DEEP LEARNING-BASED APPROACH FOR JAPANESE KEYPHRASE GENERATION

Padipat Sitkrongwong[1], Nakarin Sritrakool[1,2],
Phannakan Tengkiattrakul[1], Pongsakorn Jirachanchaisiri[1,2] and Atsuhiro Takasu[1,2]
[1]*National Institute of Informatics, Tokyo, Japan*
[2]*The Graduate University for Advanced Studies, Kanagawa, Japan*

ABSTRACT

Retrieving keywords or keyphrases from a text document is one of the essential features in many applications. For example, the dataset search systems may use those keyphrases to suggest more relevant data to users. However, obtaining the keyphrases from a Japanese text is challenging due to its unique sentence structure. Moreover, most existing methods derive an extraction approach incapable of producing new Japanese keyphrases for each document. Thus, we propose a method for generating Japanese keyphrases based on a deep learning-based text-to-text generative model. Our method can accurately generate the existing and the additional relevant keyphrases for each document text. A model architecture used in our method also helps produce the Japanese keyphrases that preserve their original meanings.

KEYWORDS

Japanese Keyphrase Generation, Dataset Search, T5

1. INTRODUCTION

Recently, open data has become an important resource for analyzing what happens in the real world. Since open data is provided by various organizations, ranging from national and local governments to citizens, search engines are essentially important for retrieving relevant data. Google launched Google Dataset Search (https://datasetsearch.research.google.com), and we are developing a search engine for open data in Japan (https://search.ckan.jp). Search by keywords is a primary approach to the dataset search, but it is not easy for users to choose effective keywords for this new type of information. This paper, therefore, aims to propose a method to generate keyphrases for enhancing the dataset search results.

There are two main approaches to obtain the list of keyphrases from a document text: extraction and generation. Keyphrase extraction (Le et al., 2013; Teng, 2021) aims to extract keyphrases that are previously mentioned in each document. In comparison, keyphrase generation (Meng et al., 2017; Shen et al., 2022) aims to generate a list of keyphrases that may or may not exist in the document text.

Most existing works proposed under those approaches failed to produce high-quality keyphrases when applied to Japanese text because the Japanese sentence structure differs from English. For example, in one sentence, the Japanese words or phrases might be written consecutively without explicit word boundary, which is different from English sentences, leading to a challenge in designing the tokenization methods that work well with Japanese text. Also, unlike English words, most Japanese words could be composed of more than one token. Separately extracting those tokens from one word might completely abandon its original meaning.

Acknowledged by these unique characteristics, some researchers proposed techniques for retrieving Japanese keywords and keyphrases (Le et al., 2013; Teng, 2021). For example, Le et al. proposed a chunk-based keyword extraction that treats Japanese keywords as chunks of text that yield each document's important contents (Le et al., 2013). However, these methods are based on the extraction approach that can extract only the mentioned keyphrases in a document. Therefore, the keyphrases for each document may not necessarily need to be matched with the phrases mentioned in a document. In many applications, the ability to generate new keywords or keyphrases for each document could be a potential advantage. For example, the dataset search system could provide more relevant search results based on the newly generated tags for each dataset.

To solve these challenges, we propose a deep learning-based approach for Japanese keyphrase generation. Instead of following the extraction approach, we derive a text-to-text generative model as our model architecture. Our main contributions are:

1. We propose a model architecture for Japanese keyphrase generation based on a deep learning-based text-to-text generation approach. This model can accurately generate keyphrases (tags) for each document (dataset description) and the additional relevant keyphrases that most keyphrase extraction methods cannot retrieve. Moreover, it can generate new keyphrases that have not existed in the keyphrase vocabulary.
2. The masking mechanism used in our derived model can mask the consecutive tokens, which helps produce the Japanese keyphrases that preserve their original meaning.

2. THE PROPOSED METHOD

2.1 Deep Learning-Based Text-to-Text Generation

We adopt a pre-trained model called T5 (Text-to-Text-Transfer-Transformer) (Raffel et al., 2020) for the Japanese keyphrase generation. T5 is an encoder-decoder model trained on a large corpus of text and metadata in multi-task learning, which makes it generalized for various natural language processing (NLP) tasks. Furthermore, it is designed based on a unified text-to-text generation framework, meaning its input and output are always in a string format.

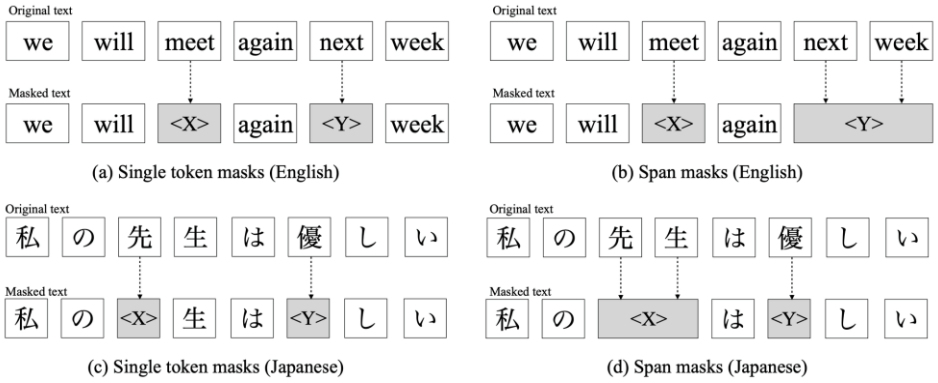

Figure 1. Applying two masking mechanisms to a text in two languages. (a) Single token masks with English text, (b) Span masks with English text, (c) Single token masks with Japanese text, and (d) Span masks with Japanese Text

Similar to Bidirectional Encoder Representations Transformers (BERT) (Devlin et al., 2019), T5 was trained with the masked language model objective, where some words of the input text are randomly masked. The representations of those mask words are then learned by leveraging the information derived from their left and right context words. The example of a masked text done by BERT is presented in Figure 1a, where the words "meet" and "next" in the original text are masked by the mask tokens <X> and <Y> before feeding into the model. However, there are two main differences between BERT and T5. Since BERT comprises only the encoder model, it aims to produce a dense input representation, which can be used mainly for text classification tasks. In contrast, T5 is an encoder-decoder model designed with the objective of text generation. The text-to-text generative nature of T5 makes it applicable to the NLP tasks that BERT cannot achieve, including machine translation or text summarization (Agarwal et al., 2020).

The masking mechanism is the other component that makes T5 different from BERT and suitable for this work. Normally, BERT randomly replaces a single-word token with a masked token, as shown in Figure 1a. This masking, however, can cause a problem when working with Japanese text where most words are composed of two or more tokens (which are obtained by applying Japanese word tokenizers such as SentencePiece (Kudo and Richardson, 2018)). If only one token from those words is masked, they might lose their original meaning. For example, Figure 1c illustrates the possible masking of BERT applied to a Japanese sentence 「私の先生

は優しい」 (My teacher is kind). In this sentence, a token「先」from a word 「先生」 (teacher) is masked, whereas the token 「生」remains unmasked. These two tokens, however, also have their individual meanings (「先」means “previous” and 「生」means “life”). Separately masking only one of them could produce a text representation far from its original meaning.

Fortunately, the masking problem can be alleviated by the masking mechanism of T5. In T5, a mask token can be used to mask more than one consecutive word token. We believe this masking scheme could work especially well for the Japanese keywords and keyphrases. Considering the same sentence but masks with T5 in Figure 1d, both tokens from the word 「先生」are now masked together. We believe that T5’s masking scheme could help preserve the original meaning of Japanese keywords, leading to a more meaningful representation and fidelity of the model.

2.2 Japanese Keyphrase Generation

In order to apply T5 to Japanese text, we use the version of T5 that was pre-trained with Japanese text corpora (https://huggingface.co/sonoisa/t5-base-japanese), such as Japanese Wikipedia and OSCAR.

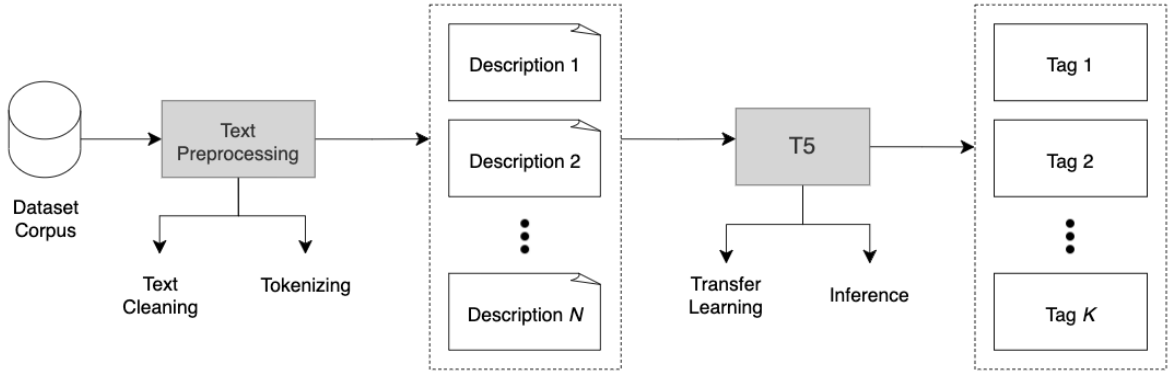

Figure 2. An overview of our Japanese keyphrase generation pipeline using T5

The pipeline of the method is in Figure 2. We first retrieve dataset metadata from the dataset corpus. We then extract only the description of each dataset (as a document) and its corresponding tags (as keyphrases) from the metadata. After that, the Japanese texts are normalized, such as full-width conversion and removing special characters, whereas English texts are converted into lowercase. Both descriptions and tags are then tokenized using SentencePiece (Kudo and Richardson, 2018). After obtaining a cleaned and tokenized dataset description and its corresponding tags, they are fed into T5 for the Japanese phrase generation task, which consists of two phases. In the transfer learning phase, we fine-tuned the pre-trained T5 by feeding the pairs of <description, tags> as input and output of the model. The main objective is to let T5 learn to generate the corresponding tags given a description as input. In the inference phase, we use the fine-tuned T5 to generate the list of tags by providing a dataset description as input. This model can generate the tag list for tagged and untagged descriptions and new tags that have never appeared before in the tag vocabulary.

3. EXPERIMENTS

3.1 Experiment Settings

The model is evaluated on the metadata obtained from our dataset search system (https://search.ckan.jp). From this metadata, a total of 27,682 dataset descriptions and 5,585 unique tags are extracted. After that, the descriptions are cleaned. Then, the input (descriptions) and labels (tags) are tokenized and truncated to 512 and 200 tokens at maximum. Next, the dataset is split into 80%, 10%, and 10% for training, validating, and testing. Finally, the dataset is used to fine-tune the T5 model (https://github.com/sonoisa/t5-japanese) with a learning rate of 1e-4, batch size of 16, training epoch of 10,000, and an early stop when the loss of validation set is not improved for 10 epochs. The other parameters are set to their default values used in T5.

3.2 Experimental Results

Table 1. Comparison of T5 with different repetition penalties on various measurements

Model	Repetition Penalty	Hit Rate	Coverage Ratio	BLEU (Unigram)	ROUGE (Unigram)
T5	1.0	0.8952	0.8940	0.8567	0.8733
T5	1.5	**0.8963**	**0.8959**	**0.8585**	**0.8744**
T5	3.0	0.8953	0.8932	0.8552	0.8732

Table 1 shows the performance comparison of T5 with different repetition penalties. The evaluation metrics include hit rate, coverage ratio, BLEU, and ROUGE. The hit rate and coverage ratio indicate how well the model can correctly identify and generate text, with higher values indicating better performance. The BLEU and ROUGE (Yang et al., 2018) are commonly used metrics for evaluating the quality of the machine-generated text, with higher values indicating better performance. Note that when the penalty equals 1, it means no penalty. The results indicate that the T5 models with penalties of 1.0, 1.5, and 3.0 have similar hit rate and coverage ratio values, with a slightly higher value for T5 1.5. Among T5 with different penalties, they have similar BLEU and ROUGE, except T5 with penalty 1.5, which performs slightly better than the others.

In addition to the accuracy performances, we also give examples of the generated keyphrases in two scenarios. Table 2 shows the results from a description related to the “cyber security project report”. As it can be seen, our model can generate not only the matched keyphrases (「クラウド」: “cloud”, 「セキュリティ」: “security”) but also the additional relevant ones such as “security”, 「セキュリティガイドライン」 (“security guideline”) or「安全」(“safety”). Moreover, Table 3 shows the results from a description related to the “energy / power-saving project report”. In this case, the model can even generate the keyphrase 「節電」 (power saving), which does not exist in the tag vocabulary.

Table 2. Example of generating additional keyphrases

Description	平成 26 年度サイバーセキュリティ経済基盤構築事業クラウドセキュリティ監査制度の見直し調 査報告 書 iso/iec 27001 及び 27002(2013 年版)に基づき、情報セキュリティ管理基準の改訂を検討し た資料、 及び、so/iec 27001 及び 27002(2013 年版)、クラウドセキュリティガイドライン、iso/iec 27017 の策 定動向に基づき、クラウド情報セキュリティ管理基準の改訂を検討した資料。(127 文字)		
Matched Keyphrases	クラウド, セキュリティ	**Additional Keyphrases**	security, セキュリティガイドライン, 企業_会社情報, 商取引, 商業, 安全, 工業, 法 律, 産業, 研究, 社会, 経済

Table 3. Example of generating a new (out of vocabulary) keyphrase

Description	平成 26 年度エネルギー使用合理化促進基盤整備事業(電力需要抑制に係る設備導入効果の実態分析調 査)調査報告書「平成 23 年度補正建築物節電改修支援事業」の補助金交付を受けた事 業者を対象とし て、事業完了後 1 年間の電力抑制実績及び目標値に対する実績値の傾向分析を 行った。		
Matched Keyphrase	エネルギー	**New Keyphrase**	節電

4. CONCLUSION

We proposed a model architecture for Japanese keyphrase generation based on a deep learning-based text-to-text generation approach. This model can generate the existing and additional relevant keyphrases for each document text and preserve their original meanings. For future works, we want to try other state-of-the-art encoder-decoder models to evaluate the performances of Japanese keyphrase generation tasks. Moreover, we want to design a way to evaluate the correctness of the newly generated keyphrases obtained by our proposed method to ensure that they are reliable when using them in the downstream tasks.

ACKNOWLEDGEMENT

This work was supported by the Cross-ministerial Strategic Innovation Promotion Program (SIP) Second Phase, "Big-Data and AI-Enabled Cyberspace Technologies" by New Energy and Industrial Technology Development Organization (NEDO).

REFERENCES

Agarwal O, Kale M, Ge H, et al. (12 2020) Machine Translation Aided Bilingual Data-to-Text Generation and Semantic Parsing. In: *Proceedings of the 3rd International Workshop on Natural Language Generation from the Semantic Web (WebNLG+)*, Dublin, Ireland (Virtual), 12 2020, pp. 125–130. Association for Computational Linguistics.

Devlin J, Chang M-W, Lee K, et al. (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)*, Minneapolis, Minnesota, June 2019, pp. 4171–4186. Association for Computational Linguistics.

Kudo T and Richardson J (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*, Brussels, Belgium, November 2018, pp. 66–71. Association for Computational Linguistics.

Le TTN, Le Nguyen M and Shimazu A (2013) Unsupervised Keyword Extraction for Japanese Legal Documents. In: *JURIX*, 2013, pp. 97–106.

Meng R, Zhao S, Han S, et al. (2017) Deep Keyphrase Generation. In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vancouver, Canada, July 2017, pp. 582–592. Association for Computational Linguistics.

Raffel C, Shazeer N, Roberts A, et al. (2020) Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of machine learning research: JMLR* 21(1).

Shen X, Wang Y, Meng R, et al. (2022) Unsupervised Deep Keyphrase Generation. *Proceedings of the AAAI Conference on Artificial Intelligence* 36(10): 11303–11311.

Teng M (2021) Using the Ship-Gram Model for Japanese Keyword Extraction Based on News Reports. *Complexity* 2021. Hindawi.

Yang A, Liu K, Liu J, et al. (2018) Adaptations of ROUGE and BLEU to better evaluate machine reading comprehension task. In: *Proceedings of the Workshop on Machine Reading for Question Answering*, Stroudsburg, PA, USA, 2018. Association for Computational Linguistics.