Identifying Explicit Discourse Connectives in German

Peter Bourgonje and **Manfred Stede**
Applied Computational Linguistics
University of Potsdam / Germany
firstname.lastname@uni-potsdam.de

Abstract

We are working on an end-to-end Shallow Discourse Parsing system for German and in this paper focus on the first subtask: the identification of explicit connectives. Starting with the feature set from an English system and a Random Forest classifier, we evaluate our approach on a (relatively small) German annotated corpus, the Potsdam Commentary Corpus. We introduce new features and experiment with including additional training data obtained through annotation projection and achieve an f-score of 83.89.

1 Introduction

A task central to the field of Discourse Processing is the uncovering of coherence relations that hold between individual (elementary) units of a text. When discourse relations are explicitly signaled in a text, the explicit markers are called *(discourse) connectives*. Connectives can be two-way ambiguous in the sense of having either a discourse or a sentential reading, and if they have a discourse reading, many can assign multiple senses. Further, connectives form a syntactically heterogeneous group and include coordinating and subordinating conjunctions, adverbials, and depending on the definition maintained, also certain prepositions. In our experiments, we adopt the definition of Pasch et al. (2003, p.331) where *X* is a connective if *X* cannot be inflected, the meaning of *X* is a two-place relation, the arguments of *X* are propositional structures and the expressions of the arguments of *X* can be sentential structures. Following Stede (2002), we include prepositions that have a discourse function.

Recent approaches toward end-to-end shallow discourse parsing (SDP) have focused on a pipeline approach where the identification of discourse connectives is the first step, followed by the extraction of the arguments of the connective and the classification of the sense. This pipeline architecture has dominated the CONLL 2015[1] and 2016[2] shared tasks on SDP. We will adopt it for our goal, viz. developing an end-to-end discourse parser for German. This paper focuses on the first step in the pipeline and introduces a connective identification module for German. We train a classifier using annotated data (Section 3), investigate and extend the feature set (Section 4), discuss and evaluate the results (Section 5) and summarize in Section 6.

2 Related Work

Early attempts at formalizing discourse parsing procedures for English are described in (Soricut and Marcu, 2003), among others. Pitler and Nenkova (2009) experiment with syntactically motivated features for the binary classification of discourse connectives (connective or non-connective reading) and report an f-score of 94.19 for the PDTB data (Prasad et al., 2008). The SDP pipeline architecture is adopted from Lin et al. (2014) and is also used in the best-scoring systems of the 2015 and 2016 CONLL shared tasks, (Wang and Lan, 2015) and (Oepen et al., 2016) respectively. Oepen et al. (2016) achieve an overall f-score of 27.77 for full SDP, but 91.79 for identifying explicit connectives. The best-scoring system for this subtask (Li et al., 2016) achieved an impressive 98.38.

A notable drawback of the pipeline architecture is the possibility of error propagation. This is addressed by (Biran and McKeown, 2015), who use

[1] http://www.cs.brandeis.edu/~clp/conll15st/
[2] http://www.cs.brandeis.edu/~clp/conll16st/

Proceedings of the SIGDIAL 2018 Conference, pages 327–331,
Melbourne, Australia, 12-14 July 2018.

a tagging-based approach and divide the task into processing intra-sentential and inter-sentential relations (as opposed to the more typical division into explicit and implicit relations) and report a final f-score of 39.33. This is based on a more lenient scoring system though, and Oepen et al. (2016) achieve 44.20 using a similar partial matching scoring system.

The main resources available for German are DiMLex, a lexicon of German discourse connectives containing 275 entries (Stede, 2002), (Scheffler and Stede, 2016) and the Potsdam Commentary Corpus (PCC) (Stede and Neumann, 2014), described in more detail in Section 3. We experiment with generating extra training data through annotation projection. This approach is inspired by Versley (2010), who attempts to disambiguate German connectives using a parallel English-German corpus. Earlier work on connective identification for German is done by (Dipper and Stede, 2006), who train the Brill Tagger using a modified tag set and consider only 9 of the 42 ambiguous entries in DiMLex, reporting an f-score of 90.20. In our present study, we deal with the full set of connectives for which we have training data.

3 Data

To the best of our knowledge, the only German corpora containing discourse annotations are the PCC[3] and a subsection of the TüBa-D/Z corpus (Versley and Gastel, 2012), complemented by a lexicon of discourse connectives; DiMLex[4]. We use the PCC, which is a corpus of 176 texts taken from the editorials page of a local German newspaper and is annotated on several layers: discourse connectives and their arguments and sense, syntax trees, Rhetorical Structure Theory trees and coreference chains.

The PCC contains in total 33,222 words and 1,176 connective instances. Because the texts were not sampled to extract targeted examples (of particular connectives or senses), they do not contain the full set of connective entries from DiMLex, but 156 unique connectives, compared to in total 275 entries in DiMLex. From this corpus we extracted 3,406 data instances (1,176 connective instances, plus 2,230 candidates with a non-connective reading). Of 156 unique connectives, 74 are unambiguous and always have discourse reading (at least in the PCC). But these 74 connectives represent only 279 instances (8% of the total data). Of the remaining 82 connectives, the distribution is heavily skewed and covers the full spectrum of possibilities; while connectives like 'Und'[5] ('and'), 'sondern' ('but/rather') and 'wenn' ('if') have a high connective ratio of 0.95, 0.93 and 0.97 respectively; 'als' ('as'), 'Wie' ('(such) as') and 'durch' ('by/through') very seldom have the connective reading (a ratio of 0.08, 0.05, and 0.06, respectively).

In comparison, the training section of the 2016 CONLL shared task data alone contains ca. 933k words and ca. 278k training instances, so we cannot expect to get results nearly as good as those that were obtained for English. In an attempt to generate additional training data, we thus experimented with annotation projection, inspired by Versley (2010). We implemented an English connective classifier using the feature set of Lin et al. (2014), classified the English part of a parallel corpus, located the German counterparts through word alignment, and used the sentences obtained as additional training data. The parallel corpus is EuroParl (Koehn, 2005) and the word alignments were obtained using MGIZA (Gao and Vogel, 2008). Filtering out input sentences of more than 100 words (due to high syntactic parsing costs for subsequent steps) and alignments to German words not present in DiMLex, this resulted in 18,853 extra data instances.

4 Method

We started with the feature set of Lin et al. (2014) (in turn based on (Pitler and Nenkova, 2009)), which is a combination of surface (token and bigram), part-of-speech and syntactic features (like path to the root node, category of the siblings, etc.). The parse trees are obtained from the NLTK implementation of the Stanford Parser for German (Rafferty and Manning, 2008). We use a Random Forest classifier (Pedregosa et al., 2011) for all experiments. All scores are the result of 10-fold cross-validation using 90% of the PCC as training data and the remaining 10% as test data (except for the setup using the additional EuroParl data; this data is added to the training data for each of

[3] http://angcl.ling.uni-potsdam.de/resources/pcc.html

[4] https://github.com/discourse-lab/dimlex

[5] Note that we make a distinction between 'Und' (uppercase U) and 'und' here.

the 10 folds). As a result of error analysis on the output when using the base feature set, we added some extra features. Because we include prepositions in our set of connectives (which additionally includes conjunctions and adverbials), we included a feature indicating the syntactic group of the connective to explicitly differentiate for five cases; the four categories above[6] plus *other* for the remaining cases (like 'um...zu' (discontinuous 'in order...to')). The value for this feature is just a more general label than connective's part-of-speech category, included to avoid sparsity. While being sentence-initial is in most cases reflected by the bigram features, we included an explicit feature that indicates whether or not the candidate is initial to a clause that starts with *S* (*S* or *S-bar*). These two features, which are directly derived from other features already present in the set, would likely not improve performance much if more training data is available, but as our experiments show, they do improve the f-score by another 2 points in our scenario in which training data is limited. Another feature that improved performance was sentence length; intuitively it makes sense that as sentences get longer, the need for explicit structuring of the propositions therein increases. Together, these added features improved the f-score (see Table 1).

5 Results & Evaluation

The results for the different setups are illustrated in Table 1. We use a micro-averaged f1 score for all experiments.

We compare performance of the classifier to a majority vote baseline, where each instance is assigned its most frequent label. Using the base feature set results in an f-score of 81.90 (second row of Table 1). Using extra training data generated through annotation projection on EuroParl yields a negative result (below the baseline) and f-score decreases considerably, to 65.98 (third row). This decrease can be explained by the susceptibility of this approach to error propagation. The English classifier, trained on the PDTB (f-score of 93.64) is applied to another domain (EuroParl), word-alignments introduce errors, and the additional German training data is again from another domain (EuroParl) than the test set (news commentary). The extra training data obtained in this way (18,853 instances) apparently does not compensate for this. We note that the scores resulting from annotation projection data are comparable to the f-score of 68.7 reported by (Versley, 2010). This may suggest an upper-limit in performance when using data obtained through annotation projection, but more research is needed to verify this.

Since the PCC has gold annotations for syntax trees, we used these for part-of-speech tag and other syntactic features, in order to establish the impact of parsing errors. As shown in the first row, this mainly impacts precision and leads to an increase of almost 5 points for the f-score (using the base feature set). However, because having access to gold parses is not feasible in an end-to-end scenario, we consider this an estimation of the impact of parsing errors and continue using automatically generated parse trees for the other experiments.

The best results were obtained using the extended feature set (see Section 4) and are displayed in the last row of Table 1.

Inspecting the individual scores, we found that in particular 'auch' ('also') and 'als' ('as/than') were difficult to classify (with f-scores of 27.03 and 28.57, respectively), despite being relatively frequent (208 and 147 examples in the PCC). Although they are not connectives in the majority of cases (ratios of 0.13 ('auch') and 0.08 ('als')), some connectives with similar ratios yet significantly lower frequencies have higher f-scores, such as 'so' ('so/thus'); frequency of 108, ratio of 0.11 and f-score of 72.00) and 'damit' ('in order to/thereby'); frequency of 30, ratio of 0.19 and f-score of 60.00. When using separate classifiers for the different syntactic categories (a setup which did not result in improved performance), the conjunctions performed best (with 91.81 for co-ordinating and 90.25 for subordinating conjunctions) and prepositions worst (51.55), but group-internally the differences were equally large, with some prepositions having above-average scores and some having scores close to 0. Further attempts at increasing the overall f-score quickly led to looking into solutions for individual connectives and came with the risk of over-fitting to the data set.

To put our score for German into perspective, we performed a set of experiments with different amounts of training data for English. Figure 1 shows the f-score (y-axis) when gradually increasing the number of training instances (x-axis). The

[6] prepositions, co-ordinating conjunctions, sub-ordinating conjunctions and adverbials

	precision	recall	f-score
majority vote baseline	73.76	87.32	79.60
base features + gold trees	*86.44*	*85.13*	*85.76*
base features + auto-generated trees	78.88	85.16	81.90
base features + EuroParl training data	74.23	59.54	65.96
extended features + auto-generated trees	**82.16**	**85.69**	**83.89**

Table 1: Results for binary connective classification on PCC for gold trees and automatically generated trees

blue line represents the curve for English, starting with 1,000 instances randomly sampled from the total of 278k instances in the 2016 CONLL shared task data. Recall that using this full set, the f-score using the same feature set and classification algorithm (RandomForest) is 93.64. The orange triangle represents performance for German, using all available instances from the PCC. While we have no explanation for the dent in the curve at 10,000 instances (and the smaller one around 20,000), we focus on the German score and note that with 81.90, this is 1.8 points higher than the corresponding score for English (80.09). This comparison suggests that the problem of connective identification is not significantly more or less challenging for German than it is for English. In fact, seeing that we also include the syntactic category of prepositions (which is not included in the PDTB connectives), and this group scored the worst in our separate-classifier setup, it suggests that for the other categories, performance is better for German than it is for English. When leaving out prepositions altogether, f-score increased to 85.99. But because it was a conscious decision to include prepositions, the most straightforward means of improving performance for the problem at hand seems to be adding more (in-domain) training data.

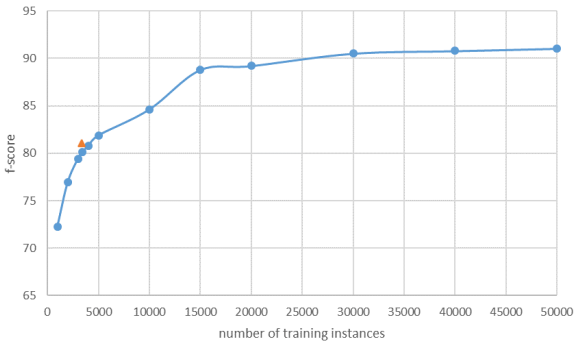

Figure 1: f-scores for varying training data volumes for English (blue line) and f-score for PCC as training data for German (orange triangle)

6 Conclusion & Outlook

We implement the first part of a pipeline for end-to-end discourse parsing for German; the identification of discourse connectives. We use a Random Forest classifier and add additional syntactic features to the base set, which is taken from a state-of-the-art system for English. Evaluating this approach on the Potsdam Commentary Corpus, we arrive at an f-score of 83.89, improving by over 4 points compared to a majority vote baseline. Generating additional training data through annotation projection on a parallel corpus does not improve performance. Our approach is best compared to Dipper and Stede (2006), who achieve a higher f-score (90.20) but only consider 9 connectives whereas we consider the full set present in the annotated data. Versley (2010) also does not limit the set of connectives but uses an annotation projection approach resulting in an f-score of 68.7.

We show that performance for German is on par with (in fact, slightly better than) English when using the same amount of training data, the same feature set and the same classifier. This may suggest that the task is not necessarily more challenging or complicated for German than it is for English, though it remains unclear what role domain plays here (news commentary in the German case vs. news in the English case). We plan to annotate more training data in the same domain, but also out-of-domain to establish domain influence. We will continue to work on the follow-up components in the pipeline (argument extraction and sense classification), but will simultaneously attempt to improve performance for this first step in the pipeline, due to the sensitivity of the architecture to error propagation.

Acknowledgments

We are grateful to the Deutsche Forschungsgemeinschaft (DFG) for funding this work in the project 'Anaphoricity in Connectives'. We would like to thank the anonymous reviewers for their helpful comments on an earlier version of this manuscript.

References

Or Biran and Kathleen McKeown. 2015. PDTB Discourse Parsing as a Tagging Task: The Two Taggers Approach. In *SIGDIAL Conference*. The Association for Computer Linguistics, pages 96–104.

Stefanie Dipper and Manfred Stede. 2006. Disambiguating potential connectives. In *Proceedings of the KONVENS Conference*. Konstanz.

Qin Gao and Stephan Vogel. 2008. Parallel implementations of word alignment tool. In *Software Engineering, Testing, and Quality Assurance for Natural Language Processing*. Association for Computational Linguistics, SETQA-NLP '08, pages 49–57.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Conference Proceedings: the tenth Machine Translation Summit*. AAMT, Phuket, Thailand, pages 79–86.

Zhongyi Li, Hai Zhao, Chenxi Pang, Lili Wang, and Huan Wang. 2016. A Constituent Syntactic Parse Tree Based Discourse Parser. In *Proceedings of the CONLL 2016 Shared Task*. Berlin, pages 60–64.

Ziheng Lin, Hwee Tou Ng, and Min-Yen Kan. 2014. A PDTB-Styled End-to-End Discourse Parser. *Natural Language Engineering* 20:151–184.

Stephan Oepen, Jonathon Read, Tatjana Scheffler, Uladzimir Sidarenka, Manfred Stede, Erik Velldal, and Lilja Øvrelid. 2016. OPT: OsloPotsdamTeesside—Pipelining Rules, Rankers, and Classifier Ensembles for Shallow Discourse Parsing. In *Proceedings of the CONLL 2016 Shared Task*. Berlin.

Renate Pasch, Ursula Brauße, Eva Breindl, and Ulrich Herrmann Waßner. 2003. *Handbuch der deutschen Konnektoren*. Walter de Gruyter, Berlin/New York.

Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Edouard Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12:2825–2830.

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*. Association for Computational Linguistics, ACLShort '09, pages 13–16.

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

Anna N. Rafferty and Christopher D. Manning. 2008. Parsing Three German Treebanks: Lexicalized and Unlexicalized Baselines. In *Proceedings of the Workshop on Parsing German*. Association for Computational Linguistics, PaGe '08, pages 40–46.

Tatjana Scheffler and Manfred Stede. 2016. Adding semantic relations to a large-coverage connective lexicon of German. In Nicoletta Calzolari et al., editor, *Proc. of the Ninth International Conference on Language Resources and Evaluation (LREC 2016)*. Portoro, Slovenia.

Radu Soricut and Daniel Marcu. 2003. Sentence level discourse parsing using syntactic and lexical information. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*.

Manfred Stede. 2002. DiMLex: A lexical approach to discourse markers. In *Exploring the Lexicon - Theory and Computation*, Edizioni dell'Orso, Alessandria.

Manfred Stede and Arne Neumann. 2014. Potsdam Commentary Corpus 2.0: Annotation for discourse research. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. European Language Resources Association (ELRA), Reykjavik, Iceland.

Yannick Versley. 2010. Discovery of Ambiguous and Unambiguous Discourse Connectives via Annotation Projection. In *Proceedings of Workshop on Annotation and Exploitation of Parallel Corpora (AEPC)*. Northern European Association for Language Technology (NEALT).

Yannick Versley and Anna Gastel. 2012. Linguistic tests for discourse relations in the TüBa-D/Z corpus of written German. *Dialogue and Discourse* pages 1–24.

Jianxiang Wang and Man Lan. 2015. A Refined End-to-End Discourse Parser. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning - Shared Task*. Association for Computational Linguistics, pages 17–24.