Domain Knowledge Distillation for Multilingual Sentence Encoders in Cross-lingual Sentence Similarity Estimation

Risa Kondo[1] **Hiroki Yamauchi**[1] **Tomoyuki Kajiwara**[1,2] **Marie Katsurai**[3] **Takashi Ninomiya**[1]
[1]Ehime University [2]The University of Osaka [3]Doshisha University
{kondo@ai., yamauchi@ai., kajiwara@}cs.ehime-u.ac.jp
katsurai@mm.doshisha.ac.jp
ninomiya.takashi.mk@ehime-u.ac.jp

Abstract

We propose a domain adaptation method for multilingual sentence encoders. In domains requiring a high level of expertise, such as medical and academic, domain-specific pre-trained models have been released in each language. However, there is no its multilingual version, which prevents application to cross-lingual information retrieval. Obviously, multilingual pre-training with developing in-domain corpora in each language is costly. Therefore, we efficiently develop domain-specific cross-lingual sentence encoders from existing multilingual sentence encoders and domain-specific monolingual sentence encoders in each language. Experimental results on translation ranking in three language pairs with different domains reveal the effectiveness of the proposed method compared to baselines without domain adaptation and existing domain adaptation methods.

1 Introduction

To obtain knowledge comprehensively from large-scale text data on the Web, cross-lingual information retrieval (Artetxe and Schwenk, 2019) is promising. For application to embedding-based cross-lingual information retrieval, multilingual sentence encoders (Conneau et al., 2020; Reimers and Gurevych, 2020; Feng et al., 2022; Wang et al., 2024) are actively researched. Although these encoders are trained on general texts such as Wikipedia and CC100 (Wenzek et al., 2020), in specialized fields that require a high-level of expertise, such as medical and academic, developing domain-specific multilingual sentence encoders would be desirable. However, while domain-specific monolingual sentence encoders (Beltagy et al., 2019; Alsentzer et al., 2019; Araci, 2019; Zhang et al., 2021; Yamauchi et al., 2022; Labrak et al., 2023) have been released, the lack of their multilingual versions has prevented application to cross-lingual information retrieval.

Therefore, we propose a method to train a domain-specific cross-lingual sentence encoder. Since developing in-domain corpora in each language and subsequently performing multilingual pre-training requires significant costs, we fine-tune pre-trained multilingual sentence encoders on a bilingual corpus, similar to the training of general-purpose multilingual sentence encoders (Reimers and Gurevych, 2020; Feng et al., 2022; Tiyajamorn et al., 2021; Kuroda et al., 2022; Wang et al., 2024). Previous studies have focused on bringing bilingual sentence embeddings closer to each other based on methods such as knowledge distillation (Reimers and Gurevych, 2020), translation ranking (Feng et al., 2022), contrastive learning (Tiyajamorn et al., 2021; Wang et al., 2024), and adversarial learning (Kuroda et al., 2022). This study, in contrast, not only brings embeddings from source and target languages close together, but also simultaneously distills domain knowledge.

Experimental results on three domains and language pairs, i.e., Academic (English-Japanese), Medical (English-French), and Financial (English-Chinese), reveal that the proposed method achieves higher cross-lingual similarity estimation performance than the baseline without domain adaptation and existing domain adaptation methods. Our detailed analysis reveals that the proposed method is effective even when the in-domain bilingual corpus has only a few thousand sentence pairs available.

2 Proposed Method

As shown in Figure 1, this study extracts language-agnostic embeddings (hereafter, meaning embeddings) from a multilingual sentence encoder through fine-tuning on an in-domain bilingual corpus, while simultaneously distilling domain knowledge from domain-specific monolingual sentence encoders in each language. Our meaning embed-

Proceedings of Recent Advances in Natural Language Processing,pages 572–577
Varna, Sep 8–10, 2025
https://doi.org/10.26615/978-954-452-098-4-067

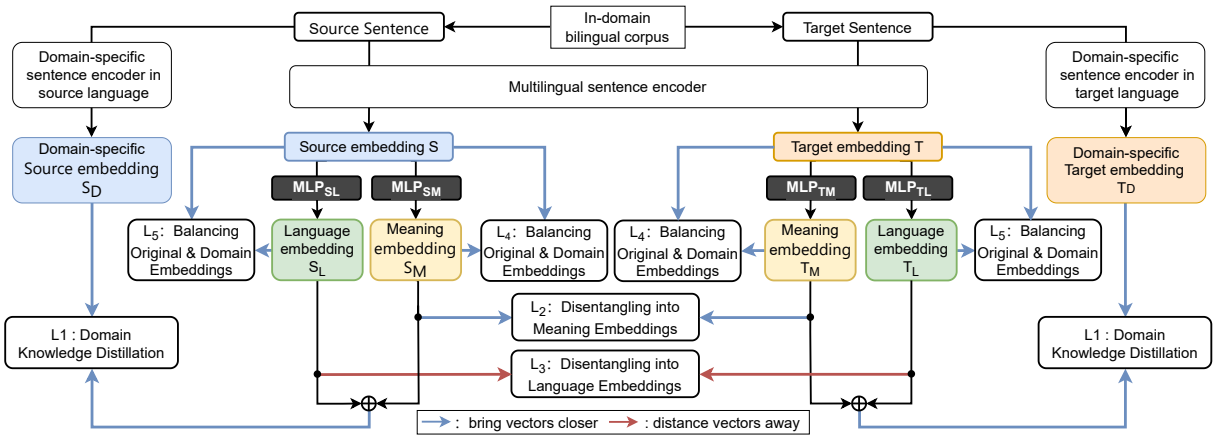

Figure 1: Overview of the proposed method. We train four MLPs to extract language-agnostic domain-specific meaning embeddings (yellow) using an in-domain bilingual corpus, a pre-trained multilingual sentence encoder, and pre-trained domain-specific monolingual sentence encoders in each language.

dings (yellow in Figure 1) can generate vectors that are similar to semantically close sentences, regardless of the input language, on a vector space suitable for the target domain. The proposed method consists of two main ideas: (1) extraction of language-agnostic embeddings and (2) adaptation to domain-specific embeddings. The former is inspired by DREAM (Tiyajamorn et al., 2021) and MEAT (Kuroda et al., 2022) that disentangle embeddings from multilingual sentence encoders into language-specific and -agnostic information; and the latter is inspired by the knowledge distillation of mSBERT (Reimers and Gurevych, 2020).

As in previous studies (Tiyajamorn et al., 2021; Kuroda et al., 2022), we disentangle embeddings from a multilingual sentence encoder into language and meaning embeddings by two multilayer perceptrons (MLPs), MLP_L and MLP_M. This is done for each of source and target sentence embeddings $S \in \mathbb{R}^d$ and $T \in \mathbb{R}^d$, respectively. We train a total of four MLPs using multi-task learning with the following five loss functions.

$$L = L_1 + L_2 + L_3 + L_4 + L_5 \quad (1)$$

2.1 Domain Knowledge Distillation

We aim to distill domain knowledge into meaning embeddings by the following loss that makes sentence embeddings, which combine language and meaning embeddings, closer to sentence embeddings from domain-specific sentence encoders.

$$L_1 = 2 - (\cos((S_L + S_M), S_D) + \cos((T_L + T_M), T_D)) \quad (2)$$

where $S_L \in \mathbb{R}^d$ and $T_L \in \mathbb{R}^d$ are language embeddings in the source and target languages, $S_M \in \mathbb{R}^d$ and $T_M \in \mathbb{R}^d$ are meaning embeddings in each language, $S_D \in \mathbb{R}^d$ and $T_D \in \mathbb{R}^d$ are sentence embeddings from domain-specific models. Here, d is the number of dimensions of each embedding.

2.2 Disentangling into Meaning Embeddings

We want to bring meaning embeddings between semantically similar sentences closer independent of their language. To achieve this, we define the following loss that brings meaning embeddings closer between bilingual sentences.

$$L_2 = 1 - \cos(S_M, T_M) \quad (3)$$

2.3 Disentangling into Language Embeddings

We want to distance language embeddings between sentences in different languages, independent of their meanings. To achieve this, we define the following loss that distances language embeddings between bilingual sentences.

$$L_3 = \max(0, \cos(S_L, T_L)) \quad (4)$$

2.4 Balancing Origin & Domain Embeddings

To make meaning embeddings language-agnostic, we want to avoid being too influenced by domain-specific sentence encoders that are language-specific. To achieve this, we define the following losses that prevent meaning and language embeddings from deviating too much from original ones.

$$L_4 = 2 - (\cos(S, S_M) + \cos(T, T_M)) \quad (5)$$

$$L_5 = 2 - (\cos(S, S_L) + \cos(T, T_L)) \quad (6)$$

2.5 Implementation Details

We use average pooling of token embeddings output from each model for sentence embeddings. All the MLPs in our model are single-layer feedforward neural networks (fully-connected layer) without activation functions. Only MLPs are fine-tuned in our method, and sentence encoders are fixed.

3 Evaluation

We evaluate the performance of the proposed method on a translation ranking task in three language pairs with different domains. Translation ranking is a task that ranks target language sentences in descending order of cross-lingual semantic similarity to a given source sentence. Our automatic evaluation metrics include ExactMatch, which evaluates only first-place candidates, and MRR@10, which evaluates the top 10 candidates. In this experiment, semantic similarity is estimated by cosine similarity.

3.1 Setting

Dataset Experiments were conducted in three domains: Academic, Medical, and Financial. In the academic domain, we used an English-Japanese corpus of project titles in Japanese research funds (KAKENHI[1]). In the medical domain, we used an English-French corpus of PubMed, which was employed in the biomedical translation task of WMT16 (Bojar et al., 2016). In the financial domain, we used an English-Chinese corpus (Turenne et al., 2022) of article titles from the Financial Times. The number of sentence pairs in these parallel corpora is shown in Table 1.

	Train	Valid	Test
Academic (En-Ja)	100,000	5,000	5,000
Medical (En-Fr)	100,000	5,000	5,000
Finance (En-Zh)	50,000	5,000	5,000

Table 1: Corpus size

Model Domain-specific monolingual sentence encoders include SciBERT[2] (Beltagy et al., 2019) in English and AcademicRoBERTa[3] (Yamauchi et al., 2022) in Japanese for the academic domain, Bio_ClinicalBERT[4] (Alsentzer et al., 2019) in English and DrBERT[5] (Labrak et al., 2023) in French for the medical domain, FinBERT[6] (Araci, 2019) in English and Mengzi-BERT[7] (Zhang et al., 2021) in Chinese for the financial domain. For multilingual models, we employed mBERT[8] (Devlin et al., 2019), LaBSE[9] (Feng et al., 2022) and mE5[10] (Wang et al., 2024), which are the commonly used multilingual sentence encoders. We used an implementation of HuggingFace Transformers (Wolf et al., 2020) and trained each model with a batch size of 512 and an Adam optimizer (Kingma and Ba, 2015) with a learning rate of 10^{-5}. Training was terminated when the loss in Equation (1) in the validation dataset did not improve by three epochs.

Comparison In addition to a baseline that uses sentence embeddings from each model as is, we evaluate three comparison methods. First, as a comparison method (1) that only distills domain knowledge, we consider the following loss L_6 using MLP_S for the source language and MLP_T for the target language instead of MLP_L and MLP_M that disentangle language and meaning embeddings.

$$L_6 = 2 - (\cos(S_S, S_D) + \cos(T_T, T_D)) \quad (7)$$

where S_S and T_T are embeddings for each language obtained through MLP_S and MLP_T.

Next, as a comparison method (2) that does not disentangle language and meaning embeddings, we consider the loss $L_6 + L_7$, which brings embeddings in the source and target languages closer in addition to distilling domain knowledge.

$$L_7 = 1 - \cos(S_S, T_T) \quad (8)$$

Finally, as a comparison method (3) without domain adaptation, we consider the loss $L - L_1 + L_8$, which reconstructs the original embeddings by

[1] https://kaken.nii.ac.jp/

[2] https://huggingface.co/allenai/scibert_scivocab_uncased

[3] https://huggingface.co/EhimeNLP/AcademicRoBERTa

[4] https://huggingface.co/emilyalsentzer/Bio_ClinicalBERT

[5] https://huggingface.co/Dr-BERT/DrBERT-7GB

[6] https://huggingface.co/ProsusAI/finbert

[7] https://huggingface.co/Langboat/mengzi-bert-base-fin

[8] https://huggingface.co/google-bert/bert-base-multilingual-uncased

[9] https://huggingface.co/sentence-transformers/LaBSE

[10] https://huggingface.co/intfloat/multilingual-e5-base

		Academic (En→Ja)		Medical (En→Fr)		Financial (En→Zh)	
		ExactMatch	MRR@10	ExactMatch	MRR@10	ExactMatch	MRR@10
	mBERT	0.025	0.087	0.696	0.747	0.083	0.127
(1)	L_6	0.000	0.002	0.000	0.002	0.000	0.002
(2)	$L_6 + L_7$	0.279	0.385	0.586	0.663	0.068	0.123
(3)	$L - L_1 + L_8$	0.403	0.492	0.872	0.889	0.175	0.244
	Ours	**0.443**	**0.531**	**0.885**	**0.901**	**0.188**	**0.259**
	LaBSE	0.908	0.927	0.943	0.949	0.476	0.537
(1)	L_6	0.000	0.002	0.000	0.002	0.000	0.002
(2)	$L_6 + L_7$	0.714	0.774	0.846	0.875	0.232	0.313
(3)	$L - L_1 + L_8$	0.914	0.933	**0.950**	**0.954**	0.497	0.558
	Ours	**0.917**	**0.936**	**0.950**	**0.954**	**0.513**	**0.575**
	mE5	0.863	0.899	0.939	0.946	0.529	0.594
(1)	L_6	0.000	0.002	0.000	0.002	0.000	0.002
(2)	$L_6 + L_7$	0.640	0.720	0.784	0.831	0.238	0.322
(3)	$L - L_1 + L_8$	0.928	0.948	0.952	0.956	0.600	0.665
	Ours	**0.935**	**0.952**	**0.954**	**0.957**	**0.607**	**0.671**

Table 2: Experimental results of translation ranking that uses English sentences as queries to retrieve sentences in other languages.

adding language and meaning embeddings.

$$L_8 = 2 - (\cos((S_L + S_M), S) + \cos((T_L + T_M), T)) \quad (9)$$

3.2 Result

Tables 2 and 3 show the experimental results. Table 2 shows the results when using English sentences as queries to retrieve sentences in other languages, and Table 3 shows the results when using sentences in other languages as queries to retrieve English sentences. The proposed method consistently outperformed the baselines for each model for all domains and language pairs. These consistent experimental results show that our method is effective for domain-specific cross-lingual sentence similarity estimation.

The fact that the comparative method (1), which only distills domain knowledge, completely loses retrievability, suggests that simple domain adaptation cannot train a cross-lingual model. The comparative method (2), which does not disentangle language embeddings and meaning embeddings, also degrades performance significantly, indicating that the extraction of language-agnostic meaning embeddings is important for cross-lingual retrieval. Although the comparative method (3), which does not distill domain knowledge, always outperformed the baseline, it consistently performed below or equal to the proposed method. These results suggest that while the performance of cross-lingual retrieval can be improved by simply extracting language-agnostic meaning embeddings, its quality can be further improved by domain adaptation.

3.3 Ablation Study

Table 4 presents an ablation analysis in LaBSE evaluating multiple combinations of losses, other than L_1, which are essential for domain adaptation. The performance drop in (a) suggests that L_2 and L_3 are important for disentangling language and meaning embeddings to achieve high cross-lingual performance. The significant performance drop in (b) suggests that L_4 and L_5, which prevent deviation from original embeddings, are essential for stable domain adaptation. (c) and (e) show that the to-English performance is worse than the proposed method when excluding losses related to meaning embeddings. In contrast, (d) and (f) show that excluding losses related to language embeddings has no significant impact.

3.4 Analysis of Sensitivity to Corpus Size

Figure 2 shows the performance of translation ranking while reducing the size of the target domain corpus for training. Even training with 5k sentence pairs revealed that our method outperforms the LaBSE baseline. Since developing a large-

		Academic (Ja→En)		Medical (Fr→En)		Financial (Zh→En)	
		ExactMatch	MRR@10	ExactMatch	MRR@10	ExactMatch	MRR@10
	mBERT	0.059	0.105	0.718	0.768	0.053	0.087
(1)	L_6	0.000	0.002	0.000	0.002	0.000	0.002
(2)	$L_6 + L_7$	0.247	0.349	0.777	0.821	0.023	0.054
(3)	$L - L_1 + L_8$	0.295	0.382	0.886	0.904	0.149	0.208
	Ours	**0.353**	**0.443**	**0.896**	**0.912**	**0.164**	**0.226**
	LaBSE	0.889	0.910	0.937	0.944	0.353	0.416
(1)	L_6	0.000	0.002	0.000	0.002	0.000	0.002
(2)	$L_6 + L_7$	0.714	0.775	0.895	0.914	0.096	0.155
(3)	$L - L_1 + L_8$	0.908	0.926	0.948	0.952	0.448	0.513
	Ours	**0.911**	**0.929**	**0.949**	**0.953**	**0.460**	**0.528**
	mE5	0.777	0.823	0.941	0.949	0.398	0.484
(1)	L_6	0.000	0.002	0.000	0.002	0.000	0.002
(2)	$L_6 + L_7$	0.633	0.714	0.874	0.899	0.101	0.161
(3)	$L - L_1 + L_8$	0.912	0.934	**0.952**	**0.956**	**0.571**	**0.637**
	Ours	**0.913**	**0.935**	**0.952**	**0.956**	**0.571**	**0.637**

Table 3: Experimental results of translation ranking that uses sentences in other languages as queries to retrieve English sentences.

	L2	L3	L4	L5	En → Ja	Ja → En
LaBSE					0.908	0.889
(a)			✓	✓	0.882	0.860
(b)	✓	✓			0.002	0.412
(c)		✓	✓	✓	0.907	0.890
(d)	✓		✓	✓	0.917	**0.911**
(e)	✓	✓		✓	**0.920**	0.898
(f)	✓	✓	✓		0.918	**0.911**
Ours	✓	✓	✓	✓	0.917	**0.911**

Table 4: Evaluation of ExactMatch in ablation analysis. Results are similar for other domains and language pairs.

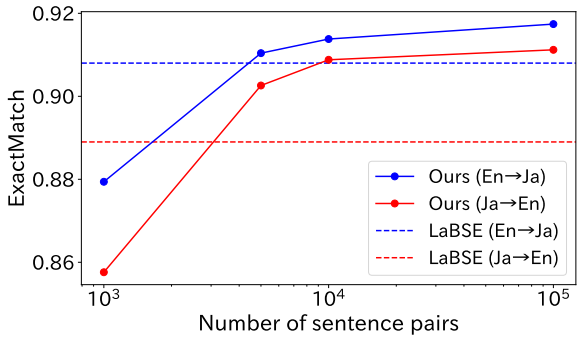

Figure 2: Sensitivity to training corpus size.

scale bilingual corpus in a specific domain is costly, the ability to train cross-lingual domain adaptation from small datasets is a strength of our method.

4 Conclusion

In this study, we address domain adaptation of multilingual sentence encoders. The proposed method uses an in-domain bilingual corpus and domain-specific monolingual sentence encoders in each language to simultaneously extract language agnostic meaning embeddings from a multilingual sentence encoder while distilling domain knowledge. Experimental results in three domains (Academic, Medical, and Financial) and three language pairs (English-Japanese, English-French, and English-Chinese) show that the proposed method can consistently improve the performance of cross-lingual similarity estimation for all domains and languages. Our method can achieve domain adaptation of multilingual sentence encoders even from an in-domain bilingual corpus of thousands of sentence pairs.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Number JP23K20410 and JST BOOST Program Japan Grant Number JPMJBY24036821.

References

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and

Matthew McDermott. 2019. Publicly Available Clinical BERT Embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78.

Dogu Araci. 2019. FinBERT: Financial Sentiment Analysis with Pre-trained Language Models. *arXiv:1908.10063*.

Mikel Artetxe and Holger Schwenk. 2019. Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A Pretrained Language Model for Scientific Text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 3615–3620.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. 2016. Findings of the 2016 Conference on Machine Translation. In *Proceedings of the First Conference on Machine Translation*, pages 131–198.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

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.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT Sentence Embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 878–891.

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

Yuto Kuroda, Tomoyuki Kajiwara, Yuki Arase, and Takashi Ninomiya. 2022. Adversarial Training on Disentangling Meaning and Language Representations for Unsupervised Quality Estimation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 5240–5245.

Yanis Labrak, Adrien Bazoge, Richard Dufour, Mickael Rouvier, Emmanuel Morin, Béatrice Daille, and Pierre-Antoine Gourraud. 2023. DrBERT: A Robust Pre-trained Model in French for Biomedical and Clinical domains. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 16207–16221.

Nils Reimers and Iryna Gurevych. 2020. Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 4512–4525.

Nattapong Tiyajamorn, Tomoyuki Kajiwara, Yuki Arase, and Makoto Onizuka. 2021. Language-agnostic Representation from Multilingual Sentence Encoders for Cross-lingual Similarity Estimation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7764–7774.

Nicolas Turenne, Nicolas Turenne, Ziwei Chen, Guitao Fan, Jianlong Li, Yiwen Li, Siyuan Wang, and Jiaqi Zhou. 2022. Mining an English-Chinese Parallel Dataset of Financial News. *Journal of Open Humanities Data*, 8(9):1–12.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual E5 Text Embeddings: A Technical Report. *arXiv:2402.05672*.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. CCNet: Extracting High Quality Monolingual Datasets from Web Crawl Data. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4003–4012.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45.

Hiroki Yamauchi, Tomoyuki Kajiwara, Marie Katsurai, Ikki Ohmukai, and Takashi Ninomiya. 2022. A Japanese Masked Language Model for Academic Domain. In *Proceedings of the Third Workshop on Scholarly Document Processing*, pages 152–157.

Zhuosheng Zhang, Hanqing Zhang, Keming Chen, Yuhang Guo, Jingyun Hua, Yulong Wang, and Ming Zhou. 2021. Mengzi: Towards Lightweight yet Ingenious Pre-trained Models for Chinese. *arXiv:2110.06696*.