Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 1152–1158
Marseille, 11–16 May 2020

Adapting BERT to Implicit Discourse Relation Classification with a Focus on Discourse Connectives

Yudai Kishimoto, Yugo Murawaki, Sadao Kurohashi
Graduate School of Informatics, Kyoto University
Yoshida-honmachi, Sakyo-ku, Kyoto, 606-8501, Japan
{kishimoto, murawaki, kuro} @nlp.ist.i.kyoto-u.ac.jp

Abstract

BERT, a neural network-based language model pre-trained on large corpora, is a breakthrough in natural language processing, significantly outperforming previous state-of-the-art models in numerous tasks. However, there have been few reports on its application to implicit discourse relation classification, and it is not clear how BERT is best adapted to the task. In this paper, we test three methods of adaptation. (1) We perform additional pre-training on text tailored to discourse classification. (2) In expectation of knowledge transfer from explicit discourse relations to implicit discourse relations, we add a task named explicit connective prediction at the additional pre-training step. (3) To exploit implicit connectives given by treebank annotators, we add a task named implicit connective prediction at the fine-tuning step. We demonstrate that these three techniques can be combined straightforwardly in a single training pipeline. Through comprehensive experiments, we found that the first and second techniques provide additional gain while the last one did not.

Keywords: discourse relation classification, BERT, explicit connective, implicit connective

1. Introduction

Discourse relation classification, a task of recognizeing the semantic relations between two text spans, is beneficial for many NLP tasks including machine translation (Meyer et al., 2015), natural language inference (Pan et al., 2018), summarization (Isonuma et al., 2019), and sentiment analysis (Saito et al., 2019). In the Penn Discourse TreeBank (PDTB) (Prasad et al., 2008), discourse relations are conventionally divided into two types: explicit and implicit. Explicit relations have strong cues named discourse connectives such as "because" and "however", while implicit relations lack these cues. For this reason, the recognition of implicit relations is the bottleneck of discourse relation classification (Xue et al., 2016; Dai and Huang, 2019).

In this paper, we investigate how to improve the performance of implicit discourse relation classification by applying BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2018). BERT is a Transformer-based neural network architecture with specialized training procedures. It is pre-trained on large corpora like Wikipedia and BooksCorpus, and fine-tuned using task-specific datasets to transfer the pre-trained representations to downstream tasks. Although BERT is conceptually simple, it significantly outperforms previous state-of-the-art models in many natural language processing tasks such as reading comprehension (Devlin et al., 2018), syntactic analysis (Goldberg, 2019), and sentiment analysis (Xu et al., 2019).

In implicit discourse relation classification, Nie et al. (2019) and Shi and Demberg (2019b) reported that BERT significantly outperformed previous state-of-the-art models. Considering the broader context of research in this area, however, we expect additional improvement to be achieved with task-specific adaptation. Table 1 summarizes recent studies and techniques employed by them. It is evident that a comprehensive investigation is needed to answer the question: How is BERT best adapted to the task?

We examine three adaptation methods in this paper. The first one is to exploit a large amount of unlabeled data from same domain text (referred to as *Domain text* in Table 1). In the context of BERT, a simple way to do this is to perform additional pre-training on the domain text (Shi and Demberg, 2019b), but the domain can be more specific to this task. In fact, Rutherford and Xue (2015) automatically collected explicit argument pairs from an unlabeled corpus. In the pre-BERT era, they had no choice but to forcibly convert them into pseudo-training data for implicit discourse relation classification. Now, such explicit argument pairs can be used for BERT's additional pre-training.

The second technique is a more direct use of explicit argument pairs (referred to as *explicit connective prediction* in Table 1). Explicit argument pairs have, by definition, (explicit) discourse connectives. Training a model to predict explicit connectives may help it learn the discourse relations (Wu et al., 2017). We call this task explicit connective prediction. For BERT, Nie et al. (2019) inserted this task between pre-training and fine-tuning (additional pre-training). While their additional pre-training only covers explicit connective prediction (referred to as *single-task pre-training* in Table 1), but a more straightforward way is to do this together with BERT's ordinary pre-training tasks (multi-task pre-training).

The third technique is to exploit implicit connectives (referred to as *implicit connective prediction* in Table 1). In PDTB, annotators inserted an implicit connective between an implicit argument pair to facilitate consistent annotation. Implicit connectives were exploited by Zhou et al. (2010), Qin et al. (2017), and Shi and Demberg (2019a) for implicit discourse relation classification. A BERT-friendly formalization of the task is a multi-task learning at the fine-tuning step: BERT is trained to predict the implicit connective as well as the discourse relation.

	Using BERT	Domain text	Explicit connective prediction	Implicit connective prediction
BERT	✓			
BERT+ICP	✓			fine-tuning
BERT+DT	✓	DC		
BERT+DT+ICP	✓	DC		fine-tuning
BERT+ECP	✓		multi-task pre-training	
BERT+ECP+ICP	✓		multi-task pre-training	fine-tuning
Rutherford and Xue (2015)		DC		
Wu et al. (2017)			non-BERT pre-training	
Lei et al. (2018)				
Bai and Zhao (2018)				
Shi and Demberg (2019a)				single-step joint learning
Nie et al. (2019)	✓		single-task pre-training	
Shi and Demberg (2019b)	✓	raw text		
Dai and Huang (2019)				

Table 1: Model configurations. DT, ECP, and ICP in the left column correspond to domain text, explicit connective prediction, and implicit connective prediction, respectively. DC means the use of a large amount of explicit argument pairs while raw text means the use of a raw text that is from the same domain as PDTB.

In this way, we show that these three techniques can be combined straightforwardly in a single training pipeline. In the experiments, we compare multiple combinations of adaptation methods. We found that the first and second techniques yield additional gain for implicit discourse relation classification while the last one does not.

2. Related Work

2.1. Implicit Discourse Relation Classification

Discourse connectives are one of the strongest cues in discourse relation classification. For this reason, some studies try to use discourse connectives for implicit discourse relations classification.

Rutherford and Xue (2015) and Wu et al. (2017) tried to use explicit connectives. Rutherford and Xue (2015) proposed a Naive Bayes classifier using a large amount of unlabeled training data. They collected explicit argument pairs with freely omissible discourse connectives which can be dropped independently of the context without changing the interpretation of the discourse relation. However, Sporleder and Lascarides (2008) argued training on explicit argument pairs was not a good strategy. They investigated how feasible it is to use explicit argument pairs in implicit discourse relation classification and reported that explicit and implicit argument pairs may be too dissimilar linguistically and that removing unambiguous discourse markers in the automatic labeling process may lead to a meaning shift in the examples. However, the recent success of neural network-based transfer learning motivates us to rethink the importance of explicit connectives. In fact, Wu et al. (2017) proposed discourse-specific word embeddings. Embeddings were learned by classifying discourse connectives taken from a large amount of explicit discourse argument pairs.

Zhou et al. (2010), Qin et al. (2017) and Shi and Demberg (2019a) used pseudo-training data in which implicit connectives are inserted into implicit argument pairs. Zhou et al. (2010) proposed a language model that can predict implicit connectives. They used implicit connectives predicted by the language model as a feature of an SVM-based implicit discourse relation classifier. Qin et al. (2017) proposed a neural network model using domain adversarial training. This model focused to transfer knowledge from the recognition model supplied with implicit connectives to the model without connectives. Shi and Demberg (2019a) proposed a sequence-to-sequence neural network model. This model tried to generate implicit connectives from implicit argument pairs, and adapted the representation of the arguments to the classification task.

2.2. BERT

Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2018), a multi-layer bidirectional Transformer encoder, is one of the breakthrough models in natural language processing. Training of BERT consists of two stages: pre-training and fine-tuning. In pre-training, BERT learns contextual information and the relationship between two sentences from a large unlabeled corpus. In fine-tuning, BERT is trained using a task-specific dataset and adapts the pre-trained representations to downstream tasks.

In the GitHub repository,[1] Devlin et al. (2018) suggest that it will likely be beneficial to run additional steps of pre-training if the downstream task has a large domain-specific corpus available. In fact, some studies showed the benefit of running domain-specific steps. Xu et al. (2019) proposed a post-training approach to enhance domain-specific and task-specific knowledge. They defined a new task named Review Reading Comprehension and modified the pre-training method to solve Review Reading Comprehension. Lee et al. (2019) proposed a pre-trained language representation model for the biomedical domain. This model runs additional pre-training on biomedical corpora, starting from the BERT pre-training model. Beltagy et al. (2019) proposed a task-specific BERT model pre-trained only on the science domain data. They reported that the proposed method outperformed vanilla BERT on named entity recognition, relation classification and dependency parsing in the scientific domain.

[1] `https://github.com/google-research/bert` (access date: Dec. 1, 2019)

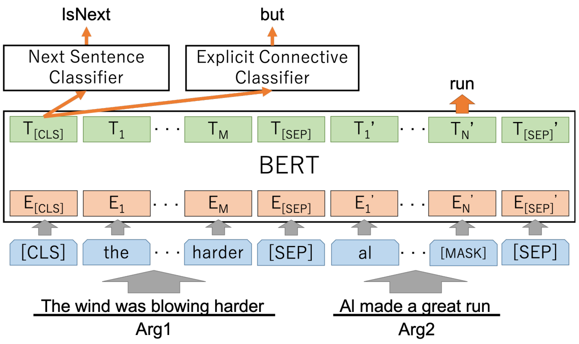
IsNext
but
Next Sentence Classifier
Explicit Connective Classifier
run
T[CLS] T1 · · · TM T[SEP] T1' · · · TN' T[SEP]'
BERT
E[CLS] E1 · · · EM E[SEP] E1' · · · EN' E[SEP]'
[CLS] the · · · harder [SEP] al · · · [MASK] [SEP]
The wind was blowing harder
Arg1
Al made a great run
Arg2

Figure 1: Overview of the additional pre-training step with the explicit connective prediction task. The input is an explicit argument pair automatically extracted from an unlabeled corpus, but the explicit connective is dropped. BERT is trained to recover the connective while performing the MaskedLM and next sentence prediction tasks.

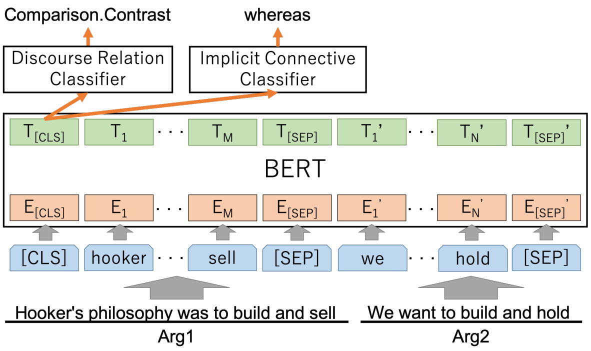
Comparison.Contrast
whereas
Discourse Relation Classifier
Implicit Connective Classifier
T[CLS] T1 · · · TM T[SEP] T1' · · · TN' T[SEP]'
BERT
E[CLS] E1 · · · EM E[SEP] E1' · · · EN' E[SEP]'
[CLS] hooker · · · sell [SEP] we · · · hold [SEP]
Hooker's philosophy was to build and sell
Arg1
We want to build and hold
Arg2

Figure 2: Overview of the fine-tuning step with implicit connective prediction. The input is an implicit argument pair randomly selected from the training data, where annotators provide an implicit connective for each pair. BERT is trained to predict the implicit connective as well as the discourse relation.

3. Proposed Method

3.1. BERT

In this paper, we use BERT (Devlin et al., 2018) as the baseline model. For implicit discourse relation classification, the input is a pair of arguments, Arg1 and Arg2, and the output is one of the pre-defined discourse relations. Arg1 and Arg2 are concatenated into a single sequence, with the special token [SEP] indicates the end of each of the arguments. The special token [CLS] is inserted at the beginning of the sequence.

In the pre-training step, BERT is trained on two unsupervised prediction tasks: MaskedLM and next sentence prediction. In the MaskedLM task, 15% of the input tokens are masked at random, and BERT is trained to recover those masked tokens. In the next sentence prediction task, BERT determines if a given pair of sentences actually occurs consecutively in this order. Using these two pre-training tasks, BERT can learn contextual information and the relationship between two sentences.

In the fine-tuning step, Devlin et al. (2018) proposed four task-specific models incorporating BERT with one additional output layer. Implicit discourse relation classification can be cast as a sequence pair classification task.

3.2. The Use of Domain Text

Devlin et al. (2018) suggest that it will likely be beneficial to run additional steps of pre-training if a downstream task has a large domain-specific corpus available. For this reason, we collect a large amount of explicit argument pairs from an unlabeled corpus and use them in the additional pre-training step.

We collect explicit argument pairs in two steps. We first locate the occurrences of discourse connectives in the unlabeled corpus. Among 100 discourse connectives defined by The PDTB Research Group (2007), we used all connectives. We then identify the spans of the corresponding argument pairs. In this paper, we use a discourse parser by Lin et al. (2014) to predict explicit argument pairs.[2] They reported that the F-measure of partial matching on argument pairs in this parser is 80.96%,[3] which we believe is sufficiently high for pretraining. Note that Wu et al. (2017) collected explicit argument pairs using a similar method. However, they only used argument pairs located within the same sentence while we do not apply this constraint.

3.3. Explicit Connective Prediction Task

We aim to learn knowledge about discourse relations from explicit connectives. Explicit connectives mark the conceptual relationship between the two sentences. For example, the discourse relation of "*I cannot buy this laptop* because **I don't have enough money.**" is Contingency.Cause because we know "because" always represents a causality relation. Note that we referred to the pair as Arg1 (in *italic*) and Arg2 (in **bold**). We expect BERT to be flexible enough to transfer knowledge from explicit argument pairs to implicit ones. Accordingly, we introduce an additional pre-training step with the explicit connective prediction task.

The additional pre-training with the explicit connective prediction task is illustrated in Figure 1. In this step, the explicit connective classifier is given the representation of the [CLS] token and outputs an explicit connective. Note that Nie et al. (2019) also adapted the explicit connective prediction task to BERT. However, they singled out this task while we perform it jointly with MaskedLM and next sentence prediction.

3.4. Implicit Connective Prediction Task

In order to enhance the performance of implicit discourse relation classification, some studies try to extract strong cues from PDTB's annotation. In PDTB, annotators assigned implicit connectives to implicit argument pairs. Because these connectives are very similar to explicit connectives, recent studies tried to learn knowledge from implicit connectives (Qin et al., 2017; Shi and Demberg, 2019a).

[2] We used a re-implementation (https://github.com/WING-NUS/pdtb-parser).

[3] Partial matching means that a parser gets a credit if there is any overlap between the verbs and nouns of the two spans.

Following these studies, we combine the implicit connective prediction task in the fine-tuning steps of BERT.

The fine-tuning step with the implicit connective prediction task is shown in Figure 2. In this task, the implicit connective classifier is given the representation of the [CLS] token and outputs an implicit connective.

4. Experiments

4.1. Setup

4.1.1. Penn Discourse TreeBank

We evaluated the performance of our models on the Penn Discourse TreeBank (PDTB) 2.0 (Prasad et al., 2008), which is the most popular and largest corpus of discourse relations in English. The annotation is done as another layer on the Wall Street Journal sections of the Penn Treebank. Each discourse relation consists of two text spans (arguments), a relation label and a discourse connective.

Relation labels are organized as a 3-level hierarchy in the PDTB. Popular experimental settings are top-level one-versus-all binary classification (Pitler et al., 2009), top-level 4-way classification (Pitler et al., 2009; Rutherford and Xue, 2015), second-level 11-way classification (Lin et al., 2009; Ji and Eisenstein, 2015), and modified second-level classification for the CoNLL 2015 Shared Task (Xue et al., 2015). We used top-level one-versus-all binary classification, top-level 4-way classification, and second-level 11-way classification in this experiment.

In top-level one-versus-all binary classification and top-level 4-way classification, we followed previous studies (Pitler et al., 2009; Rutherford and Xue, 2015); sections 2–20 as the training set, sections 0–1 as the development set, and sections 21–22 as the test set. Note that we used equal numbers of positive and negative examples in top-level one-versus-all binary classification. In second-level 11-way classification, we report results in three different settings. The first setting, PDTB-Lin, is based on Lin et al. (2009); sections 2–21 as the training set, section 22 as the development set, and section 23 as the test set. The second one, PDTB-Ji, is following Ji and Eisenstein (2015); sections 2–20 as the training set, sections 0–1 as the development set, and sections 21–22 as the test set. The last one, Cross Validation, is following Shi and Demberg (2017); 10-fold cross validation using the whole corpus of sections 0–24.

Table 2 shows a distribution of relation labels in the Cross Validation dataset. Note that although we tried to replicate the procedures described by Shi and Demberg (2017) as closely as possible, there remained slight differences in the discourse relation distribution.

4.1.2. Model Configurations

We used a pre-trained model named BERT$_{\text{BASE}}$–uncased as a baseline model. It was released with the original BERT code.[4] In fine-tuning, we set the batch size to 32 and the maximum sequence length to 128. The other hyperparameter settings were the same as those of BERT$_{\text{BASE}}$–uncased. Devlin et al. (2018) recommend choosing 3 or 4 as the number of training epochs. Accordingly, we fine-tuned for 3 epochs.

Sense	Train	Dev	Test
Comparison.Concession	175	22	22
Comparison.Contrast	1,650	206	207
Contingency.Cause	3,290	412	411
Contingency.Pragmatic cause	55	7	7
Expansion.Alternative	144	18	18
Expansion.Conjunction	2,751	344	344
Expansion.Instantiation	1,116	139	140
Expansion.List	309	38	39
Expansion.Restatement	2,486	311	310
Temporal.Asynchronous	520	65	65
Temporal.Synchrony	140	18	17
Total	12,636	1,580	1,580[5]

Table 2: The distribution of relation labels in the Cross Validation dataset.

In this experiment, we modified BERT's training procedure with 3 methods: domain text, explicit connective prediction, and implicit connective prediction. The upper half of Table 1 shows model configurations of our settings.

Domain text in Table 1 means additional pre-training using domain text. The models with +DT in Table 1 were run the additional steps of pre-training using about 26.8 million explicit argument pairs extracted from the Gigaword corpus, a large unlabeled newswire corpus.[6] In additional pre-training, we pre-trained the model for 1 epoch. The batch size was 8 and the maximum sequence length was set to 512. The other hyperparameter settings were the same as those of BERT$_{\text{BASE}}$–uncased.

Explicit connective prediction in Table 1 means the explicit connective prediction task at the additional pre-training step. Multi-task pre-training refers to the combination of three tasks while single-task pre-training only covers this task. For the models with +ECP in Table 1, we added the explicit connective prediction task to the +DT setting.

Implicit connective prediction in Table 1 means implicit connective prediction at the fine-tuning step. For the models with +ICP in Table 1, we added the implicit connective prediction task to the fine-tuning step.

4.1.3. Models for Comparison

For comparison, we collected state-of-the-art models from the literature:

- **Rutherford and Xue (2015)** A Naive Bayes classifier that was trained explicit argument pairs with freely omissible discourse connectives extracted from a large amount of unlabeled data.
- **Wu et al. (2017)** A feedforward neural network using discourse-specific word embeddings that were learned from a large amount of explicit argument pairs.
- **Lei et al. (2018)** The state-of-the-art model in the Expansion-versus-all classification task. It is a collection of Naive Bayes classifiers using many features based on linguistic properties.

[4] `https://github.com/google-research/bert`

[5] Cross-validation allows us to test on all 15,796 instances.

[6] `https://catalog.ldc.upenn.edu/LDC2011T07`

model	One-Versus-All Binary(F_1)				4-way		11-way(F_1)		
	Comp.	Cont.	Expa.	Temp.	Macro-F	Accuracy	PDTB-Lin	PDTB-Ji	Cross Varidation
BERT	72.74	71.67	70.27	76.47	55.07	61.99	52.72	51.40	49.30
BERT+ICP	73.84	70.91	69.80	76.76	55.19	62.89	51.62	52.42	49.47
BERT+DT	74.43	73.01	70.37	77.70	55.47	63.03	52.74	52.90	50.20
BERT+DT+ICP	76.60	**74.28**	70.78	77.45	54.83	62.65	52.57	52.90	50.20
BERT+ECP	75.46	73.01	72.86	**80.88**	**58.48**	**65.26**	52.39	54.32	51.68
BERT+ECP+ICP	**77.28**	73.85	**73.40**	79.41	56.40	64.02	52.74	54.09	**51.79**
Rutherford and Xue (2015)	41.00	53.80	69.40	33.30	40.50	57.10	-	-	-
Wu et al. (2017)	-	-	-	-	44.84	58.85	-	-	-
Lei et al. (2018)	43.24	57.82	72.88	29.10	47.15	-	-	-	-
Bai and Zhao (2018)	47.85	54.47	70.60	36.87	51.06	-	45.73	48.22	-
Shi and Demberg (2019a)	41.83	62.07	69.58	35.72	46.40	61.42	45.82	47.83	41.29
Nie et al. (2019)	-	-	-	-	-	-	-	**54.70**	-
Shi and Demberg (2019b)	-	-	-	-	-	-	**54.82**	53.23	49.35
Dai and Huang (2019)	45.34	51.80	68.50	45.93	52.89	59.66	-	48.23	-

Table 3: Accuracy of the implicit discourse relations classification datasets.

label	BERT+ECP+ICP			Kishimoto et al. (2018) (F_1)
	Precision	Recall	F-measure	
Comparison.Concession	0.00	0.00	0.00	0.00
Comparison.Contrast	50.21 ± 2.66	48.76 ± 2.66	49.43 ± 2.66	22.87 ± 2.9
Contingency.Cause	54.48 ± 1.58	60.56 ± 1.58	57.32 ± 1.58	48.13 ± 1.5
Contingency.Pragmatic cause	0.00	0.00	0.00	0.00
Expansion.Alternative	34.64 ± 8.17	28.33 ± 8.17	30.64 ± 8.17	0.87 ± 2.6
Expansion.Conjunction	51.49 ± 2.08	57.69 ± 2.08	54.39 ± 2.08	44.66 ± 2.3
Expansion.Instantiation	59.96 ± 4.03	58.21 ± 4.03	58.97 ± 4.03	44.98 ± 3.7
Expansion.List	47.20 ± 14.41	26.20 ± 14.41	32.53 ± 14.41	22.03 ± 9.1
Expansion.Restatement	48.08 ± 3.43	46.51 ± 3.43	47.17 ± 3.43	33.41 ± 2.7
Temporal.Asynchronous	48.54 ± 4.87	46.46 ± 4.87	47.41 ± 4.87	21.40 ± 7.7
Temporal.Synchrony	0.00	0.00	0.00	0.00
Micro Ave.	51.86 ± 1.27	51.86 ± 1.27	51.86 ± 1.27	39.80 ± 0.9

Table 4: The performance of BERT in the Cross Validation dataset.

- **Bai and Zhao (2018)** The state-of-the-art model in Comparison-versus-all classification. It is a deeper neural network model augmented by different grained text representations like character, sentence and sentence pair levels.

- **Shi and Demberg (2019a)** The state-of-the-art model in Contingency-versus-all classification and 4-way classification in terms of accuracy. It is a sequence-to-sequence neural network model trained to transform an implicit argument pair *without* an implicit connective into a pair *with* an implicit connective.

- **Nie et al. (2019)** The state-of-the-art BERT model in PDTB-Ji. They performed additional pre-training using about 3.2 million explicit argument pairs for 5 discourse connectives.

- **Shi and Demberg (2019b)** The state-of-the-art BERT model in PDTB-Lin and Cross Validation. They used the BERT-base model that was run an additional step of pre-training on the parts of the Wall Street Journal corpus.

- **Dai and Huang (2019)** The state-of-the-art model in Temporal-versus-all classification and 4-way classification in terms of Macro F-measure. They presented a paragraph-level neural network model for incorporating external event knowledge and coreference relations.

4.2. Results

Table 3 shows the results for each of the datasets. Our settings outperformed previous models in all test sets except PDTB-Lin and PDTB-Ji. Comparing *BERT+DS* with *BERT*, we can see that the domain text strategy obtained about 1.5 point gain in PDTB-Ji, and about 1.0 point gain in Cross Validation and 4-way classification in terms of accuracy. Similarly, the comparison of *BERT+ECP* with *BERT* reveals that the explicit connective prediction task yielded about 4.4 point gain in 4-way classification in terms of Macro F-measure, about 2.9 point gain in PDTB-Ji, and about 2.4 point gain in Cross Validation. In contrast, the implicit connective prediction task provided no significant gain, as *BERT+ICP* did not consistently outperform *BERT*. A breakdown of the performance in the Cross Validation dataset is shown in Table 4. Compared with a previous model that incorporated external knowledge (Kishimoto et al., 2018), our model achieved over 10% improvements in accuracy for Comparison.Contrast, Expansion.Alternative, Expansion.Instantiation, Expansion.List, Expansion.Restatement and Temporal.Asynchronous. The accuracies for minority labels, Comparison.Concession, Contingency.Pragmatic cause and Temporal.Synchrony, stuck at 0%, however.

5. Discussion

Shi and Demberg (2019b) reported that BERT outperformed the current state-of-the-art in second-level 11-way classification. We reconfirmed their report in 11-way classification and also found that BERT outperformed previous studies in top-level classifications. Surprisingly, BERT achieved over 70% F_1 score in One-Versus-All Binary classification while many previous models achieved less than 50% in Comparison-versus-all binary classification and Temporal-versus-all binary classification. We urge researchers to switch from One-Versus-All Binary classification to second-level or third-level classification tasks.

We found that the explicit connectives prediction task resulted in 2.9% gain in PDTB-Ji while the DisSent task (Nie et al., 2019), which runs the explicit connectives prediction task in the additional pre-training step, provides 2.0% gain. This result suggests that the combination of the explicit connectives prediction task and BERT's ordinary pre-training task is a better strategy than single-task pre-training.

Qin et al. (2017) and Shi and Demberg (2019a) reported implicit connectives help implicit discourse relation classification. Comparing *BERT+ICP* with *BERT* in Table 3, however, we can see that implicit connectives provided no significant gain. We conjecture that annotated data are too small for implicit connectives to be effective for BERT, for which data size is a key factor for success.

6. Conclusion

In this paper, we applied three additional training tasks to BERT, (1) additional pre-training using domain text, (2) the explicit connective prediction task at the additional pre-training step, and (3) the implicit connective prediction at the fine-tuning step to BERT. Through comprehensive experiments, we found that the first and second techniques provide additional gain while the last one did not.

While transfer learning with BERT is demonstrated to be very effective for discourse relation classification, we feel that there are non-negligible differences between explicit and implicit argument pair. It may be worthwhile to revisit the notion of freely omissible discourse connective (Rutherford and Xue, 2015) to focus on explicit argument pairs from which knowledge can be straightforwardly transferred to implicit discourse relation classification. We plan to modify freely omissible discourse connectives to fit second-level 11-way classification. Another future direction is to adapt the BERT to incorporate external knowledge. Kishimoto et al. (2018) and Dai and Huang (2019) argued that the model for discourse classification could be further improved by incorporating external event knowledge like ConceptNet (Speer and Havasi, 2012) and temporal event knowledge. We plan to combine BERT with knowledge representation learning.

7. Bibliographical References

Bai, H. and Zhao, H. (2018). Deep enhanced representation for implicit discourse relation recognition. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 571–583.

Beltagy, I., Lo, K., and Cohan, A. (2019). SciBERT: A pretrained language model for scientific text. 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 3613–3618.

Dai, Z. and Huang, R. (2019). A regularization approach for incorporating event knowledge and coreference relations into neural discourse parsing. 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 2967–2978.

Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2018). BERT: pre-training of deep bidirectional transformers for language understanding. *Computing Research Repository*, arXiv:1810.04805.

Goldberg, Y. (2019). Assessing bert's syntactic abilities. *Computing Research Repository*, arXiv:1901.05287.

Isonuma, M., Mori, J., and Sakata, I. (2019). Unsupervised neural single-document summarization of reviews via learning latent discourse structure and its ranking. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2142–2152.

Ji, Y. and Eisenstein, J. (2015). One vector is not enough: Entity-augmented distributed semantics for discourse relations. *Transactions of the Association of Computational Linguistics*, 3:329–344.

Kishimoto, Y., Murawaki, Y., and Kurohashi, S. (2018). A knowledge-augmented neural network model for implicit discourse relation classification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 584–595.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., and Kang, J. (2019). Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Lei, W., Xiang, Y., Wang, Y., Zhong, Q., Liu, M., and Kan, M.-Y. (2018). Linguistic properties matter for implicit discourse relation recognition: Combining semantic interaction, topic continuity and attribution. In *AAAI Conference on Artificial Intelligence*.

Lin, Z., Kan, M.-Y., and Ng, H. T. (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.

Lin, Z., Ng, H. T., and Kan, M.-Y. (2014). A pdtb-styled end-to-end discourse parser. *Natural Language Engineering*, 20(2):151–184.

Meyer, T., Hajlaoui, N., and Popescu-Belis, A. (2015). Disambiguating discourse connectives for statistical machine translation. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, 23(7):1184–1197.

Nie, A., Bennett, E., and Goodman, N. (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.

Pan, B., Yang, Y., Zhao, Z., Zhuang, Y., Cai, D., and He, X. (2018). Discourse marker augmented network with reinforcement learning for natural language inference. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 989–999.

Pitler, E., Louis, A., and Nenkova, A. (2009). Automatic sense prediction for implicit discourse relations in text. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 683–691.

Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A., and Webber, B. (2008). The Penn discourse treebank 2.0. In *Proceedings of LREC2008*, pages 2961–2968.

Qin, L., Zhang, Z., Zhao, H., Hu, Z., and Xing, E. (2017). Adversarial connective-exploiting networks for implicit discourse relation classification. In *Proceedings of ACL2017*, pages 1006–1017.

Rutherford, A. and Xue, N. (2015). Improving the inference of implicit discourse relations via classifying explicit discourse connectives. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 799–808.

Saito, J., Murawaki, Y., and Kurohashi, S. (2019). Minimally supervised learning of affective events using discourse relations. 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 5757–5764.

Shi, W. and Demberg, V. (2017). Do we need cross validation for discourse relation classification? In *Proceedings of EACL2017*, pages 150–156.

Shi, W. and Demberg, V. (2019a). 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.

Shi, W. and Demberg, V. (2019b). Next sentence prediction helps implicit discourse relation classification within and across domains. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5794–5800.

Speer, R. and Havasi, C. (2012). Representing general relational knowledge in conceptnet 5. In *Proceedings of LREC2012*, pages 3679–3686.

Sporleder, C. and Lascarides, A. (2008). Using automatically labelled examples to classify rhetorical relations: An assessment. *Natural Language Engineering*, 14(3):369–416.

The PDTB Research Group. (2007). The Penn Discourse Treebank 2.0 Annotation Manual. Technical report, Institute for Research in Cognitive Science, University of Pennsylvania.

Wu, C., Shi, X., Chen, Y., Su, J., and Wang, B. (2017). Improving implicit discourse relation recognition with discourse-specific word embeddings. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 269–274.

Xu, H., Liu, B., Shu, L., and Yu, P. (2019). BERT post-training for review reading comprehension and aspect-based sentiment analysis. 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 2324–2335.

Xue, N., Ng, T. H., Pradhan, S., Prasad, R., Bryant, C., and Rutherford, A. (2015). The conll-2015 shared task on shallow discourse parsing. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning - Shared Task*, pages 1–16.

Xue, N., Ng, T. H., Pradhan, S., Rutherford, A., Webber, B., Wang, C., and Wang, H. (2016). CoNLL 2016 shared task on multilingual shallow discourse parsing. In *Proceedings of CoNLL2016 shared task*, pages 1–19.

Zhou, Z.-M., Xu, Y., Niu, Z.-Y., Lan, M., Su, J., and Tan, C. L. (2010). Predicting discourse connectives for implicit discourse relation recognition. In *Coling 2010: Posters*, pages 1507–1514.