

A Side-by-side Comparison of Transformers for English Implicit Discourse Relation Classification

Bruce W. Lee^{1,2}, BongSeok Yang², Jason Hyeong-Jong Lee²

¹University of Pennsylvania - PA, USA

²LXPER AI Research - Seoul, South Korea

brucelws@seas.upenn.edu

bongseok@lxper.com

jasonlee@lxper.com

Abstract

Though discourse parsing can help multiple NLP fields, there has been no wide language model search done on implicit discourse relation classification. This hinders researchers from fully utilizing public-available models in discourse analysis. This work is a straightforward, fine-tuned discourse performance comparison of seven pre-trained language models. We use PDTB-3, a popular discourse relation annotated dataset. Through our model search, we raise SOTA to 0.671 ACC and obtain novel observations. Some are contrary to what has been reported before (Shi and Demberg, 2019b), that sentence-level pre-training objectives (NSP, SBO, SOP) generally fail to produce the best performing model for implicit discourse relation classification. Counterintuitively, similar-sized PLMs with MLM and full attention led to better performance.

1 Introduction

An utterance has multiple dimensions of meaning. Discourse relation classification identifies one such dimension: the coherence relation between clauses or sentences arising from low-level textual cues (Zhao and Webber, 2022; Webber et al., 2019). This makes the task important to several NLP fields, including multi-party dialogue analysis (Li et al., 2022), social media postings analysis (Siskou et al., 2022), and student literary writing analysis (Fiacco et al., 2022). A discourse relation is often marked with explicit connectives such as *but*, *because*, and *and*. Consider the following example:

Although Philip Morris typically tries to defend the rights of smokers, ["*this has nothing to do with cigarettes, nor will it ever,*" the spokesman says]_{Arg1}. [**But**]_{Conn} [**some anti-smoking activists disagree**]_{Arg2}, expressing anger... → Comparison.Contrast

The explicit connective, Conn (But), is informative. Hence, it is fairly easy to know that the two arguments, *Arg1* and **Arg2**, are compared, likely in a contrasting relationship rather than similarity. This task is often referred to as explicit discourse relation classification. Pitler and Nenkova (2009) achieves a 94.15% accuracy (4-way) with Naive Bayes.

Implicit discourse relation classification, on the other hand, aims to classify discourse relationships in cases without an explicit connective. It has received constant attention (Li et al., 2022) since the release of Penn Discourse Tree Bank 2.0 (PDTB-2) (Prasad et al., 2008). Consider the following:

["*Last year we probably bought one out of every three new deals,*]_{Arg1}, he says. "[**This year, at best, it's in one in every five or six.**]_{Arg2}" → Comparison.Contrast

Without an explicit connective, Conn, discourse relation classification only relies on low-level semantic cues from the arguments, *Arg1* and **Arg2**. Such "implicit" discourse relation classification is very challenging as it requires a language model to conceptualize the unstated goal the speaker is trying to achieve, not only the literal content (Shi and Demberg, 2019b; Sileo et al., 2019).

With XLNet_{large} (Yang et al., 2019) achieving ~60% accuracy (Kim et al., 2020), pre-trained language models showed promising improvements from the past studies: Maximum-Entropy Learning (~40% F1) (Lin et al., 2014), Adversarial Network (~46% ACC) (Qin et al., 2017), Seq2Seq + Memory Network (~48% ACC) (Shi and Demberg, 2019a). Implicit discourse relation classification gives relatively small textual information for a language model to infer from. Thus, pre-training large text helps establish typical relations within/across clauses and sentences (Shi and Demberg, 2019b).

Configurations	ALBERT _{large}	BART _{large}	BigBird-R.	DeBERTa _{large}	Longformer _{large}	RoBERTa _{large}	SpanBERT _{large}
Release	2019	2020	2020	2020	2020	2019	2020
Parameters	17M	406M	-	350M	435M	340M	340M
Hidden	1024	1024	-	1024	1024	1024	1024
Layers	24 (Enc)	24 (Enc+Dec)*	-	24 (Enc)	24 (Enc)	24 (Enc)	24 (Enc)
Attention Heads	16	16	-	16	16	16	16
Self-Attention	Full	Full	Block-Sparse	Full**	Global+Window	Full	Full
Max Seq. Length	512	512	4096	512	4096	512	512
Pre-train Obj.	MLM & SOP	TI & SS	-	MLM	MLM	MLM	MLM & SBO

Table 1: Tested language models and their varying configurations. *: BART follows the original encoder-decoder architecture, 12 layers allocated for each. **: DeBERTa uses disentangled attention. MLM: masked language modelling. SOP: sentence order prediction. SBO: span boundary objective. TI: text infilling. SS: sentence shuffling.

Pre-trained language models, like BERT (Devlin et al., 2018), follow transformer-type (Vaswani et al., 2017) architecture and have only been recently introduced into implicit discourse relation classification (Kishimoto et al., 2020). To the best of our knowledge, BERT and XLNet are the only pre-trained language models (fine-tuned and) evaluated for implicit discourse relation classification on PDTB-3 (Kim et al., 2020). However, language models vary in architecture, training objective, data, etc.

Instead of performing a focused study on a single model, we fine-tune seven state-of-the-art (SOTA) language models (§2). Our wider approach brings weaknesses (§5) (as we ignore some model-specific characteristics), but it allows the bird’s-eye view of several downstream performances in PDTB-3 (§3) (Webber et al., 2019) and raises SOTA (~67% ACC) on Kim et al. (2020)’s evaluation protocol. By contrasting performances, we show that certain language model characteristics can benefit implicit discourse relation classification.

Additionally, we take the best-performing language model and check if the "full-sentence(s)" setup gives better performance (§3.4). As we elaborate further in the following sections, our sanity checks on PDTB-3 hint that some argument annotations are questionable in terms of consistency and coverage. Hence, implicit discourse relation classification accuracy might improve by simply training the language model with a full sentence(s) instead of human-annotated argument spans (*Arg1* and *Arg2*). We evaluate this idea toward the end.

2 Background

The pre-train and fine-tune paradigm have been led by the remarkable downstream task performances of pre-trained language models (Kalyan et al., 2021; Devlin et al., 2018). For several NLP tasks, a pre-trained language model could have

likely done a fine job at learning syntax, semantics, and world knowledge – given enough data and model size (Wang et al., 2019).

A pre-trained language model’s competence in discourse was questionable until Shi and Demberg (2019b) proposed that BERT’s pre-training objective can benefit implicit discourse relation classification. However, Iyer et al. (2020) hints that BERT is not the language model best suited to the task.

Implicit discourse relation classification is an active area of research (Kurfali, 2022; Zhao and Webber, 2022; Kurfali and Östling, 2021b; Knaebel, 2021; Munir et al., 2021; Kurfali and Östling, 2021a; Kishimoto et al., 2020; Bourgonje and Stede, 2019; Shi and Demberg, 2019b; Bai and Zhao, 2018; Dai and Huang, 2018; Rutherford et al., 2017). However, there has been no wide-range model study on implicit discourse relation classification, limiting a researcher’s scope of model choice. This issue is further complicated by the fact that discourse task performances do not always correlate with popular semantics-based natural language understanding (NLU) scores, such as GLUE (Sileo et al., 2019). Thus, it is difficult to predict which language model can perform well without a dedicated empirical exploration.

With the a version update to Penn Discourse Tree Bank (PDTB-3) (Webber et al., 2019) and the correspondingly updated evaluation method (Kim et al., 2020), we fine-tune seven language models to implicit discourse relation classification.

The chosen language models are: RoBERTa_{large} (Liu et al., 2019), ALBERT_{large} (Lan et al., 2019), BigBird-RoBERTa_{large} (Zaheer et al., 2020), BART_{large} (Lewis et al., 2020), Longformer_{large} (Beltagy et al., 2020), SpanBERT_{large} (Joshi et al., 2020), DeBERTa_{large} (He et al., 2020a). These models are selected with diversity in mind, especially in terms of input sequence length, attention type, and pre-train objectives. These models fol-

	ALBERT _{large}	BART _{large}	BigBird-R.	DeBERTa _{large}	Longformer _{large}	RoBERTa _{large}	SpanBERT _{large}
Hyperparameters							
Learning Rate	5e-6	5e-6	5e-6	2e-6	5e-6	2e-6	5e-6
a: Argument Spans							
Accuracy	0.565	0.657	0.649	0.671	0.668	0.670	0.627
Variance	2.53e-4	2.15e-4	4.02e-4	2.70e-4	2.15e-4	3.32e-4	1.78e-4
b: Full Sentence(s)							
Accuracy	0.534	0.629	0.620	0.634	0.627	0.617	0.598
Variance	2.27e-4	4.28e-4	2.79e-4	3.75e-4	4.18e-4	3.62e-4	2.84e-4

Table 2: Language model performances (test set) on Level-2 14-way implicit discourse relation classification.

low the popular transformer architecture (Vaswani et al., 2017), and we will not review each model in detail. A brief comparison is shown in Table 1.

3 Experiments

3.1 Data Preparation

We obtained the official PDTB-3 data from the Linguistic Data Consortium¹. PDTB-3 is a large-scale resource of annotated discourse relations and their arguments over the 1 million words Wall Street Journal Corpus (Marcus et al., 1993). From a public repository², we retrieved the corresponding evaluation script (Kim et al., 2020). We describe some characteristics of the evaluation protocol below.

Cross-validation is used on the section level to preserve paragraph and document structures. Cross-validation likely solves label sparsity issue (Shi and Demberg, 2017). The 25 sections of PDTB-3 are divided into 12 folds with 2 development, 2 test, and 21 training sections in each fold. The sliding window of two sections is used, creating 12 folds.

Label set is composed of 14 senses on L2 discourse relations (see Appendix B). Only the senses with ≥ 100 instances are used. This is to produce results that are in align with Kim et al. (2020). This alignment is crucial as we directly compared our results against fine-tuned BERT from Kim et al. (2020), which is trained with next sentence prediction (NSP) objective. Multiply-annotated labels become separate training instances.

3.2 Fine-Tuning

To ensure reproducibility, we only take pre-trained language models from the now ubiquitous Huggingface (Wolf et al., 2019) transformers library. Fine-tuning was done with PyTorch (Paszke et al., 2019) and our scripts are publicly available.

During fine-tuning, each training instance is a concatenation of two arguments (= sequence of tokens in *Arg1* and *Arg2*). BERT-type models carry special tokens ([CLS], [SEP], [EOS]) for segmentation: [CLS], *Arg1*₁ ... *Arg1*_N, [SEP], *Arg2*₁ ... *Arg2*_M, [EOS]. Depending on the model, these special tokens are modified or completely removed.

As for hyperparameter searches, we mostly focus on the learning rate. We use the popular AdamW optimizer with a linear scheduler (no warm-up steps). As for the learning rate, we start from 2e-5, a value commonly used for text classification since Sun et al. (2019). We test lower learning rates of 2e-6 and 5e-6; we find that 5e-6 (which is slightly lower than what is usually used in sequence classification) performs best for almost all models. The batch size is 8 and the max input length is set at 256.

Lastly, for each experiment step (i.e. BART on fold 1), we train for 10 epochs with an early stop. The training stops if the current epoch’s validation loss (see development set §3.1) did not decrease from the previous epoch. Model training time, GPU, language model repository address, and other details on hyperparameters are in Appendix C.

3.3 Evaluation and Observations

In Table 2-a, we report the mean test set accuracy of 12 folds along with variance. This is in alignment with what was recommended by Kim et al. (2020). Development set performances are given in Table 3 to facilitate reproducibility. For multiply-annotated labels (also discussed in §3.1), the model only has to get one label correct. We reach some surprising observations, which we share below.

1) Sentence-level pre-train objectives are not necessary to create best-performing models. This is contrary to Shi and Demberg (2019b), which proposed that NSP helps implicit discourse

¹www ldc.upenn.edu

²github.com/najoungkim/pdtb3

	ALBERT _{large}	BART _{large}	BigBird-R.	DeBERTa _{large}	Longformer _{large}	RoBERTa _{large}	SpanBERT _{large}
a: Argument Spans							
Accuracy	0.566	0.663	0.653	0.673	0.669	0.670	0.629
Variance	2.94e-4	1.33e-4	1.62e-4	2.47e-4	1.68e-4	1.01e-4	2.03e-4
b: Full Sentence(s)							
Accuracy	0.567	0.660	0.645	0.656	0.661	0.652	0.639
Variance	3.92e-4	4.59e-4	3.50e-4	1.85e-4	2.56e-4	4.10e-4	3.45e-4

Table 3: Language model performances (**dev set**) on Level-2 14-way implicit discourse relation classification.

relation classification after conducting an ablation study on BERT. Their finding was intuitive as well because implicit discourse relation classification aims to find the relationship between two argument spans.

But in a more general scope, the necessity of NSP has been questioned multiple times (Yang et al., 2019; Lample and Conneau, 2019). In other words, NSP – or any other sentence-level pre-train objective for that matter – could have been only helpful in some specific ablation study of BERT-type models but not in other cases (Liu et al., 2019). We obtain supporting results in Table 2-a, where language models with sentence-level objectives performed worse than MLM-only models given similar model sizes (ALBERT is an exception).

2) Long-document modifications (mostly done by altering attention schemes of an existing model) decrease the original model performance.

At first, we postulated that long-document models could lead to performance increases because they can learn long-span discourse relations during pre-training. But using sparse or block attention mechanisms eventually led to a performance decrease.

The decrease is clearly demonstrated by BigBird-RoBERTa_{large} and Longformer_{large}. Both models start from the existing RoBERTa_{large} checkpoint and modify it to process longer sequences. Such modifications achieved performance increases in other NLP tasks like question-answering, coreference resolution, and some cases of sequence classification. But implicit discourse relation classification, which requires the model’s understanding of dense discourse relations hidden within a few tokens, long-document modification is a drawback.

3) The simplest combination of MLM and full attention is best suited for implicit discourse relation classification. We are making this argument within the scope of what we have tested. We believe that MLM and full attention (e.g., RoBERTa, DeBERTa) work best because the model has to make inferences based on a relatively small number of tokens. Hence, trivial textual cues should

not be risked being overlooked. MLM, with full attention, forces every token to attend to every other and learn the token-specific relations, likely to lose the least textual cues and nuances.

3.4 Train Full Sentence or Argument Span?

Following the aforementioned observations, we postulated that fine-tuning language models using full sentence(s) could further improve classification accuracy. By full sentence(s), we refer to the sentence(s) (usually up to two) that the annotated argument spans appeared. We had two reasons for our postulation: 1. textual cues that hint at underlying discourse relation could be spread throughout the sentence(s), 2. argument span annotation is sometimes inconsistent, especially at punctuation marks, unnecessarily confusing the language model. Implicit discourse relation classification has rarely been tested using the full sentence.

We built an argument matcher to find the source sentence of each annotated argument span. For inter-sentential relations, we only considered argument spans that came from two adjacent source sentences. We share the test set results in Table 2-b. The results bring us to our fourth observation.

4) As input, concatenating argument spans generally perform better than full sentence(s).

Opposed to our postulation, using full sentence(s) as input decreased performance on the test set. Though we see mixed results on the development set in Table 3, training full sentences as input generally decrease performance. But when it comes to implicit discourse sense classification from the raw text (that means in practical, end-to-end applications), the benefits of using argument spans must be weighed against the low accuracies (50% ~ 60%) of the available argument extractors.

4 Conclusion

Researchers often build or modify a neural network to improve task performance. While such effort is essential, this paper shows that SOTA can also be raised through extensive search and application of

existing resources. Through a side-by-side comparison of seven PLMs, we also make handy observations on pre-training objectives, long-document modifications, and full-sentence setups. Though some might consider these phenomena rather expected, nothing is scientifically conclusive until an analysis is performed at an adequate scale. We hope that our report helps researchers working towards discourse understanding, and we continue to discuss the missing details in the appendices.

5 Acknowledgement

We thank the anonymous reviewers for their crisp and realistic advices on cleaning the language of the paper and experiments. Most of the review opinions were accepted and were reflected in the paper as they were valid.

References

- Hongxiao Bai and Hai Zhao. 2018. Deep enhanced representation for implicit discourse relation recognition. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 571–583.
- Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.
- Peter Bourgonje and Manfred Stede. 2019. Explicit discourse argument extraction for german. In *International Conference on Text, Speech, and Dialogue*, pages 32–44. Springer.
- Zeyu Dai and Ruihong Huang. 2018. Improving implicit discourse relation classification by modeling inter-dependencies of discourse units in a paragraph. *arXiv preprint arXiv:1804.05918*.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- James Fiacco, Shiyang Jiang, David Adamson, and Carolyn Rosé. 2022. [Toward automatic discourse parsing of student writing motivated by neural interpretation](#). In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, pages 204–215, Seattle, Washington. Association for Computational Linguistics.
- Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020a. DeBERTa: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.
- Ruifang He, Jian Wang, Fengyu Guo, and Yugui Han. 2020b. [TransS-driven joint learning architecture for implicit discourse relation recognition](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 139–148, Online. Association for Computational Linguistics.
- Dan Iter, Kelvin Guu, Larry Lansing, and Dan Jurafsky. 2020. Pretraining with contrastive sentence objectives improves discourse performance of language models. *arXiv preprint arXiv:2005.10389*.
- Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, and Omer Levy. 2020. Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.
- Katikapalli Subramanyam Kalyan, Ajit Rajasekharan, and Sivanesan Sangeetha. 2021. Ammus: A survey of transformer-based pretrained models in natural language processing. *arXiv preprint arXiv:2108.05542*.
- Najoung Kim, Song Feng, Chulaka Gunasekara, and Luis Lastras. 2020. [Implicit discourse relation classification: We need to talk about evaluation](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5404–5414, Online. Association for Computational Linguistics.
- 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 12th Language Resources and Evaluation Conference*, pages 1152–1158.
- René Knaebel. 2021. [discopy: A neural system for shallow discourse parsing](#). In *Proceedings of the 2nd Workshop on Computational Approaches to Discourse*, pages 128–133, Punta Cana, Dominican Republic and Online. Association for Computational Linguistics.
- Murathan Kurfalı. 2022. *Contributions to Shallow Discourse Parsing: To English and beyond*. Ph.D. thesis, Department of Linguistics, Stockholm University.
- Murathan Kurfalı and Robert Östling. 2021a. Let’s be explicit about that: Distant supervision for implicit discourse relation classification via connective prediction. *arXiv preprint arXiv:2106.03192*.
- Murathan Kurfalı and Robert Östling. 2021b. Probing multilingual language models for discourse. In *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*, pages 8–19.
- Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*.
- Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.
- Jiaqi Li, Ming Liu, Bing Qin, and Ting Liu. 2022. A survey of discourse parsing. *Frontiers of Computer Science*, 16(5):1–12.
- Ziheng Lin, Hwee Tou Ng, and Min-Yen Kan. 2014. A pdtb-styled end-to-end discourse parser. *Natural Language Engineering*, 20(2):151–184.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.
- Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.
- Kashif Munir, Hongxiao Bai, Hai Zhao, and Junhan Zhao. 2021. Memorizing all for implicit discourse relation recognition. *Transactions on Asian and Low-Resource Language Information Processing*, 21(3):1–20.
- 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 H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.
- Emily Pitler and Ani Nenkova. 2009. Using syntax to disambiguate explicit discourse connectives in text. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 13–16, Suntec, Singapore. Association for Computational Linguistics.
- Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltasakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The penn discourse treebank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*.
- 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.
- Attapol Rutherford, Vera Demberg, and Nianwen Xue. 2017. A systematic study of neural discourse models for implicit discourse relation. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 281–291.
- Wei Shi and Vera Demberg. 2017. On the need of cross validation for discourse relation classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 150–156, Valencia, Spain. Association for Computational Linguistics.
- Wei Shi and Vera Demberg. 2019a. Learning to explicate 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 and Vera Demberg. 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 5790–5796.
- Damien Sileo, Tim Van-de Cruys, Camille Pradel, and Philippe Muller. 2019. Discourse-based evaluation of language understanding. *arXiv preprint arXiv:1907.08672*.
- Wassiliki Siskou, Clara Giralt Mirón, Sarah Molinara-Raith, and Miriam Butt. 2022. Automated detection and annotation for calls to action in Latin-American social media postings. In *Proceedings of the 6th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 65–69, Gyeongju, Republic of Korea. International Conference on Computational Linguistics.
- Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to fine-tune bert for text classification? In *China national conference on Chinese computational linguistics*, pages 194–206. Springer.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.
- Alex Wang, Jan Hula, Patrick Xia, Raghavendra Pappagari, R. Thomas McCoy, Roma Patel, Najoung Kim, Ian Tenney, Yinghui Huang, Katherin Yu, Shuning Jin, Berlin Chen, Benjamin Van Durme, Edouard Grave, Ellie Pavlick, and Samuel R. Bowman. 2019. Can you tell me how to get past sesame street? sentence-level pretraining beyond language modeling. In *Proceedings of the 57th Annual Meeting of*

the Association for Computational Linguistics, pages 4465–4476, Florence, Italy. Association for Computational Linguistics.

Bonnie Webber, Rashmi Prasad, Alan Lee, and Aravind Joshi. 2019. The penn discourse treebank 3.0 annotation manual. *Philadelphia, University of Pennsylvania*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface’s transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Nianwen Xue, Hwee Tou Ng, Sameer Pradhan, Rashmi Prasad, Christopher Bryant, and Attapol Rutherford. 2015. The conll-2015 shared task on shallow discourse parsing. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning-Shared Task*, pages 1–16.

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.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. 2020. Big bird: Transformers for longer sequences. *Advances in Neural Information Processing Systems*, 33:17283–17297.

Zheng Zhao and Bonnie Webber. 2022. Revisiting shallow discourse parsing in the pdtb-3: Handling intra-sentential implicits. *arXiv preprint arXiv:2204.00350*.

A “Full sentence(s)” Experiment

A.1 What Makes the Experiment Important?

This section is a continuation of §3.4. Here, we discuss implicit discourse relation classification from raw sentence(s), which we believe is the best practical example of real-world applications of the related fields. Such an *end-to-end* concept has been popularized through CoNLL-2016 (Xue et al., 2016) and CoNLL-2015 (Xue et al., 2015), and most systems develop a separate argument span identification model. Then, the identified argument spans would be fed to the discourse relation classification model for sense prediction (refer to examples given in §1) (He et al., 2020b).

Such a double-step process makes sense. Indeed, feeding the exact argument spans (that only contain the tokens that imply a certain discourse sense) will increase sense prediction performance.

But the problem arises because identifying argument spans from raw sentence(s) is a low accuracy operation (Knaebel, 2021). A wrong span identification eventually leads to error propagation, providing a discourse relation classification model that lacks textual information. We give a theoretical error propagation example and conduct a simple experiment to prove our point.

A.2 Theoretical Example of Error Propagation

1. A set of two raw sentences is given.

"Last year we probably bought one out of every three new deals," he says. "This year, at best, it's in one in every five or six."

2. Where correct argument spans are as below.

["Last year we probably bought one out of every three new deals,]_{Arg1}" he says. "[This year, at best, it's in one in every five or six.]"_{Arg2}

3. But an argument span identification model often makes wrong predictions (best system (?) at CoNLL-2016 scores 52.02 F1, for exact span match).

["Last year we probably bought one]"_{Arg1} out of every three new deals," he says. "This year, at best, [it's in one in every five or six.]"_{Arg2}

4. Now, compare the amount of textual information passed over to the implicit discourse relation classification model, under three setups. Note that setup 1 cannot be used in real-world settings because it requires PDTB-3’s gold annotations.

Setup 1) PDTB-3 (with gold annotations)

Last year we probably bought one out of every three new deals This year, at best, it's in one in every five or six.

Setup 2) A low accuracy argument span model

Last year we probably bought one it's in one in every five or six.

Fine-tuned PLM	Argument Span	
	ACC	F1
BERT _{large}	0.912	0.742

Table 4: BERT’s performance (12-folds test set) on PDTB-3’s argument spans.

Setup 3) Full sentence(s)

"Last year we probably bought one out of every three new deals," he says. "This year, at best, it’s in one in every five or six."

A.3 Experiment on Error Propagation

Though not all tokens are valuable under a full sentence(s) setup, we can notice that it is a foolproof way to input all meaningful tokens. Table 4 reports the classification performance of BERT_{large}, which was trained to identify argument spans using PDTB-3. Our argument span scoring scheme approximately matches CoNLL-16’s partial scoring scheme, essentially a relaxed version of conlleval. That means we consider a prediction correct if more than 70% of argument span tokens are identified. For implicit discourse relation classification, a sense prediction is correct if it matches any of the multiply-annotated senses.

BERT’s 0.912 ACC score implies that the model could correctly identify at least 70% of the gold argument span tokens more than 9 out of 10 times. Nonetheless, error propagation detrimentally affected implicit discourse relation classification performance in Table 5. This empirically proves our ideas in Appendix A.1.

Fine-tuned PLM	Implicit Sense	
	ACC	F1
DeBERTa _{large}	0.670	0.671
<i>with error propagation</i>	0.476	0.491
<i>full sentence(s)</i>	0.634	0.637

Table 5: DeBERTa performances (12-fold test set) on PDTB-3’s Level-2 14-way implicit discourse relation classification, but under three different pipeline setups.

B 14-way Label Set

C More on Fine-tuning Set Up

We ran all our experiments on a single NVIDIA Tesla V100 GPU. Model train time and repositories are listed below. Training times below suppose no

Label	Counts
Comparison.Cession	1494
Comparison.Contrast	983
Contingency.Cause	5785
Contingency.Cause+Belief	202
Contingency.Condition	199
Contingency.Purpose	1373
Expansion.Conjunction	4386
Expansion.Equivalence	336
Expansion.Instantiation	1533
Expansion.Level-of-detail	3361
Expansion.Manner	739
Expansion.Substitution	450
Temporal.Asynchronous	1289
Temporal.Synchronous	539

Table 6: Counts of 14-way implicit discourse senses.

early stop. The performances reported in Table 2 are obtained **with** early stop.

ALBERT_{large}

- huggingface.co/albert-large-v1
- ~2.4 days, for 12 folds × 10 epochs

BART_{large}

- huggingface.co/facebook/bart-large
- ~3.6 days, for 12 folds × 10 epochs

BigBird-RoBERTa_{large}

- huggingface.co/google/bigbird-roberta-large
- ~3.2 days, for 12 folds × 10 epochs

DeBERTa_{large}

- huggingface.co/microsoft/deberta-large
- ~4.6 days, for 12 folds × 10 epochs

Longformer_{large}

- huggingface.co/allenai/longformer-large-4096
- ~11 days, for 12 folds × 10 epochs

RoBERTa_{large}

- huggingface.co/roberta-large
- ~2.9 days, for 12 folds × 10 epochs

SpanBERT_{large}

- .../SpanBERT/spanbert-large-cased
- ~2.9 days, for 12 folds × 10 epochs