ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 5, No. 6, 1674-1682 (2020)
www.astesj.com
Special Issue on Multidisciplinary Innovation in Engineering Science & Technology

ASTES Journal
ISSN: 2415-6698

Japanese Abstractive Text Summarization using BERT

Yuuki Iwasaki [*1], Akihiro Yamashita[1], Yoko Konno[2], Katsushi Matsubayashi[1]

[1]*National Institute of Technology, Tokyo College, 193-0997, Japan*

[2]*CHOWA GIKEN Corporation, 001-0021, Japan*

ARTICLE INFO

Article history:
Received: 27 August, 2020
Accepted: 12 December, 2020
Online: 28 December, 2020

Keywords:
Abstractive text summarization
BERT
Japanese text summarization

ABSTRACT

In this study, we developed and evaluated an automatic abstractive summarization algorithm in Japanese using a neural network. We used a sequence-to-sequence encoder-decoder model for practical purposes. The encoder obtained a feature-based input vector of sentences using the bidirectional encoder representations from transformers (BERT) technique. A transformer-based decoder returned the summary sentence from the output as generated by the encoder. This experiment was conducted using the Livedoor news corpus with the above model. However, two problems were revealed. One is the repetition of a specific phrase while the model is generating text. The other is that the model can not handle out-of-vocabulary words. As solutions, we use repeat block in n-gram words and WordPiece. In addition, to evaluate the performance of the model, we compared the summarization accuracy between our model and a long short term memory based pointer-generator network. As revealed by the results, our model comprehends the meanings of sentences better than a pointer-generator network but makes more word-based mistakes.

1 Introduction

This paper is an extension of work originally presented in 2019 International Conference on Technologies and Applications of Artificial Intelligence (TAAI) [1].

Societies with a highly developed information system make it easy for citizens to obtain the information they need. However, for this reason, the information obtained often has redundancy. Therefore, skills are needed to extract only the necessary information. In other words, a summarization skill is important. Thus, we need a system that automatically takes the necessary points of a text and applies deep learning with a neural network. At the present time, models using bidirectional encoder representations from transformers (BERT) [2] have achieved the highest scores in abstractive summarization tasks [3]. Although there are already many models using BERT in English and some other languages, a model in Japanese has yet to be found. We therefore developed models using BERT in Japanese and evaluated them through a comparison with another model. The code is available at `https://github.com/IwasakiYuuki/Bert-abstractive-text-summarization`.

Text summarization is one of the natural language processing that effectively summarizing long sentences. The Algorithms of text summarization used in machine learning are mainly divided into two types: extractive and abstractive summaries. In the extractive text summarization, a summary sentence is generated by combining important sentences in the source text. On the other hand, the latter, abstract type, understands the input sentences and generates the corresponding summary sentences by itself. Although both types are similar in that they summarize the main points of input sentences, they are more flexible in this respect because abstract abstract summarization generates the corresponding summary sentence by itself, while extractive summarization can only process sentences collected from input sentences. In this study, we focused on abstract summarization.

In recent years, various models have been proposed for abstractive summarization. An author proposed an abstractive summarization model using Bidirectional Encoder Representations from Transformers (BERT) [3]. Experimentation results as reported in [3] revealed that the developed model achieved a new state-of-the-art performance on both CNN/Daily Mail and New York Times datasets. Viswani et al. proposed an abstractive summarization model called a pointer-generator network based CopyNet [4]. This pointer-generator network model has advantages in terms of both abstractive and extractive summarizations.

The model developed in this study was built with reference to a text summarization model using BERT and has two stages. In the first stage, the input text was encoded into a contextual representation using BERT, and after processing the input text with BERT, a

[*]Corresponding Author: Yuuki Iwasaki, Tokyo College, s20607@tokyo.kosen-ac.jp

summary draft was generated using a transformer-based decoder. In the second stage, the draft summary text was re-validated using BERT to generate a crisper summary text, which was then processed with BERT. In this experiment, only the first stage was employed.

In section 2, we introduce the several well-know neural network models and layers used in this study, i.e., BERT and multi-head attention (the layers in BERT), and pointer-generator network used for comparison with our model. In section 3, we describe an overview of the proposed model along with its structure. In section 4, we detail the settings and dataset applied while training our model. In section 5, the results of the experiments described in section 4, are evaluated, along with those of another model, i.e., a pointer-generator network. Based on the evaluations, we compare our model with a pointer-generator network based on scores from the Extracted places and ROUGE-N [5]. In section 6, we provide some concluding remarks regarding our experiments and discuss areas of future study.

2 Related Studies

2.1 Pointer-Generator Networks

A pointer-generator network [6] is an LSTM-based model for an abstractive summarization. Long short term memory(LSTM) [7] is a neural network model that is often used for natural language processing. Figure 1 shows a model overview of a pointer-generator network. The most significant differences between this model and our approach are the structures, i.e., the “copy mechanism” and “coverage mechanism”. A pointer-generator network can-not compute in parallel owing to the structure of the model. In addition, it requires a long time and a large amount of data for training to obtain the meanings of both words and sentences at the same time.

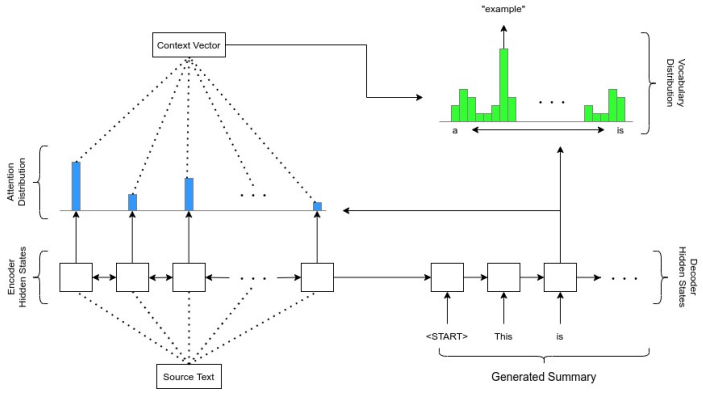

Figure 1: Overview of pointer-generator network. As shown on the left side of the figure, an ’Attention Distribution’ helps handle out-of-vocabulary words. In addition, as shown on the right side, the ’Vocabulary Distribution’ resolves the repeating problem. Referred to [6]

Copy mechanism A neural network outputs the probability of a predetermined group of words when generating them. Therefore, it cannot output proper nouns. However in sentence summaries, proper nouns are often important. Thus, a pointer-generator network dynamically allocates the words in the input text to the IDs using a copy mechanism and thus the model can handle out-of-vocabulary words.

Coverage mechanism Sentence summarization tasks occasionally have a problem of generating the same words repeatedly. A pointer-generator network incorporates a mechanism called a coverage mechanism to solve this problem. The coverage mechanism keeps the distribution of the words generated up to the t-step, and then penalizes them for producing the same word. The penalty is achieved by adding a new term to the loss function.

2.2 BERT

In recent years, pre-training models, such as BERT, have been widely incorporated into neural network models. In particular, models trained with BERT have achieved state-of-the-art performance in natural language processing tasks; BERT has been pre-trained with a large unlabeled corpus and can be fine-tuned with another corpus to achieve better performance. Figure 2 shows the process of pre-training and fine-tuning.

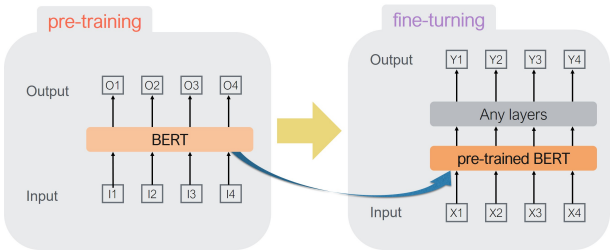

Figure 2: BERT pre-training and fine-tuning processes. A pre-training of the model is shown on the left side of this figure, and a fine-tuning is shown on the right side.

We briefly describe the structure of the BERT model with reference to [2]. BERT has several layers; each layer has a Multi-Head Attention and a linear affine with a residual connection. In our experiment, we utilize the BERT-based model which has 12-layers and 768-hidden sizes.

Pre-training In a general neural network model, the model is trained at once for the tasks. Then, the model is trained for each task at the same time as the meanings of the words and sentences are obtained. For this reason, a large amount of data is needed to train the model. However, there are two processes applied during BERT training, pre-training and fine-tuning. Pre-training tasks of BERT are called masked language model (MLM) and next sentence prediction (NSP). An MLM is a task for predicting masked words in an input text, and an NSP is a task for predicting the next sentence of an input sentence. Through a pre-training, the model obtains the meanings of the words and sentences, as well as the coherence of the texts.

Fine-turning After pre-training of the model, we re-train the model for a specified task, in this case, a summarization. A pre-trained model already has the obtained meanings of words and sentences, and thus we simply tune the model toward our task. This means it can be trained with a low data quantity. At the fine-tuning

stage, we use the pre-trained BERT model and some layers, as shown in Figure 2.

2.3 Multi-Head Attention

BERT and a transformer primarily comprise multi-head attention layers. The multi-head attention divides the attention input into multiple parts and concatenates them with multiple outputs. Multi-head attention is more accurate than attention and can be calculated through (1), (2), and (3) as described in reference [4].

$$Attention\,(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right) \tag{1}$$

$$head_i = Attention\left(QW_i^Q, KW_i^K, VW_i^V\right) \tag{2}$$

$$MultiHead\,(Q,K,V) = Concat\,(head_i,\ldots,head_n)\cdot W^O \tag{3}$$

Equation (1) represents a formula for a single input query in multi-head attention, in which the input is split into units of a head as shown in (2), and each output is concatenated as shown in (3) to generate the overall output. In practice, we compute the attention function on a set of queries simultaneously, packed together into a matrix Q. The keys and values are also packed together into matrices K and V, and W_i^Q, and W_i^K, and W_i^V are the parameter matrices. Here, d_k used in (1) is the number of dimensions of key K.

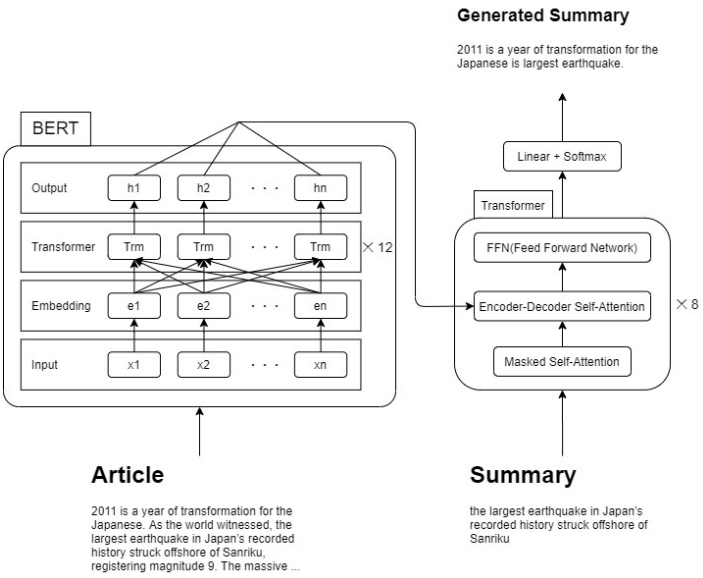

Figure 3: Overview of our text summarization model. Our model includes BERT in the encoder and a transformer in the decoder. During the training phase, we set the article body and summary into the encoder and decoder inputs, respectively.

3 Model

Figure 3 shows an overview of the model used in our experiment.

3.1 Encoder

Several pre-learning models, such as BERT, have been widely utilized in encoder-decoder models. Because BERT is efficient for fast training with high precision, it ensures a higher accuracy in existing models. In a study conducted by Zhang et al. [3], BERT is used as the encoder to achieve state-of-the-art performance for an abstract text summarization task. In this experiment we applied a pre-learning model BERT as the encoder.

3.2 Decoder

In our model, we configured a transformer-based decoder as an encoder. It was not the proposed decoder choice for BERT to send the output generated by the encoder to the input of the decoder at the same time.

Furthermore, we opted for a transformer-based decoder as opposed to a recurrent neural network (RNN) such as an LSTM [7] and GRU [8] for the following reasons: (1) Time for training - Transformer-based decoders are built using multi-head attention, which can perform parallel computations, and are, hence, faster. (2) Accuracy - Transformers are far more accurate than RNNs on machine translation tasks. (3) Long-range dependency - The attention used in the transformers makes it easier to learn long-range dependencies compared to RNNs such as an LSTM.

3.3 Abstractive summarization model

The input is denoted as $X = \{x_1, x_2, \ldots, x_n\}$, sequence representing sentence breaks is denoted as $S = \{s_1, s_2, \ldots, s_n\}$, and the corresponding summary is denoted as $A = \{a_1, a_2, \ldots, a_n\}$. We started by entering X and S into BERT.

If $f_{sen}(x)$ is assumed to be the number of sentence of x, sequence S is computed as $S = f_{sen}(X)\ mod\ 2$. The resulting BERT encoder output is denoted as H. Next, we input H and the output of the decoder at the t-th time step.

The probability of the vocabulary at the t-th time step can be obtained as shown in (4). This probability was conditioned on the decoder output until the t-th time step and the output of the encoder H.

$$P_t(w) = f_{decoder}(w \mid H, Y_{<t}) \tag{4}$$

The loss of training L, is calculated as shown in (4) using the probability of vocabulary $P_t(w)$.

$$L = -\sum_{i=0}^{n} \log P(y_t \mid H, a_{t-1}) \tag{5}$$

4 Experiments

4.1 Setting

During this experiment, we used a pre-trained model with BERT that was developed at the Kurohara and Kawahara laboratories of Kyoto College [9]. Most of the BERT hyperparameters were same as those of BERT-Base (i.e., 12-layers, 768-hidden, and 12heads) in [2]. The model was trained for 30 epochs with 1.8 billion Japanese Wikipedia corpuses. The input text was divided into sub-words with byte pair encoding (BPE [10]) using the morphological analysis system Juman++ [11].

The vocabulary size stood at 32,000 words. The decoder in our model comprises eight multi-head attention layers; the division of hidden size as 3,072; and the embedding vector as 768, similar to the encoder.

We used Adam as the model optimizer, and set the parameters as $\beta_1 = 0.9, \beta_2 = 0.999, and \epsilon = 1.0 \times 10^{-9}$. The maximum learning rate was set at 1.0×10^{-9}. In addition, the dynamic learning rate was adopted as the model learning rate in [4].

The learning rate is computed as shown in (6),

$$lr = max\_learning\_rate * \frac{min\left(cs^{-0.5}, ws^{-1.5} * cs\right)}{ws^{-0.5}} \quad (6)$$

In principle, the learning rate increases linearly up to the warmup step (ws). If the current step (cs) exceeds the warmup step, the learning rate gradually decreases. The learning rate peaks when the current step and warmup step are equal. At this point, the learning rate is denoted as $lr = max\_learning\_rate$. For our experiment, we set $warmup\_step\,(ws) = 4000$ and $max\_learning\_rate = 0.0001$ for training. The model includes BERT with 12 multi-head attention layers. The batch size is set to 4 for GPU memory because the maximum input sequence was set to 512.

Figure 4: An example of Livedoor news corpus. Every article has three sections, a title, a summary and an article body. Each section is labeled in the above figure.

4.2 Dataset

The Livedoor News corpus contains 130,000 Japanese news articles from Livedoor News, each news article is accompanied by a three-line summary. The article text and summary are set as input and output of the experiment, respectively. From the dataset, 100,000 data points were used for training and 30,000 data points were used for validation. The maximum length of the input sequence was set to 512 tokens, but some data values exceeded this limit. In such cases, only the first 512 tokens were entered into the model.

Every article on the Livedoor news website has text and a three-line summary. Figure 4 shows an example of an article body and summary. In this example, the article body at the bottom of the figure is input into the model, and the summary at the middle of the figure is output to the model.

4.3 Training

If 100,000 live news corpus data points in 1 epoch were trained in 15 epochs, the loss of each lexicon in the validation data was about 6.5 words, and the accuracy of identifying the correct lexicon was about 67%. The graphics card used for training was a Titan X (Pascal) with 12 GB of memory. The training took about three days.

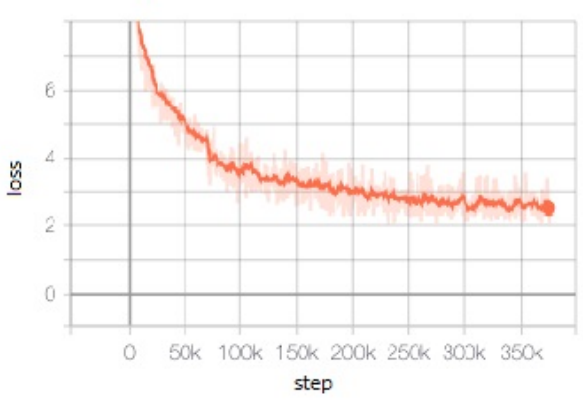

Figure 5: Outputs of loss function for every training step. A gradual decrease in the numbers of loss function outputs is implied in this figure.

5 Results and Discussions

In this section, we can see the evaluations of the generated summaries with two models, i.e., our model (BERT+Transformer) and a pointer-generator network. From the generated summaries, we evaluated the models on the following criteria:

- Grammatical Accuracy
- Vocabulary mistake
- importance level

To check the validity of the generated sentences, we focused on grammatical and vocabulary errors. We also extracted and checked which parts of the original text were important for the generation of the summary text. The above three points are evaluated subjectively. In addition, we calculated and compared their respective ROUGE-N scores.

5.1 Generated summaries

The texts in the appendix are the article bodies and their summaries that generated by the two trained models, our model and the pointer-generator network (translated from Japanese into English). The texts under "Input" are article bodies in the Livedoor news corpus, and the texts under "Output" are summaries generated by our model and the

pointer-generator network from each article body. In the appendix, we provide three article bodies and generated their summaries as examples. When generating the text, a beam search with a width of 4 was used. Prior to the training process, WordPiece [12] was used to further divide out-of-vocabulary words into multiple words. For example, the word "smoke" which is not in the vocabulary of the model, is able to represent the combination of "smo" and "ke" in the vocabulary. For the quantitative effect of WordPiece, refer to an ablation study [13].

One problem that arises during the summary sentence generation is the "repetition problem". This means that an arbitrary phrase will repeatedly appear in the summary sentence. We used a REPEAT block to get around this. The repeat block detects the repetition of a sentence in n-grams increments and returns to the (n-1)-steps before the repetition and reapplies the word generation. In this experiment, we used the repeat block of the tri-gram that had the highest ROUGE-N score.

For example, you can see the *t*-th step generated summaries by the model as follows:

summary : It is going to rain in **the(1) the(2) the(3)**

In the above example, the same three words are generated from the end of the word, and thus the approach returns to the two steps prior to "the(1)" and regenerates words other than the those previously generated.

summary : It is going to rain in **the(1) evening**

Input 1 in the appendix demonstrates that both models were able to learn correctly as, it can be seen, to some extent, that the summary text retained the key points of the input text. In addition, there are no errors in grammar or vocabulary. However, if we look at the summary of a pointer-generator network, the sentence "The Berlin International Film Festival Silver Bear Award is one of the three biggest film festivals in the world for "smoke". " is not extremely important. Therefore, in the above example, we thought the summary generated by our model was better than that of the pointer-generator network owing to the importance.

Input 2 in the appendix is regarding the risks to the human body from sitting for long periods of time. At first glance, both summaries appear to be correct, but the word "skillfully" is included in the summary of our model. This is not the word that should be used. Therefore, we felt that the pointer-generator network was able to produce a more accurate summary in this example than our approach.

Input 3 in the appendix deals with the response of a real estate company when a house is an accidental property. No erroneous words appear in either summary. However, the sentence "It is said that the building should be disliked because it is a place to live continuously." in the summary of our model, is not grammatically correct. In terms of content, however, we believe that our model is able to extract the more important parts. For example, the content regarding payment is fairly important but is not included in the summary of the pointer-generator network.

In the present example, our model extracts the more important parts, but the grammar was correct for the pointer-generator network. From the three examples provided in the appendix, it appears that the pointer-generator network has fewer vocabulary and grammatical errors, although our model is able to extract more of the important content. In addition, you can see that the locations that the pointer-generator network extract is mostly in the first half of the input text. By contrast, our model extracts from all throughout the input text. In other words, our model is more expressive, whereas the pointer-generator network is less prone to grammatical and vocabulary errors. In addition, because the pointer-generator network has a copy mechanism, we thought it might be related to the fact that the model is somewhere between abstract and extractive text.

5.2 *Extracted places*

We will now see which parts of the input text are extracted by the summary statement generated in the previous section. Figure 6, 7 and 8 shows the places that the model extracts from the input texts. Each horizontal axis represents the number of words, and the bar graph shows the word positions referenced by the models.

Figures 6, 7 and 8, indicate that the locations extracted by the pointer-generator network are mostly in the first half of the input texts. By contrast, those extracted by our model(BERT+Trasnformer) are scattered throughout the input text. This leads us to believe that our model may be able to understand the entire content of the input text better than the pointer-generator network.

Figure 6: Input 1 highlighting the extracted locations

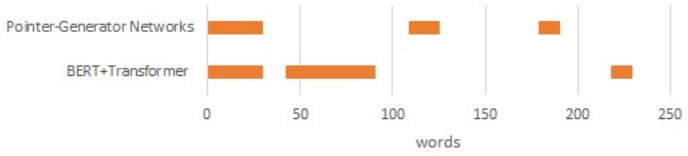

Figure 7: Input 2 highlighting the extracted locations

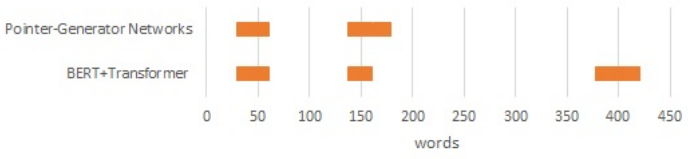

Figure 8: Input 3 highlighting the extracted locations

5.3 ROUGE-N socre

Next, we compare the pointer-generator network with our model using the ROUGE-N [5], which is an index frequently used in the evaluation of summary sentences, which represents the percentage of agreement between words in N-gram units.

As an example, let us try to calculate the ROUGE-N score of the following sentence.

summary : It is going to rain in the evening today
references : It is going to rain today

Referring to [5], the equation for obtaining the ROUGE-N score is shown below.

$$ROUGE_N = \frac{\sum_{C \in Reference} \sum_{Gram_n \in C} Count_{match}(Gram_n)}{\sum_{C \in Reference} \sum_{Gram_n \in C} Count(Gram_n)} \quad (7)$$

Considering the above example, we have the same six pairs on uni-gram, and four pairs on a bi-gram. From this, the ROUGE-1,2 score of the example could be derived as follows:

$$\begin{aligned} ROUGE_1 &= \frac{6}{9} \\ &= 0.667 \\ ROUGE_2 &= \frac{4}{8} \\ &\fallingdotseq 0.500 \end{aligned}$$

In our experiment, we calculated the ROUGE-1,2 scores with 1,000 Japanese news articles (sample size $n = 1,000$) chosen at random from the verification datasets. The standard error, *SE*, calculated using the unbiased standard deviation, u, and the sample size, n, is expressed as follows:

$$SE = \frac{u}{\sqrt{n}} \quad (8)$$

Using the above equation, the margin of error, e, for the 95% confidence interval (95% CI) is expressed as in Equation 9.

$$e = \pm 1.96 \times SE \quad (9)$$

The ROUGE-1,2 scores, the standard errors, *SE*, and the unbiased standard deviations, u, for both the models are presented in Table 1.

Table 1: ROUGE-1, 2 scores for pointer-generator networks and our model (BERT + transformer)

Model		Score	u	e
Pointer-Generator Networks	ROUGE-1	0.463	0.123	±0.008
	ROUGE-2	0.221	0.142	±0.009
BERT + Transformer	ROUGE-1	0.470	0.107	±0.007
	ROUGE-2	0.215	0.126	±0.008

Table 1 indicates that the ROUGE-1,2 scores of our model and the pointer-generator network were similar. However, our model's structure was designed for multiple tasks, whereas the pointer-generator network was a model designed for text summarization tasks because of the copy mechanism. In addition, our model generated all the words in the output sentences. Thus, we hypothesized that our model achieved a better comprehension of the text.

5.4 Word Mover's Distance

Next, we compared the pointer-generator network with our model using word mover's distance (WMD) [14] to evaluate their semantic similarity. The WMD is a novel distance function between text documents. The ROUGE scores are based on literal word overlap; hence, it is necessary to evaluate a score that is not based on literal word overlap. Among a number of recently proposed semantic similarity metrics, WMD is shown to be the most reasonable solution to measure semantic similarity in reformulated texts [15]. We evaluated the WMD for both models with 1,000 Japanese news articles chosen at random from the verification datasets and used the Japanese word embedding vectors [16] to calculate the WMD. Table 2 presents the WMD scores, unbiased standard deviation, and margin of error for the pointer-generator networks and our model.

Table 2: WMD of pointer-generator networks and our model (BERT + transformer)

Model	WMD	u	e
Pointer-Generator Networks	0.533	0.122	±0.008
BERT + Transformer	**0.521**	0.106	±0.007

As shown in Table 2, our model has a slightly lower WMD score than the pointer-generator network. This implies that the summaries generated by our model were more similar to the article's summaries than those generated by the pointer-generator network. Thus, the hypothecate that our model achieved a better comprehension of the text than the pointer-generator network was strengthened.

6 Conclusion and Future works

We conducted an experiment to demonstrate an abstractive summarization of Japanese text with a neural network model using BERT. In addition, we conducted a comparison between the qualitative and quantitative aspects of our model and a model frequently used for text summarization, i.e., a pointer-generator network. Our model is composed of a BERT encoder and a transformer-based decoder. The dataset used in this paper was the Livedoor news corpus consisting of 130,000 datapoints, of which 100,000 were used for training.

The results of the experiment revealed that the model was able to learn correctly as the summary sentence captured the key points of the text to a certain extent. However, the contents of the summary sentence were repeated, and the model could not handle unknown words. As a solution, we applied two mechanisms, a repeat block and WordPiece. The repeat block detects repeated n-grams in the generated summaries and regenerates summaries at the (t-n-1)-th step excluding the repeat word. To handle this, WordPiece further divides the out-of-vocabulary words into sub-words.

The results of evaluations of our model and the pointer-generator network revealed that their ROUGE-1,2 scores were similar. However, our model's structure was designed for multiple tasks, whereas the pointer-generator network was a model designed for text summarization tasks because of the copy mechanism. Therefore, we hypothesized that our model achieved a better comprehension of the text. In addition, the comparison of the two models to evaluate their semantic similarity, that is, the WMD, strengthened our hypothecate. While extracting phrases, the pointer-generator network extracted phrases from the first half of the input text, whereas our model extracted phrases from throughout the text. However, a qualitative evaluation of our model revealed that it made numerous grammatical and vocabulary mistakes. We believe that this problem can be solved by improving the model.

In the future, we will explore these recommendations through additional experiments and compare the results.

7 Acknowledgements

This work was supported by JSPS KAKENHI Grant Number 19K12906.

References

[1] Y. Iwasaki, A. Yamashita, Y. Konno, K. Matsubayashi, "Japanese abstractive text summarization using BERT," in 2019 International Conference on Technologies and Applications of Artificial Intelligence (TAAI), 1–5, 2019, doi:10.1109/TAAI48200.2019.8959920.

[2] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," arXiv preprint arXiv:1810.04805, 2018, doi:10.18653/v1/N19-1423.

[3] H. Zhang, J. Cai, J. Xu, J. Wang, "Pretraining-Based Natural Language Generation for Text Summarization," 789–797, 2019, doi:10.18653/v1/K19-1074.

[4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, u. Kaiser, I. Polosukhin, "Attention is All You Need," in Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17, 6000–6010, Curran Associates Inc., Red Hook, NY, USA, 2017, doi:10.5555/3295222.3295349.

[5] C.-Y. Lin, E. Hovy, "Automatic Evaluation of Summaries Using N-Gram Co-Occurrence Statistics," in Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1, NAACL '03, 71–78, Association for Computational Linguistics, USA, 2003, doi:10.3115/1073445.1073465.

[6] A. See, P. J. Liu, C. D. Manning, "Get To The Point: Summarization with Pointer-Generator Networks," in Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 1073–1083, Association for Computational Linguistics, Vancouver, Canada, 2017, doi:10.18653/v1/P17-1099.

[7] S. Hochreiter, J. Schmidhuber, "Long Short-Term Memory," Neural Comput., **9**(8), 1735–1780, 1997, doi:10.1162/neco.1997.9.8.1735.

[8] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," arXiv preprint arXiv:1406.1078, 2014, doi:10.3115/v1/D14-1179.

[9] S. Kurohashi, Y. Murawaki, "BERT pretrained model with Japanese corpus," http://nlp.ist.i.kyoto-u.ac.jp/index.php?BERT%E6%97%A5%E6%9C%AC%E8%AA%9EPretrained%E3%83%A2%E3%83%87%E3%83%AB (accessed 2019-07-16).

[10] R. Sennrich, B. Haddow, A. Birch, "Neural Machine Translation of Rare Words with Subword Units," in Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics 1, 1715–1725, Association for Computational Linguistics, Berlin, Germany, 2016, doi:10.18653/v1/P16-1162.

[11] T. Kudo, K. Yamamoto, Y. Matsumoto, "Applying Conditional Random Fields to Japanese Morphologiaical Analysis," IPSJ-NL, **161**, 89–96, 2004.

[12] M. Schuster, K. Nakajima, "Japanese and korean voice search," in 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 5149–5152, IEEE, 2012, doi:10.1109/ICASSP.2012.6289079.

[13] K. Bostrom, G. Durrett, "Byte Pair Encoding is Suboptimal for Language Model Pretraining," ArXiv, 2020, doi:abs/2004.03720.

[14] M. Kusner, Y. Sun, N. Kolkin, K. Weinberger, "From word embeddings to document distances," in International conference on machine learning, 957–966, 2015, doi:10.5555/3045118.3045221.

[15] I. Yamshchikov, V. Shibaev, N. Khlebnikov, A. Tikhonov, "Style-transfer and Paraphrase: Looking for a Sensible Semantic Similarity Metric," arXiv, 2020, doi:abs/2004.05001.

[16] M. Suzuki, "Japanese Wikipedia entity vector," http://www.cl.ecei.tohoku.ac.jp/~m-suzuki/jawiki_vector/ (accessed 2020-10-30).

Appendix

The following are three articles from Livedoor news and a summary of our model and the pointer-generator network generated as examples. The examples below were used for the discussions provided in section 5.

Input 1 Director Wayne Wang from Hong Kong, who has won the Silver Bear Award at the Berlin International Film Festival, one of the three biggest film festivals in the world in the movie "Smoke" (1995), and Japan's world-famous Beat Takeshi and Nishijima Hidetoshi. The movie "When a Woman Sleeps" (published on February 27). It was the first time in 12 years since "Blood and Bone" starring in a film other than his own work in 2004, and said "I was so worried about my acting that I held my head," he said unexpectedly. Is it because of the mysterious style that the boundaries between dreams and delusions and reality are vague? Three people talked to each other. This film is a movie based on a short story by Javier Marias, a Spanish writer, and is set in a quiet resort hotel, and a mysterious couple (Takeshi and Shiori Nana Shiori), two years older than their parents and children. It is a mystery of love and hate that stares at a writer (Nishijima) who is obsessed with a relationship that has become obsessed with and has run out of curiosity. "I thought that the main character was Nishijima," said the smiling Takeshi, and read the script, "I thought I was foolish. A normal script would tell the story and the ending." He said that he had a hard time understanding a complicated story. Nishijima, who praises himself for being more acquainted with movie expressions, relieved, "If I say "good", I'm sure there will be no mistake." I watched all the acting all the time. When I finished watching, I didn't even remember what kind of movie it was," said an unknown side of Takeshi called "Kitano of the World." Nishijima, who immediately decided to appear, "Wayne Wang is a work that shoots Beat Takeshi, so it will come out (laughs)," said "This time especially Beat Takeshi is obsessed with love, even at the shooting site. I was shocked a lot and was really moved. It was wonderful." Director Wang, who worked on Japanese movies for the first time, also said, "I wanted to work with Takeshi rather than want to shoot in Japan." Takeshi played an unusual role in keeping a video recording of a young beautiful woman (Kana) sleeping. When asked, "What do

you want to leave your dear ones?", "I've been in the entertainment world for a long time, so I thought, "I thought that the gag of Takeshi at that time, that laugh was amazing." It's okay if you can leave one," he said. (Interview and text: Megumi Shibata)

Output 1 (Pointer-generator network) The Berlin International Film Festival Silver Bear Award is one of the three biggest film festivals in the world for "smoke." For the first time in 12 years, Takeshi said, "I was so worried about my performance that I held my head." "The boundaries between dreams and delusions and reality are vague and mysterious.

Output 1 (BERT+Transformer) A movie "When a Woman Sleeps", in which Takeshi Beat and Hidetoshi Nishijima formed a tag. Beat Takeshi was said to have been attached to love. "I'm worried about my performance, and I'm worried about the finished product, too," he said.

Input 2 "Mentai Wide-Fukuoka Broadcasting" At the corner of "Special News THE Slide Show" broadcasted on February 9, 2016, we covered the risks to your body from sitting for a long time. The image is an image (taken by Andrea Arbogast, from Flickr). A study of 220,000 men and women over the age of 45 in Australia found that the risk of death was 1.4 times higher for people sitting 11 hours or more a day than for those sitting less than 4 hours a day I got the result. It seems that the risk of obesity, diabetes, cancer, etc. increased, causing death. The detailed cause is unknown, but it is also said that "myokine" that suppresses heart disease is not secreted unless the muscles are contracted. This excessive sitting puts a heavy burden on the lower back, with 1.4 times being normally seated and 1.85 times being leaned forward. Even in Fukuoka, more and more offices are incorporating "sometimes standing work". More than 90% of the desks sold in Sweden and Denmark are of the type that can be moved up and down, and by using this you can do "occasionally standing work". Even in Japan, Rakuten has changed the desks of all employees so that they can be moved up and down, to avoid sitting down. Even in Fukuoka, some companies are making their desks movable up and down and changing the height of the desks themselves. When I asked when to change the height of the desk, this was the most common opinion. Because I may feel drowsy after lunch, switch to standing work at such times. When you are tired from working on a computer or lose your concentration, start a standing job. Stand up when your shoulders, waist, or neck hurts. It seemed that it was good to stand up, and there were interesting opinions. I have less stiff shoulders, lower back pain, and neck pain. Increased communication from the surroundings. When I sat down, I was staring at my computer, but by standing up, I was able to see my surroundings better and start talking. Since I also work as a deskwork worker, I sometimes feel pain in my lower back and shoulders after sitting for a long time. For this reason, I stand up and relax once every hour, but I found it useful to have a desk like this. If you keep sitting down, you may not be able to see the surroundings because you are staring at a desk or a personal computer, so I would like you to incorporate such desks in various offices. (Writer: Piyoko).

Output 2 (Pointer-generator network) At the "Special News THE Slide Show" on the 9th, I heard about the risks to the body from sitting for a long time. It is said that "myokine" that suppresses heart disease is not secreted unless the muscles are contracted. It is said that more and more offices in Fukuoka have adopted "sometimes standing work".

Output 2 (BERT+Transformer) In the program on the 9th, we discussed the risks to the body from sitting for a long time. The risk of obesity, diabetes, cancer, etc. is high, and the risk of death is 14 times higher. In addition, he changed the man's desk to something that can be skillfully moved up and down to avoid sitting.

Input 3 Last year, an incident that involved the illegal construction of a large-scale condominium construction project in Yokohama City reminded me of the difficulty of choosing a property, but I finally got the feeling that I could get a house. If it's a so-called "accident property", can you ask the real estate agent for a refund? Lawyer Masami Takeshita answers. (Consultation). At the end of last year, I bought a second-hand house, but when I looked up the reason because the neighborhood was strange, the previous owner was at home It seems that he tried to hang himself. The real estate company that introduced this property did not tell me that it was an accident property. In this case, is it possible to request the real estate agent to refund the purchase price in full? (Answer). The fact that a building purchased for residential use commits suicide and lacks "comfortable living" means that "a building is a place to live continuously, so there are also psychological flaws due to the historical background of the building that should be disliked. (Defect)" is understood. If the buyer makes a contract without knowing that there is a defect, the seller will be liable for the defect guarantee, and if the defect does not meet the contract purpose, the contract will be canceled, and if it does not reach that level, You are obliged to compensate the buyer for damages. In addition, there are court cases that the contract can be canceled if there is suicide and it is judged that it is reasonable to feel that it is not suitable for living because it lacks comfort to live in. .. It is the case that the cancellation of the sale of condominiums, which was found to have committed suicide six years ago when the deposit was paid, was approved. On the other hand, there are some cases in which the suicide was canceled 7 years ago, and there are also cases in the sale and purchase of a plan to demolish an old house in which the degree of dislike of suicide two years ago due to the removal of a building was not a defect. In your case, depending on the time of suicide and the degree of topicality, you may be able to cancel the contract and request a full refund of the price, and even if you do not go that far, the objective evaluation of the suicide building and the purchase price You can consider the difference claim for damages. The important things that the Homebuilding Act requires an intermediary to explain are not the history of the building, such as suicide, but if the trader knows suicide, which is an important factor in determining the buyer's contract decision, There is a statutory obligation to explain and if you remain silent, you will be liable. However, it is questionable to impose accountability forever, and there seems to be a guidance that sales are 5 years and rent is 3 years. In any case, when you buy a used building, you should ask about its provenance, etc., and you can pursue the default liability

of breach of accountability to the lying seller. (Attorney Profile). Masami Takeshita: Born in Osaka in 1946. Graduated from the Faculty of Law, the University of Tokyo. Registered as a lawyer in 1971. * Weekly post February 26, 2016 issue.

Output 3 (Pointer-generator network) Attorney Masami Takeshita explained why a real estate company can request a refund if the property is an ”accident.” It seems that the building purchased for residential use may lack ”comfortable living” due to suicide. ”There are also psychological flaws due to the disgusting historical background of buildings.

Output 3 (BERT+Transformer) A lawyer explains whether you can request a refund when you purchase the accident property in 2017. It is said that the building should be disliked because it is a place to live continuously. Regarding loans, it seems that there is a possibility that you can cancel the contract and request full refund.