All-Words Word Sense Disambiguation for Historical Japanese

Soma Asada and **Kanako Komiya** and **Masayuki Asahara**
Tokyo University of Agriculture and Technology / 2-24-16 Naka-cho, Koganei-shi, Tokyo ,Japan
National Institute for Japanese Language and Linguistics / 10-2 Midoricho, Tachikawa,Tokyo, Japan
s231157v@st.go.tuat.ac.jp
kkomiya@go.tuat.ac.jp
masayu-a@ninjal.ac.jp

Abstract

This paper presents an all-words word sense disambiguation (WSD) system for historical Japanese. For historical Japanese, a WSD system for a lexical sample task, which only targets frequent words in a corpus, has been reported (Komiya et al., 2022a). However, the WSD system for a lexical sample task requires a model to be trained for each target word. We developed an all-words WSD system as a sequential labelling system, which trains a single model for all words in a corpus. In addition, we input the book ID as well as the input text to give information, from which the text was taken, into the system. We used two granularities of word senses, middle-grained and fine-grained concept IDs defined by Word List by Semantic Principles. The accuracies of our system were 87.62 % for middle-grained senses and 85.25 % for fine-grained senses and they significantly outperformed the most frequent sense baselines and simple BERT systems without book IDs in both settings. Finally, we investigated the effect of the base large language models trained with contemporary Japanese and the influence of multitask learning.

1 Introduction

Word sense disambiguation (WSD) is a process that determines a word sense of a polyseme, i.e., a word that has multiple senses. For example, “dream” mainly has two senses: (1) a series of reality-like ideas or mental images that happen in one’s mind during sleep and (2) a wish to be or to have something that is hard to achieve. The sense of a polyseme is identified by its use in a context. In the field of machine learning, the system defines the sense of a target word using contextual information such as parts of speech of surrounding words or word co-occurrence relationships. WSD can contribute to vocabulary exercises and help with reading comprehension for beginners in language learning.

In this paper, we conduct WSD on historical Japanese texts. WSD for ancient languages is beneficial for achieving a more accurate understanding of texts where introspection is not effective, by automatically assigning word senses to the ambiguous words. Historical texts reflect language usage from older times and often contain polysemes. For instance, when examining the Japanese term ‘‘為る (*suru*)” (English “do”), we find that it holds eight distinct word senses in contemporary language; however, in classical language, it encompasses twenty-five distinct word senses. Applying WSD allows readers to understand the precise meaning of polysemes used in the context of historical texts. WSD of polysemous words in historical texts also contributes to linguistic research. Understanding word usage and meaning changes in specific eras and regions deepens cultural and societal comprehension.

WSD is broadly divided into two types of task: lexical sample task and all-words WSD. Lexical sample task aims to determine the senses of specific words, in most cases, they are words that frequently appear in a corpus, and all-words WSD aims to determine senses of all words in texts. Generally, the system of lexical sample task trains a classification model for each target word while the all-words WSD system usually trains a single model using a sequential labelling approach. The accuracy of WSD for historical Japanese was lower than that for contemporary Japanese due to the small amount of word sense tagged data (Komiya et al., 2022b). However, thanks to the completion of word sense tagging of the Corpus of Historical Japanese (CHJ)[1] in 2022, much data is now available.

For historical Japanese, a WSD system for a lexical sample task has been reported (Komiya et al., 2022a,b) (see Section 2). We developed an all-words WSD system for the same corpus (see Sec-

[1] https://clrd.ninjal.ac.jp/chj/overview-en.html

tion 4). Hereby, word senses of less frequent words in the corpus could be determined. In addition, we added the book IDs as the prefix of the input sentence and gave the information, from which book the sentence was taken, to the system. For word senses, we used the concept IDs defined by Word List by Semantic Principle (WLSP) (National Institute for Japanese Language and Linguistics, 1964)[2] and experimented with two types of granularities of concept IDs (see Section 3). We used Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019) and a Robustly Optimized BERT Pretraining Approach (RoBERTa) (Liu et al., 2019) for the system (see Section 5). The experiments revealed that our system input book IDs outperformed the most frequent sense (MFS) baselines and the simple BERT systems without book ID inputs. In addition, we investigated the effect of the base language models trained with contemporary Japanese and the influence of multitask learning with a document classification task (see Section 6).

The contributions of this paper are as follows:

1. We developed an all-words WSD system from historical Japanese using large language models;

2. We proposed giving the book IDs to inform from which book the input sentence was extracted into the system;

3. The accuracy of our system outperformed the MFS baseline and the simple BERT model without book ID inputs; and

4. We analyzed the effect of the base large language models trained with contemporary Japanese and the influence of multitask learning.

2 Related Work

Komiya et al. (2022b) and Komiya et al. (2022a) reported research on WSD for historical Japanese. They tackled the lexical sample task. They used methods for diachronic domain adaptation using contemporary Japanese for historical Japanese. Komiya et al. (2022b) compared various types of features for a historical WSD system. They showed that the word embeddings (word2vec) trained with historical texts and fine-tuned with contemporary texts is effective for WSD for historical Japanese. Komiya et al. (2022a) developed a WSD system using BERT, which is a kind of large language models pre-trained with a large amount of contemporary Japanese texts, to complement the training data. They also reported the effectiveness of multitask learning with a document classification task. As far as we know, this paper is the first attempt to develop all-words WSD system for historical Japanese. Following (Komiya et al., 2022a), we used BERT pre-trained with contemporary Japanese texts.

In addition, much work has been done on WSD for contemporary Japanese including all-words WSD. For example, Suzuki et al. (2019) reported an all-words WSD system for contemporary Japanese. They have shown that a system using the Euclidean distance of embeddings between words around a target word and synonyms is effective for disambiguating the word senses in the Balanced Corpus of Contemporary Written Japanese (BCCWJ) (Maekawa et al., 2014). They adopted a knowledge-based method where no labelled data was used whereas we used a supervised method for all-words WSD. Shinnou et al. (2017) developed an all-words WSD system with a text analysis tool KyTea[3]. They also conducted experiments on BCCWJ.

Some works have been done on English all-words WSD (Iacobacci et al., 2016; Raganato et al., 2017; Navigli et al., 2007; Du et al., 2019; Blevins and Zettlemoyer, 2020; Keung et al., 2020). Du et al. (2019) performed an all-words WSD in English using BERT and showed its effectiveness for this task.

3 Data

We used 10 pieces of literature in CHJ-WLSP(Asahara et al., 2022) for the experiments, which are totally the same as (Komiya et al., 2022a). CHJ is a diachronic corpus from the Nara period to the Meiji and Taisho eras. The collection ranges in age from the 900s to the 1900s, and its genres vary from stories, essays, and textbooks. The statistics of ten pieces of the literature book are summarized in Table 1. Samples are the identifier for the literature book in CHJ. Descriptions are the title in Japanese and their (literal) English translation. Year is the year of the establishment of the literature book. Words are the word count in the

[2] https://www2.ninjal.ac.jp/past-publications/english/publication/catalogue/bunruigoihyo_2ed/index.html

[3] http://www.phontron.com/kytea/index-ja.html

annotated samples[4]. In total, 647,751 words are annotated from 11 samples, which are 10 pieces of literature, because Jinjo Shogaku Tokuhon (Textbook) has two editions.

Concept IDs of WLSP were annotated for the subset of CHJ as word sense labels. WLSP is a thesaurus that classifies and organizes words by their senses. It records semantic information such as word groups with similar meanings and inclusion relationships of concepts in a form that is easy to handle. A record in WLSP contains the following information: record IDs, lemma IDs, types of record, class IDs, division IDs, section IDs, articles, article IDs, paragraph IDs, small paragraph IDs, word IDs, lemma with explanatory notes, lemma without explanatory note, pronunciations, reverse pronunciations. We used the concept IDs as word sense in the experiments. They include information on the IDs of classes, divisions, sections, and articles. For example, when the concept ID, "2.3102," is assigned to a word "言う"(say), it means that the class ID is "2 (Verbal,)" the division ID is ".3 (Action,)" the section ID is ".31 (Language,)" and the article ID is ".3102 (Name.)" The smaller decimal place indicates the more detailed classification.

Table 2 shows an annotation example of Taketori Monogatari. The pSample and pStart columns are the offset information in the CHJ. Word segmentation is carried out on the corpus and the morphological information is annotated to it. The surface form (orthToken) and the lemma of the original texts were also included in the corpus. Although we cannot show POS tags in the table because of the page limit, the annotator can also see the POS tags and annotate the word sense labels in the concept ID column. For example,'今' (*now*) is annotated by concept ID 1.1641. Table 3 shows the label structure of WLSP.

Table 4 lists the statistics of the sense-annotated corpus, CHJ-WLSP. We experimented with two granularities of word senses. One is middle-grained, which uses the first three digits of the concept IDs defined by WLSP. The other is fine-grained, which uses all five digits of the concept IDs. As seen in Table 4, the number of target sense types remarkably declines (from 1,747 to 304) by coarsening the granularity of word senses, which means the degree of ambiguity is reduced. The number of target tokens and types also declines. As a result, the averages of word senses per word are 2.91 and 2.73 for fine-grained and middle-grained concept IDs respectively.

4 All-words WSD of Historical Japanese

We build an all-words WSD system using Japanese contemporary BERT and a RoBERTa as pre-trained models and fine-tuning them with historical Japanese, following the prior study (Komiya et al., 2022a). We developed the system as a sequential labeling system, which allocates sense tags for all input words. Unlike other sequential labeling tasks such as named entity recognition, for WSD, each target word has a different set of labels. For example, the meaning of "dog" should be selected from its sense inventory and the system does not have to consider any meanings of words other than "dog." Hence, the system referred to the sense inventory to obtain the candidate sense labels of each target word and considered the one with the highest output score as the correct label for both the training and inference steps.

In addition, we input the book IDs into the system as well as the input sentence. The meanings of words tend to vary depending on the domains of the texts and the periods when the texts were written. We intended to give this information to the system. Moreover, we implemented multitask learning of WSD and document classification. The system is required to simultaneously predict word senses in an input sentence and the book title from where the sentence was taken. These methods are inspired by (Komiya et al., 2022a), which reported that the multitask learning with document classification task was effective for the lexical sample task of WSD.

Figure 1 demonstrates the inputs of the system according to each method. The methods "-single," "-prefix," and "-multi" mean BERT-single or RoBERTa-single, BERT-prefix, and BERT-multitask, respectively. For the BERT-prefix method, we added a book ID and a [SEP] token before the input sentence. In the example, "0" is book ID of Taketori Monogatari. Because we restricted the maximum length of the input sentences to 510 tokens, the length of the input sentences is all the same regardless of the methods. Only five sentences in the dataset we used were equal to or more than the maximum length. The prefixes or book IDs are out of range of the evaluation of the WSD systems. We used the [CLS] token for document classification of the BERT-multitask method. How-

[4] The annotation of 1642虎明 is for not whole data.

Samples	Descriptions	Year	Words
0900竹取	Taketori Monogatari (*lit. The Tale of the Bamboo Cutter*)	10th century	12,757
0934土佐	Tosa Nikki (*lit. Tosa Diary*)	10th century	8,208
1100今昔	Konjaku Monogatari-shu (*lit. Anthology of Tales from the Past*)	Heian period	175,598
1212方丈	Hojoki (*lit. Square-jo Record*)	1212	5,402
1220宇治	Uji Shui Monogatari (*lit. Gleanings from Uji Dainagon Monogatari*)	13th Century	120,705
1252十訓	Jikkin-sho (*A Miscellany of Ten Maxims*)	1252	90,177
1336徒然	Tsurezuregusa (*Essays in Idleness*)	ca. 1330	40,834
1642虎明	Toraakira-bon Kyogen [a]	1642	5,448
1895太陽	Taiyo *The Sun* (Magazine) [b]	1895	46,394
1904小読	1st Jinjo Shogaku Tokuhon (Textbook) [c]	1904	45,334
1910小読	2nd Jinjo Shogaku Tokuhon (Textbook)	1910	96,894
Total			647,751

[a] https://iss.ndl.go.jp/books/R100000002-I000008304623-00
[b] https://viaf.org/viaf/184683725/
[c] https://dglb01.ninjal.ac.jp/ninjaldl/bunken.php?title=kokutei1

Table 1: The statistics of ten pieces of the literature book: from Asahara et al. (2022), Table 2

pSampleID	pStart	orthToken	lemma	Concept ID	Class	Coarse	Middle	Fine
20-竹取0900_00001	20	いま	今	1.1641	Nominal	Relation	Time	Now
20-竹取0900_00001	40	は	は					
20-竹取0900_00001	50	むかし	昔	1.1642	Nominal	Relation	Time	Past
20-竹取0900_00001	80	、	、					
20-竹取0900_00001	90	たけとり	竹取					
20-竹取0900_00001	130	の	の					
20-竹取0900_00001	140	翁	翁	1.2050	Nominal	Subject	Human	Old-Young
20-竹取0900_00001	150	と	と					
20-竹取0900_00001	160	いふ	言う	2.3102	Verbal	Action	Language	Name
20-竹取0900_00001	180	もの	者	1.2000	Nominal	Subject	Human	Human
20-竹取0900_00001	200	あり	有る	2.1200	Verbal	Relation	Existence	Existence
20-竹取0900_00001	220	けり	けり					
20-竹取0900_00001	240	。	。					

Translation: *Once upon a time, there was an old man called Taketori.*

Table 2: Annotation example of Taketori Monogatari

「今」 'now': 1.1641			
Syntactic	Semantic		
	Coarse	Middle	Fine
Class	Division	Section	Article
体	関係	時間	現在
Nominal	Relation	Time	Now
1	.1	.16	.1641

Table 3: Label structure of WLSP (Concept ID)

ever, for the BERT-single, RoBERTa-single, and BERT-prefix methods, we used the [CLS]-ignore option.

5 Experiments

In this paper, we used Japanese BERT model[5] and RoBERTa model[6] pre-trained with contemporary Japanese texts. The architectures of the two models are the same as the original ones. The Japanese BERT was trained with Japanese Wikipedia, which is 4.0 GB in total and contains approximately 30M sentences. The RoBERTa model was trained with Japanese Wikipedia (as of 2021/09/20) and the Japanese data of CC-100.

5.1 Experimental Settings

The inputs of the system are sentence-based. The sentences are separated by a Japanese punctuation mark (。) in the CHJ. Exceptionally, there are no punctuation marks in Toraakirabon Kyogen (虎明本狂言) because it is written in a script

[5] https://huggingface.co/cl-tohoku/bert-base-japanese-v2

[6] https://huggingface.co/nlp-waseda/roberta-base-japanese

	Fine-grained	Middle-grained
Tokens	647,751	647,751
Target tokens	329,109	324,952
Target types	3,878	3,672
Target sense types	1,747	304
Average of senses	2.91	2.73

Table 4: Statistics of CHJ-WLSP

Methods	Inputs														
-single			いま	は	むかし	、	たけとり	の	翁	と	いふ	もの	あり	けり	。
-prefix	0	[SEP]	いま	は	むかし	、	たけとり	の	翁	と	いふ	もの	あり	けり	。
-multi		[CLS]	いま	は	むかし	、	たけとり	の	翁	と	いふ	もの	あり	けり	。
(Fine labels)			1.1641		1.1642				1.2050		2.3102	1.2000	2.1200		
(Middle labels)			1.16		1.16				1.20		2.31	1.20	2.12		

Figure 1: Examples of input data for each method

form. Therefore, we utilized Japanese-style quotation marks (「, 」, 『, and 』) and reading marks (、) with a boundary tag assigned at sentence boundaries in the corpus as the markers of the end of sentences [7]. We used orthographic tokens, lemma, and concept IDs for input features of each word in a sentence. The inputs of the system is sentence-based and the input sentences, the examples, are shuffled in random order, regardless of periods or book titles. Since the vocabulary size of the large language model is limited, a single word in CHJ sometimes split into multiple subwords. For example, "痛み入る" (to feel sorry with gratefulness) is treated as a single word in CHJ, but the language model split the word into "痛み"(pain) and "入る" (enter). In this case, the first token is treated as a target word.

For the hyperparameters, we conducted a grid search using the values shown in Table 5. We tuned the hyperparameters using validation data. We conducted a five-fold cross-validation with a ratio of training: validation: test data as 3: 1: 1.

Granularity	Epoch number	Learning rate
Fine	5,10,15	3e-5, 1e-5, 3e-6
Middle	10, 15	3e-5, 1e-5, 3e-6

Table 5: Hyperparameters

We used Adam as the optimization function and cross-entropy loss as the loss function.

5.2 Evaluation Method

For all tokens that have multiple senses in the corpus, we calculated the accuracy using the following formula.

$$\text{Accuracy} = \frac{\text{Number of correct tokens}}{\text{Number of target tokens}} \quad (1)$$

Notably, even if a word has multiple senses in the corpus, not all of them appeared in both the training and test data. For both fine- and middle-grained senses, the number of polysemes that appeared only twice in the corpus is approximately 10% of all polysemes in the corpus. We compared our methods with the most frequent sense (MFS) baseline, which is calculated using the following formula.

$$\text{MFS} = \frac{\text{Number of tokens with MFS}}{\text{Number of all tokens}} \quad (2)$$

For example, if a word that has two senses appeared 10 times in a corpus, the word whose sense was sense 1 appeared 6 times and the word whose sense was 4 times, and the MFS baseline will be 60%.

6 Evaluation

Table 6 presents the accuracies of WSD. MFS, BERT-single, RoBERTa-single, BERT-prefix, and BERT-multi in the table indicate the most frequent sense baseline, the BERT-based system with single-task learning, the RoBERTa-based system with single-task learning, BERT-based system with single-task learning with book ID inputs as the prefix of the input sentence, and the BERT-based system with multitask learning, respectively.

According to the table, we can see that the BERT-based system with single-task learning with book

[7] We did not directly use a boundary tag because it was assigned at each sentence in complex sentences. For example, the tags were assigned at B and D in the sentence "A is B because C is D."

ID inputs as the prefix of the input sentence is the best for both middle- and fine-grained senses. The differences between the best method and the second-best method, the BERT-based system with single-task learning with and without book ID inputs, are significant by a chi-square test with a significance level of 0.05, in experiments with fine- and middle-grained senses. Therefore, we can see that the book ID contributed to the improvement of the accuracy of the systems. We believe that this is because the meanings of words tend to vary according to the domains or writing styles of texts and the periods when the texts were written. The book ID could give information about them to the system.

In Table 6, the differences between the second-best method, the BERT-based system with single-task learning, and the MFS baseline are also significant. This result is the same as that of the lexical sample task, reported by Komiya et al. (2022a). In addition, the differences between the second-best method and the RoBERTa-based system are also significant. This fact indicates that, for all-words WSD systems from historical to contemporary Japanese, BERT we used is better than RoBERTa we used as the base large language model.

Granularity	Method	Accuracy[%]
Fine-grained	MFS	81.61
	BERT-single	84.52
	RoBERTa-single	83.78
	BERT-prefix	**85.25**
	BERT-multi	80.76
Middle-grained	MFS	84.10
	BERT-single	87.11
	RoBERTa-single	86.60
	BERT-prefix	**87.62**
	BERT-multi	85.59

Table 6: Accuracies of all-words WSD

Now, let us compare our results with the results of (Komiya et al., 2022a), shown in Table 7, the work of the lexical sample task WSD of historical Japanese, although they cannot be directly compared because the target words are different.

They used 33 target words which appear more than 1,000 times in CHJ, whereas we disambiguated the senses of all polysemes in the same corpus. The accuracy of the simple BERT-based system with single-task learning for the fine-grained senses (84.52 %) is competitive with theirs (84.68 %). However, when we compare our MFS baseline of fine-grained senses with their MFS baseline, ours (81.61 %) is higher than theirs (78.29 %) by 3.32 points, which implies that less frequent words, that is, the words appeared less than 1,000 times in a corpus, tend to have a higher probability of being the most frequent sense and are easy to disambiguate.

Granularity	Method	Accuracy[%]
Fine-grained	MFS	78.29
	BERT-single	84.68
	BERT-multi	85.17

Table 7: Accuracies of lexical sample task WSD

While at the same time, less training data tends to lead to less accuracy for machine learning. Table 8 shows the accuracies of the BERT-based systems with single-task learning and the MFS baselines for less frequent words. We can see that all the accuracies of the BERT-based systems with single-task learning are substantially lower than those of the MFS baseline except for the words which appeared only two times in the corpus. They are extreme cases of words with little training data. For these words, the fallback algorithm using the MFS in the training data could be effective in the future.

	Fine-grained		Middle-grained	
# occurrence	BERT	MFS	BERT	MFS
	[%]	[%]	[%]	[%]
2	51.91	50.00	53.43	50.00
3	51.13	63.35	55.43	63.98
4	51.08	62.92	58.48	63.66
5	53.56	67.06	59.81	67.80
6	55.98	67.21	61.69	69.11
7	59.62	69.43	62.40	70.74
8	59.48	71.22	61.73	72.01
9	61.66	70.87	66.24	70.11

Table 8: Accuracies of less frequent words.

Table 9 shows WSD accuracy and MFS baseline according to the frequencies. The row frequency in the table means less than 130 times, middle frequency means equal to or more than 130 times and less than 360 times, and high frequency indicates equal to or more than 360 times. We can see that the WSD accuracy could not outperform the MFS baseline when the frequency of the target word of WSD is less than 130 times. Although they don't affect the micro-averaged WSD accuracy because

Method	Granularity	Frequency		
		Low	Middle	High
BERT-prefix	Fine-grained	72.62	80.14	90.56
	Middle-grained	75.76	82.85	92.38
MFS	Fine-grained	75.49	75.55	87.16
	Middle-grained	77.61	78.96	89.46

Table 9: WSD accuracy and MFS baseline according to the frequencies

the frequency of the word is low, the WSD accuracy of the less frequent word should be improved in the future.

In addition, as seen in Table 6, the systems with multitask learning do not yield good results in comparison to single-task learning systems for both granularities[8]. This result is opposite to (Komiya et al., 2022a), which reported that multitask learning with document classification is effective for lexical sample task WSD, when they used the same dataset as ours (CHJ-WLSP 2022). However, they reported that it was not effective for the experiments with fewer training data, CHJ-WLSP 2019. Therefore, we see that the decline in accuracies when we used multitask learning probably comes from the effects of less frequent words.

Although the multitask learning with document classification task, the book ID was effective for the improvement of the WSD accuracies. Therefore, we assume that the information itself, from where the text was extracted, was effective. Lack of training data for each WSD target words, especially for less frequent words, could be the reason why the multitask learning with document classification is not effective.

Finally, Tables 10 and 11 show the top 10 words with frequent errors of our best system, the BERT-based system with book ID inputs, using middle- and fine-grained senses, respectively. These tables display not only words but also translations, the number of senses (#senses), the number of errors (#errors), the number of tokens (#tokens), and the error rates of the words. We can see that all of the top 10 words with frequent errors except for one word with an asterisk mark are the target words for the lexical sample task in (Komiya et al., 2022a). The exception is the word "ばかり," which is an adverbial particle that means "only." This result indicates that the rare senses of frequent words should be the main problem to be solved in the future.

7 Conclusion

In this paper, we developed an all-words WSD system for CHJ. We used the concept IDs defined by WLSP as word sense and implemented WSD systems for two granularity-senses, fine-grained and middle-grained senses. We input the book IDs as well as the input text where the word senses are to be disambiguated to consider the domains of the texts and the periods when the texts were written. We compared two large language models, BERT and RoBERTa trained with contemporary Japanese and investigated the effect of multitask learning with document classification. The results show that the BERT-based system with single-task learning with book ID inputs was the best. The book IDs contributed to the improvement of the WSD accuracy. In addition, the Japanese BERT we used was better than the RoBERTa we used for all-words WSD in historical Japanese for both granularity settings and multitask learning was not effective. The results of experiments indicate that the rare senses of frequent words should be the main problem to be solved in the future.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Number 22K12145, and collaborative research projects of National Institute for Japanese Language and Linguistics.

References

Masayuki Asahara, Nao Ikegami, Tai Suzuki, Taro Ichimura, Asuko Kondo, Sachi Kato, and Makoto Yamazaki. 2022. CHJ-WLSP: Annotation of 'word list by semantic principles' labels for the corpus of historical Japanese. In *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages*, pages 31–37, Marseille, France. European Language Resources Association.

[8] The accuracies of document classification itself are 85.43 for fine-grained senses and 85.42 for middle-grained senses.

Ranks	Words	Translations	#senses	#errors	#tokens	Error rates
1	為る	do	19	2,484	7,072	0.3512
2	成る	become	14	610	1,740	0.3506
3	然る	like that	12	501	1,548	0.3236
4	物	object	7	496	2,053	0.2416
5	取る	get	15	469	1,262	0.3716
6	共	together	10	369	1,199	0.3078
7	ばかり*	only	5	351	1,078	0.3256
8	様	appearance	14	321	1,805	0.1778
9	又	and	6	317	1,373	0.2309
10	皆	every	4	312	729	0.428

Table 10: Top 10 words with frequent errors of the experiment with middle-grained senses

Ranks	Words	Translations	#senses	#errors	#tokens	Error rates
1	為る	do	26	2638	7,072	0.373
2	成る	become	17	688	1,740	0.3954
3	然る	like that	13	545	1,548	0.3521
4	物	object	8	521	2,053	0.2538
5	人	human	7	518	4142	0.1251
6	取る	get	23	502	1,262	0.3978
7	中	inside	14	399	876	0.4555
8	ばかり*	only	5	398	1,078	0.3692
9	様	appearance	16	391	1,805	0.2166
10	言う	say	7	385	6,302	0.0611

Table 11: Top 10 words with frequent errors of the experiment with fine-grained senses

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.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

Jiaju Du, Fanchao Qi, and Maosong Sun. 2019. Using bert for word sense disambiguation. *arXiv preprint arXiv:1909.08358*.

Ignacio Iacobacci, Mohammad Taher Pilehvar, and Roberto Navigli. 2016. Embeddings for word sense disambiguation: An evaluation study. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 897–907, Berlin, Germany. Association for Computational Linguistics.

Phillip Keung, Yichao Lu, Julian Salazar, and Vikas Bhardwaj. 2020. Don't use English dev: On the zero-shot cross-lingual evaluation of contextual embeddings. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 549–554, Online. Association for Computational Linguistics.

Kanako Komiya, Nagi Oki, and Masayuki Asahara. 2022a. Word sense disambiguation of corpus of historical Japanese using Japanese BERT trained with contemporary texts. In *Proceedings of the 36th Pacific Asia Conference on Language, Information and Computation*, pages 438–446, Manila, Philippines. De La Salle University.

Kanako Komiya, Aya Tanabe, and Hiroyuki Shinnou. 2022b. Diachronic domain adaptation of word sense disambiguation for corpus of historical japanese using word embeddings. *NINJAL Research Papers*, 23:59–73.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

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.

National Institute for Japanese Language and Linguistics. 1964. *Word List by Semantic Principles*. Shuuei Shuppan, In Japanese.

Roberto Navigli, Kenneth C. Litkowski, and Orin Hargraves. 2007. SemEval-2007 task 07: Coarse-grained English all-words task. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 30–35, Prague, Czech Republic. Association for Computational Linguistics.

Alessandro Raganato, Claudio Delli Bovi, and Roberto Navigli. 2017. Neural sequence learning models for word sense disambiguation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1156–1167, Copenhagen, Denmark. Association for Computational Linguistics.

Hiroyuki Shinnou, Kanako Komiya, Minoru Sasaki, and Shinsuke Mori. 2017. Japanese all-words wsd system using the kyoto text analysis toolkit. In *Proceedings of the 31st Pacific Asia Conference on Language, Information and Computation*, pages 392–399.

Rui Suzuki, Kanako Komiya, Masayuki Asahara, Minoru Sasaki, and Hiroyuki Shinnou. 2019. Unsupervised all-words wsd using synonyms and embeddings. *Journal of Natural Language Processing*, 26(2):361–379.