency	70.0	50.4 (11.7)
currently,	PDTB.Entrel	212.0	63.5 (13.5)
generally,	PDTB.Entrel	31.0	53.4 (13.5)
for instance	PDTB.Expansion	179.0	77.5 (23.4)
similarly,	PDTB.Expansion	47.0	74.6 (23.4)
then,	PDTB.Temporal	62.0	36.7 (2.4)
later,	PDTB.Temporal	44.0	31.2 (2.4)
rather,	PDTBf.Alternative	36.0	25.4 (1.1)
instead,	PDTBf.Alternative	27.0	22.5 (1.1)
then,	PDTBf.Asynchronous	60.0	38.7 (4.5)
previously,	PDTBf.Asynchronous	36.0	36.4 (4.5)
by doing this,	PDTBf.Cause	22.0	61.1 (27.0)
so,	PDTBf.Cause	38.0	55.9 (27.0)
additionally	PDTBf.Conjunction	47.0	63.5 (22.8)
meanwhile,	PDTBf.Conjunction	167.0	55.5 (22.8)
but	PDTBf.Contrast	89.0	61.4 (12.8)
by comparison,	PDTBf.Contrast	214.0	45.4 (12.8)
for instance	PDTBf.Instantiation	138.0	65.1 (8.7)
for example	PDTBf.Instantiation	32.0	51.6 (8.7)
elsewhere,	PDTBf.List	41.0	16.2 (2.4)
meanwhile,	PDTBf.List	25.0	8.3 (2.4)
specifically,	PDTBf.Restatement	100.0	67.6 (19.5)
essentially,	PDTBf.Restatement	61.0	54.5 (19.5)
separately,	PDTBf.Synchrony	21.0	2.8 (1.2)
moreover	PersuasivenessEloquence.high	21.0	46.7 (26.5)
hence,	PersuasivenessEloquence.low	21.0	84.0 (73.5)
by doing this,	PersuasivenessEloquence.low	21.0	80.8 (73.5)
undoubtedly,	PersuasivenessPremiseType.common knowledge	24.0	85.7 (100.0)

antecedents	consequents	support	confidence+prior
moreover	PersuasivenessPremiseType.common knowledge	35.0	83.3 (100.0)
for instance	PersuasivenessRelevance.high	25.0	67.6 (59.9)
moreover	PersuasivenessRelevance.high	29.0	64.4 (59.9)
undoubtedly,	PersuasivenessRelevance.low	21.0	56.8 (40.1)
especially,	PersuasivenessRelevance.low	20.0	37.0 (40.1)
for instance	PersuasivenessSpecificity.high	24.0	82.8 (45.9)
moreover	PersuasivenessSpecificity.high	21.0	72.4 (45.9)
undoubtedly,	PersuasivenessSpecificity.low	20.0	87.0 (54.1)
undoubtedly,	PersuasivenessStrength.low	20.0	87.0 (100.0)
especially,	PersuasivenessStrength.low	23.0	59.0 (100.0)
likewise,	QNLI.entailment	38.0	74.5 (50.0)
actually,	QNLI.entailment	48.0	68.6 (50.0)
regardless,	QNLI.not entailment	29.0	87.9 (50.0)
thirdly,	QNLI.not entailment	23.0	85.2 (50.0)
collectively,	QQP.duplicate	45.0	68.2 (36.9)
indeed,	QQP.duplicate	113.0	67.3 (36.9)
anyway,	QQP.not-duplicate	87.0	100.0 (63.1)
namely,	QQP.not-duplicate	50.0	100.0 (63.1)
technically,	RTE.entailment	56.0	72.7 (50.4)
in turn,	RTE.entailment	83.0	66.9 (50.4)
by comparison,	RTE.not entailment	29.0	67.4 (49.6)
incidentally,	RTE.not entailment	38.0	58.5 (49.6)
technically,	SICKE.contradiction	29.0	87.9 (14.8)
rather,	SICKE.contradiction	147.0	69.7 (14.8)
in turn,	SICKE.entailment	32.0	64.0 (28.9)
alternately,	SICKE.entailment	93.0	59.6 (28.9)
meanwhile,	SICKE.neutral	155.0	92.8 (56.4)
elsewhere,	SICKE.neutral	765.0	89.8 (56.4)
unfortunately,	SST-2.neg	240.0	96.0 (44.3)
as a result,	SST-2.neg	65.0	94.2 (44.3)

antecedents	consequents	support	confidence+prior
nonetheless	SST-2.pos	383.0	93.4 (55.7)
nevertheless	SST-2.pos	56.0	90.3 (55.7)
so,	STAC.Acknowledgement	40.0	21.3 (10.6)
absolutely,	STAC.Acknowledgement	162.0	20.3 (10.6)
so,	STAC.Clarification question	23.0	12.2 (2.9)
really,	STAC.Clarification question	54.0	11.9 (2.9)
however	STAC.Comment	91.0	48.7 (10.9)
overall,	STAC.Comment	25.0	32.5 (10.9)
otherwise,	STAC.Conditional	21.0	25.0 (1.1)
anyway,	STAC.Continuation	52.0	10.4 (6.2)
and	STAC.Continuation	21.0	8.2 (6.2)
probably,	STAC.Contrast	76.0	18.9 (3.9)
maybe,	STAC.Contrast	35.0	18.3 (3.9)
alternately,	STAC.Elaboration	22.0	59.5 (7.7)
personally,	STAC.Elaboration	23.0	17.2 (7.7)
especially,	STAC.Explanation	21.0	12.4 (4.0)
anyway,	STAC.Explanation	36.0	7.2 (4.0)
really,	STAC.Q Elab	147.0	32.5 (4.6)
or,	STAC.Q Elab	41.0	19.3 (4.6)
surprisingly,	STAC.Question answer pair	71.0	89.9 (19.7)
sadly,	STAC.Question answer pair	249.0	77.3 (19.7)
finally,	STAC.Result	130.0	46.9 (6.3)
once,	STAC.Result	70.0	44.6 (6.3)
finally,	STAC.Sequence	29.0	10.5 (0.9)
currently,	STAC.no relation	50.0	65.8 (21.3)
eventually,	STAC.no relation	74.0	59.7 (21.3)
elsewhere,	STS.dissimilar	516.0	70.0 (26.2)
meantime,	STS.dissimilar	124.0	65.3 (26.2)
in turn,	STS.similar	142.0	60.2 (34.0)
specifically,	STS.similar	25.0	51.0 (34.0)

antecedents	consequents	support	confidence+prior
presently,	SUBJ.objective	24.0	100.0 (49.8)
soon,	SUBJ.objective	159.0	99.4 (49.8)
frankly,	SUBJ.subjective	127.0	100.0 (50.2)
again,	SUBJ.subjective	65.0	100.0 (50.2)
technically,	Sarcasm.notsarcasm	34.0	72.3 (50.1)
secondly,	Sarcasm.notsarcasm	27.0	69.2 (50.1)
seriously,	Sarcasm.sarcasm	225.0	71.2 (49.9)
so,	Sarcasm.sarcasm	82.0	65.6 (49.9)
well,	SwitchBoard.Acknowledge (Backchannel)	30.0	2.8 (1.2)
seriously,	SwitchBoard.Action-directive	25.0	4.6 (2.1)
only,	SwitchBoard.Affirmative Non-yes Answers	20.0	3.0 (1.7)
actually,	SwitchBoard.Agree/Accept	64.0	17.3 (3.7)
actually,	SwitchBoard.Appreciation	58.0	15.7 (4.7)
especially,	SwitchBoard.Collaborative Completion	38.0	10.1 (2.5)
anyway,	SwitchBoard.Conventional-closing	82.0	39.4 (2.9)
surely,	SwitchBoard.Declarative Yes-No-Question	22.0	20.2 (4.0)
or,	SwitchBoard.Dispreferred Answers	24.0	1.7 (0.6)
honestly,	SwitchBoard.Hedge	24.0	19.7 (1.6)
so,	SwitchBoard.Hold Before Answer/Agreement	24.0	2.5 (1.1)
only,	SwitchBoard.Negative Non-no Answers	43.0	6.4 (0.8)
so,	SwitchBoard.Open-Question	85.0	8.8 (1.6)
well,	SwitchBoard.Other	36.0	3.4 (0.8)
or,	SwitchBoard.Other Answers	25.0	1.8 (0.5)
absolutely,	SwitchBoard.Quotation	89.0	6.2 (3.1)
especially,	SwitchBoard.Repeat-phrase	24.0	6.4 (1.3)
or,	SwitchBoard.Rhetorical-Question	48.0	3.4 (1.9)
so,	SwitchBoard.Self-talk	22.0	2.3 (0.4)
really,	SwitchBoard.Signal-non-understanding	37.0	5.6 (0.5)
luckily,	SwitchBoard.Statement-non-opinion	20.0	71.4 (15.6)
personally,	SwitchBoard.Statement-opinion	43.0	20.4 (5.2)

antecedents	consequents	support	confidence+prior
meaning,	SwitchBoard.Summarize/Reformulate	26.0	6.9 (3.0)
this,	SwitchBoard.Uninterpretable	158.0	56.0 (19.2)
realistically,	SwitchBoard.Wh-Question	48.0	33.8 (5.7)
incidentally,	SwitchBoard.Yes-No-Question	32.0	78.0 (14.5)
coincidentally,	Verifiability.experiential	20.0	80.0 (14.2)
recently,	Verifiability.experiential	23.0	76.7 (14.2)
especially,	Verifiability.non-experiential	36.0	39.1 (15.6)
unfortunately,	Verifiability.non-experiential	28.0	22.8 (15.6)
third,	Verifiability.unverifiable	23.0	100.0 (70.3)
ideally,	Verifiability.unverifiable	227.0	99.6 (70.3)

Table A.1 – Categories and most associated markers. CR.neg denotes the negative class in the CR dataset. (Supp)ort is the number of examples where the marker was predicted given a dataset. (Conf)idence is the estimated probability of the class given the prediction of the marker i.e. $P(y|m)$. The prior is $P(y)$.

BIBLIOGRAPHY

Agirre, E., Diab, M., Cer, D., and Gonzalez-Agirre, A. (2012). Semeval-2012 task 6: A pilot on semantic textual similarity. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, pages 385–393. Association for Computational Linguistics.

Allen, J. and Core, M. (1997). Draft of DAMSL: Dialog act markup in several layers.

Ampomah, I. K. E., Park, S.-b., and Lee, S.-j. (2016). A Sentence-to-Sentence Relation Network for Recognizing Textual Entailment. *World Academy of Science, Engineering and Technology International Journal of Computer and Information Engineering*, 10(12):1955–1958.

Arpit, D., Jastrzebski, S., Ballas, N., Krueger, D., Bengio, E., Kanwal, M. S., Maharaj, T., Fischer, A., Courville, A. C., Bengio, Y., and Lacoste-Julien, S. (2017). A closer look at memorization in deep networks. *ArXiv*, abs/1706.05394.

Asher, N. and Lascarides, A. (2003). *Logics of conversation*. Cambridge University Press.

Asr, F. T. and Demberg, V. (2012). Implicitness of Discourse Relations. In *COLING*.

Athiwaratkun, B. and Wilson, A. G. (2017). Multimodal word distributions. In *Conference of the Association for Computational Linguistics (ACL)*.

Austin, J. L. (1962). *How to do things with words*. William James Lectures. Oxford University Press.

Austin, J. L. (1975). *How to do things with words*. Oxford university press.

Badene, S., Thompson, K., Lorré, J.-P., and Asher, N. (2019). Data programming for learning discourse structure. In *Proceedings of the 57th Conference of the Association for Computational Linguistics*, pages 640–645, Florence, Italy. Association for Computational Linguistics.

Baroni, M., Bernardi, R., and Zamparelli, R. (2014a). Frege in space: A program of compositional distributional semantics. *LiLT (Linguistic Issues in Language Technology)*, 9.

Baroni, M., Dinu, G., and Kruszewski, G. (2014b). Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 238–247, Baltimore, Maryland. Association for Computational Linguistics.

Barzilay, R. and Lapata, M. (2008). Modeling Local Coherence: An Entity-Based Approach. *Computational Linguistics*, 34(1):1–34.

Baudiš, P., Pichl, J., Vyskočil, T., and Šedivý, J. (2016). Sentence Pair Scoring: Towards Unified Framework for Text Comprehension.

Baxter, J. (2000). A model of inductive bias learning. *Journal of artificial intelligence research*, 12:149–198.

Belinkov, Y. and Glass, J. (2019). Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 7:49–72.

Bender, E. M. (2011). On achieving and evaluating language-independence in nlp. *Linguistic Issues in Language Technology*, 6(3):1–26.

Bird, S., Klein, E., and Loper, E. (2009). *Natural Language Processing with Python*, volume 43.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Bordes, A., Glorot, X., Weston, J., and Bengio, Y. (2013a). A Semantic Matching Energy Function for Learning with Multi-relational Data. *Machine Learning*.

Bordes, A., Usunier, N., Weston, J., and Yakhnenko, O. (2013b). Translating Embeddings for Modeling Multi-Relational Data. *Advances in NIPS*, 26:2787–2795.

Bowman, S. and Zhu, X. (2019). Deep learning for natural language inference. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials*, pages 6–8, Minneapolis, Minnesota. Association for Computational Linguistics.

Bowman, S. R. (2016). *Modeling natural language semantics in learned representations*. PhD thesis.

Bowman, S. R., Angeli, G., Potts, C., and Manning, C. D. (2015a). A large annotated corpus for learning natural language inference. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing,Lisbon, Portugal, 17-21 September 2015*, (September):632–642.

Bowman, S. R., Angeli, G., Potts, C., and Manning, C. D. (2015b). A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*.

Bowman, S. R., Manning, C. D., and Potts, C. (2015c). Tree-structured composition in neural networks without tree-structured architectures. In *Proceedings of the 2015th International Conference on Cognitive Computation: Integrating Neural and Symbolic Approaches - Volume 1583*, COCO'15, pages 37–42, Aachen, Germany, Germany. CEUR-WS.org.

Bowman, S. R., Vilnis, L., Vinyals, O., Dai, A. M., Jozefowicz, R., and Bengio, S. (2016). Generating Sentences from a Continuous Space. *Iclr*, pages 1–13.

Brahma, S. (2018). Unsupervised Learning of Sentence Representations Using Sequence Consistency. *CoRR*, abs/1808.04217.

Braud, C. and Denis, P. (2016). Learning Connective-based Word Representations for Implicit Discourse Relation Identification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 203–213, Austin, Texas. Association for Computational Linguistics.

Bride, A., Van de Cruys, T., and Asher, N. (2015). A generalisation of lexical functions for composition in distributional semantics. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 281–291, Beijing, China. Association for Computational Linguistics.

Buechel, S. and Hahn, U. (2017). {E}mo{B}ank: Studying the Impact of Annotation Perspective and Representation Format on Dimensional Emotion Analysis. In *Proceedings of the 15th Conference of the {E}uropean Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 578–585, Valencia, Spain. Association for Computational Linguistics.

Carlile, W., Gurrapadi, N., Ke, Z., and Ng, V. (2018). Give Me More Feedback: Annotating Argument Persuasiveness and Related Attributes in Student Essays. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 621–631, Melbourne, Australia. Association for Computational Linguistics.

Carlson, L., Marcu, D., and Okurowski, M. E. (2001). Building a Discourse-tagged Corpus in the Framework of Rhetorical Structure Theory. In *Proceedings of the Second SIGdial Workshop on Discourse and Dialogue - Volume 16*, SIGDIAL '01, pages 1–10, Stroudsburg, PA, USA. Association for Computational Linguistics.

Caruana, R. (1997). Multitask learning. *Machine learning*, 28(1):41–75.

Cer, D., Yang, Y., Kong, S.-y., Hua, N., Limtiaco, N., John, R. S., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., Sung, Y.-H., Strope, B., and Kurzweil, R. (2018a). Universal Sentence Encoder.

Cer, D., Yang, Y., Kong, S.-y., Hua, N., Limtiaco, N., St. John, R., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., Strope, B., and Kurzweil, R. (2018b). Universal Sentence Encoder for {E}nglish. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 169–174, Brussels, Belgium. Association for Computational Linguistics.

Chang, H.-S., Wang, Z., Vilnis, L., and McCallum, A. (2018). Distributional Inclusion Vector Embedding for Unsupervised Hypernymy Detection. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 485–495, New Orleans, Louisiana. Association for Computational Linguistics.

Chen, D., Peterson, J. C., and Griffiths, T. L. (2017a). Evaluating vector-space models of analogy. Proceeding:0–5.

Chen, Q., Zhu, X., Ling, Z.-H., Wei, S., Jiang, H., and Inkpen, D. (2017b). Enhanced LSTM for Natural Language Inference. In Barzilay, R. and Kan, M.-Y., editors, *ACL (1)*, pages 1657–1668. Association for Computational Linguistics.

Chen, Y., Perozzi, B., Al-Rfou', R., and Skiena, S. (2013). The expressive power of word embeddings. *ArXiv*, abs/1301.3226.

Chung, J., Gulcehre, C., Cho, K., and Bengio, Y. (2014). Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. *arXiv*, pages 1–9.

Clark, S. (2015). *Vector Space Models of Lexical Meaning*, chapter 16, pages 493–522. John Wiley Sons, Ltd.

Collobert, R. and Weston, J. (2008). A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pages 160–167. ACM.

Conneau, A. and Bordes, A. (2017). Supervised Learning of Universal Sentence Representations from Natural Language Inference Data. pages 681–691.

Conneau, A., Kruszewski, G., Lample, G., Barrault, L., and Baroni, M. (2018a). What you can cram into a single vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136. Association for Computational Linguistics.

Conneau, A., Rinott, R., Lample, G., Williams, A., Bowman, S., Schwenk, H., and Stoyanov, V. (2018b). Xnli: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485.

Danlos, L., Colinet, M., and Steinlin, J. (2015). FDTB1: Repérage des connecteurs de discours dans un corpus français. *Discours - Revue de linguistique, psycholinguistique et informatique*, (17).

Das, D., Scheffler, T., Bourgonje, P., and Stede, M. (2018). Constructing a Lexicon of {English} Discourse Connectives. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 360–365, Melbourne, Australia. Association for Computational Linguistics.

Dasgupta, I., Guo, D., Stuhlmüller, A., Gershman, S. J., and Goodman, N. D. (2018). Evaluating Compositionality in Sentence Embeddings. (2011).

de Marneffe, M.-C., Manning, C. D., and Potts, C. (2012). Did It Happen? The Pragmatic Complexity of Veridicality Assessment. *Comput. Linguist.*, 38(2):301–333.

de Marneffe, M.-C., Simons, M., and Tonhauser, J. (2019). The commitmentbank: Investigating projection in naturally occurring discourse. *Proceedings of Sinn und Bedeutung*, 23(2):107–124.

Deleuze, G. and Guattari, F. (1988). *A thousand plateaus: Capitalism and schizophrenia*. Bloomsbury Publishing.

Dennett, D. C. (1989). *The intentional stance*.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (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 Papers)*. Association for Computational Linguistics.

Dolan, B., Quirk, C., and Brockett, C. (2004). Unsupervised Construction of Large Paraphrase Corpora: Exploiting Massively Parallel News Sources. In {*COLING*} *2004, 20th International Conference on Computational Linguistics, Proceedings of the Conference, 23-27 August 2004, Geneva, Switzerland*.

Dolan, R. J. (2002). Emotion, cognition, and behavior. *science*, 298(5596):1191–1194.

Dreyfus, H. L. (2007). Why Heideggerian AI Failed and How Fixing it Would Require Making it More Heideggerian. *Philosophical Psychology*, 20(2):247–268.

Ferreira, W. and Vlachos, A. (2016). Emergent: a novel data-set for stance classification. In *HLT-NAACL*.

Field, H. (1972). Tarski's theory of truth. *The Journal of Philosophy*, 69(13):347–375.

Firth, J. R. (1957). A synopsis of linguistic theory, 1930-1955. *Studies in linguistic analysis*.

Fraser, B. (1996). Pragmatic markers. *Pragmatics. Quarterly Publication of the International Pragmatics Association (Ipra)*, 6(2):167–190.

Frege, G. (1884). *Grundlagen der Arithmetik*. Breslau: Wilhelm Koebner.

Fu, R., Guo, J., Qin, B., Che, W., Wang, H., and Liu, T. (2014). Learning Semantic Hierarchies via Word Embeddings. *Acl*, pages 1199–1209.

Gauthier, J. and Ivanova, A. (2018). Does the brain represent words? An evaluation of brain decoding studies of language understanding. *arXiv preprint arXiv:1806.00591*.

Getoor, L. and Taskar, B. (2007). *Introduction to Statistical Relational Learning (Adaptive Computation and Machine Learning)*. The MIT Press.

Ghaeini, R., Hasan, S. A., Datla, V., Liu, J., Lee, K., Qadir, A., Ling, Y., Prakash, A., Fern, X., and Farri, O. (2018). DR-BiLSTM: Dependent reading bidirectional LSTM for natural language 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)*, pages 1460–1469, New Orleans, Louisiana. Association for Computational Linguistics.

Glockner, M., Shwartz, V., and Goldberg, Y. (2018). Breaking NLI Systems with Sentences that Require Simple Lexical Inferences. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Short Papers)*, (3):1–6.

Godfrey, J. J., Holliman, E. C., and McDaniel, J. (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, pages 517–520, Washington, DC, USA. IEEE Computer Society.

Gong, Y., Luo, H., and Zhang, J. (2018). Natural Language Inference over Interaction Space. In *International Conference on Learning Representations*.

Grave, E., Bojanowski, P., Gupta, P., Joulin, A., and Mikolov, T. (2018). Learning Word Vectors for 157 Languages. In Chair), N. C. C., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Hasida, K., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S., and Tokunaga, T., editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Green, M. S. (2000). Illocutionary Force and Semantic Content. *Linguistics and Philosophy*, 23(5):435–473.

Grice, H. P. (1975). Logic and Conversation. In Cole, P. and Morgan, J. L., editors, *Syntax and Semantics: Vol. 3: Speech Acts*, pages 41–58. Academic Press, New York.

Groenendijk, J. and Stokhof, M. (1991). Dynamic predicate logic. *Linguistics and philosophy*, 14(1):39–100.

Gururangan, S., Swayamdipta, S., Levy, O., Schwartz, R., Bowman, S. R., and Smith, N. A. (2018). Annotation Artifacts in Natural Language Inference Data.

Halliday, M. (1985). *An Introduction to Functional Grammar*. Edward Arnold Press, Baltimore.

Hochreiter, S., Bengio, Y., Frasconi, P., Schmidhuber, J., et al. (2001). Gradient flow in recurrent nets: the difficulty of learning long-term dependencies.

Hochreiter, S. and Schmidhuber, J. (1997). Lstm can solve hard long time lag problems. In *Advances in neural information processing systems*, pages 473–479.

Hooda, S. and Kosseim, L. (2017). Argument Labeling of Explicit Discourse Relations using LSTM Neural Networks.

Hovy, E. and Maier, E. (1992). Parsimonious or profligate: How many and which discourse structure relations? Technical Report RR-93-373, USC Information Sciences Institute.

Howard, J. and Ruder, S. (2018). Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, Melbourne, Australia. Association for Computational Linguistics.

Jernite, Y., Bowman, S. R., and Sontag, D. (2017). Discourse-Based Objectives for Fast Unsupervised Sentence Representation Learning.

Joshi, M., Choi, E., Levy, O., Weld, D. S., and Zettlemoyer, L. (2019). pair2vec: Compositional Word-Pair Embeddings for Cross-Sentence Inference. In *NAACL*.

Joulin, A., Grave, E., Bojanowski, P., and Mikolov, T. (2017). Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain. Association for Computational Linguistics.

Kamp, H., Van Genabith, J., and Reyle, U. (2011). Discourse representation theory. In *Handbook of philosophical logic*, pages 125–394. Springer.

Karttunen, L. (1974). Presupposition and linguistic context. *Theoretical linguistics*, 1(1-3):181–194.

Kaur, A. and Gupta, V. (2013). A survey on sentiment analysis and opinion mining techniques. *Journal of Emerging Technologies in Web Intelligence*, 5(4):367–371.

Kingma, D. and Ba, J. (2014). Adam: A Method for Stochastic Optimization. *International Conference on Learning Representations*, pages 1–13.

Kiros, J. and Chan, W. (2018). {I}nfer{L}ite: Simple Universal Sentence Representations from Natural Language Inference Data. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4868–4874, Brussels, Belgium. Association for Computational Linguistics.

Kiros, R., Zhu, Y., Salakhutdinov, R. R., Zemel, R., Urtasun, R., Torralba, A., and Fidler, S. (2015). Skip-thought vectors. In *Advances in neural information processing systems*, pages 3294–3302.

Knott, A. (1996). *A data-driven methodology for motivating a set of coherence relations*. PhD thesis, University of Edinburgh, {UK}.

Kroll, J. F. (1980). Comprehension and memory in rapid sequential reading. *Attention and performance VIII*, 8:395.

Lacroix, T., Usunier, N., and Obozinski, G. (2018). Canonical Tensor Decomposition for Knowledge Base Completion. In *ICML*.

Lahiri, S. (2015). SQUINKY! A Corpus of Sentence-level Formality, Informativeness, and Implicature. *CoRR*, abs/1506.02306.

Lascarides, A. and Asher, N. (2008). Segmented discourse representation theory: Dynamic semantics with discourse structure. In *Computing meaning*, pages 87–124. Springer.

Läubli, S., Sennrich, R., and Volk, M. (2018). Has Machine Translation Achieved Human Parity? A Case for Document-level Evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4791–4796, Brussels, Belgium. Association for Computational Linguistics.

Le, Q. and Mikolov, T. (2014). Distributed Representations of Sentences and Documents. *International Conference on Machine Learning - ICML 2014*, 32:1188–1196.

Levesque, H., Davis, E., and Morgenstern, L. (2012). The winograd schema challenge. In *Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning*.

Levy, O. and Goldberg, Y. (2014a). Linguistic Regularities in Sparse and Explicit Word Representations. *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pages 171–180.

Levy, O. and Goldberg, Y. (2014b). Neural word embedding as implicit matrix factorization. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 27*, pages 2177–2185. Curran Associates, Inc.

Levy, O., Remus, S., Biemann, C., and Dagan, I. (2015). Do Supervised Distributional Methods Really Learn Lexical Inference Relations? *Naacl-2015*, pages 970–976.

Liu, H., Wu, Y., and Yang, Y. (2017). Analogical Inference for Multi-Relational Embeddings. *Icml*.

Liu, X., He, P., Chen, W., and Gao, J. (2019). Improving Multi-Task Deep Neural Networks via Knowledge Distillation for Natural Language Understanding. *arXiv preprint arXiv:1904.09482*.

Liu, Y. P., Li, S., Zhang, X., and Sui, Z. (2016). Implicit discourse relation classification via multi-task neural networks. *ArXiv*, abs/1603.02776.

Logeswaran, L., Lee, H., and Radev, D. (2016). Sentence Ordering using Recurrent Neural Networks. pages 1–15.

Logeswaran, L., Lee, H., and Radev, D. R. (2018). Sentence Ordering and Coherence Modeling using Recurrent Neural Networks. In *Proceedings of the Thirty-Second {AAAI} Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th {AAAI} Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 5285–5292.

Malmi, E., Pighin, D., Krause, S., and Kozhevnikov, M. (2018). Automatic Prediction of Discourse Connectives. In *Proceedings of the 11th Language Resources and Evaluation Conference*, Miyazaki, Japan. European Language Resource Association.

Mann, W. C. and Thompson, S. A. (1988). Rhetorical structure theory: Toward a functional theory of text organization. *Text-interdisciplinary Journal for the Study of Discourse*, 8(3):243–281.

Mantzavinos, C. (2016). Hermeneutics.

Marcu, D. and Echihabi, A. (2002). An unsupervised approach to recognizing discourse relations. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 368–375.

May, C., Wang, A., Bordia, S., Bowman, S. R., and Rudinger, R. (2019). On measuring social biases in sentence encoders. 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)*, pages 622–628, Minneapolis, Minnesota. Association for Computational Linguistics.

McCann, B., Keskar, N. S., Xiong, C., and Socher, R. (2018). The natural language decathlon: Multitask learning as question answering. *arXiv preprint arXiv:1806.08730*.

McCoy, R. T., Pavlick, E., and Linzen, T. (2019). Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *ACL*.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. *Nips*, pages 1–9.

Mitchell, J. and Lapata, M. (2008). Vector-based models of semantic composition. In *Proceedings of ACL-08: HLT*, pages 236–244, Columbus, Ohio. Association for Computational Linguistics.

Morey, M., Muller, P., and Asher, N. (2017). How much progress have we made on RST discourse parsing? a replication study of recent results on the RST-DT. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1319–1324, Copenhagen, Denmark. Association for Computational Linguistics.

Mou, L., Men, R., Li, G., Xu, Y., Zhang, L., Yan, R., and Jin, Z. (2016). Natural Language Inference by Tree-Based Convolution and Heuristic Matching. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 130–136, Berlin, Germany. Association for Computational Linguistics.

Mueller, J. and Thyagarajan, A. (2016). Siamese Recurrent Architectures for Learning Sentence Similarity. In *AAAI*, pages 2786–2792.

Nandakumar, N., Baldwin, T., and Salehi, B. (2019). How well do embedding models capture non-compositionality? a view from multiword expressions. In *Proceedings of the 3rd Workshop on Evaluating Vector Space Representations for NLP*, pages 27–34, Minneapolis, USA. Association for Computational Linguistics.

Nangia, N. and Bowman, S. R. (2019). Human vs. muppet: A conservative estimate of human performance on the GLUE benchmark. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4566–4575, Florence, Italy. Association for Computational Linguistics.

Nickel, M., Tresp, V., and Kriegel, H.-P. (2011). A Three-Way Model for Collective Learning on Multi-Relational Data. *Icml*, pages 809–816.

Nie, A., Bennett, E., and Goodman, N. (2019). DisSent: Learning sentence representations from explicit discourse relations. pages 4497–4510.

Niven, T. and Kao, H.-Y. (2019). Probing neural network comprehension of natural language arguments. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4658–4664, Florence, Italy. Association for Computational Linguistics.

Oraby, S., Harrison, V., Reed, L., Hernandez, E., Riloff, E., and Walker, M. (2016). Creating and Characterizing a Diverse Corpus of Sarcasm in Dialogue. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 31–41. Association for Computational Linguistics.

Pan, B., Yang, Y., Zhao, Z., Zhuang, Y., Cai, D., and He, X. (2018a). Discourse Marker Augmented Network with Reinforcement Learning for Natural Language Inference. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 989–999, Melbourne, Australia. Association for Computational Linguistics.

Pan, B., Yang, Y., Zhao, Z., Zhuang, Y., Cai, D., and He, X. (2018b). Discourse marker augmented network with reinforcement learning for natural language inference. In *ACL*.

Panchenko, A., Ruppert, E., Faralli, S., Ponzetto, S. P., and Biemann, C. (2017). Building a Web-Scale Dependency-Parsed Corpus from Common Crawl. pages 1816–1823.

Pang, B. and Lee, L. (2004). A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd annual meeting on Association for Computational Linguistics*, page 271. Association for Computational Linguistics.

Park, J. and Cardie, C. (2014). Identifying appropriate support for propositions in online user comments. In *Proceedings of the first workshop on argumentation mining*, pages 29–38.

Pascanu, R., Gulcehre, C., Cho, K., and Bengio, Y. (2014). How to construct deep recurrent neural networks. In *Proceedings of the Second International Conference on Learning Representations (ICLR 2014)*.

Pennington, J., Socher, R., and Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Perone, C. S., Silveira, R., and Paula, T. S. (2018). Evaluation of sentence embeddings in downstream and linguistic probing tasks. *CoRR*, abs/1806.06259.

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018a). Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018b). Deep contextualized word representations. In *Proc. of NAACL*.

Pfeffer, J. (1981). Understanding the role of power in decision making. *Jay M. Shafritz y J. Steven Ott, Classics of Organization Theory, Wadsworth*, pages 137–154.

Phang, J., Févry, T., and Bowman, S. R. (2018). Sentence Encoders on STILTs: Supplementary Training on Intermediate Labeled-data Tasks. *CoRR*, abs/1811.01088.

Piedeleu, R., Kartsaklis, D., Coecke, B., and Sadrzadeh, M. (2015). Open system categorical quantum semantics in natural language processing. *arXiv preprint arXiv:1502.00831*.

Pitler, E., Louis, A., and Nenkova, A. (2009). Automatic sense prediction for implicit discourse relations in text. *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP Volume 2 ACLIJCNLP 09*, 2(August):683–691.

Pitler, E. and Nenkova, A. (2009). Using Syntax to Disambiguate Explicit Discourse Connectives in Text. In {*ACL*} *2009, Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the AFNLP, 2-7 August 2009, Singapore, Short Papers*, pages 13–16.

Pitler, E., Raghupathy, M., Mehta, H., Nenkova, A., Lee, A., and Joshi, A. (2008a). Easily Identifiable Discourse Relations. In *Coling 2008: Companion volume: Posters*, pages 87–90. Coling 2008 Organizing Committee.

Pitler, E., Raghupathy, M., Mehta, H., Nenkova, A., Lee, A., and Joshi, A. K. (2008b). Easily identifiable discourse relations. *Technical Reports (CIS)*, page 884.

Poliak, A., Haldar, A., Rudinger, R., Hu, J. E., Pavlick, E., White, A. S., and Van Durme, B. (2018a). Collecting Diverse Natural Language Inference Problems for Sentence Representation Evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 67–81.

Poliak, A., Naradowsky, J., Haldar, A., Rudinger, R., and Durme, B. V. (2018b). Hypothesis Only Baselines in Natural Language Inference. *Proceedings of the 7th Joint Conference on Lexical and Computational Semantics*, (1):180–191.

Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A., and Webber, B. (2008). The Penn Discourse TreeBank 2.0. In Nicoletta Calzolari (Conference Chair) Khalid Choukri, B. M. J. M. J. O. S. P. D. T., editor, *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Prasad, R., Riley, K. F., and Lee, A. (2014). Towards Full Text Shallow Discourse Relation Annotation : Experiments with Cross-Paragraph Implicit Relations in the PDTB. (2009).

Pustejovsky, J. (2012). Co-compositionality in grammar. *The Oxford handbook of compositionality*, 371:382.

Pérez, J., Marinković, J., and Barceló, P. (2019). On the turing completeness of modern neural network architectures. In *International Conference on Learning Representations*.

Radford, A. (2018). Improving Language Understanding by Generative Pre-Training.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8).

Raghu, M., Poole, B., Kleinberg, J., Ganguli, S., and Dickstein, J. S. (2017). On the expressive power of deep neural networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2847–2854. JMLR. org.

Ribeiro, E., Ribeiro, R., and de Matos, D. M. (2015). The influence of context on dialogue act recognition. *arXiv preprint arXiv:1506.00839*.

Rocktäschel, T., Grefenstette, E., Hermann, K. M., Kociský, T., and Blunsom, P. (2015). Reasoning about entailment with neural attention. *CoRR*, abs/1509.06664.

Roze, C., Danlos, L., and Muller, P. (2012). LEXCONN: A French Lexicon of Discourse Connectives. *Discours*, (10).

Rutherford, A. and Xue, N. (2015). Improving the Inference of Implicit Discourse Relations via Classifying Explicit Discourse Connectives. In {*NAACL*} {*HLT*} *2015, The 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, Colorado, USA, May 31 - June 5, 2015*, pages 799–808.

Schuster, M. and Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681.

Searle, J. (2006). Chinese room argument, the. *Encyclopedia of cognitive science*.

Searle, J. R., Kiefer, F., Bierwisch, M., and Others (1980). *Speech act theory and pragmatics*, volume 10. Springer.

Searle, J. R., Willis, S., et al. (1983). *Intentionality: An essay in the philosophy of mind*. Cambridge university press.

Seo, M. J., Kembhavi, A., Farhadi, A., and Hajishirzi, H. (2017). Bidirectional Attention Flow for Machine Comprehension. In *5th International Conference on Learning Representations, {ICLR} 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*.

Seok, M., Song, H.-J., Park, C.-Y., Kim, J.-D., and Kim, Y.-S. (2016). Named Entity Recognition using Word Embedding as a Feature 1. *International Journal of Software Engineering and Its Applications*, 10(2):93–104.

Shen, D., Wang, G., Wang, W., Renqiang Min, M., Su, Q., Zhang, Y., Li, C., Henao, R., and Carin, L. (2018). Baseline needs more love: On simple word-embedding-based models and associated pooling mechanisms. In *ACL*.

Shriberg, E., Dhillon, R., Bhagat, S., Ang, J., and Carvey, H. (2004). The ICSI meeting recorder dialog act (MRDA) corpus. In *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue at HLT-NAACL 2004*.

Shwartz, V. and Dagan, I. (2019). Still a pain in the neck: Evaluating text representations on lexical composition. *Transactions of the Association for Computational Linguistics*, 7:403–419.

Sileo, D., Van De Cruys, T., Pradel, C., and Muller, P. (2019). Mining discourse markers for unsupervised sentence representation learning. 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)*, pages 3477–3486, Minneapolis, Minnesota. Association for Computational Linguistics.

Socher, R., Chen, D., Manning, C., Chen, D., and Ng, A. (2013). Reasoning With Neural Tensor Networks for Knowledge Base Completion. In *Neural Information Processing Systems (2003)*, pages 926–934.

Sporleder, C. and Lascarides, A. (2008). Using automatically labelled examples to classify rhetorical relations: an assessment. *Natural Language Engineering*, 14(3):369–416.

Stede, M. (2002). Di{M}{L}ex: A Lexical Approach to Discourse Markers. In *Exploring the Lexicon - Theory and Computation*. Edizioni dell'Orso, Alessandria.

Subramanian, S., Trischler, A., Bengio, Y., and Pal, C. J. (2018). Learning general purpose distributed sentence representations via large scale multi-task learning. *International Conference on Learning Representations*.

Sun, Y., Wang, S., Li, Y., Feng, S., Tian, H., Wu, H., and Wang, H. (2019). Ernie 2.0: A continual pre-training framework for language understanding.

Szabó, Z. G. (2017). Compositionality. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, summer 2017 edition.

Tai, K. S., Socher, R., and Manning, C. D. (2015). Improved semantic representations from tree-structured long short-term memory networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers*, pages 1556–1566.

Tamaazousti, Y. (2018). *On The Universality of Visual and Multimodal Representations*. Theses, Université Paris-Saclay.

Tay, Y., Luu, A. T., and Hui, S. C. (2018). Compare, compress and propagate: Enhancing neural architectures with alignment factorization for natural language inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1565–1575, Brussels, Belgium. Association for Computational Linguistics.

Torabi Asr, F., Zinkov, R., and Jones, M. (2018). Querying word embeddings for similarity and relatedness. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 675–684, New Orleans, Louisiana. Association for Computational Linguistics.

Trouillon, T., Welbl, J., Riedel, S., Gaussier, É., and Bouchard, G. (2016). Complex Embeddings for Simple Link Prediction. In *Proceedings of the 33nd International Conference on Machine Learning*, volume 48.

Van de Cruys, T. (2010). *Mining for Meaning. The Extraction of Lexico-Semantic Knowledge from Text*. PhD thesis, University of Groningen, The Netherlands.

van der Maaten, L. and Hinton, G. (2008). Visualizing Data using {t-SNE}. *Journal of Machine Learning Research*, 9:2579–2605.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Vilnis, L. and McCallum, A. (2015). Word Representations via Gaussian Embedding. *Iclr*, page 12.

Wang, A., Hula, J., Xia, P., Pappagari, R., McCoy, R. T., Patel, R., Kim, N., Tenney, I., Huang, Y., Yu, K., Jin, S., Chen, B., Durme, B. V., Grave, E., Pavlick, E., and Bowman, S. R. (2019a). Can You Tell Me How to Get Past Sesame Street? Sentence-Level Pretraining Beyond Language Modeling. In *ACL 2019*.

Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. R. (2019b). SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. (2018). {GLUE}: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *Proceedings of the 2018 {EMNLP} Workshop {B}lackbox{NLP}: Analyzing and Interpreting Neural Networks for {NLP}*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. R. (2019c). {GLUE}: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *International Conference on Learning Representations*.

Wang, A., Tenney, I. F., Pruksachatkun, Y., Yu, K., Hula, J., Xia, P., Pappagari, R., Jin, S., McCoy, R. T., Patel, R., Huang, Y., Phang, J., Grave, E., Liu, H., Kim, N., Htut, P. M., F'evry, T., Chen, B., Nangia, N., Mohananey, A., Kann, K., Bordia, S., Patry, N., Benton, D., Pavlick, E., and Bowman, S. R. (2019d). `jiant` 1.2: A software toolkit for research on general-purpose text understanding models. `http://jiant.info/`.

Watanabe, S. (1985). Theorem of the ugly duckling. *Pattern Recognition: Human and Mechanical*.

Wermter, S., Riloff, E., and Scheler, G. (1996). *Connectionist, statistical and symbolic approaches to learning for natural language processing*, volume 1040. Springer Science & Business Media.

Wieting, J., Bansal, M., Gimpel, K., and Livescu, K. (2015). Towards Universal Paraphrastic Sentence Embeddings. *CoRR*, abs/1511.08198.

Williams, A., Nangia, N., and Bowman, S. (2018a). 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)*, pages 1112–1122. Association for Computational Linguistics.

Williams, A., Nangia, N., and Bowman, S. (2018b). 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)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Wittgenstein, L. (2013). *Tractatus logico-philosophicus*. Routledge.

Wolf, M. J., Miller, K., and Grodzinsky, F. S. (2017). Why we should have seen that coming: comments on microsoft's tay experiment, and wider implications. *ACM SIGCAS Computers and Society*, 47(3):54–64.

Xue, N., Demberg, V., and Rutherford, A. (2017). A Systematic Study of Neural Discourse Models for Implicit Discourse Relation. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, {EACL} 2017, Valencia, Spain, April 3-7, 2017, Volume 1: Long Papers*, pages 281–291.

Yang, M. C., Lee, D. G., Park, S. Y., and Rim, H. C. (2015). Knowledge-based question answering using the semantic embedding space. *Expert Systems with Applications*, 42(23):9086–9104.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., and Le, Q. V. (2019). XLNet: Generalized Autoregressive Pretraining for Language Understanding.

Young, T., Hazarika, D., Poria, S., and Cambria, E. (2018). Recent trends in deep learning based natural language processing. *ieee Computational intelligenCe magazine*, 13(3):55–75.

Zeldes, A. (2017). The GUM corpus: Creating multilayer resources in the classroom. *Language Resources and Evaluation*, 51(3):581–612.

Zeyrek, D. and Webber, B. (2008). A Discourse Resource for Turkish: Annotating Discourse Connectives in the {METU} Corpus. In *Proceedings of IJCNLP*.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. (2016). Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*.

Zhou, Y., Lu, J., Zhang, J., and Xue, N. (2014). Chinese Discourse Treebank 0.5 {LDC2014T21}.