Feature-augmented model for multilingual discourse relation classification

[1]**Eleni Metheniti** and [1,2,3]**Chloé Braud** and [1,3]**Philippe Muller**
[1]UT3 - IRIT ; [2]CNRS ; [3]ANITI
firstname.lastname@irit.fr

Abstract

Discourse relation classification within a multilingual, cross-framework setting is a challenging task, and the best-performing systems so far have relied on monolingual and mono-framework approaches. In this paper, we introduce transformer-based multilingual models, trained jointly over all datasets—thus covering different languages and discourse frameworks. We demonstrate their ability to outperform single-corpus models and to overcome (to some extent) the disparity among corpora, by relying on linguistic features and generic information about the nature of the datasets. We also compare the performance of different multilingual pretrained models, as well as the encoding of the relation direction, a key component for the task. Our results on the 16 datasets of the DISRPT 2021 benchmark show improvements in accuracy in (almost) all datasets compared to the monolingual models, with at best 65.91% in average accuracy, thus corresponding to a 4% improvement over the state-of-the-art.

1 Introduction

Discourse relation classification is the process of identifying the semantic-pragmatic relations between clauses or sentences, forming the discourse structure of a document. It is considered a crucial step in building knowledge graphs (Zhang et al., 2022) and NLP downstream tasks requiring textual coherence and additional context, for example, text generation (Bosselut et al., 2018) or summarization (Xu et al., 2020), text categorization (Liu et al., 2021), and question answering (Jansen et al., 2014).

These relations, also called rhetorical relations, may be considered *explicit*, when the connection is denoted by the presence of distinct words called *connectives*, or *implicit*, i.e. relations expressed without a discourse connective. For example, the *concession* relation between the two arguments is expressed with the connective *however* in the first example below, while in the second example, the relation *manner* is implicit. Most previous studies focused on implicit discourse relation classification, which is considered a harder task than the prediction of explicit relations. However, our setting requires that the system identifies both explicit and implicit relations simultaneously, a configuration that is more realistic and includes corpora with and without annotations of explicit markers.

1. [It's best to wash adults' overalls alone, especially men's.] [*However*, it is okay to wash just a few items with them, like blue jeans.] (GUM_whow_overalls)
   Label: CONCESSION

2. [The ad would have run during the World Series tomorrow,] [replacing the debut commercial of Shearson's new ad campaign, "Leadership by Example."] (wsj_2201)
   Label: EXPANSION.MANNER

Varied typologies of discourse relations have been presented in the literature and applied to annotate several corpora in different languages. In this paper, we are presenting an approach to address multilingual, multi-framework discourse relation classification. We use as a take-off point the DISRPT Shared Task on *Discourse Relation Classification across Formalisms* and its datasets covering various languages and frameworks (Zeldes et al., 2021), and compare our results to the current state-of-the-art system on the DISRPT data, which is composed of monolingual models, DisCoDisCo (Gessler et al., 2021).

Our multilingual approach is based on joint training across all available corpora, covering varied languages and discourse frameworks. We conduct experiments over 16 corpora, covering 11 languages and 3 discourse frameworks. We jointly train a classifier with all the datasets of the Shared Task,

Proceedings of the 5th Workshop on Computational Approaches to Discourse (CODI 2024), pages 91–104

and we compare different transformer-based multilingual pretrained models. We extend the feature-based approach proposed by Gessler et al. (2021) and Gessler et al. (2022), to investigate its effect within a multilingual, cross-domain setting. Each DisCoDisCo monolingual model used different features, hence we evaluate which features are more informative in our joint setting. We also enhance our models with features targeted to our multilingual, cross-framework setting. Moreover, we test the effect of relation direction to the classification process. We examine two methods of expressing the direction of the relation between two units, either by annotating it with new tokens (Gessler et al., 2021) or by switching the position to unify it across relations (Metheniti et al., 2023). We adhere to the use of pretrained models of base size and fine-tune them for the discourse relation classification task. This ensures reproducibility, and shorter training times and computational power required.

Overall, we observe that XLM-RoBERTa models perform better than BERT models and that, contrary to the monolingual models presented in (Gessler et al., 2021), for the multilingual, cross-framework settings, using all available features is the most beneficial for all models. For the encoding of the relation position, we observe that encoding with additional tokens is more beneficial than switching the argument position, and both approaches are better than none. Finally, we report state-of-the-art performance on discourse relation classification with a maximum of 65.91% in average accuracy over all the datasets, thus outperforming previous results by about 4%. The code for fine-tuning the classifiers can be found on GitLab[1].

2 Previous Work

Most of the existing literature on discourse relation classification has focused on *implicit* relations, since explicit ones are considered easier to predict, with already accuracy above 90% with simple models and features (Pitler and Nenkova, 2009). However, it has been shown that the task can be more difficult for different domains or languages associated with small datasets (Xue et al., 2016; Scholman et al., 2021; Johannsen and Søgaard, 2013).

Approaches for **implicit relation classification** have either made use of linguistic features (Lin et al., 2009) or the least ambiguous connectives as implicit connectives (Qin et al., 2017), or even explicit connectives (Shi et al., 2017; Kurfalı and Östling, 2021). More recently, several approaches have been proposed relying on transformer-based architectures and pre-trained language models, demonstrating their effectiveness for domain transfer (Shi and Demberg, 2019), or for learning effective representation of sentences for the task (Nie et al., 2019; Sileo et al., 2019), with also attempts relying on additional pre-training of language models (Kishimoto et al., 2020).

The **DISRPT Shared Tasks** were created to motivate research on challenging discourse analysis tasks, within a multilingual, cross-framework setting, by providing unified file formats for multiple discourse datasets. There have been two editions including the task on Discourse Relation Classification (Zeldes et al., 2021; Braud et al., 2023b): since not all datasets have annotations distinguishing between explicit and implicit relations, the focus is on predicting simultaneously all types of relations. This makes for a more realistic scenario, where the nature of the relation is not assumed to be known, and it corresponds to the task performed by a discourse parser.

In DISRPT 2021 (Zeldes et al., 2021), there were two submitted systems, for 16 datasets and 11 languages. **DisCoDisCo** (Gessler et al., 2021) is a system based on monolingual and corpus-specific classifiers based on pretrained BERT language models. The inputs were enriched with handcrafted features and direction annotations (described in detail in Section 3.2). It was the most successful system, with an average 61.82% accuracy. Meanwhile, **DisCRel** (Varachkina and Pannach, 2021) aimed for a hierarchical and multilingual approach. They used sentence-level embeddings made with Sentence-BERT (Reimers and Gurevych, 2019) and stacked random forest classifiers, to predict coarse-grained relations first and then fine-grained ones. They achieved 54.23% averaged accuracy.

In DISRPT 2023 (Braud et al., 2023a), three systems were submitted. Some datasets were updated from 2021, a new framework was added (DEP, Yang and Li, 2018), and 10 new datasets and 2 new languages were added, for a total of 26 datasets and 13 languages. **HITS** (Liu et al., 2023) was the system with the best performance for 2023. It employed a combination of framework-based, multilingual, and monolingual classifiers, based on large pretrained language models. To enhance performance, they also employed bootstrap aggregat-

[1] gitlab.irit.fr/melodi/andiamo/discret_feat

ing techniques and adversarial training. The average accuracy score was 62.36% overall. When the score is calculated by including only the corpora available in 2021, the average accuracy is 58.18% (Braud et al., 2023b), thus a lower score than DisCoDisCo.[2] In **DiscReT**, we (Metheniti et al., 2023) created multilingual classifiers trained jointly on all languages and corpora. We used pretrained multilingual BERT language models (Devlin et al., 2019) and adapters (Houlsby et al., 2019). We also incorporated modifications on the label distribution to reduce the total number of labels across all corpora (see Section 3.3); however, there were problems with fully reverting the labels for the evaluation process. The average accuracy was 54.44%. **DiscoFlan** (Anuranjana, 2023) used the Flan-T5 generative language model (Chung et al., 2022) and trained monolingual models. The prompts queried the model for the relation between the two units. They post-process the model's output to match the labels of each corpus label set (see Section 3.3). Accuracy was 31.2% on average.

3 Methodology

3.1 Dataset

For the multilingual, cross-framework motivation of our experiments, we use the datasets created for the DISRPT Shared Task (Zeldes et al., 2021) for Task 3: *Discourse Relation Classification across Formalisms*.[3] We are using the datasets of the 2021 edition so that our results can be directly compared to the results of Gessler et al. (2021). These datasets are made of 16 corpora, in 11 languages, annotated in one of the following theoretical frameworks: PDTB (Penn Discourse Treebank Prasad et al., 2004), RST (Rhetorical Structure Theory, Mann and Thompson, 1988) and SDRT (Segmented Discourse Representation Theory, Asher and Lascarides, 2003). In all datasets, despite the different frameworks, discourse relations are annotated between pairs of segments that are primarily clauses or at most sentences.

3.2 DisCoDisCo augmentation methodology

Gessler et al. (2021) was the winning system of the DISRPT 2021, and compared to the results of the 2023 models on the common corpora, it is still the most successful system on the relation classification task. The submitted system is composed of multiple models; each model is a classifier fine-tuning a monolingual pretrained BERT model trained on one dataset. They use the same monolingual pretrained model for datasets of the same language but train each dataset separately. They apply two methods of feature augmentation: handcrafted features in addition to the input sequence, and annotation of the relation direction between the two units.

DisCoDisCo features Regarding the additional features of the input sequence, Gessler et al. insert manually created features as a dense embedding before the encoder. The feature vector is added between the `[CLS]` token and the input sequence tokens, and it includes sequence-level information with categorical and numerical features. Categorical features are embedded whereas numerical features are log-scaled or binned and embedded, and the feature layer is padded for the leftover dimensions.

The authors create a total of 28 features for each input sequence. These features were created by exploiting existing annotations (e.g. GENRE from the GUM corpus, SPEAKER identities from STAC corpora), by calculating them (e.g. LENGTH, DISTANCE), with the help of the syntactic parses from the DISRPT 2021 Tasks 1-2 datasets, or with external libraries (e.g. SpaCy (Honnibal et al., 2020) to eliminate stop-words for the LEXICAL OVERLAP features). The full list of these features can be found in Table 1, which includes information from Gessler et al. (2021) and the system's source code. While they generate all features for all inputs and corpora, in their submitted system for the DISRPT 2021 Shared Task, for the discourse relation classification task, they only use a few of these features for each corpus-specific model. These decisions seemed to be geared toward optimizing performance rather than being based on language, framework, or human insights; for example, only using the features of one of the units. For our experiments, we are testing both the use of all features and the use of only the "common" features that were used for at least one dataset.

Unit direction annotation Discourse relations are annotated between pairs of text segments. Some relations can be directed, meaning that the order of the arguments of the relation is meaningful. This feature depends on the way relations are encoded,

[2]Note that the comparison is inequitable, because there have been changes in some corpora, e.g. English GUM.

[3]The datasets and their statistics can be found in github.com/disrpt/sharedtask2021.

Feature in JSON	Feature	Type	Example	Description	Used
nuc_children	Nucleus' Children	Num.	2	No. of discourse units in Unit 1	5
sat_children	Satellite's Children	Num.	2	No. of discourse units in Unit 2	8
genre	Genre	Cat.	reddit	Genre of a document (where available)	5
u1_discontinuous	Discontinuous	Cat.	`True`	Whether Unit 1's tokens are not all contiguous in the text	3
u2_discontinuous	Discontinuous	Cat.	`True`	Whether Unit 2's tokens are not all contiguous in the text	5
u1_issent	Is Sentence	Cat.	`True`	Whether Unit 1 is a whole sentence	3
u2_issent	Is Sentence	Cat.	`True`	Whether Unit 2 is a whole sentence	5
u1_length	Length	Num.	9	Length of Unit 1, in tokens	-
u2_length	Length	Num.	13	Length of Unit 2, in tokens	-
length_ratio	Length Ratio	Num.	0.3	Ratio of unit 1 and unit 2's token lengths	3
u1_speaker	Name of Speaker 1	Cat.	Rainbow	Name of Speaker (available only for STAC)	-
u2_speaker	Name of Speaker 2	Cat.	Markus	Name of Speaker (available only for STAC)	-
same_speaker	Same Speaker	Cat.	`True`	Whether the same speaker produced Unit 1 and Unit 2	2
u1_func	Unit Function	Cat.	root	Universal Dependencies Relation of Unit 1's Head to the Head of the input sequence	1
u1_pos	Part of speech & Morphological Tag	Cat.	VBN	Part of speech & Morphological tag of the Unit 1's Head	-
u1_depdir	Universal Part of speech Tag	Cat.	ROOT	Part of speech of the Unit 1's Head wrt. the Head of the input sequence	8
u2_func	Unit Function	Cat.	advcl	Universal Dependencies Relation of Unit 2's Head to the Head of the input sequence	8
u2_pos	Part of speech & Morphological Tag	Cat.	VB	Part of speech & Morphological tag of the Unit 2's Head	8
u2_depdir	Universal Part of speech Tag	Cat.	LEFT	Part of speech of the Unit 2's Head wrt. the Head of the input sequence	7
doclen	Document Length	Num.	214	Length of the document, in tokens	-
u1_position	Position	Num.	0.4	Position of Unit 1 in the document, between 0.0 and 1.0	9
u2_position	Position	Num.	0.4	Position of Unit 2 in the document, between 0.0 and 1.0	-
percent_distance	Percent of distance	Num.	0.05	No. of discourse units between Unit 1 and Unit 2 divided by sequence length	-
distance	Distance	Num.	7	No. of other discourse units between Unit 1 and Unit 2	9
lex_overlap_words	Lexical Overlap	Cat.	assets sold	List of overlapping non-stoplist words in Unit 1 and Unit 2	-
lex_overlap_length	Lexical Overlap	Num.	3	No. of overlapping non-stoplist words in Unit 1 and Unit 2	1
unit1_case	Uppercased letter	Cat.	cap_initial	Whether the unit starts with a capital letter or not	1
unit2_case	Uppercased letter	Cat.	other	Whether the unit starts with a capital letter or not	1

Table 1: List of all features generated by the DisCoDisCo system, in the preprocessing stage, with descriptions. "Type" refers to whether the feature is categorical or numerical. With "No. Used" we note how many corpus-specific DisCoDisCo models used said feature (out of 16 models in total).

we could have different labels with the arguments following the order of the text (e.g. *cause vs result*), or one unique label where the first argument has always the same role compared to the second regarding the semantics of the relation. All existing studies focusing on discourse relation identification consider this information as given: they present to the learning model the arguments in the order given by the annotation, thus first, then second argument of the relation. It is not the case when one performs full discourse parsing: the parser knows that two segments are attached, but not in which order, and the segments are presented in the order of the text. In order to better understand this important aspect of the task, we investigate different encodings of this information within a transformer architecture.

In the DISRPT datasets, the pairs of segments are presented in the linear order of the text, but an additional column indicates the order of the arguments for the annotated relation. Gessler et al. introduced two pseudo-tokens (not as BERT special tokens) in order to encode the direction between the two units:

- If the direction of the relation follows the linear order of the text, a case annotated as (1>2) in DISRPT data, the } token is added after the `[CLS]` token and before Unit 1 and the > token before Unit 2.
- If the direction of the relation is reversed, a case annotated as (1<2) in DISRPT data, the < token is added after Unit 1 and the { token after Unit 2.

3.3 Proposed additional augmentation

Corpus-specific features Previous approaches to training multilingual, cross-framework classifiers with all corpora and languages reported results

lower than monolingual systems. We assumed that one issue was the lack of guidance of the model, where it was hard for the model to make correlations between datasets. In order to tackle this issue, we add at the start of each sequence some additional tokens that characterize the dataset and should help the model to link samples from the same language or framework. We add as additional tokens, after the `[CLS]` token and before the input sequence tokens, the following tokens:

- Language: the language of the corpus in English (e.g. English, French, German, etc.);[4]
- Corpus: the name of the dataset in the DISRPT 2021 data (e.g. `deu.rst.pcc`, `eng.rst.rstdt`, `fra.sdrt.annodis`, etc);
- Framework: the framework name (e.g. rst, pdtb, sdrt, dep).

Feature embedding as tokens Instead of creating a dense embedding as Gessler et al. did, we are adding the additional features in the input sequence as tokens. Each feature value (numerical, categorical, and Boolean) is added to the vocabulary, in order not to be split into subwords by the tokenizer. Only the value of the feature is added, not its key, to not create an excessive amount of new tokens (e.g. all numbers encoded separately for each numerical feature). For example, the new token `0.1` does not refer to a number in the text but may refer to the feature **u1_position** or **length_ratio**, depending on its order in the input sequence. This extends the size of the vocabulary and, therefore, extends the size of the token embedding matrix of the model to match the embedding matrix of the tokenizer. This technique stays close to the process of concatenating the feature vector with the token vector while assuring reproducibility with the HuggingFace models (Wolf et al., 2020).

Unit direction unification In addition to implementing the relation direction annotation of the DisCoDisCo system (i.e. additional tokens), we are also testing the effectiveness of unifying the direction by switching unit positions. In Metheniti et al. (2023), we proposed to reorder the two units in the input sequence, to follow the order of the arguments of the relation, instead of the linear order of the text as encoded in DISRPT files. If the arguments are in the same order as in the text (1>2), then the input is unchanged, but if they are in reverse order (1<2), the units have their position switched in the input sequence of the model.

Label merging The joint training set of the 16 corpora of the DISRPT 2021 Shared Task contains 126 labels, making for a complex learning problem. These labels come from three different annotation frameworks, and sometimes overlap; for example, the labels *Expansion.Correction* in `tur.pdtb.tdb` (Turkish, PDTB) and *correction* in `eng.sdrt.stac` (English, SDRT) point to the same relation. Suggestions for unified label sets are limited to specific frameworks or do not cover all relations present in corpora (Benamara and Taboada, 2015; Braud et al., 2017; Varachkina and Pannach, 2021). We adapt the label harmonization that we proposed for the DISRPT 2023 Shared Task datasets, which implements minimal substitutions to less-frequent labels, and lower-casing (Metheniti et al., 2023). The number of our labels was reduced from 126 originally to 102 labels.

Label Filtering Multilingual, multi-framework classification models provide a probability distribution of every label included in the training set, regardless of the target language and framework. We took inspiration from the strategy of Anuranjana (2023) who addressed the problem of generating annotations that may not match the labels of the training set by filtering the output of their generative model so that it converts them to existing labels. We are also post-processing our classification model label outputs, and we keep in the predictions only labels coming from the target corpus' framework. Thus a label that is present in the combined training corpus but not in the target framework label set will not be returned, even if it were assigned a higher probability by the model.

3.4 Classification models

We fine-tune multilingual classifiers built on pretrained multilingual transformer-based models. Fine-tuning is performed with all training sets of all languages and datasets jointly, while evaluation is performed on the evaluation and test sets of each dataset individually. We used PyTorch (Paszke et al., 2019) and the HuggingFace libraries to build our classifiers, with

[4]Preliminary experiments with the language token in the corpus' original language (e.g. English, Français, Deutsch, etc.) showed the same performance as with the language token in English since the models we are using contain multilingual embeddings.

Model	DisCoDisCo 2021 (BERT)	mBERT	DistilmBERT	XLM	mBERT	DistilmBERT	XLM
Relation direction	**Add. tokens**	**Add. tokens**			**Switching units**		
No features	60.41	59.54	56.81	62.09	58.36	55.69	60.52
Common DisCoDisCo features	61.82	62.56	60.92	64.86	59.75	57.24	61.14
All DisCoDisCo features	-	63.09	60.28	64.50	62.33	59.08	**63.95**
Language, Corpus, Framework (LCF)	-	61.76	59.17	64.13	58.34	55.69	60.52
LCF + Common	-	63.46	**62.01**	**65.91**	61.12	57.75	62.88
LCF + All	-	**63.67**	61.92	65.53	**63.89**	**59.65**	63.51

Table 2: Average accuracies of the models, reported on the test set. We report the results of the DisCoDisCo system with individual models trained with or without their specific features and the DisCoDisCo relation annotation. For our multilingual models, we report models trained with the DisCoDisCo direction annotation ("Add. tokens") or the DiscReT direction normalization ("Switching units"). The models were trained with different sets of features or without. In bold are the best scores for each column, so for model and direction fixed.

the models: `bert-base-multilingual-cased`,[5] `distilbert-base-multilingual-cased`,[6] and `xlm-roberta-base`.[7] Each classification model is trained for 10 epochs, keeping the best result out of the 10 epochs, based on the development set. The fine-tuning process for these models, per epoch, was around 1 hour for DistilmBERT, 2 hours for mBERT, and 2 hours 10 minutes for XLM-RoBERTa, on a GPU cluster with 4 Nvidia Geforce GTX 1080TI graphics cards.

Multilingual BERT (mBERT) (Devlin et al., 2019) is a pretrained model based on BERT. It has been trained on Wikipedia data of the top 104 languages, with masked language modeling (MLM) and next-sentence prediction objectives. The base and cased version of the model contains 12 layers, 12 heads, and 177M parameters. We selected it, in order to compare it with the DisCoDisCo models that were built on monolingual BERT-base architectures. DistilmBERT (Sanh et al., 2019) is a multilingual distilled version of mBERT with the same training set and objectives. The base and cased model has 6 layers, 12 heads, and 134M parameters. As a lighter version of BERT, it would be interesting to compare a BERT-based model with fewer parameters. XLM-RoBERTa (Conneau et al., 2020) is a multilingual pretrained model based on RoBERTa. It is pretrained on 2.5TB of filtered CommonCrawl data in 100 languages. The base version of the model has 12 layers and 279M parameters. RoBERTa models have outperformed BERT in several datasets in the Shared Task (Liu et al., 2023), therefore we decided to include them in our experiments.

[5] huggingface.co/bert-base-multilingual-cased
[6] huggingface.co/distilbert-base-multilingual-cased
[7] huggingface.co/xlm-roberta-base

4 Results

In Table 2 we present the average accuracy for all the multilingual classification models we trained, with different pretrained models, with different combinations of features, and with different handling of relation annotation. We report the results of DisCoDisCo (Gessler et al., 2021) in the second column, and the results obtained by our system in the others. The second row indicates how the direction of the relation is encoded, based on unit direction annotation ("Add. tokens") as in Gessler et al. or by unit direction modification ("Switching units") as in Metheniti et al. In the Appendix, the results for individual test sets can be found: in Table 4 for models trained with features and direction annotation based on additional tokens, in Table 5 for models trained with features and direction unification based on switching units, and in Table 6 for models trained without features, with different direction handling (including no encoding of the direction at all).

Overall, our models outperform the state-of-the-art system DisCoDisCo in several settings, when linguistic features (i.e. "Common/All DisCoDisCo features") and/or dataset information ("LCF", Language, Corpus, Framework) are used, with at best 65.91% in average accuracy, against 61.82% for DisCoDisCo. This demonstrates that single multilingual, cross-framework models are able to leverage correlations between the different datasets, and thus take advantage of a larger amount of data if fed with additional information.

For the models most similar to DisCoDisCo, i.e. using mBERT with annotations of direction ("Add. Tokens"), our results are very close to theirs: 60.41% *vs* 59.54% ("No features") and 61.82% *vs* 62.56% ("Common features"). XLM-RoBERTa models performed better than the mBERT-based models and also surpassed the lightweight DistilmBERT models. They were the ones that steadily surpassed the DisCoDisCo system, with 62.09% and 64.86% respectively for the same configurations. Moreover, the mBERT models also performed better than the DisCoDisCo baseline, when they were provided with the LCF tokens, either when limited to "Common features" (63.46%), or when using "All features" (63.67%).

Observing the different sets of features that we used, the addition of any features improves the accuracy of multilingual classification, and the best configuration was, in most cases, the features used by Gessler et al. (2021), with the addition of corpus-specific features. When we used additional tokens to encode the direction of the relation, the most beneficial set of features for all the models was the "common" features, i.e. only the features used by at least one model in the DisCoDisCo 2021 system. We notice that the model with the highest accuracy of all is the XLM-RoBERTa model, using this encoding of the direction and the common DisCoDisCo and LCF features. However, using all features in this setting leads to very similar results (-0.4%). When the direction is encoded by switching the units, the situation is reversed: results are better when using all the features rather than only the common ones. The addition of the LCF tokens, alongside the DisCoDisCo features, showed an increase in accuracy as well. For the XLM-RoBERTa models, the presence of all features was also most beneficial, but not necessarily the presence of the LCF features.

Looking at individual datasets, our models outperformed the DisCoDisCo 2021 system in all but one dataset, the Basque `eus.rst.ert` (by 0.15%, Table 4). For some datasets, the improvement was significant (with XLM-RoBERTa models), for example up to 15.72% for `spa.rst.sctb` (Spanish) and over 8% for `fas.rst.prstc` (Farsi) and `zho.rst.sctb` (Chinese). The mBERT models trailed not far behind the XLM-RoBERTa ones, however, there was an instance where an mBERT model was more successful, mBERT with all DisCoDisCo and LCF features for `fra.sdrt.annodis` (French). Also, models with all features were more successful for the French, Portuguese, and Spanish datasets.

Comparing the performance between the two ways of handling the direction of the relation, the direction annotation based on additional tokens was the better option for most datasets when the DisCoDisCo features were used. However, for the Dutch `nld.rst.nldt` and Portuguese `por.rst.cstn` datasets, the performance was identical with either setting. Observing the effect of the direction handling without the addition of features, we note that, while the method based on additional tokens performed better overall, there were instances where switching the arguments was better (English `eng.rst.rstdt`, `eng.sdrt.stac`, Spanish `spa.rst.rststb`), and one dataset for which no change was marginally better (Portuguese `por.rst.cstn`), see Tables 4, 5 and 6 in Appendix.

5 Discussion

Our multilingual approach outperformed the monolingual approach of DisCoDisCo (Gessler et al., 2021) in all but one dataset, the Basque one. Our initial assumption was that the use of the multilingual setting would be beneficial since the use of more data is favorable to the models and the instances of less frequent labels would be higher. Indeed, for the very small datasets `spa.rst.sctb` (Spanish, 326 train sentences) and `zho.rst.sctb` (Chinese, 361 train sentences), the improvement was elevated with all models. For the largest datasets (`eng.pdtb.pdtb`, English, 44.5K train sentences; `rus.rst.rrt`, Russian, 19K train sentences; `tur.pdtb.tdb`, Turkish, 25K train sentences) there was also an improvement of $3-4\%$. In the case of `eus.rst.ert` (Basque) with 1.6K train sentences, we observed the label distribution; it has 25 unique labels and similar distributions to `spa.rst.rststb`, `spa.rst.sctb`, and `zho.rst.sctb`. We observe the classification report results (Pedregosa et al., 2011) for the most successful model in Table 3. In the 2021 edition of the data, there were a few labels in this dataset with misspellings (*motibation* instead of *motivation*), which were corrected in the 2023 edition. These labels were not changed by the DiscReT mappings and were not corrected in order to stay true to the 2021 data. Even with these errors, however, this is not the smallest, most complex, or relation-rich dataset. Therefore, the failure of the Basque dataset may be related to the language's typological dif-

	precision	recall	f1-score	support
motibation	0	0	0	2
summary	0	0	0	3
concession	0.58	0.65	0.61	17
purpose	0.87	0.80	0.83	50
joint	0	0	0	1
causation	0.51	0.51	0.51	37
interpretation	0.50	0.08	0.13	13
circumstance	0.76	0.67	0.71	48
expansion.conjunction	0.28	0.48	0.35	25
unconditional	0.50	0.25	0.33	4
evaluation	0.30	0.56	0.39	16
anthitesis	0	0	0	5
unless	0.59	0.62	0.60	21
solution-hood	0	0	0	8
result	0.50	0.53	0.51	34
background	0.52	0.76	0.62	29
means	0.69	0.68	0.68	37
conditional	1.00	0.33	0.50	9
preparation	0.90	0.85	0.87	73
elaboration	0.66	0.69	0.67	140
list	0.63	0.44	0.52	54
evidence	0.40	0.25	0.31	8
sequence	0.37	0.48	0.42	23
justify	0.42	0.62	0.50	8
restatement	0.60	0.23	0.33	13
accuracy			0.60	678
macro avg	0.41	0.37	0.37	678
weighted avg	0.62	0.60	0.60	678

Table 3: Classification report for the `eus.rst.ert` (Basque) test set, with the XLM-RoBERTa model with Common and LCF features (epoch 8).

ference from the rest, as it benefits less from the multilingual pretrained language models.

Comparing the use of different models, we observe that the XLM-RoBERTa base models are more successful, probably because of their larger number of parameters. For the original DisCoDisCo model, the use of BERT-based models was obligatory for most languages, as at the time there were fewer options available, especially for less common languages. The mBERT base models were not far less successful than the XLM-RoBERTa, with the help of features. The DistilmBERT models are far too optimized and lightweight, missing parameters that were, as is shown, necessary for the classification process.

Overall, the addition of features, even as simple as additional tokens in the input sequence, improved classification accuracy significantly. In the monolingual setting, it was possible to test different feature sets to configure which was the best, but for the multilingual setting, selecting features is not straightforward, as different corpora contain different annotations (e.g. the GUM and STAC corpora are the only ones with the SPEAKER information). The small differences in accuracy between using all features and only the ones used for the DisCoDisCo 2021 system are produced because, in a multilingual setting with all the datasets used jointly, some features that are informative for some corpora will not be for others, if the annotation does not exist. The addition of the language, framework, and corpus name was also beneficial, in order to annotate the presence of corpus-specific features, even if the information of language is not directly accessible to the model.

Finally, regarding the relation direction, human intuition is different than the way models process input. The proposal to unify all relation directions by switching the arguments (Metheniti et al., 2023) sounds beneficial in theory, especially when the same relation can be initiated in either unit. However, transformer-based models are not necessarily sensitive to word order; even though positional information is injected in them, some research suggests that they are not sensitive to permutations (Pham et al., 2021; Gupta et al., 2021). However, other research supports that not all permutations are processed equally (Sinha et al., 2021) and that the models learn structural information (Wang and Chen, 2020; Papadimitriou et al., 2022). It is, therefore, understandable that the presence of additional tokens noting the direction as in (Gessler et al., 2021) may communicate more information about the relation direction to the models, than switching unit positions. However, there was also a smaller improvement with the unification of the direction; this points to the models either being capable of constructing a rudimentary structure of the two arguments or the models not being completely insensible to word order.

6 Conclusion

In this paper, we reprised the DISRPT 2021 Shared Task on Relation Classification across Formalisms and revisited the most successful model of the last two editions, DisCoDisCo (Gessler et al., 2021). We adapted DisCoDisCo methodologies to multilingual relation classification models, with the addition of techniques and suggestions from other participating teams of the 2023 edition (Metheniti et al., 2023; Anuranjana, 2023).

We found that XLM-RoBERTa models outperform BERT models, in the multilingual setting, especially with the presence of DisCoDisCo's handcrafted features. The most successful model was

trained only with the features used by DisCoDisCo models, as opposed to all features created in the preprocessing stage—but this success was only marginal to the use of all features, and was not true for all architectures. The addition of corpus-specific tokens (language, corpus name, framework) was also beneficial in the multilingual setting. Finally, annotating the relation direction with additional tokens was more successful than unifying the position of the two arguments, due to the make of transformer-based models. It should be noted that this information proved crucial and that further studies are needed on this aspect, in particular on the possibility of predicting direction and on the heterogeneity of existing corpora with regard to its encoding.

As a future direction, we are considering using our approach on the updated DISRPT 2023 benchmark, which includes modified corpora, additional corpora in more languages, and some small validation datasets that allow for testing out-of-domain performance.

Acknowledgements

This work is supported by the AnDiaMO project (ANR-21-CE23-0020). Our work has benefited from the AI Interdisciplinary Institute ANITI. ANITI is funded by the French "Investing for the Future – PIA3" program under the Grant agreement n°ANR-19-PI3A-0004. This work is also partially supported by the SLANT project (ANR-19-CE23-0025). Chloé Braud and Philippe Muller are part of the program DesCartes and are also supported by the National Research Foundation, Prime Minister's Office, Singapore under its Campus for Research Excellence and Technological Enterprise (CREATE) program. The work was also supported by the ANR grant SUMM-RE (ANR-20-CE23-0017).

References

Kaveri Anuranjana. 2023. DiscoFlan: Instruction fine-tuning and refined text generation for discourse relation label classification. In *Proceedings of the 3rd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2023)*, pages 22–28, Toronto, Canada. The Association for Computational Linguistics.

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

Farah Benamara and Maite Taboada. 2015. Mapping different rhetorical relation annotations: A proposal. In *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics*, pages 147–152, Denver, Colorado. Association for Computational Linguistics.

Antoine Bosselut, Asli Celikyilmaz, Xiaodong He, Jianfeng Gao, Po-Sen Huang, and Yejin Choi. 2018. Discourse-aware neural rewards for coherent text generation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 173–184, New Orleans, Louisiana. Association for Computational Linguistics.

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. 2023a. 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.

Chloé Braud, Yang Janet Liu, Eleni Metheniti, Philippe Muller, Laura Rivière, Attapol Rutherford, and Amir Zeldes, editors. 2023b. *Proceedings of the 3rd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2023)*. The Association for Computational Linguistics, Toronto, Canada.

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.

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.

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.

Luke Gessler, Lauren Levine, and Amir Zeldes. 2022. Midas loop: A prioritized human-in-the-loop annotation for large scale multilayer data. In *Proceedings of the 16th Linguistic Annotation Workshop (LAW-XVI) within LREC2022*, pages 103–110, Marseille, France. European Language Resources Association.

Ashim Gupta, Giorgi Kvernadze, and Vivek Srikumar. 2021. Bert & family eat word salad: Experiments with text understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12946–12954.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python. Technical report, Zenodo.

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.

Peter Jansen, Mihai Surdeanu, and Peter Clark. 2014. Discourse complements lexical semantics for non-factoid answer reranking. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 977–986, Baltimore, Maryland. Association for Computational Linguistics.

Anders Johannsen and Anders Søgaard. 2013. Disambiguating explicit discourse connectives without oracles. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 997–1001, Nagoya, Japan. Asian Federation of Natural Language Processing.

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.

Murathan Kurfalı and Robert Östling. 2021. Let's be explicit about that: Distant supervision for implicit discourse relation classification via connective prediction. In *Proceedings of the 1st Workshop on Understanding Implicit and Underspecified Language*, pages 1–10, Online. Association for Computational Linguistics.

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.

Xin Liu, Jiefu Ou, Yangqiu Song, and Xin Jiang. 2021. Exploring discourse structures for argument impact classification. 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 3958–3969, 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.

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.

Allen Nie, Erin Bennett, and Noah Goodman. 2019. DisSent: Learning sentence representations from explicit discourse relations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4497–4510, Florence, Italy. Association for Computational Linguistics.

Isabel Papadimitriou, Richard Futrell, and Kyle Mahowald. 2022. When classifying arguments, BERT doesn't care about word order...except when it matters. In *Proceedings of the Society for Computation in Linguistics 2022*, pages 203–205, online. Association for Computational Linguistics.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in

Python. *Journal of Machine Learning Research*, 12:2825–2830.

Thang Pham, Trung Bui, Long Mai, and Anh Nguyen. 2021. Out of order: How important is the sequential order of words in a sentence in natural language understanding tasks? In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1145–1160, Online. Association for Computational Linguistics.

Emily Pitler and Ani Nenkova. 2009. Using syntax to disambiguate explicit discourse connectives in text. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 13–16, Suntec, Singapore. Association for Computational Linguistics.

Rashmi Prasad, Eleni Miltsakaki, Aravind Joshi, and Bonnie Webber. 2004. Annotation and data mining of the Penn Discourse TreeBank. In *Proceedings of the Workshop on Discourse Annotation*, pages 88–95, Barcelona, Spain. Association for Computational Linguistics.

Lianhui Qin, Zhisong Zhang, Hai Zhao, Zhiting Hu, and Eric Xing. 2017. Adversarial connective-exploiting networks for implicit discourse relation classification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1006–1017, Vancouver, Canada. Association for Computational Linguistics.

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.

Merel Scholman, Tianai Dong, Frances Yung, and Vera Demberg. 2021. Comparison of methods for explicit discourse connective identification across various domains. In *Proceedings of the 2nd Workshop on Computational Approaches to Discourse*, pages 95–106, Punta Cana, Dominican Republic and Online. Association for Computational Linguistics.

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

Wei Shi, Frances Yung, Raphael Rubino, and Vera Demberg. 2017. Using explicit discourse connectives in translation for implicit discourse relation classification. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 484–495, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Damien Sileo, Tim Van De Cruys, Camille Pradel, and Philippe Muller. 2019. Mining discourse markers for unsupervised sentence representation learning. 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 3477–3486, Minneapolis, Minnesota. Association for Computational Linguistics.

Koustuv Sinha, Prasanna Parthasarathi, Joelle Pineau, and Adina Williams. 2021. UnNatural Language Inference. 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 7329–7346, Online. Association for Computational Linguistics.

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.

Yu-An Wang and Yun-Nung Chen. 2020. What do position embeddings learn? an empirical study of pre-trained language model positional encoding. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6840–6849, Online. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. 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, 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.

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.

Hongming Zhang, Xin Liu, Haojie Pan, Haowen Ke, Jiefu Ou, Tianqing Fang, and Yangqiu Song. 2022. ASER: Towards large-scale commonsense knowledge acquisition via higher-order selectional preference over eventualities. *Artificial Intelligence*, 309:103740.

A Appendix: Full classification results

Model	Control	m	d	x	m	d	x	m	d	x	m	d	x	m	d	x
Features		**Common Features**			**All Features**			**LCF**			**LCF + Common Features**			**LCF + All Features**		
deu.rst.pcc	39.23	41.54	38.46	45.00	42.69	37.69	**48.85**	36.92	31.15	40.00	41.92	41.15	44.23	40.38	39.62	43.08
eng.pdtb.pdtb	74.44	73.64	70.71	74.48	73.68	70.14	74.97	74.61	72.66	76.83	74.48	72.18	**77.45**	74.39	72.35	76.25
eng.rst.gum	66.76	66.43	64.61	68.15	67.29	65.76	68.05	63.18	61.98	67.43	67.34	65.57	67.96	67.05	65.04	**68.87**
eng.rst.rstdt	67.10	61.48	59.49	63.48	58.75	56.52	59.95	68.26	67.66	69.98	69.10	68.17	**70.44**	69.88	67.94	68.96
eng.sdrt.stac	65.03	62.78	63.18	65.83	64.04	63.18	65.23	58.81	59.40	62.25	63.58	61.92	**66.16**	62.85	61.92	65.50
eus.rst.ert	**60.62**	56.93	56.49	57.23	58.26	56.78	56.93	55.31	54.28	57.37	57.23	58.55	60.47	56.78	56.49	57.82
fas.rst.prstc	52.53	58.11	56.08	60.64	57.26	56.08	59.97	55.91	53.72	59.80	58.11	56.25	**60.98**	58.45	55.07	58.11
fra.sdrt.annodis	46.40	50.56	45.44	48.96	**51.52**	44.16	47.20	48.80	43.20	49.28	49.28	44.48	49.28	51.36	44.48	47.84
nld.rst.nldt	55.21	51.84	51.53	57.67	53.07	52.45	**58.59**	48.16	46.01	56.75	54.29	49.39	**58.59**	54.60	51.84	57.06
por.rst.cstn	64.34	67.28	68.01	**69.85**	68.75	66.54	68.38	68.38	63.97	68.38	68.75	69.12	69.49	**69.85**	66.54	68.75
rus.rst.rrt	66.44	68.91	67.25	71.02	68.73	66.48	**71.30**	65.04	63.74	67.71	68.66	67.39	71.19	69.47	68.34	70.59
spa.rst.rststb	54.23	55.16	51.64	57.75	55.4	51.64	54.93	53.99	50.47	56.57	56.81	54.69	55.16	56.81	54.93	**59.39**
spa.rst.sctb	66.04	71.70	75.47	76.10	75.47	72.33	75.47	74.84	74.84	74.84	73.58	78.62	78.62	73.58	79.25	**81.76**
tur.pdtb.tdb	60.09	58.53	54.27	62.80	58.77	54.27	62.32	57.11	54.50	63.51	57.35	55.69	62.80	56.87	57.58	**64.22**
zho.pdtb.cdtb	86.49	87.47	85.36	**89.58**	87.86	85.62	88.79	87.73	84.30	88.65	88.13	84.83	89.45	88.52	85.22	88.52
zho.rst.sctb	64.15	68.55	66.67	69.18	67.92	64.78	71.07	71.07	64.78	66.67	66.67	64.15	**72.33**	67.92	64.15	71.70
AVERAGE	61.82	62.56	60.92	64.86	63.09	60.28	64.50	61.76	59.17	64.13	63.46	62.01	**65.91**	63.67	61.92	65.53

Table 4: Results of models with features and direction normalization based on additional tokens as in Gessler et al. (2021), for all datasets. The models are: DisCoDisCo 2021 System with features (Control), mBERT (m), DistilmBERT (d), and XLM-RoBERTa (x).

Model	Control	m	d	x	m	d	x	m	d	x	m	d	x	m	d	x
Features		**Common Features**			**All Features**			**LCF**			**LCF + Common Features**			**LCF + All Features**		
deu.rst.pcc	39.23	33.08	33.46	41.92	39.62	35.77	43.08	31.92	26.15	35.00	37.31	33.08	40.77	40.38	35.77	**43.46**
eng.pdtb.pdtb	74.44	70.98	68.37	71.78	72.66	69.34	73.37	72.44	70.05	73.90	73.55	70.45	74.52	75.01	71.42	**75.45**
eng.rst.gum	66.76	63.46	61.65	64.04	**67.53**	61.65	67.38	58.54	56.10	61.50	64.71	62.41	64.99	66.38	63.89	66.52
eng.rst.rstdt	67.10	60.56	59.54	60.70	59.63	58.42	61.21	65.89	63.62	66.73	67.80	66.87	67.89	68.82	67.70	**69.28**
eng.sdrt.stac	65.03	64.17	62.58	66.62	64.37	62.32	**67.09**	59.54	58.68	61.79	63.05	62.78	64.83	64.17	62.45	66.49
eus.rst.ert	**60.62**	53.54	52.65	53.98	57.52	56.19	57.82	50.44	45.43	51.62	54.57	51.18	53.39	59.59	53.39	56.49
fas.rst.prstc	52.53	53.21	51.18	55.24	57.43	53.38	**59.12**	50.68	49.16	53.89	53.38	51.18	56.93	58.95	55.24	57.60
fra.sdrt.annodis	46.40	48.16	42.72	47.84	48.64	40.96	48.16	47.84	43.20	48.48	**49.44**	42.24	46.88	48.96	38.24	45.28
nld.rst.nldt	55.21	50.92	45.40	51.53	53.37	48.77	**58.90**	45.40	41.72	51.23	51.23	43.87	55.21	55.21	51.53	53.68
por.rst.cstn	64.34	68.01	65.44	68.38	68.75	66.54	**70.59**	66.54	67.28	68.38	66.91	64.71	68.01	67.28	63.60	67.65
rus.rst.rrt	66.44	66.51	64.69	67.82	68.87	67.18	**70.10**	62.26	58.78	63.00	66.41	63.74	66.94	68.62	66.76	69.26
spa.rst.rststb	54.23	54.93	52.82	55.16	54.93	51.41	56.34	54.69	50.94	53.99	55.63	51.41	55.87	**56.57**	53.05	55.40
spa.rst.sctb	66.04	71.07	66.67	71.07	72.96	71.07	72.33	69.18	67.30	72.33	74.21	70.44	**79.25**	78.62	69.81	74.84
tur.pdtb.tdb	60.09	50.24	48.34	54.74	56.64	54.03	**61.61**	50.24	49.76	57.11	50.00	49.76	56.87	57.58	55.45	60.19
zho.pdtb.cdtb	86.49	84.30	83.11	85.22	86.41	84.70	87.60	84.30	81.93	86.54	84.96	83.25	85.88	**87.60**	83.91	87.20
zho.rst.sctb	64.15	62.89	57.23	62.26	67.92	63.52	**68.55**	63.52	61.01	62.89	64.78	56.60	67.92	68.55	62.26	67.30
AVERAGE	61.82	59.75	57.24	61.14	62.33	59.08	**63.95**	58.34	55.69	60.52	61.12	57.75	62.88	63.89	59.65	63.51

Table 5: Results of models with features and direction normalization based on switching units as in Metheniti et al. (2023), for all datasets. The models are: DisCoDisCo 2021 System with features (Control), mBERT (m), DistilmBERT (d), and XLM-RoBERTa (x).

Model	Control	m	d	x	m	d	x	m	d	x
Direction		**No change**			**Switching units**			**Add. tokens**		
deu.rst.pcc	33.85	31.15	28.46	35.77	31.92	26.15	35.00	38.46	33.08	**42.31**
eng.pdtb.pdtb	**75.63**	65.22	63.89	68.68	71.95	70.05	73.90	72.35	69.87	73.99
eng.rst.gum	**62.65**	51.08	47.97	54.95	60.21	56.10	61.50	57.29	53.18	60.26
eng.rst.rstdt	66.45	49.42	48.40	50.95	64.73	63.62	**66.73**	52.44	51.14	55.45
eng.sdrt.stac	59.67	53.64	53.58	57.28	57.62	58.68	**61.79**	54.70	55.30	57.62
eus.rst.ert	**59.59**	49.85	46.31	50.44	50.74	45.43	51.62	57.52	51.03	57.08
fas.rst.prstc	51.18	51.86	48.82	54.90	50.84	49.16	53.89	56.42	53.38	**58.45**
fra.sdrt.annodis	48.32	48.64	42.88	48.80	47.68	43.20	48.48	**49.28**	44.16	48.80
nld.rst.nldt	52.15	47.55	42.33	51.84	45.40	41.72	51.23	48.16	46.01	**57.98**
por.rst.cstn	67.28	66.18	64.71	**69.49**	68.01	67.28	68.38	67.65	64.34	69.12
rus.rst.rrt	65.46	59.69	57.72	62.50	62.29	58.78	63.00	65.67	63.67	**67.39**
spa.rst.rststb	54.23	52.82	51.17	51.41	52.35	50.94	53.99	53.99	50.47	**57.04**
spa.rst.sctb	61.01	60.38	63.52	59.75	71.70	67.30	**72.33**	69.81	71.07	71.07
tur.pdtb.tdb	57.58	51.66	47.87	59.48	50.71	49.76	57.11	58.06	53.79	**61.37**
zho.pdtb.cdtb	87.34	80.87	78.89	82.85	82.85	81.93	86.54	84.17	84.30	**88.13**
zho.rst.sctb	64.15	56.6	49.69	55.35	64.78	61.01	62.89	66.67	64.15	**67.30**
AVERAGE	60.41	54.79	52.26	57.15	58.36	55.69	60.52	59.54	56.81	**62.09**

Table 6: Results of models without features and different direction handling, for all datasets. The models are: DisCoDisCo 2021 System with features (Control), mBERT (m), DistilmBERT (d), and XLM-RoBERTa (x).