Comparison of methods for explicit discourse connective identification across various domains

Merel C.J. Scholman[1], **Tianai Dong**[1], **Frances Yung**[1] and **Vera Demberg**[1,2]
[1]Language Science and Technology, [2]Computer Science, Saarland University
Saarbrücken, Germany
{m.c.j.scholman,tdong,frances,vera}@coli.uni-saarland.de

Abstract

Existing parse methods use varying approaches to identify explicit discourse connectives, but their performance has not been consistently evaluated in comparison to each other, nor have they been evaluated consistently on text other than newspaper articles. We here assess the performance on explicit connective identification of four parse methods (PDTB e2e, Lin et al., 2014; the winner of CONLL2015, Wang and Lan, 2015; DisSent, Nie et al., 2019; and Discopy, Knaebel and Stede, 2020), along with a simple heuristic. We also examine how well these systems generalize to different datasets, namely newspaper text (PDTB), scientific text (BioDRB), prepared spoken text (TED-MDB) and spontaneous spoken text (Disco-SPICE). The results show that Discopy outperforms the other parse methods in all datasets, with the exception of DiscoSPICE. Moreover, performance drops significantly from the PDTB to all other datasets. We provide a more fine-grained analysis of domain differences and connectives that prove difficult to parse, in order to highlight the areas where gains can be made.

1 Introduction

Understanding the discourse relations that hold between segments in natural language text is crucial to many NLP applications, such as text generation, dialogue understanding, and question-answering systems. Shallow discourse parsers are used to uncover such relations by identifying connectives, extracting their arguments, and predicting the sense of the discourse relation. The current contribution focuses on the first step in the pipeline: finding explicit connectives in natural language. This step is not only interesting from the perspective of discourse relation classification, it can also be valuable for downstream applications, as accurate connective identification is crucial to mitigate the effect of cascaded errors downstream (Lin et al., 2014).

Connective identification is not a trivial task, as some connectives are ambiguous and may not consistently function as discourse connectives. A simple dictionary lookup would therefore not be able to distinguish between the discourse connective-usage of *yet* in Example (1), compared to the non-connective usage in (2) and (3).

(1) Julie wants to buy a house. Yet she has not found the right one.

(2) Julie wants to buy a house. She has yet to find the right one.

(3) Julie wants to buy a luxurious, yet affordable house.

In (1), *yet* expresses an adversative relation (equivalent to *nevertheless*). In (2), it functions as an adverb expressing a temporal meaning (i.e. *up until this time*). Finally, in (3), *yet* does not function in a relation with two complete discourse arguments, and would therefore not be annotated as a connective. An accurate parser would need to be able to distinguish between these fine-grained differences in the usage of connective candidates.

Further, connective usage diverges between domains. For example, *but* and *so* are often used as discourse structuring markers in spoken language, rather than discourse connectives. Existing parsers, however, have mainly been evaluated on newspaper text, since the largest discourse-annotated corpus available comes from this domain (PDTB, Prasad et al., 2008). Performance of existing discourse connective identification parsers on domains other than the written one is currently not well known.

Finally, we note that there is a lack of information on which connectives remain difficult to identify, even in state-of-the-art parsers. This is due to the tendency of studies reporting only the general accuracy, without providing detail on the accuracy on specific connectives. However, such information can provide the field with more information on what to focus on.

Proceedings of the 2nd Workshop on Computational Approaches to Discourse, pages 95–106
November 10–11, 2021. ©2021 Association for Computational Linguistics

In sum, performance of different parsing methods have not been consistently evaluated in comparison to each other, nor have they been evaluated consistently on text other than newspaper articles. In addition, previous work tends to report only overall accuracy, whereas a more fine-grained analysis of connectives might prove to be valuable for future efforts. The current contribution fills these gaps in the literature.

Specifically, we evaluate the performance on explicit connective identification of four parse methods (PDTB e2e, Lin et al., 2014; the winner of CONLL2015, Wang and Lan, 2015; DisSent, Nie et al., 2019; and Discopy, Knaebel and Stede, 2020), along with a simple heuristic as a baseline. The heuristic will identify all connectives in the data, without being able to distinguish discourse versus non-discourse usages of connectives. We include this "parse method" to provide insight into how discourse usage of connectives affects the performance of such a heuristic.

We also evaluate how well these systems generalize to different datasets from new domains, namely scientific text (BioDRB, Prasad et al., 2011), prepared spoken text (TED-MDB, Zeyrek et al., 2019) and spontaneous spoken text (DiscoSPICE, Rehbein et al., 2016).

2 Related work

Connective identification In recent years, shallow discourse parsing has received notable attention in the field. Much of this work has focused on the automatic labeling of implicit connectives, as performance there is significantly lacking (the current state-of-the-art achieves an F1 around 64% on a four-way classification, e.g. Ji et al., 2016; Lan et al., 2017; Shi and Demberg, 2019). Explicit relation identification has received less attention since prior work has reported high accuracy on explicit connectives in news articles. An open question is how these parsers perform on out-of-domain data.

Pitler and Nenkova (2009)'s work on explicit connective identification formed the basis for much subsequent research efforts in this direction. They show that syntactic features provide highly useful information for predicting whether a connective candidate functions as a discourse connective (an accuracy of 94.2% F1 on PDTB2 with a 10-fold cross validation on Sec. 02–22). Lin et al. (2014) built on this work to develop the PDTB end-to-end (PDTB e2e) discourse parser. Their parser performs at 95.4% F1 on the PDTB with a 10-fold cross validation on Sec. 02–22. In the context of the CONLL2015 shared task, Wang and Lan (2015) built on both of these approaches and presented the top-ranked system, which achieved an F1 score of 94.2% on connective identification in the PDTB2 section 23 test set and 91.9% in the CONLL 2015 blind test set. All three models rely heavily on various combinations of syntactic and lexical features extracted from texts.

More recent models have taken different approaches. DisSent uses dependency parsing and sentence embeddings to annotate discourse relations (Nie et al., 2019). They report an accuracy of 87.9% F1 on PDTB relations in determining whether a connective is present. Notably, they used sentences extracted from books as training data, rather than PDTB. This might make their approach more stable across domains. Finally, Knaebel and Stede (2020) use a neural approach that integrates contextualized word embeddings and predicts whether a connective candidate is part of a discourse relation or not. They achieve an F1 score of 97% on the PDTB2 section 23 test set.

2.1 Domain differences

In the current paper we focus on comparing the parsers' performance on two written corpora with the performance on two spoken corpora. Spoken data differs from written in a number of ways, including shorter sentence length on average and a higher rate of elliptical structures and omissions. Moreover, discourse connectives are used differently across domains. For example, compared to written data, spoken data tends to have a higher rate of explicit connectives, fewer connective types, and more non-discourse connective usage of connectives (see, e.g. Crible and Cuenca, 2017; Rehbein et al., 2016). It is also not uncommon for relations in spoken data to have an incomplete or even implicit sentence argument, as in Example (4), or to have connectives function as a discourse marker rather than connective, as in Example (5) (examples taken from DiscoSPICE):

(4) **And** uhm, bring cos uh unfortunately just she's been up in Belfast this week.

(5) And his face got really red. So I thought oh God. **So** yesterday when he he came back he said what were you saying to me about John.

In Example (4), the second argument (Arg2) for the connective *and* is not fully uttered; instead, the speaker cut off her utterance after the verb *bring* and switched to a different discourse relation. In Example (5), the speaker uses *so* in a non-discourse connective usage (the utterance cannot be paraphrased as “I thought oh God and as a result he asked me...”). It is unclear how parsers developed for the written domain would handle such cases that are more typical of the spoken domain.

Even within the written domain, differences in connective usage can occur between various types of written text. For example, Roman et al. (2016) found a higher rate of discourse connectives used in science textbooks compared to social studies textbooks. Moreover, specific connectives can occur more in one domain than another; for example, *in summary* occurs more commonly in biomedical abstracts than in the general written domain (Gopalan and Devi, 2016).

3 Method

3.1 Parse methods

Heuristic The heuristic is based on the list of 100 connectives from the PDTB2. This method simply extracts all connective candidates from the PDTB connective list, without distinguishing between usages. The heuristic will function as the baseline model.

PDTB end-to-end The PDTB end-to-end model is trained using the PDTB dataset sections 02-21, evaluated using sec 22, and tested on sec 23. To distinguish a discourse connective from its non-connective usage, Lin et al. (2014) extract a set of lexical and syntactic features for a connective and its preceding and following word. They also utilize the syntactic parse path from the connective to the root of the tree model, as well as the compressed path where adjacent identical tags are combined.

Note that a version of this parser specifically aimed at parsing BioDRB has been made available. Here, we use the general version of the parser to be able to consistently evaluate its performance on out-of-domain text across datasets.

CONLL2015 Similar to the PDTB e2e parser, the CONLL2015 winning parser is trained on PDTB sec 02-21, evaluated using sec 22, and tested on sec 23. Wang and Lan (2015) reimplemented well-established techniques from Pitler and Nenkova (2009) and Lin et al. (2014), and added the POS tags of nodes from the connective’s parent to capture more syntactic context information from the connective.

As part of the CONLL2015 shared task, competitors had access to the dependency tree, which we do not have access to. Instead, we generated parse trees using Stanford’s CoreNLP Natural Language Processing Toolkit (the same parser that was used by the e2e parser). We note that the parse tree result after running Stanford CoreNLP might be different from the one that is given by CONLL2015, which in turn might affect performance of this model.

DisSent In this model, connectives are used in the downstream task for learning sentence representation from explicit discourse relations. Nie et al. (2019) used texts from a BookCorpus (Zhu et al., 2015) to train and test their models. They identified common connectives in these texts, choosing those with a frequency greater than 1% in PDTB. As a result, the list of connectives that their parser can identify is much smaller than that used by other parsers (18 connectives compared to approximately 100 used by other parsers). Moreover, the possible dependency patters that are defined for each of these connective types is also restricted, and so this parser runs the risk of missing instances of the 18 connective types that it does aim to identify. We prepared the input for this parser using Stanford’s CoreNLP dependency parser.

Discopy The final model included is a recent neural model proposed by Knaebel and Stede (2020), which achieves state-of-the-art results in connective identification. Instead of using various combinations of syntactic or lexical features, the method implements a neural model which relies on pretrained word embeddings. The model concatenates contextualized embeddings of connectives with the embeddings of their contexts. For multi-word connectives, the contextualized embeddings of the single words are averaged. The model is then trained in a multi-task setting, to predict the connective or predict the coherence relation. They used PDTB sec 02–22 for training and sec 23-24 for testing. Here, we assess their best performing bert-based model on connective identification task, with context size of 1.

3.2 Data

Parser performance was measured on four different datasets in order to determine how well the parsers can identify connectives in various domains.

All datasets have already been annotated with discourse relations, and therefore we can use the gold connectives from these annotations.

PDTB2 sec 23 We evaluated all parsers on the Penn Discourse Treebank 2.0 (PDTB2, Prasad et al., 2008). The PDTB consists of discourse annotations on the Wall Street Journal texts. We here evaluate performance on section 23, which is commonly used as test dataset. According to the gold label annotations, the dataset contains 923 explicit connective tokens, with 62 unique types.

BioDRB We included text from the Biomedical Discourse Relation Bank (BioDRB; Prasad et al., 2011). This corpus consists of discourse annotations of 24 biomedical research articles from the GENIA corpus, using an adapted version of the PDTB2 annotation framework. These texts represent the biomedical, scientific text genre.

The BioDRB has 2636 gold explicit connective tokens with 180 unique connective types. The higher number of connective types in BioDRB compared to PDTB is mainly due to BioDRB having annotated modified connectives as unique types (e.g. *180 seconds after*, *due mainly to*), and to BioDRB annotating post-modified connectives (because of), which PDTB2 does not annotate. To make the comparison consistent across datasets, we mapped the gold connectives in BioDRB to the corresponding connective heads annotated in PDTB. We removed connectives considered to be alternative lexicalizations in PDTB. The final dataset contains 2574 connective tokens with 134 unique connective types.

TED-MDB TED-MDB (Zeyrek et al., 2019) is a resource of TED talk transcripts manually annotated for discourse relations. TED talks are highly structured speeches that are often minutely prepared and are meant to provide targeted information on various topics or ideas. The resource is multilingual, but we focus only on English in the current contribution.

TED-MDB currently consists of 6 TED talk transcripts annotated in PDTB3-style, with a total of 304 explicit connective tokens and 35 unique connective types. The reduced number of connective types compared to the PDTB can be attributed to the smaller size of this dataset as well as the genre. However, TED-MDB also includes connective types that are not included in PDTB2's connective list (some of which are part of PDTB3's connective list), such as *at, by, in* and *through*.

Disco-SPICE Disco-SPICE (Rehbein et al., 2016) is a corpus of transcribed broadcast interviews and telephone conversations from the SPICE-Ireland corpus (Kallen and Kirk, 2008), annotated in PDTB3-style. These texts represent a more informal, spontaneous spoken genre than the TED talks. This dataset contains 1163 explicit connective tokens with 50 unique connective types. Again, the reduced number of connective types can be attributed to the domain.

4 Results

4.1 Overall accuracy

Table 1 presents the accuracy of each parser. The results show that Discopy is most accurate in identifying connectives in the PDTB, BioDRB and TED-MDB, but the e2e parser displays a higher F1 score in DiscoSPICE. Moreover, we find that, across the board, performance significantly drops for datasets other than the PDTB. This emphasizes the need for out-of-domain evaluation and development. Chi-squared tests confirm that the difference in performance between the parsers and between the datasets are significant, see Appendix A.

Heuristic The heuristic uses PDTB's connective list as input to identify connectives, resulting in the highest recall on all datasets.[1] The heuristic's recall on BioDRB is lower than on other datasets because a portion of the connectives annotated in BioDRB are not included in PDTB's connective lexicon.

Its F1 score on spoken data is relatively high compared to other parsers, particularly for DiscoSPICE. This can be attributed to its insensitivity to syntactic requirements for Arg2 that are based on the written domain but are often impractical for the spoken domain. Hence, when syntactic features are too complex/inaccurate for spoken texts, connectives themselves can be used as reliable features.

Nevertheless, of all parse methods included, the heuristic's precision is lowest on all datasets. This reflects its high false positive rate, which is due to the heuristic not being able to distinguish between discourse versus non-discourse usage of connectives. Simply extracting all connectives using a heuristic might therefore be helpful for identifying

[1] Recall on PDTB is not 1 because there was one instance of *if...if...then*, where the gold standard attributed a single instance of *then* to both *if*'s.

	Heuristic			PDTB e2e			CONLL2015			DisSent			Discopy		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
PDTB	.29	.99	.46	.92	.91	.91	.92	.82	.87	.75	.32	.45	.97	.92	**.95**
BioDRB	.21	.67	.32	.80	.53	.63	.86	.31	.46	.48	.24	.32	.86	.56	**.68**
TED-MDB	.49	.92	.64	.81	.69	.74	.84	.66	.74	.64	.22	.33	.85	.71	**.77**
DiscoSPICE	.32	.88	.46	.41	.54	**.47**	.35	.37	.36	.54	.20	.29	.40	.48	.43

Table 1: Precision, Recall and F1 per corpus and parse method.

connectives, but human input or additional computational input would still be necessary to then decide on whether the candidate connective functions as a discourse connective.

PDTB e2e Performance of the PDTB e2e parser on the PDTB corpus is lower than reported in Lin et al. (2014) (91% versus 95% F1, respectively). This drop in the F1 score may be due to minor differences between our experimental set-up and theirs: we evaluated connective identification from final outputs of the end-to-end parser, while the original paper performed a separate training and testing on connective classifiers.

Applying the parser to datasets other than PDTB leads to a drop in performance. On BioDRB and TED-MDB, the e2e parser shows a high precision score but low recall score, which can be explained in part by the corpora having annotated a larger number of connective types than the PDTB2. Finally, we note low precision and recall scores on DiscoSPICE, with F1 scores on this spoken genre considerably lower than on the PDTB texts.

CONLL2015 Performance of the CONLL2015 winning parser on PDTB is lower than reported in the original paper (87% versus 94% F1, respectively, Wang and Lan, 2015). This drop in performance is likely caused by the difference in dependency tree parser used to prepare the data.

Further, this parser shows a low recall score on BioDRB, which can again be explained by the difference in connective lists. The performance on TED-MDB is in line with PDTB e2e, but performance on DiscoSPICE is relatively low. This might be attributed to the syntactic structure of the texts: Wang and Lan (2015)'s new features can capture more syntactic features from texts, but this might result in lower accuracy when the parsed structure diverges from that of the trained texts. This indicates that complex lexical and syntactic features are too restrictive and therefore not appropriate for spontaneous spoken texts like those in DiscoSPICE.

DisSent As mentioned in Section 3, DisSent targets only the 18 most frequently occurring connective types. Considering DisSent only takes a small portion of PDTB's connective types into account, it has the potential to uncover a relatively high number of connective tokens: a maximum of 86% of all gold connectives in PDTB, 43% in the BioDRB, 87% in TED-MDB, and 89% in DiscoSPICE. However, the results show that it missed a significant portion of these connectives: it identified 26% of all gold connectives in PDTB, 11% in the BioDRB, 19% in TED-MDB, and 18% in DiscoSPICE. Consequently, DisSent shows low recall scores in every corpus. In addition, it shows poor precision in BioDRB and TED-MDB specifically (but relatively high precision in DiscoSPICE compared to other parsers, which might be due to less diverse dependency patterns in DiscoSPICE). We conclude that DisSent can only be competitive if it is extended to include more connective types and syntactic patterns.

Discopy Discopy outperforms the other parsers on the PDTB, BioDRB, and TED-MDB. This indicates that word embeddings are more flexible than hand-engineered features, and can perform well without domain knowledge. The outlier for this parser is DiscoSPICE: e2e outperforms Discopy on this dataset (47% versus 43% F1). It seems that none of the parsers can parse connectives in this dataset with high accuracy, which could be explained by the features of transcribed spontaneous spoken language that are very different from written language (such as fragments, disfluencies and interjections).

The performance of the parser on PDTB is slightly lower than reported in the original paper, which could be a result of the implementation of the models.

4.2 Analyses of specific connectives

Detailed results on accuracy per connective for every dataset and parser are provided as an online appendix.[2] We here highlight some observations based on this data. We focus on results from the PDTB e2e, CONLL2015, and Discopy, since these provide the most coverage and behave similarly.

Certain connectives were identified accurately in all corpora by the parsers. These include *additionally, although, however, while, instead, meanwhile, nevertheless, therefore, unless,* and *whereas*. Many of these tend to function consistently as discourse connectives (i.e. they do not occur frequently in a non-connective usage), which might explain why they are easier to identify.

The remainder of this section presents a qualitative analysis, in which we consider in particular connectives that were included in the PDTB connective list, but were not identified correctly. We take the PDTB dataset as our reference model, and compare the other datasets to performance in this dataset. Most findings can be classified as an issue relating to the gold label, an issue relating to the parser, or an issue relating to token frequency. We end with general observations regarding DisSent. Table 4 in Appendix B presents a summary of the findings.

Performance on the PDTB dataset All three parsers show relatively poor accuracy on *or, nor, once* and *previously*. Moreover, all parsers miss some connective types altogether, such as *finally, specifically, rather, hence* (all parsers), *earlier* (e2e and CONLL), and *consequently* (CONLL and Discopy). Both of these issues are likely related to token frequency: these connectives all occur very infrequently in the test set and relatively infrequently in the training data. The poor performance might therefore be a reflection of the lack of training data; it remains a question whether the performance on these connectives will generalize if the test set would include more instances of them.

Both PDTB e2e and CONLL2015 (but not Discopy) also show poor performance on *as*, which occurs more frequently in the PDTB test set (n=40). The e2e parser shows higher precision than recall, whereas the CONLL2015 parser shows lower precision than recall. Manual inspection of the instances of *as* that were not identified did not reveal any pattern indicating why they might have been missed. As we will see in the next paragraphs, *as* proved to be difficult to identify in the other datasets as well. The divergence between the parsers on this connective is particularly interesting.

Genre-specific findings: BioDRB versus PDTB Performance on the following connectives in BioDRB was relatively poor in all parsers: *as, once, still, except,* and *after*. Note that performance for *as, once,* and *after* was also comparatively lower in PDTB (albeit higher than in BioDRB). Difficulty with these connective types is likely due to their frequent usage as non-discourse connectives. Regarding *still*, the e2e parser seems to have a specific issue: it only identified those instances of *still* in BioDRB that occurred argument-initially.

Second, we find an issue with the BioDRB gold label for *until*: the gold BioDRB dataset contains 11 instances of *until*, but all three parsers only identified one instance. For the “false negatives”, the Arg2 only contains a verb or noun phrase, as in (7). Such fragments are generally not considered to be full relational arguments and the connectives are therefore usually not annotated. It appears that BioDRB has more relaxed restrictions on what can constitute a relational argument. This can possibly explain the general trend of the parsers displaying relatively low recall compared to precision scores for BioDRB.

(7) E14.5 fetal thymic lobes were collected and stored in the TRIzol (GIBCO BRL) at -70C **until** RNA isolation.

We also find an issue with the gold label for instances of *also* in BioDRB: all parsers perform well on this connective in the PDTB but in the BioDRB, they show low precision. The gold BioDRB lists 92 instances of *also*, PDTB e2e identified 183 instances, CONLL 126 instances and Discopy 156 instances. Consider Example (8), a true positive identified by all parsers, and Example (9), a “false positive” identified by e2e and Discopy. Both relations are very similar, and so it is unclear why one instance of *also* was part of the gold dataset and the other was not.

(8) A much lower, but still significant increase was seen in fetal TN2 cells (25% increase, p < 0.01). Proliferation was **also** significantly higher at the fetal TN3 stage compared to adult (50% increase, p < 0.01).

(9) In the peripheral blood of OX35-treated rats, the percentage of CD3+ cells was significantly lower than in PBS-treated animals. The percentage of CD4+ cells in the OX35-treated

[2] https://osf.io/xzsa7/?view_only=0cc5d937fa264fb59b5fe5e035c1ac75

group was **also** significantly lower than that of the PBS control.

Similar to performance in the PDTB, the parsers show low performance on *previously* in BioDRB, which appears to be a difficult connective to identify accurately. In BioDRB, the errors can be attributed to both an issue with the gold label as well as an issue with the parsers. The gold contained no instances of *previously*, but the parsers all identified a different number of occurrences (3 to 8 instances). Some of these cases, such as Example (10), appear to be valid.

(10) In agreement with this, western blot analysis demonstrated an upregulation of Id1 protein, while the amount of Id2 and Id3 protein levels remained unchanged. **Previously**, Id1 has been considered not to be expressed in later developmental stages than pro-B cells (...).

Genre-specific findings: spoken versus written Spoken data is characterised by certain frequent connectives displaying a higher rate of non-connective usage compared to written data (e.g., *but*, *so*, *and*). As expected, these connectives show poorer performance in the spoken domain compared to the written domain.

We also find lower performance on *as*, *when* and *then* in the spoken domain. Regarding *as*, we can see this is a consistently difficult connective candidate to identify across all datasets. The poor performance of the parsers on *when* in TED-MDB can be attributed to an issue with the gold labels. Certain instances of *when* were not included in the gold TED-MDB, but the parsers were accurate in identifying these “false positives”. Example (11) presents an instance of *when* that was not part of the gold dataset but was identified accurately by the parsers.

(11) We thrive **when** we stay at our own leading edge.

Poor performance on *then* in DiscoSPICE can be attributed in part to missed instances in the gold dataset, but also to the common usage of *then* as a non-discourse connective in DiscoSPICE, as in Example (12). Such instances, where no clear arguments or relation sense can be identified for the connective, were not annotated in DiscoSPICE but the parsers did identify them.

(12) Yeah so, so hopefully just the three people are alright and they’re not. Cos **then** like with eleven people you’d be assured to have a few good people there but three people you’re just.

We also note a peculiarity for *if* in TED-MDB. The parsers identified false positives for *if* in TED-MDB, an issue which can be attributed to annotation standards. The instances were in fact rhetorical “what if” relations (see Example 13), whereby the first argument could be taken to be *what* (or rather, *what* substitutes Arg1). Annotation frameworks would likely not consider such instances as true connectives, since the *what* cannot constitute a full relational argument according to most segmentation conventions.

(13) What **if** they used that firepower to allocate more of their capital to companies working the hardest at solving these challenges or at least not exacerbating them?

Finally, we note two issues specific to transcribed spoken data. First, some connectives are sometimes spelled differently in transcribed text than in normal written text, such as *because* being transcribed as *cos*. Such phonetic spelling variants lead to a higher rate of false negatives. Second, some false negatives in DiscoSPICE can be attributed to disfluencies, interjections or fillers such as *uh* and *ehm*. For example, both PDTB e2e and CONLL (but not Discopy) missed an instance of the connective *after* in DiscoSPICE (Example 14) because the connective is immediately followed by a filler, which affected the syntactic parse of the argument. Both of these issues might be solved by additional preprocessing of the spoken text (correcting spelling variants and removing disfluencies), but a better solution would be to develop a parser for spoken data that can handle such characteristics.

(14) Fintan rang me actually right **after** *uhm* I put down the phone to you.

Connective-specific findings for DisSent Of the 18 connectives that DisSent aims to identify, it shows poor performance in most datasets on four connectives in particular: *and, so, as,* and *though.* Furthermore, DisSent fails to identify any instances of *but* in both the PDTB and BioDRB. These results can be explained by DisSent’s method:it does not identify connectives based on a heuristic search, but rather based on the syntactic pattern that the connective occurs in. The possible patterns that are provided for every connective are, in some cases, too restricted or coarse-grained, which is why the parser misses many instances of *and*, and all instances of *but*. This parser is hence extremely sensitive to the dependency parse of the dataset.

5 Discussion

Explicit connective identification can be done relatively reliably by existing parsers, but gains can still be made in this area. We therefore aimed to evaluate existing parsers and uncover more fine-grained errors. The results showed that Discopy (Knaebel and Stede, 2020) outperformed the other parsers in three out of four datasets: PDTB, BioDRB, and TED-MDB. This indicates that the contextualized embeddings used by Discopy are more flexible predictors of discourse connective usage than the syntactic and lexical features used by other parsers, even on out-of-domain data.

The exception to this is DiscoSPICE, for which the PDTB e2e parser performed best. However, even e2e's performance on this dataset was not sufficient. DiscoSPICE contains features and syntactic patterns that are specific to spoken data, such as disfluencies, incomplete sentence structures, and increased ambiguity of connectives (e.g., whether *so* is used as a connective or marker). These features can explain the low performance of all parsers on this dataset. There is still room for improvement in this area.

The performance of all parsers was lower on out-of-domain text compared to PDTB. This reaffirms earlier findings, which showed that a connective identification classifier trained on PDTB does not perform well on BioDRB even with domain adaptation techniques, compared to a classifier trained on the BioDRB alone (Ramesh and Yu, 2010; Prasad et al., 2011). Of course, these results cannot be considered surprising, given that prior work on discourse parsing has heavily focused on the written domain, with a strong bias towards newspaper text. Similar evaluations have not been done on other domains, nor have the parsers been applied frequently to other domains (but see Laali and Kosseim, 2014; Marchal et al., 2021, for an application of the e2e parser to spoken translated data). This underlines the importance of evaluation of existing parsers on other domains and the need for domain adaption of connective identification models and classifiers.

The results further indicated that performance is affected by the connectives' usage as non-discourse connectives and the connectives' frequency. One solution, as suggested by Lin et al. (2014), is to separately train a model for each highly ambiguous connective and another generic model to identify the remaining connectives. Another solution could be to provide parsers with more training data for infrequent connectives, which can be obtained via connective generation; this approach has recently been applied to address the lack of training data for implicit relations (Shi and Demberg, 2019; Kurfalı and Östling, 2021).

When analyzing connective-specific results, we found that poor performance could often be attributed to issues with the gold label. Some of the false positives seemed to be valid connectives, and might therefore actually be missed instances in the gold dataset. This highlights that in manual annotation, errors are still prevalent and maybe inevitable to some extent, which can affect performance of parsers. One way to remedy this is to support manual annotation with the output of SOTA parsers or even a simple heuristic, so that inconsistencies or false negatives would be less likely to occur.

We also found divergences between what frameworks and parsers consider relational arguments: BioDRB identified instances of *until* with only a noun phrase as the second argument, which the parsers consistently did not consider a relational connective. Conversely, Discopy consistently identified instances of *what if*-relations, which the gold standard did not consider to be relations. Discourse segmentation therefore has an impact on connective identification as well. Unfortunately, the issue of segmentation is still not entirely resolved in the field of discourse relation annotation (see, e.g., Hoek et al., 2018).

Many connectives annotated in the gold datasets could not be identified at all by the parsers because they rely on PDTB2's connective list. This list is not exhaustive and would need to be expanded for the parsers to provide more coverage. One place to start would be PDTB 3.0's connective list (Webber et al., 2019). However, even this list will not cover all connective types that might also occur in other genres (e.g. *hereafter* occurs in BioDRB but not PDTB). Extending connective lists with a general lexicon of English connectives (Das et al., 2018) can provide new connective candidates as well.

With regards to running the models, we observed a trade-off between accuracy and computational cost: the training and testing of Discopy (a bert-based model) required much more computational energy than any traditional parser (Bender et al., 2021). When comparing the success of neural methods and the other three parsers, it is important to be clear about the context in which they are used (Bender and Koller, 2020).

Finally, we note that our results did not perfectly replicate those of the original authors. This is likely due to a difference in the experimental set-up used (e.g., the dependency parse method used by Wang and Lan (2015) was not publicly available) and a difference in evaluation methods. To ensure that the lack of replicability was not due to incorrect implementation, the second author and another, unrelated researcher independently implemented all parse methods, and neither were able to perfectly replicate the results. Note that Han et al. (2020) also obtained different results than the original when replicating the e2e parser (although they observed improved scores), and attributed these divergences to minor variations between experimental set-ups in terms of implementations, hyperparameter settings and/or choice of the type of F1 score reported on. These results emphasize the general need in the field for more transparent reporting and a more consistent approach to evaluation (see also Kim et al., 2020, for implicit relation classification).

6 Conclusion

A comparison of different parse methods revealed that Discopy, a neural parser using sentence embeddings, generally outperforms parsers using syntactic and lexical features. The results also showed a severe performance drop when applying the parsers to other domains, especially spontaneous spoken discourse. This can be attributed to genre-specific syntactic structures, issues with the gold standards, and differences between connective lexicons. These results emphasise the need for out-of-domain training and evaluation, and provide insight as to where gains can be made.

Acknowledgements

This research was funded by the German Research Foundation (DFG) as part of SFB 1102 “Information Density and Linguistic Encoding”. We are grateful to Ida Novindasari and Anam Sadiq for their contributions to the implementations of the parsers.

References

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? . *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*.

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online. Association for Computational Linguistics.

Ludivine Crible and Maria-Josep Cuenca. 2017. Discourse markers in speech: characteristics and challenges for corpus annotation. *Dialogue and Discourse*, 8(2):149–166.

Debopam Das, Tatjana Scheffler, Peter Bourgonje, and Manfred Stede. 2018. Constructing a lexicon of English discourse connectives. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 360–365.

Sindhuja Gopalan and Sobha Lalitha Devi. 2016. BioDCA Identifier: A System for Automatic Identification of Discourse Connective and Arguments from Biomedical Text. In *Proceedings of the Fifth Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM2016)*, pages 89–98.

Kelvin Han, Phyllicia Leavitt, and Srilakshmi Balard. 2020. Comparing PTB and UD information for PDTB discourseconnective identification. In *Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Volume 3: Rencontre des Étudiants Chercheurs en Informatique pour le TAL*, pages 123–136.

Jet Hoek, Jacqueline Evers-Vermeul, and Ted JM Sanders. 2018. Segmenting discourse: Incorporating interpretation into segmentation? *Corpus Linguistics and Linguistic Theory*, 14(2):357–386.

Yangfeng Ji, Gholamreza Haffari, and Jacob Eisenstein. 2016. A latent variable recurrent neural network for discourse-driven language models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 332–342, San Diego, California. Association for Computational Linguistics.

Jeffrey L Kallen and John Monfries Kirk. 2008. *ICE-Ireland: A User's Guide: Documentation to Accompany the Ireland Component of the International Corpus of English (ICE-Ireland)*. Cló Ollscoil na Banríona.

Najoung Kim, Song Feng, Chulaka Gunasekara, and Luis Lastras. 2020. Implicit discourse relation classification: We need to talk about evaluation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5404–5414. Association for Computational Linguistics.

René Knaebel and Manfred Stede. 2020. Contextualized embeddings for connective disambiguation in shallow discourse parsing. In *Proceedings of the First Workshop on Computational Approaches to Discourse*, pages 65–75, Online. Association for Computational Linguistics.

Murathan Kurfalı and Robert Östling. 2021. Let's be explicit about that: Distant supervision for implicit discourse relation classification via connective prediction. *arXiv preprint arXiv:2106.03192*.

Majid Laali and Leila Kosseim. 2014. Inducing discourse connectives from parallel texts. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 610–619.

Man Lan, Jianxiang Wang, Yuanbin Wu, Zheng-Yu Niu, and Haifeng Wang. 2017. Multi-task attention-based neural networks for implicit discourse relationship representation and identification. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1299–1308, Copenhagen, Denmark. Association for Computational Linguistics.

Ziheng Lin, Hwee Tou Ng, and Min-Yen Kan. 2014. A PDTB-styled end-to-end discourse parser. *Natural Language Engineering*, 20(2):151–184.

Marian Marchal, Merel CJ Scholman, and Vera Demberg. 2021. Semi-automatic discourse annotation in a low-resource language: Developing a connective lexicon for Nigerian Pidgin. In *Proceedings of the Second Workshop on Computational Approaches to Discourse*.

Allen Nie, Erin Bennett, and Noah Goodman. 2019. DisSent: Learning sentence representations from explicit discourse relations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4497–4510, Florence, Italy. 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*, pages 13–16, Suntec, Singapore. 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 the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Rashmi Prasad, Susan McRoy, Nadya Frid, Aravind Joshi, and Hong Yu. 2011. The biomedical discourse relation bank. *BMC bioinformatics*, 12(1):1–18.

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.

Ines Rehbein, Merel Scholman, and Vera Demberg. 2016. Annotating discourse relations in spoken language: A comparison of the PDTB and CCR frameworks. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1039–1046, Portorož, Slovenia. European Language Resources Association (ELRA).

Diego Xavier Roman, Allison Briceño, Hannah Rohde, and Stephanie Hironaka. 2016. Linguistic cohesion in middle-school texts: A comparison of logical connectives usage in science and social studies textbooks. *The Electronic Journal for Research in Science & Mathematics Education*, 20(6).

Wei Shi and Vera Demberg. 2019. Learning to explicitate connectives with Seq2Seq network for implicit discourse relation classification. In *Proceedings of the 13th International Conference on Computational Semantics - Long Papers*, pages 188–199, Gothenburg, Sweden. Association for Computational Linguistics.

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*, pages 17–24, Beijing, China. Association for Computational Linguistics.

Bonnie Webber, Rashmi Prasad, Alan Lee, and Aravind Joshi. 2019. The Penn Discourse Treebank 3.0 annotation manual. *Philadelphia, University of Pennsylvania*.

Deniz Zeyrek, Amália Mendes, Yulia Grishina, Murathan Kurfalı, Samuel Gibbon, and Maciej Ogrodniczuk. 2019. Ted multilingual discourse bank (ted-mdb): a parallel corpus annotated in the pdtb style. *Language Resources and Evaluation*, pages 1–27.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27.

A Model comparisons

Chi^2 tests confirm that the difference in performance (TP and FN distribution) between the top three parsers is significant in each dataset, with the exception of TED-MDB, as displayed in Table 2. The top three parsers considered here because these were most competitive. Chi^2 tests confirm that the difference in performance (TP and FP distribution) between the datasets is significant in each parser, as shown in Table 3.

Dataset	χ^2	*df*	
PDTB	52.45	2	***
BioDRB	369.88	2	***
TED-MDB	1.74	2	
DiscoSPICE	69.25	2	***

Table 2: Chi^2 statistics to test whether there is a significant difference between the parsers per dataset.

Parser	χ^2	*df*	
e2e	879.32	3	***
CONLL	1005	3	***
Discopy	1204.8	3	***

Table 3: Chi^2 statistics to test whether there is a significant difference between the datasets per parser.

B Summary of connective-specific results

Table 4 presents the highlights of the connective-specific results per dataset.

	Poor performance	Unidentified connectives	Issues with gold label	Other observations
PDTB2	*as, or, nor, once, previously*	*finally, specifically, rather, hence*	-	PDTB e2e and CONLL2015 show low performance on *as*, but Discopy performs better.
BioDRB	*after, as, except, once, still*	*besides*	*until, also, previously*	BioDRB maintains different segmentation rules; New connective types not on PDTB2's list; PDTB e2e parser only identifies argument-initial instances of *still*.
TED-MDB	*and, as, so, then, when*	*for, on the one hand, rather*	*when*	Rhetorical "what if" relations present false positives; New connective types not on PDTB2's list.
Disco-SPICE	*and, as, but, so, then*	*after, also, finally, for, later, otherwise, still*	*then*	Phonetic spelling of connectives cannot be identified; Syntactic structures affect performance; PDTB e2e and CONLL are affected by interjections.

Table 4: Summary of connective-specific analysis per dataset.