Generalizing across Languages and Domains for Discourse Relation Classification

Peter Bourgonje and Vera Demberg
Language Science and Technology
Saarland University
peterb@lst.uni-saarland.de,vera@coli.uni-saarland.de

Abstract

The availability of corpora annotated for discourse relations is limited and discourse relation classification performance varies greatly depending on both language and domain. This is a problem for downstream applications that are intended for a language (i.e., not English) or a domain (i.e., not financial news) with comparatively low coverage for discourse annotations. In this paper, we experiment with a state-of-the-art model for discourse relation classification, originally developed for English, extend it to a multi-lingual setting (testing on Italian, Portuguese and Turkish), and employ a simple, yet effective method to mark out-of-domain training instances. By doing so, we aim to contribute to better generalization and more robust discourse relation classification performance across both language and domain.

1 Introduction

Interpreting discourse relations is an essential part of understanding a text, and has been shown to be beneficial for many down-stream tasks such as argument mining (Kirschner et al., 2015), summarization (Xu et al., 2020; Dong et al., 2021) and relation extraction (Tang et al., 2021). However, it is one of the tasks that are not easily solved by modern prompting methods that obliviate the need for training data (Chan et al., 2024; Yung et al., 2024): To date, achieving high performance still relies on high quality annotated data for training (or fine-tuning) a model. However, such discourse-annotated data is scarce and expensive to obtain. While relatively large resources exist for English newspaper texts, only small datasets (if any) are available for other languages. Additionally, recent work has shown that discourse classification performance can also be severely degraded by moving to a different domain (Gessler et al., 2021; Liu et al., 2023; Metheniti et al., 2023). In this paper, we work with the Penn Discourse Treebank (PDTB) framework (Prasad et al., 2008) and aim to systematically explore ways in which existing English data sources can be leveraged to obtain performant models in other languages and domains.

We test two different scenarios: The first one is a setting where no discourse-annotated data is available in a language at all. In this setting, one can either translate the PDTB into that language and project the labels onto the translated text, and then treat the resulting data as a dataset for that new language and fine-tune a model on it. Alternatively, one could employ a *multi-lingual* transformer (Conneau et al., 2019; Xue et al., 2021; Wei et al., 2023) that is trained on English discourse relation classification, and then simply use that model to label data in another language. In this paper we compare both of these settings.

The second scenario is one where at least a small amount of discourse-labelled data is available in the target language. For this setting, we investigate the benefit of augmenting the small original corpus with English data (or translated data with projected annotations). We specifically consider the common situation where the test data in the other language is also in a different domain than the English PDTB data from which we aim to leverage annotation.

We experiment with three other languages: Italian, Portuguese and Turkish. The selection is motivated by the fact that PDTB-annotated resources are available for these languages, thus allowing us to evaluate performance on originally annotated data and to train on target language annotated data for our low-resource setting.

Our best-performing set-up improves over state-of-the-art results for all three languages. In an attempt to better understand the characteristics of the different languages and domains represented in our data, we analyze the relation distribution and compare corpus sentence similarities.

Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 554–565
September 18–20, 2024.

2 Background and Related Work

Related work on discourse relation classification typically divides into papers focusing on explicit relations only (Pitler and Nenkova, 2009) or end-to-end approaches to discourse parsing (Lin et al., 2014; Oepen et al., 2016; Bourgonje, 2021; Knaebel, 2021) on the one hand, and implicit relations only on the other (Liu et al., 2016; Kishimoto et al., 2018; Shi and Demberg, 2019; Liu et al., 2020). Because of the relatively strong cues that explicit relations come with (in the form of discourse connectives), the former typically focuses on feature-engineering or makes use of lexical resources, whereas the latter focuses on neural approaches and methods based on contextualized embeddings. See Section 3 for more information on the PDTB and its relation types.

In this paper, we adopt the state-of-the-art implicit discourse relation classifier from Jiang et al. (2023). Because we want to evaluate this in a multilingual and multi-domain setting, we chose the corpora featured in Braud et al. (2023). In their shared task however, the relation type for PDTB corpora is not explicitly marked in the training and evaluation data, and the aim is to classify the relation between two arguments, regardless of the relation type (implicit, explicit or any other type). This means that we apply a classifier originally intended for implicit relations, to explicit (and other types of) relations as well.

With regard to zero-shot transfer learning for discourse relation classification, our work is inspired by Kurfalı and Östling (2019), who experiment with zero-shot transfer learning for implicit relation classification, by taking the model (intended for English and Chinese) from Rutherford and Xue (2016), training it on English and pooling data from different languages, and subsequently testing it on six other languages (German, Lithuanian, Polish, Portuguese, Russian and Turkish). Since 2019, LLMs with multi-lingual capabilities have become increasingly available, and we follow up on the work of Kurfalı and Östling (2019) by investigating the potential of XLM-RoBERTa-base (Conneau et al., 2019) for generalisation across languages for discourse relation classification.

By using corpora featured in the DISRPT shared task series (Zeldes et al., 2019, 2021; Braud et al., 2023), we can directly compare our results to the winning systems for the respective corpora. The submissions to the latest iteration include HITS by Liu et al. (2023), who use XLM-RoBERTa (base and large, depending on the training data size), DiscRet by Metheniti et al. (2023), who use the multi-lingual BERT base cased model (mBERT) (Devlin et al., 2019), and DiscoFlan by Anuranjana (2023), who uses Flan-T5 (Chung et al., 2022). For one of our test corpora (the Turkish Discourse Treebank), the system from Gessler et al. (2021), submitted in 2021, was not beat in 2023. Gessler et al. (2021) use a transformer-based neural classifier (a language-specific BERT model (Devlin et al., 2019)) which enhances contextualized word embeddings with hand-crafted features.

Jiang et al. (2023) improved the prior state-of-the-art in implicit relation classification for English; the Penn Discourse TreeBank version 2 and 3 (Prasad et al., 2008; Webber et al., 2019), by incorporating the hierarchical structure containing all senses, and the hierarchical sense label sequence corresponding to each instance during classification. We adopt their system architecture, and exchange RoBERTa-base for XLM-RoBERTa-base in most of our experiments.

3 Data

Penn Discourse TreeBank In our experiments, we use corpora that are annotated following the Penn Discourse TreeBank (PDTB) framework (Prasad et al., 2008; Webber et al., 2019). The PDTB comprises annotations over Wall Street Journal articles, thus represents the (financial) news domain. The PDTB paradigm, also referred to as *shallow* discourse parsing, differentiates first and foremost between different relation types, of which *explicit* and *implicit* relations are the most common.[1] The former are explicitly and lexically marked (with words or phrases like *however*, *as a result of*, *until*, also referred to as *discourse connectives*), the latter rely on the semantics of the related propositions in order to infer the relation. Relations are annotated between exactly two arguments, referred to as *arg1* and *arg2*.

Our goal is to classify relations between (pre-segmented) arguments according to the PDTB relation sense hierarchy, which first categorizes relations into four top levels (*Comparison*, *Contingency*, *Expansion* and *Temporal*), and further categorizes them into more detailed second-level senses. Although the PDTB sense hierarchy

[1] See Prasad et al. (2008, pp. 2963) and Webber et al. (2019, pp. 9) for details and other relation types.

actually specifies more unique second-level senses[2], most previous work renders second-level classification an 11-way classification problem, as that is the number of unique senses in the corresponding annotated corpora. The PDTB sense hierarchy includes a third level, but like most related work, we report classification performance on the first and second levels only. We follow the approach of Braud et al. (2023), by adopting their train/test split and predicting one label for one input sequence (we refer to Kim et al. (2020) for a detailed discussion on evaluation). The scores for 4-way accuracy and f_1 in Section 5 thus correspond to classification at the top level of the hierarchy, and the scores for 11-way accuracy and f_1 correspond to distinguishing between the eleven most frequent classes at the second level of the hierarchy.

As mentioned in Section 2, while Jiang et al. (2023) focus on implicit relations only, Braud et al. (2023) combine all relation types. To maximize comparability to related work, we thus follow Jiang et al. (2023) in using implicit relations only when data from the PDTB is concerned, and combine different relation types when the Italian, Portuguese and Turkish corpora coming from Braud et al. (2023) are concerned. We use the pre-processing script[3] from Jiang et al. (2023) to format the original PDTB data. This results in a train, dev and test split, all with implicit relations only (14,751 in total, see Table 1). For the Italian, Portuguese and Turkish corpora, we use the train, dev and test splits from Braud et al. (2023).

Translated discourse data To augment the training data available for other languages (whose corpora are much smaller than the PDTB), we translate the training, development and test sets of the English PDTB into Italian, Portuguese and Turkish using the Google Translate API[4]. In the following sections, the train, dev and test split from Jiang et al. (2023) are referred to as pdtb2, while their machine-translated versions are referred to as pdtb-it, pdtb-pt and pdtb-tr.

Italian discourse corpus: LUNA The LUNA corpus contains "Italian spontaneous speech recorded in the help-desk facility of the Consortium for Information Systems of Piedmont Regio" (Tonelli et al., 2010, pp.2084), thus represents Italian, and transcribed (but originally spoken) IT helpdesk dialogs. LUNA contains 1,188 relation instances in total (train, dev and test).

Portuguese discourse corpus: CRPC The CRPC corpus from Mendes and Lejeune (2022) contains a written subset of the Reference Corpus of Contemporary Portuguese, which in turn aims to serve as a representative sample for the Portuguese language and contains texts from many sources (literature, newspapers, magazines, science, economics, law, parliamentary debates, technical and didactic texts, pamphlets) (Généreux et al., 2012, pp.2237). CRPC contains 6,274 relation instances in total.

Turkish discourse corpus: TDB The Turkish Discourse Bank (TDB) corpus (Zeyrek and Kurfalı, 2017) contains written Turkish texts from a variety of genres (novels, stories, research surveys, travel and news articles, interviews and memoirs). TDB contains 1,809 relation instances in total.

The combination of authentic and synthetic (i.e., machine-translated) corpora enables us to experiment with different set-ups, using in-domain, out-of-domain, in-language and out-of-language configurations for training and test sets, to see how well the model generalizes across the different dimensions. Statistics of our data sets are included in Table 1. Since pdtb-it, pdtb-pt and pdtb-tr are direct translations of the relations in our pdtb2 corpus, the number of instances in those data sets are identical to the pdtb2. One of the key goals of this paper is to find out how our relation classifier generalizes across both languages and domains. While "language" is comparatively well-defined (with Turkish being a different language than Portuguese, for example), the notion of "domain" is less clear-cut. The LUNA corpus stands out in that it represents spontaneous speech in help-desk context, but both CRPC and TDB are multi-genre, include news texts and therefore could be considered not that different from the (financial news) PDTB texts. In our experiments though, we assume each of the three non-English corpora to be of a different domain than the PDTB, and get back to the discussion of domain differences in Section 6.

[2] 16 in the 2.0 version of the hierarchy (Prasad et al., 2008).

[3] https://github.com/YJiangcm/GOLF_for_IDRR/blob/master/preprocess.py

[4] Translations were obtained on February 28, 2024.

	pdtb2	LUNA	CRPC	TDB
train	12,547	728	4,869	1,348
dev	1,165	168	769	193
test	1,039	292	636	268
total	14,751	1,188	6,274	1,809

Table 1: Data statistics.

4 Method

Discourse relation classification model Our model is based on Jiang et al. (2023), who improve over prior work on implicit discourse relation classification by proposing a hierarchy-aware architecture, that takes into the account the global and local level of PDTB relation senses. We use the default hyper-parameter settings of Jiang et al. (2023), except for the number of epochs, which we set to 30. We use XLM-RoBERTa-base (Conneau et al., 2019) for most configurations, since we want to investigate the potential for generalisation across languages. For experiments where training and test data is from the same language, we use `roberta-base` (Liu et al., 2019) for English, `roberta-base-italian`[5] for Italian, `portuguese-roberta-base`[6] for Portuguese, and `roberta-base-turkish-uncased` (Aytan and Sakar, 2022) for Turkish.

State-of-the-art models for Italian, Portuguese and Turkish We compare performance of our set-up on the Italian and Portuguese corpora to that of Liu et al. (2023), and on the Turkish corpus to that of Gessler et al. (2021) (see Section 2 for details). Recall that while Jiang et al. (2023) work with implicit relations only from the PDTB, because the data featured in the 2021 and 2023 shared tasks (Zeldes et al., 2021; Braud et al., 2023) combines all relation types, for LUNA, CRPC and TDB, we train and evaluate on both implicits, explicits and other relation types.

Domain adaptation In addition to trying out different data configurations (of training and test data), we experiment with marking out-of-domain training samples at training time. This is inspired by Daumé III (2007); Kim et al. (2016), who augment the feature space that is used as input to the classifier model, thereby forcing the learning algorithm to do the adaptation. In their implementation, the dimension *domain* simply occupies a particular position in the vector representation of the input to the classifier. Similarly, we simply concatenate the final representation with a binary flag, indicating if the training sample is in-domain or out-of-domain. In the original model architecture, the vectorized representations of *arg1* and *arg2* are concatenated and used as input for the classifier. In our experiments with marking of out-of-domain data, we combine this concatenated vector with another vector of zeros if the sample is out-of-domain, and with another vector of ones if the sample is in-domain.

5 Results

The following subsections present the results for different base models and different configurations of training and test data.

5.1 Mono-lingual vs. Multi-lingual Model

We first want to test how much model performance degrades by switching to a multi-lingual instead of a mono-lingual base model. We therefore reran the original model from Jiang et al. (2023), and compare it to a version in which we replace the mono-lingual English RoBERTa-base model by the multi-lingual XLM-RoBERTa-base model. Recall that 4-way and 11-way results correspond to classification on the top and second level, respectively, of the PDTB sense hierarchy. Table 2 shows that our replication of Jiang et al. (2023) yielded slightly lower (but roughly comparable) results, but that we see a sharp drop in performance: 10 points in both accuracy and f_1[7] when exchanging the English model for the multi-lingual one. When working with English data, using a mono-lingual English model thus yields better results.

model	4-way f_1(acc.)	11-way f_1(acc.)
JZW23-orig	65.76 (72.52)	41.74 (61.16)
JZW23-reprod	64.07 (71.61)	39.63 (60.35)
XLM-R-base	54.57 (62.95)	30.61 (48.99)

Table 2: Results for a mono-lingual and multi-lingual base model on English data (pdtb2). JZW23 stands for Jiang et al. (2023); orig refers to reported numbers in their table 1; reprod refers to our results from running their code; XLM-R-base stands for the XLM-RoBERTa-base model.

[5] https://huggingface.co/osiria/roberta-base-italian

[6] https://huggingface.co/flax-community/portuguese-roberta-base

[7] All f_1-scores in this paper are macro-averaged.

5.2 Language Transfer (Zero-resource setting)

Next, we consider a setting in which no data is available in the target language and compare how well the mono-lingual model using translated English corpus data for training does compared to a setting where a multi-lingual model is trained on the English corpus and then applied to Italian/Portuguese/Turkish data (pdtb-it, pdtb-pt, pdtb-tr). We test both on the pdtb2 test set translated into the target language as well as on the test set of data that was originally annotated in the target language. Note that in the latter case, the model has to deal with both a language transfer problem and with a domain-adaptation problem, as the original corpora contain data from different domains than the English PDTB corpus. Our experiments reported in this section use the multi-lingual XLM-RoBERTa-base model; we will get back to a comparison to mono-lingual models for Italian, Portuguese and Turkish in Section 5.4 below.

Table 3 illustrates that performance on the translated pdtb2 dataset to Italian, Portuguese and Turkish remains relatively stable compared to the performance of the multi-lingual model on English (compare f_1 and accuracy scores to the last row in Table 2). This suggests that it is possible to learn discourse relations independently of language, and apply these learned representations to another language, which has not been seen during task-specific fine-tuning.

Furthermore, we can see that performance is slightly better when fine-tuning the multi-lingual model on the translated pdtb2 data compared to training it on English and then applying to the target language (compare the first two rows for each language in Table 3).

Finally, we can also observe that there is a substantial drop in performance when evaluating on the test set of the original Italian (LUNA) / Portuguese (CRPC) / Turkish (TDB) data. There might be several reasons for this: The discourse-annotated texts from the other languages are from different domains – hence, the approach not only has to generalize across languages but also across domains, a well-known notoriously difficult problem. Alternatively, it is possible that the translated data is atypical, suffering from translationese effects and thereby might hamper generalization from translated data to native data. Another factor is type of arguments the model has seen during training. Since we train on implicit relations from the PDTB2, the model has only seen examples of inter-sentential relations. In addition to the implicit vs. explicit distinction, it is also confronted with intra-sentential (explicit) relations in the test set-up. Finally, it is also possible that Italian / Portuguese / Turkish annotators took different decisions in discourse annotation compared to English annotators on PDTB, leading to a discrepancy in label usage, e.g., by using a smaller set of labels.

train	test	4-way f_1 (acc.)	11-way f_1 (acc.)
pdtb2	pdtb-it	54.79 (64.10)	31.85 (49.66)
pdtb-it	pdtb-it	56.72 (64.39)	33.63 (49.23)
pdtb-it	LUNA	43.06 (48.29)	18.03 (34.59)
pdtb2	pdtb-pt	53.24 (62.95)	31.17 (48.80)
pdtb-pt	pdtb-pt	55.03 (63.75)	31.61 (47.83)
pdtb-pt	CRPC	45.74 (57.08)	17.03 (37.26)
pdtb2	pdtb-tr	51.24 (61.50)	30.41 (46.92)
pdtb-tr	pdtb-tr	51.57 (60.73)	30.31 (45.91)
pdtb-tr	TDB	43.15 (47.01)	18.11 (32.84)

Table 3: Results for the XLM-R-base model on language transfer, testing on synthetic, translated data as well as on originally annotated data in the target language.

5.3 Domain Transfer (Low Resource Setting)

Next, we consider a setting where some target language discourse-annotated data is available for training. Our main questions are (a) how our basic setup based on the XLM-RoBERTa-base model compares to the previous state-of-the-art on the Italian, Portuguese and Turkish datasets; (b) whether performance can be improved by exploiting translated data from English; (c) whether our implementation of a domain-adaptation technique inspired by Daumé III (2007); Kim et al. (2016) helps in dealing with the domain gap between translated pdtb2 data and the target domain.

Regarding our first question, we compare our results to the best-performing systems of the 2021 and 2023 shared task iterations (Zeldes et al., 2021; Braud et al., 2023). The results are shown in Table 4. For LUNA, Liu et al. (2023) outperform our baseline setup, while for CPRC and TDB, our baseline outperforms the results of Liu et al. (2023) and Gessler et al. (2021), respectively.

Regarding our section question, we explore training on both the translated pdtb2 data and the train-

model	4-way f_1 (acc.)	11-way f_1 (acc.)
LFS23	- (65.00)	- (-)
XLM-R on LUNA	64.31 (63.70)	32.81 (52.74)
+ pdtb-it	67.64 (66.78)	41.57 (57.53)
+ pdtb-it + DA	**72.72 (71.92)**	**57.13 (62.33)**
LFS23	- (78.53)	- (-)
XLM-R on CPRC	78.71 (83.18)	76.39 (82.55)
+ pdtb-pt	**79.86 (83.96)**	73.45 (83.02)
+ pdtb-pt + DA	**79.86** (83.49)	**77.90 (83.65)**
GBLPZZ21	- (60.09)	- (-)
XLM-R on TDB	61.80 (64.55)	51.11 (59.33)
+ pdtb-tr	64.45 (68.66)	40.05 (61.57)
+ pdtb-tr + DA	63.98 (67.54)	53.92 (64.18)

Table 4: Baseline performance on Italian (LUNA), Portuguese (CRPC) and Turkish (TDB), compared to prior work; LFS23 stands for (Liu et al., 2023); GBLPZZ21 stands for (Gessler et al., 2021); DA stands for domain adaptation (adding a flag that indicates what domain each data point comes from).

ing data from the target domain, in a setup where the model is first trained on the translated pdtb2 data for 15 epochs, and then on the training section of the target-language original data for 15 more epochs. Our results (see the magenta rows in Table 4) show that using translated English data from the financial news domain as additional training data increases performance consistently for all three languages in the 4-way (top-level) classification task; however, we also observe a drop in performance on the 11-way (second-level) classification task. A more detailed analysis indicates that this might be due to different distributions of second-level labels between the corpora (we will get back to this in Section 6), hence second-level labels suffer more severely from the domain shift between PDTB and other domains.

Simple domain adaptation Finally, regarding our third question, table 4 also presents the results for explicitly marking the out-of-domain data (here: the translated pdtb2 texts) at training time. The rows marked "+DA" represent configurations where the training data is from multiple domains, but from the same language. We find that domain marking leads to improved performance in most settings, with strongest improvements obtained on the 11-way classification problems. This indicates that the distribution shift regarding second-level labels can be modelled successfully by including the domain flag. We note that our proposed method including translated English data and the simple domain adaptation technique outperform the previous state-of-the-art results consistently and by a substantial margin on all three languages. It should be noted here though that we use additional training data which was not available in the shared task, and that for a direct comparison, the winning system of the shared task should be trained with this additional data as well.

We also tested a configuration where the multilingual model is first fine-tuned on English pdtb2 data for 15 epochs (without translating that data, but in a setting that does use the domain adaptation flag), and then further fine-tuned on the target language training data. We found that this setting leads to worse results than using translated data for Italian (3 point drop) and Portuguese (1 point drop), whereas for Turkish, better results are obtained when using original, English pdtb2 data, combined with Turkish in-domain data (f_1 66.55, acc 69.78 on 4-way classification, and f_1 55.00, acc 64.93 on 11-way classification).

5.4 Multi-lingual vs. Mono-lingual Target Language Models

Because our experiments on English with a monolingual vs. a multi-lingual base model indicated a significant drop in performance moving from a mono-lingual to a multi-lingual model (see Table 2), we also used dedicated mono-lingual models (see Section 4), with translated data (pdtb-it, pdtb-pt, pdtb-tr) in combination with out-of-domain marking in an attempt to further improve performance. We found, however, that unlike the English setting, this did not improve performance, compared to using the multi-lingual model. For LUNA, 4-way f_1 and accuracy dropped from 72.72, 71.92 to 70.24, 69.86, respectively. 11-way f_1 dropped from 57.13 to 50.14, with accuracy staying at 62.33. For CRPC, 4-way f_1 and accuracy dropped significantly from 79.86 and 83.96 to 53.68 and 63.68, respectively. 11-way f_1 and accuracy dropped significantly as well, from 77.90, 83.65 to 40.15, 62.58. For TDB, the performance drop was equally significant. 4-way f_1 and accuracy dropped from 66.55, 69.78 to 55.02, 59.70. 11-way f_1 and accuracy dropped from 55.00, 64.92 to 37.96, 54.48.

6 Discussion

Overall, we obtain the best results by combining data from different domains, and marking the out-of-domain instances at training time. For Italian and Portuguese, using in-language training data yields better results, whereas for Turkish, combining English out-of-domain data with Turkish in-domain data yields better results. According to Conneau et al. (2019, Appendix A), the training data for XLM-RoBERTa-base included 20.9 GiB of Turkish, compared to 30.2 GiB for Italian, 49.1 GiB for Portuguese, and 300.8 GiB for English. This could explain the better performance when using original, English data, since the model has seen comparatively few Turkish at pre-training. However, since the difference between the amount of training data in Turkish, Italian and Portuguese is not that large, we consider more research necessary to draw conclusions on this. For Italian and Portuguese, it seems that the automatically obtained, synthetic data is good enough to improve classification performance for discourse relation classification when testing on authentic data.

The performance difference for the three corpora overall are rather large, but Table 1 indicates that performance does not correspond to the size of the corpus. Although CRPC is the largest and has the highest scores overall, LUNA has higher scores than TDB, despite LUNA being about 1.5 times smaller than TDB.

Differences in label distributions Figure 1 displays the distribution of top-level senses for the four corpora used in our experiments.[8]

From this, we can see that LUNA has a more balanced distribution than the others, possibly explaining its comparatively high scores (for our best-performing set-up), taking into account that it is by far the smallest corpus. Note that pdtb2 has a fairly imbalanced distribution, with a large proportion of expansion relations and relatively few temporal relations. This distribution can be partially attributed to specificities of the newspaper genre, and partially to the fact that for the pdtb2 corpus, we are only working with implicit relations. Temporal relations are often expressed explicitly, which may contribute to their low rate in pdtb2. We also note that CRPC is most similar to pdtb2 in the context of the number of implicit relations in the dataset, as it has the highest implicit ratio of the three non-English corpora: Table 6 of Braud et al. (2023) shows a ratio of 0.58 (711 implicit relations, divided by 1,228 total relations), compared to 0.47 and 0.32 for TDB and LUNA, respectively.

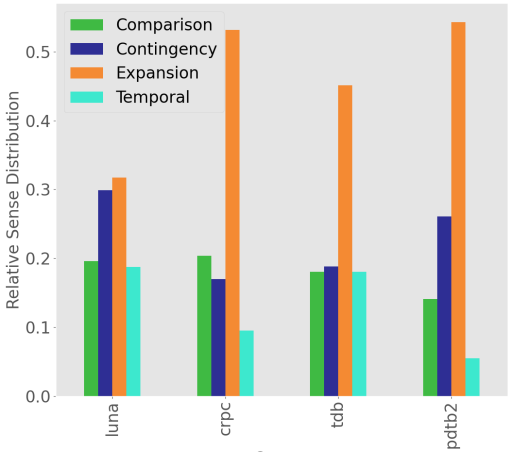

Figure 1: Top-level sense distributions.

Turning to second-level senses in Figure 2, we see that CRPC and TDB have slightly fewer unique senses (6 for both) than the other two (9 for LUNA, 11 for pdtb2). While having fewer classes often results in higher scores for multi-class classification, this does not seem to be very predictive for our task, as CRPC has a relatively high f_1-score (77.90), but TDB (with only 6 unique senses) scores 53.92, while LUNA scores 57.13 (with 9 unique senses). In this respect, it is important to point out that there might be corpus-specific biases: For all three systems submitted to the 2023 DISRPT shared task, CRPC relation classification performs significantly above the corresponding mean of the system, and two of the three systems shows second-best results on this corpus (after the Thai corpus) (Braud et al., 2023, Table 5). This might indicate that the CRPC corpus contains particularly *easy* relations, and we consider an investigation of what *easy* means in this context and important direction for future work.

Our results also showed that including pdtb2 data was detrimental to performance for 11-way classification when no domain flagging is used, and we speculated that this could be due to strong differences in the distributions of second-level senses. In Figure 2, we can indeed observe that in the expansion class, pdtb2 has many more instantiation and restatement relations, and fewer conjunctions than the other corpora.

[8] Recall that pdtb-it, pdtb-pt and pdtb-tr have the exact same distributions as the pdtb2.

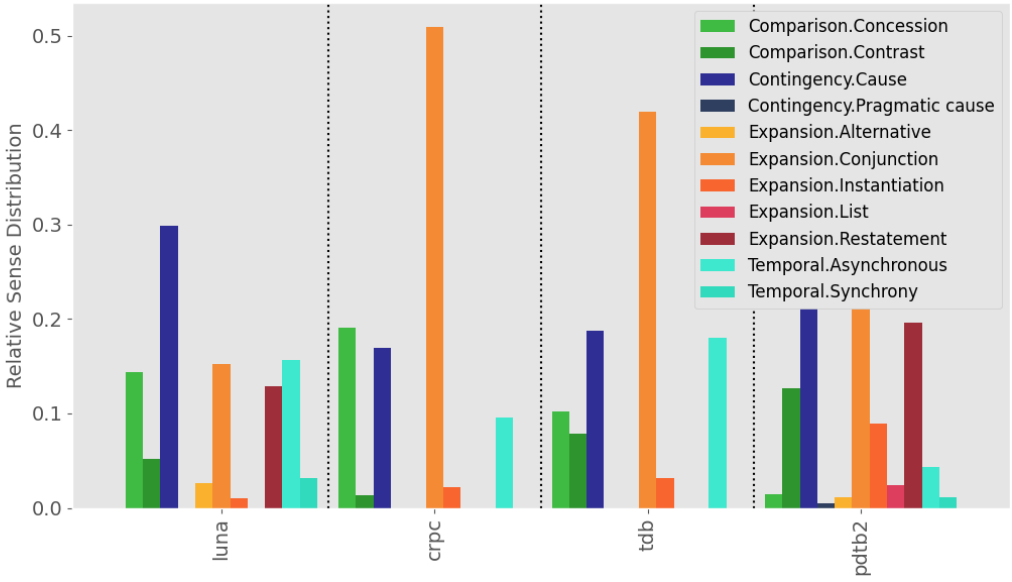

Figure 2: Second-level sense distributions.

Domain differences between corpora In an attempt to assess to what extent the corpora used in our experiments differ with respect to the actual words and phrases used, we include Figure 3. This is the result of a pair-wise comparison of all corpora, where the number expresses the average cosine similarity between all relational arguments in the corpora. To calculate cosine similarity, we encode the arguments using `stsb-xlm-r-multilingual` from sentence-transformers (Reimers and Gurevych, 2019). The sentence-transformers architecture is specifically designed to express semantic similarity at sentence level, and by using a multi-lingual model, the assumption is that a particular sentence in English shows a high degree of similarity with the (maximally faithful) translation of that sentence in, for example, Turkish. The numbers on the diagonal in Figure 3 express how diverse a corpus is: The high number for LUNA hence indicates that there is relatively little diversity in the LUNA corpus (0.34) compared to e.g., the pdtb2 corpus. The low number for LUNA vs. pdtb2 (0.092) in Figure 3 indicates that the relational arguments in LUNA tend to be very different from the relational arguments of pdtb2. From this figure, we can read that both LUNA and TDB stand out in their usage of words and phrases, as they display a higher average cosine similarity when compared intra-corpus than when compared inter-corpus. While we indeed see a significant drop when training on pdtb2 and testing on LUNA and TDB, the same drop is observed when training on pdtb2 and testing on CRPC, although pdtb2 and CRPC display a considerably lower divergence compared to pdtb2 and LUNA, and pdtb2 and TDB.

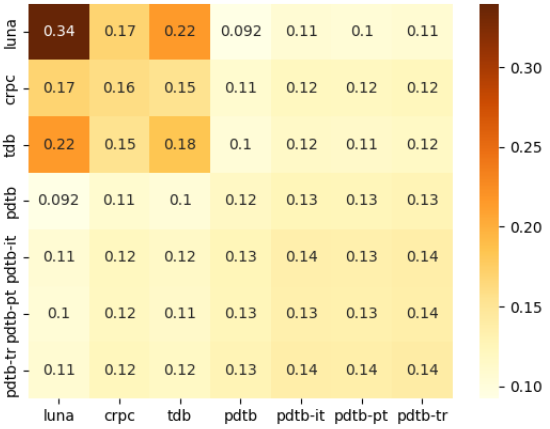

Figure 3: Corpus similarity heatmap.

Another possible explanation of performance could be the single- or multi-domain aspect of an evaluation corpus. We compared the performance of the winning system of the 2023 shared task[9] (Liu et al., 2023) along this axis, and this reveals that the average performance on single-domain cor-

[9]We use this, and not our system, to have more data points, as we do not have results for the Thai and Chinese corpora.

pora (65.00 for LUNA, 74.30 for the PDTB, 64.96 for TEDM, 95.83 for TDTB and 59.63 for CDTB (Braud et al., 2023, Table 5)) is higher than on multi-domain corpora (78.53 for CRPC and 45.50 for TDB): 71.95 vs. 62.02 for single- vs. multi-domain, respectively. This could be because a single-domain corpus is likely to be more consistent in terms of the types of discourse relations that occur in it. This observation, however, is only based on two data points for multi-domain corpora, and while an interesting direction, we consider more data points necessary before such conclusions can be drawn.

7 Conclusion

In this paper, we adopt a state-of-the-art implicit discourse relation classification model developed for English, and apply it to both implicit and non-implicit discourse relations from three corpora that differ in language and domain: An Italian corpus of transcribed IT Helpdesk dialogs, a multi-domain Portuguese corpus and a multi-domain Turkish corpus. By experimenting with different configurations of in-domain, out-of-domain, in-language and out-of-language training data, we explore to what extent the model generalizes across languages and domains. We also demonstrate the importance of using a flag to mark out-of-domain data at training time. Overall, our setup improves over prior work by just under 7 points for Italian, over 5 points for Portuguese, and over 9 points for Turkish (all based on 4-way classification accuracy scores). Our code is published on GitHub.[10]

We attempt to link the classification results to the number of training samples, label distribution and language usage. We find that the number of training samples or sentence similarity between training and test domain is not very indicative of performance, and that instead the label distribution is likely to be a more reliable indicator. In future work, we plan to delve deeper into specific label distributions of the different domains, and potentially continue this line of work by not just looking at different domains, but by also including and testing on annotated data using wholly different label sets (e.g., Rhetorical Structure Theory (Mann and Thompson, 1988) corpora).

In order to maximize comparability to related work, we tested on implicit relations only in pdtb set-ups, while we combine different relation types in the other set-ups. An interesting direction of future work would be to more systematically investigate the significance of the strict distinction between explicit and implicit relations (as it is often found in the literature), given current, state-of-the-art models for discourse relation classification.

Limitations

Our experiments rely on fine-tuning on LLMs, and benefit greatly from running on a GPU. Reproduction of our results without having access to a GPU will therefore be time-consuming. Furthermore, although XLM-RoBERTa is specifically targeted at multi-lingual use cases, the amount of training data varies per language (Conneau et al., 2019, Appendix A). For languages with relatively few GiBs of training data, performance may be significantly lower than for the languages we included in our evaluation.

Ethics Statement

Since our method relies on XLM-RoBERTa for the encoding of input, it will propagate any biases present in (the training data of) this pre-trained language model.

Acknowledgements

We thank the three anonymous reviewers for their insightful comments. This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 232722074 – SFB 1102.

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.

Burak Aytan and C Okan Sakar. 2022. Comparison of transformer-based models trained in turkish and different languages on turkish natural language processing problems. In *2022 30th Signal Processing and Communications Applications Conference (SIU)*, pages 1–4. IEEE.

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

[10] https://github.com/PeterBourgonje/GOLF_multilingual

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.

Chunkit Chan, Cheng Jiayang, Weiqi Wang, Yuxin Jiang, Tianqing Fang, Xin Liu, and Yangqiu Song. 2024. Exploring the potential of ChatGPT on sentence level relations: A focus on temporal, causal, and discourse relations. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 684–721, St. Julian's, Malta. Association for Computational Linguistics.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *Preprint*, arXiv:2210.11416.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Hal Daumé III. 2007. Frustratingly easy domain adaptation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 256–263, Prague, Czech Republic. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Yue Dong, Andrei Mircea, and Jackie Chi Kit Cheung. 2021. Discourse-aware unsupervised summarization for long scientific documents. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1089–1102, Online. Association for Computational Linguistics.

Michel Généreux, Iris Hendrickx, and Amália Mendes. 2012. Introducing the reference corpus of contemporary Portuguese online. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2237–2244, Istanbul, Turkey. European Language Resources Association (ELRA).

Luke Gessler, Shabnam Behzad, Yang Janet Liu, Siyao Peng, Yilun Zhu, and Amir Zeldes. 2021. DisCoDisCo at the DISRPT2021 shared task: A system for discourse segmentation, classification, and connective detection. In *Proceedings of the 2nd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2021)*, pages 51–62, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yuxin Jiang, Linhan Zhang, and Wei Wang. 2023. Global and local hierarchy-aware contrastive framework for implicit discourse relation recognition. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8048–8064, Toronto, Canada. Association for Computational Linguistics.

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, Online. Association for Computational Linguistics.

Young-Bum Kim, Karl Stratos, and Ruhi Sarikaya. 2016. Frustratingly easy neural domain adaptation. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 387–396, Osaka, Japan. The COLING 2016 Organizing Committee.

Christian Kirschner, Judith Eckle-Kohler, and Iryna Gurevych. 2015. Linking the thoughts: Analysis of argumentation structures in scientific publications. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 1–11, Denver, CO. Association for Computational Linguistics.

Yudai Kishimoto, Yugo Murawaki, and Sadao Kurohashi. 2018. A knowledge-augmented neural network model for implicit discourse relation classification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 584–595, Santa Fe, New Mexico, USA. 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.

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.

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.

Xin Liu, Jiefu Ou, Yangqiu Song, and Xin Jiang. 2020. On the importance of word and sentence representation learning in implicit discourse relation classification. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 3830–3836.

Yang Liu, Sujian Li, Xiaodong Zhang, and Zhifang Sui. 2016. Implicit discourse relation classification via multi-task neural networks. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, page 2750–2756. AAAI Press.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

William Mann and Sandra Thompson. 1988. Rhetorical Structure Theory: Towards a functional theory of text organization. *TEXT*, 8:243–281.

Amália Mendes and Pierre Lejeune. 2022. Crpc-db a discourse bank for portuguese. In *Computational Processing of the Portuguese Language: 15th International Conference, PROPOR 2022, Fortaleza, Brazil, March 21–23, 2022, Proceedings*, page 79–89, Berlin, Heidelberg. Springer-Verlag.

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.

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.

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

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Attapol Rutherford and Nianwen Xue. 2016. Robust non-explicit neural discourse parser in English and Chinese. In *Proceedings of the CoNLL-16 shared task*, pages 55–59, Berlin, Germany. Association for Computational Linguistics.

Wei Shi and Vera Demberg. 2019. Next sentence prediction helps implicit discourse relation classification within and across domains. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5790–5796, Hong Kong, China. Association for Computational Linguistics.

Jialong Tang, Hongyu Lin, Meng Liao, Yaojie Lu, Xianpei Han, Le Sun, Weijian Xie, and Jin Xu. 2021. From discourse to narrative: Knowledge projection for event relation extraction. 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 732–742, Online. Association for Computational Linguistics.

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).

Bonnie Webber, Rashmi Prasad, Alan Lee, and Aravind Joshi. 2019. The Penn Discourse Treebank 3.0 Annotation Manual.

Xiangpeng Wei, Haoran Wei, Huan Lin, Tianhao Li, Pei Zhang, Xingzhang Ren, Mei Li, Yu Wan, Zhiwei Cao, Binbin Xie, Tianxiang Hu, Shangjie Li, Binyuan Hui, Bowen Yu, Dayiheng Liu, Baosong Yang, Fei Huang, and Jun Xie. 2023. Polylm: An open source polyglot large language model. *Preprint*, arXiv:2307.06018.

Jiacheng Xu, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Discourse-aware neural extractive text summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5021–5031, Online. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Frances Yung, Mansoor Ahmad, Merel Scholman, and Vera Demberg. 2024. Prompting implicit discourse relation annotation. In *Proceedings of The 18th Linguistic Annotation Workshop (LAW-XVIII)*, pages 150–165, St. Julians, Malta. 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 and Murathan Kurfalı. 2017. TDB 1.1: Extensions on Turkish discourse bank. In *Proceedings of the 11th Linguistic Annotation Workshop*, pages 76–81, Valencia, Spain. Association for Computational Linguistics.