Diachronic Domain Adaptation of Word Sense Disambiguation in Corpus of Historical Japanese Using Word Embeddings

国立国語研究所論集 (NINJAL Research Papers) 23: 59–73 (2022)
ISSN 2186-1358

Diachronic Domain Adaptation of Word Sense Disambiguation in Corpus of Historical Japanese Using Word Embeddings

KOMIYA Kanako[a] TANABE Aya[b] SHINNOU Hiroyuki[c]

[a]Tokyo University of Agriculture and Technology / Project Collaborator, NINJAL
[b]Ibaraki University
[c]Ibaraki University / Project Collaborator, NINJAL

Abstract
There have been many studies on word sense disambiguation (WSD) in contemporary Japanese. However, it is difficult to achieve high performance of WSD in historical Japanese because of the lack of sense-tagged corpora. Therefore, diachronic adaptation using contemporary Japanese could be a solution. We investigated the effectiveness of the fine-tuning of word embeddings for WSD in historical Japanese. A variety of fine-tuning scenarios are examined, including the case where the word embeddings of contemporary Japanese (NWJC2vec) are fine-tuned with historical Japanese and the case where the word embeddings trained with historical Japanese are fine-tuned with contemporary Japanese. Moreover, when NWJC2vec was fine-tuned with a historical corpus, the case where the word embeddings were gradually fine-tuned in the order of time was also tested. The word embeddings of two words before and after the target word are used as the features for the support vector machine, which is a classifier of WSD. The following three scenarios are compared: (1) all the examples from the contemporary Japanese corpus and 80% examples from the historical corpus are used as the training data for the test of the remaining 20% examples from the historical corpus, (2) 5-fold cross validation of the examples of the historical Japanese corpus, and (3) all the examples from the contemporary corpus are used as the training data for test examples from the historical corpus. The best accuracy was achieved when we used word embeddings trained from a historical corpus and fine-tuned with a contemporary corpus in the 5-fold cross validation scenario.*

Keywords: domain adaptation, historical corpus, diachronic adaptation, word sense disambiguation, word embeddings

1. Introduction

Word sense disambiguation (WSD) involves identifying the senses of words in documents. Usually, the lexical sample task of WSD is solved with supervised learning using a large amount of training data. However, when the domain of the test data is different from that of the training data, the performance of WSD would be lower. In addition, it is impractical to prepare a large corpus of every domain because the annotation of the word senses is time-consuming. Therefore, much research has been carried out on WSD with a special focus on domain adaptation (Komiya

* This manuscript represents the research results of the NINJAL collaborative research project 'Basic Research on Corpus Annotation—Extension, Integration and Machine-aided Approaches' (project leader: Masayuki Asahara) and 'The Construction of Diachronic Corpora and New Developments in Research on the History of Japanese' (project leader: Toshinobu Ogiso). This work was supported by JSPS KAKENHI Grant Number 17H00917. This work was presented in the 230th NINJAL Salon on 2nd November 2021.

and Okumura 2011, Komiya et al. 2018, Yaginuma et al. 2018). Generally, the domain differences treated in the research were the differences among the topics of documents or styles of writing, and the experiments were conducted using contemporary text corpora. However, the WSD of historical texts has the same problem: the WSD model exhibits poorer performance because of the lack of historical corpora. Therefore, we propose the use of a domain adaptation method for the WSD of historical texts using a contemporary corpus.

There are three types of approaches for domain adaptation depending on the information to be learned: supervised, semi-supervised, and unsupervised approaches (Daumé 2007, Daumé et al. 2010). In a supervised approach, a model is trained with labeled data in both the source and target domains. In a semi-supervised approach, it is developed from the labeled data of the source and target domains, and from the unlabeled data of the target domain. Finally, an unsupervised approach is developed from labeled source data and unlabeled target data. We performed experiments in these three approaches using sense-tagged data from historical and contemporary corpora. In addition, we used not only unlabeled target data but also unlabeled source data to create word embeddings for the features of the source and target data. Moreover, we fine-tuned the features using the source and target data for diachronic adaptation. In this study, we investigated the best scenario for diachronic adaptation of Japanese WSD in historical texts using word embeddings.

2. Related work

WSD can be categorized into two groups: lexical sample task and all-words WSD. The lexical sample task targets frequent words in a dataset (Okumura et al. 2010, Komiya and Okumura 2011, Iacobacci et al. 2016), and all-words WSD disambiguates all words in a corpus (Iacobacci et al. 2016, Raganato et al. 2017a, Raganato et al. 2017b, Shinnou et al. 2017b, Suzuki et al. 2018). There have been a number of studies on WSD in contemporary Japanese in both groups.

In addition, much research has been conducted on domain adaptation of Japanese WSD such as Okumura et al. (2010), Komiya and Okumura (2011), Shinnou et al. (2017b), and Komiya et al. (2018). This research is the first diachronic domain adaptation of historical Japanese WSD using the ordinary domain adaptation method.

In addition, there have been some studies of 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 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 was also proposed in Kim et al. (2014), as the input to their system and showed that fine-tuned word embeddings improved the translation performance. Kim et al. (2014) automatically detected changes in language over time through a chronologically trained neural language model using diachronic fine-tuning. They obtained word embeddings specific to each year and demonstrated that some words had changed in their meanings. Based on their research, we believe that diachronically domain-adapted word embeddings can capture changes in language meanings over time. In the current study, we used diachronically fine-tuned word embeddings for the WSD task. Related to the 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

the dictionary of contemporary Japanese. However, they did not perform the WSD of historical Japanese itself. This research is also related to the methods used to capture the change in meanings. Kobayashi et al. (2021) used the BERT model (Devlin et al. 2019) and Aida et al. (2021) used PMI-SVD (Pointwise Mutual Information and Singular Value Decomposition) joint learning to capture the change in the meaning of modern and contemporary Japanese.

In recent years, the use of word embeddings—for example, via word2vec (Mikolov et al. 2013a, Mikolov et al. 2013b, Mikolov et al. 2013c)—has become a fundamental technology in natural language processing (NLP). Word embeddings are vector representations of meanings, which are calculated based on their contexts and used to examine similarities in the meaning of two individual language units. Komiya et al. (2018) carried out domain adaptation for contemporary texts by fine-tuning word embeddings. We used the same method for diachronic domain adaptation. As for the fine-tuning of word embeddings, Komiya and Shinnou (2018) investigated the parameters for effective fine-tuning of word embeddings using a small corpus. Schnabel et al. (2015) proposed tuning of word embeddings depending on each task. Shinnou et al. (2017c) showed that fine-tuning for each domain was effective even if the original word embeddings were trained with an immense amount of data.

3. Data

We used both sense-tagged corpora and plain text corpora for the experiments. Sense-tagged corpora were used as the training data for WSD. Plain text corpora were used for training historical word embeddings and fine-tuning for historical and contemporary word embeddings. We used pre-trained word embeddings (NWJC2vec) (Shinnou et al. 2017a) for the contemporary data.

3.1 Sense-tagged corpora

The main task of our research is WSD, which involves predicting the word senses in sentences according to a dictionary. We use sense-tagged corpora to generate the training examples for WSD models.

We used the Corpus of Historical Japanese (CHJ) (National Institute for Japanese Language and Linguistics 2021) and the Balanced Corpus of Contemporary Written Japanese (BCCWJ) (Maekawa et al. 2014) as Japanese historical and contemporary sense-tagged corpora, respectively. The sense-tagged part of BCCWJ is released as BCCWJ-WLSP. They were manually tokenized using the UniDic delimitation standard, a Japanese dictionary compiled by the National Institute for Japanese Language and Linguistics. BCCWJ is tokenized with contemporary UniDic (Maekawa et al. 2010) and CHJ is tokenized with historical UniDic (Ogiso et al. 2012).

Table 1 summarizes the domains, numbers of word tokens, periods, and styles of sub-corpora from the BCCWJ used for the experiments.

Table 1. Domain, number of word tokens, period, and style of sub-corpora used from BCCWJ

Domain	Number of Word Tokens	Period	Style
PB	111,983	2001–2005 Contemporary	Book
PN	117,543	2001–2005 Contemporary	Newspaper
PM	117,568	2001–2005 Contemporary	Magazine

Table 2 shows the example of a sentence piece from BCCWJ, "わが身が雑巾になり切れないような修行 (Ascetic training that makes you feel you cannot be like a cleaning rag.)" The original sentence was "わが身が雑巾になり切れないような修行では恥ずかしいと思いませんか。 (Would you not be ashamed if ascetic training makes you feel you cannot be like a cleaning rag?)"

Table 3 displays the book titles, English book titles, numbers of word tokens, periods, and styles of sub-corpora from CHJ used for the experiments.

Table 4 shows the example of a sentence piece from CHJ, "かつ消えかつ結びて，久しくとどまりたるためしなし (disappear and appear and there is no example that they remain for long)". Original sentence was "よどみに浮かぶうたかたは，かつ消えかつ結びて久しくとどまりたるためしなし。(Bubbles float across stagnancy and they appear and disappear and never remain for long.)"

Table 2. Example of records in BCCWJ

Orthographic Token	Pronunciation	Lemma	English Translation	Concept Number (WLSP)
わが	Waga	我が	My	3.1040
身	Mi	身	Body	1.5600
が	Ga	が	(Case particle)	(None)
雑巾	Zoukin	雑巾	Cleaning rag	1.4541
に	Ni	に	(Case particle)	(None)
成り	Nari	成る	Become	2.1500
切れ	Kire	切る	Accomplish	2.1571
ない	Nai	ない	Not	(None)
よう	You	様	Like	3.1130
な	Na	だ	(Auxiliary verb)	(None)
修行	Shugyo	修行	Ascetic training	1.3050

Table 3. Book titles, number of word tokens, period, and style of books in CHJ

Book Title	Word Tokens	Period	Style
Taketori Monogatari (The Tale of the Bamboo Cutter)	12,757	Heian (Around 900)	Fictional prose narrative
Tosa Nikki (Tosa Diary)	8,208	Heian (934)	Poetic diary
Hōjōki (Square-jō record)	5,402	Kamakura (1212)	Essay
Tsurezuregusa (Essays in Idleness)	40,834	Kamakura (1332)	Essay
Toraakira-bon Kyogen	5,448	Edo (1642)	Kyogen (Traditional Japanese comic theater)

Table 4. Example of records in CHJ

Orthographic Token	Pronunciation	Lemma	English Translation	Concept Number (WLSP)
かつ	Katsu	且つ	And	4.1110
消え	Kie	消える	Disappear	2.1250
かつ	Katsu	且つ	And	4.1110
結び	Musubi	結ぶ	Appear	2.1220
て	Te	て	(Conjunctive particle)	(None)
,	(None)	,	(Punctuation)	(None)
久しく	Hisashiku	久しい	Long	3.1600
とどまり	Todomari	留まる	Remain	2.1503
たる	Taru	たり	(Auxiliary verb)	(None)
ためし	Tameshi	例	Example	1.1100
なし	Nashi	無い	No	3.1200

We used the Word List by Semantic Principles (WLSP) (National Institute for Japanese Language, 2004), which is a Japanese thesaurus of contemporary words, as a contemporary Japanese dictionary. In WLSP, the article numbers or concept numbers indicate shared synonyms. The article numbers could be used as the word senses to generate the training examples. In the WLSP thesaurus, words are classified and organized according to their meanings. Each WLSP record contains the following fields: record ID number, lemma number, record type, class, division, section, article, article number (concept number), paragraph number, small paragraph number, word number, lemma (with explanatory note), lemma (without explanatory note), reading (pronunciation), and reverse reading. Each record has an article number, which represents four fields: class, division, section, and article. 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 use the article numbers in WLSP with words as word senses since we can treat a pair of concepts and words as word senses. We have BCCWJ and CHJ with the article numbers (Miyajima et al. 2014), which are word-sense-tagged corpora that are in their infancy (Asahara et al. 2018, Kato et al., 2018), and the same were used for the experiments.

3.2 Word embeddings

We used NWJC2vec for word embeddings in contemporary Japanese. These word embeddings were generated from the NWJC-2014-4Q dataset (Asahara et al. 2014), which is an enormous Japanese Web corpus, using the word2vec (Mikolov et al. 2013a, Mikolov et al. 2013b, Mikolov et al. 2013c) toolkit Gensim. Tables 5 and 6 present summary statistics for the NWJC-2014-4Q data and the parameters used to generate the word embeddings, respectively. Please note that although we used the BCCWJ for the sense-tagged corpus, we did not fine-tune NWJC2vec with the BCCWJ but directly used NWJC2vec as the word embeddings for contemporary Japanese.

Table 5. Statistics for the NWJC-2014-4Q_dataset

Number of URLs collected	83,992,556
Number of sentences (Some are overlapped)	3,885,889,575
Number of sentences (No overlapping)	1,463,142,939
Number of words (tokens)	25,836,947,421

Table 6. Parameters used to generate NWJC2vec

Description	Parameter	Value
CBOW or skip-gram	-cbow	1
Dimensionality	-size	200
Number of surrounding words	-window	8
Number of negative samples	-negative	25
Hierarchical softmax	-hs	0
Minimum sample threshold	-sample	1.00E-04
Number of iterations	-iter	15

We followed (Yaginuma et al. 2018) for the parameters for fine-tuning NWJC2vec and historical word embeddings (see Table 7).[1] The other parameters were set to default settings.

The Complete Collection of Japanese Classical Literature published by Shogakukan was used as a plain text corpus. The statistics for the historical corpus are presented in Table 8. When we created the historical word embeddings, the dimensionality was set to 200 or 300, and the window size was set to 2. The other parameters were the same as the default settings of the Gensim toolkit. We used only 200-dimensional word embeddings when we fine-tuned them due to memory limitations.

Table 7. Parameters used to fine-tune NWJC2vec and historical word embeddings

Description	Parameter	Value
CBOW or skip-gram	-cbow	1
Dimensionality	-unit	200
Number of surrounding words	-window	5
Number of negative samples	-negative	5
Batch size	-batchsize	1000
Number of iterations	-iter	1

Table 8. Statistics for the historical corpus

Period	Number of Sentences	Vocabulary Size	Num. of Words
Modern	22,485	25,584	544,293
Muromachi	12,640	14,931	386,101
Kamakura	35,020	29,062	933,190
Heian	59,744	29,520	1,543,102
Nara	4,832	6,013	112,094
Total	134,721	105,110	3,518,780

[1] First, we tried to use retrofitting tool in (Faruqui et al, 2015) but it cannot be used for Japanese.

4. Diachronic adaptation for WSD

We used fine-tuning of word embeddings for the diachronic adaptation of WSD. Word embeddings are vector representations of meanings that are calculated based on their context. Because they can be obtained from plain texts, domain adaptation can be carried out even if no tagged corpora are available. Additionally, fine-tuning was also carried out with only unlabeled corpora. It is an approach for transfer learning, in which an additional corpus is used to tune the learned word embeddings. We have four types of corpora: the sense-tagged corpora of historical and contemporary Japanese and unlabeled corpora of historical and contemporary Japanese. Therefore, both the unlabeled corpora of historical and contemporary Japanese could be used for both initial training and additional training of word embeddings.

In addition, we have some scenarios in which contemporary Japanese and historical Japanese are used as examples of WSD itself. In our study, we compared combinations of scenarios and features.

4.1 Scenarios

We tested three scenarios of diachronic domain adaptation. We used a contemporary corpus as the source data and a historical corpus as the target data.

Both scenario: For this scenario, we used both source and target data for training. Whole source data and 80% of target data were used for training, and 20% of the target data were used for testing. We used 5-fold cross validation.

Target Only scenario: For this scenario, we used only the target data for training. Eighty percent of the target data were used for training, and 20% of the target data were used for testing. We used 5-fold cross validation. Generally, this scenario is not a domain-adaptation scenario if no source data are used for training. However, we used the features generated using unlabeled source data. These methods are domain adaptation methods, even in Target Only scenario.

Source Only scenario: For this scenario, we used only the source data for training. Whole source data were used for training, and all test data were used for testing.

Figure 1 shows examples of three scenarios assuming that there were 150 examples of a contemporary Japanese corpus (BCCWJ) and 50 examples of a historical Japanese corpus (CHJ).

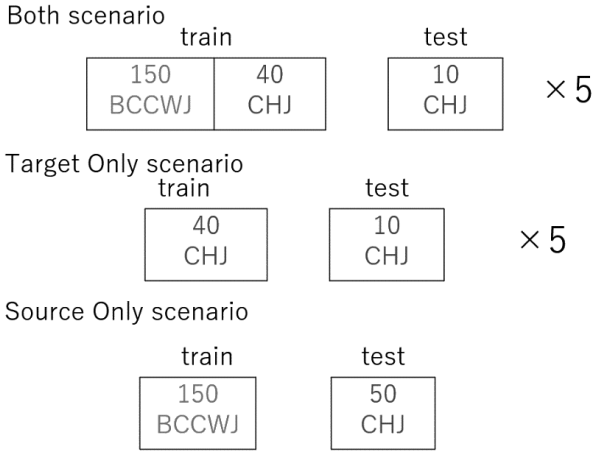

Figure 1. Examples of three scenarios

4.2 Features and fine-tuning

We tested five types of features: (1) historical features, (2) contemporary features, (3) historical features fine-tuned with contemporary corpus, (4) contemporary features fine-tuned with historical corpus, and (5) contemporary features fine-tuned in the order of time, for diachronic domain adaptation of historical Japanese. For word embeddings, the first two methods are baselines, and the last three methods are diachronic domain adaptations. However, at certain times the methods using **historical features** or **contemporary features** could be domain adaptation as we also used sense-tagged data.

Historical features: Word embeddings trained with CHJ were used for the features.

Contemporary features: NWJC2vec were used for the features.

FT historical features with contemporary corpus: Word embeddings trained with CHJ were fine-tuned with BCCWJ and used for the features.

FT contemporary features with historical corpus: NWJC2vec were fine-tuned with CHJ at one time and were used for the features.

FT contemporary features in the order of time: NWJC2vec fine-tuned with CHJ in the order of time, were used for the features.

4.3 Data and categories of domain adaptation approach

Table 9 summarizes the data used for the diachronic adaptation that we tested. S and T represent the source and target data of the sense-tagged corpus, and s and t denote the source and target data of the unlabeled corpus, respectively. Additionally, Table 10 shows the categories of domain adaptation approaches according to Daumé et al. (2010). S, Semi-s, and Uns are supervised, semi-supervised, and unsupervised approaches, respectively. Historical features in Both scenario

is a conventional semi-supervised approach, and historical features in Source Only scenario is an unsupervised approach. Historical features in Target Only approach and contemporary features in Source Only approach are not domain adaptation approaches because they do not use both source and target data. Contemporary features in Both scenario is a supervised approach, but also use an unlabeled source corpus. Fine-tuned features in Both scenario are semi-supervised approaches that use an unlabeled source corpus. Contemporary features and fine-tuned features in Target Only scenario cannot be categorized into conventional categories, but they are domain adaptation methods. The nearest approach can be semi-supervised. Fine-tuned features in Source Only scenario are unsupervised approaches that use an unlabeled source corpus.

Table 9. Data for the diachronic domain adaptation

Scenarios	Both	Target Only	Source Only
Historical features	S, T, t	T, t	S, t
Contemporary features	S, T, s	T, s	S, s
FT historical features with contemporary corpus	S, T, s, t	T, s, t	S, s, t
FT contemporary features with historical corpus	S, T, s, t	T, s, t	S, s, t
FT contemporary features in the order of time	S, T, s, t	T, s, t	S, s, t

Table 10. Categories of domain adaptation approaches according to Daumé et al. (2010)

Scenarios	Both	Target Only	Source Only
Historical features	Semi-s	No	Uns
Contemporary features	(S)	(Semi-s?)	No
FT historical features with contemporary corpus	(Semi-s)	(Semi-s?)	(Uns)
FT contemporary features with historical corpus	(Semi-s)	(Semi-s?)	(Uns)
FT contemporary features in the order of time	(Semi-s)	(Semi-s?)	(Uns)

5. Experiments

The target words of WSD are 58 words, which are 為る (Suru, do), 成る (Naru, become), 物 (Mono, object), 月 (Tsuki, moon), 方 (Hou, direction), 内 (Uchi, inside), 彼 (Kare, he), 見える (Mieru, see), 共 (Tomo, together), 身 (Mi, body), 居る (Iru, stay), 様 (Sama, appearance), 時 (Toki, time), 此れ (Kore, this), 万 (Man, ten thousand), 上 (Ue, up), 知る (Shiru, know), 他 (Hoka, other), 皆 (Mina, every), 或る (Aru, a certain), 一 (Ichi, one), 人 (Hito, human), 見る (Miru, look), 思う (Omou, think). 又 (Mata, and), 間 (Aida, between), 作る (Tsukuru, make), 女 (Onna, woman), 唯 (Tada, only), 読む (Yomu, read), 言う (Iu, say), 年 (Toshi, year), 行く (Iku, go), 良い (Yoi, good), 今 (Ima, now), 聞く (Kiku, listen), 国 (Kuni, country), 書く (Kaku, write), 道 (Michi, way), 返る (Kaeru, return), 有る (Aru, there is), 日 (Hi, day), 中 (Naka, inside), 所 (Tokoro, place), 家 (Ie, house), 取る (Toru, get), 置く (Oku, put on), 心 (Kokoro, heart), 立つ (Tatsu, stand), 事 (Koto, thing), 来る (Kuru, come), 何 (Nani, what), 後 (Ato, after), 持つ (Motsu, hold), 入る (Hairu, enter), 男 (Otoko, man), 付ける (Tsukeru, put on), and 下 (Shita, under). The translations shown here are candidates of multiple meanings. They were selected because they appeared 50 times or more in both CHJ and BCCWJ. The number of polysemous words that appeared 50 times or more in CHJ were 82. The total number of polysemous words was 981, of which 725 appeared in BCCWJ.

Table 11 shows the micro- and macro-averaged most frequent sense percentages and the average number of word senses of BCCWJ and CHJ. Micro and macro in Table 11 indicate the

micro- and macro-averaged most frequent sense percentages. Word Senses in the same table are the average number of word senses. This table shows that the number of word senses of CHJ is greater than that of BCCWJ. In addition, the most frequent sense percentages of historical Japanese are lower than those of contemporary Japanese. These facts indicate that WSD of historical Japanese is more difficult than that of contemporary Japanese.

Table 11. The micro- and macro-averaged most frequent percentages and the averaged number of word senses of BCCWJ and CHJ

Corpora	Micro	Macro	Word Senses
BCCWJ	84.56%	74.49%	4.71
CHJ	75.54%	69.90%	5.48

We used a support vector machine (SVM) as a WSD classifier. We used LIBLINEAR 2.30 as an SVM tool. The word embeddings of the two words before and after the target word were used as the features. We used zero vectors for the begging or ending of sentences and beyond.

6. Results

Table 12 shows the micro- and macro-averaged accuracies of WSD according to the three scenarios and six feature types. Micro and macro in this table indicate the macro- and micro-averaged accuracies of WSD, respectively. We have two types of historical features, **historical 300 features** and **historical 200 features**. Please note that although historical 300 features yielded better results, we used 200-dimentional features when we fine-tuned the historical features owing to memory limitations. The numbers in bold indicate that they are the best methods for each scenario.

According to the table, the best micro- and macro-averaged accuracies were achieved when **FT historical features with contemporary corpus** were used in Target Only scenario. In Both and Source Only scenarios, **FT historical features with contemporary corpus** were the best for micro-averaged accuracy, and **Contemporary features** were the best for macro-averaged accuracy.

Table 12. Results of WSD

Scenarios	Both		Target Only		Source Only	
Features	Micro	Macro	Micro	Macro	Micro	Macro
Historical 300 features	69.56	63.12	74.28	70.57	59.24	48.20
Historical 200 features	69.41	63.10	73.94	70.41	59.05	48.51
Contemporary features	71.27	**66.69**	73.41	69.79	61.05	**51.21**
FT historical features with contemporary corpus	**72.03**	66.27	**74.83**	**70.80**	**61.75**	50.64
FT contemporary features with historical corpus	70.45	66.01	72.42	68.33	57.58	47.08
FT contemporary features in the order of time	70.76	66.27	73.39	69.35	56.35	47.87

7. Discussion

When we compare the three scenarios, Both, Target Only, and Source Only scenarios, according to Table 12, we can see that the ranks of the results are always the same. That is, Target Only scenario is the best, Both scenario is the second best, and Source Only scenario is the worst. We think that this is due to the priors of the word senses of the two corpora, BCCWJ and CHJ that are different from each other. Additionally, the number of examples of BCCWJ was more than

three times greater than that of CHJ. Therefore, the priors of the word senses changed according to BCCWJ. The labeled source data were not effective for the diachronic adaptation of WSD.

Next, we compared the six features. Table 12 shows that **FT historical features with contemporary corpora** tend to be better in every scenario. In addition, in Both and Source Only scenarios, features created with contemporary corpus, that is, **contemporary features**, **FT historical features with contemporary corpus**, **FT contemporary features with historical corpus**, and **FT contemporary features in the order of time**, tend to achieve better results, while in Target Only scenario, features generated from historical features, that is, **historical 300 features**, **historical 200 features**, and **FT historical features with contemporary corpus**, tend to show better results. **Historical 300 features** outperformed **historical 200 features** but features with fine-tuning were based on **historical 200 features** owing to memory limitations.

In fact, we anticipated that the final features, **FT contemporary features in the order of time** would be better features, as this method achieved good results (Takaku et al. 2020). Unfortunately, this was not the case. According to Takaku et al. (2020), although the initialization of the word embedding layer of the translation system was effective, the best result was when the ensemble method was used with fine-tuned features, and **FT contemporary features in the order of time** itself could not be the best method for translating texts of each period. Table 12 shows that, in every scenario, **FT contemporary features with historical corpus** and **FT contemporary features in the order of time** are worse than **contemporary features**, which means that fine-tuning of contemporary features decreased the accuracy of WSD. In contrast, **FT historical features with contemporary corpus** outperformed the original **historical 200 features**, which means that the fine-tuning of historical features increased WSD accuracies.

Now, let us discuss the difference in the types of approach for domain adaptation, namely, supervised, semi-supervised, and unsupervised approaches. In general, it is said that WSD accuracy will increase when more data are available. Therefore, a semi-supervised approach, which uses labeled data of source and target domains and unlabeled data of target domain, should surpass a supervised approach, which uses labeled data of source and target domains, and a supervised approach should outperform an unsupervised approach, which uses labeled source data and unlabeled target data. In our case, we also used unlabeled source data for all three approaches.

The method using **contemporary features** in Both scenario is the only supervised approach in our experiments (see Table 10). For a supervised approach, the most frequent sense percentage is a strong baseline, because we assume that all the data, including labeled target data, could be used as the baseline. If we have the labeled target data, in most cases, we can determine the most frequent senses. The micro- and macro-averaged accuracies of **contemporary features** in Both scenario were 71.27% and 66.69%, respectively, whereas the most frequent sense percentages were 75.54% and 69.90%, respectively. Unfortunately, our supervised approach method could not exceed the baseline.

The best method among semi-supervised-like approaches is the method using **FT historical features with contemporary corpus** in Target Only scenario, which is the best method in our experiments. This method used labeled target data, unlabeled source data, and unlabeled target data (see Table 9). Therefore, this is a semi-supervised approach that uses unlabeled source data and no labeled source data. In this case, the most frequent sense percentage was a strong baseline. The micro- and macro-averaged accuracies of **FT historical features with contemporary corpus** in Target Only scenario are 74.83 % and 70.80%, respectively. Although our method could not

achieve the micro-averaged accuracy, it outperformed the macro-averaged accuracy. These results imply that the word sense prior to words with many examples tends to be biased. Moreover, in the case where there are many examples in the target domain, the domain adaptation method hardly works.

The best method among unsupervised approaches is the method using **FT historical features with contemporary corpus** in Source Only scenario. For unsupervised approaches, we assume that we cannot label the target data at all, so the most frequent sense percentage is no longer a baseline. Instead, **contemporary features** in Source Only scenario are a baseline. The micro- and macro-averaged accuracies of **FT historical features with contemporary corpus** in Source Only scenario are 61.75% and 50.64%, respectively, whereas those of **Contemporary features** in Source Only scenario are 61.05% and 51.21%. Our method surpassed the baseline for the micro-averaged accuracy, but not for the macro-averaged accuracy.

8. Conclusions

We performed diachronic adaptation of WSD in a historical Japanese corpus using a contemporary Japanese corpus and examined the effects of the scenarios, features, and labeled and unlabeled data. Using labeled and unlabeled corpora of both historical and contemporary Japanese corpora, we tested three types of domain adaptation: supervised, semi-supervised, and unsupervised. Word embeddings were used for the features of SVM and the classifiers for WSD. Fifty-eight frequent polysemous words that appeared in both historical and contemporary corpora were used for the lexical sample task. We tested three scenarios, that is, Both, Target Only, and Source Only scenarios and six kinds of features: (1) historical 300 features, (2) historical 200 features, (3) contemporary features, (4) FT historical features with contemporary corpus, (5) FT contemporary features with historical corpus, and (6) FT contemporary features in the order of time. The best scenario was Target Only, which is the case where only examples of historical corpus were used for the training data, and (4) FT historical features with contemporary corpus were the best feature type. The method using (4) FT historical features with contemporary corpus was the best feature type in Target Only scenario was the best among all the methods. The method belongs to the semi-supervised approach, and it outperforms the most frequent sense baseline for the macro-averaged accuracy. In the group of unsupervised approaches, the best method was the method using (4) FT historical features with contemporary corpus in Source Only scenario. This surpassed the contemporary feature baseline for the micro-averaged accuracy.

References

Aida, Taichi, Mamoru Komachi, Toshinobu Ogiso, Hiroya 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]. *Proceedings of the NLP2021*, 712–717 (in Japanese).

Asahara, Masayuki, Sachi Kato, Tai Suzuki and Nao Ikegami (2018) 'Nihongo rekishi kopasu' 4 sakuhin ni taisuru bunruigoihyo bangou huyo to sono bunseki. ['Corpus of historical Japanese': Annotation and analysis of article numbers of word list by semantic principles towards 4 literatures]. *Proceedings of the Conference of Society for Japanese Linguistics Autumn* 2018 (in Japanese).

Asahara, Masayuki, Kikuo Maekawa, Mizuho Imada, Sachi Kato and Hikari Konishi (2014) Archiving and analysing techniques of the ultra-large-scale web-based corpus project of NINJAL, Japan. *Alexandria: The Journal of National and International Library and Information Issues* 25 (1-2): 129–148.

Daumé III, Hal (2007) Frustratingly easy domain adaptation. *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL 2007)*, 256–263.

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

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

Faruqui, Manaal, Jesse Dodge, Sujay Kumar Jauhar, Chris Dyer, Eduard Hovy and Noah A. Smith (2015) Retrofitting word vectors to semantic lexicons. *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2015)*, 1606–1615.

Hoshino, Sho, Yusuke Miyao, Shunsuke Ohashi and Akiko Aizawa (2014) Taisho kopasu wo mochiita kobun no gendaigo kikaihonyaku. [Machine translation of historical texts into contemporary texts using a parallel corpus]. *Proceedings of the NLP2014*, 816–819 (in Japanese).

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

Kato, Sachi, Masayuki Asahara and Makoto Yamazaki (2018) Annotation of word list by semantic principles labels for the balanced corpus of contemporary written Japanese. *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation (PACLIC 2018)*, 247–253.

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

Kobayashi, Kazuma, Taichi Aida and Mamoru Komachi (2021) BERT wo shiyou shita nihongo no tango no tsuujiteki na imihenka no bunseki. [Analysis of diachronic changes in meanings of Japanese words using BERT]. *Proceedings of the NLP2021*, 952–956 (in Japanese).

Komiya, Kanako and Manabu Okumura (2011) Automatic determination of a domain adaptation method for word sense disambiguation using decision tree learning. *Proceedings of 5th International Joint Conference on Natural Language Processing (IJCNLP 2011)*, 1107–1115.

Komiya, Kanako and Hiroyuki Shinnou (2018) Investigating effective parameters for fine-tuning of word embeddings using only a small corpus. *Proceedings of the Workshop on Deep Learning Approaches for Low-resource NLP(ACL 2018)*, 60–67.

Komiya, Kanako, Minoru Sasaki, Hiroyuki Shinnou and Manabu Okumura (2018) Domain adaptation using word embeddings for word sense disambiguation. *Journal of Natural Language Processing* 25 (4): 463–480.

Maekawa, Kikuo, 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. *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010)*, 1483–1486.

Maekawa, Kikuo, 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.

Mikolov, Tomas, Kai Chen, Greg Corrado and Jeffrey Dean (2013a) Efficient estimation of word representations in vector space. *Proceedings of the International Conference on Learning Representations Workshop* 2013, 1–12.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg Corrado and Jeffrey Dean (2013b) Distributed representations of words and phrases and their compositionality. *Proceedings of Conference on Neural Information Processing Systems 2013 (NIPS 2013)*, 1–9.

Mikolov, Tomas, Wen-Tau Yih and Geoffrey Zweig (2013c) Linguistic regularities in continuous space word representation. *Proceedings of the 2013 Conference of the North American Chapter of the Association*

for Computational Linguistics: Human Language Technologies (NAACL-HLT 2013), 746–751.

Miyajima, Tatsuo, Hisao Ishii, Seiya Abe and Tai Suzuki (2014) *Nihon koten taisho bunrui goi hyo* [*Word list by semantic principles referring to Japanese classics*]. Tokyo: Kasama Shoin (in Japanese).

National Institute for Japanese Language (2004) *Bunruigoihyo zouho kaitei ban* [*Word list by semantic principles, revised and enlarged edition*]. Tokyo: Dainippon Tosho (in Japanese).

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

Okumura, Manabu, Kiyoaki Shirai, Kanako Komiya and Hikaru Yokono (2010) SemEval-2010 task: Japanese WSD. *Proceedings of the SemEval-2010 (ACL2010)*, 69–74.

Raganato, Alessandro, Jose Camacho-Collados and Roberto Navigli (2017a) SemEval-2007 task 07: Coarse-grained English all-words task. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2017)*, 99–110.

Raganato, Alessandro, Claudio Delli Bovi and Roberto Navigli (2017b) Neural sequence learning models for word sense disambiguation. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP 2017)*, 1156–1167.

Schnabel, Tobias, Igor Labutov, David Mimno and Thorsten Joachims (2015) Evaluation methods for unsupervised word embeddings. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 298–307.

Shinnou, Hiroyuki, Masayuki Asahara, Kanako Komiya and Minoru Sasaki (2017a) Nwjc2vec: Word embedding data constructed from NINJAL web Japanese corpus. *Journal of Natural Language Processing* 24 (5): 705–720 (In Japanese).

Shinnou, Hiroyuki, Kanako Komiya and Minoru Sasaki (2017b) Fine-tuning for nwjc2vec. *Proceedings of the Language Rresource Workshop 2017 (LRW2017)*, 116–121 (in Japanese).

Shinnou, Hiroyuki, Kanako Komiya, Minoru Sasaki and Shinsuke Mori (2017c) Japanese all-words WSD system using the Kyoto text analysis toolkit. *Proceedings of the 31st Pacific Asia Conference on Language*, *Information and Computation (PACLIC 2017)*, 392–399.

Suzuki, Rui, Kanako Komiya, Masayuki Asahara, Minoru Sasaki and Hiroyuki Shinnou (2018) All-words word sense disambiguation using concept embeddings. *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018)*, 1006–1011.

Takaku, Masashi, Tosho Hirasawa, Mamoru Komachi and Kanako Komiya (2020) Neural machine translation from historical Japanese to contemporary Japanese using diachronically domain-adapted word embeddings. *Proceedings of the 34th Pacific Asia Conference on Language*, *Information and Computation (PACLIC 2020*), 534–541.

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

Yaginuma, Daiki, Kanako Komiya and Hiroyuki Shinnou (2018) WSD of domain adaptation by distributed representation of fine tuning. *IPSJ SIG Technical Reports (NL)* 1: 1–5 (in Japanese).

Related websites

Asahara, Masayuki. *BCCWJ-WLSP*. https://github.com/masayu-a/BCCWJ-WLSP (accessed October 2021).

GENSIM. *Word2vec embeddings*. https://radimrehurek.com/gensim/models/word2vec.html (accessed October 2021).

Machine Learning Group at National Taiwan University, *LIBLINEAR—A Library for Large Linear Classification*. https://www.csie.ntu.edu.tw/~cjlin/liblinear/(accessed November 2021).

National Institute for Japanese Language and Linguistics. *Balanced Corpus of Contemporary Written Japanese*. https://ccd.ninjal.ac.jp/bccwj/ (accessed December 2021).

National Institute for Japanese Language and Linguistics (2021) *Corpus of Historical Japanese.* (Version 2021.3, Chunagon Version 2.5.2) https://ccd.ninjal.ac.jp/chj/ (accessed October 2021).

Shogakukan, Inc., *The Complete Collection of Japanese Classical Literature (New Edition) Shogakukan, Inc.* https://japanknowledge.com/en/contents/koten/ (accessed October 2021).

分散表現を利用した日本語歴史コーパスにおける語義曖昧性解消の通時適応

古宮嘉那子[a]　田邊　絢[b]　新納浩幸[c]

[a] 東京農工大学／国立国語研究所 共同研究員

[b] 茨城大学

[c] 茨城大学／国立国語研究所 共同研究員

要旨

語義タグ付きコーパスを用いた現代日本語の語義曖昧性解消の研究は数多い。しかし，入手可能なタグ付きコーパスが少ないため，日本語の古典語の語義曖昧性解消を高性能に行うことは難しい。そのため，現代日本語文を用いて通時的な領域適応を行うことは，古典語の語義曖昧性解消の性能を高めるひとつの解決方法であると考えられる。本研究では，日本語の古典語の語義曖昧性解消において，領域適応手法のひとつである，分散表現の fine-tuning の効果について調べる。現代文の分散表現である NWJC2vec の古典語による fine-tuning や，古典語によって作成した分散表現の現代文による fine-tuning など，様々な fine-tuning のシナリオを検証した。さらに，NWJC2vec を古典語で fine-tuning する際には，時代順に段階的に分散表現を fine-tuning する手法についても試した。語義曖昧性解消の対象語の前後二語ずつの単語の分散表現を素性とし，Support Vector Machine の分類器に用いて分類を行った。シナリオは（1）現代文のコーパスの全用例と古典語のコーパスの用例 8 割を訓練事例とし，残りの 2 割の古典語の用例をテストとして利用する場合，（2）古典語の用例だけを利用して五分割交差検定を行った場合，（3）現代文のコーパスの全用例を訓練事例とし，古典語全用例をテストする場合の三通りを比較した。最高の精度となったのは，（2）古典語の用例だけを利用したシナリオで，古典語によって作成した分散表現に現代文による fine-tuning を行った場合であった。

キーワード：領域適応，歴史コーパス，通時適応，語義曖昧性解消，分散表現