BERT fine-tuning for Japanese Author Attribution using Stylometric Features

Abstract

Authorship Attribution (AA) is the task of identifying the author of a text. It marks a novel effort in applying deep learning techniques for AA in Japanese, a field where such approaches have been limited. Historically, AA studies in Japanese have predominantly employed Random Forests and SVMs, focusing on small author groups and facing a scarcity of datasets for author identification. Our work diverges by fine-tuning a pre-trained BERT model to assess its efficacy in both text-only scenarios and when incorporating Japanese-specific stylistic features. Key findings reveal an 84% accuracy rate for identifying five authors using only text data, and a notable 82% accuracy with an expanded set of 80 authors, highlighting the potential of deep learning for managing larger author pools. However, the addition of stylistic features for a set of 25 authors resulted in reduced accuracy (53%). The study further achieved 97% accuracy in distinguishing Japanese speakers and 61% in nationality prediction. These outcomes emphasize the viability of deep learning-based AA in Japanese, presenting a significant advancement in the domain.

1 Introduction

Author Attribution (AA) is to identify the author of given texts. The method of authorization is divided into three parts: literary methods, statistical methods, and machine learning. Machine learning methods are on the rise. The main topic of the machine learning method is the author's name identification of short message. Statistical methods use long texts for analysis and use for similarities (Zheng & Jin, 2023). The author's identification was also used in criminal investigations, and in Japan, Jin et al. has analyzed documents related to the actual case (Zaitsu & Jin, 2015).

As far as I know, in the field of AA in Japanese, there are numerous studies that utilize statistical methods rather than machine learning techniques. There are issues with the corpora used in research on Japanese author identification, such as not being open source or not being updated. For example, in (MinZhe, 2013), compositions from 11 university students at a certain private university in a local area are used, and since there is no mention of how they were obtained, it is considered not to be open source. Moreover, in (Huang & Jin, 2020), 20 works from 20 Japanese authors were downloaded from Aozora Bunko and used for analysis. However, as of July 31, 2023, the works that have been published in Aozora Bunko and whose copyrights have expired were all by authors who died before 1967, which means that most of the works used old character forms and historical kana orthography (Liu & Jin, 2023).

In this paper, we conducted fine-tuning for AA targeting Japanese texts using BERT. Experiments were conducted under two conditions: using only text data, and using text data with added stylometric features of Japanese. These experiments were performed for various numbers of authors, including 5, 10, 25, 50, 75, 100, and for all authors. Additionally, since the corpus used in this study included Japanese learners, experiments were conducted separately for Japanese native speakers and Japanese learners. Other experiments included classifying whether a speaker is native Japanese and predicting nationality. To the best of our knowledge, this is the first attempt at AA using a deep learning-based approach targeting Japanese.

The next section discusses related approaches. Section 3 presents the model and architecture used, along with a description of the stylistic features of Japanese. Section 4 provides a detailed explanation of the experiments, and Section 5 discusses the factors behind the success and failure of these experiments, as well as future work directions. Finally, Section 7 concludes the paper.

2 Related Work

2.1 Author Attribution except than Japanese

In Western languages such as English and French, the most typical stylometry is used by characters and words. Sharma et al. classified short online texts obtained from the WhatsApp messaging application for features such as word n-grams and character n-grams using Nave Bayes, SVM, conditional tree, and random forest algorithms. The results showed that SVM achieved up to 95% (Sharma, et al., 2018). In addition, Wen et al. Propose the new ensemble model that uses the translation embedded method to predict the author relationship between the author and online news, and the best result is the accuracy. Showed that 93% achieved (Wen, et al., 2020).

2.2 Japanese Authorship Attribution

A typical language difference between Japanese and English is that there is no space between words and the types of characters used are different. Since there is no space between words, the way to the token is divided depending on the tool used. One of the methods of the author's identification is that the author who has a similar meaning but uses different expressions is sometimes distinguished. In the case of Japanese, in addition to the same meaning and reading, it is one way to determine the writing style of the author, such as using kanji or using hiragana or katakana.

The length of the sentence is often used as a characteristic of the writer. The reason is that, especially in languages that are not divided into words, such as Chinese and Japanese, are easy to calculate. However, the length of the sentence length can be used as a descriptor of the writer's style, but in Japanese, the length of the sentence is not always a powerful descriptor (Zheng & Jin, 2023).

Functional words (prepositions, particle, limited lyrics, adverbs, etc.) are characterized by functioning as conjunctions of other words, providing clues to grammar structure, and being relatively stable. These clues are independent of the topic, capturing the writer's purely unconscious style. In addition, due to the high incidence of particles, it is effective even in sentences with a small number of characters such as diary (about 500 characters) and essay (about 1000 characters) (Zheng & Jin, 2023).

(Zaitsu & Jin, 2018) used the four stylistic features, Usage rate of non-independent words, bigrams of parts of speech, bigrams of particles, how to use commas (words). These stylistic features are the advantages of not depending on the content, in addition to the high identification features (Zaitsu & Jin, 2018).

(Zaitsu & Jin, 2023) used Random Forest for classifying human written text and AI generated text. They used about 1000 characters for 36 people, the author of the Japanese psychology thesis, and about 1000 characters for 144 texts. This study uses a Japanese stylometry similar to (Zaitsu & Jin, 2018) for classification. As a result, the maximum value of the performance level (accuracy, reproduction rate, accuracy, F1 score) was 100%.

(Liu & Jin, 2023) proposed a novel feature, Nucleus Bunsetsu (NBS), by decomposing sentences into phrase units according to their dependency structure and then expanding them into a tree-like structure. They defined the root phrase and the phrases directly connected to it as nucleus phrases and proposed the patterns extracted from these segments as the new feature NBS. Using works from 10 novelists to construct a corpus, they demonstrated the effectiveness of their method through binary and ten-group discrimination simulations. In binary discrimination, the performance of NBS closely matched that of the comparative phrase pattern Type B, and in ten-group discrimination, it showed a significant advantage with a 2 point difference in accuracy.

(Sun & Jin, 2018) focused on phonemes, the smallest unit of sound, as a preliminary step towards utilizing phonological features for stylistic analysis. They clarified the position of phoneme information in Japanese native speakers AA through comparison with existing stylistic features, including punctuation habits, morphological tag bi-grams, and phrase pattern features. Although phonemes did not perform as well as the comparison features, in a four-group classification of four authors, they achieved accuracies of 0.84 using SVM and 0.85 using RF, respectively.

(Huang, et al., 2018) focused on sentence structure and proposed sentence patterns, considered as a part of the sentence's features, for AA. They validated its effectiveness using 400 novels (20×20) from 20 authors. Both RF and SVM showed the highest accuracy with the

combination of phrase patterns + sentence patterns, achieving 99.20% and 97.89%, respectively.

3 Method

3.1 Fine-tuning BERT

We utilized the Japanese pre-trained BERT model available from the Transformer library, named "cl-tohoku/bert-large-japanese-v2" [1]. Fine-tuning is straightforward owing to the self-attention mechanism in the Transformer, which allows BERT to model various downstream tasks—regardless of whether they involve a single text or pairs of texts—by simply swapping the appropriate inputs and outputs (Devlin, et al., 2018).

Figure 1 illustrates the architecture for fine-tuning with only text data, while Figure 2 shows the architecture when stylistic features are added. The first architecture is quite simple, where text data extracted from the training dataset is fed into the pre-trained Japanese BERT model. This model encodes the text data and extracts contextual features, which are then used to output class probabilities.

The second architecture integrates a BERT model with custom features. Initially, text data and four types of custom features are extracted from the training dataset. The text data is fed into the BERT model, and the additional features are encoded separately before being combined with the output of the BERT model. These combined features are then supplied to a custom-defined class named "Bert with Custom Features Class," which outputs the final class probabilities. This approach enables the model to make classifications considering not only the information obtained from the text data but also the linguistic characteristics of Japanese.

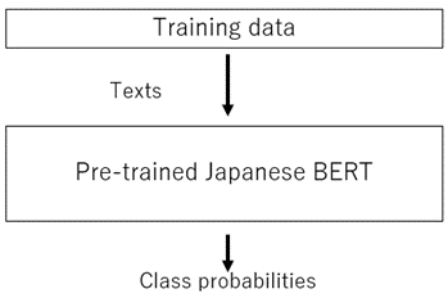

Figure1: BertAA with only texts architecture.

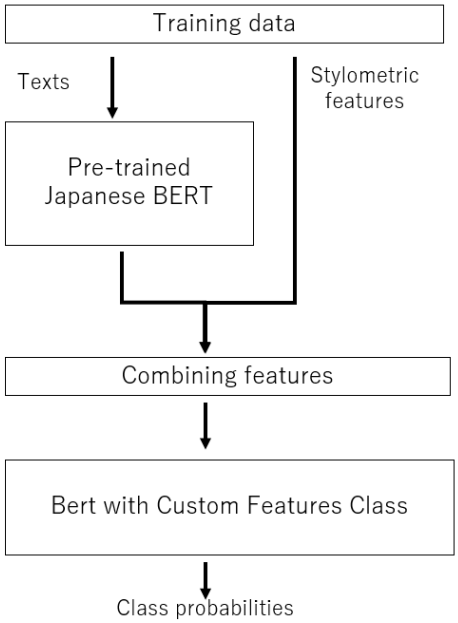

Figure2: BertAA + stylometric features architecture.

3.2 Stylometric features

In this study, we utilized four stylistic features of Japanese as proposed by (Zaitsu & Jin, 2018): the usage rate of content words, part-of-speech bigrams, particle bigrams, and the manner of comma placement (preceding word). The usage rate of function words was analyzed based on the frequency of occurrence of words such as adverbs, auxiliary verbs, particles, verbs (excluding content words), nouns (excluding content words, but including pronouns), adjectives (excluding content words), attributive words, interjections, symbols, conjunctions, and prefixes. Particle bigrams focused on pairs of particle words, and the manner of comma placement extracted the word immediately preceding the comma. All stylistic features were pre-extracted using GiNZA (Matsuda, 2020) and incorporated into the dataset.

4 Experiments

4.1 Dataset

In this study, we utilized the "Composition Bilingual Database" published by the National Institute for Japanese Language and Linguistics[2]. This database comprises four components: compositions in Japanese written by learners of Japanese, translations of these compositions into the authors' native languages by the authors themselves, corrections of the compositions by

[1] https://huggingface.co/tohoku-nlp/bert-large-japanese-v2

[2] https://mmsrv.ninjal.ac.jp/essay/essay_05.html

Japanese language teachers or similar experts, and information on the linguistic backgrounds of the composition authors and correctors. For our research, we only used the Japanese composition data and information about the authors. Graph 3 shows the nationalities of all the authors, with 83 being Japanese. All composition data were converted into line units, totaling 34,187 lines. The average number of texts per author is 16, and after processing noise, the average number of tokens per line in each text is 20.

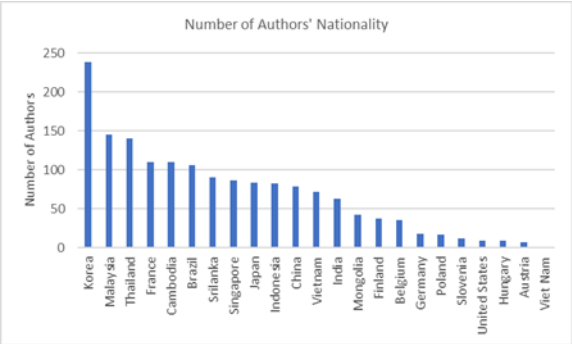

Figure 3: Breakdown of authors' nationality of the composition translation database

4.2 Parameter Settings

For the implementation of our model, we utilized Python 3.10.12. All experiments were conducted on Google Colab, with CUDA Version 12.2. Our training process included the use of a single GPU, a Tesla T4. We trained the model for either 3 or 5 epochs. The dataset was divided into training and testing sets with a split ratio of 8:2. The maximum length for tokenization was set to 512. The random selection of authors was controlled by setting the random seed to 42. Other parameters were consistent with those used for a BERT base model.

4.3 Baselines and Results

In this study, we included texts from Japanese learners in addition to those from native Japanese speakers, hence we performed fine-tuning separately for Japanese natives and non-natives. Initially, fine-tuning was conducted using only text data, followed by fine-tuning using a combination of text data and stylistic features. Additionally, we conducted training to classify whether a given text was written by a Japanese native speaker.

We based our experiments on the BertAA and +Style models from the Enron dataset by (Fabien, et al., 2020). Experiments were conducted with datasets for Japanese native speakers and Japanese learners, respectively, across different numbers of authors. The results are presented in Table 1, Graph 4, and Graph 5. The epoch was set to 5. Authors were selected based on the highest number of texts per author for each dataset.

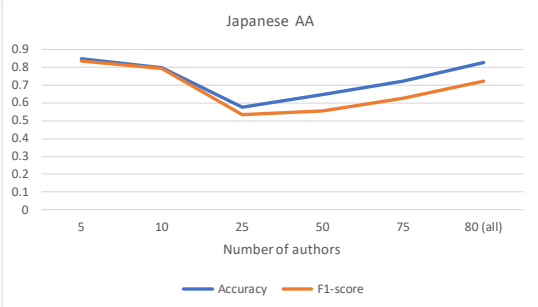

Figure 4: AA for Japanese native speakers when fine-tuning using text data

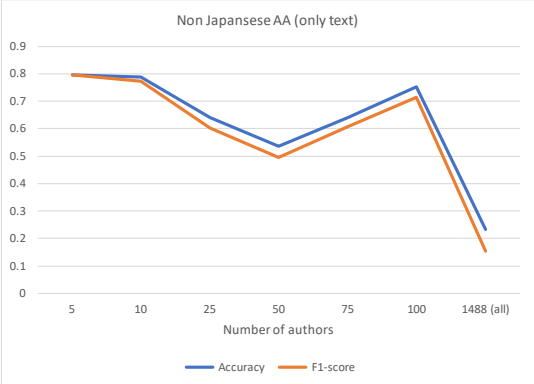

Figure 5: AA for Japanese leaners when fine-tuning using text data

For the database of Japanese native speakers, since the maximum number of authors after removing noise was 80, experiments were conducted with 5, 10, 25, 50, 75, and 80 authors. In contrast, for the database of Japanese learners, where the maximum number of authors after noise removal was 1488, experiments were also conducted with 100 and 1488 authors. These results are shown in Table 2, Graphs 4 and 5. A trend was observed where fewer authors resulted in higher accuracy, surprisingly, this also applied when experimenting with all Japanese authors. As shown in Table 2, while the highest accuracy reported by (Fabien, et al., 2020) was 99.95% for 5 authors, our best result was 84.6% for a similar number of Japanese speakers only. The accuracy for the database of Japanese learners was generally

Nation of people	Number of authors	Baseline Fabien et al.	Token Count Mean	Token Count Max	Token Count Min	Texts per Author Mean	Texts per Author Max	Texts per Author Min	Accuracy	F1-score
Japan	**5**	99.95	21.4	77	6	26.0	33	24	**84.6**	83.3
	10	99.1	22.2	77	6	24.4	33	22	79.6	79.2
	25	98.7	23.5	77	6	21.4	33	18	57.4	53.4
	50	98.1	25.9	109	6	18.7	33	15	64.7	55.6
	75	97.6	26.9	109	6	16.9	33	12	72.3	62.4
	80 (all)	-	27.1	109	6	16.5	33	10	**82.6**	72.1
Non Japan	**5**	99.95	20.5	61	6	59.2	66	54	**79.7**	79.5
	10	99.1	20.7	71	6	54.4	62	48	78.8	77.3
	25	98.7	19.8	71	6	49.12	68	38	64.2	60.2
	50	98.1	20.2	71	6	44.48	63	35	53.7	49.5
	75	97.6	20.4	74	6	41.57	68	28	64.1	60.7
	100	97.0	20.2	68	6	39.26	64	25	75.3	71.5
	1488 (all)	-	20.9	132	6	17.55	66	4	23.2	15.3

Table 1: Fine-tuning results for text data and statistical information on author selection

Nation of people	Number of authors	Baseline Fabien et al.	Token Count Mean	Token Count Max	Token Count Min	Texts per Author Mean	Texts per Author Max	Texts per Author Min	Accuracy	F1-score
Japan	**5**	99.95	23.5	40	15	18.3	25	16	18.2	12.1
	10	99.1	24.1	41	15	16.2	25	14	32.1	21.6
	25	98.7	24.6	41	15	11.9	25	6	**53.2**	46.5
	50	98.2	25.4	41	15	12.3	25	7	33.9	31.4
	75	97.5	25.8	41	15	14.0	25	11	30.5	26.0
	80 (all)	-	25.8	41	15	12.3	25	7	36.4	28.1
Non Japan	**5**	99.95	18.9	31	11	61.4	68	55	0.0	0.0
	10	99.1	19.2	31	11	56.1	68	50	8.0	1.8
	25	98.7	19.2	31	11	50.2	68	43	3.4	1.5
	50	98.2	19.4	31	11	45.0	68	38	0.0	0.0
	75	97.5	19.5	31	11	41.8	68	33	0.0	0.0
	100	97.0	19.4	31	11	39.1	68	30	1.0	0.9
	1488 (all)	-	19.2	31	11	16.5	68	5	0.1	0.0

Table 2: Fine-tuning results of text data and stylometric features, and statistical information of author selection

lower than that for the database of Japanese native speakers. Comparing the number of texts per author across different author counts, the database of Japanese learners showed greater variability and a larger change in median than the database of Japanese speakers.

Results of fine-tuning with added stylistic features to text data are presented in Graphs 4, 5, and Table 2. Both databases showed drastically low accuracies, indicating the experiments were unsuccessful. Variability in the number of texts per author was observed in the database of Japanese learners. While (Fabien, et al., 2020) showed high accuracy across all author counts, our dataset did not maintain stable accuracy. Furthermore, while (Fabien, et al., 2020) achieved the highest accuracy with 5 authors, our experiments yielded the highest accuracy with 25 Japanese speakers.

	Accuracy	F1 Score
Japanese or not	97.0	80.6
Which country	61.8	52.0

Table 3: Additional experimental results

Additionally, two more experiments were conducted, both utilizing only text as input. The results are shown in Table 3. The first experiment involved classifying whether an individual is a Japanese native speaker, setting the epoch to 3 and fine-tuning the pre-trained BERT model, which resulted in an accuracy of 97% and an F1 Score of

80%. The second experiment, set with an epoch of five, aimed to classify the nationality of authors, resulting in an accuracy of 61% and an F1 Score of 52%.

5 Discussion

When using only text data, we observed a tendency for accuracy to increase with a smaller number of authors, a phenomenon also reported by (Fabien, et al., 2020), where an increase in the number of authors tends to lower the accuracy. The highest accuracy achieved in our experiment, which classified whether a speaker was Japanese, can be attributed to it being a binary classification. One reason for the lower fine-tuning accuracy for texts by Japanese learners could be the numerous grammatical and lexical errors present in their compositions, which introduced noise and potentially hindered accurate classification. The bias in the number of texts per author and the limited amount of text data per author in the Japanese corpus used in this study are also considered factors that prevented accuracy from exceeding 90%.

(Zheng & Jin, 2023) states, "A model with high bias fails to learn from the training and testing data due to being too simple, leading to significant errors on both. A low-variance model overfits the training data and does not generalize to the testing data, resulting in low error rates on training data but high error rates on testing data. In supervised learning, underfitting occurs when the model cannot capture the underlying patterns of the data. Such models, usually due to insufficient training data or an inappropriate classifier for the data structure, exhibit high bias and low variance." This explanation aligns with the phenomena observed in our study, especially the part about "low error rates on training data but high error rates on testing data". Therefore, considering the bias-variance trade-off and determining how the learning model operates is essential for future work.

Directly fine-tuning BERT with stylometric features might have been one reason for the low accuracy. When conducting experiments with stylometric features, we incrementally tested from one to four features. However, extremely low accuracy was consistently observed across all patterns, suggesting potential issues with the stylometric features themselves, as well as the methods of input and integration. Previous studies have classified authors using stylometric features with classifiers like random forests and SVMs. Although high accuracies were achieved using stylometric features with random forests and hierarchical cluster analysis, I have not found studies classifying Japanese stylometric features using a deep learning-based approach. Further research is needed on the selection and input methods of stylometric features for Japanese authorship attribution using deep learning, including model settings and feature selection.

The main difference between the two architectures—using only text and adding stylometric features—is the presence of additional features and the resulting complexity of the model. The first architecture uses only text data, providing a simple yet powerful BERT-based classification. In contrast, the second architecture can deepen the model's understanding by incorporating custom features, though this comes with added preprocessing and complexity in model definition. The architecture for authorship attribution heavily depends on the task requirements and the nature of the data. Therefore, further exploration into preprocessing stylometric features and defining the model is necessary.

6 Conclusion

In this study, we conducted fine-tuning of a pre-trained BERT model for Japanese AA. This represents one of the first attempts to analyze the performance of fine-tuning a domain-specific pre-trained language model for AA in Japanese. While stable and high accuracy was achieved using only text data, the addition of Japanese stylometric features resulted in failure across all author count patterns. Future efforts should focus on identifying Japanese stylometric features that are suitable for fine-tuning BERT. Additionally, although we used a Japanese pre-trained BERT model in this study, we will consider experimenting with other pre-trained Japanese models.

References

Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K., 2018. *Bert: Pre-training of deep bidirectional transformers for language understanding,* s.l.: arXiv preprint arXiv:1810.04805.

Fabien, M., Villatoro-Tello, E., Parida & Parida, S., 2020. *BertAA: BERT fine-tuning for Authorship*

Attribution, s.l.: Proceedings of the 17th International Conference on Natural Language Processing (ICON).

Huang, S. & Jin, M., 2020. *Author identification using function phrases as stylometric features (in Japanese),* s.l.: Information Society of Information Society 30.3: 390-400.

Huang, S., Liu, Y. & Jin, M., 2018. *CA12-5 Comparative analysis of the effectiveness of sentence pattern features in author identification,* s.l.: Japanese Behaviormetric Society Conference Abstracts 46. Japanese Behaviormetric Society.

Liu, Y. & Jin, M., 2023. *Proposal of stylistic features using core clause related information for author identification (in Japanese),* s.l.: Theory and application of data analysis 12.1 : 33-46.

Matsuda, H., 2020. *Practical Japanese analysis using GiNZA-Universal Dependencies (in Japanese),* s.l.: Natural Language Processing 27.3 : 695-701..

MinZhe, J., 2013. *Identification of the writer based on the sentence pattern,* s.l.: Activity measurement studies 40.1: 17-28.

Sharma, A., Ananya, N. & Reetika, R., 2018. *An investigation of supervised learning methods for authorship attribution in short hinglish texts using char & word n-grams,* s.l.: arXiv preprint arXiv:1812.10281.

Sun, H. & Jin, M., 2018. *CA12-4 Japanese author identification using phonemes as stylistic features (in Japanese),* s.l.: Japanese Behaviormetric Society Conference Abstracts 46. Japanese Behaviormetric Society.

Wen, W., Li1, Q. & Zhang, X., 2020. *Predicting Online News Authorship by an Authorship Embeddings Space Method,* s.l.: 2020 5th IEEE International Conference on Big Data Analytics (ICBDA). IEEE.

Zaitsu, W. & Jin, M., 2015. *Author identification of documents related to crime using text mining (in Japanese),* s.l.: Journal of the Japanese Society of Forensic Science and Technology 20.1: 1-14.

Zaitsu, W. & Jin, M., 2018. *Standardization of the accuracy of the author's identification and judgment procedure by text mining (in Japanese),* s.l.: Activity measuring 45.1: 39-47.

Zaitsu, W. & Jin, M., 2023. *Distinguishing ChatGPT (-3.5,-4)-generated and human-written papers through Japanese stylometric analysis,* s.l.: arXiv preprint arXiv:2304.05534.

Zheng, W. & Jin, M., 2023. *A review on authorship attribution in text mining,* s.l.: Wiley Interdisciplinary Reviews: Computational Statistics 15.2: e1584.