Contextualized and Generalized Sentence Representations by Contrastive Self-Supervised Learning: A Case Study on Discourse Relation Analysis

Hirokazu Kiyomaru and **Sadao Kurohashi**
Graduate School of Informatics, Kyoto University
Yoshida-honmachi, Sakyo-ku, Kyoto, 606-8501, Japan
{kiyomaru, kuro}@nlp.ist.i.kyoto-u.ac.jp

Abstract

We propose a method to learn contextualized and generalized sentence representations using contrastive self-supervised learning. In the proposed method, a model is given a text consisting of multiple sentences. One sentence is randomly selected as a target sentence. The model is trained to maximize the similarity between the representation of the target sentence with its context and that of the masked target sentence with the same context. Simultaneously, the model minimize the similarity between the latter representation and the representation of a random sentence with the same context. We apply our method to discourse relation analysis in English and Japanese and show that it outperforms strong baseline methods based on BERT, XLNet, and RoBERTa.

1 Introduction

Understanding the meaning of a sentence is one of the main interests of natural language processing. In recent years, distributed representations are considered to be promising to capture the meaning of a sentence flexibly (Conneau et al., 2017; Arora et al., 2017; Kiros et al., 2015).

One typical way to obtain distributed sentence representations is to learn a task that is somehow related to sentence meaning. For example, sentence representations trained to solve natural language inference (Bowman et al., 2015; Williams et al., 2018) are known to be helpful for many language understanding tasks such as sentiment analysis and semantic textual similarity (Conneau et al., 2017; Wieting and Gimpel, 2018; Cer et al., 2018; Reimers and Gurevych, 2019).

However, there is an arbitrariness in the choice of tasks used for training. Furthermore, there is a size limitation on manually annotated data, which makes it hard to learn a wide range of language expressions.

A solution to these problems is self-supervised learning, which has been used with great success (Mikolov et al., 2013; Peters et al., 2018; Devlin et al., 2019). For example, inspired by skip-grams (Mikolov et al., 2013), Kiros et al. (2015) proposed to train a sequence-to-sequence model to generate sentences before and after a sentence, and use the encoder to compute sentence representations. Inspired by masked language modeling in BERT, Zhang et al. (2019) and Huang et al. (2020) presented methods to learn contextualized sentence representations through the task of restoring a masked sentence from its context.

In self-supervised sentence representation learning, sentence generation is typically used as its objective. Such an objective aims to learn a sentence representation specific enough to restore the sentence, including minor details. On the other hand, in case we would like to handle the meaning of a larger block such as paragraphs and documents (which is often called context analysis) and consider sentences as a basic unit, a more abstract and generalized sentence representation would be helpful.

We propose a method to learn contextualized and generalized sentence representations by contrastive self-supervised learning (van den Oord et al., 2019; Chen et al., 2020). In the proposed method, a model is given a text consisting of multiple sentences and computes their contextualized sentence representations. During training, one sentence is randomly selected as a *target sentence*. The model is trained to maximize the similarity between the representation of the target sentence with its context, to which we refer as s_{pos}, and the representation of the masked target sentence with the same context, to which we refer as s_{anc}. Simultaneously, the model is trained to minimize the similarity between the latter representation s_{anc} and the representation of a random sentence with the same context as the target sentence, to which we refer as s_{neg}.

Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 5578–5584
June 6–11, 2021. ©2021 Association for Computational Linguistics

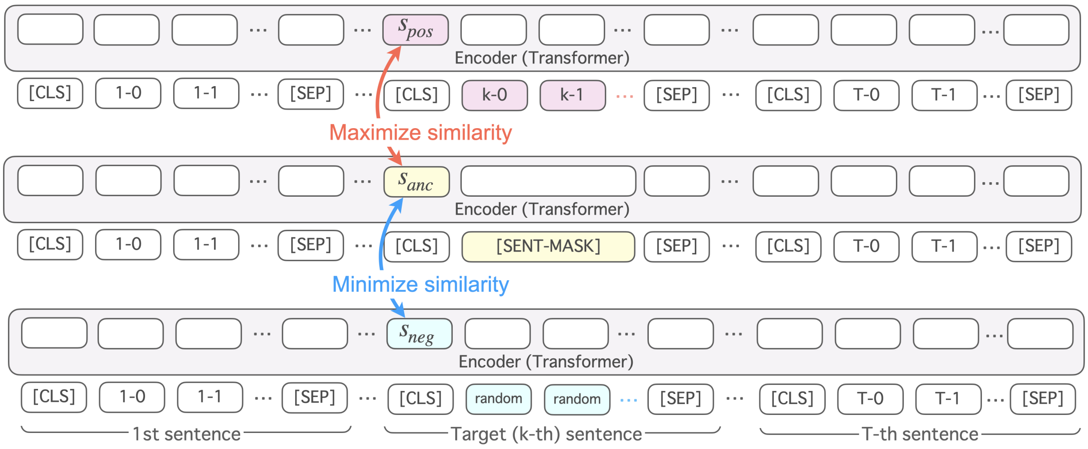

Figure 1: Overview of our method.

From the viewpoint of optimizing s_{anc}, this can be seen as a task to capture a generalized meaning that contextually valid sentences commonly have, utilizing s_{pos} and s_{neg} as clues. From the viewpoint of optimizing s_{pos}, this can be seen as a task to generalize the meaning of a sentence to the level of s_{anc}.

We show the effectiveness of the proposed method using discourse relation analysis as an example task of context analysis. Our experiments on English and Japanese datasets show that our method outperforms strong baseline methods based on BERT (Devlin et al., 2019), XLNet (Yang et al., 2019), and RoBERTa (Liu et al., 2019).

2 Learning Contextualized Sentence Representations

Figure 1 illustrates the overview of our method. The encoder takes an input text consisting of T (> 1) sentences and computes their contextualized sentence representations. The encoder is trained by contrastive self-supervised learning.

2.1 Encoder

The encoder is a Transformer (Vaswani et al., 2017) with the same architecture as BERT (Devlin et al., 2019). Following Liu and Lapata (2019), we insert the ⟨CLS⟩ and ⟨SEP⟩ tokens at the beginning and the end of each sentence, respectively. The representation of the ⟨CLS⟩ token is used as the sentence representation of its following sentence.

2.2 Contrastive Objective

We propose a contrastive objective to learn contextualized sentence representations, aiming to capture sentences' generalized meaning.

We first randomly select one sentence from the input text as a *target sentence*. In Figure 1, the k-th sentence ($1 \leq k \leq T$) is selected as a target sentence. We refer to the representation of the target sentence as s_{pos}. We then create another input text by masking the target sentence with the ⟨SENT-MASK⟩ token. We refer to the representation of the masked sentence as s_{anc}. We finally create yet another input text by replacing the target sentence with a random sentence. We refer to the representation of the replaced random sentence as s_{neg}.

Our contrastive objective is to maximize the similarity between s_{pos} and s_{anc} while minimizing the similarity between s_{neg} and s_{anc}. We use the dot product as the similarity measure. When using N random sentences per input text, the contrastive loss $\mathcal{L}$ is calculated as follows:

$$\mathcal{L} = -\log \frac{\exp(\langle s_{pos}, s_{anc} \rangle)}{\sum_{s \in \mathcal{S}} \exp(\langle s, s_{anc} \rangle)}, \quad (1)$$

where $\langle \cdot, \cdot \rangle$ is the dot product and $\mathcal{S} = \{s_{pos}, s_{neg}^1, \cdots, s_{neg}^N\}$.

To optimize s_{anc}, the model needs to capture a generalized meaning that contextually valid sentences commonly have, using s_{pos} and s_{neg} as clues. On the other hand, to optimize s_{pos}, the model needs to generalize the meaning of a sentence to the level of s_{anc}.

The encoder is trained by optimizing the contrastive loss and the standard masked language modeling loss (Devlin et al., 2019) jointly.

2.3 Generative Objective

For comparison, we train the encoder through the task of generating a masked sentence from its context. We first mask a sentence in the input text with the ⟨SENT-MASK⟩ token. Given the text, the encoder computes the representation of the masked sentence. Then, given the representation, a decoder generates the masked sentence in an autoregressive manner. The decoder's architecture is almost the same as the encoder, but it has an additional layer on the top to predict a probability distribution over words. We use teacher forcing and compute the generative loss by summing cross-entropy at each generation step.

The encoder and decoder are trained by optimizing the generative loss and the standard masked language modeling loss jointly.

2.4 Implementation Details

2.4.1 English

We use an English Wikipedia dump and BookCorpus (Zhu et al., 2015)[1] to create input texts. We first split texts into sentences using spacy (Honnibal et al., 2020). We then extract as many consecutive sentences as possible so that the length does not exceed the maximum input length of 128. When a sentence is so long that an input text including the sentence cannot be created while meeting the length constraint, we give up using the sentence. The number of sentences in an input text T was 4.91 on average. After creating input texts, we assign random sentences to each of them. Random sentences are extracted from the same document. We assigned three random sentences per input text, i.e., $N = 3$.

We initialize the encoder's parameters using the weights of RoBERTa$_{\text{BASE}}$ (Liu et al., 2019). The other parameters are initialized randomly. We train the model for 10,000 steps with a batch size of 512. We use the Adam optimizer (Kingma and Ba, 2015) with a learning rate of 2e-5, $\beta_1 = 0.9$, $\beta_2 = 0.999$, linear warmup of the learning rate over the first 1,000 steps, and linear decay of the learning rate.

2.4.2 Japanese

We use a Japanese Wikipedia dump to create input texts. We split the texts into clauses using KNP, a widely used Japanese syntactic parser (Kawahara and Kurohashi, 2006). We create input texts and assign random sentences to them in the same way as in Section 2.4.1. The number of sentences (clauses) in an input text T was 6.42 on average.

We initialize the encoder's parameters with BERT$_{\text{BASE}}$, pretrained on a Japanese Wikipedia dump[2]. The other details are the same as in Section 2.4.1.

3 Discourse Relation Analysis

We show the effectiveness of the proposed method using discourse relation analysis as a concrete example of context analysis. Discourse relation analysis is a task to predict the logical relation between two arguments. An argument roughly corresponds to a sentence or a clause. We conduct experiments on English and Japanese datasets.

3.1 Datasets

3.1.1 Penn Discourse Tree Bank (PDTB) 3.0

PDTB 3.0 is a corpus of English newspaper with discourse relation labels (Prasad et al., 2018). We focus on implicit discourse relation analysis, where no explicit discourse marker exists. Following Kim et al. (2020), we use the Level-2 labels with more than 100 examples and use 12-fold cross-validation.

3.1.2 Kyoto University Web Document Leads Corpus (KWDLC)

KWDLC is a Japanese corpus consisting of leading three sentences of web documents with discourse relation labels (Kawahara et al., 2014; Kishimoto et al., 2018). As KWDLC does not discriminate between implicit discourse relations and explicit discourse relations, we target both. KWDLC has seven types of discourse relations, including NoRelation. The evaluation protocol is 5-fold cross-validation. Following Kim et al. (2020), each fold is split at the document level rather than the individual example level.

3.2 Model

We train two types of models; one uses the context of arguments, and the other does not.

When a model uses context, the model is given the paragraph that contains arguments of interest. In this setting, first, the paragraph is split into sentences. Arguments are treated as a single sentence, and their context is split in the way described in Section 2.4. Then, an encoder computes the representation of each sentence in the same manner as

[1] Because the original BookCorpus is no longer available, we used a replica created by a publicly available crawler (https://github.com/soskek/bookcorpus).

[2] Available at https://alaginrc.nict.go.jp/nict-bert/index.html.

Context	Encoder	Acc
Unused	BERT$_{\text{BASE}}$ (Kim et al., 2020)	57.60
	XLNet$_{\text{BASE}}$ (Kim et al., 2020)	60.78
	RoBERTa$_{\text{BASE}}$	61.68 ± 1.63
Used	BERT$_{\text{BASE}}$	56.83 ± 1.43
	RoBERTa$_{\text{BASE}}$	62.25 ± 1.47
	RoBERTa$_{\text{BASE}}$ + Gen	62.19 ± 1.33
	RoBERTa$_{\text{BASE}}$ + Con (ours)	**63.30 ± 1.42**

Table 1: Results of implicit discourse relation analysis on PDTB 3.0 using the Level-2 label set (Kim et al., 2020). **Gen** and **Con** indicate that the encoder is further pretrained by optimizing the generative objective and the contrastive objective, respectively. The scores are the mean and standard deviation over folds.

in Section 2.1. Given the concatenation of the arguments' representations, a relation classifier predicts the discourse relation. As a relation classifier, we employ a multi-layer perceptron with one hidden layer and ReLU activation.

When a model does not use context, the model is given arguments of interest only. In this setting, we use the sentence pair classification method proposed by Devlin et al. (2019).

Our proposed method is introduced to a context-using model by initializing its encoder's parameters using our sentence encoder. In experiments, we report a difference in performance depending on models used for initialization.

3.3 Implementation Details

Input texts are truncated to the maximum input length of 512, which is long enough to hold almost all inputs. We train models for up to 20 epochs. At the end of each epoch, we compute the performance for the development data and adopt the model with the best performance. If the performance does not improve for five epochs, we stop the training. We use the Adam optimizer with a learning rate of 2e-5, $\beta_1 = 0.9$, $\beta_2 = 0.999$. We update all the parameters in models, i.e., pretrained sentence encoders are fine-tuned to solve discourse relation analysis.

3.4 Results

Table 1 shows the result for PDTB 3.0. The evaluation metric is accuracy. The highest performance was achieved by the proposed method. To our knowledge, this is the state-of-the-art performance among models with the same parameter size as BERT$_{\text{BASE}}$. The model that optimized the generative objective was inferior not only to the proposed method but also to vanilla RoBERTa with context.

Table 2 shows the result for KWDLC. The evaluation metrics are accuracy and micro-averaged precision, recall, and F1[3]. The highest performance was again achieved by the proposed method. The decrease in performance by optimizing the generative objective is consistent with the experimental results on PDTB 3.0.

3.5 Qualitative Analysis

We show an example of discourse relation analysis in KWDLC.

(1) ⟨$_{\text{Arg1}}$ 新潟県にある国営公園・越後丘陵公園へ、１泊で遊びに出掛けようと⟩ ⟨$_{\text{Arg2}}$ 思い立ちました。⟩
⟨$_{\text{Arg1}}$ I want to go to a government-managed park in Niigata Prefecture for an overnight visit,⟩ ⟨$_{\text{Arg2}}$ I came up with that.⟩
Label: NoRelation

Arguments are enclosed in ⟨ and ⟩. The models except ours erroneously predicted the discourse relation of Purpose between Arg1 and Arg2. This is probably because the Japanese postpositional particle "と" can be a discourse marker of Purpose. For example, if Arg2 was "荷造りを始めた (I started packing)," the prediction would be correct. However, in this case, the postpositional particle "と" is used to construct a sentential complement. That is, Arg1 is the object of Arg2. It is not possible to distinguish between the two usages from its surface form. Our model correctly predicted the discourse relation of NoRelation, which implies that our method understood that Arg1 is a sentential complement.

We show another example of implicit discourse relation analysis in KWDLC.

(2) ⟨$_{\text{Arg1}}$ 以前から計画していたホームページを開設することができ、⟩ ⟨$_{\text{Arg2}}$ 嬉しいかぎりである。⟩
⟨$_{\text{Arg1}}$ I was able to launch the website that I had planned for a while,⟩ ⟨$_{\text{Arg2}}$ I'm happy.⟩
Label: Cause/Reason

While most models predicted the discourse relation of NoRelation between Arg1 and Arg2, the

[3] As examples with the discourse relation of NoRelation accounts for more than 80% of the dataset, precision, recall, and F1 are calculated without examples with NoRelation to make performance difference intelligible.

Context	Encoder	Acc	Prec	Rec	F1
Unused	BERT$_{\text{BASE}}$	80.68 ± 1.59	45.90 ± 4.06	**41.42 ± 8.35**	43.37 ± 6.28
Used	BERT$_{\text{BASE}}$	84.36 ± 2.05	62.55 ± 10.26	39.13 ± 7.38	47.67 ± 6.68
	BERT$_{\text{BASE}}$ + Gen	84.16 ± 1.60	57.84 ± 8.51	40.13 ± 0.42	47.21 ± 2.68
	BERT$_{\text{BASE}}$ + Con (ours)	**85.02 ± 1.85**	**63.51 ± 5.90**	41.04 ± 4.24	**49.74 ± 4.11**

Table 2: Results of discourse relation analysis on KWDLC. The scores are the mean and standard deviation over folds.

Query:	⟨The Beatles were an English rock band formed in Liverpool in 1960.⟩ ⟨**The group, whose best-known line-up comprised John Lennon, Paul McCartney, George Harrison and Ringo Starr, are regarded as the most influential band of all time.**⟩
Retrieved:	1) ⟨Britney Jean Spears (born December 2, 1981) is an American singer, songwriter, dancer, and actress.⟩ ⟨**She is credited with influencing the revival of teen pop during the late 1990s and early 2000s, for which she is referred to as the "Princess of Pop".** ⟩ ... 2) ⟨Dynasty was an American band, based in Los Angeles, California, created by producer and SOLAR Records label head Dick Griffey, and record producer Leon Sylvers III.⟩ ⟨**The band was known for their dance/pop numbers during the late 1970s and 1980s.**⟩ ... 3) ⟨Lu Ban (–444BC) was a Chinese structural engineer, inventor, and carpenter during the Zhou Dynasty.⟩ ⟨**He is revered as the Chinese god (patron) of builders and contractors.**⟩ ... 10) ⟨Stacey Park Milbern (May 19, 1987 – May 19, 2020) was an American disability rights activist.⟩ ⟨**She helped create the disability justice movement and advocated for fair treatment of people with disabilities.**⟩ ... 20) ⟨The National Action Party (, PAN) is a conservative political party in Mexico founded in 1938.⟩ ⟨**The party is one of the four main political parties in Mexico, and, since the 1980s, has had success winning local, state, and national elections.**⟩ ...

Table 3: Results of sentence retrieval based on the cosine similarity between sentence representations computed by our method. ⟨·⟩ indicates a sentence. The query and retrieved sentences are marked in bold, and their contexts are shown together. The numbers indicate the rank of sentence retrieval.

proposed model correctly recognized the discourse relation of CAUSE/REASON. We speculate that the models other than ours failed to understand Arg1 at the level of "a happy event occurred."

4 Sentence Retrieval

To investigate what is learned by our contrastive objective, we did sentence retrieval based on the similarity between sentence representations. For targets, we randomly sampled 500,000 sentences with context from input texts used for training. For a query, we used a sentence with context in a Wikipedia article. Computing the sentence representations for the targets and query, we searched the closest sentences based on their cosine similarity.

Table 3 shows an example. In addition to the top-ranked sentences, we also picked up some highly-ranked sentences. The top two sentences were very similar to the query sentence regarding the topic, meaning, and context. While the sentences of lower rank had different topics from the query sentence, they all described a positive aspect of an entity and had a similar context in terms of that an entity is introduced in their preceding sentences. We confirmed that almost the same results were obtained in Japanese. We leave quantitative evaluation of sentence retrieval for future work.

5 Conclusion

We proposed a method to learn contextualized and generalized sentence representations using contrastive self-supervised training. Experiments showed that the proposed method improves the performance of discourse relation analysis both in English and Japanese. We leave an in-depth analysis of the level of abstraction trained by the proposed method for future work.

References

Sanjeev Arora, Yingyu Liang, and Tengyu Ma. 2017. A simple but tough-to-beat baseline for sentence embeddings. In *Proceedings of the Fifth International Conference on Learning Representations (ICLR)*.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages

632–642, Lisbon, Portugal. Association for Computational Linguistics.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Brian Strope, and Ray Kurzweil. 2018. Universal sentence encoder for English. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 169–174, Brussels, Belgium. Association for Computational Linguistics.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680, Copenhagen, Denmark. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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

Yichen Huang, Yizhe Zhang, Oussama Elachqar, and Yu Cheng. 2020. INSET: Sentence infilling with INter-SEntential transformer. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2502–2515, Online. Association for Computational Linguistics.

Daisuke Kawahara and Sadao Kurohashi. 2006. A fully-lexicalized probabilistic model for Japanese syntactic and case structure analysis. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 176–183, New York City, USA. Association for Computational Linguistics.

Daisuke Kawahara, Yuichiro Machida, Tomohide Shibata, Sadao Kurohashi, Hayato Kobayashi, and Manabu Sassano. 2014. Rapid development of a corpus with discourse annotations using two-stage crowdsourcing. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 269–278, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

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.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the Third International Conference on Learning Representations (ICLR)*.

Ryan Kiros, Yukun Zhu, Russ R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. In *Advances in Neural Information Processing Systems*, volume 28, pages 3294–3302. Curran Associates, Inc.

Yudai Kishimoto, Shinnosuke Sawada, Yugo Murawaki, Daisuke Kawahara, and Sadao Kurohashi. 2018. Improving crowdsourcing-based annotation of Japanese discourse relations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. 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 3730–3740, Hong Kong, China. Association for Computational Linguistics.

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*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, volume 26, pages 3111–3119. Curran Associates, Inc.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Rashmi Prasad, Bonnie Webber, and Alan Lee. 2018. Discourse annotation in the PDTB: The next generation. In *Proceedings 14th Joint ACL - ISO Workshop on Interoperable Semantic Annotation*, pages 87–97, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2019. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

John Wieting and Kevin Gimpel. 2018. ParaNMT-50M: Pushing the limits of paraphrastic sentence embeddings with millions of machine translations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 451–462, Melbourne, Australia. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems*, volume 32, pages 5753–5763. Curran Associates, Inc.

Xingxing Zhang, Furu Wei, and Ming Zhou. 2019. HIBERT: Document level pre-training of hierarchical bidirectional transformers for document summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5059–5069, Florence, Italy. Association for Computational Linguistics.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *The IEEE International Conference on Computer Vision (ICCV)*.