Domain Adaptation using Word Embeddings for Word Sense Disambiguation

Kanako Komiya†, Minoru Sasaki†, Hiroyuki Shinnou† and Manabu Okumura††

In this paper, we propose domain adaptation using word embeddings for word sense disambiguation (WSD). The validity for WSD of word embeddings derived from a huge corpus such as Wikipedia had already been shown, but their validity in a domain adaptation framework has not been previously discussed. If word embeddings are valid in this new context, the impact of the document type of the corpora on WSD is still unknown. Therefore, we investigate the performances of domain adaptation in WSD using word embeddings from the source, target and general corpora and examine (1) whether the word embeddings are valid for domain adaptation of WSD and (2) if they are, the effects of the document type of the corpora from which the word embeddings are derived. We used three corpora of distinct document types and performed domain adaptation experiments using the document types as the domains. The experiments, conducted using Japanese corpora, revealed that the accuracy of WSD was highest when we used the word embeddings obtained from the target corpora together with a general corpora.

Key Words: Word Sense Disambiguation, Domain Adaptation, Word Embeddings, Word2vec

1 Introduction

Word embeddings (Mikolov, Chen, Corrado, and Dean 2013a; Mikolov, Sutskever, Chen, Corrado, and Dean 2013b; Mikolov, tau Yih, and Zweig 2013c) are the vector representations of word meanings that have compositionality. They are effective for word sense disambiguation (WSD) tasks (Sugawara, Takamura, Sasano, and Okumura 2015) since they are low-dimensional vectors and the sparseness of the meaning representations can be greatly alleviated through their use.

Domain adaptation involves adapting a classifier that has been trained from data in one domain (source domain) to data in another domain (target domain). Domain adaptation, including domain adaptation in WSD (see Section 2), has been studied extensively. First, we propose using word embeddings for domain adaptation of WSD (see Section 3). Since domain adaptation

---

† Ibaraki University
‡ Tokyo Institution of Technology
suffers from data sparseness caused by shifting between source and target domains, we suppose that the use of word embeddings can improve the performance of WSD in a domain adaptation framework.

However, one problem with domain adaptation is the shift of the priors of the word senses in the texts (Shinnou, Onodera, Sasaki, and Komiya 2015a), which means that the word meanings in one domain could be different from those in another domain.

Therefore, second, we investigate the performances of domain adaptation in WSD using the word embeddings obtained from the source, target and general corpora, and we examine the effects according to the document types. Our experiments were with Japanese WSD, and the experiments (described in Section 4) revealed that the accuracy of WSD was highest when we used word embeddings obtained from the corpus of the target domain and those obtained from the general domain together. The observed dependence on the corpus indicated to us that not only did the size of the corpus used to generate the word embeddings affect the results but so did the document type of the corpus. Finally, we discuss the results (Section 5) and conclude this paper (Section 6).

We have three sense-tagged corpora of different document types: Q & A sites, newspapers and white papers, so we performed the domain adaptation experiments regarding these document types as the domains. Considering that, the same topic can be described in both Q & A sites and newspapers, the main difference among these corpora is not the topic of the documents but rather the manner of writing.

2 Related Work

The use of word embeddings is a considerably active area of current and recent research. Some important examples of this research include the works of Sugawara et al. (2015) and Taghipour and Ng (2015). The former study investigated the impact of the use of word embeddings for WSD, whereas the latter study proposed semi-supervised WSD using word embeddings and revealed that the word embeddings of a general or specific domain improved the performance of WSD.

Research on domain adaptation is not limited to the area of natural language processing. The domain adaptation problem can be categorized into three types depending on the information to be learned, namely, supervised, semi-supervised and unsupervised approaches. According to Daumé (Daumé III, Kumar, and Saha 2010), in a supervised approach, a classifier is developed from a large quantity of labelled source data and a small amount of labelled target data. In a semi-supervised approach, a classifier is developed from a large amount of labelled source data,
Komiya et al. Domain Adaptation using Word Embeddings for Word Sense Disambiguation

Given a triple of the target word type of WSD, source data and target data, Komiya and Okumura (2011, 2012b) determined an optimal method of domain adaptation using decision tree learning. They discussed which features influenced how the best method was determined and proposed determining the optimal training data set for each instance using the degree of confidence for supervised domain adaptation in WSD (Komiya and Okumura 2012a).

That work which is closest to ours is that of Sugawara et al. (2015), which investigated the context representation using word embeddings for WSD. Iacobacci, Pilehvar, and Navigli (2016) also employed word embeddings for word sense disambiguation, but they did not attempt domain adaptation. Yamaki, Shinnou, Komiya, and Sasaki (2016) also proposed a method that employed context word embeddings for supervised word sense disambiguation. We propose using...
the context representation derived from word embeddings for domain adaptation of WSD and furthermore show that word embeddings can improve the performance of WSD in a domain adaptation framework. Additionally, we investigate the performances of domain adaptation in WSD using the word embeddings obtained from the source, target and general corpora and examine the effects according to their document types.

3 Domain Adaptation in WSD using Word Embeddings

In addition to the baseline features in both the source and target data, we carry out domain adaptation for WSD by concatenating the features of the word embeddings of words surrounding the target word of WSD. We followed Sugawara et al. (2015) and used Context-Word-Embeddings; a Context-Word-Embedding is a vector formed by the concatenation of the real-valued vectors consisting of the words in the context window of the target word of WSD. As Sugawara et al., described, Context-Word-Embeddings of a target word of WSD is a vector concatenating $v_{w-N}, \ldots, v_{w-1}, v_{w+1}, \ldots, v_{w+N}$ when the window size is $N$ and the words appearing in the context window are $w_{-N}, \ldots, w_{-1}, w_{+1}, \ldots, w_{+N}$, where $v_w$ represents an embedding of word $w$. If the dimension of each word embedding is $d$, the size of this feature vector is $2N \times d$. Table 1 and Figure 1 show a simple example of Context-Word-Embeddings. Figure 1 shows the Context-Word-Embeddings of an instance vector for the word “大きい” (big) in the context of the phrase “効果が大きいため、” (Because the effect is big,) as shown in Table 1. In this example, $N$ is two and $d$ is three.

We investigated the following nine methods of domain adaptation which vary by the document types of the corpora used to generate the word embeddings.

Table 1  Simple example of Context-Word-Embeddings

| Japanese word | Translation | Word embedding |
|---------------|-------------|----------------|
| 効果          | effect      | 0.1 0.2 0.3    |
| が             | is          | 0.4 0.1 0.7    |
| 大きい         | big         | 0.5 0.2 0.1    |
| ため          | because     | -0.2 0.3 0.1   |
| .             | .           | -0.7 0.8 0.3   |

1 Sugawara et al. (2015) reported that Context-Word-Embeddings improved the results of WSD more than Average-Word-Embeddings, which are the averages of vector representations of words in the context window of the target word of WSD.
Add Target Word embeddings obtained from the target data are used.
Add Source Word embeddings obtained from the source data are used.
Add Wiki Word embeddings obtained from Wikipedia, a huge, general corpus, are used.
Add Target & Source Word embeddings obtained from the source and those obtained from
the target data are used together by concatenating them.
Add Target & Wiki Word embeddings obtained from the target data and those obtained
from the huge, general corpus are used together by concatenating them.
Add Target Large Word embeddings obtained from the large target data are used.
Add Source Large Word embeddings obtained from the large source data are used.
Add Target Large & Source Large Word embeddings obtained from the large target data
and those obtained from the large source data are used together by concatenating them.
Add Target Large & Wiki Word embeddings obtained from the large target data and those
obtained from the huge, general corpus are used together by concatenating them.

4 Experiment

We used word2vec\(^2\) (Mikolov et al. 2013a, 2013b, 2013c) to generate the word embeddings. To generate the word embeddings, the vector size (which is \(d\) in Section 3) and the window size were set to 200 and 5, respectively. We used a skip-gram algorithm. The window size of each target word (which is \(N\) in Section 3) was set to 2. We used default settings for other parameters. Libsvm (Chang and Lin 2001), which supports multi-class classification, was used as the classifier for WSD.\(^3\) A linear kernel was used in accordance with the results obtained from preliminary experiments. Twenty features, which were the same as (Komiya and Okumura 2011) and (Komiya and Okumura 2012a), were introduced as baseline features to train the classifier.

- Morphological features used include:
  - Bag-of-words,
  - Part-of-speech (POS), and

\(^2\) https://code.google.com/archive/p/word2vec/
\(^3\) We used the -b option of libsvm.
This document describes baseline syntactic and semantic features, and presents data used in experiments. 

Baseline syntactic features are as follows:

- If the POS of a target word is a noun, the verb that the target word modifies is used.
- If the POS of a target word is a verb, the case element of ‘ヲ’ (wo, objective) for the verb is used.

The semantic feature of the baseline was

- Semantic classification code.

Morphological features and a semantic feature were extracted from the surrounding words (two words to the right and left) of the target word and from the target word itself. POS and the finer subcategory of POS could be obtained by using a morphological analyser. We used Mecab\(^4\) as a morphological analyser, CaboCha\(^5\) as a syntactic parser and the Word List by Semantic Principles (National Institute for Japanese Language and Linguistics 1964) for semantic classification codes. (For example, the code of the ‘program’ was 1.3162.)

4.1 Data

Three labelled data sets were used for the experiments: (1) the sub-corpus of white papers in the Balanced Corpus of Contemporary Japanese (BCCWJ) (Maekawa, Yamazaki, Ogiso, Maruyama, Ogura, Kashino, Koiso, Yamaguchi, Tanaka, and Den 2014), (2) the sub-corpus of documents from a Q & A site on the WWW in the BCCWJ and (3) Real World Computing (RWC) text databases (newspaper articles) (Hashida, Isahara, Tokunaga, Hashimoto, Ogino, and Kashino 1998). Domain adaptation was conducted in six directions in accordance with the possible combinations of source and target data. Word senses were annotated in these corpora in accordance with a Japanese dictionary, that is, the Iwanami Kokugo Jiten (Nishio, Iwabuchi, and Mizutani 1994). It has three levels for sense IDs, and we used the middle-level sense in the experiments. Multi-sense words that appeared at least 50 times in each data set were selected as the target words in the experiment; 36-word types were used in the experiments. Table 2 lists the minimum, maximum and average number of instances of each word type for each corpus. Table 3 shows the percentage of word types in the target corpus unknown in the source corpus. For example, 40.82% of the words appearing in white papers did not appear in the Q & A site corpus.

Table 4 summarizes the list of target word types. The first column labelled ‘No. of senses’ is the number of the senses in the dictionary of each word type. For example, the word type “考え

---

\(^4\) https://github.com/jordwest/mecab-docs-en
\(^5\) http://sourceforge.net/projects/cabocha/
Table 2  Minimum, maximum and average number of instances of each word type for each corpus

| Document types | Min. | Max.  | Sum   | Avg. |
|----------------|------|-------|-------|------|
| White papers   | 60   | 8,691 | 76,889| 2,136|
| Q & A site     | 158  | 17,387| 112,320| 3,120|
| Newspaper      | 92   | 1,046 | 9,959 | 277  |

Table 3  Percentage of word types in the target corpus unknown in the source corpus

| Target          | Source          | Known words | All words | Unknown word ratio |
|-----------------|-----------------|--------------|-----------|-------------------|
| White papers    | Q & A site      | 4,624        | 7,813     | 40.82%            |
| White papers    | Newspaper       | 2,633        | 7,813     | 66.30%            |
| Q & A site      | White papers    | 4,624        | 17,164    | 73.06%            |
| Q & A site      | Newspaper       | 3,772        | 17,164    | 78.02%            |
| Newspaper       | White papers    | 2,633        | 5,469     | 51.86%            |
| Newspaper       | Q & A site      | 3,772        | 5,469     | 31.03%            |

Table 4  List of target word types

| No. of senses | Target words (in Japanese) | Sense example in English | No. of senses | Target words (in Japanese) | Sense example in English |
|---------------|----------------------------|--------------------------|---------------|----------------------------|--------------------------|
| 2             | 考える                      | think                    | 3             | 地方                        | area                     |
|               | 技術                        | technique                |               | 出る                        | go out                   |
|               | 経済                        | economy                  |               | 入る                        | enter                    |
|               | 現在                        | present                  |               | 開く                        | open                     |
|               | 子供                        | child                    |               | 前                          | before                   |
|               | 自分                        | self                     | 4             | 時間                        | time                     |
|               | 情報                        | information              |               | 進める                      | move forward             |
|               | 高い                        | high                     |               | 出す                        | emit                     |
|               | 場合                        | in case                  |               | 車る                        | ride                     |
|               | ほか                        | other                    |               | 計る                        | plan                     |
| 3             | 言う                        | say                      | 5             | 良い                        | good                     |
|               | 今                          | now                      |               | 持つ                        | have                     |
|               | 入れる                      | put in                   | 6             | 合う・会う                   | meet                     |
|               | 大きい                      | big                      |               | 見る                        | see                      |
|               | 関係                        | connection               |               |                             |                          |
|               | 聴く                        | hear                     | 7             | 手                          | hand                     |
|               | 社会                        | society                  |               |                             |                          |
|               | 進む                        | proceed                  | 8             | 上げる                      | raise                    |
|               |                             |                          |               | 取る                        | take                     |
“(think) has two senses in the dictionary. Please note that there is no guarantee that all the senses in the dictionary will appear in the corpora.

In addition, we used seven types of unlabelled data to generate the word embeddings: (1) a large collection of white papers in BCCWJ, (2) a large collection of documents from a Q & A site on the WWW in the BCCWJ, (3) a large collection of the newspaper articles of the Mainichi Shimbun 1994 including RWC text databases, (4) collections of the Diet minutes and Budgets and Audits at the National Diet from 1947 to 2012, (5) a subset of Q & A sites on the WWW, specifically, “Yahoo! Chiebukuro data (2nd edition)”, (6) a large collection of the newspaper articles of the Mainichi Shimbun from 1991 to 2005 including RWC text databases, and (7) dumped Japanese Wikipedia data (2015/10/02). Of these, corpora (1), (2) and (3), are the unlabelled corpora that is larger than the labelled data for WSD. It includes the corresponding labelled data. Corpora (4), (5) and (6) are used as larger corpora that correspond to the first three. Technically, the texts used to form corpus (4), the Diet minutes and Budgets and Audits, are not white papers, and corpus (4) does not include corpus (1). However, we used corpus (4) as a larger corpus corresponding to corpus (1) because we could not find a larger collection of white papers and we believe that the two classes of documents are similar to one another. As for Q & A site corpora, we used two distinct larger corpora; corpus (5a) a subset of Q & A sites on the WWW that includes corpus (2) as a subset, and corpus (5b) is another subset of the Q & A sites on the WWW that excludes corpus (2). This allowed us to investigate the impact of the inclusion of the unlabelled data of labelled data. Table 5 shows the number of word tokens and sentences according to the corpora from which the word embeddings generated, the number of word types of the word embeddings and their ratios to those of Wikipedia in accordance with each document type. The numbers of types in the table are different from the number of word types of each corpus because word2vec generated the meaning representation vector if and only if a word appeared in the corpus at least five times.

Table 6 shows the corpora for word embeddings according to the method, the source corpus and target corpus. For example, (5) a large collections of documents from a Q & A site on the WWW in the BCCWJ was used for the domain adaptation from (1) the sub-corpus of white papers in the BCCWJ to (2) the sub-corpus of documents from a Q & A site on the WWW in the BCCWJ by the method Add Target Large. Table 7 shows the ratio of unknown word

---

6 http://www2.ninjal.ac.jp/lrc/index.php
7 http://www.nii.ac.jp/dsc/idr/en/yahoo/chiebkr2/Y_chiebukuro.html
8 https://dumps.wikimedia.org/jawiki/
9 Half of them are overlapped each other.
Table 5  Number of word tokens, sentences and word types of each document type and their ratios to that of Wikipedia

| No. | Corpus          | Word tokens | Word types | Sentences |
|-----|-----------------|-------------|------------|-----------|
|     |                 | No.         | No.        | No.       |
|     |                 | Ratio       | Ratio      | Ratio     |
| (1) | White papers    | 5,496,566   | 14,303     | 139,474   | 0.47      |
| (2) | Q & A site      | 13,122,971  | 37,893     | 681,118   | 1.11      |
| (3) | Newspaper       | 32,420,197  | 69,719     | 693,033   | 2.75      |
| (4) | White papers Large | 297,419,109 | 71,609     | 6,410,020 | 25.22     |
| (5a)| Q & A site Large incl. (2) | 827,685,343 | 352,209    | 10,579,871 | 70.18     |
| (5b)| Q & A site Large excl. (2) | 1,000,576,962 | 378,746    | 13,104,701 | 84.84     |
| (6) | Newspaper Large | 537,960,323 | 201,421    | 11,416,678 | 45.62     |
| (7) | Wikipedia       | 1,179,313,609 | 1,078,930 | 89,116,263 | 100.00    |

Table 6  Corpora for word embeddings according to method and source and target corpora

| Source → Target              | (1) → (2) | (1) → (3) | (2) → (1) | (2) → (3) | (3) → (1) | (3) → (2) |
|-----------------------------|-----------|-----------|-----------|-----------|-----------|-----------|
| Add Target                  | (2)       | (3)       | (1)       | (3)       | (1)       | (2)       |
| Add Source                  | (1)       | (1)       | (2)       | (2)       | (3)       | (3)       |
| Add Wiki                    | (7)       | (7)       | (7)       | (7)       | (7)       | (7)       |
| Add Target and Source       | (1)+(2)   | (1)+(3)   | (1)+(2)   | (2)+(3)   | (1)+(3)   | (2)+(3)   |
| Add Target and Wiki         | (2)+(7)   | (3)+(7)   | (1)+(7)   | (3)+(7)   | (1)+(7)   | (2)+(7)   |
| Add Target Large incl. (2)  | (5a)      | (6)       | (4)       | (6)       | (4)       | (5a)      |
| Add Target Large excl. (2)  | (5b)      | (6)       | (4)       | (6)       | (4)       | (5b)      |
| Add Source Large            | (4)       | (4)       | (5b)      | (5b)      | (6)       | (6)       |
| Add Target Large and Source Large | (4)+(5b) | (4)+(6)   | (4)+(5b)  | (5b)+(6)  | (4)+(6)   | (5b)+(6)  |
| Add Target Large and Wiki   | (5b)+(7)  | (6)+(7)   | (4)+(7)   | (6)+(7)   | (4)+(7)   | (5b)+(7)  |

types of the tagged corpus in the word embedding vectors. For example, 12.31% of the words in white papers did not have word embeddings. See line (1), white papers, in Table 5.\(^{10}\) The average accuracies over six directions \(^{11}\) (Table 8) and the average accuracies over two directions according to the target corpora \(^{12}\) (Table 9) were evaluated in Section 4.2.

### 4.2 Results

In what follows, the micro-average is the average over word tokens, and the macro-average is that over word types. Table 8 lists the micro- and macro-averaged accuracies of WSD for the whole

---

\(^{10}\) Although distribution differences of word embeddings among corpora are helpful to analyse the effects of our method for each corpus, we could not map the word embeddings onto the same feature space because we created them separately.

\(^{11}\) Six directions are (1) → (2), (1) → (3), (2) → (1), (2) → (3), (3) → (1), and (3) → (2).

\(^{12}\) For example, the accuracies of column ‘Newspaper’ are the averages over accuracies of domain adaptation from white papers to newspaper ((1) → (3)) and from Q & A site to newspaper ((2) → (3)) except for ‘Self’, which is the result of (3) → (3).
Table 7  Unknown word type ratio of the tagged corpus in the word embedding corpus

| Tagged corpus | W2V corpus | Known words | All words | Unknown word ratio |
|---------------|------------|-------------|-----------|--------------------|
| White papers (1) | (1) | 6,851 | 7,813 | 12.31% |
| White papers (1) | (2) | 6,190 | 7,813 | 20.77% |
| White papers (1) | (3) | 7,234 | 7,813 | 7.41% |
| White papers (1) | (4) | 7,451 | 7,813 | 4.63% |
| White papers (1) | (5a) | 7,612 | 7,813 | 2.57% |
| White papers (1) | (5b) | 7,621 | 7,813 | 2.46% |
| White papers (1) | (6) | 7,648 | 7,813 | 2.11% |
| White papers (1) | (7) | 7,659 | 7,813 | 1.97% |
| Q & A site (2) | (1) | 6,844 | 17,164 | 60.13% |
| Q & A site (2) | (2) | 15,307 | 17,164 | 10.82% |
| Q & A site (2) | (3) | 14,467 | 17,164 | 15.71% |
| Q & A site (2) | (4) | 13,668 | 17,164 | 20.37% |
| Q & A site (2) | (5a) | 16,956 | 17,164 | 1.21% |
| Q & A site (2) | (5b) | 16,936 | 17,164 | 1.33% |
| Q & A site (2) | (6) | 16,371 | 17,164 | 4.62% |
| Q & A site (2) | (7) | 1,6501 | 17,164 | 3.86% |
| Newspaper (3) | (1) | 3,511 | 5,469 | 35.80% |
| Newspaper (3) | (2) | 4,707 | 5,469 | 13.93% |
| Newspaper (3) | (3) | 5,334 | 5,469 | 2.47% |
| Newspaper (3) | (4) | 5,067 | 5,469 | 7.35% |
| Newspaper (3) | (5a) | 5,357 | 5,469 | 2.05% |
| Newspaper (3) | (5b) | 5,367 | 5,469 | 1.87% |
| Newspaper (3) | (6) | 5,441 | 5,469 | 0.51% |
| Newspaper (3) | (7) | 5,409 | 5,469 | 1.10% |

Table 8  Summary of accuracies

| Method | Macro Avg. | Micro Avg. |
|--------|------------|------------|
| Baseline features | 77.90% | 79.79% |
| Add Target | 78.55%+ | 80.01%+* |
| Add Source | 78.26%+ | 79.91%+ |
| Add Wiki | 78.60%+ | 79.73%− |
| Add Target & Source | 79.27%+ | 80.79%+* |
| Add Target & Wiki | 79.41%+ | **80.92%+** |
| Add Target Large incl. (2) | 79.13%+ | 80.76%+* |
| Add Target Large excl. (2) | 79.16%+ | 80.78%+* |
| Add Source Large | 78.81%+ | 79.95%+* |
| Add Target Large and Source Large | 79.14%+ | 79.05%−* |
| Add Target Large and Wiki | **79.53%+** | 80.91%+* |
| MFS | 77.54% | 82.05% |
| Self | 89.95% | 91.80% |
Table 9  Average accuracies of WSD according to corpora and methods of domain adaptation

| Target data   | Newspaper     | Q & A site    | White papers |
|---------------|---------------|---------------|--------------|
| Method        | Macro | Micro | Macro | Micro | Macro | Micro |
| **Baseline features** | 78.84% | 78.15% | 74.01% | 76.59% | 80.86% | 84.70% |
| **Add Target** | 78.92%+ | 78.43%+ | 74.58%+ | 77.30%+* | 82.15%+ | 84.17%−* |
| **Add Source** | 78.45%− | 77.91%− | 74.30%+ | 77.14%+* | 82.04%+ | 84.20%−* |
| **Add Wiki** | 79.19%+ | 78.52%+ | 74.45%+ | 77.10%+* | 82.15%+ | 83.73%−* |
| **Add Target & Source** | 79.27%+ | 78.74%+ | **76.56%+** | 77.58%+* | 83.07%+ | 85.75%+* |
| **Add Target & Wiki** | **79.65%+** | 79.08%+* | 75.43%+ | 77.63%+* | 83.15%+ | **85.96%+** |
| **Add Target L incl. (2)** | 79.17%+ | 78.69%+ | 75.14%+ | 77.73%+* | 83.07%+ | 85.46%+* |
| **Add Target L excl. (2)** | 79.17%+ | 78.69%+ | 75.22%+ | **77.77%+** | 83.07%+ | 85.46%+* |
| **Add Source L** | 79.20%+ | 78.71%+ | 74.98%+ | 77.34%+* | 82.26%+ | 83.91%−* |
| **Add Target L and Source L** | 79.59%+ | **79.14%+** | 75.69%+ | 77.61%+* | 82.13%+ | 81.13%−* |
| **Add Target L & Wiki** | 79.59%+ | 79.07%+* | 75.66%+ | 77.72%+* | **83.33%+** | 85.81%+* |
| **MFS** | 73.20% | 72.64% | 77.71% | 78.47% | 81.72% | 88.51% |
| **Self** | 84.31% | 85.23% | 90.58% | 89.84% | 94.74% | 97.76% |

data set by method of domain adaptation, and Table 9 summarizes the micro- and macro-averaged accuracies of WSD by the corpus and method of domain adaptation. We tested Self, which is standard supervised learning with the whole target data, by five-fold cross-validation, assuming that fully annotated data were obtained and could be used for learning with word embedding features,13 MFS, which is the most frequent sense of the target corpus14, and Baseline features, which is standard supervised learning without word embedding features, as references. Please note that all the methods except for Self, including Baseline features and MFS, are performed in the domain adaptation scenario, where the domains of the training and test data are different.

Self was an upper bound, and Baseline features features was a baseline. Excluding Self and MFS, the highest accuracies for each corpus are shown in bold typeface in Tables 8 and 9. A plus signlus or a minus sign after the percentage indicate that the accuracy of the proposed method is higher or lower, respectively, than that of Baseline features. The asterisk is used to indicate that the difference between accuracies of the proposed method and Baseline features is statistically significant according to a chi-square test. The level of significance in the test was 0.05.

13 “Self” was learned using word embeddings from the target corpus itself. For example, “Self” for (1) white papers was learned using word embeddings from (1) white papers.

14 Note that we cannot know the most frequent sense in the target corpus without the labeled target data.
5 Discussion

First, Table 8 demonstrates that the macro-averaged accuracies of all the methods outperformed Baseline features features, which indicates that the word embedding features are effective in the domain adaptation framework. In addition, the same table shows that the micro-averaged accuracies of all methods except Add Wiki and Add Target Large and Source Large (namely, Add Target, Add Target, Add Source, Add Target and Source, Add Target and Wiki, two kinds of Add Target Large, Add Source Large and Add Target Large and Wiki) outperformed Baseline features. Moreover, the differences in micro-averaged accuracies between Baseline features and all methods except Add Source and Add Wiki are statistically significant.

When we used Add Wiki, macro-averaged accuracy increased but the micro-averaged accuracy decreased. Table 9 suggests that the reason for the decline is the micro-averaged accuracy of white papers. We think this is because the senses in the corpus of white papers are very biased as shown in Table 9 and for these cases the word embedding features generated from the general corpus cannot improve the accuracy. Table 9 also shows that not only Add Wiki but also Add Target, Add Source, Add Source Large and Add Target Large and Source Large could not beat Baseline features for white papers. We believe this to be for the same reason.

The decline of micro-averaged accuracy in Add Target Large and Source Large also came from white papers. The WSD accuracy of a word “技術” (technique), which had more than 7,000 instances, dropped considerably compared with Baseline features when domain adaptation from newspaper to white paper was conducted.

Next, Table 8 shows that both micro- and macro-averaged accuracies of Add Target Large and the micro-averaged accuracy of Add Target outperformed Add Wiki. In addition, the differences between Add Target and Add Wiki or the differences between Add Target Large and Add Wiki in micro-averaged accuracies are statistically significant. These results indicate that we should use the word embeddings obtained from the target corpus rather than the general corpora. In particular, the ratios of the vocabulary size of the target corpora to that of Wikipedia range from 1% to 7% and those of the larger target corpora are from 6% to 36%, according to Table 5. This indicates that Add Target or Add Target Large could improve the accuracies of WSD even though the number of word types appearing in the corpora is much smaller than the number of word types in Wikipedia.

On the other hand, Table 8 shows that in macro-averaged accuracies the best method is Add Add Target Large and Wiki. It also shows that the best method in micro-averaged accuracies is Add Target and Wiki. The same table indicates that the second-best method in micro-averaged
accuracy is Add Target Large and Wiki and Add Target and Wiki in macro-averaged accuracy. Therefore, we should use the word embeddings obtained from the target corpus and the general corpora together.

Finally, we will discuss why the word embedding features obtained from the target corpora are effective for domain adaptation. We think that the word embedding features are effective for WSD tasks because their use greatly alleviates the sparseness of the meaning representations. Our experiments show that the results of Add Target Large are always better than those of Add Target, which indicates that the word embeddings obtained from the larger corpora did improve the performance of WSD when the document types were the same. However, we cannot explain all the improvements of the results in a domain adaptation framework only through the alleviation of the sparseness alone. Table 10 shows the ratios of the unknown word types of each method; the ratio of the unknown words of Add Target (8.53%) is greater than that of Add Wiki (2.31%) in the target data.

This demonstrates that the improvement in the results comes from not only the alleviation of the sparseness, that is, the size of the word embeddings, but also the corresponding document type. In particular, when we compared Add Target with Add Wiki, the experimental results show that the document type of the corpus used to generate the word embeddings affected the results more than the size of the corpus.

Next, let us investigate the effect of inclusion of the unlabelled data with labelled data. According to Table 8, Add Target Large excl. (2) is higher than that of Add Target Large incl. (2), which indicates that the accuracy was higher when the unlabelled data were not included. Table 5 demonstrates that the number of word tokens and sentences of the corpora from which

| Table 10 | Ratio of unknown word types |
|----------|----------------------------|
| Method   | Ratio      |
| Add Target | 8.53%  |
| Add Source | 25.63% |
| Add Wiki   | 2.31%     |
| Add Target & Source | 5.24% |
| Add Target & Wiki | 1.16% |
| Add Target Large incl. (2) | 2.12% |
| Add Target Large excl. (2) | 2.16% |
| Add Source Large | 6.46% |
| Add Target Large and Source Large | 1.03% |
| Add Target Large and Wiki | 0.85% |
the word embeddings were generated and the number of word types of the word embeddings of Add Target Large excl. (2) are greater than those of Add Target Large incl. (2). In addition, Table 10 shows that ratio of unknown word types of Add Target Large excl. (2) (2.12%) is smaller than that of Add Target Large incl. (2) (2.16%). Again, these results indicate that the word embeddings obtained from larger corpora improved the performance of WSD when their document type was the same.

Moreover, Table 9 shows that accuracy of Add Target Large was better than that of Baseline features and Add Wiki for white papers. This indicates that collections of the Diet minutes and Budgets and Audits at the National Diet were useful as target corpus for white papers. This suggests that minor differences of document types do not affect the accuracies of domain adaptation for word sense disambiguation.

Finally, we should note that it has yet to be determined if our method is applicable to domain adaptation for corpora of other document types. It is also unclear more generally under what conditions our method is valid. It is difficult to analyse that because one should take into consideration not only the differences among the document types but also the differences among the distributions of the word senses of each target word. Therefore, we treat that problem as future work. However, we believe that our method is generally applicable to corpora of other document types because most domain adaptation methods are not effective for the three corpora we used when the averaged accuracies are evaluated, whereas ours was effective. We also believe that this method is applicable to domain adaptation for English word sense disambiguation because the original method, Context-Word-Embeddings, is effective for English word sense disambiguation.

6 Conclusions

We proposed the use of word embedding features generated through word2vec and showed their validity for domain adaptation in Japanese WSD tasks. We investigated the performances of domain adaptation in WSD using various methods, namely, Add Target, Add Source, Add Wiki, Add Target & Source, Add Target & Wiki, Add Target Large, Add Source Large, Add Target Large & Source Large, and Add Target Large & Wiki. Furthermore, we examined the effects of the types of document in the corpus used to generate the word embeddings. Our experiments demonstrated that word embeddings that were obtained from the target corpus together with a huge, general corpus improved the accuracies of WSD more than any other method. They also revealed that improvement in the results of domain adaptation came not only from the alleviation of the sparseness, that is, the size of the word embeddings, but also from the document types of
the corpus used to generate the word embeddings.

Acknowledgement

This paper is the revised version of Komiya, Suzuki, Sasaki, Komiya, Suzuki, Sasaki, Shinnou, and Okumura (2017), which is published in the proceedings of the CICLING 2017. We would like to thank the Yahoo Japan Corporation and the National Institute of Informatics that provide us the “Yahoo! Chiebukuro data (2nd edition).” This work was partially supported by JSPS KAKENHI Grant Number 15K16046.

Reference

Agirre, E. and de Lacalle, O. L. (2008). “On Robustness and Domain Adaptation using SVD for Word Sense Disambiguation.” In Proceedings of COLING 2008, pp. 17–24.

Agirre, E. and de Lacalle, O. L. (2009). “Supervised Domain Adaption for WSD.” In Proceedings of EACL 2009, pp. 42–50.

Blitzer, J., McDonald, R., and Pereira, F. (2006). “Domain Adaptation with Structural Correspondence Learning.” In Proceedings of EMNLP 2006, pp. 120–128.

Chan, Y. S. and Ng, H. T. (2006). “Estimating Class Priors in Domain Adaptation for Word Sense Disambiguation.” In Proceedings of COLING-ACL 2006, pp. 89–96.

Chan, Y. S. and Ng, H. T. (2007). “Domain Adaptation with Active Learning for Word Sense Disambiguation.” In Proceedings of ACL 2007, pp. 49–56.

Chang, C.-C. and Lin, C.-J. (2001). LIBSVM: a library for support vector machines. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Daumé III, H. (2007). “Frustratingly Easy Domain Adaptation.” In Proceedings of ACL 2007, pp. 256–263.

Daumé III, H., Kumar, A., and Saha, A. (2010). “Frustratingly Easy Semi-Supervised Domain Adaptation.” In Proceedings of the 2010 Workshop on Domain Adaptation for Natural Language Processing, ACL 2010, pp. 23–59.

Hashida, K., Isahara, H., Tokunaga, T., Hashimoto, M., Ogino, S., and Kashino, W. (1998). “The RWC Text Databases.” In Proceedings of the 1st International Conference on Language Resource and Evaluation, pp. 457–461.

Iacobacci, I., Pilehvar, M. T., and Navigli, R. (2016). “Embeddings for Word Sense Disambiguation: An Evaluation Study.” In Proceeding of ACL 2016, pp. 897–907.
Jiang, J. and Zhai, C. (2007). “Instance Weighting for Domain Adaptation in NLP.” In *Proceedings of ACL 2007*, pp. 264–271.

Komiya, K. and Okumura, M. (2011). “Automatic Determination of a Domain Adaptation Method for Word Sense Disambiguation Using Decision Tree Learning.” In *Proceedings of IJCNLP 2011*, pp. 1107–1115.

Komiya, K. and Okumura, M. (2012a). “Automatic Domain Adaptation for Word Sense Disambiguation Based on Comparison of Multiple Classifiers.” In *Proceedings of PAACLIC 2012*, pp. 77–85.

Komiya, K. and Okumura, M. (2012b). “Automatic Selection of Domain Adaptation Method for WSD using Decision Tree Learning.” *Journal of NLP (In Japanese)*, 19 (3), pp. 143–166.

Komiya, K., Suzuki, S., Sasaki, M., Shinnou, H., and Okumura, M. (2017). “Domain Adaptation for Word Sense Disambiguation using Word Embeddings.” In *Proceedings of CICLING 2017*, no. 57.

Kouno, K., Shinnou, H., Sasaki, M., and Komiya, K. (2015). “Unsupervised Domain Adaptation for Word Sense Disambiguation using Stacked Denoising Autoencoder.” In *Proceedings of PAACLIC-29*, pp. 224–231.

Kunii, S. and Shinnou, H. (2013). “Combined Use of Topic Models on Unsupervised Domain Adaptation for Word Sense Disambiguation.” In *Proceedings of PAACLIC-27*, pp. 224–231.

Maekawa, K., Yamazaki, M., Ogiso, T., Maruyama, T., Ogura, H., Kashino, W., Koiso, H., Yamaguchi, M., Tanaka, M., and Den, Y. (2014). “Balanced Corpus of Contemporary Written Japanese.” In *Language Resources and Evaluation*, Vol. 48, pp. 345–371.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). “Efficient Estimation of Word Representations in Vector Space.” In *Proceedings of ICLR Workshop 2013*, pp. 1–12.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013b). “Distributed Representations of Words and Phrases and their Compositionality.” In *Proceedings of NIPS 2013*, pp. 1–9.

Mikolov, T., tau Yih, W., and Zweig, G. (2013c). “Linguistic Regularities in Continuous Space Word Representations.” In *Proceedings of NAACL 2013*, pp. 746–751.

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

Nishio, M., Iwabuchi, E., and Mizutani, S. (1994). *Iwanami Kokugo Jiten Dai Go Han*. Iwanami Publisher, In Japanese.

Shinnou, H., Onodera, Y., Sasaki, M., and Komiya, K. (2015a). “Active Learning to Remove Source Instances for Domain Adaptation for Word Sense Disambiguation.” In *Proceedings
Komiya et al. Domain Adaptation using Word Embeddings for Word Sense Disambiguation

of PACLING-2015, pp. 224–231.

Shinnou, H., Sasaki, M., and Komiya, K. (2015b). “Learning under Covariate Shift for Domain Adaptation for Word Sense Disambiguation.” In Proceedings of PACLIC-29, pp. 215–223.

Sugawara, H., Takamura, H., Sasano, R., and Okumura, M. (2015). “Context Representation with Word Embeddings for WSD.” In Proceedings of PACLING 2015.

Taghipour, K. and Ng, H. T. (2015). “Semi-Supervised Word Sense Disambiguation Using Word Embeddings in General and Specific Domains.” In Proceedings of NAACL-HLT2015, pp. 314–323.

Yamaki, S., Shinnou, H., Komiya, K., and Sasaki, M. (2016). “Supervised Word Sense Disambiguation with Sentences Similarities from Context Word Embeddings.” In Proceedings of PACLIC 2016. Y16-1010.

Kanako Komiya: received her Ph.D. degree in Tokyo University of Agriculture and Technology in 2009. After being a postdoctoral fellow in Tokyo Institute of Technology and an assistant professor in Tokyo University of Agriculture and Technology, she is currently a lecturer in Ibaraki University. She is interested in natural language processing. She is a member of IPSJ, JSAI, and ANLP.

Minoru Sasaki: received his Ph.D. degree in Tokushima University. After being a research assistant in Ibaraki University, he is currently a lecturer in Ibaraki University. His research interests include information retrieval and natural language processing. He is a member of IPSJ and ANLP.

Hiroyuki Shinnou: received his Ph.D. degree in Tokyo Institute of Technology. After being a researcher in Fuji Xerox Co., Ltd. and Panasonic Corporation and a research assistant, a lecturer, and an associate professor in Ibaraki University, he is currently a professor in Ibaraki University. His research interests include Bayes statistics, machine learning and natural language processing. He is a member of IPSJ, JSAI, and ANLP.

Manabu Okumura: was born in 1962. He received B.E., M.E. and Dr. Eng. from Tokyo Institute of Technology in 1984, 1986, and 1989 respectively. He was an assistant at the Department of Computer Science, Tokyo Institute of Technology from 1989 to 1992, and an associate professor at the School of Information Science, Japan Advanced Institute of Science and Technology from 1992 to 2000. He is currently a professor at Institute of Innovative Re-
search, Tokyo Institute of Technology. His current research interests include
natural language processing, especially text summarization, computer assisted
language learning, sentiment analysis and text data mining.

(Received December 27, 2017)
(Revised April 25, 2018)
(Accepted June 18, 2018)