Zero-shot learning for multilingual discourse relation classification

Eleni Metheniti, Philippe Muller, Chloé Braud, Margarita Hernández-Casas

Zero-shot learning for multilingual discourse relation classification

[1]Eleni Metheniti [1,3]Philippe Muller [1,2,3]Chloé Braud

[4]Margarita Hernández-Casas

[1]UT3 - IRIT ; [2]CNRS ; [3]ANITI ; [4]Accenture
firstname.lastname@irit.fr

Abstract

Classifying discourse relations is a hard task: discourse-annotated data is scarce, especially for languages other than English, and there exist different theoretical frameworks that affect textual spans to be linked and the label set used. Thus, work on transfer between languages is very limited, especially between frameworks, while it could improve our understanding of some theoretical aspects and enhance many applications. In this paper, we propose the first experiments on zero-shot learning for discourse relation classification and investigate several paths in the way source data can be combined, either based on languages, frameworks, or similarity measures. We demonstrate how difficult transfer is for the task at hand, and that the most impactful factor is label set divergence, where the notion of underlying framework possibly conceals crucial disagreements.

1 Introduction

Discourse analysis examines the representation of information at a document level, by finding sentences or sentence segments that are logically and/or structurally connected. These connections are called *rhetorical relations* and may be *explicit* (with the presence of distinct words called *connectives*) or *implicit* (without distinct connectives):

1. [**Since** these statistics are encoded as dense continuous features,]$_1$ [it is not trivial to combine these features]$_2$
   cause(1,2), `eng.dep.scidtb`

2. [**Tras** obtener el soporte informático con la totalidad de los textos en ambos idiomas,]$_1$ [hemos procedido a confrontar y paralelizar las dos versiones,]$_2$
   After obtaining the computer support with all the texts in both languages, we proceeded to compare and parallelize the two versions,
   sequence(1,2), `spa.rst.sctb`

3. [Sözümü bitirmiştim.]$_1$ [Muammer'den bir su istedim.]$_2$
   I had finished my speech. I asked Muammer for a glass of water.
   temporal.asynchronous(1,2), `tur.pdtb.tdb`

Even though discourse analysis has been thoroughly studied and has brought improvements on many NLP downstream tasks (e.g. summarization (Xu et al., 2020), machine translation (Chen et al., 2020)) the domain suffers from several limitations: (1) most of the existing work focuses on specific data, namely a few corpora in English; (2) there are distinct theoretical frameworks, with different definitions of what discourse units are, and different choices of relation typology. Concerning multilinguality, the introduction of discourse-annotated corpora, in different frameworks, has stimulated work on multilingual discourse analysis, e.g. Braud et al. (2017); Liu et al. (2021) in the RST framework (Rhetorical Structure Theory, Mann and Thompson, 1988).

As an attempt to study discourse relations across languages and frameworks, the DISRPT Shared Task has hosted systems for discourse segmentation (the task of locating discourse units) and relation identification (labeling pairs of discourse units that are known to be related). DISRPT allows for the exploration of approaches that can be applied to more varied contexts with regard to languages, frameworks, and textual genres. While the motivation of the Shared Task is unification, the most successful systems are composed of monolingual models or small corpora subgroups with annotation homogeneity (Braud et al., 2023). There have been attempts to unify framework and corpora annotations, e.g. Benamara and Taboada (2015), but they have not been adopted in practice.

Discourse analysis should not, however, be limited to the currently available annotated datasets. It is important to examine how to efficiently transfer a system to languages with fewer or no resources.

This is a standard topic in different domains of NLP, with the development of cross-lingual benchmarks used to evaluate few-shot or zero-shot transfer, such as XNLI (Conneau et al., 2018), XQuad (Artetxe et al., 2020) or Wikiann (Pan et al., 2017).

Our work presents experiments on discourse relation classification across languages and formalisms, inspired by the related DISRPT Shared Task. Our motivation is to maintain a multilingual, multi-framework approach as closely as possible while exploring different contexts of zero-shot transfer. We observe the framework issues discussed above; the heterogeneity of annotations, annotation overlap, and theoretical definitions of a discourse unit, as well as the practical problem of varying corpora sizes. We train and evaluate jointly-trained multilingual multi-framework models, based on multilingual pretrained language models (of different sizes and transformer architectures), and compare them to monolingual approaches. We also create zero-shot models to test whether generalization to an unobserved language is possible, from training with the same language family, framework, or corpora with similar label spaces.

2 Previous Work

Discourse relation prediction can be divided into two tasks: (a) *shallow discourse parsing*, where relations occur in the same sentence, or between two neighboring sentences, and (b) *full discourse parsing*, where relations form a structure, usually a tree, covering a whole document. Work on shallow discourse parsing is performed with the Penn Discourse TreeBank framework (PDTB, Prasad et al., 2014) and focuses on relation classification, specifically on implicit ones (Example (3), Section 1), that is relations not triggered by a discourse connective, such as *since* or *tras/after* in Examples (1) and (2). Full discourse parsing consists of various kinds of structure predictions, plus labeling of the structure, with the use of other discourse frameworks (see Section 4.1).

Work on shallow discourse parsing has focused predominantly on English, from feature-based approaches (Pitler et al., 2009; Lin et al., 2009) to finetuning pretrained models in order to capture interactions between argument contextual embeddings (Liu et al., 2020; Wu et al., 2022). A popular approach is the use of connective prediction as an auxiliary task (Kishimoto et al., 2020; Wu et al., 2023; Liu and Strube, 2023). Recent work has also leveraged prompt tuning (Zhao et al., 2023). PDTB relations are defined with a hierarchy of subsenses with 3 levels, and the most recent work focuses on the finer-grain levels. The creation of corpora in other languages led to the CoNLL shared tasks (Xue et al., 2015, 2016) however limited to Mandarin and English. Most work assumes relation arguments are already known, and only the relation label is to be predicted.

For full discourse parsing (RST or SDRT frameworks), relation prediction is either done jointly with structure prediction, e.g. (Zhang et al., 2021; Yu et al., 2022) or as the last stage of processing (Wang et al., 2017). It is challenging to compare with PDTB approaches, since work on full parsing rarely evaluates the relation classification model independently. In addition, it is difficult to assess the contribution of relation prediction to the main parsing task. Finally, discourse parsing in a realistic setting should make no distinction between explicit or implicit relations, since all relations have to be labeled to form a covering structure. In our work, we also adhere to this setting.

Most approaches address English data, with only a few attempts to leverage joint, multilingual settings, and only on a subset of existing corpora. Regarding full discourse parsing with the RST framework, Braud et al. (2017) created a feature-based approach that is generalized to a set of languages to evaluate transfer abilities. Liu et al. (2020) equipped various translation strategies to train one model in a general dataset and produce predictions in different languages. These approaches rely on an extensive mapping of discourse relations to enable transfer and reduce label sets as much as possible; in our work, we opt for as few conversions as possible, only when needed (e.g. a unique label in one dataset).

The development of the DISRPT Shared Task was another step toward standardizing evaluations of discourse processing methodologies (Zeldes et al., 2021; Braud et al., 2023). The Shared Tasks provided a unified text format for multiple discourse annotation frameworks, and included a task on Discourse Relation Classification since the 2021 edition, for a variety of languages. In the first campaign, only two systems addressed this task, the most successful being DisCoDisCo (Gessler et al., 2021), with separate models for each language, built on finetuned monolingual pretrained models, enriched with handcrafted linguistic and

non-linguistic features. (Varachkina and Pannach, 2021) used stacked random forest classifiers, on top of sentence-level embeddings made with Sentence-BERT (Reimers and Gurevych, 2019), to predict coarse relations first and fine-grain relations in a second step.

Three systems competed in the 2023 edition of the discourse relation classification task, on an extension of the 2021 data. The best-performing system on this edition, HITS (Liu et al., 2023), used a combination of monolingual and multilingual framework-based finetuned classifiers, built mainly on large pretrained models.They also used adversarial training and bootstrap aggregating strategies to improve performance.

DiscReT (Metheniti et al., 2023) also used pretrained models (mBERT base cased, Devlin et al., 2019) and jointly trained all the corpora of the task. They also used adapters (Houlsby et al., 2019) trained on the same task and with frozen layers. This approach tried to reduce the large joint label space by creating reversible label mappings in cases of label overlap among frameworks.

Finally, DiscoFlan (Anuranjana, 2023) relied on the Flan-T5 generative language model (Chung et al., 2022) to generate relation labels, by querying with a prompt of the two arguments. Models were trained separately for each language, and the output was processed to match labels from each corpus label set, with a high variance between datasets.

Regarding **multilingual classification** tasks, one of the most recent and notable sources is the XNLI Dataset (Conneau et al., 2018), an evaluation corpus for language transfer and cross-lingual sentence classification in 15 languages, with 112.5k annotated pairs. This dataset has been used as a benchmark for downstream tasks such as natural language inference. Common approaches to NLI are multi-modal and are motivated by multilinguality; for example, performing machine translation between languages, using parallel corpora for enhancing the training set, or cross-lingual templates for enhancing the masked language modeling objective (Qi et al., 2022).

A recent approach on **zero-shot multilingual transfer** with a low-resource motivation has been proposed under the scope of the AmericasNLI dataset (Kann et al., 2022). For NLI, Ebrahimi et al. (2022) used multilingual pretrained models in a few-shot/zero-shot setting for low-resource languages, and proposed model adaptation via continued pretraining. They also observe that translation as a preprocessing step improves NLI results.

3 Methodology

3.1 Multilingual discourse relation classification across formalisms

As a take-off point for our experiments, we reprise the Discourse relation classification Shared Task, gathering inspiration from submitted systems. We are using multilingual transformer-based architectures for our experiments that have already been tested by the participating teams in the Shared Task (mBERT and XLM-RoBERTa), or not (DistilmBERT). We aimed for reproducibility rather than state-of-the-art, thus we use exclusively base-sized pretrained models and propose optimizations to bring them on par with large models. The objective is also to compare the impact of different changes, irrespective of the model size.

3.2 Zero-shot discourse relation classification

Our main motivation is to study the capacity of models for zero-shot adaptation to a new language, i.e. predicting discourse relation labels in a language, while trained on a model that has not seen that language (but has been trained on the given task in other languages). The goal is to observe under what conditions a model can adapt to new but similar data on which it has not been trained. We evaluate different scenarios of languages, frameworks, and label similarity with the Jaccard similarity coefficient [1].

Formally, given a set of corpora C, in which each corpus c belongs to a language $l(c)$ and a framework $f(c)$ and has a label set $A(c)$, we let $s(L) = \{c \in C | l(c) = L\}$ the set of corpora in a language L, and we train a model on a set:

- $S_{LF} \subseteq C$ of corpora in the same language family, where we remove successively corpora in a language L and test on $s(L)$.
- similarly on a set of corpora from the same framework $S_F = \{c \in C | f(c) = F\}$, for each language L: train on $S_F / s(L)$, test on $s(L) \cap S_F$.
- similarly on sets of corpora $S = \{c_1, .., c_k\}$ where $\forall (c, c') \in S^2, JC(A(c), A(c')) > t$, with JC the Jaccard coefficient, and t an *a priori* threshold.

[1] The Jaccard similarity coefficient (Jaccard, 1912) is a statistical method to calculate the similarity/diversity of sample sets. For two sets A and B, it is the ratio $|A \cap B| / |A \cup B|$.

3.3 Feature augmentation with tokens

We are training the classification models with a group of training sets joint and shuffled, across corpora and frameworks. To help the training process, we inject various information as prefixes to the input: either the input language name in English (e.g. German), the name of the corpus the input is taken from (e.g. `deu.rst.pcc`), and/or the framework (e.g. rst).

3.4 Classification label filtering

A classification model, in the prediction stage, returns a probability distribution of all labels in the training set. Anuranjana (2023) used a generative LLM that produces a human-readable string of text as label output, which may not belong to existing training set labels. Thus they proposed to *filter* the LLM output and select only outputs that exist as labels. Inspired by this, we are also post-processing the outputs of our classification models to pick the most probable label that belongs to the framework of the target corpus. This prevents the prediction of a label present in the merged training corpus but not in the target framework label set. We are filtering based on the framework and not on corpus-specific labels, in order to better examine the knowledge transfer between corpora of the same framework in the joint training process.

4 Experimental Settings

4.1 Data

DISRPT Benchmark We use the datasets (published with a unified text format) from the 2023 edition of the DISRPT Shared Task (Braud et al., 2023) for Task 3: *Discourse Relation Classification across Formalisms*.[2] The data is composed of 26 datasets for 13 languages covering 4 theoretical frameworks: PDTB (Penn Discourse Treebank Prasad et al., 2004), RST (Rhetorical Structure Theory, Mann and Thompson, 1988), DEP (Dependency structures, Yang and Li, 2018), or SDRT (Segmented Discourse Representation Theory, Asher and Lascarides, 2003). The datasets are presented in the Appendix, in Table 12, with the size of the label set, after the harmonization explained below, given in the last column.

Label harmonization When looking at the union of the label sets, we obtain a very large list of 163 distinct labels, making for a difficult learning problem. However, the proposals for unified label sets are limited to specific frameworks or not cover all relations present in corpora (Benamara and Taboada, 2015; Braud et al., 2017; Varachkina and Pannach, 2021). Moreover, within the shared task, the goal was to remain as faithful as possible to the original annotation and to produce predictions on the original label sets. We follow the harmonization proposed by one participating system based on (reversible) label substitutions and lower-casing, reducing the label set from 163 to 136 (Metheniti et al., 2023). In addition, we observed that GUM (`eng.rst.gum`) and GCDT (`zho.rst.gcdt`) have similar label sets in the DISRPT data, but the former included high-level classes of sense (e.g. *adversative*), while the latter used fine-grained, level 2 senses (e.g. *adversative-antithesis*). Therefore, we decided to use level 2 senses for GUM, a reversible mapping that reduces the label set from 136 to 128.

Finally, we decided to reorder the segment pairs, when necessary, in order to unify the relation direction in all inputs as *cause(1,2)*. The DISRPT data included segment pairs where the relation direction was cause-to-result (*cause(1,2)*) or result-to-cause *cause(2,1)*, even for the same relation. We switched the input order of pairs with the *cause(2,1)* direction, in accordance with previous studies.

4.2 Pretrained multilingual models

Multilingual BERT (mBERT) was introduced alongside the BERT architecture (Devlin et al., 2019) and is a pretrained model, trained on Wikipedia data of the top 104 languages, and with a masked language modeling (MLM) objective. The base and cased version of the model contains 12 layers, 12 heads, and 177M parameters. XLM-RoBERTa (Conneau et al., 2020) is a multilingual language model built on the RoBERTa architecture and is pretrained on 2.5TB of filtered CommonCrawl data of 100 languages. The base version of the model has 12 layers and 279M parameters. DistilmBERT (Sanh et al., 2019) is a multilingual distilled version of mBERT, also trained on Wikipedia data in 104 languages. The base and cased model has 6 layers, 12 heads, and 134M parameters.

The specific models used for the experiments in this paper are: bert-base-multilingual-cased, distilbert-base-multilingual-cased, and xlm-roberta-base. We built the classification models on

[2] https://github.com/disrpt/sharedtask2023/releases/tag/v1.0

Pytorch, and we trained each classification model for 10 epochs, keeping the best result out of the 10 epochs, based on a development set evaluation. For **zero-shot classification**, the models were also built with mBERT, with feature augmentation (language/corpus name/framework tokens at the beginning of each input sequence) and label filtering.

4.3 Reference models

As a **monolingual baseline**, we trained an mBERT classifier, individually for each corpus with a training set. For the four out-of-domain datasets (marked with an *asterisk) that do not have a training set, we trained a monolingual classifier with the dataset closest to the OOD model's label set, according to the Jaccard similarity coefficient (see Table 13). The `eng.dep.covdt` dataset was evaluated with `eng.rst.rstdt`, and the `eng.pdtb.tedm`, `por.pdtb.tedm`, and `tur.pdtb.tedm` datasets were evaluated with `eng.pdtb.pdtb`.

5 Results and Discussion

5.1 Multilingual discourse relation classification across formalisms

First, we experiment with a multilingual model trained over all the datasets jointly, in order to investigate language model performance, as well as the usefulness of the two enhancements proposed: filtering the output labels to ensure that predicted labels pertain to the target framework, and feature augmentation to inform the model about the nature of the source and target corpora.

5.1.1 Models with label filtering

In Table 1 we present the results for the three transformer architectures we tested, comparing their results before and after filtering the predicted labels per framework. For comparison, we also downloaded and trained the most successful system of DISRPT 2023, of the HITS team (Liu et al., 2023), with lowercased labels. We also compare our results with the reference monolingual classifiers (explained in Section 4.3). As it can be seen, discourse relation classification is a hard task, with rather low performance in general: 0.62 in accuracy at best on average, and around only 0.50 for a third of the corpora. Note that the high accuracy for `thai.pdtb.tdtb` comes from the fact that only explicit relations (triggered by a connective) are annotated in this corpus.

The HITS model outperforms our multilingual models in most corpora, but, in general, only for a 1-3% improvement. It should be noted that HITS is trained with larger specific pretrained language models and optimizations.

On the other hand, the multilingual models perform better than the monolingual ones in most cases, except for the larger English datasets and some Chinese datasets. Smaller datasets benefit moderately (e.g. `ita.pdtb.luna`, `rus.rst.rrt`) or significantly (e.g. `fra.sdrt.annodis`, `spa.rst.sctb`, `zho.rst.sctb`) from the multilingual setting, even with different frameworks present. We also note that some datasets have uniformly low accuracies with all models, such as `deu.rst.pcc` and `nld.rst.nldt`, a problem that is consistent with DISRPT 2023 results.

Regarding filtering per framework, for some frameworks there is no discernible improvement, meaning that the model was able to predict framework-related labels. However, for frameworks and corpora less represented in the data, we notice a large improvement (e.g. eng.dep.covdtb, eng.pdtb.tedm, zho.dep.scidtb).

Comparing the pretrained models we used, overall mBERT slightly outperformed XLM-RoBERTa (XML-R), the latter outperforming the former on certain corpora. DistilmBERT (DmBERT), even with its smaller parameter size, was still on par with the other two models and greatly benefited from label filtering. This finding supports our use of base models, instead of the large models used for the Shared Task, allowing for better reproducibility and interpretability.

5.1.2 Models with feature augmentation

The results for classification models with feature augmentation are presented in brief in Table 2. Models with feature augmentation outperform the baseline in all corpora except for `eng.sdrt.stac` and `*eng.dep.covdtb`. The features overall improve performance compared to models without features (see Section 5.1.1 and Table 1). These models also came even closer to the performance of the HITS system but did not outperform it.

The most successful configuration was the presence of all three tokens, language-corpus name-framework, especially for mBERT which was the most successful model overall, almost equalling the performance of HITS. XLM-RoBERTa benefited from the presence of any feature tokens. Experiments with DistilmBERT (which are omitted for brevity) showed that the smaller model benefited

Model	HITS	mBERT	DmBERT		mBERT		XLM-R	
		monol.	No F.	Filt.	No F.	Filt.	No F.	Filt.
deu.rst.pcc	**0.40**	0.32	0.31	0.31	0.35	0.35	0.37	0.37
***eng.dep.covdtb**	**0.69**	*0.63	0.25	0.41	0.18	0.47	0.12	0.30
eng.dep.scidtb	**0.75**	0.72	0.68	0.71	0.71	0.74	0.69	0.71
eng.pdtb.pdtb	**0.75**	0.73	0.69	0.71	0.71	0.73	0.71	0.73
***eng.pdtb.tedm**	**0.61**	*0.52	0.17	0.37	0.25	0.41	0.20	0.35
eng.rst.gum	**0.64**	0.54	0.39	0.41	0.44	0.45	0.42	0.43
eng.rst.rstdt	**0.67**	0.64	0.46	0.54	0.49	0.57	0.47	0.54
eng.sdrt.stac	**0.62**	**0.62**	0.58	0.61	0.58	0.60	0.59	0.60
eus.rst.ert	**0.51**	0.42	0.42	0.42	0.45	0.45	0.48	0.48
fas.rst.prstc	0.54	0.52	0.53	0.53	0.54	0.54	**0.55**	**0.55**
fra.sdrt.annodis	**0.55**	0.46	0.46	0.47	0.51	0.52	0.46	0.46
ita.pdtb.luna	**0.65**	0.52	0.53	0.53	0.55	0.56	0.51	0.53
nld.rst.nldt	**0.49**	0.43	0.45	0.45	0.46	0.46	0.47	0.47
por.pdtb.crpc	**0.74**	0.66	0.67	0.67	0.68	0.69	0.65	0.67
***por.pdtb.tedm**	0.46	*0.44	0.49	0.50	0.53	**0.54**	0.49	0.51
por.rst.cstn	0.63	0.57	0.59	0.59	0.61	0.62	**0.64**	**0.64**
rus.rst.rrt	**0.62**	0.59	0.59	0.59	0.60	0.60	0.60	0.60
spa.rst.rststb	**0.65**	0.56	0.58	0.59	0.63	0.63	0.62	0.62
spa.rst.sctb	0.61	0.43	0.61	0.61	**0.66**	**0.66**	0.55	0.55
tha.pdtb.tdtb	**0.96**	0.94	0.93	0.93	0.94	0.94	0.95	0.95
tur.pdtb.tdb	0.46	0.41	0.40	0.41	0.43	0.43	**0.49**	**0.49**
***tur.pdtb.tedm**	**0.48**	*0.35	0.42	0.42	0.46	**0.46**	0.46	0.46
zho.dep.scidtb	**0.68**	0.55	0.58	0.61	0.64	0.66	0.54	0.58
zho.pdtb.cdtb	**0.85**	0.83	0.72	0.79	0.72	0.80	0.76	0.82
zho.rst.gcdt	**0.61**	0.60	0.57	0.57	0.59	0.59	0.59	0.59
zho.rst.sctb	**0.55**	0.46	0.51	0.52	0.49	0.49	0.40	0.44
AVERAGE	**0.62**	0.56	0.52	0.55	0.55	0.58	0.53	0.56

Table 1: Classification results of multilingual classifiers, compared to the best system of DISRPT 2023 (**HITS**) and multiple monolingual mBERT classifiers (**mBERT monol.**). **No F.** is the accuracy score of the model before filtering and **Filt.** after filtering predicted labels per framework. Training was performed without feature augmentation.

Tokens	L		L+C		L+C+F	
Model (filt.)	**mBERT**	**XLM-R**	**mBERT**	**XLM-R**	**mBERT**	**XLM-R**
deu.rst.pcc	0.32	0.36	0.32	**0.37**	0.35	0.36
***eng.dep.covdtb**	0.49	0.34	0.26	0.23	0.24	0.22
eng.dep.scidtb	0.73	0.72	0.69	0.74	**0.75**	0.73
eng.pdtb.pdtb	0.73	0.74	0.74	0.75	0.73	**0.76**
***eng.pdtb.tedm**	0.46	0.33	0.54	0.52	**0.59**	0.52
eng.rst.gum	0.46	0.44	0.52	0.54	**0.57**	0.55
eng.rst.rstdt	0.54	0.55	0.64	0.64	**0.65**	0.65
eng.sdrt.stac	0.60	0.58	0.58	0.60	0.61	0.61
eus.rst.ert	0.45	0.46	0.43	0.47	**0.51**	0.45
fas.rst.prstc	0.53	0.52	0.49	0.53	**0.54**	0.50
fra.sdrt.annodis	0.50	0.48	0.44	0.47	**0.51**	**0.51**
ita.pdtb.luna	**0.60**	0.57	0.54	0.59	**0.60**	0.57
nld.rst.nldt	0.47	0.47	0.45	**0.49**	**0.49**	0.46
por.pdtb.crpc	0.69	0.69	0.69	0.69	**0.74**	0.71
***por.pdtb.tedm**	0.52	0.54	0.52	0.53	**0.59**	0.53
por.rst.cstn	0.64	0.63	0.60	0.62	**0.67**	0.62
rus.rst.rrt	0.60	0.61	0.58	0.61	**0.62**	0.60
spa.rst.rststb	0.63	0.59	0.56	0.61	**0.66**	0.63
spa.rst.sctb	0.68	0.60	0.69	0.65	**0.70**	0.64
tha.pdtb.tdtb	0.94	**0.96**	0.93	0.95	0.95	0.95
tur.pdtb.tdb	0.46	0.47	0.39	0.47	**0.52**	0.47
***tur.pdtb.tedm**	0.45	0.47	0.42	0.45	**0.48**	0.42
zho.dep.scidtb	0.65	0.60	0.62	0.64	**0.68**	0.68
zho.pdtb.cdtb	0.81	0.83	0.83	**0.84**	**0.84**	**0.84**
zho.rst.gcdt	0.58	0.56	0.58	0.59	0.60	**0.62**
zho.rst.sctb	0.51	0.41	0.64	0.60	**0.67**	0.61
AVERAGE	0.58	0.56	0.57	0.58	**0.61**	0.58

Table 2: Classification results for mBERT/XLM-RoBERTa models with label filtering and feature augmentation. The additional tokens at the start of the sequence are **L** (language in English), **C** (name of the corpus), and **F** (name of the framework). The entirety of the results are presented in Appendix, Table 14.

from either the presence of the language token or the presence of all three tokens. Feature augmentation has been greatly explored in discourse relation classification (e.g. with syntactic information for the DISRPT task by Gessler et al., 2021), and has proven to improve accuracy with all types of models. Our proposed approach is very simple, at this current stage, and does not require calculated features, yet it improves results and supports our use of base models over models with more parameters and optimizations.

5.2 Zero-shot discourse relation classification

Our goal is to investigate under which conditions transfer to a language not present during fine-tuning could be successful for the task at hand. Since mBERT was the most successful in previous multilingual experiments, we keep this model for the zero-shot setting with feature augmentation and label filtering. In the Tables presenting the results, we also report the score of Jaccard similarity between the corpus (i.e. the target) and the group of corpora used at training time (source): this score is an indication of the label set overlap between the source and target data.

5.2.1 Zero-shot within language families

For the first set of zero-shot learning experiments, we wanted to test prediction with a model trained on languages of the same family, omitting the target language. The corpora of DISRPT 2023 contain 13 languages, with great typological variety. In order to maintain the motivation of multilingualism and variation, but also ensure enough data for finetuning, we looked for language families significantly present. It is the case for the Germanic family, with German, English, and Dutch corpora, and for the Romance languages with French, Italian, Portuguese, and Spanish corpora. An example of zero-shot learning per language is: a model is trained on all Germanic language corpora except for all the English ones, thus predictions on English corpora are zero-shot.

In Table 3 we present the zero-shot results for languages of the Germanic family. We observe an expected steep drop in accuracy for most corpora. Some English corpora almost had zero accuracy, which is expected, since the corpora labels never existed in the training set. The `eng.rst.rstdt` and `nld.rst.nldt` are the only ones whose loss in accuracy is not catastrophic, because they are the ones with the less variation in labels and their labels

	Monolingual	Zero-Shot	Jac. Similar.
deu.rst.pcc	0.32	0.15	0.20
***eng.dep.covdtb**	*0.63	0.52	0.12
eng.dep.scidtb	0.72	0.06	0.20
eng.pdtb.pdtb	0.73	0.03	0.04
***eng.pdtb.tedm**	*0.52	0.02	0.04
eng.rst.gum	0.54	0.05	0.10
eng.rst.rstdt	0.64	0.40	0.20
eng.sdrt.stac	0.62	0.09	0.11
nld.rst.nldt	0.43	0.26	0.22

Table 3: Classification results for zero-shot models and Germanic languages.

	Monolingual	Zero-Shot	Jac. Similar.
fra.sdrt.annodis	0.46	0.23	0.11
ita.pdtb.luna	0.52	0.20	0.15
por.pdtb.crpc	0.66	0.04	0.16
***por.pdtb.tedm**	*0.44	0.05	0.15
por.rst.cstn	0.57	0.29	0.38
spa.rst.rststb	0.56	0.25	0.32
spa.rst.sctb	0.43	0.35	0.29

Table 4: Classification results for zero-shot models and Romance languages.

exist in the German and Dutch RST datasets. The `eng.dep.covdtb` had a relatively high accuracy because it has a high occurrence of the *elaboration* label, making prediction easier for models trained on a few labels.

The zero-shot results for languages of the Romance family are presented in Table 4. Similarly, accuracy is expectedly very low for the Portuguese PDTB corpora that have unique labels. The rest of the corpora demonstrate lower accuracies, with part of the problem being their smaller dataset sizes and low label similarity.

5.2.2 Zero-shot within frameworks

These experiments were conducted with corpora of the same framework. For example, the zero-shot model for the `spa.rst.rststb` corpus is trained on all the other RST corpora, except the Spanish `spa.rst.sctb` corpus.

The results for zero-shot classification for PDTB corpora are presented in Table 5. For most corpora, zero-shot predictions have a lower accuracy; accuracy drops significantly for `tha.pdtb.tdtb` (a corpus with mostly explicit relations, compared to the mix of implicit/explicit relations in other corpora) and `zho.pdtb.cdtb` (a corpus with smaller variation). However, for `*por.pdtb.tedm` and `*tur.pdtb.tedm`, two of the OOD datasets, there is a performance improvement when the classifier is trained with all PDTB corpora (except for the target language), compared to the monolingual

	Monolingual	Zero-Shot	Jac. Similar.
eng.pdtb.pdtb	0.73	0.55	0.54
***eng.pdtb.tedm**	*0.52	0.55	0.50
ita.pdtb.luna	0.52	0.42	0.27
por.pdtb.crpc	0.66	0.48	0.46
***por.pdtb.tedm**	*0.44	0.45	0.51
tha.pdtb.tdtb	0.94	0.57	0.49
tur.pdtb.tdb	0.41	0.37	0.51
***tur.pdtb.tedm**	*0.35	0.40	0.59
zho.pdtb.cdtb	0.83	0.47	0.22

Table 5: Classification results for zero-shot models of the PDTB framework.

	Monolingual	Zero-Shot	Jac. Similar.
deu.rst.pcc	0.32	0.20	0.28
eng.rst.gum	0.54	0.10	0.40
eng.rst.rstdt	0.64	0.42	0.21
eus.rst.ert	0.42	0.33	0.35
fas.rst.prstc	0.52	0.40	0.21
nld.rst.nldt	0.43	0.30	0.37
por.rst.cstn	0.57	0.49	0.37
rus.rst.rrt	0.59	0.40	0.24
spa.rst.rststb	0.56	0.46	0.33
spa.rst.sctb	0.43	0.60	0.32
zho.rst.gcdt	0.60	0.01	0.40
zho.rst.sctb	0.46	0.48	0.33

Table 6: Classification results for zero-shot models of the RST framework.

	Monolingual	Zero-Shot	Jac. Similar.
eng.sdrt.stac	0.62	0.19	0.48
fra.sdrt.annodis	0.46	0.24	0.48

Table 7: Classification results for zero-shot models of the SDRT framework.

	Monolingual	Zero-Shot	Jac. Similar.
***eng.dep.covdtb**	*0.63	0.11	0.29
eng.dep.scidtb	0.72	0.35	0.79
zho.dep.scidtb	0.55	0.41	0.79

Table 8: Classification results for zero-shot models of the DEP framework.

`eng.pdtb.pdtb` classifier. The `eng.pdtb.pdtb` corpus was chosen because it has the closest label set overlap with these corpora, but the larger and more varied zero-shot training set was beneficial for the target predictions.

Regarding zero-shot classification for RST corpora (see Table 6), results vary. For most corpora, the same small deterioration was observed as in the PDTB corpora. The two Spanish RST corpora showed improvement compared to the monolingual model; the presence of many common labels was beneficial to the zero-shot setting. Corpora with unique label sets had accuracies close to zero: `eng.rst.gum` and `zho.rst.gcdt` are very dissimilar to any other corpora, even after label harmo-

nization, and eng.rst.rstdt is only similar to the much smaller fas.rst.prstc.

The results for the SDRT corpora can be found in Table 7. Given that these corpora are much smaller, it is expected that accuracy would be quite low, despite their similar label sets (0.48 on the Jaccard index).

For DEP corpora (Table 8), eng.dep.scidtb has a bigger drop in zero-shot accuracy compared to zho.dep.scidtb, due to the smaller size of the Chinese dataset. The *eng.dep.covdtb dataset shows the same low accuracy as in many of the multilingual settings.

5.2.3 Zero-shot within groups with similar label sets

Our previous experiments on zero-shot learning demonstrated that the best results came from combinations of corpora with similar label sets, regardless of languages or annotation frameworks. To confirm this observation, we calculated the Jaccard correlation coefficient between pairs of corpus label sets and created groups with at least 0.4 similarity [3]. We created three groups with the required similarity and adequate training data [4], and we train without including the target language.

The first group is composed of PDTB corpora, as seen in Table 9. As with the framework zero-shot models, we observe a large drop in accuracy in the Thai corpus, because of its explicit relations. The rest of the corpora show slightly lower accuracy, even without the presence of the language in the training set. Additionally, two of the OOD corpora, the Portuguese and Turkish, show improvement in the zero-shot setting; the larger training sets and label sets are beneficial, compared to only training with English corpora.

The second group includes many of the RST corpora (see Table 10). While the Spanish and Chinese models showed improvement or no significant loss, the German, Dutch, and Russian models had lower performance. These corpora have been hard to classify in other monolingual and multilingual settings as well, and further investigation into the annotation quality may be required.

The third group is composed of DEP corpora and two RST corpora (see Table 11), not part of the previous group: eng.rst.rstdt and fas.rst.prstc. For this group, Jaccard similarities were slightly lower than for the other groups, given that there are two frameworks and varied training sizes. All accuracies are quite low, even for the English corpora, and there was no improvement for DEP corpora with the addition of the RST corpora (as seen in Table 8) or vice versa.

[3]Note that it is not a purely zero-shot setting, since we use information about the target corpus label set.

[4]The corpora not belonging to any of these groups are eng.rst.gum, eng.sdrt.stac, fra.sdrt.annodis, ita.pdtb.luna, zho.pdtb.cdtb, and zho.rst.gcdt.

	Monolingual	Zero-Shot	Jac. similar.
eng.pdtb.pdtb	0.73	0.55	0.71
***eng.pdtb.tedm**	*0.52	0.55	0.67
por.pdtb.crpc	0.66	0.47	0.55
***por.pdtb.tedm**	*0.44	0.46	0.74
tha.pdtb.tdtb	0.94	0.58	0.65
tur.pdtb.tdb	0.41	0.38	0.68
***tur.pdtb.tedm**	*0.35	0.42	0.79

Table 9: Classification results for zero-shot models of the Jaccard PDTB group.

	Monolingual	Zero-Shot	Jac. similar.
deu.rst.pcc	0.32	0.18	0.47
eus.rst.ert	0.42	0.36	0.57
nld.rst.nldt	0.43	0.31	0.62
por.rst.cstn	0.57	0.46	0.60
rus.rst.rrt	0.59	0.31	0.40
spa.rst.rststb	0.56	0.49	0.55
spa.rst.sctb	0.43	0.61	0.54
zho.rst.sctb	0.46	0.51	0.55

Table 10: Classification results for zero-shot models of the Jaccard RST group.

	Monolingual	Zero-Shot	Jac. similar.
eng.dep.scidtb	0.72	0.40	0.73
eng.rst.rstdt	0.64	0.37	0.55
fas.rst.prstc	0.52	0.46	0.50
zho.dep.scidtb	0.55	0.43	0.69

Table 11: Classification results for zero-shot models of the Jaccard DEP-RST group.

6 Conclusion

In this paper, we presented our work toward zero-shot classification of discourse relations. Our goal was to adhere closely to a multilingual, multi-framework approach, even if it would not outperform the current state-of-the-art. We first explored the relation classification systems of the DISRPT Shared Task, to find an adequate solution for multilingual multi-framework classification. We found out that a classifier based on mBERT performs on the same level as monolingual approaches with large models, for most corpora, with the addition of feature augmentation and label filtering.

We proceeded with our zero-shot experiments,

testing knowledge transfer with a multilingual pretrained model among language families, languages of the same framework, and languages of similar corpus label sets. Zero-shot learning was challenging as expected, but gave interesting results. It worked best for models trained with similar label sets and an adequate amount of data, and the multilingual embeddings were capable of handling the exclusion of the target language. This is a hopeful finding for research in this direction, for the future introduction of under-represented languages into discourse analysis, and for the integration of discourse analysis into multi-task architectures.

References

Stergos Afantenos, Nicholas Asher, Farah Benamara, Anaïs Cadilhac, Cédric Degremont, Pascal Denis, Markus Guhe, Simon Keizer, Alex Lascarides, Oliver Lemon, Philippe Muller, Soumya Paul, Verena Rieser, and Laure Vieu. 2012. Developing a corpus of strategic conversation in the settlers of catan. In *Proceedings of the workshop on Games and NLP (GAMNLP)*.

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.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, Online. Association for Computational Linguistics.

Nicholas Asher, Julie Hunter, Mathieu Morey, Benamara Farah, and Stergos Afantenos. 2016. Discourse structure and dialogue acts in multiparty dialogue: the STAC corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2721–2727, Portorož, Slovenia. European Language Resources Association (ELRA).

Nicholas Asher and Alex Lascarides. 2003. *Logics of Conversation*. Cambridge University Press.

Farah Benamara and Maite Taboada. 2015. Mapping different rhetorical relation annotations: A proposal. In *Proceedings of Starsem*.

Chloé Braud, Maximin Coavoux, and Anders Søgaard. 2017. Cross-lingual RST discourse parsing. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 292–304, Valencia, Spain. Association for Computational Linguistics.

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.

Shuyuan Cao, Iria da Cunha, and Mikel Iruskieta. 2018. The RST Spanish-Chinese treebank. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 156–166, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Paula Christina Figueira Cardoso, Erick Galani Maziero, Maria Lucía del Rosario Castro Jorge, M. Eloize, R. Kibar Aji Seno, Ariani Di Felippo, Lucia Helena Machado Rino, Maria das Graças Volpe Nunes, and Thiago Alexandre Salgueiro Pardo. 2011. CSTNews - a discourse-annotated corpus for single and multi-document summarization of news texts in Brazilian Portuguese. In *Proceedings of the 3rd RST Brazilian Meeting*, pages 88–105, Cuiabá, Brazil.

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurovsky. 2001. Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory. In *Proceedings of the Second SIGdial Workshop on Discourse and Dialogue*.

Junxuan Chen, Xiang Li, Jiarui Zhang, Chulun Zhou, Jianwei Cui, Bin Wang, and Jinsong Su. 2020. Modeling discourse structure for document-level neural machine translation. In *Proceedings of the First Workshop on Automatic Simultaneous Translation*, pages 30–36, Seat-

tle, Washington. Association for Computational Linguistics.

Yi Cheng and Sujian Li. 2019. Zero-shot Chinese discourse dependency parsing via cross-lingual mapping. In *Proceedings of the 1st Workshop on Discourse Structure in Neural NLG*, pages 24–29, Tokyo, Japan. Association for Computational Linguistics.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv 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. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Iria da Cunha, Juan-Manuel Torres-Moreno, and Gerardo Sierra. 2011. On the development of the RST Spanish Treebank. In *Proceedings of the Fifth Linguistic Annotation Workshop (LAW V)*, pages 1–10, Portland, OR.

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.

Abteen Ebrahimi, Manuel Mager, Arturo Oncevay, Vishrav Chaudhary, Luis Chiruzzo, Angela Fan, John Ortega, Ricardo Ramos, Annette Rios, Ivan Vladimir Meza Ruiz, Gustavo Giménez-Lugo, Elisabeth Mager, Graham Neubig, Alexis Palmer, Rolando Coto-Solano, Thang Vu, and Katharina Kann. 2022. AmericasNLI: Evaluating zero-shot natural language understanding of pretrained multilingual models in truly low-resource languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6279–6299, Dublin, Ireland. Association for Computational Linguistics.

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.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In *International Conference on Machine Learning*, pages 2790–2799. PMLR.

Mikel Iruskieta, María Jesús Aranzabe, Arantza Diaz de Ilarraza, Itziar Gonzalez-Dios, Mikel Lersundi, and Oier Lopez de Lacalle. 2013. The RST Basque TreeBank: An online search interface to check rhetorical relations. In *4th Workshop on RST and Discourse Studies*, pages 40–49, Fortaleza, Brasil.

Paul Jaccard. 1912. The distribution of the flora in the alpine zone. 1. *New phytologist*, 11(2):37–50.

Katharina Kann, Abteen Ebrahimi, Manuel Mager, Arturo Oncevay, John E Ortega, Annette Rios, Angela Fan, Ximena Gutierrez-Vasques, Luis Chiruzzo, Gustavo A Giménez-Lugo, et al. 2022. AmericasNLI: Machine translation and natural language inference systems for indigenous languages of the americas. *Frontiers in Artificial Intelligence*, 5:266.

Yudai Kishimoto, Yugo Murawaki, and Sadao Kurohashi. 2020. Adapting BERT to implicit discourse relation classification with a focus

on discourse connectives. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1152–1158, Marseille, France. European Language Resources Association.

Ziheng Lin, Min-Yen Kan, and Hwee Tou Ng. 2009. Recognizing implicit discourse relations in the Penn Discourse Treebank. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 343–351, Singapore. Association for Computational Linguistics.

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.

Wei Liu and Michael Strube. 2023. Annotation-inspired implicit discourse relation classification with auxiliary discourse connective generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15696–15712, Toronto, Canada. Association for Computational Linguistics.

Zhengyuan Liu, Ke Shi, and Nancy Chen. 2020. Multilingual neural RST discourse parsing. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6730–6738, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Zhengyuan Liu, Ke Shi, and Nancy Chen. 2021. DMRST: A joint framework for document-level multilingual RST discourse segmentation and parsing. In *Proceedings of the 2nd Workshop on Computational Approaches to Discourse*, pages 154–164, Punta Cana, Dominican Republic and Online. Association for Computational Linguistics.

William C. Mann and Sandra A. Thompson. 1988. Rhetorical Structure Theory: Toward 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.

Noriki Nishida and Yuji Matsumoto. 2022. Out-of-domain discourse dependency parsing via bootstrapping: An empirical analysis on its effectiveness and limitation. *Transactions of the Association for Computational Linguistics*, 10:127–144.

Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. Cross-lingual name tagging and linking for 282 languages. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1946–1958, Vancouver, Canada. Association for Computational Linguistics.

Emily Pitler, Annie Louis, and Ani Nenkova. 2009. Automatic sense prediction for implicit discourse relations in text. In *Proceedings of ACL-IJCNLP*.

Rashmi Prasad, Eleni Miltsakaki, Aravind Joshi, and Bonnie Webber. 2004. Annotation and data mining of the penn discourse treebank. In *Proceedings of the ACL Workshop on Discourse Annotation*.

Rashmi Prasad, Bonnie Webber, and Aravind Joshi. 2014. Reflections on the penn discourse treebank, comparable corpora and complementary annotation. *Computational Linguistics*.

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

Kunxun Qi, Hai Wan, Jianfeng Du, and Haolan Chen. 2022. Enhancing cross-lingual natural language inference by prompt-learning from cross-lingual templates. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*,

pages 1910–1923, Dublin, Ireland. Association for Computational Linguistics.

Gisela Redeker, Ildikó Berzlánovich, Nynke van der Vliet, Gosse Bouma, and Markus Egg. 2012. Multi-layer discourse annotation of a Dutch text corpus. In *Proceedings of LREC 2012*, pages 2820–2825, Istanbul, Turkey.

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 and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Victor Sanh, Thomas Wolf, and Sebastian Ruder. 2019. A hierarchical multi-task approach for learning embeddings from semantic tasks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6949–6956.

Sara Shahmohammadi, Hadi Veisi, and Ali Darzi. 2021. Persian Rhetorical Structure Theory. *arXiv preprint arXiv:2106.13833*.

Manfred Stede and Arne Neumann. 2014. Potsdam commentary corpus 2.0: Annotation for discourse research. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 925–929, Reykjavik, Iceland. European Language Resources Association (ELRA).

Svetlana Toldova, Dina Pisarevskaya, Margarita Ananyeva, Maria Kobozeva, Alexander Nasedkin, Sofia Nikiforova, Irina Pavlova, and Alexey Shelepov. 2017. Rhetorical relations markers in Russian RST treebank. In *Proceedings of the 6th Workshop on Recent Advances in RST and Related Formalisms*, pages 29–33, Santiago de Compostela, Spain. 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).

Hanna Varachkina and Franziska Pannach. 2021. A unified approach to discourse relation classification in nine languages. In *Proceedings of the 2nd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2021)*, pages 46–50, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yizhong Wang, Sujian Li, and Houfeng Wang. 2017. A two-stage parsing method for text-level discourse analysis. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 184–188, Vancouver, Canada. Association for Computational Linguistics.

Changxing Wu, Liuwen Cao, Yubin Ge, Yang Liu, Min Zhang, and Jinsong Su. 2022. A label dependence-aware sequence generation model for multi-level implicit discourse relation recognition. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 11486–11494. AAAI Press.

Hongyi Wu, Hao Zhou, Man Lan, Yuanbin Wu, and Yadong Zhang. 2023. Connective prediction for implicit discourse relation recognition via knowledge distillation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5908–5923, Toronto, Canada. Association for Computational Linguistics.

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.

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.

An Yang and Sujian Li. 2018. SciDTB: Discourse dependency TreeBank for scientific abstracts. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 444–449, Melbourne, Australia. Association for Computational Linguistics.

Cheng Yi, Li Sujian, and Li Yueyuan. 2021. Unifying discourse resources with dependency framework. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 1058–1065, Huhhot, China. Chinese Information Processing Society of China.

Nan Yu, Meishan Zhang, Guohong Fu, and Min Zhang. 2022. RST discourse parsing with second-stage EDU-level pre-training. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4269–4280, Dublin, Ireland. Association for Computational Linguistics.

Amir Zeldes. 2017. The GUM corpus: Creating multilayer resources in the classroom. *Language Resources and Evaluation*, 51(3):581–612.

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.

Deniz Zeyrek, Amália Mendes, Yulia Grishina, Murathan Kurfalı, Samuel Gibbon, and Maciej Ogrodniczuk. 2020. TED Multilingual Discourse Bank (TED-MDB): a parallel corpus annotated in the PDTB style. *Language Resources and Evaluation*, 54:587–613.

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).

Longyin Zhang, Fang Kong, and Guodong Zhou. 2021. Adversarial learning for discourse rhetorical structure parsing. 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 3946–3957, Online. Association for Computational Linguistics.

Haodong Zhao, Ruifang He, Mengnan Xiao, and Jing Xu. 2023. Infusing hierarchical guidance into prompt tuning: A parameter-efficient framework for multi-level implicit discourse relation recognition. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6477–6492, Toronto, Canada. Association for Computational Linguistics.

Yuping Zhou, Jill Lu, Jennifer Zhang, and Nianwen Xue. 2014. Chinese discourse treebank 0.5 ldc2014t21. *Web Download. Philadelphia: Linguistic Data Consortium*.

A Appendix: Datasets

A.1 Data size

Corpus	Source	Language	Framework	Train set	Dev. set	Test set	Num. of relations
deu.rst.pcc	Stede and Neumann (2014)	German	RST	2164	241	260	26
*eng.dep.covdtb	Nishida and Matsumoto (2022)	English	DEP	0	2399	2586	11
eng.dep.scidtb	Yang and Li (2018)		DEP	6060	1933	1911	24
eng.pdtb.pdtb	Prasad et al. (2019)		PDTB	43920	1674	2257	23
*eng.pdtb.tedm	Zeyrek et al. (2018)		PDTB	0	178	351	20
eng.rst.gum	Zeldes (2017)		RST	19496	2617	2575	31
eng.rst.rstdt	Carlson et al. (2001)		RST	16002	1621	2155	17
eng.sdrt.stac	Asher et al. (2016)		SDRT	9580	1145	1510	16
eus.rst.ert	Iruskieta et al. (2013)	Basque	RST	2533	614	678	27
fas.rst.prstc	Shahmohammadi et al. (2021)	Persian	RST	4100	499	592	17
fra.sdrt.annodis	Afantenos et al. (2012)	French	SDRT	2185	528	625	18
ita.pdtb.luna	Tonelli et al. (2010)	Italian	PDTB	955	209	380	15
nld.rst.nldt	Redeker et al. (2012)	Dutch	RST	1608	331	325	30
por.pdtb.crpc	Mendes and Lejeune (2022)	Portuguese	PDTB	8797	1285	1248	21
*por.pdtb.tedm	Zeyrek et al. (2018)		PDTB	0	190	364	20
por.rst.cstn	Cardoso et al. (2011)		RST	4148	573	272	32
rus.rst.rrt	Toldova et al. (2017)	Russian	RST	28868	2855	2843	22
spa.rst.rststb	da Cunha et al. (2011)	Spanish	RST	2240	383	426	27
spa.rst.sctb	Cao et al. (2018)		RST	439	94	159	25
tha.pdtb.tdtb	Braud et al. (2023)	Thai	PDTB	8278	1243	1344	21
tur.pdtb.tdb	Zeyrek and Kurfalı (2017)	Turkish	PDTB	2451	312	422	23
*tur.pdtb.tedm	Zeyrek et al. (2020)		PDTB	0	213	364	23
zho.dep.scidtb	Cheng and Li (2019)	Chinese (Mandarin)	DEP	802	281	215	23
zho.pdtb.cdtb	Zhou et al. (2014)		PDTB	3657	855	758	9
zho.rst.gcdt	Yi et al. (2021)		RST	6454	1006	953	31
zho.rst.sctb	Cao et al. (2018)		RST	439	94	159	26

Table 12: A comprehensive list of the datasets used for the DISRPT 2023 Shared Task. Languages with an asterisk were out-of-domain (no training set). **Num. of relations** is the relations count in the dataset.

A.2 Jaccard similarities between DISRPT 2023 dataset label sets

	deu.rst.pcc	*eng.dep.covdtb	eng.dep.scidtb	eng.pdtb.pdtb	*eng.pdtb.tedm	eng.rst.gum	eng.rst.rstdt	eng.sdrt.stac	eus.rst.ert	fas.rst.prstc	fra.sdrt.annodis	ita.pdtb.luna	nld.rst.nldt	por.pdtb.crpc	*por.pdtb.tedm	por.rst.cstn	rus.rst.rrt	spa.rst.rststb	spa.rst.sctb	tha.pdtb.tdtb	tur.pdtb.tdb	*tur.pdtb.tedm	zho.dep.scidtb	zho.pdtb.cdtb	zho.rst.gcdt	zho.rst.sctb
deu.rst.pcc	1	0.19	0	0.04	0.05	0.08	0.19	0.14	0.71	0.19	0.16	0.05	0.56	0.04	0.05	0.38	0.41	0.71	0.65	0.04	0.04	0.04	0.2	0.17	0.08	0.68
***eng.dep.covdtb**	0.12	1	0.3	0	0	0	0.56	0.13	0.15	0.47	0.16	0.04	0.14	0.07	0	0.19	0.22	0.15	0.16	0	0	0	0.31	0.11	0	0.16
eng.dep.scidtb	0.19	0.3	1	0	0	0.02	0.41	0.11	0.24	0.41	0.14	0.03	0.17	0.05	0	0.19	0.24	0.24	0.26	0	0	0	0.88	0.18	0.02	0.25
eng.pdtb.pdtb	0.04	0	0	1	0.87	0.02	0	0	0.04	0	0	0.27	0.04	0.63	0.87	0	0	0.04	0.04	0.76	0.84	0.92	0	0.03	0.02	0.04
***eng.pdtb.tedm**	0.05	0	0	0.87	1	0.02	0	0	0.04	0	0	0.3	0.04	0.71	0.9	0	0	0.04	0.05	0.78	0.79	0.87	0	0.04	0.02	0.05
eng.rst.gum	0.08	0	0.02	0.02	0.02	1	0	0	0.12	0	0.02	0.02	0.11	0.02	0.02	0.09	0.06	0.12	0.12	0.02	0.02	0.02	0.02	0	1	0.12
eng.rst.rstdt	0.19	0.56	0.41	0	0	0	1	0.18	0.26	0.89	0.21	0.03	0.21	0.06	0	0.26	0.3	0.26	0.27	0	0	0	0.43	0.18	0	0.26
eng.sdrt.stac	0.14	0.13	0.11	0	0	0	0.18	1	0.13	0.18	0.48	0.03	0.1	0.03	0	0.12	0.12	0.16	0.14	0	0.03	0	0.11	0.14	0	0.14
eus.rst.ert	0.71	0.15	0.24	0.04	0.04	0.12	0.26	0.13	1	0.26	0.13	0.05	0.78	0.04	0.04	0.55	0.48	0.93	0.86	0.04	0.04	0.04	0.25	0.16	0.12	0.89
fas.rst.prstc	0.19	0.47	0.41	0	0	0	0.89	0.18	0.26	1	0.21	0.03	0.21	0.06	0	0.26	0.3	0.26	0.27	0	0	0	0.43	0.18	0	0.26
fra.sdrt.annodis	0.16	0.16	0.14	0	0	0.02	0.21	0.48	0.13	0.21	1	0.03	0.09	0	0	0.14	0.14	0.15	0.16	0	0	0	0.14	0.13	0.02	0.16
ita.pdtb.luna	0.05	0.04	0.03	0.27	0.3	0.02	0.03	0.03	0.05	0.03	0.03	1	0.05	0.33	0.3	0.04	0.06	0.08	0.05	0.33	0.27	0.31	0.03	0.14	0.02	0.05
nld.rst.nldt	0.56	0.14	0.17	0.04	0.04	0.11	0.21	0.1	0.78	0.21	0.09	0.05	1	0.04	0.04	0.68	0.41	0.73	0.67	0.04	0.04	0.04	0.18	0.11	0.11	0.7
por.pdtb.crpc	0.04	0.07	0.05	0.63	0.71	0.02	0.06	0.03	0.04	0.06	0	0.33	0.04	1	0.64	0.02	0.02	0.04	0.05	0.56	0.63	0.63	0.05	0.07	0.02	0.04
***por.pdtb.tedm**	0.05	0	0	0.87	0.9	0.02	0	0	0.04	0	0	0.3	0.04	0.64	1	0	0	0.04	0.05	0.71	0.79	0.87	0	0.04	0.02	0.05
por.rst.cstn	0.38	0.19	0.19	0	0	0.09	0.26	0.12	0.55	0.26	0.14	0.04	0.68	0.02	0	1	0.46	0.51	0.5	0	0	0	0.2	0.08	0.09	0.53
rus.rst.rrt	0.41	0.22	0.24	0	0	0.06	0.3	0.12	0.48	0.3	0.14	0.06	0.41	0.02	0	0.46	1	0.48	0.47	0	0	0	0.25	0.15	0.06	0.5
spa.rst.rststb	0.71	0.15	0.24	0.04	0.04	0.12	0.26	0.16	0.93	0.26	0.15	0.08	0.73	0.04	0.04	0.51	0.48	1	0.86	0.04	0.04	0.04	0.25	0.2	0.12	0.89
spa.rst.sctb	0.65	0.16	0.26	0.04	0.05	0.12	0.27	0.14	0.86	0.27	0.16	0.05	0.67	0.05	0.05	0.5	0.47	0.86	1	0.05	0.04	0.04	0.26	0.17	0.12	0.96
tha.pdtb.tdtb	0.04	0	0	0.76	0.78	0.02	0	0	0.04	0	0	0.33	0.04	0.56	0.71	0	0	0.04	0.05	1	0.69	0.83	0	0.07	0.02	0.04
tur.pdtb.tdb	0.04	0	0	0.84	0.79	0.02	0	0.03	0.04	0	0	0.27	0.04	0.63	0.79	0	0	0.04	0.04	0.69	1	0.84	0	0.03	0.02	0.04
***tur.pdtb.tedm**	0.04	0	0	0.92	0.87	0.02	0	0	0.04	0	0	0.31	0.04	0.63	0.87	0	0	0.04	0.04	0.83	0.84	1	0	0.07	0.02	0.04
zho.dep.scidtb	0.2	0.31	0.88	0	0	0.02	0.43	0.11	0.25	0.43	0.14	0.03	0.18	0.05	0	0.2	0.25	0.25	0.26	0	0	0	1	0.19	0.02	0.26
zho.pdtb.cdtb	0.17	0.11	0.18	0.03	0.04	0	0.18	0.14	0.16	0.18	0.13	0.14	0.11	0.07	0.04	0.08	0.15	0.2	0.17	0.07	0.03	0.07	0.19	1	0	0.17
zho.rst.gcdt	0.08	0	0.02	0.02	0.02	1	0	0	0.12	0	0.02	0.02	0.11	0.02	0.02	0.09	0.06	0.12	0.12	0.02	0.02	0.02	0.02	0	1	0.12
zho.rst.sctb	0.68	0.16	0.25	0.04	0.05	0.12	0.26	0.14	0.89	0.26	0.16	0.05	0.7	0.04	0.05	0.53	0.5	0.89	0.96	0.04	0.04	0.04	0.26	0.17	0.12	1

Table 13: Jaccard similarity of the label sets of two datasets, calculated between pairs of datasets from the DISRPT 2023 Shared Task.

B Appendix: Results

B.1 Full feature augmentation

	Tokens:	Language						Language, Name						Language, Name, Framework					
Model	**Monolingual**	**DistilmBERT**		**mBERT**		**XLM-R**		**DistilmBERT**		**mBERT**		**XLM-R**		**DistilmBERT**		**mBERT**		**XLM-R**	
Corpus	**No F.**	**No F.**	**F.**	**No F.**	**F.**	**No F.**	**F.**	**No F.**	**F.**	**No F.**	**F.**	**No F.**	**F.**	**No F.**	**F.**	**No F.**	**F.**	**No F.**	**F.**
deu.rst.pcc	0.32	0.28	0.28	0.32	0.32	0.36	0.36	0.18	0.18	0.32	0.32	**0.37**	**0.37**	0.34	0.34	0.35	0.35	0.36	0.36
***eng.dep.covdtb**	**0.63**	0.16	0.38	0.24	0.49	0.16	0.34	0.22	0.38	0.26	0.26	0.19	0.23	0.24	0.24	0.24	0.24	0.20	0.22
eng.dep.scidtb	0.72	0.62	0.66	0.71	0.73	0.70	0.72	0.35	0.36	0.69	0.69	0.74	0.74	**0.73**	0.73	0.75	**0.75**	0.73	0.73
eng.pdtb.pdtb	0.73	0.67	0.69	0.72	0.73	0.73	0.74	0.36	0.39	0.74	0.74	0.75	0.75	0.73	0.73	0.73	0.73	**0.76**	**0.76**
***eng.pdtb.tedm**	0.52	0.16	0.37	0.29	0.46	0.18	0.33	0.01	0.02	0.54	0.54	0.52	0.52	0.51	0.51	**0.59**	**0.59**	0.52	0.52
eng.rst.gum	0.54	0.36	0.37	0.44	0.46	0.43	0.44	0.16	0.17	0.52	0.52	0.54	0.54	0.52	0.52	**0.57**	**0.57**	0.55	0.55
eng.rst.rstdt	0.64	0.46	0.53	0.46	0.54	0.48	0.55	0.42	0.48	0.64	0.64	0.64	0.64	0.64	0.64	**0.65**	**0.65**	0.65	0.65
eng.sdrt.stac	**0.62**	0.50	0.53	0.59	0.60	0.56	0.58	0.53	0.53	0.58	0.58	0.60	0.60	0.61	0.61	0.61	0.61	0.61	0.61
eus.rst.ert	0.42	0.40	0.40	0.45	0.45	0.46	0.46	0.37	0.37	0.43	0.43	0.47	0.47	0.41	0.41	**0.51**	**0.51**	0.45	0.45
fas.rst.prstc	0.52	0.52	0.52	0.53	0.53	0.52	0.52	0.40	0.40	0.49	0.49	0.53	0.53	0.53	0.53	**0.54**	**0.54**	0.50	0.50
fra.sdrt.annodis	0.46	0.31	0.32	0.50	0.50	0.48	0.48	0.32	0.32	0.44	0.44	0.47	0.47	0.44	0.44	**0.51**	**0.51**	**0.51**	**0.51**
ita.pdtb.luna	0.52	0.52	0.52	**0.60**	**0.60**	0.56	0.57	0.36	0.36	0.54	0.54	0.59	0.59	0.57	0.57	**0.60**	**0.60**	0.57	0.57
nld.rst.nldt	0.43	0.42	0.42	0.47	0.47	0.47	0.47	0.31	0.31	0.45	0.45	**0.49**	**0.49**	0.43	0.43	**0.49**	**0.49**	0.46	0.46
por.pdtb.crpc	0.66	0.65	0.66	0.69	0.69	0.68	0.69	0.22	0.23	0.69	0.69	0.69	0.69	0.65	0.65	**0.74**	**0.74**	0.71	0.71
***por.pdtb.tedm**	0.44	0.46	0.47	0.52	0.52	0.54	0.54	0.13	0.13	0.52	0.52	0.53	0.53	0.52	0.52	**0.59**	**0.59**	0.53	0.53
por.rst.cstn	0.57	0.56	0.56	0.63	0.64	0.62	0.63	0.30	0.30	0.60	0.60	0.62	0.62	0.59	0.59	**0.67**	**0.67**	0.62	0.62
rus.rst.rrt	0.59	0.57	0.57	0.60	0.60	0.61	0.61	0.43	0.43	0.58	0.58	0.61	0.61	0.58	0.58	**0.62**	**0.62**	0.60	0.60
spa.rst.rststb	0.56	0.48	0.48	0.63	0.63	0.59	0.59	0.37	0.37	0.56	0.56	0.61	0.61	0.58	0.58	**0.66**	**0.66**	0.63	0.63
spa.rst.sctb	0.43	0.57	0.57	0.68	0.68	0.60	0.60	0.49	0.49	0.69	0.69	0.65	0.65	0.66	0.66	**0.70**	**0.70**	0.64	0.64
tha.pdtb.tdtb	0.94	0.92	0.92	0.94	0.94	**0.96**	**0.96**	0.49	0.49	0.93	0.93	0.95	0.95	0.93	0.93	0.95	0.95	0.95	0.95
tur.pdtb.tdb	0.41	0.39	0.39	0.46	0.46	0.47	0.47	0.34	0.34	0.39	0.39	0.47	0.47	0.43	0.43	**0.52**	**0.52**	0.47	0.47
***tur.pdtb.tedm**	0.35	0.37	0.37	0.45	0.45	0.47	0.47	0.22	0.22	0.42	0.42	0.45	0.45	0.45	0.45	**0.48**	**0.48**	0.42	0.42
zho.dep.scidtb	0.55	0.51	0.53	0.63	0.65	0.59	0.60	0.41	0.43	0.62	0.62	0.64	0.64	0.62	0.62	**0.68**	**0.68**	0.68	0.68
zho.pdtb.cdtb	0.83	0.68	0.77	0.75	0.81	0.78	0.83	0.33	0.42	0.83	0.83	**0.84**	**0.84**	0.8	0.8	**0.84**	**0.84**	**0.84**	**0.84**
zho.rst.gcdt	0.60	0.52	0.52	0.58	0.58	0.56	0.56	0.40	0.40	0.58	0.58	0.59	0.59	0.59	0.59	0.60	0.60	**0.62**	**0.62**
zho.rst.sctb	0.46	0.40	0.40	0.51	0.51	0.39	0.41	0.39	0.39	0.64	0.64	0.60	0.60	0.54	0.54	**0.67**	**0.67**	0.61	0.61
AVERAGE	0.56	0.48	0.51	0.55	0.58	0.54	0.56	0.33	0.34	0.57	0.57	0.58	0.58	0.56	0.56	**0.61**	**0.61**	0.58	0.58

Table 14: Full classification results for **DistimBERT**, **mBERT**, and **XLM-RoBERTa** models with label filtering and feature augmentation. **No F.** is the accuracy score of the model before filtering and **Filt.** after filtering predicted labels per framework.