Projecting Annotations for Discourse Relations: Connective Identification for Low-Resource Languages

Peter Bourgonje and **Pin-Jie Lin**
Saarland University, Saarbrücken, Germany
{peterb,pinjie}@lst.uni-saarland.de

Abstract

We present a pipeline for multi-lingual Shallow Discourse Parsing. The pipeline exploits machine translation and word alignment, by translating any incoming non-English input text into English, applying an English discourse parser, and projecting the found relations onto the original input text through word alignments. While the purpose of the pipeline is to provide rudimentary discourse relation annotations for low-resource languages (for which no annotations exist at all), in order to get an idea of performance, we evaluate it on the sub-task of discourse connective identification for several languages for which gold data are available. We experiment with different setups of our modular pipeline architecture and analyze intermediate results. Our code is made available on GitHub.

1 Introduction

Uncovering coherence relations in texts, also referred to as *discourse parsing*, is a complex task. It is comparably difficult and time-consuming for humans to annotate such relations, and as a result, relatively little training data is available for machines to train a system on. Most of this data is in English. Although recent shared tasks (Zeldes et al., 2019, 2021; Braud et al., 2023) have had a strong multilingual focus and included up to 13 different languages, there is still a large variety of languages that are seriously under-resourced when it comes to research on discourse and coherence.

In this paper, we attempt to address this issue by presenting an end-to-end, multi-lingual discourse parser. Our parser essentially consists of a processing pipeline that exploits machine translation, an English discourse parser, and a word aligner, to project discourse relation annotations onto any non-English input text, without a need for any language-specific training data. The goal of our pipeline is to kick-start the annotation of discourse relations in languages for which little to no resources are available, or to provide rudimentary discourse relation annotations for downstream applications where accuracy is not the main concern.

To get an idea of performance, we experiment with various different configurations of our modular architecture and evaluate on the sub-task of connective identification. We compare our results against a lexicon-based baseline that needs no training data either, and a state-of-the-art connective identification system trained specifically on the language and domain. Our pipeline mostly outperforms the lexicon-based baseline, by a factor of up to 2.7, and while a system specifically trained on the task outperforms our pipeline for all languages and corpora for which training data is available, we retain up to 81% of performance for some of those corpora. We analyze the (intermediate) results from different system configurations, in order to investigate which components of our processing pipeline are the most error-prone. We hope that our system proves to be a useful tool for researchers working on automated approaches to Shallow Discourse Parsing for languages for which little to no gold data is available.

The rest of this paper is organized as follows: Section 2 discusses related work, focusing mainly on discourse parsing. Section 3 explains our system architecture. Section 4 presents the results, which are discussed in Section 5. Finally, Section 6 sums up our main contributions and discusses directions for future work.

2 Related Work

In 2015 and 2016, two consecutive CoNLL shared tasks (Xue et al., 2015, 2016) caused a spark in interest in the discourse parsing task. The 2015 iteration worked with English only, the 2016 iteration was multi-lingual by adding Chinese. Both followed the Shallow Discourse Parsing paradigm proposed by the Penn Discourse TreeBank (PDTB,

Proceedings of the 5th Workshop on Computational Approaches to Discourse (CODI 2024), pages 39–49

Prasad et al. (2008, 2019)). This approach is often referred to as *Shallow Discourse Parsing* since contrary to other discourse parsing frameworks such as Rhetorical Structure Theory (RST, Mann and Thompson (1988)) or Segmented Discourse Representation Theory (SDRT, Asher and Lascarides (2003)), it makes no commitment to overall text structure, and deals with coherence relations on a local level.

PDTB parsing is often done in end-to-end fashion, with plain text as input and a list of discourse relations as output, where each relation consists of a relation type (*explicit, implicit, alternative lexicalization* or other), arguments and relation sense. Since the introduction of a pipeline architecture by Lin et al. (2014), many systems adopted this setup (Wang and Lan, 2015; Oepen et al., 2016; Knaebel, 2021). The majority of systems work on English, with some systems focusing on Chinese (Kang et al., 2016; Kong and Zhou, 2017; Chuan-An et al., 2018). Beyond that, to the best of our knowledge, the only other supported language for end-to-end parsing is German (Bourgonje, 2021).

With a series of shared tasks, the Discourse Relation Parsing and Treebanking workshops (DISRPT, Zeldes et al. (2019, 2021); Braud et al. (2023)) strongly encouraged a multi-lingual approach and moreover, attempted to converge work on different parsing paradigms, by including corpora following the annotation guidelines from both the PDTB, RST and SDRT. The shared task setup moved away from an end-to-end approach, and system submissions (Liu et al., 2023; Metheniti et al., 2023; Anuranjana, 2023) focused on particular sub-tasks.

Our contribution aims to enable discourse parsing for an even larger variety of languages, without the need for any language-specific discourse annotations. We hope that this opens up research into discourse parsing for seriously under-resourced languages. We integrate the end-to-end PDTB parser from Knaebel (2021), but in principle, an end-to-end RST parser (Joty et al., 2015; Heilman and Sagae, 2015; Ji and Eisenstein, 2014) could be plugged in as well. The components we implemented for both machine translation and word alignment were mostly selected because of their user-friendly APIs. However, our system architecture is modular by design, and systems focusing on particular, low-resource languages can easily be plugged in. A good example for machine translation is presented by Lin et al. (2023), whereas good examples for word alignment are provided by Procopio et al. (2021); Chen et al. (2021).

Using annotation projection for (sub-tasks of) discourse parsing is not novel. Laali and Kosseim (2017) use annotation projection from English to French on a parallel corpus (Europarl) and improve f1 score for discourse connective identification in French by 15 points. Sluyter-Gäthje et al. (2020) employ machine translation in combination with word alignment, in order to create a German corpus automatically annotated for discourse relations. However, in contrast to Laali and Kosseim (2017), we include machine translation and thereby dynamically enable discourse parsing for any language. In contrast to Sluyter-Gäthje et al. (2020), we focus on the pipeline itself and make that available, instead of focusing on curating and publishing the output of the process (e.g., a corpus annotated for discourse relations in a particular language).

3 Pipeline Architecture

The following subsections explain the three different components of our pipeline to annotate any non-English text with discourse relations, following the PDTB framework. We use a modular setup, such that individual components can be swapped out for alternatives that perform better for particular languages or domains. The system architecture is illustrated in Figure 1. The rounded boxes on the right depict the individual modules. The listed components are the ones we implemented, but for every module, additional components can easily be integrated. For machine translation, using a custom model, trained specifically for a low-resource language (pair) can improve performance. For the discourse parsing module, relevant alternatives that work end-to-end can be integrated. For word alignment, tools that can be trained on or tuned for specific language pairs might return better results. See Section 2 for some suggestions. As long as these components accept and return input/output in the same format, they can easily be interchanged. A detailed description of the modules and the components that we integrated into our pipeline is provided in the following subsections.

3.1 Machine Translation

The first step is translating any non-English input text into English. We integrated both the DeepL[1] and Google Tranlate[2] APIs. At time of writing

[1] https://github.com/DeepLcom/deepl-python
[2] https://pypi.org/project/googletrans/

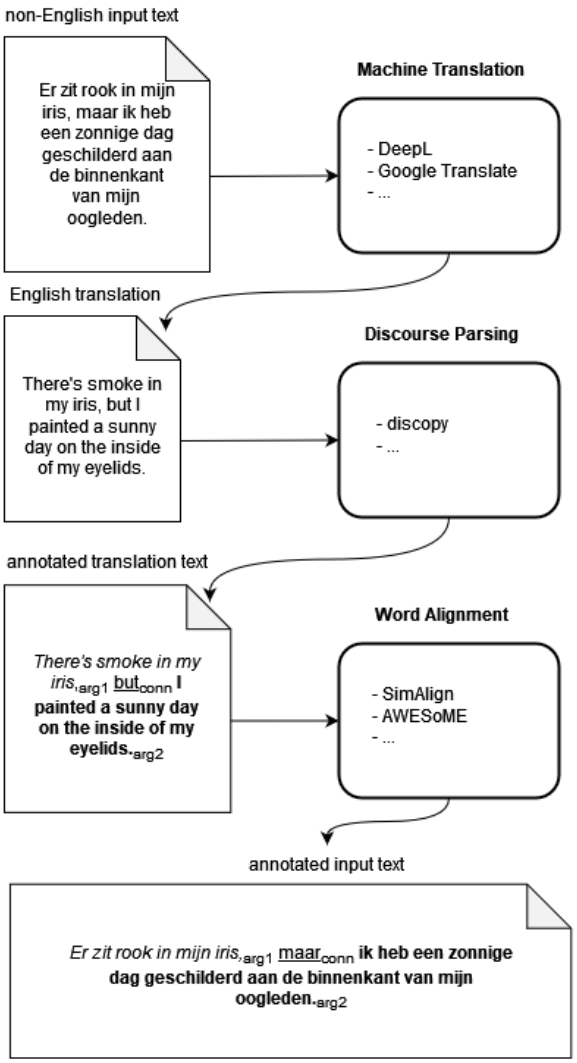

Figure 1: System architecture.

this, DeepL and Google Translate offer translations from/to 30 and 133 languages, respectively. For languages not included in either of those, or for domains where a specialized machine translation engine performs better, this module can easily be replaced by a custom machine translation engine.

Both the input and output format of this first module are a list of sentences; the original input text must be split into sentences, translation is then done sentence-by-sentence, and the output is a list of English sentences. The length of both input and output lists has to be identical.

The reason for translating sentence-by-sentence is that 1) the performance of word alignment is expected to be better when done sentence-based as opposed to text-based, and that 2) doing word alignment on longer texts rapidly leads to memory issues or long execution times. The drawback is that the English translation might be less fluent in cases where it might come more naturally to either merge or split up multiple sentences during translation.

3.2 Discourse Parsing

The second step in our pipeline is applying an end-to-end discourse parser on the English equivalent of the original input. We opted for English as an intermediate language because most training data annotated with PDTB-style discourse relations is available in English. For particular language pairs, if an end-to-end discourse parser is available in a language that is syntactically closer, using that might make sense, as word alignment can be expected to perform better in such a scenario. In our pipeline, we integrated discopy (Knaebel, 2021), because of its state-of-the-art performance and ease of use, accepting pre-tokenized input and running as a Docker container.

```
[
  ['There', "'s", 'smoke', 'in', 'my', 'iris', '.'],
  ['But', 'I', 'painted', 'a', 'sunny', 'day', 'on',
       'the', 'inside', 'of', 'my', 'eyelids', '.']
]
```

Listing 1: Example of discopy input format.

```
{"relations": [
    {
      "Arg1": {
        "CharacterSpanList": [
      [0, 25]
    ],
        "RawText": "There's smoke in my iris ,",
        "TokenList": [0, 1, 2, 3, 4, 5, 6]
      },
      "Arg2": {
        "CharacterSpanList": [
          [30, 80]
        ],
        "RawText": "I painted a sunny day on the
            inside of my eyelids.",
        "TokenList": [8, 9, 10, 11, 12, 13, 14, 15,
            16, 17, 18, 19]
      },
      "Connective": {
        "CharacterSpanList": [
          [26, 29]
        ],
        "RawText": "but",
        "TokenList": [7]
      },
      "DocID": -2650724294676803157,
      "ID": 0,
      "Sense": [
        "Comparison.Contrast"
      ],
      "Type": "Explicit"
    }
  ]
}
```

Listing 2: Example of discopy output format.

This module takes the translated and tokenized text as input. The input must be a list of sentences, which in turn consist of lists of tokens. Sentences

are already segmented in the previous step. Tokenization can be done with whatever method is most convenient to the user (e.g., spaCy, Stanza, UDPipe). An example of the required input format is included in Listing 1.

The output of this module is a JSON object, indicating where in the (English) input text, discourse relations have been found, indicated through both character offsets and token indices (based on the pre-tokenized input). An example is included in Listing 2.

3.3 Annotation Projection

The third and final step is that of projecting discourse relation annotations back onto the original input text. We integrated SimAlign (Jalili Sabet et al., 2020) and AWESoME (Dou and Neubig, 2021), but any word aligner that accepts sentence-segmented input and returns output in "Pharaoh format" can be used here. The Pharaoh format indicates which token in the source text corresponds to which token in the target text, and the example displayed in Figure 2 would be represented as follows:

```
[(0, 0), (1, 1), (2, 6), (3, 3), (4, 4), (5, 5)]
```

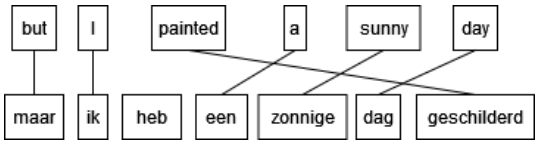

Figure 2: Word alignment example.

In this third and final step, we combine the results of the previous steps, with annotations all based on token indices and character offsets, to project the discourse relation annotations for the English translation back onto the original input text. For this, we use the same JSON format that discopy uses (see Listing 2), but now the annotations are on the original, non-English input text.

4 Results

Our pipeline is specifically targeted at low-resource languages for which no discourse relation annotations exist at all. However, without *any* gold data, we cannot get any idea of performance of our setup. So, in order to assess this across various languages and domains, we use the PDTB-style corpora featured in the 2023 DISRPT shared task (Braud et al., 2023) as gold data to evaluate our pipeline.

The shared task includes two sub-tasks, one focusing on segmentation and another focusing on relation sense classification. For PDTB-style corpora, the segmentation task is essentially about identifying connectives. The sense classification task assumes gold annotations for connective and relation arguments. While our pipeline returns discourse relations, fully specified with a connective, arguments and a relation sense, we decided to evaluate only on the segmentation task, i.e., connective identification, for now, as there is significant room for error propagation in our pipeline and we first want to get an idea of performance on the most upstream and comparably simpler task.

The segmentation sub-task for PDTB-style corpora includes five (non-English) languages (Italian, Portuguese, Turkish, Thai and Chinese), distributed over seven different corpora. An overview is included in Table 1.

Corpus	Domain
ita.pdtb.luna	IT helpdesk dialogs
por.pdtb.crpc	news, fiction
por.pdtb.tedm	TED talks
tha.pdtb.tdtb	news
tur.pdtb.tdb	news, fiction
tur.pdtb.tedm	TED talks
zho.pdtb.cdtb	news

Table 1: Overview of evaluation corpora.

In the task setup, the participants were provided with a train, dev and test set. Systems could therefore be trained and tuned for the relevant language using the train and dev sets. Since we do not train our system in any way for a particular language or domain, we do not expect to match performance of the trained systems that participated in the task, but for comparison, we do include results for DisCut (Kamaladdini Ezzabady et al., 2021; Metheniti et al., 2023), as this is the only system that submitted results for connective identification in the plain track, a setup that most resembles ours. We consider this the upper-bound of expected performance. To compare against a reasonable baseline that also does not require any pre-training and is aimed at low-resource/no-resource languages, we use a lexicon-based approach. This comprises simple pattern-matching using connective lexicons bundled on a dedicated platform[3]. Lexicons for all evaluation languages except Thai are available

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

on this platform. Similarly, because DeepL does not support Thai, the corresponding results are not included either. For all (other) corpora, we experiment with different configurations for the individual modules and their integrated components. We calculate precision, recall and f1 scores for all corpora, based on the *.test.conllu files from the shared task[4]. Since our system needs no training data, we could in principle evaluate against all available data (including train and dev sets), but to make a direct comparison to DisCut's performance possible, we evaluate on the test sets only. The results are included in Table 2.

5 Discussion

As illustrated by the performance of DisCut, with f1 scores generally in the 80s to 90s, given the availability of training data, identifying connectives is relatively easy, at least when compared to other subtasks in discourse parsing. We included the lexicon- and pattern-matching-based baseline, performing considerably worse, to indicate performance when no training data is used at all, since this much more resembles the targeted application scenario of our pipeline.

The mid section of Table 2 represents the results of experimenting with different system configurations. Overall, we can see that our annotation projection approach performs considerably better than the baseline, except for on **zho.pdtb.cdtb** and **ita.pdtb.luna**. However, a trained classifier performs significantly better still. Based on this relatively small set of languages and corpora, there does not seem to be a trend with regard to individual languages performing better or worse, as the difference within languages (46 and 64 for the two Portuguese corpora, and 42 and 48 for the two Turkish corpora, both for **deepl-discopy-awesome**) does not seem to be significantly smaller than the difference between languages.

The following sections discuss the influence of different system configurations with regard to machine translation and word alignment.

5.1 Machine Translation

By looking at the pairs for **deepl-discopy-simalign** and **googletrans-discopy-simalign** first, and **deepl-discopy-awesome** and **googletrans-discopy-awesome** second, we can see the influence of a difference in machine translation alone. For all languages except Chinese, the setup using DeepL performs better than the setup using Google Translate, with the difference in final f1 score ranging from 1 point (42 for **deepl-discopy-awesome** vs. 41 for **googletrans-discopy-awesome** on **tur.pdtb.tdb**), to 6 points (64 for **deepl-discopy-awesome** vs. 58 for **googletrans-discopy-awesome** on **por.pdtb.tedm**). For Chinese, the setup using Google Translate outperforms the setup using DeepL by up to 4 points. Recall that DeepL does not support Thai, hence no results using this in the setup can be provided for **tha.pdtb.tdtb**.

As noted in Section 3.1, the machine translation module only accepts input that is already split into sentences, and translation proceeds on a sentence-by-sentence basis. Translating sentences in isolation is likely to have a negative impact on translation output quality, since it will be less context-aware. This is particularly unfortunate as we are dealing with coherence relations, which are often realized beyond sentence boundaries. We consider it an important next step in the development of our system to feed the sentence-segmented input into the machine translation engine in batch-wise fashion. In this way, we can take context into account, but still force it to return the same number of output sentences as are present in the input, to allow for sentence-based word alignment.

5.1.1 Implicitation and Explicitation

Since we are translating discourse relations and are evaluating on the sub-task of connective identification, an issue known from the literature (Meyer and Webber, 2013; Lapshinova-Koltunski and Carl, 2022; Lapshinova-Koltunski et al., 2022; Yung et al., 2023) to take into account is *implicitation* and *explicitation*, where discourse connectives are either removed (explicit relations in the source text become implicit relations in the target text) or added (vice-versa) during translation. Especially implicitation has a negative effect on performance, as discourse connectives just disappear. Explicitation presumably does not affect performance that much in our evaluation setup, as in most cases, word alignment will not find any tokens in the source text that align to the newly added connectives in the target text. Both implicitation and explicitation are known to play out differently, depending on whether text is translated by machines or by humans, i.e., Meyer and Webber (2013) find an implicitation rate of up to 18% in human trans-

[4] https://github.com/disrpt/sharedtask2023data

	ita.pdtb.luna			por.pdtb.crpc			por.pdtb.tedm			tha.pdtb.tdtb		
	p	r	f1	p	r	f1	p	r	f1	p	r	f1
baseline	28	**58**	**38**	58	13	22	48	15	23	-	-	-
deepl-discopy-simalign	50	30	**38**	68	33	44	**85**	**52**	**64**	-	-	-
deepl-discopy-awesome	**51**	30	**38**	73	**34**	**46**	**85**	**52**	**64**	-	-	-
googletrans-discopy-simalign	48	29	36	72	28	41	81	46	58	52	20	29
googletrans-discopy-awesome	48	29	36	**75**	28	41	81	46	58	**61**	**21**	**32**
DisCut	*66*	*78*	*72*	*78*	*81*	*79*	*75*	*85*	*79*	*85*	*59*	*70*

	tur.pdtb.tdb			tur.pdtb.tedm			zho.pdtb.cdtb		
	p	r	f1	p	r	f1	p	r	f1
baseline	27	18	22	49	13	21	51	**30**	**38**
deepl-discopy-simalign	37	35	36	69	33	45	46	26	33
deepl-discopy-awesome	42	**42**	**42**	**79**	**34**	**48**	50	24	33
googletrans-discopy-simalign	36	34	35	64	21	42	57	28	37
googletrans-discopy-awesome	**45**	38	41	71	31	43	**58**	27	36
DisCut	*90*	*92*	*91*	*51*	*89*	*65*	*92*	*89*	*90*

Table 2: Results of four system configurations on seven non-English corpora. We compare our parser with a lexicon-based **baseline** and language-specific, trained system (**DisCut**). The reported scores are in percentages (%).

lations, and up to 8% in machine translations.

To investigate to what extent this effect may have negatively impacted performance of our system, we select one corpus where our pipeline did not outperform the baseline (**ita.pdtb.luna**) and one where it outperformed the baseline by quite some margin (**por.pdtb.tedm**). We look at implicitation, by selecting sentences that contain one or more connectives, and then checking if their English translation contains a *potential* connective, using Eng-DiMLex, an inventory of English discourse connectives (Das et al., 2018). If there is no match, we consider this a case of potential implicitation, and manually investigate further.

In **ita.pdtb.luna**, there are 202 sentences containing one or more connectives (of 1.304 sentences in total). Using the procedure described above, we find 23 instances of possible implicitation. Out of these 23 instances, 8 are cases where the input is too short to return a reasonable translation. Because the corpus consists of IT helpdesk dialogs, these include (possibly interrupted) turns in a dialog, such as *ma tanto noi* ("but we") and *perchè* ("why"). 4 instances contain *cioè*, which is consistently translated to "i.e." in English, which is not in Eng-DiMLex. This is basically a design decision (to not include abbreviations), since the semantically identical *for example* is included in Eng-DiMLex and would be annotated according to PDTB guidelines. Of the remaining 12 cases, 7 are cases of actual implicitation, and the other 5 are originating from the fact that the connective in Italian is one word, and corresponding candidates in English are phrasal. A frequent example is *che*, where the English equivalent *that* is present in the translation. But although Eng-DiMlex includes *given that*, *so that* and *after that*, for example, it does not include *that* in isolation.

In **por.pdtb.tedm**, there are 122 sentences containing one or more connectives (of 246 sentences in total), and 7 instances of possible implicitation. Upon manual investigation, we found that this includes 4 cases of actual implicitation, with the remaining 3 being border line cases, which according to the English, PDTB annotation guidelines (Eng-DiMLex is largely extracted from the PDTB) are not considered connectives. An example is *Agora podem vê# -la a desenrolar.* ("Now you can watch it unfold."), where *Agora* is annotated as a connective, whereas "Now" would probably not be annotated according to PDTB guidelines.

In all corpora except for **tur.pdtb.tdb**, recall is considerably lower than precision. We suspect that the reason for this is that we can "lose" connectives in our processing pipeline (which negatively impacts recall), but we can never "gain" new connectives to compensate for this. If discopy finds new connectives in the English translations (i.e., explicitation), they will not be projected back onto the original text, because they are implicit there. Upon investigation, we found that for **tur.pdtb.tedm**, with 247 connective tokens, only 119 were found

in the English translations, resulting in an upper-bound (if all instances found are correct) of 48% for recall. The subsequent annotation projection step actually retained all 119 instances. This suggests that the largest source of error is running discopy on the English translations.

Ultimately, it might be more relevant to consider a more holistic evaluation, focusing on which *relations* (including arguments and senses) have been found, instead of which *connectives* have been found. As explained earlier though, we first want to get an idea of performance of comparably simpler tasks, before we move to such a more abstract evaluation.

5.2 Discourse Parsing & Word Alignment

In our current setup, we only include one discourse parser, hence cannot experiment with different setups for this module. By investigating the rows **deepl-discopy-simalign** and **deepl-discopy-awesome** first, and **googletrans-discopy-simalign** and **googletrans-discopy-awesome** second, we can see the influence of a difference in word alignment alone. The setup using AWESoME outperforms the setup using SimAlign on all data sets except **zho.pdtb.cdtb**, where only in the setup using Google Translate, SimAlign returns better results. AWESoME outperforming SimAlign overall is in line with the findings of Dou and Neubig (2021), who compare their results to SimAlign as well.

In an attempt to isolate the effect of discourse parsing and word alignment quality on our final f1 score, we zoom in on one document from one particular corpus. We select talk_1976 from **por.pdtb.tedm** and investigate the best-performing setup for this corpus (**deepl-discopy-awesome**). Talk_1976 contains 59 connectives in its gold annotation. In the English translation of this document, discopy finds 38 relations, 34 of which are explicit (i.e., contain a connective). We found that all 34 connectives were true positives, and they were correctly aligned to the source connective. This is in line with the relatively high precision (85) for this corpus. During manual analysis, we noticed that most instances of explicit relations that were found, featured fairly frequent connectives like *e* ("and"), *mas* ("but"), *se* ("if") and *porque* ("because"), but less frequent connectives were missed by the parser. A case in point is *Agora* ("now") in *Agora podem vê# -la a desenrolar.* ("Now you can watch it unfold."), which was missed by discopy, although as mentioned in the previous section, this might actually not be considered a connective, according to the PDTB guidelines, and recall that discopy is trained on the PDTB.

Another example of this kind, resulting from the fact that the corpus discopy is trained on, might use a different definition than the corpus it is applied on, is *Bem, imaginemos que pegamos no Telescópio Espacial Hubble e o rodamos e o deslocamos para a órbita de Marte.* ("Well, let's imagine we take the Hubble Space Telescope and rotate it and move it into orbit around Mars."), where *imaginemos* ("let's imagine") is annotated as a connective. Similarly, in *Seria o mesmo se erguesse o meu polegar e bloqueasse o ponto luminoso à frente de_ os meus olhos* ("It would be the same if I raised my thumb and blocked the light spot in front of my eyes"), *Seria o mesmo* ("It would be the same") is annotated as a connective. Such examples would most likely be annotated as alternative lexicalizations in the English PDTB, but other corpora might have different definitions. We refer to Danlos et al. (2018) for a detailed discussion, and furthermore note that because in this paper, we are evaluating on connectives specifically, this issue is particularly challenging. For users interested in discourse parsing in general (without specifically looking at connectives), it might be less important whether some relation is found as an Explicit or as an AltLex type relation, as long as it is found.

5.3 Domain Transfer

Based on the 7 corpora, distributed over 5 different languages, we do not observe a significantly larger variance in f1 score across languages, compared to within languages. The two best-scoring corpora are both from the TED Multilingual Discourse Bank (Zeyrek et al., 2018). This raises the question as to whether expected performance is determined by original language or, rather, by original domain. Discopy has been trained on the original, English PDTB corpus, which represents the financial news domain (Wall Street Journal articles). The two TED corpora **por.pdtb.tedm** and **tur.pdtb.tedm** contain *"prepared, formal monologues (...) delivered to a live audience"* (Zeyrek et al., 2018, pp.1915), on a variety of topics. At first glance, this does not necessarily resemble WSJ articles. While one of the corpora for which our pipeline does not outperform the baseline, **ita.pdtb.luna**, is from an even less similar genre (spoken dialogs from the IT helpdesk domain (Tonelli et al., 2010)), the other corpus, **zho.pdtb.ctdb** consists of newswire text

(Zhou et al., 2014), which at first glance seems very similar to the domain discopy was trained on. In the 2023 shared task, **por.pdtb.tedm** and **tur.pdtb.tedm** corresponded to the "Out of Domain" setting. For Turkish, this seems to have had a major impact on a system trained on a different domain, as demonstrated by the performance drop from 91 (**tur.pdtb.tdb**) to 65 (**tur.pdtb.tedm**) for DisCut. However, such a drop is not observed for DisCut's performance on Portuguese, with both corpora having the same f1 score.

6 Conclusion & Future Work

We present a multi-lingual Shallow Discourse Parsing pipeline that makes use of machine translation, an English discourse parser and word alignment to project annotations onto the original, non-English input text. We specifically aim to support low-resource scenarios and make rudimentary discourse parsing possible for languages without any available training data, since our pipeline needs no training data at all. Our code is made available online.[5]

We evaluate our approach on the sub-task of connective identification and compare different configurations of our pipeline to a lexicon-based baseline, and to a system specifically designed for the task and trained on in-language, in-domain data. Our system outperforms the baseline in most cases, and for individual corpora improves f1 score by a factor of 2.7. We find that a trained system still performs considerably better, but for the best-scoring corpus, we retain 81% of the upper-bound f1 score.

In our current architecture, translation is done sentence-by-sentence, so as to keep sentences aligned for better word alignment performance. We consider more context-aware translation (Herold and Ney, 2023) the most important piece of future work. In addition, further investigation of error propagation, as well as the effect of domain transfer, are promising venues for future work. In this paper, we evaluate our approach on the sub-task of connective identification only. Our pipeline returns fully specified relations (with a type, arguments and relation sense), and we leave it to future work to evaluate on more than just connective identification. Relevant related work in this respect is represented by Kurfalı and Östling (2019), who work on implicit relation classification without exploiting any (language-specific) training data, and we consider it an important next step to experiment with zero-shot transfer (Kurfalı and Östling, 2019, 2021) for other sub-tasks of discourse parsing.

Our system architecture is modular by design, with relatively common exchange formats (Pharaoh for word alignments, PDTB-style JSON for discourse relations) across modules, and where individual components fine-tuned to a particular language are available, these can easily plugged in. Furthermore, our current architecture includes only a PDTB parser and another possible extension is the integration of RST parsers.

7 Limitations

In our pipeline, we integrated two alternatives for machine translation, and two alternatives for word alignment. Due to the limited availability of end-to-end Shallow discourse parsers, we only include one such parser in our setup and evaluation. Since we see systematic differences in performance for both machine translation and word alignment, depending on which module is used, integrating more components would provide a broader perspective. Especially since both alignment components are designed to work out-of-the-box, without any fine-tuning, which most likely means that they will work best on languages not too dissimilar to English.

Since we use a discourse parser trained on one specific English corpus, from one domain (financial news), we consider this the most prominent limitation of our system. While through this very work, we attempt to open up discourse research to under-resourced languages, we recognize that we may actually end up enforcing principles and paradigms that happen to work well for English onto languages where discourse relations may be realized in different ways. We already observe and discuss examples of this kind in Section 5.2. While we believe that our work may support the creation of corpora in other languages, it is important to keep this in mind and attempt to minimize bias when using the output of our system in annotation campaigns.

Acknowledgments

This research was funded by the German Research Foundation (DFG), Project-ID 232722074 – SFB 1102: Information Density and Linguistic Encoding. We thank the reviewers for their comments on an earlier version of this manuscript.

[5] https://github.com/PeterBourgonje/projan-disco/

References

Kaveri Anuranjana. 2023. DiscoFlan: Instruction Fine-tuning and Refined Text Generation for Discourse Relation Label Classification. In *Proceedings of the 3rd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2023)*, pages 22–28, Toronto, Canada. The Association for Computational Linguistics.

Nicholas Asher and Alex Lascarides. 2003. *Logics of Conversation*. Studies in Natural Language Processing. Cambridge University Press.

Peter Bourgonje. 2021. *Shallow Discourse Parsing for German*. Doctoral thesis, Universität Potsdam.

Chloé Braud, Yang Janet Liu, Eleni Metheniti, Philippe Muller, Laura Rivière, Attapol Rutherford, and Amir Zeldes. 2023. The DISRPT 2023 Shared Task on Elementary Discourse Unit Segmentation, Connective Detection, and Relation Classification. In *Proceedings of the 3rd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2023)*, pages 1–21, Toronto, Canada. The Association for Computational Linguistics.

Chi Chen, Maosong Sun, and Yang Liu. 2021. MaskAlign: Self-Supervised Neural Word Alignment. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4781–4791, Online. Association for Computational Linguistics.

Lin Chuan-An, Hen-Hsen Huang, Zi-Yuan Chen, and Hsin-Hsi Chen. 2018. A Unified RvNN Framework for End-to-End Chinese Discourse Parsing. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 73–77, Santa Fe, New Mexico. Association for Computational Linguistics.

Laurence Danlos, Katerina Rysova, Magdalena Rysova, and Manfred Stede. 2018. Primary and Secondary Discourse Connectives: Definitions and Lexicons. *Dialogue and Discourse*, 9(1):50–78.

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, Melbourne, Australia. Association for Computational Linguistics.

Zi-Yi Dou and Graham Neubig. 2021. Word Alignment by Fine-tuning Embeddings on Parallel Corpora. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2112–2128, Online. Association for Computational Linguistics.

Michael Heilman and Kenji Sagae. 2015. Fast rhetorical structure theory discourse parsing. *CoRR*, abs/1505.02425.

Christian Herold and Hermann Ney. 2023. Improving long context document-level machine translation. In *Proceedings of the 4th Workshop on Computational Approaches to Discourse (CODI 2023)*, pages 112–125, Toronto, Canada. Association for Computational Linguistics.

Masoud Jalili Sabet, Philipp Dufter, François Yvon, and Hinrich Schütze. 2020. SimAlign: High Quality Word Alignments without Parallel Training Data using Static and Contextualized Embeddings. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 1627–1643, Online. Association for Computational Linguistics.

Yangfeng Ji and Jacob Eisenstein. 2014. Representation Learning for Text-level Discourse Parsing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13–24, Baltimore, Maryland. Association for Computational Linguistics.

Shafiq Joty, Giuseppe Carenini, and Raymond T. Ng. 2015. CODRA: A Novel Discriminative Framework for Rhetorical Analysis. *Computational Linguistics*, 41(3):385–435.

Morteza Kamaladdini Ezzabady, Philippe Muller, and Chloé Braud. 2021. Multi-lingual Discourse Segmentation and Connective Identification: MELODI at Disrpt2021. In *Proceedings of the 2nd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2021)*, pages 22–32, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xiaomian Kang, Haoran Li, Long Zhou, Jiajun Zhang, and Chengqing Zong. 2016. An End-to-End Chinese Discourse Parser with Adaptation to Explicit and Non-explicit Relation Recognition. In *Proceedings of the CoNLL-16 shared task*, pages 27–32, Berlin, Germany. Association for Computational Linguistics.

René Knaebel. 2021. discopy: A Neural System for Shallow Discourse Parsing. In *Proceedings of the 2nd Workshop on Computational Approaches to Discourse*, pages 128–133, Punta Cana, Dominican Republic and Online. Association for Computational Linguistics.

Fang Kong and Guodong Zhou. 2017. A CDT-Styled End-to-End Chinese Discourse Parser. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 16(4).

Murathan Kurfalı and Robert Östling. 2019. Zero-shot transfer for implicit discourse relation classification. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 226–231, Stockholm, Sweden. Association for Computational Linguistics.

Murathan Kurfalı and Robert Östling. 2021. Probing multilingual language models for discourse. In *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*, pages 8–19, Online. Association for Computational Linguistics.

Majid Laali and Leila Kosseim. 2017. Improving Discourse Relation Projection to Build Discourse Annotated Corpora. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 407–416, Varna, Bulgaria. INCOMA Ltd.

Ekaterina Lapshinova-Koltunski and Michael Carl. 2022. Using Translation Process Data to Explore Explicitation and Implicitation through Discourse Connectives. In *Proceedings of the 3rd Workshop on Computational Approaches to Discourse*, pages 42–47, Gyeongju, Republic of Korea and Online. International Conference on Computational Linguistics.

Ekaterina Lapshinova-Koltunski, Christina Pollkläsener, and Heike Przybyl. 2022. Exploring Explicitation and Implicitation in Parallel Interpreting and Translation Corpora. *The Prague Bulletin of Mathematical Linguistics*, 119:5–22.

Pin-Jie Lin, Muhammed Saeed, Ernie Chang, and Merel Scholman. 2023. Low-Resource Cross-Lingual Adaptive Training for Nigerian Pidgin. In *Proceedings of INTERSPEECH 2023*, pages 3954–3958.

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

Wei Liu, Yi Fan, and Michael Strube. 2023. HITS at DISRPT 2023: Discourse Segmentation, Connective Detection, and Relation Classification. In *Proceedings of the 3rd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2023)*, pages 43–49, Toronto, Canada. The Association for Computational Linguistics.

William Mann and Sandra Thompson. 1988. Rhetorical Structure Theory: Towards a Functional Theory of Text Organization. *TEXT*, 8:243–281.

Eleni Metheniti, Chloé Braud, Philippe Muller, and Laura Rivière. 2023. DisCut and DiscReT: MELODI at DISRPT 2023. In *Proceedings of the 3rd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2023)*, pages 29–42, Toronto, Canada. The Association for Computational Linguistics.

Thomas Meyer and Bonnie Webber. 2013. Implicitation of Discourse Connectives in (Machine) Translation. In *Proceedings of the Workshop on Discourse in Machine Translation*, pages 19–26, Sofia, Bulgaria. Association for Computational Linguistics.

Stephan Oepen, Jonathon Read, Tatjana Scheffler, Uladzimir Sidarenka, Manfred Stede, Erik Velldal, and Lilja Øvrelid. 2016. OPT: Oslo–Potsdam–Teesside—Pipelining Rules, Rankers, and Classifier Ensembles for Shallow Discourse Parsing. In *Proceedings of the 20th Conference on Computational Natural Language Learning: Shared Task (CoNLL Shared Task 2016)*, pages 20–26, Berlin.

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)*, pages 2961–2968, Marrakech, Morocco. European Language Resources Association (ELRA).

Rashmi Prasad, Bonnie Webber, Alan Lee, and Aravind Joshi. 2019. Penn Discourse Treebank Version 3.0, LDC2019T05.

Luigi Procopio, Edoardo Barba, Federico Martelli, and Roberto Navigli. 2021. MultiMirror: Neural Cross-lingual Word Alignment for Multilingual Word Sense Disambiguation. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 3915–3921. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Henny Sluyter-Gäthje, Peter Bourgonje, and Manfred Stede. 2020. Shallow Discourse Parsing for Under-Resourced Languages: Combining Machine Translation and Annotation Projection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1044–1050, Marseille, France. European Language Resources Association.

Sara Tonelli, Giuseppe Riccardi, Rashmi Prasad, and Aravind Joshi. 2010. Annotation of Discourse Relations for Conversational Spoken Dialogs. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Jianxiang Wang and Man Lan. 2015. A Refined End-to-End Discourse Parser. In *Proceedings of the 19th Conference on Computational Natural Language Learning: Shared Task (CoNLL Shared Task 2015)*, pages 17–24. Association for Computational Linguistics.

Nianwen Xue, Hwee Tou Ng, Sameer Pradhan, Rashmi Prasad, Christopher Bryant, and Attapol Rutherford. 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. Association for Computational Linguistics.

Nianwen Xue, Hwee Tou Ng, Sameer Pradhan, Attapol Rutherford, Bonnie Webber, Chuan Wang, and Hongmin Wang. 2016. CoNLL 2016 Shared Task on Multilingual Shallow Discourse Parsing. In *Proceedings of the CoNLL-16 shared task*, pages 1–19, Berlin, Germany. Association for Computational Linguistics.

Frances Yung, Merel Scholman, Ekaterina Lapshinova-Koltunski, Christina Pollkläsener, and Vera Demberg. 2023. Investigating Explicitation of Discourse Connectives in Translation using Automatic Annotations. In *Proceedings of the 24th Annual Meeting of the*

Special Interest Group on Discourse and Dialogue, pages 21–30, Prague, Czechia. Association for Computational Linguistics.

Amir Zeldes, Debopam Das, Erick Galani Maziero, Juliano Antonio, and Mikel Iruskieta. 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. Association for Computational Linguistics.

Amir Zeldes, Yang Janet Liu, Mikel Iruskieta, Philippe Muller, Chloé Braud, and Sonia Badene. 2021. The DISRPT 2021 Shared Task on Elementary Discourse Unit Segmentation, Connective Detection, and Relation Classification. In *Proceedings of the 2nd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2021)*, pages 1–12, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Deniz Zeyrek, Amália Mendes, and Murathan Kurfalı. 2018. Multilingual Extension of PDTB-Style Annotation: The Case of TED Multilingual Discourse Bank. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Yuping Zhou, Jill Lu, Jennifer Zhang, and Nianwen Xue. 2014. Chinese Discourse Treebank 0.5 LDC2014T21.