Interspeech 2024
1-5 September 2024, Kos, Greece

Utilization of Text Data for Response Timing Detection in Attentive Listening

Yu Watanabe[1], *Koichiro Ito*[1], *Shigeki Matsubara*[1,2]

[1]Graduate School of Informatics, Nagoya University, Japan
[2]Information Technology Center, Nagoya University, Japan

watanabe.yu.x3@s.mail.nagoya-u.ac.jp, ito.koichiro.z5@f.mail.nagoya-u.ac.jp,
matsubara.shigeki.z8@f.mail.nagoya-u.ac.jp

Abstract

Conversational agents are expected to play the role of listening to narratives instead of humans. To be recognized as narrative listeners, these agents are required to generate responses indicating attentive listening at appropriate times. However, narrative and response data are not always well accumulated to develop statistical models. One solution to this issue is to utilize other data besides response data. In this study, we utilize text data to train a response timing detection model inspired by the relationship between punctuation marks and attentive listening responses. Specifically, the model was trained on a punctuation insertion task using text data before being trained on the response data. A response timing detection experiment was conducted to evaluate the effect of utilizing text data in terms of the amount of response data. The results showed that the utilization of text data enhanced the performance of the response timing detection, especially with limited data.

Index Terms: attentive listening, punctuation marks

1. Introduction

In Japan, increasing single-person households, particularly among the elderly [1], has led to fewer opportunities for people to narrate. Narrating is a basic human need, and losing opportunities to narrate is a social problem. One solution to this issue is that conversational agents provide opportunities for people to narrate instead of humans. To be recognized as narrative listeners, these agents are required to show an attitude of **attentive listening**. An explicit means of achieving this is to respond to the other's narrative. In this paper, utterances in response to narratives with the purpose of showing attentive listening are referred to as **attentive listening responses**.

Attentive listening responses convey an attitude of listening to the narrative and indicate understanding to the speaker, and generating them at appropriate times can help motivate the speaker to narrate. However, generating them at inappropriate times can be counterproductive since it interrupts the narration [2]. To act as narrative listeners, conversational agents are required to be able to detect appropriate response timing.

To enable response timing detection, it is practical to acquire a statistical model from narrative and response data. However, such of data to train a response timing detection model are not always well accumulated. One solution to this problem is to utilize other data besides the response data to train the model. A previous study proposed that the method involves utilizing punctuation in text data and sentiment polarity [3]. However, it is unclear how punctuation marks relate to attentive listening responses and how much the amount of response data affects the utilization effect.

In this study, we investigate the relationship between the punctuation marks and responses, and evaluate the effect of utilizing punctuation marks in terms of the response data volume. Previous studies have shown that, for spoken language, both responses and punctuation marks tend to be generated and inserted after a pause in the narrative [4, 5, 6]. In a preliminary investigation, we compared the occurrence position of responses and the insertion position of punctuation marks in narratives. The investigation showed that they exhibited certain similarities.

According to our preliminary investigation, our model was trained on a punctuation insertion task using text data before training on the response timing detection task using response data. In other words, a punctuation insertion task was introduced as an intermediate task to the response timing detection task, which is the target task. By training on the punctuation insertion task, valuable features for response timing detection are expected to be acquired.

A response timing detection experiment was conducted to evaluate the effect of utilizing text data in terms of the amount of response data used for training. The experimental results showed that training on the punctuation insertion task enhanced the performance on the response timing detection task, especially with limited response data.

2. Attentive Listening Responses and Punctuation Marks

2.1. Attentive Listening Responses

Attentive listening responses are utterances that convey an attitude of listening to the narrative and indicate understanding to the speaker. Figure 1 shows examples of both narrative utterances and attentive listening responses. If the listener generates attentive listening responses at appropriate times, the speaker's motivation to narrate can be enhanced. In contrast, generating attentive listening responses at inappropriate times interrupts the speaker's narrative, which can be counterproductive [2].

For generation timing detection of backchannels (BC), a typical example of attentive listening responses, several methods have been proposed, e.g., rule-based methods [7, 8] and methods that employ n-gram models [9], finite state transducers [10], decision trees [11, 12], CRFs [13], SVMs [14, 15], and LSTMs [16]. Recently, a method that employs pre-trained models [17] has also been proposed.

2.2. Relationship between Attentive Listening Responses and Punctuation Marks

To avoid interrupting the speaker's narration, attentive listening responses are likely to be generated within or after a pause in utterances. In fact, it has been stated that pauses in a narrative

10.21437/Interspeech.2024-1892

Narrative utterance		Attentive listening response	
Japanese	English translation	Japanese	English translation
20 代の後半に 仲良し 3 人組で 安い宿を使って 軽井沢に 行こうと	In my late twenties with two of my close friends using inexpensive lodging to Karuizawa I was going	ええ 軽井沢に	Uh-huh To Karuizawa

Figure 1: *Example of narrative utterances and attentive listening responses. The narrative in English is "In my late twenties, I was going to Karuizawa with two of my close friends using inexpensive lodging." The "ええ（Uh-huh)" is a short response that indicates successful listening. The "軽井沢に（to Karuizawa）" is a response that repeats part of the words in the narrative.*

Narrative (translation)	Response	Punctuation
今まで (Until now)		
憧れてた (I had longed)	✓	
ことを (things)	✓	
実際に (actually)		
見たり (see)	✓	✓
聞いたり (and hear)	✓	✓

Figure 2: *Example of responses and punctuation marks assigned to the narrative. The narrative in English is "Until now, I actually see and hear things that I had longed."*

trigger the generation of BCs [4].

In addition, the position of punctuation marks in transcriptions of spoken language has been analyzed. A study that attempted to use punctuation marks to estimate speech translation units [5] found that a large proportion of punctuation marks occur immediately after a pause. A study attempting to insert punctuation marks into the transcriptions of speeches [6] has pointed out that the occurrence of pauses provides an effective clue for the prediction of punctuation mark positions.

As mentioned above, previous studies have noted that attentive listening responses and punctuation marks occur with high probability after a pause in the narrative.

2.3. Comparison of Response and Punctuation Positions

This section quantitatively evaluates the degree of match between the occurrence timing of responses and the insertion positions of punctuation marks.

The responses of one listener in the Attentive Listening Response Corpus [18] were used in this evaluation. The corpus contains responses generated in reply to the Japanese Elder's Language Index Corpus (JELiCo) [19]. JELiCo contains approximately 8 hours and 40 minutes of narratives by 30 elderly people. The narratives in JELiCo were divided into *bunsetsus*[1] by the Japanese dependency structure analyzer CaboCha [20]. Each response was mapped to a bunsetsu boundary closest to the responsive utterance start time. Punctuation marks were inserted into the narrative using ChatGPT[2] because there was no punctuation in the narrative data. Figure 2 shows examples of the responses and punctuation marks assigned to the narrative.

We then compared the response timing and the punctuation mark position. Figure 3 shows a Venn diagram illustrating the

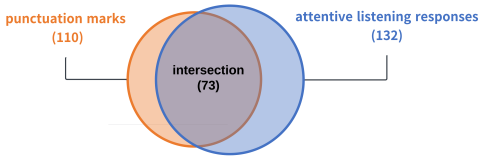


Figure 3: *The degree of match the response timings and positions of punctuation marks.*

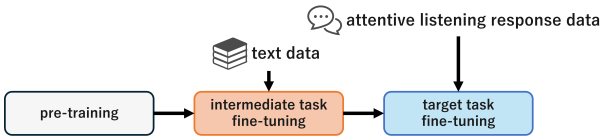


Figure 4: *The pipeline of intermediate-task transfer learning.*

degree of match between the response timing and the punctuation mark position. The Simpson coefficient was 0.664. The score shows a certain degree of similarity between the response timing and the punctuation position.

3. Model Training

Based on our analysis in Section 2, we utilized text data to train a response timing detection model.

3.1. Intermediate-task Transfer Learning

The intermediate-task transfer learning is a technique that first fine-tunes the pre-trained language models on an intermediate task before training the models on the target task [21, 22]. Figure 4 shows the pipeline of intermediate-task transfer learning in this study. Intermediate-task transfer learning can improve the performance of the models on the target task. Regarding the intermediate task, it is necessary to select a task that is relevant to the target task [23, 24]. Based on the similarity between the responses and punctuation marks in Section 2, we introduce a punctuation insertion task into the text as an intermediate task for the response timing detection task.

3.2. Intermediate Task

The intermediate task is a punctuation insertion task. Figure 5 shows an example of a punctuation insertion task into text. The model classified each bunsetsu boundary in the text into one of three classes, i.e., "Comma," "Period," or placement without punctuation marks "None." The input for the classification is the character strings before the bunsetsu boundary to be classified.

[1] Bunsetsu is a linguistic unit in Japanese that roughly corresponds to a basic phrase in English. A bunsetsu comprises a single independent word and zero or more ancillary words.

[2] https://openai.com/blog/chatgpt

Table 1: *Class distribution on the Kyoto University Text Corpus and the Attentive Listening Response Corpus.*

	Kyoto University Text Corpus			Attentive Listening Response Corpus		
	Comma	Period	None	BC	*Not-BC*	None
training	22,786	16,731	123,846	5,227	3,324	8,866
development	1,648	1,172	8,540	1,703	1,215	3,352
test	2,444	1,836	13,014	1,657	1,152	3,350

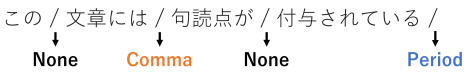


Figure 5: *Example of punctuation insertion task. "/" is a bunsetsu boundary. This Japanese sentence means "Punctuation is given in this sentence."*

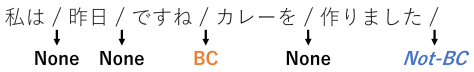


Figure 6: *Example of response timing detection task. "/" is a bunsetsu boundary. This Japanese sentence means "I made a curry yesterday."*

3.3. Target Task

The target task is a response timing detection task. Figure 6 shows an example of a response timing detection task. In this task, the model classifies whether or not each bunsetsu boundary of a narrative is a response timing.

There are response types including BC, admiration, echoic response, and others [25]. BCs are the most typical response and are short feedback to the speaker. The occurrence frequency of BCs is high because they are less dependent on the content of the speaker's utterance and can be generated more generically than other responses. In contrast, many of responses other than BCs (*Not-BC*s) depend on the content of the speaker's utterance. Thus, the generation frequency is less, and the generation timing is also considered to differ from that of BCs. Based on this difference in generation timing, each bunsetsu boundary is classified into one of three classes, i.e., "BC" timing, "*Not-BC*" timing, or unsuitable response generation timing ("None" timing). The input for the classification is character strings of the narrative before the boundary to be classified.

4. Experiment

In this experiment, we investigated the effectiveness of text data utilization mainly in terms of the amount of response data.

4.1. Experimental Overview

We evaluated the effect of introducing the intermediate task in terms of the amount of training data for the target task. The training data for the target task was set to a minimum of 200 and a maximum of 16,000, with the data increasing by 200 up to 2,000 and by 2,000 after that. The baseline is the model fine-tuned on only the target task. The intermediate-task model is the model fine-tuned on both tasks. Training and evaluation were conducted three times with different seeds, and the average of three macro-F_1 was used as the final evaluation value.

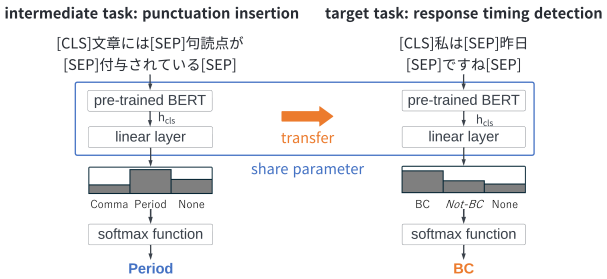


Figure 7: *Model overview. Examples of input sentences are the same as those in Figures 5 and 6.*

4.2. Experimental Data

In the intermediate task, we used the Kyoto University Text Corpus [26] as text data. This corpus contains articles from Mainichi Shimbun, a Japanese newspaper, annotated with various linguistic information. In this experiment, article data from January 1 to 17 were used. Each sentence was divided into bunsetsus by referencing the annotations in the corpus. Then, punctuation marks were removed from the text. We defined a bunsetsu boundary with a comma as "Comma," a bunsetsu boundary with a period as "Period," and a bunsetsu boundary without punctuation marks as "None." This text corpus was divided into training, development, and test sets based on the split for predicate-argument structures analysis. Each set contained 163,363, 11,360, and 17,294 bunsetsu boundaries. Table 1 shows the data size classified according to the presence or absence of punctuation marks.

In the target task, we used the Attentive Listening Response Corpus [18]. In this experiment, the manual transcriptions of the speaker's narrative and the response data of one listener were used. Response timing was classified according to the response type corresponding to a bunsetsu boundary. We defined a bunsetsu boundary with only BC as "BC" timing, a bunsetsu boundary with one or more responses other than BC as "*Not-BC*" timing, and a bunsetsu boundary with no responses as "None" timing. The corpus was divided into training, development, and test sets at a ratio of 3:1:1. Each set contained 17,417, 6,270, and 6,159 bunsetsu boundaries. Table 1 shows the data size for each type of response timing.

4.3. Implementation

In this experiment, a three-class classification layer was added to the pre-trained BERT[3] [27] to perform punctuation insertion and response timing detection. Figure 7 shows a model overview. The input to BERT consisted of a [CLS] token at the beginning and the strings of the three bunsetsus immediately be-

[3] https://huggingface.co/cl-tohoku/bert-base-japanese-v3

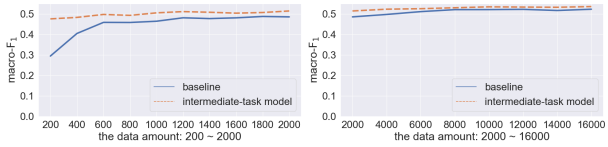


Figure 8: *Performance of the target task (macro-F_1).*

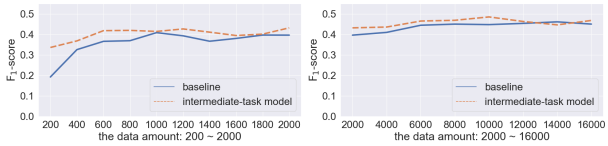


Figure 9: *Performance of BC timing detection.*

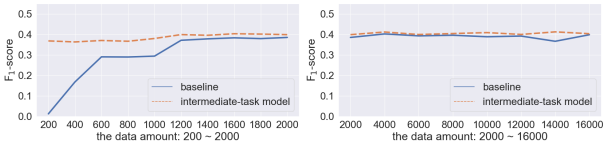


Figure 10: *Performance of Not-BC timing detection.*

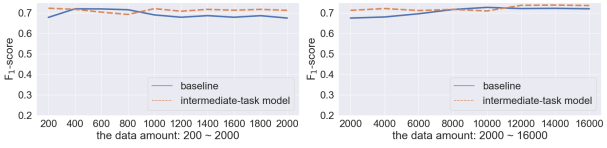


Figure 11: *Performance of None timing detection.*

fore the bunsetsu boundary to be classified. A [SEP] token was inserted between each bunsetsu. The outputs in both tasks corresponded to "Comma" and "BC" timing, "Period" and "*Not-BC*" timing, and "None" and "None" timing. The loss function was Cross Entropy Loss, and the optimizer was AdamW [28]. In the intermediate task and the target task, the maximum training epochs were 30 and 20, and the learning rates were 1e-5 and 5e-6, respectively. The gradient accumulation steps were 4 and 1, and batch sizes were 32 and 16. The patience for early stopping was set to 5 for both tasks. Training was performed using four GPUs on the intermediate task and one GPU on the target task. These hyperparameters and values were determined by the performance of each task on the development data.

Training was continued on the target task for the model of the epoch with the best macro-F_1 in the development data of the intermediate task. In other words, the training of the target task continued to update all parameters, including the classification layer, using the parameters learned in the intermediate task as initial values. The model of the epoch with the best macro-F_1 on the development data was used to evaluate the performance of the target task.

4.4. Experimental Results

We report the result for the target task[4]. Figure 8 shows the performance of the target task. The intermediate-task model outperformed the baseline across all amounts of training data[5]. This result indicates that training in the punctuation insertion task was effective in improving the performance of the response timing detection task. In addition, the improvement over the baseline tended to be significant with a small amount of training data. This result indicates that training on the punctuation insertion task was especially effective with limited response data. Thus, we conclude that utilizing the text data enabled the model to learn response timing features efficiently.

4.5. Discussion

This section discusses the effect of text data utilization on each class of the target task. Figure 9 shows the BC timing detection performance. The intermediate-task model outperformed the baseline up to approximately 12,000 training data and even fell below the baseline after that. These results indicate that training on the punctuation insertion task had a positive effect on the BC with comparatively small data for the target task, and the effect was limited with comparatively large data.

Figure 10 shows the *Not-BC* timing detection performance. The intermediate-task model significantly outperformed the baseline up to approximately 1,000 training data. Although the difference became smaller after 1,000, the performance improvement was observed. From these results, there was a difference in the range of the data amount where the intermediate task was effective between the *Not-BC* and BC. In addition, when the punctuation insertion task had a positive effect, the improvement over the baseline tended to be greater than that of the BC. Thus, the effectiveness of utilizing punctuation marks differs between the BC and *Not-BC* classes.

Figure 11 shows the None timing detection performance. In contrast to the results for the BC and *Not-BC* classes, no significant difference was observed between both models across all the amount of training data. This result indicates that the punctuation insertion task improved the performance of both the BC and *Not-BC* maintaining the performance of the None class.

5. Conclusion and Limitations

This paper described the utilization of text data for response timing detection in attentive listening. Specifically, we used punctuation marks for response timing detection based on our analysis that the responses and punctuation marks were similar. A response timing detection experiment was conducted to investigate its effectiveness. Experimental results showed that the utilization of text data enabled the model to efficiently learn the features of response timing.

In this study, we used data in Japanese, for which response data has been relatively well accumulated. The findings of this study may be particularly effective in languages where the accumulation is not advanced. In the future, we would like to consider adaptation to such languages.

There are some limitations of this study. First, only newspaper article data were used as text data. It is also possible to use chat data from SNS and subtitle data from lectures as text data. Second, a punctuation insertion task was introduced only as an intermediate task. This task also can be introduced as a subtask in multi-task learning.

[4]The macro-F_1 for the test set on the intermediate task was 0.800. From this result, we confirmed that the model could learn the features of punctuation insertion positions appropriately.

[5]A 3×3 McNemar-Bowker test ($\alpha = 0.05$) was conducted, applying Bonferroni correction. The statistically significant differences were observed three times with all amounts of data.

6. Acknowledgements

The narrative corpus was provided by the Social Computing Laboratory of Nara Institute of Science and Technology, Japan. The computation was carried out using the General Projects on supercomputer "Flow" at Information Technology Center, Nagoya University.

7. References

[1] Ministry of Health, Labor and Welfare, "Summary report of comprehensive survey of living conditions 2019," 2020, https://www.mhlw.go.jp/english/database/db-hss/dl/report_gaikyo_2019.pdf.

[2] Y. Ohya, *Taijin Kyugo no Genba de Tsukaeru Kiku Tsutaeru Kyokansuru Gijutsu Benricho*. Shoeisha, 2017, (in Japanese).

[3] R. Cho, K. Ochi, K. Inoue, and T. Kawahara, "Prediction of backchannels in spoken dialogue by pre-training with large-scale text data," in *Proceedings of the 84th National Convention of IPSJ*, 2022, pp. 305–306, (in Japanese).

[4] S. K. Maynard, *Japanese conversation : self-contextualization through structure and interactional management*, ser. Advances in discourse processes, 1989.

[5] T. Shimizu, S. Nakamura, and T. Kawahara, "Effect of punctuation marks for speech translation unit boundary detection," *IPSJ SIG Technical Reports*, vol. 108, no. 338, SP2008-100, pp. 127–131, 2008, (in Japanese).

[6] Y. Akita and T. Kawahara, "Automatic comma insertion of lecture transcripts based on multiple annotations," in *Proceedings of the 12th Annual Conference of the International Speech Communication Association (INTERSPEECH-2011)*, 2011, pp. 2889–2892.

[7] N. Ward and W. Tsukahara, "Prosodic features which cue back-channel responses in English and Japanese," *Journal of Pragmatics*, vol. 32, no. 8, pp. 1177–1207, 2000.

[8] R. Poppe, K. P. Truong, D. Reidsma, and D. Heylen, "Backchannel strategies for artificial listeners," in *Proceedings of the 10th International Conference on Intelligent Virtual Agents (IVA-2010)*. Springer, 2010, pp. 146–158.

[9] N. Cathcart, J. Carletta, and E. Klein, "A shallow model of backchannel continuers in spoken dialogue," in *Proceedings of the 10th Conference on European Chapter of the Association for Computational Linguistics (EACL-2003)*, vol. 1, 2003, pp. 51–58.

[10] S. Fujie, K. Fukushima, and T. Kobayashi, "A conversation robot with back-channel feedback function based on linguistic and non-linguistic information," in *Proceedings of the 2nd International Conference on Autonomous Robots and Agents (ICARA-2004)*, 2004, pp. 379–384.

[11] H. Noguchi and Y. Den, "Prosody-based detection of the context of backchannel responses," in *Proceedings of the 5th International Conference on Spoken Language Processing (ICSLP-1998)*, 1998, pp. 487–490.

[12] N. Kitaoka, M. Takeuchi, R. Nishimura, and S. Nakagawa, "Response timing detection using prosodic and linguistic information for human-friendly spoken dialog systems," *Transactions of the Japanese Society for Artificial Intelligence*, vol. 20, no. 3, pp. 220–228, 2005.

[13] L.-P. Morency, I. de Kok, and J. Gratch, "A probabilistic multimodal approach for predicting listener backchannels," *Journal of Autonomous Agents and Multi-Agent Systems*, vol. 20, no. 1, pp. 70–84, 2010.

[14] T. Yamaguchi, K. Inoue, K. Yoshino, K. Takanashi, N. G. Ward, and T. Kawahara, "Analysis and prediction of morphological patterns of backchannels for attentive listening agents," in *Proceedings of the 7th International Workshop on Spoken Dialogue Systems (IWSDS-2016)*, 2016, pp. 1–12.

[15] Y. Kamiya, T. Ohno, S. Matsubara, and H. Kashioka, "Construction of back-channel utterance corpus for responsive spoken dialogue system development," in *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC-2010)*, 2010, pp. 2414–2419.

[16] R. Ruede, M. Müller, S. Stüker, and A. Waibel, "Enhancing backchannel prediction using word embeddings," in *Proceedings of the 18th Annual Conference of the International Speech Communication Association (INTERSPEECH-2017)*, 2017, pp. 879–883.

[17] J. Y. Jang, S. Kim, M. Jung, S. Shin, and G. Gweon, "BPM_MT: Enhanced backchannel prediction model using multi-task learning," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP-2021)*, 2021, pp. 3447–3452.

[18] K. Ito, M. Murata, T. Ohno, and S. Matsubara, "Construction of responsive utterance corpus for attentive listening response production," in *Proceedings of the 13th Language Resources and Evaluation Conference (LREC-2022)*, 2022, pp. 7244–7252.

[19] K. Aramaki, "Japanese elder's language index corpus v2," 2016, https://figshare.com/articles/dataset/Japanese_Elder_s_Language_Index_Corpus_v2/2082706/1.

[20] T. Kudo and Y. Matsumoto, "Japanese dependency analysis using cascaded chunking," in *Proceedings of the 6th Conference on Natural Language Learning (CoNLL-2002)*, 2002, pp. 63–69.

[21] J. Phang, T. Févry, and S. R. Bowman, "Sentence encoders on STILTs: Supplementary training on intermediate labeled-data tasks," *arXiv preprint arXive:1811.01088*, 2019.

[22] T.-Y. Chang and C.-J. Lu, "Rethinking why intermediate-task fine-tuning works," in *Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2021*, 2021, pp. 706–713.

[23] C. Poth, J. Pfeiffer, A. Rücklé, and I. Gurevych, "What to pre-train on? Efficient intermediate task selection," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP-2021)*, 2021, pp. 10 585–10 605.

[24] Y. Pruksachatkun, J. Phang, H. Liu, P. M. Htut, X. Zhang, R. Y. Pang, C. Vania, K. Kann, and S. R. Bowman, "Intermediate-task transfer learning with pretrained language models: When and why does it work?" in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL-2020)*, 2020, pp. 5231–5247.

[25] Nihongo Kijutsu Bunpo Kenkyukai, *Gendai Nihongo Bunpo 7*. Kuroshio Shuppan, 2009, (in Japanese).

[26] S. Kurohashi and M. Nagao, "Building a Japanese parsed corpus while improving the parsing system," in *Proceedings of the 1st International Conference on Language Resources and Evaluation (LREC-1998)*, 1998, pp. 719–724.

[27] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "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 (NAACL-2019)*, 2019, pp. 4171–4186.

[28] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proceedings of the 7th International Conference on Learning Representations (ICLR-2019)*, 2019.