Proceedings of the 13th Workshop on Building and Using Comparable Corpora, pages 22–28
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

Automatic Creation of Correspondence Table of Meaning Tags from Two Dictionaries in One Language Using Bilingual Word Embedding

Teruo Hirabayashi*, Kanako Komiya*, Masayuki Asahara†, Hiroyuki Shinnou*
*Ibaraki University
4-12-1 Nakanarusawa, Hitachi, Ibaraki JAPAN
{18nm736g, kanako.komiya.nlp, hiroyuki.shinnou.0828}@vc.ibaraki.ac.jp

†National Institute for Japanese Language and Linguistics
10-2 Midoricho, Tachikawa, Tokyo, JAPAN
masayu-a@ninjal.ac.jp

Abstract

In this paper, we show how to use bilingual word embeddings (BWE) to automatically create a corresponding table of meaning tags from two dictionaries in one language and examine the effectiveness of the method. To do this, we had a problem: the meaning tags do not always correspond one-to-one because the granularities of the word senses and the concepts are different from each other. Therefore, we regarded the concept tag that corresponds to a word sense the most as the correct concept tag corresponding the word sense. We used two BWE methods, a linear transformation matrix and VecMap. We evaluated the most frequent sense (MFS) method and the corpus concatenation method for comparison. The accuracies of the proposed methods were higher than the accuracy of the random baseline but lower than those of the MFS and corpus concatenation methods. However, because our method utilized the embedding vectors of the word senses, the relations of the sense tags corresponding to concept tags could be examined by mapping the sense embeddings to the vector space of the concept tags. Also, our methods could be performed when we have only concept or word sense embeddings whereas the MFS method requires a parallel corpus and the corpus concatenation method needs two tagged corpora.

Keywords: Bilingual Word Embedding, Concept Embeddings, Word Embeddings, Dictionary

1. Introduction

Recently, corpora that have tags from more than one tag set are increasing. For example, “The Balanced Corpus of Contemporary Written Japanese” (BCCWJ) (Maekawa et al., 2014) is tagged with concept tags from “Word List by Semantic Principles” (WLSP) (National Institute for Japanese Language and Linguistics, 1964) after tagged with sense tags from “Iwanami Kokugo Jiten (Nishio et al., 1994).”

Because these tags are tagged referring to different dictionaries, the word senses of a word are different from each other. However, both tagging schemes are common in a way, that is, a unique meaning is given to every word in the corpus. Wu (Wu et al., 2019) created a corresponding table of word senses from Iwanami Kokugo Jiten and concept numbers form WLSP manually. If we could this process automatically, tagging of corpora would be much easier. Therefore, in this paper, we describe how to utilize bilingual word embeddings (BWE) to automatically create a corresponding table of meaning tags from two dictionaries in one language, Japanese, and examine the effectiveness of the method.

2. Related Work

BWE is classified into four groups according to how to make cross-lingual word embeddings [1]. First approach is monolingual mapping. These approaches initially train monolingual word embeddings and learn a transformation matrix that maps representations in one language to those of the other language. Mikolov et al. (Mikolov et al., 2013b) have shown that vector spaces can encode meaningful relations between words and that the geometric relations that hold between words are similar across languages. Because they do not assume the use of specific language, their method can be used to extend and refine dictionaries for any language pairs. Second approach is pseudo-cross-lingual. These approaches create a pseudo-cross-lingual corpus by mixing contexts of different languages. Xiao and Guo (Xiao and Guo, 2014) proposed the first pseudo-cross-lingual method that utilized translation pairs. They first translated all words that appeared in the source language corpus into the target language using Wiktionary. Then they filtered out the noises of these pairs and trained the model with this corpus in which these pairs are replaced with placeholders to ensure that translations of the same word have the same vector representation. Third approach is cross-lingual training. These approaches train their embeddings on a parallel corpus and optimize a cross-lingual constraint between embeddings of different languages that encourages embeddings of similar words to be close to each other in a shared vector space. Hermann and Blunsom (Hermann and Blunsom, 2014) trained two models to output sentence embeddings for input sentences in two different languages. They retrained these models with sentence embeddings using a least-squares method. Final approach is joint optimization. They not only consider a cross-lingual constraint, but also jointly optimize mono-lingual and cross-lingual objectives. Klementiev et al. (Klementiev et al., 2012) was the first research using joint optimization. Zou (Zou et al., 2013) used a matrix factorization approach to learn cross-lingual word representations for English and Chinese and utilized the representa-

[1] http://ruder.io/cross-lingual-embeddings/

tions for machine translation task. In this paper, we train BWE model by monolingual mapping and create a correspondence table of meaning tags using the model. To our knowledge, this research is the first research that uses BWE to find correspondences of meaning tags in one language.

3. Methods

Usually, BWE is used for cross-lingual applications, e.g., machine translation. The word embeddings trained from a parallel corpus, a comparable corpus, or two monolingual corpora are necessary for BWE. On the other hand, the number of corpora that were tagged by more than one tag sets is increasing. One corpus could have tags of part of speeches, word senses, named entities, and so on. We can regard a corpus that was tagged with two tag sets as a parallel corpus. For example, a corpus that was tagged with the meaning tags of two dictionaries in one language would be regarded as a parallel corpus of the meaning tag sets of two dictionaries.

In this research, we show how to utilize BWE to automatically find the correspondences of meaning tags in one language and investigate the effectiveness of the method. We generated two sets of word embeddings from a corpus with two meaning tags from different dictionaries. After that, we find correspondences of the meanings from two dictionaries using BWE. We used BCCWJ with concept tags from WLSP and sense tags from Iwanami Kokugo Jiten for the experiments. Both the word sense of Iwanami Kokugo Jiten and the concept number of WLSP represent a meaning of words and both of them are classified using a tree structure. The meaning tags do not always correspond one-to-one because the granularities of the word senses and the concepts are different from each other. However, the final purpose of this research is to automatically create a correspondence table between the word senses and the concept tags. We regarded the concept tag that corresponds to a word sense the most as the correct concept tag corresponding the word sense.

3.1. Sense Tags from Iwanami Kokugo Jiten

Iwanami Kokugo Jiten is a Japanese monolingual dictionary. In Iwanami Kokugo Jiten, each word sense has a sense tag such as "17877-0-0-1-0", composed of "headline ID"-"compound word ID"-"large classification ID"-"medium classification ID"-"small classification ID." When word sense has no corresponding ID, it would be 0. For example, the word senses and their corresponding sense tags of a word "子供 (child or children)" are listed in Table 1 [2].

Table 1: Word Senses and Their Corresponding Sense Tags of "子供 (Child or Children)" from Iwanami Kokugo Jiten[4]

Sense Tag	Word Sense
17877-0-0-1-0	<1> 幼い子。児童。 Young person. Someone who is not yet an adult. Kid.
17877-0-0-2-0	<2> 自分のもうけた子。むすこ、むすめ。子。 Son/daughter. A son or daughter of any age.

Figure 1 shows the tree structure of Iwanami Kokugo Jiten. In this research, we used Annotated Corpus of Iwanami Japanese Dictionary Fifth Edition 2004, which is BCCWJ tagged with Iwamnami Kokugo Jiten, provided Gengo Shigen Kyokai, or Language Resource Academy [6].

3.2. Concept Tags from WLSP

WLSP is a Japanese thesaurus in which a word is classified and ordered according to its meaning. One record is composed of the following elements, record ID number, lemma number, type of record, class, division, section, article, concept number, paragraph number, small paragraph number, word number, lemma with explanatory note, lemma without explanatory note, reading and reverse reading. Concept number consists of a category, a medium item and a classification item. We used concept numbers as the concept tags. For example,"子供 (child or children)" is polyseme and two concepts are registered in WLSP, which are "1.2050" and "1.2130" (Table 2). This paper utilizes a corpus that is in its infancy, namely BCCWJ annotated with concept tags or concept numbers of WLSP.

The goal of our research is to find the correspondences of the meaning tags from two dictionaries. In the example of "子供 (child or children)," we think that the word senses "17877-0-0-1-0" and "17877-0-0-2-0" in Iwanami Kokugo Jiten respectively correspond to concepts "1.2050" and "1.2130" in WLSP, however, please note that the meaning tags do not always correspond one-to-one. We utilized only two sets of meaning tag from BCCWJ and did not use the reference source: the dictionaries.

Figure 2 shows the tree structure of WLSP.

3.3. Bilingual Word Embeddings

We used monolingual mapping. Monolingual mapping consists of two steps. First, monolingual word embeddings are trained for each language. In our research, one language corresponds to one meaning tag set in Japanese. After that, they are mapped to a common vector space so that word embeddings of the words whose meanings are similar to each other in two languages can be brought closer. Because the geometrical relations that hold between words are similar across languages, it is possible to transform a vector space of a language to that of another language using a linear projection. In this research, we adapted two methods of BWE, linear transformation matrix and VecMap. A

[2]We eliminated the compound words from the dictionary.

[4]English translations in Table 1 are quoted from Longman Dictionary of Contemporary English [5].

[6]https://www.gsk.or.jp/catalog/gsk2010-a/

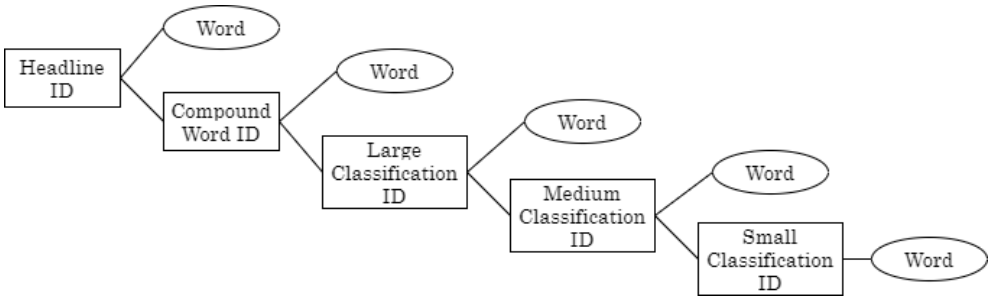

Figure 1: Tree Structure of Iwanami Kokugo Jiten

Table 2: Concept-tags and Their Corresponding Class, Division, Section of "子供 (child or children)" from WLSP

Concept number	Class	Division	Section	Article
1.2050	Nominal words	Agent	Human	Young or old
1.2130	Nominal words	Agent	Family	Child or descendant

linear projection matrix W was learned when we used a linear transformation matrix. VecMap is an implementation of a framework of Artetxe et al. to learn cross-lingual word embedding mappings (Artetxe et al., 2017)(Artetxe et al., 2018a)(Artetxe et al., 2018b).

4. Experiment

4.1. Experimental Setting

We utilized BCCWJ tagged with word senses of Iwanami Kokugo Jiten and BCCWJ tagged with concepts of WLSP. Table 3 shows the number of word tokens, unique words, unique word senses, and unique concepts.

Table 3: Statistic Data of BCCWJ

Number of Word tokens	340,995
Number of Unique Words	25,321
Number of Unique Word Senses	26,713
Number of Unique Concepts	3,164

The settings of word2vec are shown in Table 4. We used C-Bow algorithm and we set the number of dimensions as 200, the window size as 5, the number of iterations as 5, the batch size as 1,000, and the min-count as 1, respectively. We set the min-count as 1 because the corpus size was small.

Table 4: Settings of word2vec

Parameters	Settings
Dimensionality	200
Learning Algorithm	C-BoW
Window Size	5
Number of Epochs	5
Batch Size	1,000
min-count	1

4.1.1. Linear Transformation Matrix

When a linear project matrix is learnt, we conduct experiments as follows.

1. Generate a word-sense-tag and concept-tag corpora respectively, and learn word-sense or concept embeddings for each corpus from them using word2vec [7] (Mikolov et al., 2013a; Mikolov et al., 2013c; Mikolov et al., 2013d) (cf. Figure 3).

2. Learn a linear projection matrix W from the vector space of the word-senses to that of the concepts using pairs of the embeddings for monosemous common nouns, which are generated in the last step.

3. Apply the matrix W to the word-sense embeddings and obtain the projected concept embeddings for them.

We defined a monosemous word as a word that meets two conditions, which are, (1) it has only one sense in Iwanami Kokugo Jiten and (2) it does not have any concept number in WLSP. We chose them because the concepts in WLSP are like synsets in English WordNet; many words share a concept. Therefore, if a word has a concept number, we cannot treat the word as monosemous word because we generated word embeddings for each concept number. We used 104 monosemous common nouns as seed words of our experiments. We randomly extracted ten words for evaluation data and used other 94 words for the training data to obtain the number of epochs that minimize the loss. We iterated this operation for 20 times and used the average number of epochs for the number of epochs of the final experiment.
Table 5 shows learning parameters of the linear transformation matrix.

Table 5: Learning Parameters of Linear Transformation Matrix

Parameters	Settings
Dimensionality	200 × 200
Optimization Algorithm	Adam
Number of Epochs	118

[7] https://code.google.com/archive/p/word2vec/

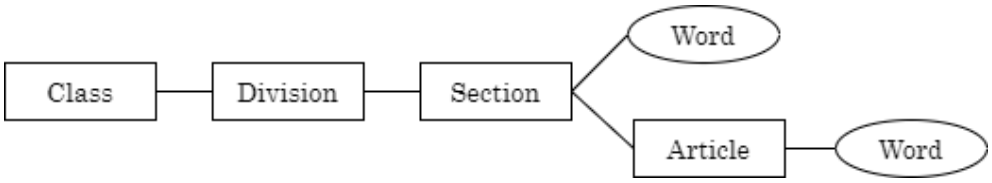

Figure 2: Tree Structure of WLSP

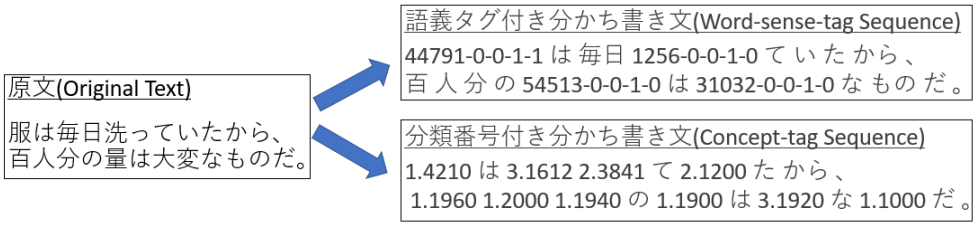

Figure 3: Word-sense-tag and Concept-tag Sentences

4.1.2. VecMap

VecMap [8] is used for the second method of BWE. When we used a linear transformation matrix, we projected the vector space of word senses of Iwanami Kokugo Jiten into that of concepts of WLSP. However, VecMap projects both the vector spaces of word senses and concepts into a new vector space. The three options, supervised, semi-supervised, and identical, were compared. Supervised and semi-supervised VecMap utilize the specified words but Identical VecMap uses identical words in two languages as the seeds of the projection. Therefore, the seed words of supervised and semi-supervised VecMap are the same as the linear transformation matrix but that of identical VecMap is different from it. The seed words of identical VecMap is monosemous words whereas those of supervised or semi-supervised VecMap is monosemous common nouns. The number of monosemous words, the seed words of identical VecMap, is 2,015. We used default settings for the tool of VecMap for each option. Table 6 lists the default settings of the parameters of each specific option and the general default settings of them.

4.1.3. Evaluation

We evaluated the correspondences of the meaning tags as follows.

1. Calculate the cosine similarities between the projected concept embeddings and the embeddings of the concepts from the target word.
2. Choose the concepts that have the highest similarities to the projected concept embeddings as the corresponding concepts for the word senses.
3. Calculate the accuracy.

We targeted at polysemous nouns that appeared equal to or more than 50 times in the corpus. They were nine words, which were, “関係 (relationship)”, “技術 (technology)”, “現場 (field)”, “子供 (child)”, “時間 (time) ”, “市場 (market)”, “電話 (phone)”, “場所 (place)”, and “前 (before)” and their word senses were 25 in total. We regarded an estimated concept tag to be correct when it is the same to the tag aligned with its corresponding sense tag most frequently in the tagged corpus. We evaluated the most frequent sense (MFS) accuracy for comparison. For MFS, the most frequent concept from WLSP for each word type in a corpus was regarded as the corresponding concept number for all the word senses for the word from Iwanami Kokugo Jiten. Also, we tested another comparative method, which is “concatenation corpus method;” a concept sequence corpus and a word sense sequence corpus are concatenated, and the concept embeddings and the word sense embeddings were generated together at the same time.

[8] https://github.com/artetxem/vecmap#publications

4.2. Results

Table 7 shows the accuracies of the corresponding meanings. Thirteen out of 25 word senses were aligned with the correct concept tags by a linear transformation matrix, and the accuracy was 52.0%. The results of VecMap were 36.0%, 48.0%, and 48.0% when supervised, semi-supervised, and identical options were used. In the comparative experiment, 16 out of 25 word senses were aligned with the correct concept tags by both the MFS and corpus concatenation methods, and the accuracy was 64.0%. The accuracy of the random baseline, which is the method where each word sense was chosen at random, was 41.5%. The list of concept tags estimated by the linear transformation matrix, i.e., the best method of BWE in Table 7, the MFS and corpus concatenation methods, and the oracle for 25 word senses of 9 words are shown in Table 8. The correct concept tags are shown in bold. “X-X-X-X” in the word sense of “前 (before)” means a new word sense not listed in a dictionary, and in this research, it was considered as one of the word sense of the experiment.

5. Discussion

According to Table 7, the accuracies of the proposed the methods were lower than the accuracy of the MFS method and the corpus concatenation method. However, as mentioned above, in reality, one concept tag does not always

Table 6: Parameters of VecMap

Option	Parameter	Default Setting of Specific Option	General Default Setting
Supervised	Batch size	1000	10000
Semi-supervised	Self-Learning	TRUE	FALSE
Semi-supervised	Vocabulary_cutoff	200,000	0
Semi-supervised	csls_neibourhood	10	0
Identical	Self-Learning	TRUE	FALSE
Identical	Vocabulary_cutoff	200,000	0
Identical	csls_neibourhood	10	0

Table 7: Accuracies of Each Method

Method	Accuracy
Linear Transformation Matrix	52.0 %
VecMap Supervised	36.0 %
VecMap Semi-supervised	48.0 %
VecMap Identical	48.0 %
MFS	64.0 %
Corpus Concatenation	64.0 %
Random	41.5 %

correspond one sense tag. Sometimes one concept tag corresponds to plural sense tags and vice versa. We chose to make one-to-one correspondence for simplicity. From this perspective, the proposed methods have an advantage: the relations of the sense tags corresponding to a concept tag can be examined by mapping the sense embeddings to the vector space of the concept tags. Since the corpus concatenation method also uses word2vec, it also examine the relations of the sense tags but our method could be performed when we have only concept or word sense embeddings and do not have any tagged corpora.

Also, in this research, we conducted the experiments using a corpus where two kinds of meaning tags are assigned. However, it is possible to use two different corpora for two meaning tag sets for our proposed methods, the use of BWE. In other words, we can conduct the experiments using two corpora, for example, a corpus assigned with concept tags from WLSP and another corpus assigned with word senses from Iwanami Kokugo Jiten. In that case, comparable corpora would be better than two monolingual corpora for BWE because the meanings of words should be similar to each other. Also, the accuracies may be lower when we use different two corpora because words do not share the contexts in two monolingual corpora. Furthermore, it is desirable to use a relatively large corpus for the experiments in this research because only the concepts or word senses of words appeared in the corpus are able to have a corresponding meaning.

In this research, the experiments were performed on words that appeared 50 times or more in the corpus, but when the number of occurrences for each word sense was counted, there were four word senses that appeared only once. Since we used word2vec tool, it is preferable to use a corpus where all the meanings appear more than the threshold value [9]. We had a hypothesis that relatively large number of examples are required to generate meaning embeddings. Therefore, we examined how the correspondence accuracies between the word senses and the concepts differ depending on the occurrences of the word senses in the corpus. Figure 4 shows correct and incorrect numbers of the examples according to the occurrences of the word senses. For this figure, 25 word senses were grouped by occurrences so that each group has 5 word senses. The numbers of correct and incorrect answers are plotted on the vertical axis for each group and these groups are shown in order of the decreasing occurrences. The label of the bar graph in Figure 4 indicates "minimum number of occurrences in each group" − "maximum number of occurrences in each group".

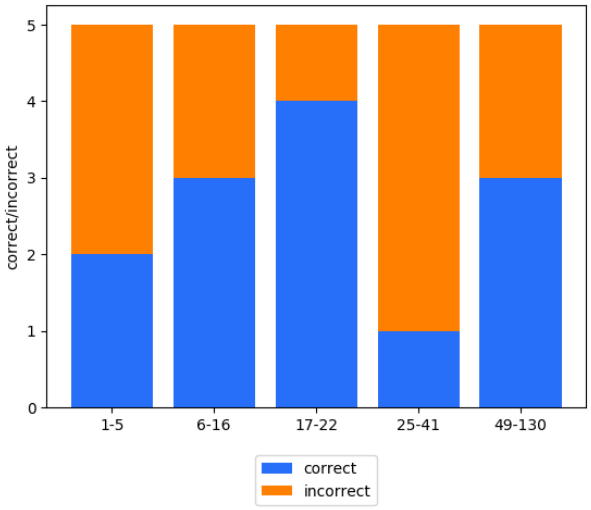

Figure 4: Numbers of correct and incorrect answers according to occurrences of the word senses

Despite our hypothesis, according to Figure 4, there was no correlation between the occurrences of the word senses and the correspondence accuracies in this research.

Because both the concept tags and the word sense tags were manually annotated on BCCWJ, the accuracies of annotations are very high. However, since there are still few corpora with which two or more types of tags are assigned, we plan to use a tagger to automatically tag one type of meaning tags on a corpus with another type of meaning tags for the preprocessing of the proposed method for future work. However, in this case, the performance of the

[9] Word2vec generates vectors only for the word (word senses or concepts in this research) that appeared equal to or more than a threshold value. Default setting is five. We set this value to one to acquire meaning vectors for the words that appeared only once.

Table 8: Correspondence Table of "Iwanami Kokugo Jiten" and "WLSP"

Words	Word Numbers	Word Senses	Concept Numbers			
			Linear transformation Matrix	MFS	Corpus Concatenation	Oracle
関係 (relationship)	9667	0-0-1-0	**1.1110**	**1.1110**	**1.1110**	1.1110
		0-0-2-0	1.3500	**1.1110**	**1.1110**	1.1110
		0-0-3-0	**1.1110**	**1.1110**	**1.1110**	1.1110
技術 (technology)	10703	0-0-1-0	1.3421	**1.3850**	**1.3850**	1.3850
		0-0-2-0	**1.3421**	1.3850	1.3850	1.3421
現場 (field)	15615	0-0-1-0	1.2620	1.2620	**1.1700**	1.1700
		0-0-2-0	**1.2620**	**1.2620**	**1.2620**	1.2620
子供 (child)	17877	0-0-1-0	1.2130	**1.2050**	1.2130	1.2050
		0-0-2-0	**1.2130**	1.2050	**1.2130**	1.2130
時間 (time)	20676	0-0-1-0	**1.1600**	**1.1600**	1.1962	1.1600
		0-0-2-0	**1.1962**	1.1600	**1.1962**	1.1962
		0-0-3-0	**1.1600**	**1.1600**	**1.1600**	1.1600
		0-0-4-0	1.1962	**1.1600**	**1.1600**	1.1600
市場 (market)	21128	0-0-1-0	1.2600	1.2600	**1.2640**	1.2640
		0-0-2-0	**1.2600**	**1.2600**	1.2640	1.2600
		0-0-3-0	**1.2600**	**1.2600**	1.2640	1.2600
電話 (phone)	35881	0-0-1-0	1.4620	**1.3122**	**1.3122**	1.3122
		0-0-2-0	**1.4620**	1.3122	**1.4620**	1.4620
場所 (place)	41150	0-0-1-0	1.3833	**1.1700**	**1.1700**	1.1700
		0-0-2-0	1.3833	**1.1700**	1.3833	1.1700
前 (before)	48488	0-0-1-1	**1.1740**	1.1670	**1.1740**	1.1740
		0-0-2-0	1.1650	1.1670	**1.1740**	1.1740
		0-0-2-1	1.1635	**1.1670**	1.1635	1.1670
		0-0-2-2	1.1635	**1.1670**	1.1740	1.1670
		X-X-X-X	**1.1635**	1.1670	1.1650	1.1635

tagger should be considered to guarantee the quality of the automatic tagged corpus.

6. Conclusion

In this research, we described how to utilize bilingual word embeddings to obtain the correspondences of meanigs from two dictionaries in one language and investigated the effectiveness of the method. We used BCCWJ with concept tags from WLSP and sense tags from Iwanami Kokugo Jiten for the experiments. The experiments showed that the correspondence accuracies of the proposed methods were lower than MFS baseline or the corpus concatenation method. However, because our method utilizes the embedding vectors of the word senses, the relation of the sense tags corresponding to concept tags can be examined by mapping the sense embeddings to the vector space of the concept tags. Also, our method could be performed when we have only concept or word sense embeddings. However, it is necessary to expand the corpus for the further evaluation because the proposed method uses one corpus for both the training and the test and only the word senses or the concepts that appeared in the corpus are able to have correspondence. In addition, we would like to investigate further how the accuracy of this study changes when the corpus is expanded.

Acknowledgements

This work was supported by JSPS KAKENHI Grants Number 18K11421, 17H00917, and a project of the Center for Corpus Development, NINJAL.

References

Artetxe, M., Labaka, G., and Agirre, E. (2017). Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 451–462.

Artetxe, M., Labaka, G., and Agirre, E. (2018a). Generalizing and improving bilingual word embedding mappings with a multi-step framework of linear transformations. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, pages 5012–5019.

Artetxe, M., Labaka, G., and Agirre, E. (2018b). A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 789–798.

Hermann, K. M. and Blunsom, P. (2014). Multilingual models for compositional distributed semantics. *arXiv preprint arXiv:1404.4641*.

Klementiev, A., Titov, I., and Bhattarai, B. (2012). Inducing crosslingual distributed representations of words. *Proceedings of COLING 2012*, pages 1459–1474.

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. *Language resources and evaluation*, 48(2):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*, pages 1–12.

Mikolov, T., Le, Q. V., and Sutskever, I. (2013b). Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013c). Distributed representations of words and phrases and their compositionality. In *Proceedings of NIPS 2013*, pages 1–9.

Mikolov, T., tau Yih, W., and Zweig, G. (2013d). Linguistic regularities in continuous space word representations. In *Proceedings of NAACL 2013*, pages 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.

Wu, P., Kondo, M., Moriyama, N., Ogiwara, A., and Asahara, M. (2019). Alignment table between ‘word list by semantic principles’ and ‘annotated corpus of iwanami japanese dictionary fifth edition 2004’. In *Proceedings of the Language Resources Workshop 2019*, pages 337–342.

Xiao, M. and Guo, Y. (2014). Distributed word representation learning for cross-lingual dependency parsing. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pages 119–129.

Zou, W. Y., Socher, R., Cer, D., and Manning, C. D. (2013). Bilingual word embeddings for phrase-based machine translation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1393–1398.