Construction of a Japanese Word Similarity Dataset

Yuya Sakaizawa and Mamoru Komachi
Tokyo Metropolitan University
6-6 Asahigaoka
Hino city, Tokyo 191-0065, Japan
{sakaizawa-yuya@ed.,komachi@}tmu.ac.jp

Abstract

An evaluation of distributed word representation is generally conducted using a word similarity task and/or a word analogy task. There are many datasets readily available for these tasks in English. However, evaluating distributed representation in languages that do not have such resources (e.g., Japanese) is difficult. Therefore, as a first step toward evaluating distributed representations in Japanese, we constructed a Japanese word similarity dataset. To the best of our knowledge, our dataset is the first resource that can be used to evaluate distributed representations in Japanese. Moreover, our dataset contains various parts of speech and includes rare words in addition to common words.

Keywords: word embeddings, distributed representation, word similarity

1. Introduction

Traditionally, a word is represented as a sparse vector indicating the word itself (one-hot vector) or the context of the word (distributional vector). However, both the one-hot notation and distributional notation suffer from data sparseness since dimensions of the word vector do not interact with each other. Distributed word representation addresses the data sparseness problem by constructing a dense vector of a fixed length, wherein contexts are shared (or distributed) across dimensions. Distributed word representation is known to improve the performance of many NLP applications such as machine translation (Chen and Guo, 2015) and sentiment analysis (Tai et al., 2015) to name a few. The task to learn a distributed representation is called representation learning.

However, evaluating the quality of learned distributed word representation itself is not straightforward. In language modeling, perplexity or cross-entropy is widely accepted as a de facto standard for intrinsic evaluation. In contrast, distributed word representations include the additive (or compositional) property of the vectors, which cannot be assessed by perplexity. Moreover, perplexity makes little use of infrequent words; thus, it is not appropriate for evaluating distributed presentations that try to represent them.

Therefore, a word similarity task and/or a word analogy task are generally used to evaluate distributed word representations in the NLP literature. The former judges whether distributed word representations improve modeling contexts, and the latter estimates how well the learned representations achieve the additive property. However, such resources other than for English (e.g., Japanese) seldom exist. In addition, most of these datasets comprise high-frequency nouns so that they tend not to include other parts of speech. Hence, previous data fail to evaluate word representations of other parts of speech, including content words such as verbs and adjectives.

To address the problem of the lack of a dataset for evaluating Japanese distributed word representations, we propose to build a Japanese dataset for the word similarity task.

The main contributions of our work are as follows:

- To the best of our knowledge, it is the first work that constructs a Japanese word similarity dataset.
- The dataset contains various parts of speech and includes rare words in addition to common words.

2. Related Work

In general, distributed word representations are evaluated using a word similarity task. For instance, WordSim353 (Finkelstein et al., 2002), MC (Miller and Charles, 1991), RG (Rubenstein and Goodenough, 1965), and SCWS (Huang et al., 2012) have been used to evaluate word similarities in English. Moreover, Baker et al. (2014) built a verb similarity dataset (VSD) based on WordSim353 because there was no dataset of verbs in the word-similarity task. Recently, SimVerb-3500 was introduced to evaluate human understanding of verb meaning (Gerz et al., 2016). It provides human ratings for the similarity of 3,500 verb pairs so that it enables robust evaluation of distributed representation for verbs. However, most of these datasets include English words only. There has been no Japanese dataset for the word-similarity task.

Apart from English, WordSim353 and SimLex-999 (Hill et al., 2015) have been translated and rescored in other languages: German, Italian and Russian (Leviant and Reichart, 2015). SimLex-999 has also been translated and rescored in Hebrew and Croatian (Mrksic et al., 2017). SimLex-999 explicitly targets at similarity rather than relatedness and includes adjective, noun and verb pairs. However, this dataset contains only frequent words.

In addition, the distributed representation of words is generally learned using only word-level information. Consequently, the distributed representation for low-frequency words and unknown words cannot be learned well with conventional models. However, low-frequency words and unknown words are often comprise high-frequency morphemes (e.g., unkingly → un + king + ly). Some previous studies take advantage of the morphological information to provide a suitable representation for low-frequency words and unknown words (Luong et al., 2013; Soricut and Och,

Currently at JustSystems Corporation.

Sentence	I don't think it is likely to not include these people, or [exclude] まさかこういった方々を対象としない、[排除する]わけではないと思いますが				
Paraphrase	ignore 無視する	ostracize 排斥する	avoid 敬遠する	exclude 排除する	remove 除外する

Figure 1: An example of the dataset from a previous study (Kodaira et al., 2016).

Frequency	1-	101-	1001-	10001-
Verb	239	539	710	598
Adjective	183	322	523	350
Noun	15	63	172	258
Adverb	23	75	80	81

Table 1: The number of parts of speech classified into each frequency.

word 1		word 2		sim.
EN	JA	EN	JA	
close	瞑る	close	つぶる	10
erase	拭き取る	wipe	拭う	8
mopey	塞ぎ込んだ	sick	病んだ	5
investigate	手探る	go	行く	2
fly	とばせる	control	制御できる	0

Table 2: Example of the degree of similarity when we requested annotation at Lancers.

2015). Morphological information is particularly important for Japanese since Japanese is an agglutinative language.

3. Construction of a Japanese Word Similarity Dataset

What makes a pair of words similar? Most of the previous datasets do not concretely define the similarity of word pairs. The difference in the similarity of word pairs originates from each annotator's mind, resulting in different scales of a word. Thus, we propose to use an example-based approach (Table 2) to control the variance of the similarity ratings. We remove the context of word when we extracted the word. So, we consider that an ambiguous word has high variance of the similarity, but we can get low variance of the similarity when the word is monosemous.

For this study, we constructed a Japanese word similarity dataset[1]. We followed the procedure used to construct the Stanford Rare Word Similarity Dataset (RW) (Luong et al., 2013).

We extracted Japanese word pairs from the Evaluation Dataset of Japanese Lexical Simplification (Kodaira et al., 2016). It targeted content words (nouns, verbs, adjectives, adverbs). It included 10 contexts about target words annotated with their lexical substitutions and rankings. Figure 1 shows an example of the dataset. A word in square brackets in the text is represented as a target word of simplification. A target word is not only recorded in the lemma form but also in the conjugated form. We built a Japanese similarity dataset from this dataset using the following procedure.

Word selection: First, paraphrase candidates were extracted from this dataset. Because the construction process of the simplification dataset was divided into a paraphrase acquisition phase and a simplification ranking phase, we simply discarded the simplification rankings from the dataset to obtain paraphrase candidates. Table 1 shows the frequency of extracted words in the Japanese Wikipedia as of May 2015. As shown in the table, low-frequency words are included in the dataset.

Pair construction: Because extracted words are annotated with their paraphrase candidates, we picked up each pair from the candidate as a word pair. Consequently, we acquired 5,051 verb pairs, 4,033 adjective pairs, 1,528 noun pairs and 902 adverb pairs. To balance the numbers of verb and adjective pairs with other parts of speech, we extracted samples at random for verbs and adjectives. Finally, we obtained 1,464 verb pairs and 960 adjective pairs.

We observed that the similarity of the pairs extracted from the dataset of Kodaira et al. (2016) was low without providing contexts; thus, we did not augment the dataset by inserting pseudo-negative instances from WordNet's synsets, as was done in the RW corpus. Another reason why we did not employ the synset from the Japanese WordNet (Isahara et al., 2008) was because its quality was not as good as the English WordNet except for concrete nouns[2].

Human judgment: We opted to use the crowd-sourcing service (Lancers[3]) to hire native Japanese speakers. We asked annotators to assign the degree of similarity for each pair using the same 10-point scale[4]. We used only those annotators who were able to complete at least 95% of their previous assignments correctly. We collected similarity rating for each word pair from ten annotators and defined the average of their annotations as the similarity of the pairs.

Although (Kodaira et al., 2016) gave the annotators the context during annotation, we removed the context and gave only pairs to annotators. We did so because the previous datasets such as VSD and RW did not present any context during annotation[5]. To improve the quality of the annotation, we presented an example of the degree of similarity

[1] https://github.com/tmu-nlp/JapaneseWordSimilarityDataset

[2] It might be because it was translated from the English WordNet. This is why we decided not to translate the existing English word similarity dataset to create a Japanese version.

[3] http://www.lancers.jp

[4] In a crowdsourcing request, we indicated that a similarity of pairs with different notations, such as "write（書いた）" and "write（かいた）" is 10.

[5] Another reason why we did not do so is because the SCWS has a very high variance even though it is annotated with contexts (Table 5).

POS	verb	adj	adv	noun
IAA	0.69	0.67	0.61	0.56

Table 3: Inter-annotator agreements of each POS.

of the pairs during annotation (Table 2). Consequently, we collected 4,851 pairs overall. Table 4 shows an example of a pair from our dataset. Inter-annotator agreements (IAA) of each POS are shown in Table 3. The inter-annotator agreement is the average Spearman's ρ between a single annotator and the average of all others.

4. Discussion

4.1. Comparison to Other Datasets

Table 5 shows how several resources vary. WordSim353 comprises high-frequency words and so the variance tends to be low. In contrast, RW includes low-frequency words, unknown words, and complex words composed of several morphemes; thus, the variance is large. VSD has many polysemous words, which increase the variance. Despite the fact that our dataset, similar to the VSD and RW datasets, contains low-frequency and ambiguous words, its variance is 3.00. The variance level is low compared with the other corpora. We considered that the examples of the similarity in the task request reduced the variance level.

We did not expect SCWS to have the largest variance in the datasets shown in Table 5 because it gave the context to annotators during annotation. At the beginning, we thought the context would serve to remove the ambiguity and clarify the meaning of word; however after looking into the dataset, we determined that the construction procedure used several extraordinary annotators. It is crucial to filter insincere annotators and provide straightforward instructions to improve the quality of the similarity annotation like we did. To gain better similarity, each dataset should utilize the reliability score to exclude extraordinary annotators. For example, for SCWS, an annotator rating the similarity of pair of "CD" and "aglow" assigned a rating of 10. We assumed it was a typo or misunderstanding regarding the words. To address this problem, such an annotation should be removed before calculating the true similarity. All the datasets except for RW simply calculated the average of the similarity, but datasets created using crowdsourcing should consider the reliability of the annotator.

4.2. Analysis

We present examples of a pair with high variance of similarity as shown below:

Aspect of relatedness. (e.g., a pairing of "fast（速い）" and "early（早い）".)

Although they are similar in meaning with respect to the time, they have nothing in common with respect to speed; Annotator A assigned a rating of 10, but Annotator B assigned a rating of 1.

Another example, the pairing of "be eager（懇願する）" and "request（頼む）". Even though the act indicated by the two verbs is the same, there are some cases where they express different degrees of feeling. Compared with "request", "eager" indicates a stronger feeling. There were two annotators who emphasized the similarity of the act itself rather than the different degrees of feeling, and vice versa. In this case, Annotator A assigned a rating of 9, but Annotator B assigned a rating of 2.

Although it was necessary to distinguish similarity and semantic relatedness (Mrksic et al., 2016) and we asked annotators to rate the pairs based on semantic similarity, it was not straightforward to put paraphrase candidates onto a single scale considering all the attributes of the words. This limitation might be relaxed if we would ask annotators to refer to a thesaurus or an ontology such as Japanese Lexicon (Ikehara et al., 1997).

Comparing spell[6]**.** (e.g., a pairing of "slogan（スローガン）" and "slogan（標語）".)

In Japanese, we can write a word using hiragana, katakana, or kanji characters; however because hiragana and katakana represent only the pronunciation of a word, annotators might think of different words. In this case, Annotator A assigned a rating of 8, but Annotator B assigned a rating of 0. Similarly, we confirmed the same thing in other parts of speech. Especially, nouns can have several word pairs with different spellings, which results in their IAA became too low compared to other parts of speech.

Frequency or time expressions. (e.g., a pairing of "often（しばしば）" and "frequently（しきりに）".)

We confirmed that the variance becomes larger among adverbs expressing frequency. This is due to the difference in the frequency of words that annotators imagines. In this case, Annotator A assigned a rating of 9, but Annotator B assigned a rating of 0. Similarly, we confirmed the same thing among adverbs expressing time.

5. Conclusion

In this study, we constructed the first Japanese word similarity dataset. It contains various parts of speech and includes rare words in addition to common words. Crowdsourced annotators assigned similarity to word pairs during the word similarity task. We gave examples of similarity in the task request sent to annotators, so that we reduced the variance of each word pair. However, we did not restrict the attributes of words, such as the level of feeling, during annotation. Error analysis revealed that the notion of similarity should be carefully defined when constructing a similarity dataset.

As a future work, we plan to construct a word analogy dataset in Japanese by translating an English dataset to Japanese. We hope that a Japanese database will facilitate research in Japanese distributed representations.

6. Bibliographical References

Baker, S., Reichart, R., and Korhonen, A. (2014). An Unsupervised Model for Instance Level Subcategorization Acquisition. In *Proceedings of the 2014 Conference*

[6]We indicated these pair's similarity is 10. However, some annotators ignored this instruction. It would be necessary to clean the spellings of paraphrase candidates before requesting similarity annotation.

word 1	EN	follow	exclude	challenge	storm	elucidate	wander
	JA	受け継ぐ	除外する	チャレンジする	しける	明白になる	迷う
word 2	EN	inherit	remove	wish	rough	reflect	stop
	JA	継承する	除去する	望む	あれる	反映される	止める
similarity		9.3	7.3	6.0	5.7	2.7	1.7

Table 4: Examples of verb pairs in our dataset. The similarity rating is the average of the ratings from ten annotators.

Dataset	Variance
WordSim353	3.16
VSD	4.76
RW	5.70
SCWS	8.60
JWSD (our dataset)	3.00

Table 5: Variance of each dataset.

on Empirical Methods in Natural Language Processing (EMNLP), pages 278–289.

Chen, B. and Guo, H. (2015). Representation Based Translation Evaluation Metrics. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, pages 150–155.

Isahara, H., Bond, F., Uchimoto, K., Utiyama, M., and Kanzaki, K. (2008). Development of the Japanese WordNet. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC)*, pages 2420–2423.

Luong, M.-T., Socher, R., and Manning, C. D. (2013). Better Word Representations with Recursive Neural Networks for Morphology. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning (CoNLL)*, pages 104–113.

Miller, G. A. and Charles, W. G. (1991). Contextual Correlates of Semantic Similarity. *Language and Cognitive Processes*, 6(1):1–28.

Mrksic, N., Séaghdha, D. Ó., Thomson, B., Gasic, M., Rojas-Barahona, L. M., Su, P., Vandyke, D., Wen, T., and Young, S. J. (2016). Counter-fitting Word Vectors to Linguistic Constraints. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computatinal Linguistics: Human Language Technologies (NAACL-HLT)*, pages 142–148.

Rubenstein, H. and Goodenough, J. B. (1965). Contextual Correlates of Synonymy. *Communications of the ACM*, 8(10):627–633.

Soricut, R. and Och, F. (2015). Unsupervised Morphology Induction Using Word Embeddings. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 1627–1637.

Tai, K. S., Socher, R., and Manning, C. D. (2015). Improved Semantic Representations From Tree-Structured Long Short-Term Memory Networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, pages 1556–1566.

7. Language Resource References

Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., and Ruppin., E. (2002). Placing Search in Context: The Concept Revisited. *ACM Transactions on Information Systems (TOIS)*, 20(1):116–131.

Gerz, D., Vulic, I., Hill, F., Reichart, R., and Korhonen, A. (2016). SimVerb-3500: A Large-Scale Evaluation Set of Verb Similarity. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2173–2182.

Hill, F., Reichart, R., and Korhonen, A. (2015). Simlex-999: Evaluating Semantic Models with (Genuine) Similarity Estimation. *Computational Linguistics*, 41(4):665–695.

Huang, E. H., Socher, R., Manning, C. D., and Ng, A. Y. (2012). Improving Word Representations via Global Context and Multiple Word Prototypes. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 873–882.

Ikehara, S., Miyazaki, M., Shirai, S., Yokoo, A., Nakaiwa, H., Ogura, K., Ooyama, Y., and Hayashi, Y. (1997). *A Japanese Lexicon*. Iwanami Shoten.

Kodaira, T., Kajiwara, T., and Komachi, M. (2016). Controlled and Balanced Dataset for Japanese Lexical Simplification. In *Proceedings of the ACL 2016 Student Research Workshop*, pages 1–7.

Leviant, I. and Reichart, R. (2015). Judgment Language Matters: Multilingual Vector Space Models for Judgment Language Aware Lexical Semantics. *CoRR*, abs/1508.00106.

Luong, M.-T., Socher, R., and Manning, C. D. (2013). Better Word Representations with Recursive Neural Networks for Morphology. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning (CoNLL)*, pages 104–113.

Mrksic, N., Vulic, I., Séaghdha, D. Ó., Leviant, I., Reichart, R., Gasic, M., Korhonen, A., and Young, S. J. (2017). Semantic Specialisation of Distributional Word Vector Spaces using Monolingual and Cross-Lingual Constraints. *Transactions of the Association for Computational Linguistics*, 5:309–324.