Detection of Polysemy and Ambiguity in Japanese Adjectives Using Corpora

Takumi Osawa
Graduate School of Engineering
Takushoku University
24m303@st.takushoku-u.ac.jp

Takehiro Teraoka
Department of Computer Science
Faculty of Enginnering
Takushoku University
tteraoka@cs.takushoku-u.ac.jp

Abstract

In this work, we utilize different categories of modifiers to detect whether an adjectival expression is polysemous. Current disambiguation tasks focus only on words that have previously been determined as polysemous, and therefore require prior knowledge. An increase or decrease in a word's sense does not constitute polysemy in the conventional dictionary-based system and is thus not subject to word sense disambiguation. In this study, using a blog-based dataset and the Mainichi Newspaper Corpus, we detected polysemy and ambiguity by focusing on the difference between adjectives in sentences in which the adjectives are used. Our experimental results showed that the F-measure for polysemy detection and for ambiguity detection was 0.87 and 0.72, respectively, thus demonstrating the effectiveness of our method.

1 Introduction

Adjectives, adjectival verbs, and other adjectival expressions can sometimes have ambiguous meanings. As some of them are used in both positive and negative senses, it is vital to determine which sense they are used in. One example is the adjective 適当だ 'appropriate', which can be used both in the affirmative, as in "it fits well," and in the negative, as in "it is not good enough." While it was typically used in the positive sense in the past, these days it has increasingly been used in the negative sense. The polysemy of adjectival expression and the ability to accurately judge ambiguous adjectival expression used in both positive and negative forms is one of the most important factors in higher-level contextual understanding and emotional analysis today.

(1) 彼の掃除は適当だから部屋が汚い。(in Japanese)
kare-no souji-ha tekitou-da-kara heya-ga kitanai.
"His room is dirty because it is not well cleaned."

(2) その空欄に適当な語を埋める。
sono kuuran-ni tekitou-na go-wo umeru.
"Fill in the blanks with the appropriate words."

In the case of sentence (1) above, the term 適当だ 'not well' is used in the negative sense, i.e., "not quite right". In the case of sentence (2), the word 適当だ [1] 'appropriate' is used in the positive sense, such as "moderately appropriate," making it difficult to distinguish between the two. Various studies have been conducted on word sense disambiguation tasks to address this challenge. However, most prior works have targeted only words with prior ambiguity, and cannot handle cases in which the presence or absence of ambiguity is unknown. Therefore, the objective of this study is to detect polysemy and ambiguity in adjectives without prerequisite knowledge.

In recent years, ChatGPT has become widely utilized in various fields of natural language processing because it can generate sentences as if it were talking to a person. It is also easy to use, even for people who are unfamiliar with natural language processing, and its popularity among the regular population has therefore grown. Most recently, the GPT-4omni model (GPT-4o) has been launched and is attracting more and more attention, with additional target languages and improved performance over the previous GPT4 model. However, it has not been possible to make distinctions and judgements on the meaning of Japanese adjectives, which is the subject of this study. Below are some examples in which ChatGPT, using the GPT-4o model, was unable to distinguish between various adjectives. Specifically, the adjectives were not polysemous but ChatGPT judged them to be such, and the meanings assigned to them were not necessary to distinguish between them in the eyes of the people.

[1] 適当だ is the basic form.

- 過酷だ ‘Harsh’
  1. Very strictly forbidding (Harsh environment)
  2. Harsh conditions (Harsh working conditions)
- 清楚だ ‘Neat’
  1. Elegant (Woman who is neat and tidy)
  2. Pure (Having a neat image)

The above examples demonstrate that even in large language models (LLMs), there are cases where hallucinations occur and correct decisions cannot be made.

2 Related Works

Word Sense Disambiguation (WSD) is a topic that has been studied in many languages using a variety of supervised and semi-supervised learning methods. Yuan et al. (2016) based their WSD approach on a long short term memory (LSTM) language model and reported that the algorithm showed excellent results on many all-word tasks in SemEval. Thanks to its ability to take word order into account, the accuracy was significantly better than the algorithm based on Word2vec, especially for verbs. Le et al. (2018) replicated the unpublished model of Yuan et al. and confirmed that SemEval2 and SemEval2013 could achieve comparable performances using a corpus that was two orders of magnitude smaller. This suggests that a very large unannotated dataset is not necessary to improve the performance of all-word WSD. (Laba et al., 2023) conducted a WSD study for Ukrainian and showed that the context embedding required for WSD is best achieved by sentenceBERT (Reimers and Gurevych, 2019) using the multilingual model PMMBv2.

Rui et al. (2019)'s Japanese WSD study utilized embedded word representations obtained from BERT as the feature vectors of target words to perform word sense disambiguation. In conventional word sense disambiguation tasks, feature vectors are created and trained using a one-hot-vector and the part-of-speech, lexical, affix, and thesaurus information surrounding the target word as features. Since the embedded representation of each word is context-dependent, the representation obtained from BERT denotes the meaning of the word. In the experiment, word senses were discriminated for 50 target words.

In another approach, (Gumizawa and Yamamoto, 2018) created a topic-based classification dictionary for word sense disambiguation by assigning categories to words in consideration of the topic of the sentence. To improve the accuracy of word sense disambiguation by unsupervised learning, (Tabuchi and Osawa, 2022) examined features using the relations between superordinate and subordinate words defined in the Japanese WordNet. (Hashiguti and Sasaki, 2023) aimed to improve the accuracy of word sense disambiguation by replacing word sense labels with the estimated lexicographer categories.

The above studies are based on the assumption that the target words are polysemous, and do not take into account the increase or decrease in the number of senses of a word. In addition, nouns were often chosen as target words, and adjectives were rarely targeted.

3 Proposed Method

3.1 Dataset Construction

In this work, we assume that the different categories of modifiers indicate polysemy for a particular adjectival expression.

(3) あの山は高い。
ano yama-ha takai.
“That mountain is high.”

(4) あの財布は高い。
ano saihu-ha takai.
“That purse is expensive.”

There is no difference between sentences (3) and (4) in Japanese except for the modifier, and the word used for the adjective is the same in both sentences. 高い ‘High’ is an adjectival expression with multiple meanings, such as “located above a reference point such as the ground,” “high price,” and “a high frequency of sound vibration.” Therefore, the meaning of the word in the adjectival expressions of (3) and (4) is different. This suggests that differences in the categories of the modifiers create differences in the word sense of the adjectives.

Here, we construct the dataset by replacing the qualified terms with categorical terms. Sentences containing adjectives were extracted from the Hatena Blog Corpus[2] and the Mainichi Newspaper Corpus[3]. A classified vocabulary table was used to replace the modifiers with categorical words.

[2]https://hatenablog.com/
[3]http://mainichi.jp/contents/edu/03.html

(5) 友人と合流し，適当な店へ。

yuujin-to gouryuu-si tekitou-na mise-he.

"I met up with my friend and went to a suitable restaurant."

(6) 友人と合流し，適当な社会へ。

yuujin-to gouryuu-si tekitou-na syakai-he.

"I met up with my friend and joined a suitable society."

Above, (5) is the original sentence, and "restaurant", the modifier of "appropriate", belongs to "society" in the lexical category list, so the replacement occurs as in (6). For words that are not listed in the classified vocabulary list, Word2Vec is utilized to vectorize the meanings of the words. The vector representation obtained in this way was used to calculate the cosine similarity between words, and the word with the highest similarity was treated as the category word.

Words listed in the middle item of the Japanese Bunrui database (NINJAL, 2004) were used as category words in order to replace modifiers with category words. The Classified Lexicon is a database created by the National Institute for Japanese Language and Linguistics (NINJAL), in which words are classified according to their meanings. The number of records is 101,070, and the components of a record include the heading number, record type, middle item, and reading. There are a total of 49 types of entries, including "language," "food," "space," "use," "land," etc.

3.2 Determination of Polysemy

In our approach, we assume that the low cosine similarity between the modifiers in sentences in which a particular adjectival expression was used means that the target adjectival expression was used as a different sense of the word.

Under this assumption, by calculating the similarity between the modifiers and the variance of the similarity, we can determine the variation of the similarity for a single adjectival expression.

For example, we calculate the similarity of the modifiers of the sentences in which the adjectival expression "it's appropriate" is used in a round-robin manner. The similarity of the modifiers of the sentences in which a particular adjectival expression is used is calculated on a random basis, so the differences in the meaning of the adjectival expression will result in differences in the similarity of the modifiers.

4 Evaluation

4.1 Differences in the Classification of Modifiers

Since our approach is based on the assumption that "differences in the category of the adjectives indicate polysemy," it is necessary to verify whether there is a difference between adjectives with polysemy and adjectives without polysemy. For this purpose, the cosine similarity between the adjectives in a given sentence in the dataset is calculated on a random sample basis using BERT's (Devlin et al., 2018) variance representation, which can take the context into account. The values are then compiled into a heatmap. This allows us to visually identify the differences between adjectives with and without polysemy.

4.2 Detecting Polysemy

The dataset are assigned a label of 1 for ambiguity and 0 for non-ambiguity. The model is then evaluated by building the model with SVM. We utilize 10-fold cross-validation to ensure that the accuracy of the machine learning does not vary depending on the split test data.

The presence or absence of polysemy is determined by using the Digital Daijisen, and a word is considered to have polysemy if it has more than one sense. For SVM features, the variance of similarity of the modifiers, the minimum similarity, and the BertScore (Zhang et al., 2020) are used. The SVM attributes we used are listed in Table 1.

The similarity variance of the modifiers represents the difference between the categories of the modifiers. When there is polysemy, the similarity is scattered and the variance increases. In contrast, when there is no polysemy, the variance is small. The minimum value of the similarity varies depending on the ambiguity of the adjectives. The BertScore is a measure of how close the meanings of sentences are by using the embedded expressions in BERT. The baseline is a version of the disambiguation method used in the related study (Rui et al., 2019), extended to determine whether an adjectival expression has polysemy. In addition, we added the GPT-4o model ChatGPT LLM as a baseline as well, where we give the ChatGPT a list of target words and ask it to "divide these words into polysemy words with multiple senses and non-polysemy words. If the word is polysemous, please also specify which sense it has." I entered the above as a prompt.

Attribute	Description	Value
Variance of similarity of the modifiers	The similarity of the qualifiers is calculated by summing the similarity and taking the variance.	Continuous
Minimum similarity	The smallest value of the similarity of the modifier is calculated by round-robin.	Continuous
BERTScore	How close it is to the meaning of the sentence is determined.	Continuous

Table 1: Attributes and values with SVM.

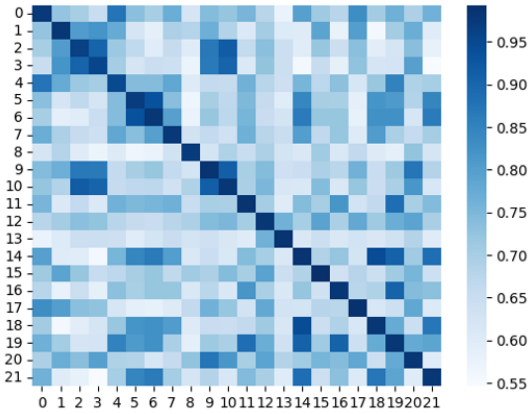

Figure 1: Polysemous, 適当だ 'Appropriate.'

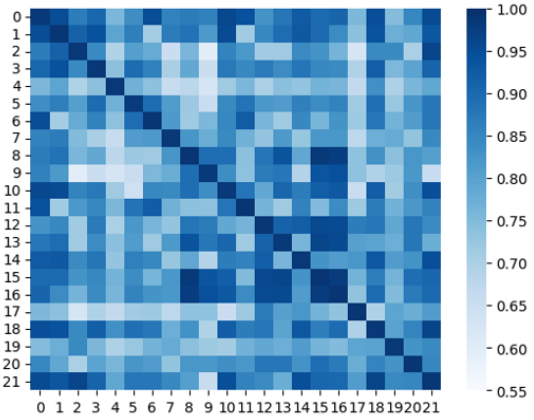

Figure 2: Not polysemous, 容易だ 'Easy.'

4.3 Detecting Ambiguity

We examine whether or not the adjectives in the target sentences that have been judged to have polysemy are ambiguous, with positive or negative usage. By replacing the adjectival expression with a synonym, we presume that an adjectival expression with ambiguity will make a difference in the meaning of the sentence. Therefore, the sentences before and after the replacement are processed to make a judgment.

Among the polysemy items, the ones used in both positive and negative senses were assigned a label of 1, and the others were assigned a label of 0. We then evaluated the model by building a model with SVM and used 10-fold cross-validation for the ambiguity detection. As in the case of polysemy detection, the Digital Daijisen is used for ambiguity detection. The BertScore, which was also used for polysemy detection, is used for the features. For the baseline, the polarity values calculated by Transformers are used as features.

4.4 Result

4.4.1 Differences in the Classification of Modifiers

Figures 1, 2, 3, and 4 respectively show cosine similarity heatmaps of the adjectival expressions "it's appropriate," "it's easy," "it's natural," and "it's huge" having polysemy and non-polysemy. The cosine similarity was calculated for each of the several sentences in which these adjectives were used, and the value of the diagonal line is 1.00. Figures 1 and 3 show that the adjectival expressions "it's appropriate" and "it's natural," which have a polysemous meaning, exhibit many light blue spots, indicating that the similarity is low in each of the sentences. In contrast, Figure 2 and 4 show that the adjectival expressions "it's easy" and "it's huge", which do not have polysemy, have more similarity than "it's appropriate" and "it's natural" because the dark blue color is scattered throughout the sentences. The high similarity of the adjectives means that they are used in the same sense. The similarity of the adjectives depends on their polysemy, which can be used as a feature to determine the polysemy of the adjectives.

4.4.2 Detecting Polysemy

Table 2 lists the number of adjective expressions and sentences for each corpus. The results of the evaluation experiment are shown in Table 3. In contrast to the baseline results using a neural network and adapted to the Hatena Blog Corpus, where both the percentage of correct answers and

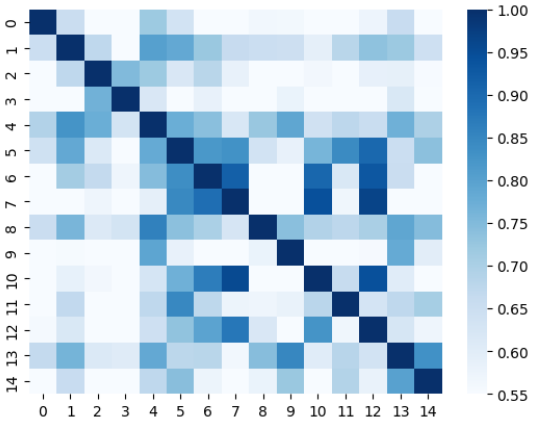

Figure 3: Polysemous, 当たり前だ 'Natural.'

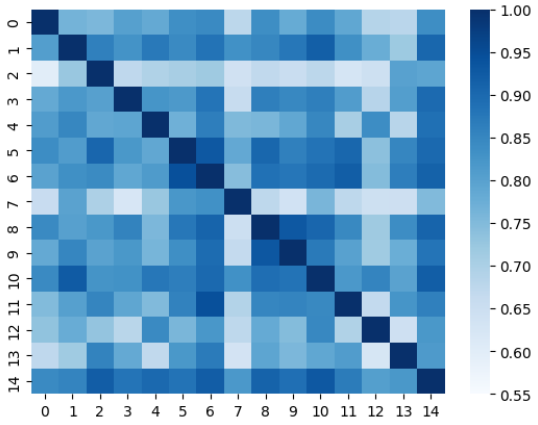

Figure 4: Not polysemous, 巨大だ 'Huge.'

Corpus	No. of words	No. of sentences
Hatena Blog Corpus	120	1,932
Mainichi Newspaper Corpus 2019	129	1,935
Mainichi Newspaper Corpus 2020	147	2,205
Mainichi Newspaper Corpus 2021	150	1,932

Table 2: Breakdown of each corpus.

the conformance rate were 70

In ChatGPT, polysemy was detected for the same adjectives as in the Hatena blog. The results showed that for words with polysemy, the detection was relatively accurate. However, for those without polysemy, many were incorrectly determined. ChatGPT determined that for a given adjectival expression, there were seven words without polysemy, but of these, only five were actually correct.

The above results indicate that focusing on the modifiers is suitable for detecting the ambiguity of the adjectives. However, the reproducibility of the method decreased compared to the baseline. This is presumably because there are more sentences that use adjectives with polysemy and more cases where it is impossible to judge if the adjective is not polysemous or not.

4.5 Detecting Ambiguity

Table 4 lists the results of the evaluation experiment using the Hatena Blog Corpus. The number of adjectives with polysemy is 67 and the number of sentences is 1,295. Compared to the baseline using polarity values, the reproduction rate of our method decreased, but the other evaluation indices increased. This resulted in more overtakes, but fewer false positives. Replacing adjectives with synonyms and using the difference between before and after replacement were some of the better elements for detecting ambiguity. However, this alone is not sufficient as a feature, and further improvement in accuracy is required. Another reason for the low baseline values is that many of the calculated polarity values were negative.

5 Discussion

5.1 Differences in the Classification of Modifiers

Figures 1 and 2 show the scatter plots of the cosine similarity between the modifiers. In Figure 1, many of the words are light blue, indicating that the cosine similarity values are generally low. Adjectival expressions with multiple meanings, such as "it's appropriate," can be used in multiple ways, and each meaning has a different category of modifier. As for Figure 2, in contrast, many of the words are colored dark blue, indicating that the cosine similarity score is higher than that of the other words.

Adjectival expressions such as "it's easy" that do not have polysemy have only one sense, so the category of the modifier does not change. The above results indicate that the degree of similarity of the modifiers is a useful feature to determine the presence or absence of polysemy.

5.2 Detecting Polysemy

We were able to detect the polysemy of adjectives with an accuracy of more than 80

	Accuracy	Precision	Recall	F-measure
ChatGPT (GPT-4o)	0.61	0.61	0.96	0.75
Baseline	0.72	0.72	1.00	0.85
Hatena Blog Corpus	0.81	0.80	0.97	0.87
Mainichi Newspaper Corpus 2019	0.70	0.73	0.80	0.75
Mainichi Newspaper Corpus 2020	0.70	0.76	0.58	0.64
Mainichi Newspaper Corpus 2021	0.68	0.69	0.94	0.79
Hatena Blog Corpus + Mainichi Newspaper Corpus 2019	0.79	0.77	0.98	0.86
Hatena Blog Corpus + Mainichi Newspaper Corpus 2021	0.80	0.81	0.97	0.87

Table 3: Polysemy detection results.

	Accuracy	Precision	Recall	F-measure
Baseline	0.44	0.50	0.78	0.61
Proposed Method	0.82*	0.80	0.65	0.72

Table 4: Ambiguity detection results. Asterisk (*) indicates a significant difference between the baseline and our proposed method, as verified by a sign test (p<0.01).

However, because the Hatena Blog Corpus is made up of blog-based content, many of the sentences are colloquial. Therefore, adjectives such as 甘い 'sweet' and 古い 'old,' which are polysemous, were judged to be non-polysemous. In the case of 甘い 'sweet,' some of the blogs obtained by scraping were food reports, and many words that expressed sweetness in terms of taste, such as "it tastes like sugar or honey," were found. Therefore, examples of the expressions "lack of harshness" and "pleasantly enchanting" were not present in the corpus. In the case of 古い 'old,' the meanings of "a long time has passed since it was in that state," "outdated," and "not fresh" were all present and used in the corpus. However, all of them were considered to be polysemous by our method, since there was no difference between them.

The Mainichi Newspaper Corpus is one of the strictest written corpora in terms of written expression, and as a result, there is often a single use of the word sense of an adjectival expression. Therefore, compared to the Hatena Blog Corpus, it was sometimes difficult to correctly determine whether a word had multiple meanings or not. Among them, the Mainichi Newspaper Corpus 2020 had a lower evaluation index than the other Mainichi Newspaper corpora. Therefore, among the adjectives that were judged to have polysemy but not polysemy, those with a variance value of less than 0.01 and a minimum value of 0.50 or more were excluded and re-detected, and the results are shown in Table 5.

Hallucination occurred in the LLM ChatGPT, which judged most words as having polysemy for adjectival expressions that did not have polysemy. Our method is better at detecting polysemy, as it was able to correctly judge some adjectival expressions as having no polysemy, which ChatGPT incorrectly detected.

5.3 Detecting Ambiguity

In terms of the ambiguity detection, the accuracy of the proposed method was significantly higher than that of the baseline method, which used polarity values as features. These polarity values were mostly negative, and there were almost no sentences that were judged to be positive. Therefore, there was no difference between adjectives with and without ambiguity, and the values of the evaluation index were calculated to be low across the board. In contrast, the proposed method replaced words in the adjectives with synonyms and looked at the relationship between the words before and after the synonyms, so it was not affected by the polarity value.

However, although the proposed method is more accurate than the baseline method, there is still room for improvement. The ambiguity detection had corpus-dependent problems, which were more pronounced than in the case of polysemy detection. Two examples are the words 適当だ 'not well' and 微妙だ 'subtle.' In Japanese, the word 適当だ 'not well' has two types of usage: positive (e.g., "moderately applicable") and negative (e.g., "not good enough"). However, in the blog-based corpus, where many colloquial expressions are used, the negative usage of "irresponsible" was often found. In addition, 微妙だ 'subtle' is often used negatively as a "euphemism for a negative mood," and less frequently as a positive expression of "tasteful, indescribable beauty or flavor."

	Accuracy	Precision	Recall	F-measure
Mainichi Newspaper Corpus 2020	0.75	0.79	0.75	0.73
Mainichi Newspaper Corpus 2019 + 2020 + 2021	0.72	0.70	0.98	0.82
Hatena Blog corpus + Mainichi Newspaper Corpus 2020	0.68	0.69	0.94	0.79

Table 5: Polysemy detection results.

As described above, the bias in the sense of the word used for one adjective may have resulted in the low recurrence rate. Therefore, it is necessary to consider not only the BertScore before and after the substitution but also the co-occurrence information of the sentences and distributed expressions.

6 Conclusion

6.1 Summary

In this work, we aimed to detect ambiguity by determining the polysemy of an adjectival expression using the difference in the categories of the modifiers, replacing the adjectival expression with a synonym, and analyzing the difference between the sentences before and after the replacement. The assumption was made that the difference in the meanings of adjective expressions was the difference in the category of the modifier, so we also investigated whether this assumption was correct or not. The results of evaluation experiments visually showed from the heatmap that the difference in the category of the modifier is effective in determining whether an adjectival expression is polysemous or not. The differences in the categories of the modifiers were used to determine the polysemy of the adjectives. The variance of the cosine similarity, the BERTScore, and the minimum value of the cosine similarity were used for the features, and an F value of 0.87 was obtained, which is high accuracy. In judging ambiguity, the F value was 0.72, which was not very accurate because there were cases in which there was a difference in colloquial or written expressions between the positive and negative meanings of a word.

6.2 Future Work

In this study, two levels of detection were used: whether the adjectival expression has polysemy or ambiguity. One of the common problems in both detection methods is that the accuracy depends on the dataset: namely, some adjectives with polysemy and ambiguity are more likely to be used as colloquial expressions, while others are more likely to be used as written expressions. In the blog-based dataset we used, many of the meanings of adjectives were used as colloquial expressions, while those used as written expressions were less common. Even in the Mainichi Shimbun corpus, which includes written expressions, there were words for which the univocality of the meaning was observed and the polysemy of the adjectival expression could not be judged well.

The results of this study showed that, while the accuracy of detecting ambiguity was good, it was not as high as that of detecting polysemy. Therefore, we believe that not only looking at the difference between adjectives and synonyms but also considering the sentences before and after the adjectives and using distributed expressions may improve the accuracy. In addition, to improve the accuracy of dialogue systems, it is necessary to determine the meaning of the ambiguous adjectives detected.

Acknowledgement

This work was supported by JSPS KAKENHI Grant Number JP22K00646.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Yuki Gumizawa and Kazuhide Yamamoto. 2018. Japanese word sense disambiguation based on topics. *Proceedings of the Twenty-fourth Annual Meeting of the Association for Natural Language Processing*, pages 248–251. (in Japanese).

Takuya Hashiguti and Minoru Sasaki. 2023. Wordnet lexicographer category estimation for word meaning size contraction for word meaning disambiguation. *Proceedings of the Twenty-nineth Annual Meeting of the Association for Natural Language Processing*, pages 1038–1042. (in Japanese).

Yurii Laba, Volodymyr Mudryi, Dmytro Chaplynskyi, Mariana Romanyshyn, and Oles Dobosevych. 2023. Contextual embeddings for ukrainian: A large language model approach to word sense disambiguation. In *Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP)*, pages 11–19.

Minh Le, Marten Postma, Jacopo Urbani, and Piek Vossen. 2018. A deep dive into word sense disambiguation with lstm. In *Proceedings of the 27th international conference on computational linguistics*, pages 354–365.

NINJAL(2004). 2004. Word list by semantic principles(『分類語彙表増補改訂版データベース』(ver1.0)).

Niles Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing,*, pages 3982–3992.

Cao Rui, Hirotaka Tanaka, Bai Jing, Ma Wen, and Hiroyuki Shinnou. 2019. Word sense disambiguation using supervised learning with bert. In *Proceedings of Language Resources Workshop*, volume 4, pages 273–279. (in Japanese).

Tomoaki Tabuchi and Ei-Ichi Osawa. 2022. Features for improving the accuracy of unsupervised learning for word sense disambiguation. In *The 36th Annual Conference of the Japanese Society for Artificial Intelligence, 2022*, pages 2B5GS601–2B5GS601. The Japanese Society for Artificial Inteligence. (in Japanese).

Dayu Yuan, Julian Richardson, Ryan Doherty, Colin Evans, and Eric Altendorf. 2016. Semi-supervised word sense disambiguation with neural models. *arXiv preprint arXiv:1603.07012*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. *ICLR*.