Disambiguating explicit discourse connectives without oracles

Anders Johannsen Anders Søgaard
University of Copenhagen
{ajohannsen,soegaard}@hum.ku.dk

Abstract

Deciding whether a word serves a discourse function in context is a prerequisite for discourse processing, and the performance of this subtask bounds performance on subsequent tasks. Pitler and Nenkova (2009) report 96.29% accuracy (F_1 94.19%) relying on features extracted from gold-standard parse trees. This figure is an average over several connectives, some of which are extremely hard to classify. More importantly, performance drops considerably in the absence of an oracle providing gold-standard features. We show that a very simple model using only lexical and predicted part-of-speech features actually performs slightly better than Pitler and Nenkova (2009) and not significantly different from a state-of-the-art model, which combines lexical, part-of-speech, and parse features.

1 Introduction

Discourse relations structure text by linking segments together in functional relationships. For instance, someone might say "Saber-toothed tigers are harmless *because* they're extinct", making the second part of the sentence serve as an explanation for the first part. In the example the discourse connective *because* functions as a lexical anchor for the discourse relation. Whenever an anchor is present we say that the discourse connective is *explicit*.

Complicating the matter, phrases used as discourse connectives sometimes appear in a non-discourse function. For instance, "and" may be either a simple conjunction, as in "sugar and salt", or a discourse relation suggesting a temporal relationship between events, for instance "he struck the match and went away". The Penn Discourse Treebank (PDTB) (Prasad et al., 2008) distinguish 100 types of explicit connectives—a subset of these are listed in Table 2. The type of relationship is selected from a hierachial structure where the four top-level categories are Comparison, Contingency, Temporal, and Expansion.

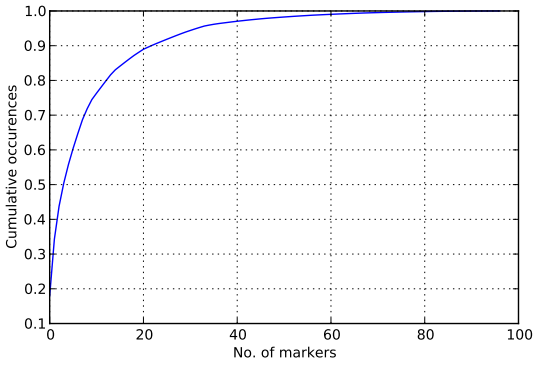

Figure 1: A picture of the problem. 10% of connectives account for roughly 75% of occurrences

Discourse relations are important for many applications and, since the PDTB was released, much effort has gone into developing tools for recreating the annotations of the resource automatically. Recently two ambitious end-to-end parsers have appeared which transform plain text to full PDTB-style annotations (Lin et al., 2010; Ghosh, 2012). Both systems share a pipelined architecture in which the output of one component becomes the input to the next. A crucial first step in their processing is correctly identifying explicit discourse connectives; when unsuccessful subsequent steps fail.

An accuracy in the high ninetees seems to suggest that the problem is almost solved. For the task of discourse connective disambiguation this unfortunately does not hold true, because, as we argue here, the task benefits from being seen and evaluated as a number of smaller tasks, one for each

International Joint Conference on Natural Language Processing, pages 997–1001,
Nagoya, Japan, 14-18 October 2013.

connective type. Figure 1 shows why: the distribution of connectives follows a power law such that the majority of occurrences comes from relatively few but highly frequent connective types. If we do not take into account the uneven sizes of the categories, our performance figure ends up saying very little about how well we are doing on most of the connectives, because it is being dominated by the performance on a few high-frequency items.

In this paper we look in more detail on the evaluation of the discourse connective disambiguation task, in particular how two commonly used feature models perform on individual discourse connectives. The models are Pitler and Nenkova (2009) (P&N), and its extension by Lin et al. (2010) (Lin). Motivated by our findings we advocate the use of macro-averaging as a necessary supplement to micro-averaging. Additionally, we perform our experiments in a more realistic setting where access to oracle gold-standard annotations is not assumed. The observed performance drop from oracle to predicted parses leads us to propose a new model, which approximates the syntactical information of the parse trees with part-of-speech tags. Although these features are less powerful in theory, the model has comparable macro-average performance in realistic evaluation.

The rest of the paper is structured as follows. In the next section we give reasons why low-frequency connectives should not be overlooked. Section 3 describes our experiments, and Section 4 reports on the results. The discussion is in Section 5, followed by a review of related work in Section 6. Section 7 concludes the paper.

2 The importance of the long tail

Are there any compelling reasons to pay attention to the lower-frequency connectives when high-frequency connectives overwhelmingly dominate? As noted in the caption to Figure 1, the top 10 account for above 75% of the occurrences and top 20 for above 90%. So why should we care?

It turns out that the low-frequency connectives are quite evenly distributed among texts. In the Wall Street Journal part of the Penn Treebank, 70% of articles that contain explicit markers contain at least one marker not in the top 10. Not counting very short texts (having only two or fewer explicit connectives of any type), the number rises to 87%. While low performance on less frequent connectives does not hurt a token-level macro-average much, it still means that you are likely to introduce errors in something like 70% of all WSJ articles. These errors percolate leading to erroneous text-level discourse processing.

In Webber and Joshi (2012) the prime example of a discourse application is automatic text simplification. Here, ignoring the long tail of discourse connectives would be out of the question, because it is precisely those less familiar expressions — which people encounter rarely and have weaker intuitions about — that would benefit the most from a rewrite. Two other examples, also cited in Webber and Joshi (2012), are automatic assessment of student essays (Burstein and Chodorow, 2010), and summarization (Thione et al., 2004). In student essays we encourage clear argumentative structure and rich vocabulary; failing to recognize that in an automatic system would not qualify as fair evaluation. And summarization is often performed over news wire, which, as shown in the PDTB, has a high per-article incidence of connectives not in top 10. Additionally, some low-frequency connectives like ”ultimately” and ”in particular” are strong cues for text selection.

Another reason to suspect that low-frequency connectives are important comes from an observation about the distribution of connectives in biomedical text. Ramesh and Yu (2010) report an overlap of only 44% between the connectives found in the The Biomedical Discourse Relation Bank (Prasad et al., 2011), a 24 article subset of the GENIA corpus (Kim et al., 2003), and the PDTB. The intersection contains high-frequency connectives, such as ”and”, ”however,” ”also,” and ”so”. Connectives specific to the biomedical domain include ”followed by,” ”due to,” and ”in order to”, and the authors speculate that the unique connectives encode important domain specific knowledge.

3 Experiments

Our experiments are designed to shed light on three aspects of discourse connective disambiguation: 1) error distribution wrt. connective type; uneven performance builds a strong case for averaging over connective types instead of averaging over data points; 2) performance loss in the absence of an oracle; and 3) performance of simple model based on cheaper and more reliable annotations.

We experiment with three different feature sets,

all of which model syntactical aspects of the discourse connective.

The P&N and Lin feature sets are chosen to represent state-of-the-art. The high accuracy of P&N at 96.29% is frequently cited as an encouraging result, see Huang and Chen (2011; Alsaif and Markert (2011; Tonelli and Cabrio (2012; Zhou et al. (2010). Besides discourse parsing P&N has been used for tasks as diverse as measuring text coherence (Lin et al., 2011) and improving machine translation (Meyer and Popescu-Belis, 2012). The POS+LEX feature set is proposed as an alternative model. The baseline always predicts the majority class.

P&N This feature set derives from parse trees and replicates the features of Pitler and Nenkova (2009). Starting from the potential discourse connective, the features include the highest category in the tree subsuming only the connective called the self-category, the parent of that category, the left sibling of the self-category, and the right sibling of the self-category. A feature fires when the right sibling contains a VP, and another if there is a trace node below the right sibling. Note that the trace feature will never fire outside of the gold parse setting since state-of-the-art parsers do not predict trace nodes.

Importantly, there is a feature for the identity of the connective and interaction features between the connective and the syntactical features in effect allowing the model to fit parameters specific to each connective. Furthermore, combinations of the syntactical features are allowed, but they cannot be connective-specific.

Lin The feature set augments P&N with part-of-speech and string features for the tokens adjacent to the connective, as well as the part-of-speech of the connective itself. The part-of-speech features for the adjacent tokens interact with the part-of-speech of the connective, and the string features interact with the indicator feature for the connective. It also adds a syntax feature: the path to the root of the parse tree.

POS+LEX The simple feature set builds on part-of-speech tags and tokens. Part-of-speech tags are captured using a window of two tokens around the marker, and the lexical features are the same as Lin. Like P&N there is a feature for the identity of the connective as well as interaction features between the identity feature and other features.

Model	Micro		Macro	
	Oracle	Pred.	Oracle	Pred.
Baseline	72.7	72.7	53.9	53.9
P&N	93.0	90.7	85.3	80.7
Lin	95.2	92.9	86.7	83.6
POS+LEX	89.7	89.7	82.5	83.5

Table 1: Comparing F_1 score on oracle and predicted features using macro and micro averaging. A Wilcoxon signed rank test shows that the macro-averaged difference between POS+LEX and Lin10 using predicted features is not signficant at $p < 0.01$.

In keeping with Pitler and Nenkova (2009) our learner is a maximum entropy classifier trained on sections 2-22 of the WSJ using ten-fold cross-validation.

3.1 Parsing Wall Street Journal

To obtain a version of the WSJ corpus containing fully predicted parses we use the Stanford Parser[1] training a separate model for each section. To parse a specific section we train on everything but that section (e.g. for parsing section 5 the training set is section 0-4 and 6-24). Average F_1 on all sections is 85.87%. Although the very best state-of-the-art parsers[2] report F_1 of above 90%, our parsing score greatly exceeds typical performance on real-life data, which is almost always out-of-domain with respect to 1980s WSJ. Thus this setting still compares favourably to performance in the wild.

4 Results

A summary of the results is found in Table 1. For a subset of frequent and less frequent connectives, Table 2 lists individual F_1 scores. In all of the feature sets we see a marked drop moving from micro-average (average over instances) to macro-average (average over connective types)—P&N, for instance, goes from 93.0% to 85.3%. This shows that the scores of less frequent connectives are somewhat lower than frequent ones. When

[1] `http://nlp.stanford.edu/software/lex-parser.shtml`, 2012-11-12 release with the 'goodPCFG' standard settings

[2] `http://aclweb.org/aclwiki/index.php\?title=Parsing_(State_of_the_art)`

	Oracle		Pred.		Disc.
	Lin	P+L	Lin	P+L	
but	98.6	96.1	97.6	96.1	78.9
and	94.9	77.0	89.0	77.0	14.7
also	97.0	97.3	97.5	97.2	93.5
if	93.4	93.1	92.3	93.0	82.6
when	89.9	88.5	89.3	88.4	65.5
because	99.5	99.4	99.4	99.5	63.4
while	97.6	97.7	97.5	97.4	91.9
as	89.8	63.1	78.1	63.0	13.0
after	93.7	74.0	87.9	72.9	42.4
however	98.7	98.4	98.5	98.4	95.7
...					
ultimately	43.2	30.3	36.4	29.4	37.5
rather	84.8	83.9	80.0	83.9	8.2
in other words	97.1	94.4	91.4	94.4	89.5
as if	84.8	84.8	71.0	88.2	66.7
earlier	76.9	66.7	74.1	69.6	2.1
meantime	80.0	76.5	82.4	80.0	71.4
in particular	89.7	85.7	85.7	80.0	48.4
in contrast	100.0	100.0	100.0	100.0	50.0
thereby	95.7	95.7	100.0	95.7	100.0
...					

Table 2: F_1 score per connective. The table is sorted by the number of actual discourse connectives in the PDTB. After the break the table continues from position 50. The last column gives the percentage of discourse connectives.

features are derived from predicted parses performance also fall, from 93.0% to 90.7% with micro-average, and even more dramatically with macro-average, where it goes from 85.3% to 80.8%. Given that we are interested in real life performance this last figure is the most interesting.

5 Discussion

In NLP applications we cannot assume the existence of oracles providing us with gold-standard features. Often switching to predicted features introduces greater uncertainty. If the parser often confuses two non-terminals that are important for connective disambiguation we loose predictive power. Thus, on the P&N model, the average conditional entropy per feature given the class (how surprising the feature is when we know the answer) increases by 8.8% when the oracle is unavailable. In contrast there is almost no difference between the conditional entropy of the POS model with oracle features and without, indicating that the errors made by the tagger are not confusing in the disambiguation task.

Predicted parse features are associated with uncertainty even when used in combination with words and part of speech. Comparing the number of times the Lin model changes an incorrect prediction of POS+LEX to a correct one and the number of times it introduces a new error by changing a correct prediction to an incorrect one, we observe that corrections almost always come with a substantial number of new errors. In fact, 58 connectives have at least as many new errors as corrections.

Predicted parse features also contribute to feature sparsity, because of the greater variability of automatic parses. On the other hand, they are more expressive than part of speech, and in the example below, where only Lin correctly identifies 'and' as a discourse connective, part of speech simply does not contain enough information.

"A whole day goes by **and** no one even knows they're alive.

6 Related work

Atterer and Schütze (2007) present similar experiments for prepositional phrase attachment showing that approaches assuming gold-standard features suffer a great deal when they are evaluated on predicted features. Spitkovsky et al. (2011) also caution against the use of gold-standard features, arguing that for unsupervised dependency parsing using induced parts of speech is superior to relying on gold-standard part-of-speech tags.

This work also relates to Manning (2011) who point out that even though part-of-speech tagging accuracy is above 97% the remaining errors are not randomly distributed but in fact occur in just the cases we care most about.

7 Conclusion

Discourse connective disambiguation is an important subtask of discourse parsing. We show that when realistic evaluation is adopted — averaging over connective types and not relying on oracle features — performance drops markedly. This suggests that more work on the task is needed. Moreover, we show that in realistic evaluation a simple feature model using part-of-speech tags and words performs just as well as a much more complex state-of-the-art model.

Acknowledgements

We wish to thank the ESICT project for partly funding this work. The ESICT project is supported by the Danish Council for Strategic Research.

References

Amal Alsaif and Katja Markert. 2011. Modelling discourse relations for Arabic. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 736–747, Stroudsburg, PA, USA. Association for Computational Linguistics.

Michaela Atterer and Hinrich Schütze. 2007. Prepositional phrase attachment without oracles. *Computational Linguistics*, 33(4):469–476.

Jill Burstein and Martin Chodorow. 2010. Progress and New Directions in Technology for Automated Essay Evaluation. In R. Kaplan, editor, *The Oxford Handbook of Applied Linguistics, 2nd Edition*, pages 487–497. Oxford University Press.

Sucheta Ghosh. 2012. *End-to-End Discourse Parse using Cascaded Structured Prediction*. Phd thesis, University of Trento.

Hen-Hsen Huang and Hsin-Hsi Chen. 2011. Chinese discourse relation recognition. In *Proceedings of 5th International Joint Conference on Natural Language Processing (IJCNLP 2011)*, pages 1442–1446.

J.-D. Kim, T. Ohta, Y. Tateisi, and J. Tsujii. 2003. GENIA corpus–a semantically annotated corpus for biotextmining. *Bioinformatics*, 19(Suppl 1):i180–i182, July.

Ziheng Lin, Hwee Tou Ng, and Min-yen Kan. 2010. A PDTB-Styled End-to-End Discourse Parser. Technical Report 2004, School of Computing, National University of Singapore, Singapore.

Ziheng Lin, Hwee Tou Ng, and Min-Yen Kan. 2011. Automatically evaluating text coherence using discourse relations. pages 997–1006, June.

Christopher D. Manning. 2011. Part-of-Speech Tagging from 97% to 100%: Is It Time for Some Linguistics? In AlexanderF. Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing SE - 14*, volume 6608 of *Lecture Notes in Computer Science*, pages 171–189. Springer Berlin Heidelberg.

Thomas Meyer and Andrei Popescu-Belis. 2012. Using Sense-labeled Discourse Connectives for Statistical Machine Translation. In *Proceedings of the Joint Workshop on Exploiting Synergies Between Information Retrieval and Machine Translation (ESIRMT) and Hybrid Approaches to Machine Translation (HyTra)*, EACL 2012, pages 129–138, Stroudsburg, PA, USA. Association for Computational Linguistics.

Emily Pitler and Ani Nenkova. 2009. Using syntax to disambiguate explicit discourse connectives in text. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, ACLShort '09, pages 13–16, Stroudsburg, PA, USA. Association for Computational Linguistics.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn Discourse TreeBank 2.0. *In Proceedings of LREC*.

Rashmi Prasad, Susan McRoy, Nadya Frid, Aravind Joshi, and Hong Yu. 2011. The biomedical discourse relation bank. *BMC Bioinformatics*, 12(1):188.

Balaji Polepalli Ramesh and Hong Yu. 2010. Identifying discourse connectives in biomedical text. In *AMIA Annual Symposium Proceedings*, volume 2010, page 657. American Medical Informatics Association.

Valentin I Spitkovsky, Hiyan Alshawi, Angel X Chang, and Daniel Jurafsky. 2011. Unsupervised dependency parsing without gold part-of-speech tags. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1281–1290. Association for Computational Linguistics.

Gian Lorenzo Thione, Martin Van Den Berg, Livia Polanyi, and Chris Culy. 2004. Hybrid text summarization: Combining external relevance measures with structural analysis. In *Proceedings ACL Workshop Text Summarization Branches Out. Barcelona*.

Sara Tonelli and Elena Cabrio. 2012. Hunting for Entailing Pairs in the Penn Discourse Treebank. In *Proceedings of COLING 2012*, pages 2653–2668, Mumbai, India, December. The COLING 2012 Organizing Committee.

Bonnie Webber and Aravind Joshi. 2012. Discourse Structure and Computation: Past, Present and Future. In *Association for Computational Linguistics*, page 42.

Zhi Min Zhou, Man Lan, Zheng Yu Niu, Yu Xu, and Jian Su. 2010. The effects of discourse connectives prediction on implicit discourse relation recognition. In *Proceedings of the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, SIGDIAL '10, pages 139–146, Stroudsburg, PA, USA. Association for Computational Linguistics.