Word Sense Disambiguation of Corpus of Historical Japanese Using Japanese BERT Trained with Contemporary Texts

Kanako Komiya
Tokyo University of Agriculture and Technology
2-24-16 Naka-cho Koganei
Tokyo Japan 184-8588
kkomiya@go.tuat.ac.jp

Nagi Oki
Tokyo University of Agriculture and Technology
2-24-16 Naka-cho Koganei
Tokyo Japan 184-8588
s182739s@st.go.tuat.ac.jp

Masayuki Asahara
National Institute for Japanese Language and Linguistics
10-2 Midoricho Tachikawa
Tokyo Japan 190-0014
masayu-a@ninjal.ac.jp

Abstract

Diachronic adaptation for word sense disambiguation (WSD) in Historical Japanese is known as a difficult problem because the most frequent sense (MFS) baseline is hard to beat. However, this paper reports that the model using BERT trained with contemporary texts significantly outperforms the MFS baseline. We also showed the effectiveness of multitask learning of WSD and document classification. We conducted the experiments using two sets of a sense-tagged corpus, Corpus of Historical Japanese sense-tagged with Word List by Semantic Principles: the datasets developed until 2019 and 2022. Finally, we discuss the reason why the diachronic adaptation for WSD of historical Japanese using BERT trained with contemporary Japanese is effective.

Introduction

Word sense disambiguation (WSD) involves the task of identifying the senses of words in documents. There have been a number of studies on WSD of Contemporary Japanese using sense-tagged corpora. However, due to the limitation of the sense-tagged corpora, it was difficult to achieve high performance for WSD of historical Japanese. To alleviate this problem, diachronic adaptation using contemporary Japanese has been tried for WSD of historical Japanese. However, the prior work shows that the most frequent sense (MFS) baseline is hard to beat for conventional methods that used examples from contemporary corpora and/or word embeddings trained with contemporary texts in addition to historical texts (Tanabe, 2020).

Meanwhile, Bidirectional Encoder Representations from Transformers (BERT) Devlin et al., (2019) substantially improved the state of the art of tasks of natural language processing, including WSD (Blevins and Zettlemoyer, 2020; Loureiro and Camacho-Collados, 2020). First, in this paper, we will show that WSD of historical Japanese using BERT significantly outperforms the MFS baseline and conventional methods. As BERT is trained mostly with contemporary Japanese texts (Japanese Wikipedia), WSD of historical Japanese using BERT is considered as a form of diachronic adaptation. Next, we tried multitask learning of

WSD and document classification using a sense-tagged corpus, Corpus of Historical Japanese (CHJ) [1](see Section 3).

We conducted the experiments using two sets of CHJ, the datasets developed until 2019 and 2022 (see Sections 4 and 5). Finally, we discuss why BERT is effective for WSD of historical Japanese in Sections 6 and 7.

The contributions of this paper are listed as follows:

(1) We show that the diachronic adaptation using BERT trained with contemporary Japanese substantially outperformed the MFS baseline and conventional methods for WSD of historical Japanese,

(2) We show that multitask learning of WSD and document classification of the text where the target word of WSD was taken from is effective when large amounts of training data was used, and

(3) We discuss what kind of information contributed to the diachronic adaptation for WSD of historical Japanese.

Related Work

WSD has two categories: lexical sample task and all-words WSD. Lexical sample task targets frequent words in a dataset (Iacobacci et al., 2016; Okumura et al., 2010; Komiya and Okumura, 2011) and all-word WSD disambiguates all words in a corpus (Raganato et al., 2017a; Shinnou et al., 2017; Iacobacci et al., 2016; Suzuki et al., 2018; Raganato et al., 2017b; Blevins and Zettlemoyer, 2020; Loureiro and Camacho-Collados, 2020). There have been a number of studies on WSD of contemporary Japanese of both categories. This paper focuses on the lexical sample task.

In addition, there have been some studies on historical Japanese texts. Hoshino et al. (2014) proposed translating historical Japanese to contemporary Japanese using a statistical machine translation system trained with a corpus obtained by their method using sentence alignment. Takaku et al. (2020) employed neural machine translation for translation from historical Japanese to contemporary Japanese. They used word embeddings diachronically fine-tuned with historical corpora, including word embeddings gradually fine-tuned in the order of time, which is proposed in Kim et al. (2014), for the input to their system and showed the fine-tuned word embeddings improved the translation performances.

Tanabe (2020) used the diachronically fine-tuned word embeddings for the WSD task including those trained following methods used by Takaku et al. (2020). According to (Daumé III et al., 2010; Daumé III, 2007), there are three types of approaches for domain adaptation depending on the information to be learned, namely, supervised, semi-supervised and unsupervised approaches. Tanabe (2020) used not only sense-tagged corpora but also unlabeled texts for diachronic adaptation, in three scenarios including all three types of domain adaptation approaches.

Related to WSD of historical Japanese, Tanabe et al. (2018) proposed a system to classify the word senses of words in a Japanese historical corpus to determine the word senses that are not listed in a dictionary of contemporary Japanese. However, they did not perform the WSD of historical Japanese itself.

This research is also related to the methods to capture the change of meanings. Kulkarni et al. (2015), Hamilton et al. (2016b), and Hamilton et al. (2016a) have shown the effectiveness of distributional semantics for this task. Kobayashi et al. (2021) used the BERT model and Aida et al. (2021) used PMI and SVD joint learning to capture the change of meaning of modern and contemporary Japanese.

Diachronic Adaptation Using BERT Trained with Contemporary Japanese Texts

As mentioned above, WSD of historical Japanese using BERT trained with contemporary Japanese texts is considered as a form of diachronic adaptation. We show that the method using the BERT model outperforms the MFS baseline and the conventional methods using word embeddings proposed by Tanabe (2020).

In addition, we attempted multitask learning of WSD and document classification. For the multitask learning of WSD and document classification, we simultaneously predicted not only word senses but also the literature the input

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

sentence was taken from. The motivation behind this method is to capture the diversity of the periods when each literature was written. We process all the historical Japanese texts at one time but the word sense in very old literature, say the work written in the 900s, should be different from that in relatively new literature like the work written in the 1600s. We also anticipated that the frequent senses vary depending on the literature work the input sentence was taken from.

Data

We used Corpus of Historical Japanese sense-tagged with Word List by Semantic Principles (CHJ-WLSP) (Asahara et al., 2022). Word List by Semantic Principles (WLSP) (National Institute for Japanese Language and Linguistics, 1964) is a Japanese thesaurus. In the WLSP, the article numbers or concept numbers indicate shared synonyms. In the WLSP thesaurus, words are classified and organized by their meanings. We can use the article numbers in WLSP with words as word senses. For example, the word “犬” (inu, meaning spy or dog) has two records in the WLSP, and therefore has two article numbers, 1.2410 and 1.5501, indicating that the word is polysemous. We can also use the historical version of WLSP (Miyajima et al., 2014).

We conducted the experiments using two sets of CHJ-WLSP, the datasets developed until 2019 and 2022. First, for comparison, we used the same data as (Tanabe, 2020), CHJ-WLSP developed until 2019. We refer to this data as CHJ-WLSP 2019. The literature was 5 works, that is, Taketori-monogatari (The Tale of the Bamboo Cutter), Tosa Nikki (Tosa Diary), Hōjōki (Square-jō record), Tsurezuregusa (Essays in Idleness), and Toraakira-bon Kyogen. Following (Tanabe, 2020), we used 58 words for the target words of WSD. They were selected because they appeared 50 times or more in CHJ and the Balanced Corpus of Contemporary Written Japanese (BCCWJ) (Maekawa et al., 2014).

For the second experiments, we used data from 10 works, CHJ-WLSP developed until 2022 for this experiment. We refer to this data as CHJ-WLSP 2022. The additional literature was Konjaku Monogatarishū (Anthology of Tales from the Past), Jikkinsyo, Uji Shūi Monogatari, Taiyo magazine, and Kokutei textbook. We used 33 words, which are the words appeared more than 1,000 times in CHJ, for the target words of WSD. Table 1 displays the book title, number of word tokens, period, style of sub-corpora from CHJ we used for the experiments. Table 2 shown as an appendix lists the words, pronunciation, translation, and developed year of the data (2019 or 2022). Both in the table means the words are used for both experiments of CHJ-WLSP 2019 and CHJ-WLSP 2022. The translations shown in the table are just examples because the words are polysemous.

Book Title	Word Tokens	Period	Style
Taketori Monogatari (The Tale of the Bamboo Cutter)	12,757	Around 900	Fictional prose narrative
Tosa Nikki (Tosa Diary)	8,208	934	Poetic diary
Hōjōki (Square-jō record)	5,402	1212	Essay
Tsurezuregusa (Essays in Idleness)	40,834	1336	Essay
Toraakira-bon Kyogen	5,448	1642	Kyogen (Traditional theater)

Konjaku Monogatarishū (Anthology of Tales from the Past)	175,598	1100
Uji Shūi Monogatari	120,705	1220
Jikkinsyo	90,177	1252
Taiyo Magazine	46,394	1895-1
Kokutei Textbook	154,955	1910

Table 1 Book Titles, Number of Word Tokens, Period, Style of Books in CHJ of 5 works (CHJ-WLSP 2019) and 10 works (CHJ-WLSP 2022)

Experiments

We used bert-base-japanese-whole-word-masking, Japanese BERT mostly trained with contemporary Japanese texts (Japanese Wikipedia)[2], from transformers library.

For the simple BERT model, we used fine-tuning of BERT model added one layer. The input of the final layer is the output vector of the BERT model for the target word of WSD. We used softmax and cross entropy loss for the last layer. Stochastic Gradient Descent was used for the optimization function.

For the multitask learning, two final layers were added in parallel to the BERT model. One is for WSD, and another is for document classification. The input of the final layers is the same for the two final layers: the output vector of the BERT model for the target word of WSD.

Because we adopted the lexical sample task, a model is trained for each target word type of WSD. In other words, we prepared 58 models or 33 models for each method. The input of the system is sentence-based, and an example includes a target word token. Because of the length limitation of the BERT input, a sentence beyond 512 tokens is shorted to 512 tokens. In addition, if the target word of WSD was appeared in the omitted part of the sentence, we did not use the example for the experiments.

Mostly, we used Japanese period marks for dividing the sentences, but only for Toraakira-bon Kyoken, we used blank marks and boundary marks in CHJ to divide the sentences because it included no period marks. Tables 3 summarizes the maximum, minimum, and average number of data points for each target word of WSD.

	CHJ-WLSP 2019	CHJ-WLSP 2022
Max	419	7,072
Min	50	1,009
Avg.	167.83	2,255.15

Tables 3 The maximum, minimum, and average number of data points for each WSD target word

For fair comparison with Tanabe (2020) in terms of the number of training data points, we conducted experiments without development data using CHJ-WLSP 2019. Here, the ratio of the training and test data was set to 4:1. The data split was performed using random sampling. We set epoch number as 20 and learning rate as 0.00005 according to the preliminary experiments. The results are an average of three trials.

In addition, for CHJ-WLSP 2019 and 2022, we used grid search for the hyper-parameters, i.e., the epoch number and learning rate using five-fold cross validation with development data. The options of the epoch number were from 1 to 30 and those of the learning rate were 0.00001, 0.0001, and 0.001. Here, the ratio of the training, test, and development data was set to 3:1:1. Please note that, when we performed cross validation, because we used a development set and Tanabe (2020) did not, the amount of training data of the BERT model is smaller than that of the model of the prior work.

Table 4 shows the number of training data points according to the experiment. Random sampling 2019 in the table means data split with random sampling for CHJ-WLSP 2019, and cross validation 2019 and 2022 indicate the cross validation using CHJ-WLSP 2019 and CHJ-WLSP 2022.

Experiments	Num of data
Random sampling 2019	134.26
Cross validation 2019	100.70
Cross validation 2022	1,353.09

Table 4 the number of test data points according to the experiment

1.1 Baseline

We used the method of Tanabe (2020) for the baseline. The best method in Tanabe (2020) is the method using fine-tuning contemporary features with a historical corpus in the Target Only scenario, i.e., the scenario where no example from contemporary corpus was used. She used NWJC2VEC, the word embeddings generated from the NWJC-2014-4Q dataset (Asahara et al., 2014) as pretrained word embeddings of contemporary Japanese texts. She used plain texts of CHJ to fine-tune word embeddings. She also used BCCWJ for fine-tuning and BCCWJ-WLPS for a sense-tagged contemporary corpus. BCCWJ is tokenized with contemporary UniDic (Den et al., 2010) and CHJ is tokenized with historical UniDic (Ogiso et al., 2012).

Tanabe (2020) used a scikit-learn library of support vector machine using NWJC2VEC fine-

[2] https://huggingface.co/cl-tohoku/bert-base-japanese-whole-word-masking

tuned with CHJ. When generating the word embeddings of historical texts for a comparison, she used word2vec and the dimensionality was set to 200 and the window size was set to 2. The other parameters were the same as the default settings of the Gensim toolkit. She used five-fold cross validation without a development set. Please note that our MFS of the test data is slightly different from that of Tanabe (2020). We believe that is because the data split of cross validation is different.

Results

Table 5 shows the WSD accuracies of the models trained with CHJ-WLSP 2019. Hereinafter, Micro and Macro in the tables are micro- and macro averaged accuracies.

Model	Micro	Macro
Simple BERT model	**77.50%**	**72.82%**
Multitask learning	77.24%	72.55%
MFS of random sample	73.79%	69.81%
Tanabe (2020)	74.83%	70.80%
MFS of Tanabe (2020)	75.54%	70.00%

Table 5 WSD accuracies of the models trained with CHJ-WLSP 2019

This table shows that we substantially outperformed the MFS baseline. In addition, although our MFS of the test data is lower than that of the prior work (They are 75.54% and 73.79% when the micro-averaged MFS is compared), our BERT model significantly outperforms the result of the prior work according to the chi square test. The level of the significance in the test was 0.01. Additionally, the difference between our MFS and the BERT model was also significant.

In addition, Table 6 displays the WSD accuracies of the models trained with CHJ-WLSP 2019 using cross validation. Tables 5 and 6 indicate that multitask learning of WSD and document classification was not effective for CHJ-WLSP 2019. The differences between the simple BERT model and multitask learning were not significant. However, the two models using BERT outperformed the MFS baseline again.

Model	Micro	Macro
Simple BERT model	**76.85%**	**69.81%**
Multitask learning	76.70%	69.64%
Our MFS	74.20%	69.09%

Table 6 WSD accuracies of the models trained with CHJ-WLSP 2019 using cross validation

Table 7 shows the WSD accuracies of the models trained with CHJ-WLSP 2022 using cross validation. The table shows that multitask learning of WSD and document classification surpassed the simple BERT model. It was significant according to the chi square test. The level of the significance in the test was 0.01. In addition, the differences between the MFS and two BERT based models were also significant.

Model	Micro	Macro
Simple BERT model	84.68%	84.25%
Multitask learning	**85.17%**	**84.45%**
MFS	78.29%	78.20%

Table 7 WSD accuracies of the models trained with CHJ-WLSP 2022 using cross validation

Discussion

According to the three tables in Section 6, we can see that BERT trained from contemporary Japanese texts is effective for WSD of historical Japanese texts.

We believe that the reason why the BERT model outperformed the prior work that used word2vec is not the amount of training data that trained the BERT model or word2vec, because Tanabe (2020) used NWJC2VEC, the word embeddings generated from the NWJC-2014-4Q dataset, which included more than a billion sentences.

As Japanese BERT is trained with Japanese Wikipedia, which consists of approximately 17M sentences, the training data cannot be the reason. Therefore, even if we cannot know the concrete information that provided the improvement of the WSD performances, the network architecture could be the reason. In addition, as some studies reported BERT can capture various language information including syntactic structures (Jawahar et al. 2019), this property of BERT could be the reason of the success of the diachronic adaptation of historical Japanese.

We feared that the unknown words of the BERT model adversely affected the WSD accuracies, but they were at least not serious. Table 8 shows the percentage of unknown word tokens ([UNK]) and subword tokens that begins with a sharp mark (#) in the input tokens of the system. It means the tokens of input sentences are counted, excluding tokens beyond 512 tokens per an input. [UNK] and subword # in the table mean the percentage of unknown word tokens and that of the subword tokens that begin with a sharp mark, respectively.

Data	[UNK]	Subword #
CHJ-WLSP 2019	0.60%	11.71%
CHJ-WLSP 2022	1.52%	9.58%

Table 8 The percentage of unknown word tokens ([UNK]) and subword tokens that begin with a sharp mark (#)

According to Table 8, we can see that the unknown word tokens are very rare in CHJ-WLSP 2019 and CHJ-WLSP 2022 Because of subword tokens, most of input tokens could be interpreted by the BERT model trained with contemporary texts.

Next, let us discuss the multitask learning of WSD and document classification. When we used CHJ-WLSP 2019, this method did not work but when we used CHJ-WSLP 2022, it was significantly effective. The first reason to explain this fact that we can think of is the amount of the training data points, as shown in Table 4. The second reason could be the difference of the variety of the literature of two datasets. Tables 9 and 10 display the average number of test data in CHJ-WLSP 2019 and CHJ-WLSP 2022, respectively.

Literature	Num of test data
Taketori Monogatari	350.6
Tosa Nikki	264.8
Hōjōki	134.8
Tsurezuregusa	1,104
Toraakira-bon Kyogen	92.6

Table 9 Average number of test data in CHJ-WLSP 2019 disaggregated by the literature

Literature	Num of test data
Taketori Monogatari	312.2
Tosa Nikki	224.4
Hōjōki	120.6
Tsurezuregusa	1,011.2
Toraakira-bon Kyogen	22.8
Konjaku Monogatarishū	4,796
Uji Shūi Monogatari	3,167
Jikkinsyo	1,848.8
Taiyo Magazine	1,013.2
Kokutei Textbook	2,375.2

Table 10 Average number of test data in CHJ-WLSP 2022 disaggregated by the literature

According to Table 9, in CHJ-WLSP 2019, more than half of the test data came from only one literature, Tsurezuregusa. On the other hand, as shown in Table 10, the data balance is more balanced compared to CHJ-WLSP 2019.

Moreover, Tables 11 and 12 show the accuracies of document classification of CHJ-WLSP 2019 and CHJ-WLSP 2022. The improvement of the accuracy, that is the difference between accuracies of multitask learning and most frequent document in CHJ-WSLP 2022 was considerably higher than that in CHJ-WSLP 2019. This fact indicates that the document classification task was learned better using CHJ-WLSP 2022.

Model	Micro	Macro
Multitask learning	66.17%	63.95%
Most frequent document	58.15%	55.69%

Table 11 Accuracies of document classification of CHJ-WLSP 2019

Model	Micro	Macro
Multitask learning	63.90%	69.22%
Most frequent document	35.08%	36.53%

Table 12 Accuracies of document classification of CHJ-WLSP 2019

Finally, comparing the experiments using CHJ-WLSP 2019 and CHJ-WLSP 2022, we can see that the obvious factor to improve the WSD accuracies is the amount of in-domain labeled data. The amount of training data points of CHJ-WLSP 2022 was approximately 13 times more than that of CHJ-WLSP 2019 as shown in Table 4. This research showed that, when we use more than 1,000 data points for training data, the WSD accuracy is around 85%, which is considerably higher than the MFS baseline.

In the future, we plan to develop an all-words WSD system for historical Japanese texts. In addition, we plan to explore properties other than data amount to improve the performance of WSD.

Conclusions

We reported that BERT trained with contemporary Japanese texts considerably improved the WSD accuracies of historical Japanese texts. This method can be considered as a form of diachronic adaptation. Because the amount of training data of BERT model cannot account for the improvement of WSD accuracies, the network architecture of the BERT itself could be the reason why diachronic adaptation using BERT trained with contemporary text worked. In addition, because of the subword tokens, unknown word tokens are rare in historical texts. We performed experiments using two sets of a corpus, which are CHJ-WLSP 2019 and CHJ-WLSP 2022. We also showed the effectiveness of the multitask learning of WSD and classification of the sentence that was included the target word taken from, when we used CHJ-WLSP 2022. We also showed that the WSD accuracies substantially improved as the in-domain labeled data for training increased.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Number 17H00917, 17KK0002, 18K11421 and 22K12145, and projects of the Center for Corpus Development, National Institute for Japanese Language and Linguistics.

References

Taichi Aida, Mamoru Komachi, Toshinobu Ogiso, iroya Takamura, and Daichi Mochihashi. 2021. Tsujiteki na tango no imihenka wo toraeru tango bunsanhyougen no ketsugogakusyu. [joint learning of word embeddings capturing diachronic changes of meanings of words]. In Proceedings of the NLP2021, (In Japanese), pages 712–717.

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, (To appear)

Masayuki Asahara, Kikuo Maekawa, Mizuho Imada, Sachi Kato, and Hikari Konishi. 2014. Archiv ing and analysing techniques of the ultra-large-scale web-based corpus project of ninjal. Alexandria: The journal of national and international library and in formation issue, 25(1-2):129–148.

Terra Blevins and Luke Zettlemoyer. 2020. Moving down the long tail of word sense disambiguation with gloss informed bi-encoders. Proceedings of ACL.

Hal Daumé III. 2007. Frustratingly easy domain adaptation. In Proceedings of ACL 2007, pages 256–263.

Hal Daumé III, Abhishek Kumar, and Avishek Saha. 2010. Frustratingly easy semi-supervised domain adaptation. In Proceedings of the 2010 Workshop on Domain Adaptation for Natural Language Processing, ACL 2010, pages 23–59.

Yasuharu Den, Junpei Nakamura, Toshinobu Ogiso, Hideki Ogura. 2008. A Proper Approach to Japanese Morphological Analysis: Dictionary, Model, and Evaluation, In Proceedings of the sixth international conference on Language Resources and Evaluation (LREC 2008), pages.1019-1024.

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

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016a. Cultural shift or linguistic drift? comparing two computational measures of semantic change. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 2116–2121.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016b. Diachronic word embeddings reveal statistical laws of semantic change. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1489–1501.

Sho Hoshino, Yusuke Miyao, Shunsuke Ohashi, Akiko Aizawa, and Hikaru Yokono. 2014. Machine translation from historical Japanese to contemporary Japanese using parallel corpus. In Proceedings of the NLP2014, (In Japanese), pages 816–819.

Ignacio Iacobacci, Mohammad Taher Pilehvar, and Roberto Navigli. 2016. Embeddings for word sense disambiguation: An evaluation study. In Proceedings of ACL 2016, pages 897—-907.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In Proceedings of the ACL 2019, pages 3651–3657.

Yoon Kim, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde, and Slav Petrov. 2014. Temporal analysis of language through neural language models. In Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science, pages 61–65.

Kazuma Kobayashi, Taichi Aida, and Mamoru Komachi. 2021. Bert wo shiyouu shita nihongo no tango no tsuujiteki na imihenka no bunseki. [analysis of diachronic changes in meanings of Japanese words using bert]. In Proceedings of the NLP2021, (In Japanese), pages 952–956.

Kanako Komiya and Manabu Okumura. 2011. Automatic determination of a domain adaptation method for word sense disambiguation using decision tree learning. In Proceedings of IJCNLP 2011, pages 1107–1115.

Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2015. Statistically significant detection of linguistic change. In Proceedings of the 24th International Conference on World Wide Web, pages 625–635.

Daniel Loureiro and Jose Camacho-Collados. 2020. Don't neglect the obvious: On the role of unambiguous words in word sense disambiguation. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 544 pages 3514–3520.

Kikuo Maekawa, Makoto Yamazaki, Takehiko Maruyama, Masaya Yamaguchi, Hideki Ogura, Wakako Kashino, Toshinobu Ogiso, Hanae Koiso, and Yasuharu Den. 2010. Design, compilation, and preliminary analyses of Balanced Corpus of Contemporary Written Japanese. In Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010), pages 1483–1486.

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(2):345–371.

Tatsuo Miyajima, Hisao Ishii, Seiya Abe, and Tai Suzuki. 2014. Nihon koten taisho bunrui goi hyo [Word list by semantic principles refereeing to Japanese classics]. Kasama Shoin, In Japanese.

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

Toshinobu Ogiso, Mamoru Komachi, Yasuharu Den, and Yuji Matsumoto. 2012. Unidic for early middle Japanese: a dictionary for morphological analysis of classical Japanese. In Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012), pages 911–915.

Manabu Okumura, Kiyoaki Shirai, Kanako Komiya, and Hikaru Yokono. 2010. Semeval-2010 task: Japanese wsd. In Proceedings of the SemEval-2010, ACL2010, pages 69–74.

Alessandro Raganato, Claudio Delli Bovi, and Roberto Navigli. 2017a. Neural sequence learning models for word sense disambiguation. In Proceedings of EMNLP 2017, pages 1156–1167.

Alessandro Raganato, Jose Camacho-Collados, and Roberto Navigli. 2017b. Semeval-2007 task 07: Coarse-grained english all-words task. In Proceedings of EACL 2017, pages 99–110.

Hiroyuki Shinnou, Kanako Komiya, Minoru Sasaki, and Shinsuke Mori. 2017. Japanese all-words wsd system using the kyoto text analysis toolkit. In Proceedings of PACLIC 2017, pages 392–399.

Rui Suzuki, Kanako Komiya, Masayuki Asahara, Minoru Sasaki, and Hiroyuki Shinnou. 2018. All words word sense disambiguation using concept embeddings. Proceedings of LREC 2018, pages 1006–1011.

Masashi Takaku, Tosho Hirasawa, Mamoru Komachi, and Kanako Komiya. 2020. Neural machine translation from historical japanese to contemporary japanese using diachronically domain-adapted word embeddings. In Proceedings of the 34th Pacific Asia Conference on Language, Information and Computation (PACLIC 2020), page no. 22.

Aya Tanabe. 2020. Domain adaptation of word sense disambiguation of corpus of historical japanese using contemporary japanese corpus. Master theses of Graduate Schools of Science and Engineering, Ibaraki University of 2019 academic year (in Japanese).

Aya Tanabe, Kanako Komiya, Masayuki Asahara, Minoru Sasaki, and Hiroyuki Shinnou. 2018. Detecting unknown word senses in contemporary japanese dictionary from corpus of historical japanese. In Proceedings of the 8th Conference of Japanese Association for Digital Humanities (JADH 2018), pages 169–170.

Appendix A. Target Words

Word	Pronunciation	Translation	Year
為る	Suru	do	Both
一	Ichi	one	Both
居る	Iru	stay	Both
見る	Miru	look	Both
言う	Iu	say	Both
行く	Iku	go	Both
此れ	Kore	this	Both
今	Ima	now	Both

思う	Omou	think	Both
事	Koto	thing	Both
時	Toki	time	Both
取る	Toru	get	Both
所	Tokoro	place	Both
心	Kokoro	heart	Both
人	Hito	human	Both
成る	Naru	become	Both
知る	Shiru	know	Both
日	Hi	day	Both
物	Mono	object	Both
聞く	Kiku	listen	Both
又	Mata	and	Both
有る	Aru	there is	Both
様	Sama	appearance	Both
来る	Kuru	come	Both
或る	Aru	a certain	2019
下	Shita	under	2019
何	Nani	what	2019
家	Ie	house	2019
皆	Mina	every	2019
間	Aida	between	2019
共	Tomo	together	2019
月	Tsuki	moon	2019
見える	Mieru	see	2019
後	Ato	after	2019
国	Kuni	country	2019
作る	Tsukuru	make	2019
持つ	Motsu	hold	2019
書く	Kaku	write	2019
女	Onna	woman	2019
上	Ue	up	2019
身	Mi	body	2019
他	Hoka	other	2019
男	Otoko	man	2019
置く	Oku	put on	2019
中	Naka	inside	2019
道	Michi	way	2019
読む	Yomu	read	2019
内	Uchi	inside	2019
入る	Hairu	enter	2019
年	Toshi	year	2019
彼	Kare	he	2019
付ける	Tsukeru	put onand	2019
返る	Kaeru	return	2019
方	Hou	direction	2019
万	Man	ten thousand	2019
唯	Tada	only	2019
立つ	Tatsu	stand	2019
良い	Yoi	good	2019
我	Ware	I	2022
者	Mono	person	2022
出でる	Ideru	go out	2022
申す	Mousu	say	2022
是	Kore	this	2022
然る	Saru	like that	2022
其	Sore	that	2022
程	Hodo	level	2022
無い	Nai	no	2022

Table 2 Target words of WSD in CHJ-WLSP 2019 and CHJ-WLSP 2022