A BERT-Based Approach for Multilingual Discourse Connective Detection

Thomas Chapados Muermans and Leila Kosseim

Computational Linguistics at Concordia (CLaC) Laboratory
Department of Computer Science and Software Engineering
Concordia University, Montréal, Québec, Canada
thomas.chapadosmuermans@mail.concordia.ca
leila.kosseim@concordia.ca

Abstract. In this paper, we report on our experiments towards multilingual discourse connective (or DC) identification and show how language specific BERT models seem to be sufficient even with little task-specific training data. While some languages have large corpora with human annotated DCs, most languages are low in such resources. Hence, relying solely on discourse annotated corpora to train a DC identification system for low resourced languages is insufficient. To address this issue, we developed a model based on pretrained BERT and fine-tuned it with discourse annotated data of varying sizes. To measure the effect of larger training data, we induced synthetic training corpora with DC annotations using word-aligned parallel corpora. We evaluated our models on 3 languages: English, Turkish and Mandarin Chinese in the context of the recent DISRPT 2021 Task 2 shared task. Results show that the F-measure achieved by the standard BERT model (92.49%, 93.97%, 87.47% for English, Turkish and Chinese) is hard to improve upon even with larger task specific training corpora

Keywords: Discourse Analysis · Multilingual Discourse Connective Identification · Corpus Creation.

1 Introduction

Identifying discourse connectives (or DCs), such as "but" or "if", is fundamental to discourse analysis, which itself is useful to improve many downstream NLP tasks such as text generation, dialog systems and summarization, where understanding how textual elements are related to each other is crucial. Several datasets and formalisms have been proposed to study different aspects of computational discourse analysis, such as the Penn Discourse Treebank (PDTB) [15], Segmented Discourse Representation Theory (SDRT) [1] and Rhetorical Structure Theory (RST) [12]. For the task of DC identification, the PDTB is the most widely used resources, as it annotates lexical elements as DCs, whether they are explicit or not.

P. Rosso et al. (Eds.): NLDB 2022, LNAI 13286, pp. 1–12, 2022.
https://doi.org/10.1007/978-3-031-08473-7_41

The recent 2019 and 2021 DISRPT shared tasks[1] aimed at the identification of multilingual DCs in three languages: English, Turkish and Chinese. In general most participating systems achieved higher performance for English than for Turkish and Chinese, whose discourse annotated training data are much smaller. In order to address this issue, this paper investigates methods to improve DC annotation of low resource languages. In particular, we developed a BERT based model DC annotation and fine-tuned it with synthetic corpora of discourse connective annotations developed using parallel corpora. Results show that the F-measure achieved by the standard BERT model (92.49%, 93.97%, 87.47%) is hard to improve upon even with larger task specific training corpora.

2 Related Work

The Penn Discourse Treebank (PDTB) [15] is the largest English corpus manually curated with discourse annotations. These annotations fall into two categories: explicit and non-explicit relations. The former are expressed linguistically by well-defined lexical elements called discourse connectives or DCs (e.g., "but", "since"); while the latter signal the relation by other means such as an alternative lexicalization to a DC (i.e. an *AltLex*) or an entity-relation (i.e. an *EntRel*). Due to its relatively straightforward annotations formalism, the PDTB framework has been adopted for the creation of similar corpora in many languages, notably Chinese (CDTB) [24] and Turkish (TDB) [22] [23]. However, manually creating such high quality corpora is time consuming and expensive, hence very few languages have such data sets.

The earliest attempt at identifying DCs automatically using the PDTB dates back to [14] who used extracted features from gold-standard Penn Treebank parses and a maximum entropy classifier and obtained an F-measure of 94.19. Later, [7] showed that a simple logistic regression model could achieve better results without relying on gold-standard parse trees, but using only lexical features and part-of-speech tags. More recent approaches use neural methods which are more flexible for multilingual settings. In particular [13] employed multilingual BERT and bi-directional LSTMs and achieved F-measures of 88.60, 69.85, 79.32 in English, Turkish and Chinese respectively. These results seem to correlate with the size of the training set of 44k, 24k and 2k respectively. The most recent attempt for the detection of DCs, [4], used transformer models in addition to many handcrafted input features and a conditional random field as a final classifier instead of a linear output layer. This model achieved the best performance at DISRPT 2021 with an F-measure of 92.02%, 94.11%, 87.52% for English, Turkish and Chinese respectively; leading to a significant improvement in the state of the art in all three languages.

Corpus augmentation has been shown to improve many NLP tasks where annotated data sets are scarce. In particular, annotation projection has shown its usefulness for many tasks, such as part-of-speech tagging [20], word sense disambiguation [2], dependency parsing [17] and discourse relations identification [9].

[1] `https://sites.google.com/georgetown.edu/disrpt2021/home`

Since they are semantic and rhetoric in nature, it is often assumed that discourse annotations can be projected from one language to another through word alignment. In particular, [8] created a PDTB styled discourse corpus for French, by projecting discourse annotation from English (the PDTB) to French and using statistical word-alignment to identify unsupported annotations that should not be projected. The resulting corpus improved the performance of their French DC parser by 15%. Given the success of annotation projection for discourse analysis, we investigated its use to create synthetic corpora for DC annotation in Turkish and Chinese.

3 Methodology

Our work was done using the data sets of the 2021 DISRPT Task 2 *discourse connective identification across languages*[2] shared task. The task aimed at identifying DCs in three languages: English, Turkish and Chinese. For example, given the sentence:

(1) In addition, of course, some of the Japanese investments involved outright purchase of small U.S. firms.

Systems had to tag DCs using one of three tags: `B-Conn` (beginning of a DC), `I-Conn` (inside a DC) and `None` (not a DC). For example, the expected output for sentence (1) is shown in Figure 1.

In	*addition*	*,*	*of*	*course*	*,*	*some*
`B-Conn`	`I-Conn`	`None`	`None`	`None`	`None`	`None`
of	*the*	*Japanese*	*investments*	*involved*	*outright*	*purchase*
`None`	`None`	`None`	`None`	`None`	`None`	`None`
of	*small*	*U.S.*	*firms*	*.*		
`None`	`None`	`None`	`None`	`None`		

Fig. 1: Expected output of the 2021 DISRPT Task 2 for the sentence *In addition, of course, some of the Japanese investments involved outright purchase of small U.S. firms.*

3.1 The DISRPT-2021 Dataset

To train systems, the DISRPT organizers provided three PDTB styled annotated corpora: the English PDTB [15], the Turkish Discourse Treebank (TDB) [22] [23] and the Chinese Discourse Treebank (CDTB) [24]. Table 1 shows statistics of

[2] `https://sites.google.com/georgetown.edu/disrpt2021/home`

Corpus	Language	# of Train Sentences	% tok `B-Conn` + `I-Conn`	# of Test Sentences
PDTB	English	44,563	2.671	2,364
TDB	Turkish	24,960	1.900	3,289
CDTB	Chinese	2,049	2.249	404
ZHO-AG	Chinese	21,934	4.645	–
ZHO-PJ	Chinese	2,848	2.312	–
TUR-AG	Turkish	27,827	1.254	–
TUR-PJ	Turkish	4,468	4.191	–

Table 1: Statistics of the training and test data.
PDTB, TDB, and CDTB were provided by the organisers of DISRPT 2021; while ZHO-AG, ZHO-PJ, TUR-AG, and TUR-PJ were created by our methods (see Section 3.3)

these corpora. As Table 1 shows, the PDTB is the largest with approximately 44k training instances, far exceeding the number of training instances available for Chinese ($\approx$ 2k) and Turkish ($\approx$ 25k). Overall, in all 3 corpora the percentage of `B-Conn` and `I-Conn` labels is around 2% as most words are labelled as `None`.

3.2 The Base BERT DC Detection Model

To perform multilingual DC annotation, we developed a basic BERT DC annotation model using Pytorch and Huggingface [19]. The Huggingface tokenizer is used on the input sentences to produce sequences of word pieces, which are then fed to the model. As shown in Figure 2, the model is composed of a language specific BERT embedding [3], which can be found on Huggingface [19] (`bert-base-cased` for English, `bert-base-chinese` for Chinese, `dbmdz/bert-baseturkishcased` for Turkish). The output is then fed to a dropout unit which is then fed to a linear layer that produces a score for each of the 3 labels (`B-Conn`, `I-Conn`, and `None`) based on the BERT embeddings only. These scores are then fed to a conditional random field (CRF) which produces the most likely final tags for each word given the whole sentence.

We trained one model for each language for a maximum of 40 epochs using early stopping with a patience of 20 epochs on each corpora given by the organizers (see Section 3.1). This led to F-measures of 92.49, 93.97, 87.47 on the tests sets for English, Turkish and Chinese respectively[3]. The lower performance on Chinese seemed to be directly attributable to the lower number of training instances (see Table 1), hence we attempted to increase the training corpus for Chinese by creating synthetic annotated corpus. We did the same for Turkish to see if the increased data would provide any benefit to this task.

[3] Using the official DISRPT 2021 scorer available at `https://github.com/disrpt/sharedtask2021`

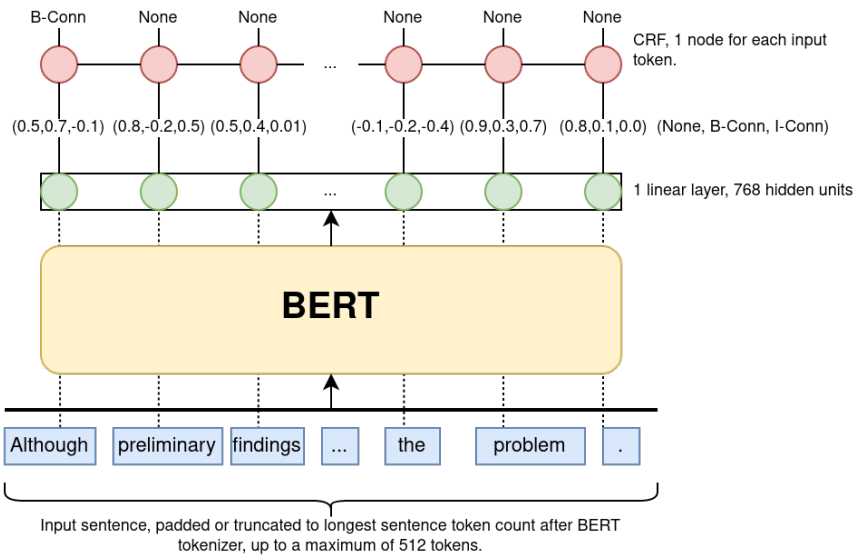

Fig. 2: Model overview of the BERT-base DC annotation model for English

3.3 Synthetic Corpora

To create the synthetic corpora annotated with DCs for Chinese and Turkish, we used two methods based on annotation projection and agreement on English-Chinese and English-Turkish parallel. To do so we first generated a list of English, Turkish and Chinese DCs from the provided PDTB, TDB, and CDTB respectively (see Section 3.1). This resulted in a list of 1160 English, 279 Turkish, and 195 Chinese connectives. The English connectives include the 100 DCs from the list of PDTB [15] plus 1060 AltLex that were labeled as DCs in the DISRPT training corpus.

Mandarin Chinese Synthetic Corpus For Chinese, we used the Tsinghua alignment evaluation set version 2 [11] [10], which contains 40,716 manually word aligned sentences. The Chinese and English sentences were already tokenized; and word alignment was already provided. We began by training our BERT based DC identification model (see Section 3.2) on the PDTB and CDTB corpora, then used it to identify DCs on both sides of the Tsinghua parallel corpus. Then, we created two synthetic datasets: one based on annotation projection, favoring recall; and one based on annotation agreement, favoring precision. As shown in Figure 3, the projection method is applied when the annotation model has identified a DC in a source language sentence, but the target language model did not identify a DC in the aligned words. In that case, the DC annotation is projected onto the aligned words in the target language. On the other hand, the agreement method favors precision by comparing the annotated parallel sentences and retaining them in the synthetic corpus only if the annotations of the

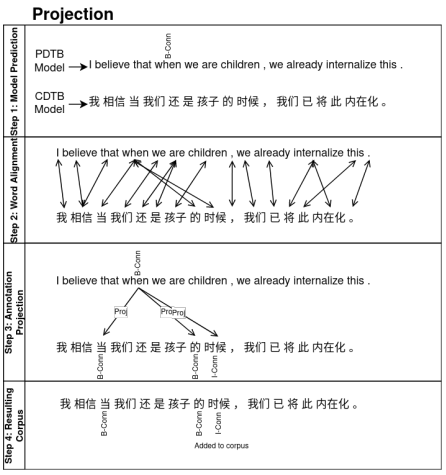

Fig. 3: Annotation projection of English discourse connectives onto a Chinese sentence.

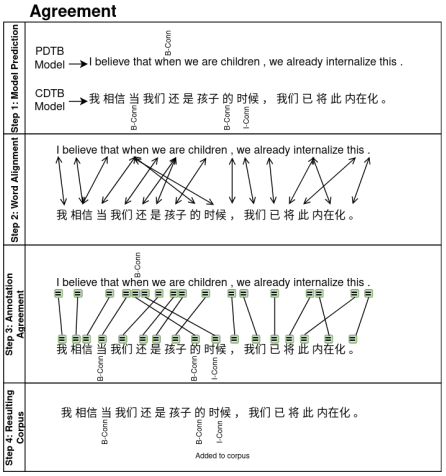

Fig. 4: Annotation agreement of English discourse connectives with a Chinese discourse connectives.

aligned words match; this is shown in Figure 4. Finally, in both methods, in order to create corpora with a similar `B-Conn` / `I-Conn` / `None` balance as the DISRPT corpora, we dropped:

1. all sentences that contain no potential DC – i.e. no word in the sentence is part of the language specific list of DCs (see Section 3.3), and
2. 50% of the sentences with at least one potential DC marked as non discourse usage – i.e. at least one word in the sentence is part of the language specific list of DCs but is labeled as `None`.

As shown in Table 1, the two resulting corpora (ZHO-AG and ZHO-PJ) contain 21,943 and 2,848 sentences respectively. ZHO-PJ contains the same ratio of `B-Conn` + `I-Conn` as the CDTB ($\approx 2\%$); but the ZHO-AG contains $\approx 4\%$.

Turkish The Turkish synthetic corpus is based on the seTimes [16] English-Turkish parallel corpus, which contains 207,677 aligned sentences. As the corpus was not word aligned, we tokenized it using Spacy [5] to split the words and punctuation. We then used SimAlign [6] to generate the word alignments with probabilities for each sentence. SimAlign can provide different alignment outputs based on three different algorithms: itermax, argmax and match. Based on the results of [6], itermax seems to perform better than the other methods; hence, we used the itermax alignments. In order to ensure that we have high quality alignment, we kept only the sentences with an average word alignment probability above 85%. Then, we proceeded the same way as we did for Chinese (see Section 3.3). As shown in Table 1, this produced a corpus of 27,827 sentences

for the agreement method (TUR-AG), effectively doubling the original corpus, and 4,468 sentences for the projection method (TUR-PJ).

4 Evaluation of the Base BERT Model

We evaluated the base BERT DC annotator (see Section 3.2) by training it with and without the new synthetic data for Turkish and Chinese, and testing it with the DISRPT 2021 test set (see Table 1) and the official DISRPT evaluation script. Table 13 shows the performance of each model for Turkish, and Table 15 shows the performance for Mandarin Chinese.

ID	Training set	Test set	Precision	Recall	F-measure
1	TDB	TDB	92.81 (±1.07)	95.17 (±0.77)	**93.97** (±0.34)
2	TUR-AG	TDB	87.62 (±2.32)	88.14 (±0.95)	87.86 (±0.71)
3	TUR-PJ	TDB	33.95 (±3.39)	52.10 (±4.18)	40.94 (±2.13)
4	TUR-AG + TUR-PJ	TDB	70.59 (±0.91)	86.85 (±1.21)	77.87 (±0.53)
5	TDB + TUR-AG	TDB	91.80 (±1.27)	94.53 (±2.75)	93.12 (±0.69)
6	TDB + TUR-PJ	TDB	89.90 (±1.17)	94.54 (±0.72)	92.16 (±0.34)
7	TDB + TUR-AG + TUR-PJ	TDB	90.97 (±0.68)	93.80 (±0.30)	92.37 (±0.49)

Table 2: Performance of the base Turkish BERT model (`dbmdz/bertbaseturkish-cased`) on the TDB test set.

ID	Training set	Test set	Precision	Recall	F-measure
1	CDTB	CDTB	88.78 (±1.60)	86.11 (±0.49)	**87.47** (±0.95)
2	ZHO-AG	CDTB	88.25 (±0.27)	85.90 (±0.85)	87.06 (±0.57)
3	ZHO-PJ	CDTB	62.34 (±4.62)	55.98 (±4.45)	58.86 (±2.97)
4	ZHO-AG + ZHO-PJ	CDTB	85.90 (±0.60)	85.26 (±0.85)	85.58 (±0.59)
5	CDTB + ZHO-AG	CDTB	89.16 (±1.03)	85.15 (±0.49)	87.10 (±0.46)
6	CDTB + ZHO-PJ	CDTB	87.52 (±0.50)	83.87 (±1.22)	85.65 (±0.51)
7	CDTB + ZHO-AG + ZHO-PJ	CDTB	88.15 (±0.85)	85.79 (±1.48)	86.95 (±1.03)

Table 3: Performance of the base Chinese BERT model (`bert-base-chinese`) on the CDTB test set.

As shown in Tables 13 and 15, the best models are the ones trained only on the original TDB and CDTB datasets (rows 1). Using the synthetic datasets, on their own (rows 2-4) or in combination with TDB or CDTB (rows 5-7) only seems to decrease the performance of the BERT model. The projection method (rows 3) seems to lead to significantly lower results for both languages (F-measure of 40.94 for Turkish, and 58.86 for Chinese). This is in line with the fact that annotation projection maximizes recall, hence the resulting corpora may contain more noise. The agreement method (rows 2) optimises precision and leads to a lower decrease in performance. The observed performance for the model trained with on the ZHO-AG (row 2) (87.06%) had a difference of only 0.36% to the model trained on CDTB (row 1) (87.42%) alone which suggests that this dataset is of high

quality and could easily be improved by human curation. While the F-measure with TUR-AG (row 2) (87.86%) had a difference of 6.11% with TDB (93.97), we believe this that this decrease in performance is due to error accumulation in the tokenization and word alignment steps (see Section 3.3). Further investigation would be required to verify the dataset.

For English we experimented with both the base and the large cased BERT embeddings [3]. As Table 12 shows, the performance of BERT-large did not lead to a significant increase in performance compared to BERT-base (93.12 versus 92.49), but did require more computational resources.

Model	Training set	Test set	Precision	Recall	F-measure
`bert-base-cased`	PDTB	PDTB	92.69 (±1.16)	92.31 (±0.41)	**92.49** (±0.77)
`bert-large-cased`	PDTB	PDTB	93.46 (±0.95)	92.79 (±0.69)	93.12 (±0.49)

Table 4: Performance of base English BERT model on the PDTB test set – `bertbasecased` versus `bertlargecased`.

5 Additional Experiments

In order to better understand the behavior of the model, we performed additional experiments with different configurations.

5.1 Multilingual BERT Embeddings

We attempted to use a multilingual model to use cross-lingual training. For this experiment we used multilingual BERT [3] embeddings, which are pre-trained on 104 different languages including English, Turkish and Chinese. To do so, we used two configurations. In the first configuration, we trained the model with all of the training data from the PDTB, TDB and CDTB (all data). In the second configuration, because the PDTB and TDB contain more training instances than the CTDB, we extracted a subset of the PDTB and TDB to balance the three corpora. The subsets we created are composed of 5% of the original PDTB and 10% of the original TDB, which were than used with the full CTDB to train a balanced model (balanced data).

Table 5 shows the results obtained using both settings: all data and balanced data, on the test sets of the PDTB, the TDB and the CDTB, when the model is trained on the PDTB, the TDB and the CTDB training sets simultaneously. As Table 5 shows, the multilingual BERT consistently under-performed when compared to the models pre-trained on the pertinent language, even when accounting for the differences in the corpora size. This this in line with the findings of [18] who showed that multilingual BERT was not sufficient to outperform a

BERT model pre-trained only on the language of study on a variety of NLP tasks such as POS tagging, named entity recognition, dependency parsing, and text classification.

Training set	Test set	All data (M-BERT)			Balanced data (M-BERT)		
		Precision	Recall	F-measure	Precision	Recall	F-measure
PDTB + TDB + CDTB	PDTB	91.53 (±0.78)	92.64 (±0.33)	92.08 (±0.25)	88.06 (±1.12)	88.31 (±0.64)	88.18 (±0.76)
PDTB + TDB + CDTB	TDB	89.29 (±1.60)	93.05 (±0.62)	91.12 (±0.82)	84.24 (±1.53)	88.17 (±1.31)	86.16 (±1.21)
PDTB + TDB + CDTB	CDTB	84.25 (±2.30)	85.19 (±1.76)	84.71 (±1.81)	85.05 (±1.39)	86.28 (±1.05)	85.65 (±0.93)

Table 5: Performance of the multilingual BERT model (`bertbasemultilingualcased`) while training on all languages simultaneously.

5.2 Linguistic Features

We investigated the use of additional linguistic information to see how the base BERT model would react. In particular, we added handcrafted features such as gold-standard universal part-of-speech tags, language-specific part-of-speech tag and universal dependency relations (provided in the DISRPT data set). These features only lowered the performance on the test set.

5.3 Bidirectional LSTM

Finally, we tried to pass the BERT output into 2 bidirectional LSTM layers instead of a CRF. Again, this either degraded the performance or took more epochs to get to a similar performance as using a CRF.

5.4 Comparison with the state-of-the-art

Given that the base-BERT model could not be improved upon, we compared it to the state-of-the-art models. Table 16 shows the results of our base BERT model using the official scorer and datasets of the DISRPT-2021 shared task, while Table 7 shows the performance of the participating systems. As Table 16 shows, our base BERT model performs as well as, if not better than the top performing model, DiscoDisco [4], for all three languages; while being significantly simpler in terms of linguistic and computational resources. The DiscoDisco approach used a collection of handcrafted features including 3 sentence embeddings (2 trainable/fine-tuned, and 1 static), a variety of grammatical and textual features (UPOS, XPOS, universal dependency relations, head distance, sentence type, and sentence length), and also a representation of the context via neighboring sentences. On the other hand, our model is less resource-intensive as it consists of only a language specific BERT-base + CRF and only uses the current sentence as context. This seems to show that the language specific BERT-base model contains sufficient information to accomplish this task, and feeding the model with additional information is redundant and only increases its complexity without significant performance gain.

Test set	Our Model base BERT		
	P	**R**	**F1**
PDTB	92.69	92.31	**92.49**
TDB	92.81	95.17	**93.97**
CDTB	88.78	86.11	**87.42**
macro average	91.43	91.20	91.29
micro average	92.49	93.45	92.96

Table 6: Performance of our base BERT models (`bert-base-cased` for PDTB, `dbmdz/bert-base-turkish-cased` for PDTB and `bert-base-chinese` for CDTB) with the official DISRPT 2021 Task 2 scorer.

Test set	Participating System											
	TMVM			**DiscoDisco**			**disCut**			**SegFormers**		
	P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
PDTB	85.98	65.54	74.38	92.32	91.15	**92.02**	93.32	88.67	90.94	89.73	92.61	91.15
TDB	80.00	24.14	37.10	93.71	94.53	**94.11**	90.55	86.93	88.70	90.42	91.17	90.79
CDTB	30.00	0.96	1.86	89.19	85.95	**87.52**	84.43	66.03	74.10	85.05	87.50	86.26
macro average	65.33	30.21	37.78	91.74	90.54	91.22	89.43	80.54	84.58	88.40	90.43	89.40
micro average	79.00	38.75	49.30	92.87	92.64	92.85	91.22	86.22	88.60	89.79	91.49	90.63

Table 7: Performance of the participating systems with the official DISRPT 2021 Task 2 scorer, taken from [21].

6 Conclusion and Future Work

This paper has described our experiments to improve a BERT-base model for discourse connective annotation in a multilingual setting. We described two methods to induce discourse annotated corpora and proposed a simple BERT-base model that is capable of achieving similar results to the best performing model at the DISRPT 2021 task 2. Our experiments with additional data, different models architectures and different input features, suggest that language specific BERT models with a CRF output and small amount of data is all that is needed to achieve a strong performance on the task of multilingual discourse connective identification.

Future work is required to evaluate the quality of the synthetic datasets created and the inclusion of additional information such as the POS tags, dependency relations, and the type of discourse relation being signaled; this would make these datasets useful for other natural language processing tasks.

The synthetic corpora are publicly available at `https://github.com/CLaC-Lab`.

7 Acknowledgment

The authors would like to thank the anonymous reviewers for their valuable comments on an earlier version of this paper. This work was financially supported by the Natural Sciences and Engineering Research Council of Canada (NSERC).

Corpus	Language	# of Sentences	# Types	# tok B-Conn	# tok B-Conn + I-Conn	% tok B-Conn + I-Conn
PDTB-train	English	44,563	1160	23,850	28,349	2.671
PDTB-dev	English	1,703	132	953	1,112	2.796
PDTB-test	English	2,364	167	1,245	1,483	2.665
TDB-train	Turkish	24,960	277	7,063	7,572	1.900
TDB-dev	Turkish	2,948	99	831	888	1.777
TDB-test	Turkish	3,289	96	854	919	1.919
CDTB-train	Chinese	2,049	195	1034	1,171	2.249
CDTB-dev	Chinese	438	113	314	398	3.560
CDTB-test	Chinese	404	114	312	354	3.514

Table 8: Statistics of the training, validation and test data at DISRPT 2021.

ID	Training set	Test set	Precision	Recall	F-measure
1	TDB	TDB	92.81 (±1.07)	95.17 (±0.77)	**93.97** (±0.34)

Table 9: Performance of the base Turkish BERT model (`dbmdz/bertbaseturkish-cased`) on the TDB test set.

ID	Training set	Test set	Precision	Recall	F-measure
1	CDTB	CDTB	88.78 (±1.60)	86.11 (±0.49)	**87.47** (±0.95)

Table 10: Performance of the base Chinese BERT model (`bert-base-chinese`) on the CDTB test set.

References

1. Asher, N., Lascarides, A.: Logics of conversation. Cambridge University Press (2003)
2. Bentivogli, L., Pianta, E.: Exploiting parallel texts in the creation of multilingual semantically annotated resources: The MultiSemCor Corpus. Natural Language Engineering **11**(3), 247–261 (2005)
3. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. CoRR **abs/1810.04805** (2018), `http://arxiv.org/abs/1810.04805`
4. Gessler, L., Behzad, S., Liu, Y.J., Peng, S., Zhu, Y., Zeldes, A.: Discodisco at the DISRPT2021 shared task: A system for discourse segmentation, classification, and connective detection. CoRR **abs/2109.09777** (2021), `https://arxiv.org/abs/2109.09777`

Row	Model #	Model	Training Dataset	PDTB Test Set		
				Test Precision	**Test Recall**	**Test F-measure**
1	Model 4	bert-base-cased + CRF	PDTB	92.69 (±1.16)	92.31 (±0.41)	92.49 (±0.77)
2	Model 4	bert-base-cased + CRF	PDTB-75%	92.99 (±0.56)	91.73 (±1.12)	92.35 (±0.48)
3	Model 4	bert-base-cased + CRF	PDTB-50%	91.77 (±0.90)	91.61 (±0.56)	91.69 (±0.33)
4	Model 4	bert-base-cased + CRF	PDTB-25%	92.04 (±0.88)	90.81 (±0.89)	91.41 (±0.13)
5	Model 4	bert-base-cased + CRF	PDTB-10%	91.71 (±0.83)	89.66 (±0.67)	90.67 (±0.19)
6	Model 4	bert-base-cased + CRF	PDTB-05%	90.66 (±1.33)	87.53 (±0.74)	89.06 (±0.45)
7	Model 4	bert-large-cased + CRF	PDTB	93.46 (±0.95)	92.79 (±0.69)	**93.12 (±0.49)**
8	—	baseline	PDTB	65.52	29.00	40.20

Table 11: Performance of various model configurations on the PDTB (English) test sets. Performance indicates the average score over five runs ± the standard deviation.

Row	Model #	Model	Training Dataset	PDTB Test Set		
				Test Precision	Test Recall	Test F-measure
1	Model 4	bert-base-cased + CRF	PDTB	92.69 (±1.16)	92.31 (±0.41)	92.49 (±0.77)
2	Model 4	bert-base-cased + CRF	PDTB-75%	92.99 (±0.56)	91.73 (±1.12)	92.35 (±0.48)
3	Model 4	bert-base-cased + CRF	PDTB-50%	91.77 (±0.90)	91.61 (±0.56)	91.69 (±0.33)
4	Model 4	bert-base-cased + CRF	PDTB-25%	92.04 (±0.88)	90.81 (±0.89)	91.41 (±0.13)
5	Model 4	bert-base-cased + CRF	PDTB-10%	91.71 (±0.83)	89.66 (±0.67)	90.67 (±0.19)
6	Model 4	bert-base-cased + CRF	PDTB-05%	90.66 (±1.33)	87.53 (±0.74)	89.06 (±0.45)
2	Model 4	bert-large-cased + CRF	PDTB	93.46 (±0.95)	92.79 (±0.69)	**93.12 (±0.49)**
3	Model 5	RoBERTa + CRF	PDTB	92.02 (±1.08)	90.97 (±0.83)	91.49 (±0.47)
4	Model 6	GPT2 + CRF	PDTB	79.64 (±2.19)	82.86 (±1.21)	81.20 (±1.17)
5	Model 1	bert-base-cased	PDTB	93.03 (±0.65)	92.95 (±0.42)	92.99 (±0.34)
6	Model 2	bert-base-cased + Bi-LSTM	PDTB	92.68 (±0.62)	92.77 (±0.57)	92.72 (±0.52)
7	Model 3	bert-base-cased+ Bi-GRU	PDTB	92.99 (±0.72)	92.56 (±0.65)	92.77 (±0.38)
13	Model 4	bert-base-multilingual-cased + CRF	PDTB + TDB + CDTB	92.29 (±1.45)	91.63 (±0.54)	91.95 (±0.54)
14	Model 1	bert-base-multilingual-cased	PDTB + TDB + CDTB	91.53 (±0.78)	92.64 (±0.33)	92.08 (±0.25)
15	Model 1	bert-base-multilingual-cased	PDTB-05% + TDB-10% + CDTB	88.06 (±1.12)	88.31 (±0.64)	88.18 (±0.76)
8	—	baseline	PDTB	65.52	29.00	40.20

Table 12: Performance of various model configurations on the PDTB (English)
test sets. Performance indicates the average score over five runs ± the standard deviation.

Row	Model #	Model	Training Dataset	TDB Test Set Test Precision	TDB Test Set Test Recall	TDB Test Set Test F-measure
1	Model 4	dbmdz/bert-base-turkish-cased + CRF	TDB	93.19 (±1.08)	94.87 (±0.72)	94.01 (±0.31)
2	Model 4	dbmdz/bert-base-turkish-cased + CRF	TDB-75%	92.45 (±0.92)	95.01 (±1.01)	93.70 (±0.32)
3	Model 4	dbmdz/bert-base-turkish-cased + CRF	TDB-50%	90.73 (±1.42)	94.44 (±1.11)	92.54 (±0.41)
4	Model 4	dbmdz/bert-base-turkish-cased + CRF	TDB-25%	90.48 (±0.83)	94.16 (±0.95)	92.28 (±0.49)
5	Model 4	dbmdz/bert-base-turkish-cased + CRF	TDB-10%	88.98 (±0.72)	91.02 (±1.61)	89.98 (±0.67)
6	Model 4	dbmdz/bert-base-turkish-cased + CRF	TDB-05%	85.84 (±1.27)	86.29 (±1.91)	86.04 (±0.41)
7	Model 4	dbmdz/bert-base-turkish-cased + CRF	TUR-AG	88.22 (±1.87)	87.80 (±0.84)	87.99 (±0.54)
8	Model 4	dbmdz/bert-base-turkish-cased + CRF	TUR-PJ	34.04 (±2.40)	52.39 (±3.71)	41.15 (±1.68)
9	Model 4	dbmdz/bert-base-turkish-cased + CRF	TUR-AG + TUR-PJ	70.34 (±1.04)	85.68 (±2.55)	77.23 (±1.00)
10	Model 4	dbmdz/bert-base-turkish-cased + CRF	TDB + TUR-AG	91.67 (±1.49)	95.10 (±1.60)	93.34 (±1.00)
11	Model 4	dbmdz/bert-base-turkish-cased + CRF	TDB + TUR-PJ	89.81 (±1.45)	94.77 (±0.75)	92.21 (±0.48)
12	Model 4	dbmdz/bert-base-turkish-cased + CRF	TDB + TUR-AG + TUR-PJ	90.77 (±0.62)	94.02 (±1.38)	92.36 (±0.61)
13	Model 1	dbmdz/bert-base-turkish-cased	TDB	93.63 (±0.18)	95.22 (±0.87)	**94.42 (±0.37)**
14	—	baseline	TDB	47.64	33.22	39.14

Table 13: Performance of various model configurations on the TDB (Turkish) test sets.

Row	Model #	Model	Training Dataset	CDTB Test Set		
				Test Precision	Test Recall	Test F-measure
1	Model 4	bert-base-chinese + CRF	CDTB	88.43 (±1.67)	86.54 (±0.72)	**87.47 (±0.96)**
2	Model 4	bert-base-chinese + CRF	CDTB-75%	88.33 (±1.34)	86.80 (±0.53)	87.55 (±0.88)
3	Model 4	bert-base-chinese + CRF	CDTB-50%	84.42 (±1.17)	85.39 (±0.54)	84.89 (±0.67)
4	Model 4	bert-base-chinese + CRF	CDTB-25%	81.16 (±2.97)	81.28 (±0.74)	81.20 (±1.62)
5	Model 4	bert-base-chinese + CRF	CDTB-10%	73.13 (±2.89)	81.60 (±1.05)	77.10 (±1.56)
6	Model 4	bert-base-chinese + CRF	CDTB-05%	69.48 (±2.35)	76.73 (±1.48)	72.89 (±1.09)
7	Model 4	bert-base-chinese + CRF	ZHO-AG	87.00 (±1.75)	85.64 (±0.73)	86.31 (±1.10)
8	Model 4	bert-base-chinese + CRF	ZHO-PJ	62.47 (±3.27)	53.27 (±4.87)	57.37 (±2.94)
9	Model 4	bert-base-chinese + CRF	ZHO-AG + ZHO-PJ	85.48 (±0.83)	84.55 (±1.39)	85.01 (±1.07)
10	Model 4	bert-base-chinese + CRF	CDTB + ZHO-AG	89.87 (±1.35)	85.20 (±0.57)	87.46 (±0.79)
11	Model 4	bert-base-chinese + CRF	CDTB + ZHO-PJ	86.50 (±1.61)	83.97 (±0.88)	85.21 (±0.81)
12	Model 4	bert-base-chinese + CRF	CDTB + ZHO-AG + ZHO-PJ	87.79 (±0.92)	84.81 (±1.93)	86.27 (±1.20)
13	Model 1	bert-base-chinese	CDTB	87.74 (±2.12)	85.77 (±1.86)	86.71 (±1.00)
14	—	baseline	CDTB	56.90	54.17	55.50

Table 14: Performance of various model configurations on the CDTB (Chinese) test sets.

Row	Model #	Model	Training Dataset	CDTB Test Set Test Precision	Test Recall	Test F-measure
1	Model 4	bert-base-chinese + CRF	CDTB	88.43 (±1.67)	86.54 (±0.72)	**87.47 (±0.96)**
2	Model 1	bert-base-chinese	CDTB	87.74 (±2.12)	85.77 (±1.86)	86.71 (±1.00)
3	—	baseline	TDB	56.90	54.17	55.50

Table 15: Performance of various model configurations on the CDTB (Chinese) test sets.

	Our Model base BERT		
Test set	**P**	**R**	**F1**
PDTB	93.46	92.79	**93.12**
TDB	93.63	95.22	**94.42**
CDTB	88.43	86.54	**87.47**
macro average	91.84	91.52	91.67
micro average	93.22	93.69	93.45

Table 16: Performance of our base BERT models (`bert-base-cased` for PDTB, `dbmdz/bert-base-turkish-cased` for PDTB and `bert-base-chinese` for CDTB) with the official DISRPT 2021 Task 2 scorer.

5. Honnibal, M., Montani, I., Van Landeghem, S., Boyd, A.: spaCy: Industrial-strength Natural Language Processing in Python (2020), `https://spacy.io/`, available at `https://spacy.io/`
6. Jalili Sabet, M., Dufter, P., Yvon, F., Schütze, H.: SimAlign: High quality word alignments without parallel training data using static and contextualized embeddings. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, (EMNLP-2020). pp. 1627–1643. Punta Cana, Dominican Republic (Nov 2020), `https://www.aclweb.org/anthology/2020.findings-emnlp.147`
7. Johannsen, A., Søgaard, A.: Disambiguating explicit discourse connectives without oracles. In: Proceedings of the Sixth International Joint Conference on Natural Language Processing, (IJCNLP 2013). pp. 997–1001. Nagoya, Japan (Oct 2013), `https://aclanthology.org/I13-1134`
8. Laali, M.: Inducing discourse resources using annotation projection (November 2017), `https://spectrum.library.concordia.ca/983791/`, phD Thesis, Concordia University, `https://spectrum.library.concordia.ca/983791/`
9. Laali, M., Kosseim, L.: Improving discourse relation projection to build discourse annotated corpora. In: Proceedings of the International Conference Recent Advances in Natural Language Processing, (RANLP 2017). pp. 407–416. Varna, Bulgaria (Sep 2017). https://doi.org/10.26615/978-954-452-049-6$_0$54
10. Liu, Y., Liu, Q., Lin, S.: Log-linear models for word alignment. In: Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL-2005). pp. 459–466. Ann Arbor, Michigan (Jun 2005). https://doi.org/10.3115/1219840.1219897, `https://aclanthology.org/P05-1057`
11. Liu, Y., Sun, M.: Contrastive unsupervised word alignment with non-local features. In: Proceedings of the Twenty-Ninth Association for the Advancement of Artificial Intelligence Conference on Artificial Intelligence, (AAAI-2015). p. 2295–2301 (2015), `http://arxiv.org/abs/1410.2082`
12. Mann, W.C., Thompson, S.A.: Rhetorical structure theory: A framework for the analysis of texts. IPrA Papers in Pragmatics **1**, 79–105 (1987)
13. Muller, P., Braud, C., Morey, M.: ToNy: Contextual embeddings for accurate multilingual discourse segmentation of full documents. In: Proceedings of the Workshop on Discourse Relation Parsing and Treebanking 2019. pp. 115–124. Minneapolis, MN (Jun 2019). https://doi.org/10.18653/v1/W19-2715, `https://aclanthology.org/W19-2715`

14. Pitler, E., Nenkova, A.: Using syntax to disambiguate explicit discourse connectives in text. In: Proceedings of the Association for Computational Linguistics and International Joint Conference on Natural Language Processing (ACL-IJCNLP 2009). pp. 13–16. Suntec, Singapore (Aug 2009), https://aclanthology.org/P09-2004
15. Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A., Webber, B.: The Penn Discourse TreeBank 2.0. In: Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008). pp. 2961–2968. Marrakech, Morocco (May 2008), http://www.lrec-conf.org/proceedings/lrec2008/pdf/754_paper.pdf
16. Tiedemann, J.: Parallel data, tools and interfaces in opus. In: Chair), N.C.C., Choukri, K., Declerck, T., Dogan, M.U., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S. (eds.) Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC 2012). p. 2214–2218. Istanbul, Turkey (may 2012), http://www.lrec-conf.org/proceedings/lrec2012/pdf/463_Paper.pdf
17. Tiedemann, J.: Improving the cross-lingual projection of syntactic dependencies. In: Proceedings of the 20th Nordic Conference of Computational Linguistics (NODALIDA 2015). pp. 191–199. Vilnius, Lithuania (May 2015), https://aclanthology.org/W15-1824
18. Virtanen, A., Kanerva, J., Ilo, R., Luoma, J., Luotolahti, J., Salakoski, T., Ginter, F., Pyysalo, S.: Multilingual is not enough: BERT for finnish. CoRR **abs/1912.07076** (2019), http://arxiv.org/abs/1912.07076
19. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Brew, J.: Huggingface's transformers: State-of-the-art natural language processing. CoRR **abs/1910.03771** (2019), http://arxiv.org/abs/1910.03771
20. Yarowsky, D., Ngai, G., Wicentowski, R.: Inducing multilingual text analysis tools via robust projection across aligned corpora. In: Proceedings of the First International Conference on Human Language Technology Research (HLT 2001). pp. 1–8. San Diego, California (Mar 2001), https://aclanthology.org/H01-1035
21. Zeldes, A., Liu, J.: Disrpt 2021 task 2 results (2021), https://sites.google.com/georgetown.edu/disrpt2021/results#h.gb445xshqmt7, available at https://sites.google.com/georgetown.edu/disrpt2021/results
22. Zeyrek, D., Kurfalı, M.: TDB 1.1: Extensions on Turkish discourse bank. In: Proceedings of the 11th Linguistic Annotation Workshop. pp. 76–81. Valencia, Spain (Apr 2017). https://doi.org/10.18653/v1/W17-0809, https://aclanthology.org/W17-0809
23. Zeyrek, D., Kurfalı, M.: An assessment of explicit inter- and intra-sentential discourse connectives in Turkish discourse bank. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). pp. 4023–4029. Miyazaki, Japan (May 2018), https://aclanthology.org/L18-1634
24. Zhou, L., Gao, W., Li, B., Wei, Z., Wong, K.f.: Cross-lingual identification of Ambiguous discourse Connectives for resource-poor Language. In: Proceedings of the 24th International Conference on Computational Linguistics: Technical Papers (COLING 2012). pp. 1409–1418. Mumbai (Dec 2012), https://aclanthology.org/C12-2138