Exploiting a lexical resource for discourse connective disambiguation in German

Peter Bourgonje and **Manfred Stede**
Applied Computational Linguistics
Universität Potsdam
Potsdam, Germany
{bourgonje|stede}@uni.potsdam.de

Abstract

In this paper we focus on connective identification and sense classification for explicit discourse relations in German, as two individual sub-tasks of the overarching Shallow Discourse Parsing task. We successively augment a purely-empirical approach based on contextualised embeddings with linguistic knowledge encoded in a connective lexicon. In this way, we improve over published results for connective identification, achieving a final F_1-score of 87.93; and we introduce, to the best of our knowledge, first results for German sense classification, achieving an F_1-score of 87.13. Our approach demonstrates that a connective lexicon can be a valuable resource for those languages that do not have a large PDTB-style-annotated coprus available.

1 Introduction

An important difference between a text and a random collection of sentences is the amount of coherence it exhibits. In a text, sentences, or propositions therein, are connected through particular relations that can be, for example, *causal*, *temporal* or *contrastive*. Such relations can be obvious from the semantics of the involved text segments alone, as in (1), where the contrastive relation between male and female worker's earnings is easily inferred. Or they can be explicitly signaled by *(discourse) connectives*. Connectives are usually understood to be a closed class of syntactically heterogeneous words and phrases, and are known to be ambiguous in two different ways. In (2), there are no two particular propositions that are related to each other by the potential connective *once*, thus it is said to have *sentential* reading, as opposed to its *discourse* reading in (3) and (4). Furthermore, *once* is ambiguous with regard to the particular relations it can express, and signals a temporal relation in (3) and a conditional relation in (4).[1]

(1) *Earnings of year-round, full-time male workers fell 1.3% in 1988 after adjusting for higher prices. Earnings of female workers were unchanged.* (wsj_1815)

(2) *Once again the company's future looked less than rosy.* (wsj_0564)

(3) *Once it gets there, a company can do with it what it wishes.* (wsj_0989)

(4) *Normally, once the underlying investment is suspended from trading, the options on those investments also don't trade.* (wsj_1962)

Differentiating between sentential and discourse reading is often referred to as *connective identification*, and classifying the particular sense of a connective, or the relation it is involved in, is often referred to as *sense classification*. Both are sub-tasks of *discourse parsing*, which in turn has applications, for example, in text summarisation (Schilder, 2002; Yoshida et al., 2014), machine translation (Meyer and Popescu-Belis, 2012; Joty et al., 2014; Sim Smith, 2017) and argumentation mining (Eckle-Kohler et al., 2015).

The availability of annotated data for the task of discourse parsing as a whole, and consequently its sub-tasks, is limited. With the PDTB being by far the largest corpus annotated for discourse relations,

[1] All examples are taken from the PDTB (Prasad et al., 2008).

Proceedings of the 28th International Conference on Computational Linguistics, pages 5737–5748
Barcelona, Spain (Online), December 8-13, 2020

containing ~53k annotated relations in its 3.0 version (Prasad et al., 2019), any language other than English can be considered a low-resource language. However, in recent years, many connective lexicons, listing all connectives of the respective language and some of their properties, have become available for several languages.[2]

The main contribution of this paper is to investigate to what extent linguistic knowledge encoded in such lexicons can augment a purely-empirical approach to connective disambiguation in a low-resource scenario. Particularly, at first we fine-tune BERT (Devlin et al., 2019) to the task of both connective identification and sense classification for German, and then attempt to improve over this approach by exploiting DiMLex (Scheffler and Stede, 2016), a German connective lexicon. Additionally, we experiment with syntactically inspired features in the tradition of Pitler and Nenkova (2009). Our results demonstrate that exploiting a connective lexicon can improve performance both within-domain and across-domain in situations where limited training data is available.

The rest of this paper is structured as follows: Section 2 provides a more detailed definition of connectives and discusses related work on connective identification and sense classification. Section 3 explains the resources we use in our experiments. Section 4 outlines our method of combining BERT with information from the connective lexicon, and also syntactically inspired features for connective disambiguation. Section 5 presents and discusses the results, including the question of generalizing the approach to other languages. Finally, Section 6 sums up the key findings and provides an outlook on future work.

2 Background & Related Work

2.1 Background

In the operationalisation of discourse relations in the PDTB framework, explicit relations are those signaled by a connective, meaning that the definitions of explicit relations and connectives largely align. Not all frameworks dealing with coherence relations distinguish implicit from explicit relations as clearly as the PDTB. The notion is absent in Rhetorical Structure Theory (Mann and Thompson, 1988), for example. And even in the PDTB, the boundary is slightly faded by the inclusion of *AltLex* (for alternative lexicalisation) instances, indicating relations signaled by, in principle, open-class words and phrases, such as *the reason is that*. More recently, in the PDTB3, *AltLexC* instances are added, where a particular syntactic (as opposed to lexical) construction signals the discourse relation.

With regard to connectives, their long tradition of research (Schiffrin, 1987; Redeker, 1991; Knott and Dale, 1994; Degand et al., 2013), has recently been discussed by Danlos et al. (2018), who, in addition, differentiate between *primary* and *secondary* connectives, following Rysova and Rysova (2014).

In the DiMLex approach, the definition is based on Pasch et al. (2003), and then follows Stede (2002) by including certain prepositions. We thus adopt the characterisation that a lexical item *X* is a connective when:

- *X* is not inflectable,
- *X* expresses some specific, two-place semantic relation,
- the arguments of the relational meaning of *X* are propositional structures,
- the verbalisations of the arguments of the relational meaning of *X* can be clauses.

Note that this definition does not include any syntactic categorisation, and following this definition, connectives are a heterogeneous group of adverbials, sub-ordinating and co-ordinating conjunctions and prepositions. Furthermore, connectives can be single words or phrases (*as long as*), which in turn can be discontinuous (*either...or*, *if...then*). Some connectives (like *although* and *in spite of*) always have discourse reading, rendering their identification–in principle–a case of mere pattern matching. Others, however, may show a heavily skewed distribution, with a conjunction like *and* often having sentential reading, and a phrase like *on the other hand* rarely having sentential reading. The same holds for senses, where many connectives can signal only one particular relation sense, rendering their sense classification

[2]The connective lexicon platform `connective-lex.info` currently contains freely-available lexicons for Arabic, Bangla, Czech, Dutch, English, French, German, Italian, Portugese and Ukrainian.

redundant. Others may signal several different senses, with their particular sense distribution again being heavily skewed (see Section 3.1 for examples).

State of the art results in end-to-end shallow discourse parsing (Wang and Lan, 2015; Oepen et al., 2016) have been achieved using a pipeline architecture, introduced by Lin et al. (2014). In this pipeline, connective identification is the first, and explicit relation sense classification the third component (after argument extraction once a particular connective is located). Errors made here are propagated down the pipeline. Improving performance for these two sub-tasks can thus have major impact on end-to-end performance. The following subsection provides an overview of performance for these two tasks on different corpora and languages.

2.2 Related Work

Because of the data situation, most work on (shallow) discourse parsing is done on English. The 3.0 version of the PDTB (Prasad et al., 2019) contains ~53k annotated relations, compared to just over 1k explicit relations in the German corpus we use (see Section 3.1 for details). For most other languages that have corpora annotated for discourse relations (see Zeldes et al. (2019) for an overview), the number of available annotations is equally low, yet connective lexicons may exist for these languages (see Stede et al. (2019) for an overview). Thus, our approach, applied to German, is potentially useful for other languages for which the required lexicon exists.

Two consecutive shared tasks on end-to-end shallow discourse parsing (Xue et al., 2015; Xue et al., 2016) have spiked interest in the overarching task, which includes connective identification and sense classification. The two year's winning systems (Wang and Lan, 2015; Oepen et al., 2016) report F_1-scores of 94.16 and 94.4, respectively, for connective identification. For German, our language of interest, early work on connective identification is described in Dipper and Stede (2006), who use a subset of nine connectives and report an F_1-score of 93.95 for the functional disambiguation task on this subset. In earlier work (Bourgonje and Stede, 2018), we include all connectives present in the Potsdam Commentary Corpus and report an F_1-score of 83.89, by extending the syntactically inspired features of Pitler and Nenkova (2009). In addition, we discuss the effect of training data volumes for the connective identification task for English by iteratively down-sampling training data size and reporting the results. The work reported on in this paper improves upon Bourgonje and Stede (2018) by combining the syntactically inspired approach with a contextualised vector-based approach for connective identification (for German), and by introducing results for sense classification for German.

Regarding sense classification for explicit relations, Meyer and Popescu-Belis (2012) attempt to improve Machine Translation performance by disambiguating connectives. They report an F_1-score of 75 when classifying a subset of 13 temporal and contrastive connectives, using syntactic features, WordNet relations and candidate translations. For state-of-the-art performance in explicit sense classification on English, we turn to Wang and Lan (2015) and Oepen et al. (2016), who report scores of 90.79 and 90.01, respectively.[3] For German, earlier work (Kunz and Lapshinova-Koltunski, 2014) investigates connectives (referred to as cohesive conjunction strategies) in both German and English in both written and spoken language. The authors present statistics for sense distribution in their GECCo corpus, but do not attempt to classify the annotated instances and to the best of our knowledge, no prior work on sense classification for German exists.

The combination of low volumes of training data and neural network architectures for discourse parsing has been discussed in (Rutherford et al., 2017) for English and Chinese, but in our case, volumes are considerably lower still. The idea of augmenting neural approaches with external knowledge for the purpose of implicit sense classification is explored by Rutherford and Xue (2014), who use Brown cluster pairs and coreference patterns, and Kishimoto et al. (2018), who use ConceptNet in combination with coreference resolution.

[3] These numbers are without error propagation.

3 Data

3.1 Potsdam Commentary Corpus

To train and evaluate our approach, we use the Potsdam Commentary Corpus (PCC) in its 2.2 version (Bourgonje and Stede, 2020). The PCC is a corpus of 176 news commentary articles, comprising ~33k words and 1,120 explicit discourse relations. It is a multi-layer corpus, annotated for, among others, coreference chains, information structure, RST trees, and sentential syntax. For the syntactic features used in our experiments though, we use automatically produced parse trees instead of gold syntax trees. We argue that this provides a more realistic impression of performance on new, incoming text, and it allows us to directly compare results based on additional data (see subsection 3.2), for which we have no gold syntax trees available. Thus, we only use the layer of discourse relations following the PDTB definition. See Bourgonje and Stede (2020) for more details on this annotation layer.

The PCC contains 175 connective types. If we add the cases of sentential reading (i.e., cases like (2) in Section 1), we have a total of 2,677 instances to train and evaluate our approach on; 1,120 connective tokens and 1,557 non-connective tokens. Of the 175 connective types, 47 are singletons and only 66 occur more than 5 times, illustrating the rather long tail with low-frequent examples.

94 of the 175 connective types always have discourse reading in the PCC, meaning that for these cases the task of connective identification in theory could be handled by simple pattern matching (but our classifier has to learn them nonetheless, and so they are included in the train and test data). This group, however, only comprises 337 instances (13% of all data). Among the other 87%, distribution is heavily skewed. Connectives like *Und*[4] ("and"), *sondern* ("but/rather") and *wenn* ("if") have a high connective ratio, of 0.95, 0.93 and 0.97, respectively. On the other hand, connectives like *als* ("as"), *Wie* ("(such) as") and *durch* ("through/by") very seldom have connective reading, with a ratio of 0.08, 0.05 and 0.06, respectively. With regard to sense distributions, 75 connective types express one sense only in the PCC, rendering their sense classification again a simple case of pattern matching. However, this set makes up only 164 instances (6% of all data). The most ambiguous connective in the PCC is *dann* ("then"), expressing up to 5 different senses. In terms of surface form, the majority of connectives in the PCC are single words; there are 140 single word connective types (96% of all data), vs. 35 phrasal connective types (4% of all data). Of these 35 phrasal types, 17 are discontinuous (2% of all data).

The connective identification task uses binary labels. The sense classification task, using the PDTB3 sense hierarchy as labels, means doing 24-way[5] classification. See (Prasad et al., 2019) for more information on the inventory of senses being used.

3.2 Wikipedia & News Data

In order to establish potential domain effects and generally improve coverage, we sampled data from two additional corpora. The seed set we use for sourcing examples are connectives that do not appear at all in the PCC (but are present in DiMLex) and connectives that have been shown as hard to disambiguate, i.e., those with an F_1-score of less than 0.70 in the earlier connective identification experiments of Bourgonje and Stede (2018). Starting from this seed set, we sampled texts from Wikipedia and news articles. For Wikipedia, we used a dump from February 2018, while the news data originates from a German-English parallel corpus[6].

For all connectives in our seed set, we sampled up to 20 instances in total from both corpora; 10 from the news texts and 10 from Wikipedia. Because the Wikipedia articles were considerably larger in volume, if no 10 instances of a particular connective could be found in the news texts, we selected more from Wikipedia to arrive at 20 instances. For every connective candidate, we included the sentence it appeared in, the five previous sentences and the two following sentences. This maximises likelihood of capturing both arguments,[7] while minimising the amount of text the annotator had to read. These seg-

[4]Note that we make a distinction between upper- and lower-case here.

[5]Not all of the 30 classes in the PDTB3 sense hierarchy are represented in the PCC.

[6]`http://homepages.inf.ed.ac.uk/pkoehn/publications/de-news/`

[7]In the PCC, <0.4% of external arguments are in a sentence more than 4 sentences prior to the connective's sentence, while none of the external arguments are in a sentence more than 1 sentence after the connective's sentence.

ments were then annotated according to whether or not the candidate has connective or non-connective reading, and for the particular sense in case of the former.

The sampled texts were annotated by a single annotator using the same procedure as done by Stede and Neumann (2014), and a subset was annotated by a second annotator to calculate agreement. In 64.6% of cases, both annotators agreed on the candidate's function (being either a connective or not a connective). Working with just the PCC, Stede and Neumann (2014) had reported an agreement of 74.5% for this annotation task. As these (and various other) authors pointed out, in German, the problem of connective ambiguity is more severe than in many other languages due to German having a rich inventory of discourse particles, many of which also have a connective reading; and in many contexts the distinction is hard to pin down.

This small annotation campaign lead to an additional 3,124 instances (940 connectives, 2,184 non-connectives). Due to targeted sampling, the number of connective types (210) is higher than in the PCC. Of these, 78 always have discourse reading, making up 9% of all data. The number of connectives that are semantically unambiguous is larger than in the PCC, with 180 connective types (18% of all data) only ever expressing one particular sense.

3.3 DiMLex

At the core of our approach to improving performance for the two classification tasks by means of linguistic knowledge is the connective lexicon. Attempts to construct such a lexicon, exhaustively listing all connectives of a language in a both human- and machine-readable way (i.e., in an XML format), started with DiMLex (Stede, 2002), a lexicon for German connectives. In the last two decades, eight other languages followed, and the multi-lingual platform[8] hosting these lexicons is described by Stede et al. (2019). DiMLex itself as been further developed, and considerable improvements have been made by Scheffler and Stede (2016), in which the authors use four different strategies of expanding and completing the lexicon. Since only using the annotated senses from an annotated corpus, such as the PCC, would result in a circular procedure, the three other strategies consist of exploiting lexicons from different languages, consulting traditional linguistic or lexicographic literature, and sampling additional data to explore potential new senses for known connectives (see Scheffler and Stede (2016) for more details). Furthermore, validating and developing DiMLex by exploiting lexicons from different languages has been explored; in combination with a parallel corpus by Bourgonje et al. (2017); in combination with machine translation by Sluyter-Gäthje et al. (2020).

DiMLex contains 274 entries, of which 41 are indicated to always have discourse reading and 153 are semantically unambiguous (i.e., always express one particular sense). Each entry contains several attributes carrying additional information. In addition to the root form of the entry, orthographical variants are included, varying from casing difference (i.e., *Deswegen* vs. *deswegen*) to alternative spelling (i.e., *überhaupt* vs. *ueberhaupt*) and style differences (i.e., *sowohl...als auch* vs. *sowohl...wie auch*). Additionally, each entry carries its syntactic type, a specification of whether or not it can have sentential reading, corpus examples, and possible senses it can express. For more information we refer to the publicly available lexicon and its documentation[9]. Again, for the purposes of the present paper it is important to note that the the PCC annotations and the DiMLex entries have originally been created independently of each other.

The resources used in this paper are summarised in Tables 1 and 2, illustrating the key characteristics of the data sets (Table 1) and the connective lexicon (Table 2). The next section explains how DiMLex is exploited to improve performance for our two disambiguation tasks on these two data sets.

[8] http://connective-lex.info/

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

	PCC	WN
number of words	33,222	75,587
connective tokens	1,120	940
non-connective tokens	1,557	2,184
connective types	175	210

Table 1: Key characteristics of the Potsdam Commentary Corpus (PCC) and Wikipedia & News Data (WN).

	DiMLex
Entries	274
Potentially non-connective	233
Multiple senses	121

Table 2: Key characteristics of DiMLex.

4 Method

4.1 Connective Identification

BERT Baseline Our baseline system uses all connective types in the PCC as candidate items to be classified. Whenever a candidate is encountered in the text, we extract the entire sentence the candidate connective appears in. If the candidate is sentence-initial, we take its previous sentence as well, to provide more (preceding) context. For this textual input, we retrieve the BERT embedding. This is then concatenated with the candidate's single-word embedding. The reason for including the isolated embedding separately is to differentiate between candidates appearing in the same sentence. Consider the example sentence *"But traders took profits and focused on crude oil inventories once that factor was eliminated."* (wsj_1932), where *and* is annotated as having sentential reading and *once* is annotated as having discourse reading. Including the candidate separately prevents feeding the classifier two identical samples with different labels. Since we use the base version of a German BERT model,[10] this returns a 2304-dimensional vector.[11] This is then fed as input to a MultiLayer Perceptron classifier, following earlier work on similar problems (Ostendorff et al., 2019; Bai and Zhao, 2018; Pacheco et al., 2016).

BERT + DiMLex (surface form only) This system is essentially the same as the baseline system, but instead of using all connectives in the PCC as candidates, we now use all entries (plus their orthographical variants) of DiMLex as candidates. With the connectives in DiMLex being a superset of those in the PCC, in the PCC setup, this effectively only adds negative examples to the data; particular candidates occurring in the PCC, but never with discourse reading are now considered too. This adds 638 items to our data set. The main motivation for using all DiMLex entries as the candidate list is so that for other corpora, connective candidates not appearing in the PCC will also be considered for connective identification.

BERT + DiMLex ambiguity info Since information on whether or not a particular connective can have sentential reading is available in DiMLex (recall that 41 connectives in DiMLex always have discourse reading, see Section 3.3), we exploit this information by simply overruling the classifier prediction, as a post-processing step, in case it predicts a sentential reading when this does not correspond to its relevant DiMLex attribute. In addition, this setup assigns discourse reading for the relevant candidates from DiMLex, also if the candidate did not appear in the training data.

BERT + DiMLex + Syntactic features In an attempt to further improve upon this, we combine the previous setup (i.e., **BERT + DiMLex ambig.**) with a set of manually crafted features. We use the feature set from Bourgonje and Stede (2018), which in turn is based on the syntactic features from Pitler and Nenkova (2009) that are widely used in connective identification, extended by Lin et al. (2014). This feature set includes surface level and part-of-speech bigrams, the categories of the connective's parent node and that of its left and right siblings, whether or not the right sibling contains a VP, and the path to the root node. The values for these features are based on constituency trees obtained from the German

[10] `https://deepset.ai/german-bert`

[11] The first 786 positions are set to a default if the candidate is not sentence-initial.

Stanford LexParser (Rafferty and Manning, 2008). Since both the BERT and the syntactic feature sets contain information of a different kind, and crucially have different dimensions, we combine predictions from the MultiLayer Perceptron classifier and a RandomForest classifier (following Bourgonje and Stede (2018), who use this for connective identification) for the additional features, and average their predictions, assigning the same weight to both classifier predictions.

For the PCC, all numbers are the result of 10-fold cross-validation. Because we want to establish domain influence, for the Wikipedia & News setup (WN), training is done on the PCC, testing on the WN and results are averaged over 10 executions. We use weighted averaging for (individual) precision, recall and F_1-scores.

4.2 Sense Classification

Sense classification has two up-stream tasks in an end-to-end discourse parsing pipeline (i.e., connective identification and argument extraction). In this study however, we use the gold connectives directly from our annotations, and do not rely on the output of our connective identification component. We also use the gold argument spans from the annotations, and instead of using the connective's sentence and, potentially, its previous sentence, for the sense classification task we use both argument spans (*Arg1* and *Arg2*, which are fed to BERT in order of appearance). Another major difference is the number of labels; instead of binary classification, sense classification with the PCC as training data means a 24-way classification problem. Other than that, the procedure is mostly comparable to the connective identification setup.

BERT Baseline The baseline uses all connectives in the corpus, retrieves the 2304-dimensional vector for both argument spans from BERT and uses this as input to the MultiLayer Perceptron classifier. Since we use gold connectives in this step, using a different candidate list (i.e., **BERT + DiMLex (surface form only))** has no effect here, which is why the corresponding cells in Table 3 are empty.

BERT + DiMLex senses Instead of using the attribute in DiMLex stating whether or not a connective can have sentential reading, we extract the list of possible senses and if a sense not in this list is assigned, we overrule the classifier prediction with its most likely sense (i.e., most frequent in the training data).

BERT + DiMLex + Syntactic features Similar to the connective identification setup, we combine the **BERT + DiMLex senses** setup with the same set of features as for connective identification. Moreover, because these features pertain to the connective (and not to the the sentence it appears in, or its arguments), the feature values are identical to the connective identification scenario (though here, the labels are not binary, but one of 24 possibilities).

5 Results & Discussion

5.1 Results: Connective Identification

In (Bourgonje and Stede, 2018), a majority vote baseline of 79.60 is reported on the PCC, and a final classification result of 83.89[12]. Our baseline, which exploits BERT representations for the potential connective's current and, if applicable, previous sentence, improves over this majority vote baseline by almost 2 points, but is outperformed by the classifier described in (Bourgonje and Stede, 2018) by over 2 points in F_1-score. Adding all entries from DiMLex to our candidate list improves F_1-score by about 4.5 points for the PCC. For the WN setup, we equally see a jump in performance; an increase in F_1-score of about 5 points (from 62.37 to 67.55). As stated in Section 4, this procedure adds only negative samples in the PCC setup (i.e., candidates with sentential reading), but the classifier performs better on this larger data set. In the WN setup, it adds only few positive cases (only 27 samples), but we see a comparable increase in performance.

[12]Note that these results are obtained using an earlier version of the corpus though, and minor modifications to relevant annotations have been made, as reported on in (Bourgonje and Stede, 2020).

		Connective Identification		Sense Classification	
		PCC	WN	PCC	WN
BERT Baseline	Precision	81.78	75.56	82.13	42.52
	Recall	81.62	62.25	83.19	35.70
	F_1-score	81.53	62.81	81.32	33.73
BERT + DiMLex (surface form only)	Precision	86.24	81.63	—	—
	Recall	86.11	62.33	—	—
	F_1-score	86.14	67.55	—	—
BERT + DiMLex	Precision	86.64	81.60	**88.52**	**61.00**
	Recall	86.40	62.80	**88.13**	**49.85**
	F_1-score	86.33	67.96	**87.13**	**49.55**
BERT + DiMLex + Syntactic features	Precision	**88.06**	**81.87**	87.52	60.43
	Recall	**88.00**	**66.73**	87.28	49.53
	F_1-score	**87.93**	**71.12**	86.34	49.44

Table 3: Results for connective identification and sense classification.

Including additional information from DiMLex (i.e., whether or not particular connectives can have sentential reading) further improves performance, but the difference is small. Despite there being 41 connectives that always have discourse reading listed in DiMLex (see Section 4), in the PCC setup it occurs only 11 times that an incorrect prediction (i.e., sentential reading prediction for one of those 41 entries) is overruled and set to discourse reading during post-processing. This happens 29 times (average over 10 runs) in the WN setup, leading to a minor F_1-score improvement of 0.19 (which is not statistically significant ($p>0.02$)) and 0.41, in the PCC and WN setup, respectively.

Finally, adding in manually engineered and syntactically inspired features further improves performance for the PCC by about 1.5 points to a final F_1-score of 87.93 for connective identification in the PCC. In the WN setup, we obtain a final F_1-score of 71.12. This demonstrates that for connective identification, with the low number of training instances we have available for German, large-scale[13] neural approaches can be augmented with external knowledge encoded into lexicons and manually crafted and syntactically inspired features to improve performance.

5.2 Results: Sense Classification

Performance on sense classification in the PCC is very similar to performance for connective identification, despite the larger number of labels. For the WN setup, we see an even larger drop in performance, due to domain effects, i.e., a baseline of 42.52 for sense classification compared to 75.56 for connective identification. Using the sense information from DiMLex and, as a post-processing step, correcting all assigned labels that do not match with possible labels for the corresponding connective in DiMLex, results in a ~6 point increase in the PCC setup, and a considerably larger increase of ~16 points in the WN setup. Exploiting sense information from DiMLex thus seems to be a particularly effective way to counter-act domain effects. Adding syntactic features does not improve performance for sense classification, though the difference between the setup with and without syntactic features is not statistically significant ($p>0.02$) in either of the setups.

Since in the pipeline architecture mentioned in Section 2, i.e., Lin et al. (2014), sense classification has two upstream tasks (connective identification and argument extraction), we implemented the components described in this paper in an end-to-end German shallow discourse parser that is currently under development. The numbers included in Table 3 for sense classification rely on gold annotations for connectives and arguments. When using predicted connectives and argument spans from the parser that is being developed, we report an F_1-score of 60.41 (compared to 87.13 when using gold connectives and argument spans) for the PCC, and 32.55 (compared to 49.55) for the WN data, demonstrating the

[13]Despite the fact that BERT is pre-trained on huge amounts of data, and meant to be fine-tuned to a particular task using significantly less data, we argue that the amounts we have available are still very low.

severity of error propagation from upstream tasks. This underscores the importance of high performance especially for tasks early in the pipeline, and small improvements for connective identification can have major (positive) downstream impact.

5.3 Discussion: Lexicon Generation

The experiments reported in this paper rely on the existence of an external knowledge base in the form of a lexicon that (ideally) exhaustively lists the connectives of a particular language, augmented with additional information on ambiguity and potential senses. As pointed out earlier, such lexicons are already available for ten languages in a web-based database. For a language that does not have a connective lexicon, its creation is a relatively labour-intense task, but it can be sped up in various ways.

Provided that a corpus with the required annotations is available, a lexicon can be at least semi-automatically extracted from the annotations. This approach was taken by Das et al. (2018), who extracted an English lexicon from the PDTB and the RST Signaling Corpus (Das and Taboada, 2018).

If the required annotations are not available, or not on a scale supporting the extraction of a list anywhere near exhaustive, parallel corpora or machine translation, in combination with annotation projection can speed up development. An approach based on a parallel corpus is explored in Bourgonje et al. (2017). The combination of machine translation and annotation projection is described in Sluyter-Gäthje et al. (2020), who foremost aim to create German annotations, but populating (or extending) a lexicon can be a by-product of this approach. In addition to these data-oriented approaches, the intuition of native speakers who are familiar with discourse research can help in further completing such a lexicon. Such a combined, i.e., data-oriented and intuition-based, approach was described in (Bourgonje et al., 2018).

Furthermore, we note that the syntactic features in our best-performing setup for connective identification are relatively straightforwardly adapted to a different language if a constituency parser for the target language is available, and we refer to Bourgonje et al. (2018) for more information on this.

6 Conclusion & Future Work

We reported our experiments on two sub-tasks of discourse parsing: connective identification and sense classification (for explicit relations). Working on German, which in this respect can be considered a low-resource language, we use a state-of-the-art language modeling tool (BERT) that has proven to be successful in a variety of tasks for which typically considerably more training data is available. We demonstrate the effectiveness of augmenting this robust approach with linguistic knowledge as encoded in a connective lexicon, and with manually crafted and syntactically inspired features. In the best-scoring setup, we obtain an F_1-score of 87.93 for connective identification, and 87.13 for sense classification for explicit relations. This improves over previously published results for German connective identification and sets a first benchmark for German sense classification (for which, to the best of our knowledge, no published results are available).

Furthermore, we move beyond the German corpus that is regularly used in this kind of work (PCC) and evaluate our approach also on annotated data samples from Wikipedia and news texts and report a significant increase in performance. Due to the relatively low agreement scores on this annotated data though, we consider improving the quality of the annotations an important piece of future work.

Our experiments are done in the context of developing an end-to-end system for German discourse parsing. The vast majority of such end-to-end systems use a pipeline architecture, in which connective identification is the first, and (explicit relation) sense classification the third component. This paper presents sense classification results using gold connectives only, and we consider it an important piece of future work to put the individual components together to get an idea of performance in a pipeline setup, with error propagation.

Finally, we note that many discourse relations in a text are not signaled explicitly (at least not by connectives); in the PDTB2, explicit relations make up ~45% of all annotated relations. The processing of non-explicit relations is not covered in this paper, and is equally an important piece of future work. The recent literature using the English PDTB corpus has usually tackled this with neural approaches, but for low-resource languages, other solutions will probably have to be found.

Acknowledgements

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - 323949969. We would like to thank the anonymous reviewers for their helpful comments on an earlier version of this manuscript.

References

Bai, H. and Zhao, H. (2018). Deep enhanced representation for implicit discourse relation recognition. *CoRR*, abs/1807.05154.

Bourgonje, P. and Stede, M. (2018). Identifying explicit discourse connectives in German. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 327–331, Melbourne, Australia. Association for Computational Linguistics.

Bourgonje, P. and Stede, M. (2020). The potsdam commentary corpus 2.2: Extending annotations for shallow discourse parsing. In *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC 2020)*, Paris, France, May. European Language Resources Association (ELRA).

Bourgonje, P., Grishina, Y., and Stede, M. (2017). Toward a bilingual lexical database on connectives: Exploiting a German/Italian parallel corpus. In *Proceedings of the Fourth Italian Conference on Computational Linguistics*, Rome, Italy, December.

Bourgonje, P., Hoek, J., Evers-Vermeul, J., Redeker, G., Sanders, T., and Stede, M. (2018). Constructing a Lexicon of Dutch Discourse Connectives. *Computational Linguistics in the Netherlands Journal*, 8:163–175, 12/2018.

Danlos, L., Rysova, K., Rysova, M., and Stede, M. (2018). Primary and secondary discourse connectives: definitions and lexicons. *Dialogue and Discourse*, 9(1):50–78.

Das, D. and Taboada, M. (2018). RST Signalling Corpus: A Corpus of Signals of Coherence Relations. *Lang. Resour. Eval.*, 52(1):149–184, March.

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.

Degand, L., Cornillie, B., and Pietrandrea, P. (2013). Discourse markers and modal particles: Two sides of the same coin? In L. Degand, et al., editors, *Discourse markers and modal particles: Categorization and description*, pages 1–18. John Benjamins Publishing Company, Amsterdam/Philadelphia.

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 and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Dipper, S. and Stede, M. (2006). Disambiguating potential connectives. In Miriam Butt, editor, *Proc. of KONVENS '06*, pages 167–173, Konstanz.

Eckle-Kohler, J., Kluge, R., and Gurevych, I. (2015). On the role of discourse markers for discriminating claims and premises in argumentative discourse. In *Proc. Empirical Methods in Natural Language Processing*, pages 2236–2242.

Joty, S., Guzmán, F., Màrquez, L., and Nakov, P. (2014). DiscoTK: Using discourse structure for machine translation evaluation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 402–408, Baltimore, Maryland, USA, June. Association for Computational Linguistics.

Kishimoto, Y., Murawaki, Y., and Kurohashi, S. (2018). A Knowledge-Augmented Neural Network Model for Implicit Discourse Relation Classification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 584–595, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.

Knott, A. and Dale, R. (1994). Using linguistic phenomena to motivate a set of coherence relations. *Discourse Processes*, 18:35–62.

Kunz, K. A. and Lapshinova-Koltunski, E. (2014). Cohesive conjunctions in english and german: Systemic contrasts and textual differences. In Lieven Vandelanotte, et al., editors, *Recent Advances in Corpus Linguistics*, pages 229–262. Language and Computers - Studies in Practical Linguistics 78. Amsterdam/New York: Rodopi.

Lin, Z., Ng, H. T., and Kan, M.-Y. (2014). A PDTB-Styled End-to-End Discourse Parser. *Natural Language Engineering*, 20:151–184.

Mann, W. and Thompson, S. (1988). Rhetorical Structure Theory: Towards a Functional Theory of Text Organization. *TEXT*, 8:243–281.

Meyer, T. and Popescu-Belis, A. (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, page 129–138, USA. Association for Computational Linguistics.

Oepen, S., Read, J., Scheffler, T., Sidarenka, U., Stede, M., Velldal, E., and Øvrelid, L. (2016). OPT: Oslo–Potsdam–Teesside—Pipelining Rules, Rankers, and Classifier Ensembles for Shallow Discourse Parsing. In *Proceedings of the CONLL 2016 Shared Task*, Berlin.

Ostendorff, M., Bourgonje, P., Berger, M., Moreno-Schneider, J., and Rehm, G. (2019). Enriching BERT with Knowledge Graph Embeddings for Document Classification. In Steffen Remus, et al., editors, *Proceedings of the GermEval Workshop 2019 – Shared Task on the Hierarchical Classification of Blurbs*, Erlangen, Germany, 10. 8 October 2019.

Pacheco, M. L., Lee, I.-T., Zhang, X., Zehady, A. K., Daga, P., Jin, D., Parolia, A., and Goldwasser, D. (2016). Adapting event embedding for implicit discourse relation recognition. In *Proceedings of the CoNLL-16 shared task*, pages 136–142, Berlin, Germany, August. Association for Computational Linguistics.

Pasch, R., Braußе, U., Breindl, E., and Waßner, U. H. (2003). *Handbuch der deutschen Konnektoren*. Walter de Gruyter, Berlin/New York.

Pitler, E. and Nenkova, A. (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. Association for Computational Linguistics.

Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A., and Webber, B. (2008). The Penn Discourse Treebank 2.0. In *Proc. of the 6th International Conference on Language Resources and Evaluation (LREC)*, Marrakech, Morocco.

Prasad, R., Webber, B., Lee, A., and Joshi, A. (2019). Penn Discourse Treebank Version 3.0. In *LDC2019T05*. Philadelphia: Linguistic Data Consortium.

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

Redeker, G. (1991). Linguistic markers of discourse structure. *Linguistics*, 26:1139–1172.

Rutherford, A. and Xue, N. (2014). Discovering implicit discourse relations through brown cluster pair representation and coreference patterns. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 645–654, Gothenburg, Sweden, April. Association for Computational Linguistics.

Rutherford, A., Demberg, V., and Xue, N. (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: Volume 1, Long Papers*, pages 281–291, Valencia, Spain, April. Association for Computational Linguistics.

Rysova, M. and Rysova, K. (2014). The centre and periphery of discourse connectives. In *Proceedings of the28th Pacific Asia Conference on Language, Information and Computing (PACLIC)*, Bangkok, Thailand.

Scheffler, T. and Stede, M. (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)*, Portorož, Slovenia, may. European Language Resources Association (ELRA).

Schiffrin, D. (1987). *Discourse Markers*. Studies in Interactional Sociolinguistics. Cambridge University Press.

Schilder, F. (2002). Robust discourse parsing via discourse markers, topicality and position. *Natural Language Engineering*, 8(3):235–255, June.

Sim Smith, K. (2017). On Integrating Discourse in Machine Translation. In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 110–121, Copenhagen, Denmark, September. Association for Computational Linguistics.

Sluyter-Gäthje, H., Bourgonje, P., and Stede, M. (2020). Shallow discourse parsing for under-resourced languages: Combining machine translation and annotation projection. In *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC 2020)*, Paris, France, May. European Language Resources Association (ELRA).

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

Stede, M., Scheffler, T., and Mendes, A. (2019). Connective-lex: A web-based multilingual lexical resource for connectives. *Discours. Revue de linguistique, psycholinguistique et informatique*.

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

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

Xue, N., Ng, H. T., Pradhan, S., Prasad, R., Bryant, C., and Rutherford, A. (2015). The CoNLL-2015 shared task on shallow discourse parsing. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning - Shared Task*, pages 1–16, Beijing, China, July. Association for Computational Linguistics.

Xue, N., Ng, H. T., Pradhan, S., Rutherford, A., Webber, B., Wang, C., and Wang, H. (2016). CoNLL 2016 shared task on multilingual shallow discourse parsing. In *Proceedings of the CoNLL-16 shared task*, pages 1–19, Berlin, Germany, August. Association for Computational Linguistics.

Yoshida, Y., Suzuki, J., Hirao, T., and Nagata, M. (2014). Dependency-based discourse parser for single-document summarization. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1834–1839, Doha, Qatar, October. Association for Computational Linguistics.

Zeldes, A., Das, D., Maziero, E. G., Antonio, J., and Iruskieta, M. (2019). The DISRPT 2019 Shared Task on Elementary Discourse Unit Segmentation and Connective Detection. In *Proceedings of the Workshop on Discourse Relation Parsing and Treebanking 2019*, pages 97–104, Minneapolis, MN, June. Association for Computational Linguistics.