UPPSALA
UNIVERSITET

Improving Sense Classification in Shallow Discourse Parsing

Jimmy Callin

Uppsala University
Department of Linguistics and Philology
Master’s Programme in Language Technology
Master’s Thesis in Language Technology

February 4, 2016

Supervisors:
Joakim Nivre

Abstract

TODO

Contents

Preface

1 Introduction

Despite the progress in Natural Language Processing (NLP), one assumption remains steadfast: sentences are to be seen as atomical units. Machine Translation systems rarely look outside of its local context for translation clues, sentiment analysis modules have trouble dealing with juxtapositions, and information extraction systems often limit extracted units of fact to the immediate sentence. While this obviously is a faulty assumption, it has been difficult to release this constraint from the equations; the increasing recall has rarely been worth the much faster decreasing precision.

Lately, it has come to the attention that we might have started to reach a convergence for certain tasks if we do not start to look into the role of sentences in their environment. This is what *discourse parsing* is trying to achieve, by revealing an inherent structure between sentences within documents. That is, how does one text unit relate to the another text unit? Take for instance:

(1) *Boeing would make cost-of-living adjustments projected to be 5 for each year of the contract* <u>though</u> **the union has called the offer insulting**.

The connective unit *though* in Example 1 can be analyzed as a sort of binding block between the italized unit and the bolded, working as a *contrastive comparison* between them. This is an example of a *sense* of a discourse relation which labels each relation according to a given sense taxonomy. That said, not all connective units are necessarily as explicit:

(2) *No wonder*. **We were coming down straight into their canal.**

Here, despite the lack of a connective unit the two sentences clearly share a connection. The bolded sentence explains the previous one as *the reason* for the lack of wonder. We will expand upon the difference between implicit and explicit relations in section 2.1.1, as well as how we can possibly classify the sense of a relation.

1.1 The CoNLL-2016 Shared Task on Shallow Discourse Parsing

At CoNLL 2015 they introduced a new shared task on *Shallow Discourse Parsing*. In this task, a system is fed a piece of newswire text as input and returns discourse relations similar to our previous examples. Such a system needs to locate both explicit or implicit discourse connectives, identify the spans of its two relations, and finally predict the sense of the discourse connective.

The shared task for the Twentieth Conference on Computational Natural Language Learning (CoNLL-2016) continues upon the work of last year's shared task while introducing Chinese as an alternative evaluation language. A new component is the introduction of sense classification where correct argument dependencies and connective are already given, leaving out the sense of the discourse connective. The purpose of this is to allow more focus to be put into solely studying the properties of discourse connectives without having to worry about the pure parsing. In this work, we are first and foremost interested in the sense classification task.

1.2 Purpose

The purpose of this thesis is to explore methods of improving sense classification in English Shallow Discourse Parsing. We will explore the following research questions:

- How can we use continuous semantic representations to increase performance in such a task?
- What can we learn from other work in Natural Language Inference?

2 Background

2.1 Discourse

Explain what discourse is.

2.1.1 Shallow Discourse Parsing

Explain what shallow discourse parsing is.

Discourse connectives

See Table 2.1 for discourse type distribution. See Table 2.2 for sense hierarchy.

Explicit vs implicit discourse connectives

It is not always possible to consider implicit discourse relations simply as explicit relations with the connective removed.

(3) I want to go to New York, but I already booked a flight. (I am not going to New York.)

(4) I want to go to New York, so I already booked a flight. (I am going to New York.)

Here, we have either a COMPARISON.Contrast, or a CONTINGENCY.Cause relation depending on what form the sense takes. If we were to remove the connective token, our linguistic intuition tells us to default for the meaning of Example 4, that is, CONTINGENCY.Cause. How can we use this to our advantage when classifying implicit connectives? (Honest question, still don't know.)

2.2 Related work

Up until this task little focus had been given to this topic from a parsing perspective, something which reflected the work in the final contributions: all papers built upon the same discourse parser as presented by Lin et al. (2014) with little variance in the form of alternative architectures. Given the complexity of the task this is not surprising, since this allows contributors to work within a focus area. Lin et al. (2014) is the first PDTB-styled end-to-end discourse parser with a parsing pipeline that closely reflects the annotation pipeline of PDTB. We will have a closer look at the components of the parser in Section **??**.

Add a brief overview of results from CoNLL 2015.

Type	Frequency	Ratio
Explicit	14722	45.25%
Implicit	13156	40.44%
EntRel	4133	12.7%
AltLex	524	1.61%

Table 2.1: Discourse relation type distribution.

TEMPORAL	CONTINGENCY	COMPARISON	EXPANSION
Asynchronous Synchronous	Cause Condition	Concession Contrast	Alternative Conjunction Exception Instantiation List Restatement

Table 2.2: Sense hierarchy

2.3 Current plan

I want to focus on sense classification, mostly due to the fact that this is provided as a separate task with provided training and test data. If there is time left, I could implement this module into the previous year's best performing end-to-end discourse parser. After having studied the previous year's systems, I have come to the conclusion that there is a gap in representation learning approaches. While I probably will not be alone in this decision (just look at last year's EMNLP), I might be able to contribute with some non-obvious ideas.

There were only two systems that used semantic representations, both from University of Dublin, and both performing relatively weakly in the task. The other submissions are only using ”shallow” features. Okita et al. (2015) focuses on using paragraph vectors along with some additional features to improve sense classification. They obtain reasonable F-score for explicit senses (.88), while the implicit sense classification is quite bad (.11). The implicit sense classification is mainly bad due to low recall (.15), while the precision is on par with explicit sense classification. Furthermore, they find that the classification task struggles with classes with explicit connectives where the surface form is occurring in more than one class. The second paper by Wang et al. (2015) is not focusing on sense classification and thus not provide much detail about its results there. Overall, its results are somewhat disappointing.

Beyond last year's submissions, Ji and Eisenstein (2014) has implemented a parser using representation learning for the RST treebank which uses another discourse framework compared with PDTB. This one shows significant improvements compared to previous SOTA on the treebanks, but the results are in no way comparable with the PDTB. Still, it is the first paper (that I could find) that uses representation learning in a discourse parsing context, and it does so successfully which makes me want to at least look into their manually selected features later on. I am honestly somewhat surprised that none of the other systems looked into using this for the task, especially since the authors of the overview papers considered its results interesting enough to briefly mention it.

What I am most excited about is the idea of applying findings from architectures for the Stanford Natural Language Inference Corpus (Bowman et al., 2015). This is a classification task that I consider to be fairly similar to the sense classification task, and there are several interesting model implementations with strong results I want to further explore. Their website[1] especially mentions Mou et al. (2015) as a particularly efficient recursive neural network model with strong results. Cheng et al. (2016) is a LSTM network model, albeit not as efficient, but receives the highest score so far. Unfortunately, neither of these seem to have published their code which would mean some reimplementation efforts. I don't mind since I would like to learn TensorFlow anyway, and Google's Mat Kelcey has already some code I could work off of, albeit the performance is not as good[2].

That said, there are some things to keep in mind. First of all, neural network approaches are mainly known to be good when there are a lot of training data. That is not necessarily the case here. If the approaches above fail, I will look into if it is possible to use other, more sparse data friendly, architectures with the semantic representations. Also, Braud and Denis (2015) compared shallow features with some dense vector representations and found that shallow features still receive SOTA for implicit discourse relation identification. They do not seem to base any of their findings on word2vec vectors though, which is the kind of features that is allowed in the closed shared task. Also, their concatenation methods differ from the kind of architectures I have in mind. I will have to read this paper more carefully, and see what kind of dead ends I might be able to avoid.

Last minute update: I just found out about Zhang et al. (2015) from EMNLP 2015 which is a shallow convolutional neural network which reaches SOTA on implicit discourse recognition (which is not entirely the same task) on PDTB, making some of the previous worries about neural network approaches less worrisome. I am going to get a good understanding of their approach, read up on the NLI papers, and see how I can combine their efforts.

But first things first: I need to set up a baseline. Here are some suggestions:

- Try to break out the sense classifier from previous year's best performing paper (Wang and Lan, 2015). They use two separate classifiers for explicit and non-explicit senses, based on the original architecture in Lin et al. (2014), while adding a few extra features to boost the performance. The code is freely available, so it should be doable. Although it is unclear if this is actually the best model for sense classification.

- Do the same thing, but for the original architecture. (Lin et al., 2014). The code base is also available.

- Try to figure out which of the systems perform best on

- Implement some other simple system, e.g. an SVM bag-of-words model, potentially with brown clusters due to them

I'm leaning towards the having both the first suggestion, using Wang and Lan (2015), as well as the SVM. The first suggestion due to me wanting to potentially

[1] `http://nlp.stanford.edu/projects/snli/`
[2] `https://github.com/matpalm/snli_nn_tf`

incorporate my classifier into their system, and SVM for its usage as baseline in Zhang et al. (2015) that I just found.

2.3.1 Questions!

- Is this a good plan?
- Is using Wang and Lan (2015) as a baseline a good idea?
- Am I fooling myself by believing that I can learn anything from studying NLI architectures?
- The sense classes are rather imbalanced (see Table 3). Do you know of any general methods on how to deal with this?
- What am I missing?

3 Data

For training we will be using the Penn Discourse Treebank (PDTB). PDTB follows the lexically grounded predicate-argument approach as proposed in Webber (2004). It covers the subset containing Wall Street Journal articles from the Penn Treebank, making up approximately one million tokens. When a connective explicitly appears, it will be syntactically connected to the *Arg2* argument of the discourse structure. *Arg1* is the other one. Due to *Arg2* being syntactically bounded to connective, it is easy to automatically classify *Arg2*.

PDTB annotates each structure with types of discourse relations according to a three level hierarchy as seen in Table 2.2, where the first level is made up of four classes: TEMPORAL, CONTINGENCY, COMPARISON, EXPANSION. Each class has a non-obligatory second level of in total 16 types to provide a more fine-grained classification. Due to the third level being considered too fine-grained, it is ignored in this work.

Furthermore, we have some additional resources at our disposal:

- Brown clusters from the RCV1 corpus.
- MPQA subjectivity lexicon.
- Skip-gram Neural Word Embeddings trained on 100 billion words from the Google News dataset.
- VerbNet 3.2.

senses	connective_token	ratio
(Expansion.Conjunction,)	7355	0.226064
(Contingency.Cause,)	4969	0.152728
(Comparison.Contrast,)	4521	0.138958
(EntRel,)	4133	0.127032
(Expansion.Restatement,)	2640	0.081143
(Temporal.Asynchronous,)	2044	0.062825
(Expansion.Instantiation,)	1392	0.042785
(Comparison.Concession,)	1268	0.038973
(Contingency.Condition,)	1114	0.034240
(Temporal.Synchrony,)	985	0.030275
(Comparison,)	484	0.014876
(Expansion.Alternative,)	435	0.013370
(Temporal.Asynchronous, Contingency.Cause)	148	0.004549
(Temporal.Synchrony, Contingency.Cause)	113	0.003473
(Temporal.Synchrony, Expansion.Conjunction)	109	0.003350
(Expansion.Conjunction, Contingency.Cause)	106	0.003258
(Expansion,)	96	0.002951
(Expansion.Conjunction, Temporal.Synchrony)	92	0.002828
(Temporal.Synchrony, Comparison.Contrast)	51	0.001568
(Expansion.Conjunction, Comparison.Contrast)	47	0.001445

Table 3.1: Frequency of sense labels in training data. Right now connective_token is simply the frequency. I should change this name. Ratio is the frequency ratio.

Bibliography

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*, 2015.

Chloé Braud and Pascal Denis. Comparing word representations for implicit discourse relation classification. In *Empirical Methods in Natural Language Processing (EMNLP 2015)*, 2015.

Jianpeng Cheng, Li Dong, and Mirella Lapata. Long Short-Term Memory-Networks for Machine Reading. *arXiv preprint arXiv:1601.06733*, 2016.

Yangfeng Ji and Jacob Eisenstein. Representation Learning for Text-level Discourse Parsing. In *ACL (1)*, pages 13–24, 2014.

Ziheng Lin, Hwee Tou Ng, and Min-Yen Kan. A PDTB-styled end-to-end discourse parser. *Natural Language Engineering*, 20(02):151–184, 2014.

Lili Mou, Men Rui, Ge Li, Yan Xu, Lu Zhang, Rui Yan, and Zhi Jin. Recognizing Entailment and Contradiction by Tree-based Convolution. *arXiv preprint arXiv:1512.08422*, 2015.

Tsuyoshi Okita, Longyue Wang, and Qun Liu. The DCU Discourse Parser: A Sense Classification Task. *CoNLL 2015*, page 71, 2015.

Jianxiang Wang and Man Lan. A refined end-to-end discourse parser. *CoNLL 2015*, page 17, 2015.

Longyue Wang, Chris Hokamp, Tsuyoshi Okita, Xiaojun Zhang, and Qun Liu. The dcu discourse parser for connective, argument identification and explicit sense classification. *CoNLL 2015*, page 89, 2015.

Biao Zhang, Jinsong Su, Deyi Xiong, Yaojie Lu, Hong Duan, and Junfeng Yao. Shallow Convolutional Neural Network for Implicit Discourse Relation Recognition. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2230–2235, 2015.