A Japanese Masked Language Model for Academic Domain

Hiroki Yamauchi[1] **Tomoyuki Kajiwara**[1] **Marie Katsurai**[2]
Ikki Ohmukai[3,4] **Takashi Ninomiya**[1]
[1]Ehime University [2]Doshisha University [3]University of Tokyo
[4]National Institute of Informatics
{yamauchi@ai.,kajiwara@,ninomiya@}cs.ehime-u.ac.jp
katsurai@mm.doshisha.ac.jp, i2k@l.u-tokyo.ac.jp

Abstract

We release a pretrained Japanese masked language model for an academic domain. Pretrained masked language models have recently improved the performance of various natural language processing applications. In domains such as medical and academic, which include a lot of technical terms, domain-specific pretraining is effective. While domain-specific masked language models for medical and SNS domains are widely used in Japanese, along with domain-independent ones, pretrained models specific to the academic domain are not publicly available. In this study, we pretrained a RoBERTa-based Japanese masked language model on paper abstracts from the academic database CiNii Articles. Experimental results on Japanese text classification in the academic domain revealed the effectiveness of the proposed model over existing pretrained models.

1 Introduction

Academic papers in various fields and languages are accumulating daily on the Web. For example, more than 76k papers in the field of natural language processing (NLP) are currently available on the ACL Anthology.[1] Since the cost for humans to exhaustively learn from these large numbers of academic papers is immeasurable, scholarly document processing by NLP (Cohan and Goharian, 2015; Singh et al., 2018; Mohammad, 2020) is promising.

In NLP based on deep learning, which is currently the mainstream, supervised learning with a large-scale labeled corpus is effective. However, in domains where technical terms are frequently used, such as in academic fields, hiring professional annotators is very expensive. Therefore, the low-resource problem is a serious issue in various languages, domains, and tasks.

In recent NLP, finetuning of pretrained masked language models on large-scale raw corpora, such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019), has been widely employed to address the low-resource problem. Especially in domains such as medical (Alsentzer et al., 2019) and academic (Beltagy et al., 2019; Lee et al., 2020), the effectiveness of domain-specific pretraining has been reported. Similar to these previous studies in English, domain-specific masked language models for medical (Kawazoe et al., 2021) and SNS[2] are widely used in Japanese, along with domain-independent masked language models.[3,4] However, there are no publicly available pretrained Japanese models that are specific to the academic domain.

In this study, we pretrained a RoBERTa-based Japanese masked language model (Liu et al., 2019) using 6.28M sentences of paper abstracts from a scholarly article database CiNii Articles[5] to improve the performance of scholarly document processing in Japanese. Experimental results on Japanese text classification in the academic domain revealed the effectiveness of the proposed model, which is specific to the academic domain, compared to the domain-independent masked language models. Our model (Academic RoBERTa) will be available on GitHub[6] when this paper is published.

2 Related Work

Finetuning of pretrained Transformer (Vaswani et al., 2017) achieves excellent performance on many NLP tasks (Wang et al., 2018). BERT (Devlin et al., 2019), a typical pretraining model, trains the Transformer encoder by multi-task learning of masked language modeling and next sentence pre-

[1] https://aclanthology.org/

[2] https://github.com/hottolink/hottoSNS-bert

[3] https://huggingface.co/cl-tohoku/bert-base-japanese

[4] https://huggingface.co/nlp-waseda/roberta-base-japanese

[5] https://ci.nii.ac.jp/

[6] https://github.com/hirokiyamauch/AcademicRoBERTa

Proceedings of the Third Workshop on Scholarly Document Processing, pages 152–157
October 12–17, 2022.

diction. RoBERTa (Liu et al., 2019) outperforms BERT by pretraining only masked language modeling, through dynamic masking and increasing batch size and number of training steps. This study conducts powerful RoBERTa-based pre-training to develop a Japanese masked language model specific to the academic domain.

The effectiveness of domain-specific pretraining to address technical terms and style-specific expressions has been reported. In English, domain-specific masked language models are publicly available for various domains, including medical (Alsentzer et al., 2019), academic (Beltagy et al., 2019; Lee et al., 2020), and SNS (Nguyen et al., 2020). Domain-specific masked language models have also been developed in Japanese, such as UTH-BERT (Kawazoe et al., 2021) for the medical domain and hottoSNS-BERT[2] for the SNS domain. However, no pretraining Japanese model specific to the academic domain has been released.

3 Methods

To improve the performance of scholarly document processing in Japanese, we release a Japanese masked language model specific to the academic domain. First, in Section 3.1, we create a Japanese corpus consisting of paper abstracts. Then, in Section 3.2, we use this corpus to conduct pretraining based on RoBERTa (Liu et al., 2019).

3.1 Corpus

We use CiNii Articles,[5] a scholarly article database, to create a Japanese corpus specific to the academic domain. We extracted 1.27 million abstracts of academic papers included in CiNii Articles as of March 2022, containing Japanese characters (hiragana or katakana). Then, a corpus of approximately 6.28 million sentences (about 180 million words) was created by applying the five-step preprocessing shown in Table 1.

Deletion of Fixed Expressions Paper abstracts extracted from CiNii Articles contain noise due to automatic information extraction, such as "論文タイプ‖研究ノート" (paper type ‖ research notes). To exclude these fixed expressions from the corpus, we remove them when the same document appears more frequently than a threshold. Since there were cases where the same document appeared 5 or 6 times due to ID registration errors, we set the threshold as 7 or more times.

Preprocess	Corpus size
Number of paper abstracts	1.27 M docs.
1. Deletion of fixed expressions	1.15 M docs.
2. Segmentation into sentences	7.31 M sents.
3. Extraction of Japanese sentences	6.68 M sents.
4. Deletion of duplicate sentences	6.33 M sents.
5. Limitation of sentence length	6.28 M sents.

Table 1: Change in corpus size due to preprocessing.

Segmentation into Sentences For 1.15 million documents obtained by the previous preprocessing, sentence segmentation is performed. Approximately 7.31 million sentences were obtained by rule-based sentence segmentation.[7]

Extraction of Japanese Sentences To clean our Japanese corpus, we remove sentences written in languages other than Japanese. Since technical terms are often expressed in other languages, sentences in which the characters above the threshold are Japanese (hiragana or katakana or kanji) are extracted. In this study, this threshold was set at 50%, resulting in about 6.68 million Japanese sentences.

Deletion of Duplicate Sentences To prevent bias caused by high-frequency expressions, sentences that occur frequently in specific fields, such as "下腹部痛を主訴に来院。" (Visited the hospital with a chief complaint of lower abdominal pain.) and fixed form sentences in academic papers, such as "その結果を以下に示す。" (The results are shown below.) are removed. In the case of sentence duplication, the sentence was left in the corpus only once and the others were deleted, resulting in a corpus of about 6.33 million unique sentences.

Limitation of Sentence Length Finally, extremely short and long sentences are removed to completely eliminate errors in fixed expressions and sentence segmentation. Sentences of less than 10 characters often contained expressions such as "（編集委員会作成）" (prepared by the editorial board) that would not be included in the actual paper abstracts. Therefore, in this study, we created a corpus of approximately 6.28 million sentences by extracting sentences with between 10 and 200 characters.

[7] https://github.com/wwwcojp/ja_sentence_segmenter

3.2 Pretraining

The corpus created in Section 3.1 is used to pretrain masked language modeling equivalent to RoBERTa (Liu et al., 2019). Subword segmentation by SentencePiece[8] (Kudo and Richardson, 2018) with a vocabulary size of $32,000$ was performed for tokenization. Our model is a Transformer (Vaswani et al., 2017) with the same structure as the `roberta-base`, implemented by the fairseq toolkit.[9] (Ott et al., 2019) That is, our masked language model consists of 12 layers of 768 dimensions with 12 self-attention heads. We set the maximum number of tokens per input instance to 512, the batch size to 64 sentences, and the dropout rate to 0.1. We used Adam (Kingma and Ba, 2015) with learning rate scheduling by polynomial decay as the optimizer and we set the maximum learning rate to 0.0001 and the warmup step to $10,000$. The number of training steps was set to $700,000$ for a fair comparison with a previous study.[4] Our model was pretrained on two CPUs (Intel Xeon GOLD 5115) with 192 GB RAM and four GPUs (RTX A6000 48 GB).

4 Evaluation

To evaluate the effectiveness of our masked language model (Academic RoBERTa) specific to the academic domain, we empirically compare our model with existing domain-independent masked language models through experiments on Japanese text classification in the academic domain.

4.1 Baselines

In this experiment, BERT (Tohoku BERT)[3] and RoBERTa (Waseda RoBERTa)[4], which are domain-independent masked language models for Japanese, are employed as baseline models. Both baselines are Transformer models (Vaswani et al., 2017) with the same structure as Academic RoBERTa and have the same size vocabulary. However, they differ in the corpus used for pretraining, its preprocessing, and the hyperparameters during pretraining. We used HuggingFace Transformers (Wolf et al., 2020) to implement our baseline models.

Tohoku BERT is a BERT model (Devlin et al., 2019) pretrained on Japanese Wikipedia. Morphological analysis with MeCab (IPADIC) (Kudo et al., 2004) and subword segmentation with WordPiece (Wu et al., 2016) were used as preprocessing. The maximum number of tokens per input instance is 512, the batch size is 256 sentences, and 1 million steps of pretraining is performed.

Waseda RoBERTa is a RoBERTa model (Liu et al., 2019) pretrained on both Japanese Wikipedia and the Japanese part of CC100 (Wenzek et al., 2020). Morphological analysis with Juman++ (Tolmachev et al., 2020) and subword segmentation with SentencePiace (Kudo and Richardson, 2018) are used as preprocessing. The maximum number of tokens per input instance is 128, the batch size is 256 sentences ($\times 8$ GPUs), and $700,000$ steps of pretraining is performed.

4.2 Tasks

As evaluation tasks in the academic domain, we experiment with two types of Japanese text classification on the titles of research projects funded by Grants-in-Aid for Scientific Research (KAKENHI). KAKENHI is a competitive research fund in Japan that covers scientific research in all fields. For this experiment, we collected $73,000$ KAKENHI proposals from 2013 to 2017. We designed two evaluation tasks: an author identification task to estimate whether the principal investigator is the same or not from pairs of research project titles, and a category classification task to estimate the research fields from research project titles. In both tasks, each masked language model is automatically evaluated by the accuracy of its classification.

Author Identification This task is a sentence-pair classification task that performs a binary classification of whether the principal investigators of two research projects are identical or not. In this experiment, a total of $120,000$ pairs, $50,000$ positive examples consisting of research project titles proposed by the same principal investigator and $70,000$ negative examples consisting of those proposed by different principal investigators, were paired and randomly split for training, validation, and evaluation as shown in the top row of Table 2. Two sentences were input simultaneously into the masked language model with a special token of `[SEP]` in between.

Category Classification This task is a sentence classification task to estimate research fields from the titles of research projects. KAKENHI employs a four-level hierarchical structure of research fields, which include 4, 14, 77, and 318 categories, in

[8] https://github.com/google/sentencepiece

[9] https://github.com/facebookresearch/fairseq

	Author indentification	Category classificaton			
# examples for Train/Valid/Test	100k / 10k / 10k	70k / 1.5k / 1.5k			
# classes	2	4	14	77	318
Tohoku BERT	95.1	83.7	69.6	53.3	40.3
Waseda RoBERTa	97.1	83.9	71.9	55.4	42.7
Academic RoBERTa	**98.7**	**84.7**	**72.9**	**58.8**	**44.6**

Table 2: Accuracy of academic text classification in Japanese.

descending order from the largest categories. In this experiment, each level of classification was performed independently. That is, the classification results for the larger categories do not affect the classification of the smaller categories.

4.3 Finetuning

The corpus described in Section 4.2 was used to finetune the masked language models. As a preprocessing, subword segmentation was performed for each model using the same settings as in the pretraining. For finetuning, the batch size was 256 sentences, the dropout rate was set to 0.1, and Adam (Kingma and Ba, 2015) was used as the optimizer with a maximum learning rate of $5e^{-5}$. Finetuning was terminated when the accuracy in the validation dataset did not improve for 10 epochs as early stopping.

4.4 Results

Table 2 shows the experimental results. RoBERTa consistently achieved better performance than BERT, and Academic RoBERTa, which is specific to the academic domain, showed the best performance on all tasks. In particular, the proposed method showed significant performance improvement in classifying minor categories (*i.e.*, 77-class and 318-class classifications), which require more detailed expertise than major categories (*i.e.*, 4-class and 14-class classifications).

There is no difference in model structure or number of training steps between Waseda RoBERTa and Academic RoBERTa. In addition, since Tohoku BERT and Waseda RoBERTa are pretrained using corpora of approximately 17 million and 4 billion sentences, respectively, our approximately 6.28 million sentences have no advantage in terms of corpus size. Therefore, the performance improvement of our model can be attributed only to its specialization in the academic domain.

4.5 Discussion

We analyze the vocabulary of the domain-specific model. We found that 49.4% of the tokens in Academic RoBERTa's vocabulary are not included in that of existing masked language models.[10] Examples of characteristic tokens that only Academic RoBERTa has include phrases that frequently appear in academic papers in any field, such as "であることが確認された" (It was confirmed that the ...) and technical terms that frequently appear in certain fields, such as "ニューラルネットワーク" (neural networks). Our model may have achieved high performance for text classification in the academic domain because our vocabulary includes many such domain-specific tokens.

5 Conclusion

In this study, we released Academic RoBERTa, a Japanese masked language model specific to the academic domain, pretrained on abstracts of academic papers included in CiNii Articles. Experimental results on Japanese text classification in the academic domain revealed that our model consistently outperforms existing domain-independent masked language models. Detailed analysis confirmed the effectiveness of domain-specific pretraining, as many domain-specific expressions were included in the vocabulary and the accuracy of text classification improved significantly for more detailed categories requiring more expertise.

Our future work includes making Japanese text generation models such as GPT-2 (Radford et al., 2019) and BART (Lewis et al., 2020) specific to the academic domain. These models could contribute to summarization and grammatical error correction in the academic domain.

[10] The vocabulary of the existing masked language model refers to the following union sets: the vocabulary of Tohoku BERT, the vocabulary of Waseda RoBERTa, and the vocabulary when training the subword segmentation of SentencePiece on Japanese Wikipedia with a vocabulary size of 32,000.

Acknowledgment

This work was supported by JSPS KAKENHI Grant Number JP20H04484 and ROIS NII Open Collaborative Research 2020 (20S0405).

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.

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.

Arman Cohan and Nazli Goharian. 2015. Scientific Article Summarization Using Citation-Context and Article's Discourse Structure. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 390–400.

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.

Yoshimasa Kawazoe, Daisaku Shibata, Emiko Shinohara, Eiji Aramaki, and Kazuhiko Ohe. 2021. A Clinical Specific BERT Developed Using a Huge Japanese Clinical Text Corpus. *PLOS ONE*, 16(11):1–11.

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

Taku Kudo and John Richardson. 2018. SentencePiece: A Simple and Language Independent Subword Tokenizer and Detokenizer for Neural Text Processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71.

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.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: A Pre-trained Biomedical Language Representation Model for Biomedical Text Mining. *Bioinformatics*, 36(4):1234–1240.

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.

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*.

Saif M. Mohammad. 2020. NLP Scholar: A Dataset for Examining the State of NLP Research. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 868–877.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. BERTweet: A Pre-trained Language Model for English Tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 9–14.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A Fast, Extensible Toolkit for Sequence Modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners. *Technical Report, OpenAI*, pages 1–24.

Mayank Singh, Pradeep Dogga, Sohan Patro, Dhiraj Barnwal, Ritam Dutt, Rajarshi Haldar, Pawan Goyal, and Animesh Mukherjee. 2018. CL Scholar: The ACL Anthology Knowledge Graph Miner. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 16–20.

Arseny Tolmachev, Daisuke Kawahara, and Sadao Kurohashi. 2020. Design and Structure of The Juman++ Morphological Analyzer Toolkit. *Journal of Natural Language Processing*, 27(1):89–132.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355.

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 12th 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.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *arXiv:1609.08144*.