Paraphrase-based Contrastive Learning for Sentence Pair Modeling

Seiji Sugiyama Risa Kondo Tomoyuki Kajiwara Takashi Ninomiya
Graduate School of Science and Engineering, Ehime University, Japan
sugiyama@ai.cs.ehime-u.ac.jp kondo@ai.cs.ehime-u.ac.jp
kajiwara@cs.ehime-u.ac.jp ninomiya.takashi.mk@ehime-u.ac.jp

Abstract

To improve the performance of sentence pair modeling tasks, we propose an additional pre-training method, also known as transfer fine-tuning, for pre-trained masked language models. Pre-training for masked language modeling is not necessarily designed to bring semantically similar sentences closer together in the embedding space. Our proposed method aims to improve the performance of sentence pair modeling by applying contrastive learning to pre-trained masked language models, in which sentence embeddings of paraphrase pairs are made similar to each other. While natural language inference corpora, which are standard in previous studies on contrastive learning, are not available on a large-scale for non-English languages, our method can construct a training corpus for contrastive learning from a raw corpus and a paraphrase dictionary at a low cost. Experimental results on four sentence pair modeling tasks revealed the effectiveness of our method in both English and Japanese.

1 Introduction

Sentence pair modeling (Lan and Xu, 2018), which estimates the relationship between two texts, is an important technique for various natural language processing tasks, from semantic textual similarity estimation (Cer et al., 2017) and recognizing textual entailment (Bowman et al., 2015) to information retrieval (Wang et al., 2024) and question answering (Zhang et al., 2023). For sentence pair modeling tasks, surface matching such as bag-of-words and word embeddings such as word2vec (Mikolov et al., 2013) have traditionally been used, followed by task-specific neural networks (He and Lin, 2016; Chen et al., 2017), and recently fine-tuning pre-trained masked language models such as BERT (Devlin et al., 2019) has become the de facto standard. However, training in masked language modeling does not necessarily bring semantically similar sentences closer together in the embedding space (Li et al., 2020). Therefore, to maximize the effectiveness of fine-tuning for sentence pair modeling tasks, it is useful to follow the pre-training of masked language modeling with additional pre-training to estimate the semantic relationships between texts, also known as transfer fine-tuning (Arase and Tsujii, 2019).

One such method recently been attracting attention is contrastive learning. Contrastive learning for sentence embeddings, like SimCSE (Gao et al., 2021; Chuang et al., 2022; Liu et al., 2023), uses annotated corpora of natural language inference (NLI) to bring embeddings of entailing and entailed sentences closer together and to separate embeddings of contradictory sentence pairs. However, while NLI corpora with hundreds of thousands of sentence pairs, such as Stanford NLI (SNLI) (Bowman et al., 2015) and Multi-Genre NLI (MNLI) (Williams et al., 2018), are available for English, there are no large-scale NLI corpora for other languages, making it difficult to obtain high-quality sentence embeddings by contrastive learning for languages other than English.

To improve the performance of sentence pair modeling in various languages, we propose a method of contrastive learning that does not rely on the NLI corpus. Our method uses a raw corpus and a paraphrase dictionary to automatically generate a large-scale training corpus for contrastive learning at a low cost. Since paraphrase dictionaries are available in many languages,[1] this method is widely applicable.

Experimental results in English and Japanese revealed that the proposed method could improve the performance of the masked language models in four types of sentence pair modeling tasks (product retrieval, similarity estimation, recognizing textual entailment, and paraphrase identifi-

[1] For example, the Multilingual PPDB (Ganitkevitch and Callison-Burch, 2014) collects millions to hundreds of millions of paraphrase pairs in 23 languages.

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop), pages 400–407

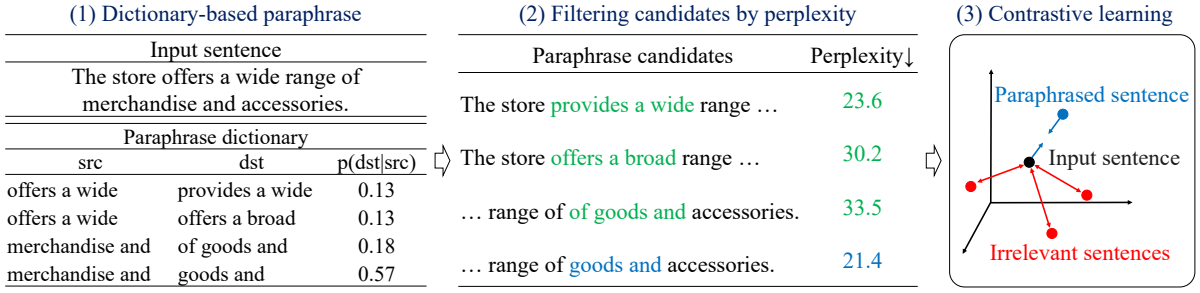

Figure 1: Overview of our paraphrase-based contrastive learning.

cation). Regarding the average performance of all tasks, the proposed method achieved the best performance for both English and Japanese compared to existing methods that learn paraphrases but do not use contrastive learning (Arase and Tsujii, 2019), contrastive learning with raw corpus or NLI corpus (Gao et al., 2021), and state-of-the-art RankCSE (Liu et al., 2023).

2 Related Work

2.1 Contrastive Learning

Contrastive learning is a method that brings semantically close data closer together in vector space and separates semantically distant data apart in vector space. Methods for acquiring sentence embeddings by applying contrastive learning to pre-trained masked language models have been actively studied in recent years (Gao et al., 2021; Chuang et al., 2022; Liu et al., 2023).

Previous studies of sentence embedding based on contrastive learning have relied on the NLI corpus (Bowman et al., 2015), in which sentence pairs are labeled with semantic relations of entailment, contradiction, and neutral, for training. However, annotating such corpora in non-English languages at high-quality and on a large-scale is expensive, so this study proposes a lower-cost alternative.

2.2 Paraphrasing for Additional Training

Paraphrasing is the task of generating text that is semantically equivalent to the input text. This technique can be applied to pre-editing (Mehta et al., 2020; Miyata and Fujita, 2021) and data augmentation (Effendi et al., 2018; Okur et al., 2022) to improve the performance of various natural language processing applications.

One such promising application of paraphrasing is additional training of pre-trained models. Pre-trained encoders can be additionally trained on the paraphrase identification task to increase the fine-tuning performance of similarity estimation and recognizing textual entailment (Arase and Tsujii, 2019). Similarly, pre-trained encoder-decoder models can be additionally trained on the paraphrase generation task to enhance the fine-tuning performance of style transfer and text simplification (Kajiwara et al., 2020). This study combines paraphrasing and contrastive learning to further improve additional training for pre-trained encoders.

3 Proposed Method

We improve the performance of sentence pair modeling with masked language models by contrastive learning that does not rely on the NLI corpus. As shown in the following steps, we boost the effectiveness of fine-tuning by conducting additional training between pre-training and fine-tuning.

1. Pre-training: masked language modeling
2. Our contrastive learning
3. Fine-tuning: supervised learning on the target task of sentence pair modeling

As shown in Figure 1, our contrastive learning uses paraphrase sentence pairs instead of entailment pairs in the NLI corpus. (1) Paraphrase an input sentence from the raw corpus based on the dictionary, (2) Select the most fluent paraphrase among the candidates, and (3) Conduct contrastive learning, which brings embeddings of the input sentence and the paraphrased sentence closer together and separates embeddings of the input sentence from the rest of the sentences in the batch.

3.1 Paraphrase-based Contrastive Learning

Although the proposed method employs the same contrastive learning loss as SimCSE (Gao et al., 2021), we use paraphrase sentence pairs (described in § 3.2) instead of entailment sentence pairs as positive instances for contrastive learning, and other

	Train	Dev	Test
Shopping Queries	1,254,438	138,625	425,762
STS-B	5,749	1,500	1,379
SICK	4,439	495	4,906
SNLI	549,367	9,842	9,824
PAWS	49,401	8,000	8,000

Table 1: Number of sentence pairs for English datasets.

	Train	Dev	Test
Shopping Queries	294,874	32,272	118,907
JSTS	11,205	1,246	1,457
JSICK	4,500	500	4,927
JNLI	18,065	2,008	2,434
PAWS-X	49,401	2,000	2,000

Table 2: Number of sentence pairs for Japanese datasets.

sentences in the batch instead of contradictory sentence pairs as negative instances. Since this study assumes no semantic relationship between sentences in a batch, the other sentences x_j in the batch work as negative instances that are semantically unrelated to the input sentence x_i. The paraphrase of the input sentence is x_i^+ and embeddings of the paraphrase pair are $\mathbf{h}_i$ and $\mathbf{h}_i^+$, respectively, and we train to minimize the loss function in Equation (1):

$$\ell_i = -\log \frac{e^{\mathrm{sim}(\mathbf{h}_i, \mathbf{h}_i^+)/\tau}}{\sum_{j=1}^{N} e^{\mathrm{sim}(\mathbf{h}_i, \mathbf{h}_j^+)/\tau}}, \quad (1)$$

where N is the batch size, τ is the temperature parameter and $\mathrm{sim}(\cdot)$ is the cosine similarity between sentence embeddings.

3.2 Generating Paraphrase Sentence Pairs

We automatically generate paraphrase sentence pairs to be used as positive instances for contrastive learning, from a raw corpus and a paraphrase dictionary.[2] Our paraphrase dictionary consists of three pairs of source phrase s, target phrase d, and paraphrase probability $p(d|s)$. In this study, we employ only paraphrase pairs $\{(s, d) \mid p(d|s) \geq \theta\}$ that have a paraphrase probability above a threshold θ. The paraphrase dictionary is applied to the input sentence from the raw corpus $x_i \in \mathcal{D}$, substituting phrases s into d to generate paraphrase candidates.

Here, as shown in Figure 1 (2), paraphrase candidates may include ungrammatical expressions. To avoid the negative effects from such ungrammatical sentences, we select the most fluent candidate with minimum perplexity to use as positive instances for contrastive learning.

[2]For paraphrase generation, we can also employ methods based on machine translation (Hu et al., 2019; Kajiwara et al., 2020) or large language models (Witteveen and Andrews, 2019). Comparison with them is left for our future work. In this study, we employ a dictionary-based paraphrase method that is computationally inexpensive and highly interpretable.

4 Experiments

We evaluated the effectiveness of the proposed method for four types of sentence pair modeling tasks in both English and Japanese.

4.1 Tasks

Our evaluation tasks are product retrieval, similarity estimation, recognizing textual entailment (RTE), and paraphrase identification. Statistics for each dataset are shown in Tables 1 and 2.

Retrieval Product retrieval is a four-class classification task of the relationships between product titles and their search queries, and we employed both English and Japanese versions of the Shopping Queries dataset[3] (Reddy et al., 2022).

Similarity Similarity estimation is a regression task that estimates the semantic similarity between two sentences, and we employed datasets of STS-B[4] (Cer et al., 2017) and SICK[5] (Marelli et al., 2014) for English and JSTS[6] (Kurihara et al., 2022) and JSICK[7] (Yanaka and Mineshima, 2022) for Japanese.

RTE RTE is a three-class classification task of semantic relationships between two sentences, and we employed datasets of SNLI[8] (Bowman et al., 2015) and SICK for English and JNLI[6] (Kurihara et al., 2022) and JSICK for Japanese.

Paraphrase Paraphrase identification is a two-class classification task of synonymity between two sentences, and we employed datasets of PAWS[9] (Zhang et al., 2019) for English and PAWS-X[9] (Yang et al., 2019) for Japanese.

[3]https://github.com/amazon-science/esci-data

[4]http://ixa2.si.ehu.es/stswiki/index.php/STSbenchmark

[5]https://zenodo.org/records/2787612

[6]https://github.com/yahoojapan/JGLUE

[7]https://github.com/verypluming/JSICK

[8]https://nlp.stanford.edu/projects/snli/

[9]https://github.com/google-research-datasets/paws

For evaluation metrics, we used the micro-f1 score for the retrieval tasks, Spearman's rank correlation coefficient for the similarity tasks, and the macro-f1 score for both the RTE and the paraphrase tasks.

4.2 Implementation Details

We fine-tuned English BERT[10] (Devlin et al., 2019) and Japanese RoBERTa[11] (Liu et al., 2019) on the sentence pair modeling tasks in Section 4.1. We want to evaluate whether the performance of each task can be improved by applying additional training of the proposed method or comparative methods before fine-tuning.

Pre-processing We used Wikipedia text from Wiki-40B[12] (Guo et al., 2020) for our contrastive learning. As pre-processing, we applied sentence segmentation and word segmentation with Moses[13] (Koehn et al., 2007) for English, and sentence segmentation with ja_sentence_segmenter[14] and word segmentation with MeCab (IPAdic)[15] (Kudo et al., 2004) for Japanese. In addition, language identification by langdetect[16] was performed, and only sentences with a confidence level of 99% or higher were used in each corpus for English and Japanese. Finally, we excluded both short sentences of 5 words or less and long sentences of 50 words or more.

Paraphrase For paraphrase dictionary, we used PPDB 2.0[17] (Pavlick et al., 2015) for English and EhiMerPPDB[18] for Japanese. These dictionaries cover phrases of up to six words in English and seven words in Japanese. To filter paraphrase candidates, perplexity was calculated with English[19] or Japanese[20] models of GPT-2 (Radford et al., 2019).

Hyperparameters The learning rate was set to 5×10^{-5}, temperature to $\tau = 0.05$, batch size to 64 sentence pairs, and Adam (Kingma and Ba, 2015) was used as our optimization method, and training was terminated when the loss on the Dev set did not improve for 3 consecutive epochs. In addition, we selected the threshold for paraphrase probability $\theta \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$ and the number of sentences for additional training $|\mathcal{D}| \in \{10k, 20k, 40k, 80k, 160k\}$ to maximize metrics on the Dev set among these combinations.

4.3 Comparative Methods

To evaluate the effectiveness of paraphrase-based contrastive learning, we compare the proposed method to existing methods that employ paraphrase but not contrastive learning (Transfer Fine-Tuning) (Arase and Tsujii, 2019), contrastive learning without paraphrase (both unsupervised and supervised SimCSE (Gao et al., 2021) and state-of-the-art RankCSE (Liu et al., 2023)), and fine-tuning without additional training.

Transfer Fine-Tuning (Arase and Tsujii, 2019) is a method for additional training to identify phrasal paraphrases on approximately 30 million paraphrase pairs. Since we use the official trained model[21] in English, it is compared only in English experiments. Unsupervised SimCSE (Gao et al., 2021) is dropout-based contrastive learning with raw corpora, and we replicate it with Wikipedia in the same settings as in § 4.2. Supervised SimCSE (Gao et al., 2021) is contrastive learning with NLI corpora, and we replicate it with SNLI (Bowman et al., 2015) and MNLI (Williams et al., 2018) for English, and with JSNLI,[22] a Japanese translation of SNLI, for Japanese. RankCSE (Liu et al., 2023) is a state-of-the-art contrastive learning that incorporates ranking consistency and ranking distillation, and we replicate it using the English[23] or Japanese[24] SimCSE as a teacher model. The hyperparameters of SimCSE and RankCSE are the same as those of the proposed method in § 4.2.[25]

4.4 Results

Experimental results are shown in Table 3. Our method achieved performance better than the base-

[10] https://huggingface.co/google-bert/bert-base-uncased
[11] https://huggingface.co/rinna/japanese-roberta-base
[12] https://www.tensorflow.org/datasets/catalog/wiki40b
[13] https://github.com/moses-smt/mosesdecoder/
[14] https://github.com/wwwcojp/ja_sentence_segmenter
[15] https://taku910.github.io/mecab/
[16] https://pypi.org/project/langdetect/
[17] http://paraphrase.org/#/download
[18] https://github.com/EhimeNLP/EhiMerPPDB
[19] https://huggingface.co/openai-community/gpt2
[20] https://huggingface.co/rinna/japanese-gpt2-medium
[21] https://github.com/yukiar/TransferFT
[22] https://nlp.ist.i.kyoto-u.ac.jp/?%E6%97%A5%E6%9C%AC%E8%AA%9ESNLI%28JSNLI%29%E3%83%87%E3%83%BC%E3%82%BF%E3%82%BB%E3%83%83%E3%83%88
[23] https://huggingface.co/princeton-nlp/unsup-simcse-bert-base-uncased
[24] https://huggingface.co/cl-nagoya/unsup-simcse-ja-base
[25] Since JSNLI has less than $160k$ sentence pairs, we set the maximum number for additional training to $140k$ pairs.

English	Retrieval	Similarity		RTE		Paraphrase	
	Shopping Queries	STS-B	SICK	SNLI	SICK	PAWS	Avg.
w/o Additional Training	0.654	0.824	0.815	**0.904**	0.858	0.913	0.828
Transfer Fine-Tuning	0.652	0.854	0.821	0.902	0.860	0.901	0.832
Unsupervised SimCSE	**0.655**	0.830	0.806	**0.904**	**0.868**	**0.918**	0.830
RankCSE	0.652	**0.858**	0.821	0.903	0.854	0.912	0.833
Supervised SimCSE	**0.655**	0.857	0.824	0.901	0.865	0.913	0.836
Ours	**0.655**	0.841	**0.842**	**0.904**	0.866	**0.918**	**0.838**
Japanese	Retrieval	Similarity		RTE		Paraphrase	
	Shopping Queries	JSTS	JSICK	JNLI	JSICK	PAWS-X	Avg.
w/o Additional Training	0.576	0.859	0.890	0.785	0.839	0.793	0.790
Unsupervised SimCSE	**0.587**	**0.861**	0.886	0.781	0.837	0.790	0.790
RankCSE	0.574	0.855	0.893	0.829	0.838	0.779	0.795
Supervised SimCSE	0.576	0.825	0.886	**0.843**	0.843	**0.800**	0.796
Ours	**0.587**	**0.861**	**0.896**	0.828	**0.856**	0.791	**0.803**

Table 3: Evaluation of four sentence pair modeling tasks. Retrieval is a product retrieval task and reports Micro-F1. Similarity is a semantic textual similarity estimation task and reports Spearman correlation. RTE and Paraphrase are tasks of recognizing textual entailment and paraphrase identification, respectively, and report Macro-F1.

line w/o additional training on all tasks in English, and better than the baseline on all tasks except the paraphrase identification task in Japanese. Here, PAWS-X focuses on word reordering, which may be incompatible with our paraphrase, which does not reorder but only substitutes phrases. Nevertheless, our method is effective for many other tasks.

Compared to the existing methods, our method achieved the best average performance in both English and Japanese. Our method has the advantage of achieving higher performance at a lower cost than traditional contrastive learning because it does not require expensive annotation like NLI corpora.

4.5 Analysis: Paraphrase Quality

The quality and quantity of paraphrases may affect the performance of our contrast learning. There is a trade-off between quality and quantity of paraphrases, which can be controlled using the paraphrase probabilities listed in the dictionary. In other words, if only paraphrases with high probability are targeted, a high quality and small quantity of paraphrases will be used.

The average performance of the sentence pair modeling tasks on the Dev set for each paraphrase probability threshold is shown in Figure 2. We found that the best performance was achieved by using only paraphrases with a probability of 0.4 or higher in both English and Japanese.

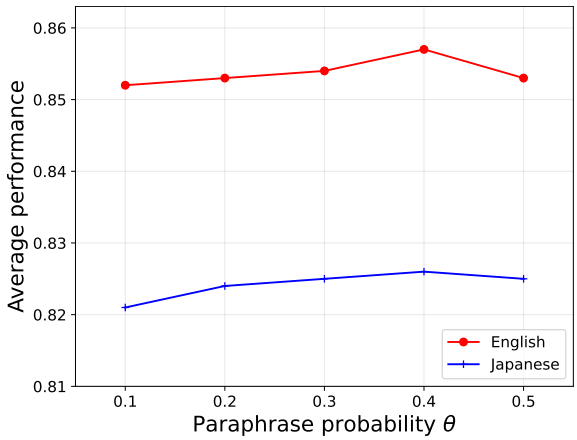

Figure 2: Relationship between paraphrase probability and average performance of sentence pair modeling. The larger the probability, the higher quality and smaller quantity of paraphrases we use in our training.

5 Summary and Future Work

In this study, we proposed paraphrase-based contrastive learning to improve the performance of sentence pair modeling. Our method can achieve high performance from automatically generated corpora, even though it is freed from the expensive annotation of NLI corpora that traditional contrastive learning relies on. Experimental results reveal performance improvements in a wide range of tasks, including product retrieval, similarity estimation, recognizing textual entailment, and paraphrase identification, in both English and Japanese.

Our future work includes both further improvement of positive instances and negative instances. Especially for positive instances, paraphrase generation could be based on machine translation or large language models.

Limitations

Language Dependency: While our method does not require expensive annotation like the NLI corpus, it relies on a raw corpus and a paraphrase dictionary. Even though paraphrase dictionaries already exist for many languages, they vary in size and quality. Since our experiments are conducted in two languages, English and Japanese, we cannot necessarily guarantee the effectiveness of the proposed method in other languages.

Training Time: We added a new training step between pre-training and fine-tuning of masked language models. This requires about 30 minutes of additional training time when running on a single NVIDIA RTX A6000 GPU with 48 GB memory.

Acknowledgments

The authors wish to thank Mercari Inc. and the Mercari R4D team for funding this research.

References

Yuki Arase and Jun'ichi Tsujii. 2019. Transfer Fine-Tuning: A BERT Case Study. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 5393–5404.

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.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Crosslingual Focused Evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation*, pages 1–14.

Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, Hui Jiang, and Diana Inkpen. 2017. Enhanced LSTM for Natural Language Inference. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 1657–1668.

Yung-Sung Chuang, Rumen Dangovski, Hongyin Luo, Yang Zhang, Shiyu Chang, Marin Soljacic, Shang-Wen Li, Scott Yih, Yoon Kim, and James Glass. 2022. DiffCSE: Difference-based Contrastive Learning for Sentence Embeddings. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4207–4218.

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*, pages 4171–4186.

Johanes Effendi, Sakriani Sakti, Katsuhito Sudoh, and Satoshi Nakamura. 2018. Multi-paraphrase Augmentation to Leverage Neural Caption Translation. In *Proceedings of the 15th International Conference on Spoken Language Translation*, pages 181–188.

Juri Ganitkevitch and Chris Callison-Burch. 2014. The Multilingual Paraphrase Database. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, pages 4276–4283.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple Contrastive Learning of Sentence Embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910.

Mandy Guo, Zihang Dai, Denny Vrandečić, and Rami Al-Rfou. 2020. Wiki-40B: Multilingual Language Model Dataset. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2440–2452.

Hua He and Jimmy Lin. 2016. Pairwise Word Interaction Modeling with Deep Neural Networks for Semantic Similarity Measurement. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 937–948.

J. Edward Hu, Rachel Rudinger, Matt Post, and Benjamin Van Durme. 2019. PARABANK: Monolingual Bitext Generation and Sentential Paraphrasing via Lexically-constrained Neural Machine Translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6521–6528.

Tomoyuki Kajiwara, Biwa Miura, and Yuki Arase. 2020. Monolingual Transfer Learning via Bilingual Translators for Style-sensitive Paraphrase Generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8042–8049.

Diederik P. Kingma and Jimmy Lei Ba. 2015. Adam: A Method for Stochastic Optimization. In *Proceedings of the 3rd International Conference for Learning Representations*.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180.

Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. 2004. Applying Conditional Random Fields to Japanese Morphological Analysis. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 230–237.

Kentaro Kurihara, Daisuke Kawahara, and Tomohide Shibata. 2022. JGLUE: Japanese General Language Understanding Evaluation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2957–2966.

Wuwei Lan and Wei Xu. 2018. Neural Network Models for Paraphrase Identification, Semantic Textual Similarity, Natural Language Inference, and Question Answering. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3890–3902.

Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li. 2020. On the Sentence Embeddings from Pre-trained Language Models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 9119–9130.

Jiduan Liu, Jiahao Liu, Qifan Wang, Jingang Wang, Wei Wu, Yunsen Xian, Dongyan Zhao, Kai Chen, and Rui Yan. 2023. RankCSE: Unsupervised Sentence Representations Learning via Learning to Rank. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 13785–13802.

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:1907.11692*.

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. 2014. A SICK Cure for the Evaluation of Compositional Distributional Semantic Models. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, pages 216–223.

Sneha Mehta, Bahareh Azarnoush, Boris Chen, Avneesh Saluja, Vinith Misra, Ballav Bihani, and Ritwik Kumar. 2020. Simplify-then-Translate: Automatic Preprocessing for Black-Box Machine Translation. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence*, pages 8488–8495.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *arXiv:1301.3781*.

Rei Miyata and Atsushi Fujita. 2021. Understanding Pre-Editing for Black-Box Neural Machine Translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1539–1550.

Eda Okur, Saurav Sahay, and Lama Nachman. 2022. Data Augmentation with Paraphrase Generation and Entity Extraction for Multimodal Dialogue System. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4114–4125.

Ellie Pavlick, Pushpendre Rastogi, Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2015. PPDB 2.0: Better Paraphrase Ranking, Fine-grained Entailment Relations, Word Embeddings, and Style Classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages 425–430.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models Are Unsupervised Multitask Learners.

Chandan K. Reddy, Lluís Màrquez, Fran Valero, Nikhil Rao, Hugo Zaragoza, Sambaran Bandyopadhyay, Arnab Biswas, Anlu Xing, and Karthik Subbian. 2022. Shopping Queries Dataset: A Large-Scale ESCI Benchmark for Improving Product Search. *arXiv:2206.06588*.

Jiajia Wang, Jimmy Xiangji Huang, Xinhui Tu, Junmei Wang, Angela Jennifer Huang, Md Tahmid Rahman Laskar, and Amran Bhuiyan. 2024. Utilizing BERT for Information Retrieval: Survey, Applications, Resources, and Challenges. *ACM Computing Surveys*, 56(7):1–33.

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*, pages 1112–1122.

Sam Witteveen and Martin Andrews. 2019. Paraphrasing with Large Language Models. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 215–220.

Hitomi Yanaka and Koji Mineshima. 2022. Compositional Evaluation on Japanese Textual Entailment and Similarity. *Transactions of the Association for Computational Linguistics*, 10:1266–1284.

Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. 2019. PAWS-X: A Cross-lingual Adversarial Dataset for Paraphrase Identification. In *Proceedings of the 2019 Conference on Empirical*

Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pages 3687–3692.

Qin Zhang, Shangsi Chen, Dongkuan Xu, Qingqing Cao, Xiaojun Chen, Trevor Cohn, and Meng Fang. 2023. A Survey for Efficient Open Domain Question Answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 14447–14465.

Yuan Zhang, Jason Baldridge, and Luheng He. 2019. PAWS: Paraphrase Adversaries from Word Scrambling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1298–1308.