Output Trend Analysis in Semantic Classification of Katakana Words Using a Large Language Model

Kazuki Kodaki and **Minoru Sasaki**
Ibaraki University / Japan
{24nm724g, minoru.sasaki.01}@vc.ibaraki.ac.jp

Abstract

In semantic classification of katakana words using a large language model (LLM), semantic divergences from the meanings of original English words such as Wasei-Eigo(Japanese-made English) may affect the accuracy of the model. In order to accurately capture the meaning of foreign words, we fine-tuned the LLM using data extracted from the BCCWJ(Balanced Corpus of Contemporary Written Japanese), analyzed the current accuracy and output trend of semantic classification for katakana words, and explored ways to improve the accuracy. The results of several experiments showed that fine-tuning was not effective for zero-shot learning, but in contrast, fine-tuning improved accuracy by about 10% for few-shot learning. Further analysis of the visualized data suggests trends related to words and meanings that the model struggles to classify correctly.

1 Introduction

Currently, research related to the Large Language Model (LLM) is being conducted in the field of natural language processing. LLM, such as those provided by ChatGPT, are trained primarily on English data, and the percentage of Japanese data seems to be relatively small. Japanese katakana words often include foreign words derived from English and Wasei-Eigo(Japanese-English words) with meanings unique to Japan. This makes semantic interpretation difficult for LLM, due to sense divergence and a lack of relevant training data. Therefore, this paper aims to improve the accuracy of the semantic classification task by adding Japanese data to existing LLM to understand the usage trends of katakana words and their senses. To solve this task, we constructed a dataset by extracting sentences containing katakana words from the Balanced Corpus of Contemporary Written Japanese (BCCWJ)(Maekawa et al., 2014) provided by the National Institute for Japanese Language and Linguistics (NINJAL), and we used and fine-tuned OpenAI's gpt-4o-mini-2024-07-18 model to analyze its output trends.

2 Related Work

Word Sense Disambiguation (WSD) research generally follows two main approaches: supervised learning and knowledge-based methods. Supervised methods classify ambiguous words using models trained on corpora annotated with correct senses by humans. Recent models use pre-trained language models, such as BERT (Devlin et al., 2019) and RoBERTa (Zhuang et al., 2021), to embed words and sentences for sense prediction (Maru et al., 2022; Blevins and Zettlemoyer, 2020). Knowledge-based methods rely on external resources such as dictionaries and ontologies instead of annotated corpora. Some approaches vectorize word definitions to learn sense relationships (Mizuki and Okazaki, 2023), while others use synonym relationships to derive meaningful sense vectors (Wang and Wang, 2020). Recent attempts have also been made to use LLM-based generative AI such as ChatGPT for WSD(Kocoń et al., 2023; Kang et al., 2024), and evaluation results show promising performance, but not yet up to the state-of-the-art models.

3 Experiment

3.1 Extraction-of-Foreign Words

For small-scale experiments, the dataset was constructed using only the BCCWJ sections PB, PM, PN, OC, OW and OY. BCCWJ was used because it is the only balanced corpus on Japanese. From these folders, sentences containing words whose short units (the smallest unit of semantic segmentation of a sentence or word) were labeled as foreign words were extracted, resulting in 32,226 sentences.

Proceedings of Recent Advances in Natural Language Processing,pages 567–571
Varna, Sep 8–10, 2025
https://doi.org/10.26615/978-954-452-098-4-066

From these, we considered four choice sentences and one question sentence to be the minimum number of occurrences, calculated the frequency of occurrence, and set the condition that there must be at least five sentences, and 10,750 sentences and 795 words were extracted.

3.2 Data Set Creation

3.2.1 Selecting target words

This section outlines the method used to select target words and sentences containing them in the data. Digital Daijisen, the source of the goo-dictionary [1] provided by NTT DOCOMO, was used as the word sense category for the target words. Considering the four choices and one question sentence as described in the previous section, we searched the 795 target words in the web-based Digital Daijisen and selected words with at least four examples of word use in the BCCWJ and at least five sentences. Non-katakana words such as foreign articles (e.g., la) were excluded. As a result, 40 words and 1,143 sentences were extracted.

3.2.2 Creation of data sets

To create the dataset, 1143 sentences containing 40 target words extracted in the previous section were used. Their meanings were manually labeled by first author, based on the semantic categories in the Digital Daijisen. Sentences were labeled as follows: 0 if the sentence consists only of the target word and the exact meaning cannot be determined, 1 if the target word is part of some other word, 2 if the context allows multiple interpretations, and 3 if the meaning of the word cannot be referenced in Digital Daijisen. 246 sentences labeled 0-3 were excluded from the dataset, leaving 897 usable sentences. This data is divided into a training and a test set, which are used for the 8 experiments described below. In experiments 1 and 7, the test set consisted of 120 randomly selected sentences (3 per each of the 40 target words), with the other 777 used for training. In experiments 2 through 6 and 8, the test data included 56 sentences, each representing a unique word sense in Digital Daijisen (i.e., appearing only once in BCCWJ), with the remaining 841 sentences used for training.

3.3 Fine-tuning and generating responses

The training data was fine-tuned using OpenAI API with the gpt-4o-mini-2024-07-18 model.

[1] https://dictionary.goo.ne.jp/

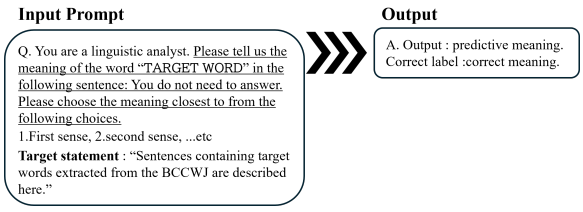

Figure 1: Used Prompts

All settings during the study, such as the learning rate and number of epochs, were done by "auto". For each experimental word sense prediction, responses were generated using ChatGPT before and after fine-tuning. The accuracy and output trends were evaluated by manually judging whether the predictions perfectly matched the semantic categories of the Digital Daijisen assigned to the test data. The prompt presented the semantic categories (scraped from the goo-dictionary) as answer choices and instructed the model to select the meaning closest to the target word. The actual prompt used is shown in the following figure1. Eight experiments were conducted by having responses generated as described above and changing the training and test data during fine-tuning. The details of each experiment and dataset are described in Sections 3.4.1 (Experiment 1), 3.4.2 (Experiments 2-6), and 3.4.3 (Experiments 7-8).

3.4 Experiments

3.4.1 Few-shot learning

The purpose of this experiment is to confirm the increase or decrease of output accuracy due to fine-tuning in few-shot learning. The test data in experiment 1 consisted of 120 sentences (40 target words x 3 sentences) randomly selected from 3 sentences from each target word, and the training data consisted of 777 sentences excluding them.

3.4.2 Zero-shot learning

These experiments aim to evaluate the effectiveness of fine-tuning in zero-shot learning. In each of the experiments in this section, the test data consisted of 56 sentences containing targets with a BCCWJ semantic frequency of 1, which was common to all. To investigate the effect of data bias by giving only examples other than the correct word sense annotated in the test data for the target word, the following experiments 2-6 were conducted. In Experiment 2, 841 training sentences were used, excluding the test data. In Experiment 3, 596 sentences were extracted from the 841 sentences of

training data, including only the target words in the test data. In Experiment 4, we randomly selected 298 sentences, half of the 596 sentences in Experiment 3, to test whether accuracy changes as the amount of data decreases. In Experiment 5, 245 sentences were used from the 841 sentences in the training data, excluding the 596 sentences in Experiment 3 (no sentences in the training data contained an target words). In Experiment 6, we extracted 122 sentences, half of the data from Experiment 5, with the same objectives as in Experiment 4.

3.4.3 Extractive Sense Comprehension and Chain-of-Thought

In Experiment 7, following ESC (Extractive Sense Comprehension)(Barba et al., 2021), target words in the training data of Experiment 1 were enclosed with $< t >< /t >$ tags, Fine-tuning was performed as in the other experiments, and responses were generated using the test data of Experiment 1. This was expected to improve output accuracy by enclosing the target words with tags. In experiment 8, we perform zero-shot CoT (Chain-of-Thought)(Kojima et al., 2022), prompting the model with "Let's think step by step" to guide multi-step inference. This was expected to improve output accuracy in zero-shot learning. Since experiment 3 was the least accurate of experiments 1-6, the data set was used directly in experiment 8 to evaluate the potential for improvement.

3.5 Result

Table 1 shows the output accuracy for each experiment. Accuracy improved in Experiments 1 and 7 after fine-tuning, but decreased in the others.

Table 1: Output Accuracy for Tuning

experiment	Fine-tuning	
	before	after
No.1	0.592 (71/120)	**0.708(85/120)**
No.2	0.679 (38/56)	0.464 (26/56)
No.3	0.679 (38/56)	0.357 (20/56)
No.4	0.679 (38/56)	0.500 (28/56)
No.5	0.679 (38/56)	0.607 (34/56)
No.6	0.679 (38/56)	0.643 (36/56)
No.7	0.600 (72/120)	**0.708 (85/120)**
No.8	0.643(36/56)	0.357(20/56)

4 Analysis

4.1 Disscussion

In this section, we present the results of analyzing the output trend of the experiments and the accompanying discussion. As shown in Table 1 and described in Section 3.6, fine-tuning improved accuracy in Experiments 1 and 7, while decreased it in others. This suggests fine-tuning is effective for few-shot settings but not for zero-shot in the semantic inference task for katakana words. The decline in zero-shot learning accuracy may stem from training data bias and the rarity of certain word senses. The former may be attributed to the data extraction method. We did not consider the frequency of word senses when extracting from the BCCWJ. Common meanings are likely to dominate the training set. The latter is because the test data are composed of sentences containing words with meanings that have a frequency of occurrence of 1 in BCCWJ, and it is considered that there are few situations in which the meaning of the word is actually used in documents. For example, in the zero-shot learning test data, the correct answer to the word "course" is "a dish of Western cuisine, served in order". The correct answer for the word "course" in the zero-Shot Learning test data was not found to be correct in all cases. The model or corpus likely lacks sufficient examples of the word, as it appears primarily in meal-related contexts. Experiment 7 did not show an improvement in precision compared to experiment 1, and the effectiveness of enclosing tags was not confirmed. Experiment 8 failed to improve upon experiment 3, suggesting that zero-shot CoT was ineffective, possibly due to Japanese training data or the nature of the semantic inference task. Experiments 4 and 6, which used half of the training data selected at random from experiments 3 and 5 , showed better accuracy—probably due to increased data bias. Furthermore, among experiments 2, 3, and 5, experiment 3 had the lowest accuracy and experiment 5 the highest accuracy. Since experiment 3 consists only of sentences that include the target word in the test data and experiment 5 excludes sentences that include the target word in the test data from the training data of experiment 2, it seems that the accuracy of semantic inference with zero-shot learning decreases when a large number of meanings of the target word are given in the training data. This suggests that including many meanings of a target word in training data reduces the accuracy of zero-shot learning inference.

4.2 Output Trend Analysis

We note several output trends. In experiments 1 and 7, the word "minus" is an example of a word

that did not produce the correct answer choice before and after fine-tuning. Digital Daijisen defines this word as having nine semantic categories, divided into detailed or limited meanings. There was little improvement before and after the fine-tuning, which may be due to the detailed semantic categories in the Digital Daijisen. Still, fine-tuning slightly improved accuracy for meanings like "not good, a bad aspect.". For experiments 2-6 and experiment 8 (zero-shot learning), we have an example where the output was more nuanced than the provided correct answer choice. The correct answer for "home" as part of a home improvement store is "Home. -bar, and my-," but the model before fine-tuning said that this is used differently from the usual "home" because it is precisely the name of a store called "home improvement store". However, fine-tuning made the output the labeled answer. Further, in experiments 2-6, 10 word senses below had one thing in common: neither of the pre-tuning nor post-tuning models output the correct answer choice. "In Western dressmaking, to cut fabric. Cutting. : cut", "In western cooking, a dish served in sequence. : course", "A style of architecture, art, music, etc. Mold. : style", "To assemble a tool, machine, etc. so that it can be used. To set up. : set", "To examine and prevent the entry of anything untoward. : check", "Negative electricity. Also, its symbol. : minus", "To subtract. : minus", "Adverse. Disadvantage. Loss. : minus", "In golf, a target hole cut on the green. : hole", "In a business organization, an organization above or below, such as a bureau, department, division, or section. : line". All of the senses are of limited use, and "minus" was particularly inaccurate.

4.3 Data Visualization

The 897 sentences extracted from BCCWJ created in section 3.2.2 were embedded into a numerical vector using OpenAI API, compressed into two dimensions, further divided into 21 clusters, and visualized in two-dimensional coordinates. UMAP (Uniform Manifold Approximation and Projection) (McInnes et al., 2018) was used to compress the dimensions. This is a dimensionality reduction method that efficiently compresses high-dimensional data to a lower dimension (two dimensions in this paper). Compared to t-SNE (Van der Maaten and Hinton, 2008), UMAP better preserves inter-cluster distances and allows parameter customization, making it suitable for this research. Therefore, we adopted it as the dimensionality reduction method in this paper. For clustering, we adopted HDBSCAN (Campello et al., 2013) due to its ability to handle complex data and accurately detect noise. This helps in visually capturing cluster tendencies. In HDBSCAN, clusters require at least 8 points; smaller groups and isolated points are treated as noise (338 points, 37.7%). Analysis of the plotted clusters revealed that many of the sentences share a similar context. For example, a certain cluster included many PC-related sentences, regardless of the target word. Therefore, we speculate that accuracy could be improved by including not only sentences with the same word senses as the target words, but also those belonging to the same cluster during fine-tuning. Furthermore, for sentences whose senses showed little effect from fine-tuning (experiments 2–6 in Section 4.2), we examined their cluster assignments. "In Western dressmaking, to cut fabric(cut).", "In western cooking, a dish served in sequence (course).". These two word senses were processed as Noise. The points treated as Noise (isolated) in plotted cluster likely lack similar sentence structures. This suggests that the BCCWJ lacks sufficient examples of sentences with the above "cut" and "course" meanings.We speculate that supplementing the dataset with sentences containing katakana words from other corpora that reflect these word meanings could improve precision.

5 Conclusion

Multiple experiments on the semantic inference of katakana words showed improved output accuracy in the gpt-4o-mini-2024-07-18 model when fine-tuned with few-shot learning. We suggest that for few-shot learning scenarios, fine-tuning with carefully selected data—including examples of hard-to-infer word senses, can enhance performance. In contrast, zero-shot accuracy declined, particularly for polysemous and context-specific words. Wasei-eigo requires a broader semantic interpretation, as its meaning often diverges from the original English source. Therefore, fine-tuning may reconstruct a semantic space biased toward specific usages, reducing zero-shot performance, particularly for Wasei-eigo with meanings that deviate from standard English. We hope that these results obtained in this research will be used to contribute to WSD in future research.

References

Edoardo Barba, Tommaso Pasini, and Roberto Navigli. 2021. Esc: Redesigning wsd with extractive sense comprehension. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4661–4672.

Terra Blevins and Luke Zettlemoyer. 2020. Moving down the long tail of word sense disambiguation with gloss informed bi-encoders. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1006–1017, Online. Association for Computational Linguistics.

Ricardo JGB Campello, Davoud Moulavi, and Jörg Sander. 2013. Density-based clustering based on hierarchical density estimates. In *Pacific-Asia conference on knowledge discovery and data mining*, pages 160–172. Springer.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Haoqiang Kang, Terra Blevins, and Luke Zettlemoyer. 2024. Translate to disambiguate: Zero-shot multilingual word sense disambiguation with pretrained language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1562–1575, St. Julian's, Malta. Association for Computational Linguistics.

Jan Kocoń, Igor Cichecki, Oliwier Kaszyca, Mateusz Kochanek, Dominika Szyd, Joanna Baran, Julita Bielaniewicz, Marcin Gruza, Arkadiusz Janz, Kamil Kanclerz, Anna Kocoń, BartKoptyra, Wiktoria Mieleszczenko-Kowszewicz, Piotr Mi, Marcin Oleksy, Maciej Piasecki, Radliński, Konrad Wojtasik, StanisWoźniak, and PrzemysKazienko. 2023. Chatgpt: Jack of all trades, master of none. *Information Fusion*, 99:101861.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Kikuo Maekawa, Makoto Yamazaki, Toshinobu Ogiso, Takehiko Maruyama, Hideki Ogura, Wakako Kashino, Hanae Koiso, Masaya Yamaguchi, Makiro Tanaka, and Yasuharu Den. 2014. Balanced corpus of contemporary written japanese. *Language resources and evaluation*, 48:345–371.

Marco Maru, Simone Conia, Michele Bevilacqua, and Roberto Navigli. 2022. Nibbling at the hard core of Word Sense Disambiguation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4724–4737, Dublin, Ireland. Association for Computational Linguistics.

Leland McInnes, John Healy, and James Melville. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Sakae Mizuki and Naoaki Okazaki. 2023. Semantic specialization for knowledge-based word sense disambiguation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3457–3470, Dubrovnik, Croatia. Association for Computational Linguistics.

Ming Wang and Yinglin Wang. 2020. A synset relation-enhanced framework with a try-again mechanism for word sense disambiguation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6229–6240, Online. Association for Computational Linguistics.

Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. 2021. A robustly optimized BERT pre-training approach with post-training. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 1218–1227. Chinese Information Processing Society of China.