ICIC Express Letters
Part B: Applications
Volume **13**, Number **7**, July **2022**

ICIC International ©2022 ISSN 2185-2766
pp. **697–703**

INTERPRETING BERT ATTENTION TRAINED FOR JAPANESE DIFFICULTY CLASSIFICATION FROM THE VIEWPOINT OF GRAMMATICAL FEATURES

Eri Maekawa* and Hajime Murao

Graduate School of Intercultural Studies
Kobe University
1-2-1 Tsurukabuto, Nada-ku, Kobe 657-8501, Japan
*Corresponding author: eri.maekawa@stu.kobe-u.ac.jp; murao@i.cla.kobe-u.ac.jp

Received November 2021; accepted January 2022

Abstract. *Text simplification is often tackled by neural machine translation with parallel corpora, where BLEU and SARI are used for evaluation. They are not methods that directly evaluate the difficulty level of texts but evaluate by comparing generated texts with corresponding simplified texts. Consequently, they cannot measure the difficulty without corresponding texts. In a preliminary study, we applied BERT to evaluating the difficulty level of Japanese texts only from the texts themselves, and it showed a pretty good performance. In this paper, we try to investigate how BERT estimates the difficulty level of Japanese texts. For this purpose, we use grammatical features such as lexical level and syntactic complexity, which are used to calculate the difficulty of Japanese texts in statistical methods. We examine how well "attention" in BERT captures the grammatical features when measuring the Japanese text difficulty. To select the grammatical features, we applied Permutation Importance to Random Forest. As a result, we showed that parts of texts related to the grammatical features got strong attention. This result means that BERT somehow utilized the grammatical features to evaluate the difficulty level of texts.*

Keywords: BERT, Permutation Importance, Attention, Random Forest

1. **Introduction.** In recent years, "Simple Japanese" has been focused on because of the increase of foreigners living in Japan and many studies using machine translation techniques have been published to generate "Simple Japanese" automatically [1, 2]. In machine translation, accuracy is often evaluated by BLEU [3]. BLEU and SARI [4] calculate the score by comparing the generated sentence and the reference one but do not directly measure the difficulty level of the text. To evaluate text simplification methods or algorithms, we require a mechanism to directly measure the difficulty of the text.

Many of the studies on difficulty evaluation of Japanese sentences had used grammatical features based on lexical difficulty and syntactic complexity. Jae-Ho [5] defined a readability formula to measure the difficulty level of sentences. The average sentence length, Chinese word rate, Japanese word rate, verb rate, and particle rate, which are the characteristics of the difficulty level, were weighed by multiple regression analysis. In order to validate the readability formula, he conducted a difficulty test on reading comprehension questions from the old Japanese Language Proficiency Test and asserted that the readability formula could capture the levels from Level 1 to Level 4.

After word embeddings have been proposed [6, 7], machine learning models based on word embeddings have been focused on. BERT [8] is a pre-training model with a Transformer [9]-based attention mechanism that operates as a classification model. BERT has shown that modification of the attention weights affects prediction accuracy in text classification [10]. The purpose of this study is to investigate how BERT estimates the difficulty

DOI: 10.24507/icicelb.13.07.697

level of Japanese text. To do so, we extract features by referring to the previous study [11], and investigate whether they are represented in the attention weights.

The process of the study is shown in Figure 1. The details are described in Chapter 3.

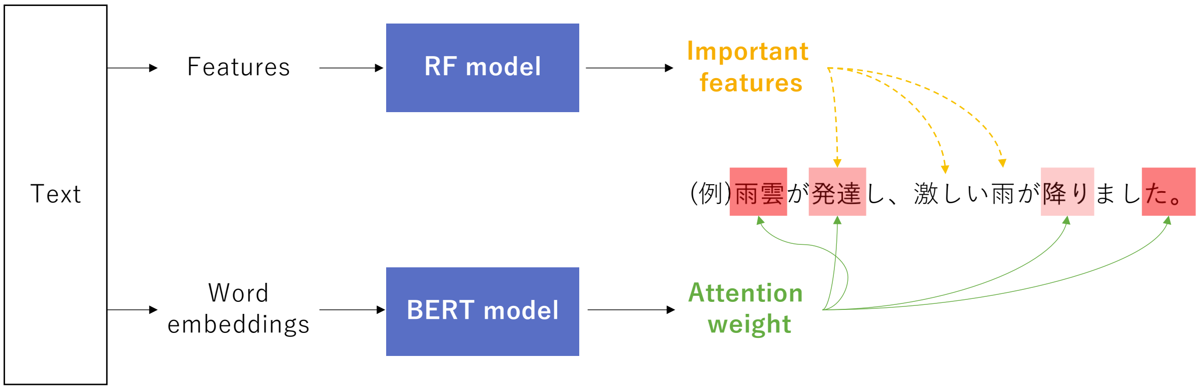


Figure 1. It extracts features from a text and trains Random Forest to determine its difficulty level. The word embedding is also extracted from the text, and is used to train BERT. The relationship between attention weight in BERT and features is examined.

2. **Data Collection and Preparation.** As normal texts, we collected texts by scrapping the news site named NHK NEWS WEB[1]. As simple texts, we collected texts by scrapping the news site named NEWS WEB EASY[2]. This site is where Japanese language teachers and reporters rewrite and publish articles from NHK NEWS WEB for foreigners. The number of articles is 722. From the collected texts, training data of 10,000 sentences and test data of 1,200 sentences were randomly chosen. They are summarized in Table 1.

Table 1. Simple texts and normal texts were collected from NEWS WEB EASY and NHK NEWS WEB respectively.

	Simple texts	**Normal texts**
Source	NEWS WEB EASY	NHK NEWS WEB
Articles	722 articles (July 2020-July 2021)	
Sentences	Training data: 10,000/Test data: 1,200	

3. Methodologies.

3.1. **Extracting word embeddings.** The text was split into morphemes using MeCab [12] and word embeddings were obtained from a pre-trained BERT model. In this study, we used BERT-base published by Tohoku University[3]. The model consists of 12 layers with 768 dimensions of hidden state and 12 attention heads. It was pre-trained using a corpus with 30M sentences generated from the Japanese version of Wikipedia Cirrussearch dump file as of Aug. 31 2020. We obtained word embeddings from a trained Japanese BERT model and fine-tuned the model to predict the difficulty level using data described in Section 2.

3.2. **Choosing features.** The features were calculated from the texts according to the lexicon and grammar. We have selected 13 features. Table 2 shows these 13 features and their basic statistics while Table 3 for normal texts.

[1] https://www3.nhk.or.jp/news/
[2] https://www3.nhk.or.jp/news/easy/
[3] https://github.com/cl-tohoku/bert-japanese

Table 2. Basic statistics of features of simple texts

Features	mean	max.	min.	stddv.
number of words	24.074	71.000	4.000	8.200
kanji rate	0.268	0.721	0.000	0.095
katakana rate	0.029	0.273	0.000	0.040
passive rate	0.001	0.100	0.000	0.007
verbal noun rate	0.026	0.231	0.000	0.034
adverb rate	0.009	0.250	0.000	0.021
reading point rate	0.046	0.429	0.000	0.032
negative rate	0.009	0.167	0.000	0.021
maximum frequency	12876.890	34705.000	3.000	8621.321
mean frequency	2231.809	21512.500	2.500	1662.133
mean distance of dependency	2.006	4.500	0.000	0.561
maximum distance of dependency	6.438	28.000	0.000	3.504
maximum depended number	2.768	8.000	0.000	0.915

Table 3. Basic statistics of features of normal texts

Features	mean	max.	min.	stddv.
number of words	38.468	179.000	3.000	16.683
kanji rate	0.355	0.750	0.000	0.097
katakana rate	0.022	0.231	0.000	0.031
passive rate	0.010	0.125	0.000	0.017
verbal noun rate	0.070	0.500	0.000	0.046
adverb rate	0.010	0.167	0.000	0.019
reading point rate	0.050	0.250	0.000	0.031
negative rate	0.007	0.154	0.000	0.016
maximum frequency	18526.886	34705.000	2.000	8492.536
mean frequency	2722.712	15336.200	1.500	1683.291
mean distance of dependency	2.467	9.472	0.000	0.805
maximum distance of dependency	10.931	64.000	0.000	6.663
maximum depended number	3.474	20.000	0.000	1.275

The features that differ in basic statistics are “the number of words”, “the kanji rate”, “passive rate”, “verbal noun rate”, and “the maximum distance of dependency”. Most of the features of normal text are larger than those of simple text while “the katakana rate” and “negative rate” do not differ significantly. The features were labeled with difficulty levels and trained with Random Forest. Although there is a large difference in the range of values, Random Forest [13] does not compare different features, and it is not necessary to normalize them.

We apply Random Forest in Python scikit-learn package to classifying feature sets into corresponding difficulty levels. We used the default parameters except for the number of decision trees 100.

3.3. **Permutation Importance.** Permutation Importance [14] is a method to measure the importance of features. The importance of each feature is how much it contributes to the prediction accuracy of the model. The method to calculate the importance is as follows: Generate a base-model with a normal data set and calculate the accuracy rate. Then, generate a model with a data for which the order of elements in a single column of features is randomized and calculate the accuracy rate. The randomly sorted features do not work as explanatory variables for prediction. The lower the accuracy of the model, the more important the permuted feature.

4. Results.

4.1. **Accuracy of classification models.** The estimation accuracy of Random Forest was 82.6%. The accuracy of BERT was 96.4%. It was proved that the BERT model with embedded words is more accurate than the Random Forest classification with features.

4.2. **Important features.** Figure 2 shows the Permutation Importance of the 13 types of features. The red dots show the averaged accuracy decrease when randomizing corresponding feature values, where the larger drop means the greater importance. The bars indicate the 95% confidence interval. The chart shows that there is a large difference in the characteristics of "verbal noun rate", "number of words", "passive rate", and "kanji rate". We can say that they have a significant impact on estimating text difficulty.

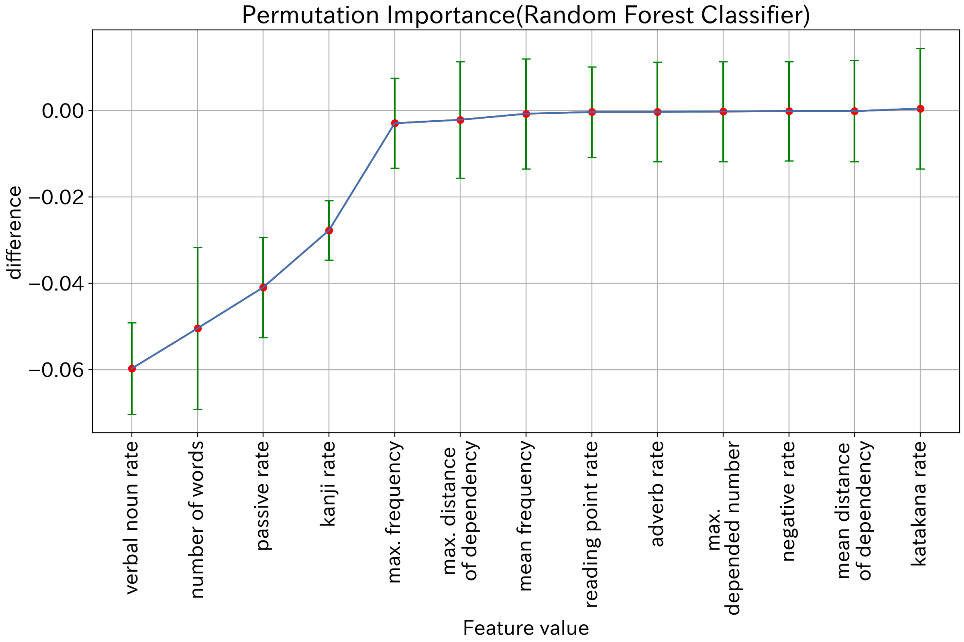


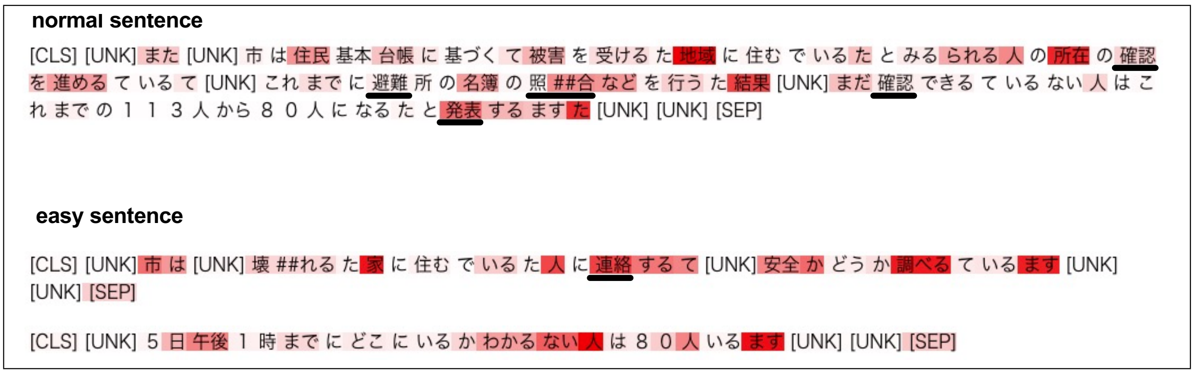
FIGURE 2. Permutation Importance shows verbal noun rate, number of words, passive rate and kanji rate are important.

4.3. **Attention weight.** BERT has shown that modification of the attention weights affects prediction accuracy in text classification [10]. We examine the relationship between the Permutation Importance and the attention weights of BERT.

Verbal Noun Rate. Figure 3 shows the sum of all 12 attentions. Words with higher attentions are shown in darker color. We consider these words influenced the reasoning. The underlines indicate the verbal noun. Most of the "verbal noun rates" get higher attention. It is possible to argue that BERT's classification model focuses on "verbal nouns rates" in order to make inferences.

Number of Words. Figure 4 shows the 11th attention. The last words in the sentence are underlined.

Each attention has unique characteristics. In the 11th layer, that the last words have strong attention suggests that attention may be focusing on the number of words in the sentence.



normal sentence

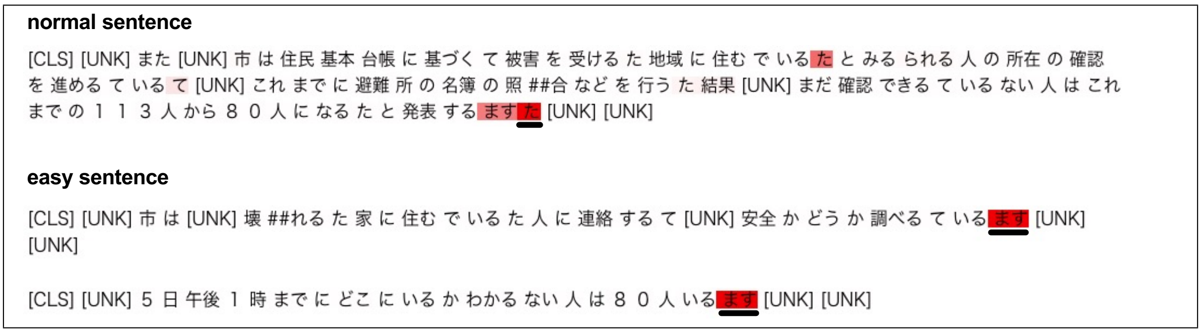
[CLS] [UNK] また [UNK] 市 は 住民 基本 台帳 に 基づく て 被害 を 受ける た 地域 に 住む で いる た と みる られる 人 の 所在 の 確認 を 進める て いる て [UNK] これ まで に 避難 所 の 名簿 の 照 ##合 など を 行う た 結果 [UNK] まだ 確認 できる て いる ない 人 は これ まで の 1 1 3 人 から 8 0 人 に なる た と 発表 する ます た [UNK] [UNK] [SEP]

easy sentence

[CLS] [UNK] 市 は [UNK] 壊 ##れる た 家 に 住む で いる た 人 に 連絡 する て [UNK] 安全 か どう か 調べる て いる ます [UNK] [UNK] [SEP]

[CLS] [UNK] 5 日 午後 1 時 まで に どこ に いる か わかる ない 人 は 8 0 人 いる ます [UNK] [UNK] [SEP]

Figure 3. Sum of all 12 layers is shown. Most of the attention in the verbal noun rate gets attention.



normal sentence

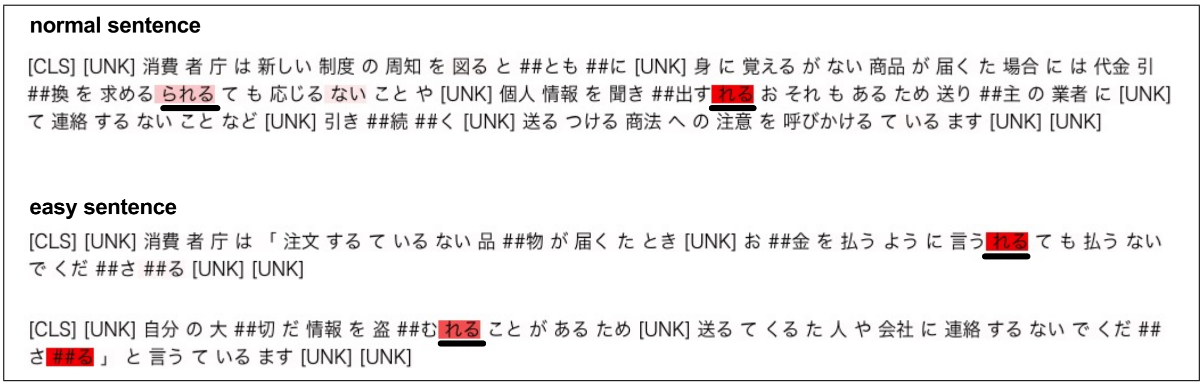
[CLS] [UNK] また [UNK] 市 は 住民 基本 台帳 に 基づく て 被害 を 受ける た 地域 に 住む で いる た と みる られる 人 の 所在 の 確認 を 進める て いる て [UNK] これ まで に 避難 所 の 名簿 の 照 ##合 など を 行う た 結果 [UNK] まだ 確認 できる て いる ない 人 は これ まで の 1 1 3 人 から 8 0 人 に なる た と 発表 する ます た [UNK] [UNK]

easy sentence

[CLS] [UNK] 市 は [UNK] 壊 ##れる た 家 に 住む で いる た 人 に 連絡 する て [UNK] 安全 か どう か 調べる て いる ます [UNK] [UNK]

[CLS] [UNK] 5 日 午後 1 時 まで に どこ に いる か わかる ない 人 は 8 0 人 いる ます [UNK] [UNK]

Figure 4. The eleventh attention is shown. The last words are underlined.



normal sentence

[CLS] [UNK] 消費 者 庁 は 新しい 制度 の 周知 を 図る と ##とも ##に [UNK] 身 に 覚える が ない 商品 が 届く た 場合 に は 代金 引 ##換 を 求める られる て も 応じる ない こと や [UNK] 個人 情報 を 聞き ##出す れる お それ も ある ため 送り ##主 の 業者 に [UNK] て 連絡 する ない こと など [UNK] 引き ##続 ##く [UNK] 送る つける 商法 へ の 注意 を 呼びかける て いる ます [UNK] [UNK]

easy sentence

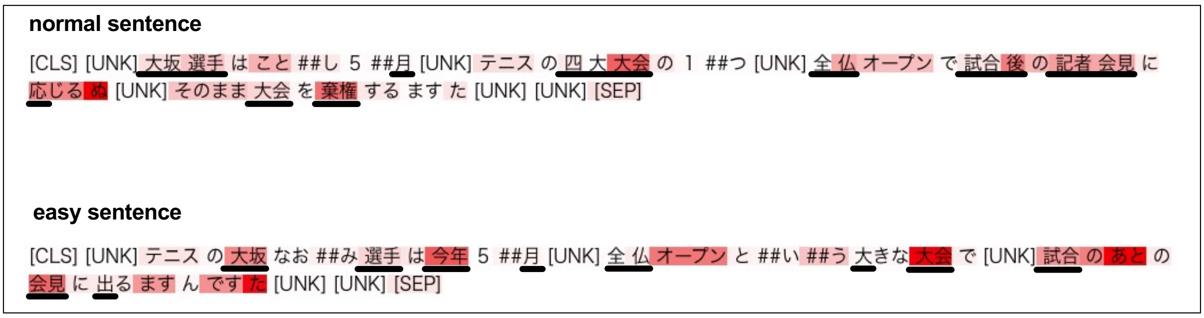
[CLS] [UNK] 消費 者 庁 は 「 注文 する て いる ない 品 ##物 が 届く た とき [UNK] お ##金 を 払う よう に 言う れる て も 払う ない で くだ ##さ ##る [UNK] [UNK]

[CLS] [UNK] 自分 の 大 ##切 だ 情報 を 盗 ##む れる こと が ある ため [UNK] 送る て くる た 人 や 会社 に 連絡 する ない で くだ ##さ ##る 」 と 言う て いる ます [UNK] [UNK]

Figure 5. The sixth attention is shown. The passive words are underlined. Since the feature and attention weights are consistent, we can say that the interpretability is high.

Passive Rate. Figure 5 shows the sixth attention. The underlined words are passive. It shows that attention is stronger for all the underlined words. We can say that "the passive rate" has an impact on the estimation of difficulty level.

Kanji Rate. Figure 6 shows the sum of all 12 attentions. Kanji characters are underlined. Kanji characters are in units of one character each, so they do not match the token units. Therefore, it is not possible to see exactly whether attention is focusing on Kanji characters or not. However, we can see that the weight is stronger for tokens that contain most Kanji.



normal sentence

[CLS] [UNK] 大坂 選手 は こと ##し 5 ##月 [UNK] テニス の 四 大 大会 の 1 ##つ [UNK] 全 仏 オープン で 試合 後 の 記者 会見 に 応じる ぬ [UNK] そのまま 大会 を 棄権 する ます た [UNK] [UNK] [SEP]

easy sentence

[CLS] [UNK] テニス の 大坂 なお ##み 選手 は 今年 5 ##月 [UNK] 全 仏 オープン と ##い ##う 大きな 大会 で [UNK] 試合 の あと の 会見 に 出る ます ん です た [UNK] [UNK] [SEP]

FIGURE 6. Sum of all 12 attention is shown. Most of the words with kanji get attention.

5. **Conclusions.** In order to investigate the relationship between attention of BERT and grammatical features, we built a Random Forest classifier using grammatical features and a BERT classifier using word embeddings. We extracted grammatical features and estimated the difficulty level with Random Forest. The accuracy was 82.6%. We assigned a difficulty level to the BERT embeddings retrieved from the text and built a classifier to guess the label. The accuracy was 96.4%. In the current experiment, word embeddings using BERT outperformed by 13.8%.

We searched for important features in Permutation Importance. We found that the following features are important: “verbal noun rate”, “number of words”, “passive rate”, and “kanji rate”.

In BERT, each layer of attention has been found to be of different importances in predicting attention. As the result of the experiments, the important features were represented in the BERT attention. The 11th layer of attention was clearly weighed at the number of words. The 6th layer of attention was clearly weighed at the “passive rate”. In addition, the “kanji rate” and “verbal noun rate” tended to be weighed more strongly in attention, which was the sum of all layers. The kanji is a smaller token than the word, so it is difficult to use the attention weight to interpret it. Previous studies have shown that words containing kanji tend to be more difficult [15], and we suggest that word embeddings using BERT capture the same feature.

The accuracy of the BERT classifier is high, and it was suggested that it captures grammatical features when considering attention. As mentioned earlier, the evaluation of text plainness is based on the similarity to the parallel corpus, so the difficulty level is not directly evaluated. We suggest that this study can contribute to the evaluation of the difficulty of the generated text.

Acknowledgement. This work was partially supported by JSPS KAKENHI Grant Number 19K12247.

REFERENCES

[1] A. Katsuta and K. Yamamoto, Improving text simplification by corpus expansion with unsupervised learning, *2019 International Conference on Asian Language Processing (IALP)*, pp.216-221, 2019.
[2] T. Kajiwara and K. Yamamoto, Evaluation dataset and system for Japanese lexical simplification, *Proc. of the ACL-IJCNLP 2015 Student Research Workshop*, pp.35-40, 2015.
[3] K. Papineni, S. Roukos, T. Ward and W.-J. Zhu, BLEU: A method for automatic evaluation of machine translation, *Proc. of the 40th Annual Meeting on Association for Computational Linguistics – ACL'02*, pp.311-318, 2002.
[4] W. Xu, C. Napoles, E. Pavlick, Q. Chen and C. Callison-Burch, Optimizing statistical machine translation for text simplification, *Transactions of the Association for Computational Linguistics*, vol.4, pp.401-415, 2016.
[5] L. Jae-Ho, Readability research for Japanese language education, *Waseda Studies in Japanese Language Education*, vol.21, pp.1-16, 2016.

[6] P. Bojanowski, E. Grave, A. Joulin and T. Mikolov, Enriching word vectors with subword information, *CoRR*, vol.abs/1607.04606, 2016.

[7] T. Mikolov, K. Chen, G. Corrado and J. Dean, Efficient estimation of word representations in vector space, *arXiv.org*, arXiv: 1301.3781, 2013.

[8] J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, *NAACL-HLT*, 2019.

[9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser and I. Polosukhin, Attention is all you need, *Advances in Neural Information Processing Systems*, pp.5998-6008, 2017.

[10] S. Vashishth, S. Upadhyay, G. S. Tomar and M. Faruqui, Attention interpretability across NLP tasks, *arXiv.org*, arXiv: 1909.11218, 2019.

[11] E. Maekawa and H. Murao, Feature analysis on the difficulty of Japanese, *The 27th Annual Meeting of the Natural Language Processing Society (NLP2021)*, https://www.anlp.jp/proceedings/annual_meeting/2021/pdf_dir/C6-1.pdf, 2021.

[12] T. Kudo, MeCab: Yet another part-of-speech and morphological analyzer, *http://mecab.sourceforge.jp*, https://ci.nii.ac.jp/naid/10027284215/, 2006.

[13] L. Breiman, Random forests, *Machine Learning*, vol.45, no.1, pp.5-32, 2001.

[14] A. Fisher, C. Rudin and F. Dominici, All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously, *Journal of Machine Learning Research*, vol.20, no.177, pp.1-81, 2019.

[15] Y. Mizutani, D. Kawahara and S. Kurohashi, An attempt to estimate the difficulty of Japanese words, *Proc. of the 24th Annual Conference of the Association for Natural Language Processing*, 2018.