Segmentation-free compositional n-gram embedding (supplementary material)

Geewook Kim and **Kazuki Fukui** and **Hidetoshi Shimodaira**
Department of Systems Science, Graduate School of Informatics, Kyoto University
Mathematical Statistics Team, RIKEN Center for Advanced Intelligence Project
{geewook, k.fukui}@sys.i.kyoto-u.ac.jp, shimo@i.kyoto-u.ac.jp

A Experiment details

A.1 Hyperparameters tuning

For skipgram, we performed a grid search over $(h, \gamma) \in \{1, 5, 10\} \times \{0.01, 0.025\}$, where h is the size of context window and γ is the initial learning rate. For sisg, we performed a grid search over $(h, \gamma, n_{\min}, n_{\max}) \in \{1, 5, 10\} \times \{0.01, 0.025\} \times \{1, 3\} \times \{4, 8, 12\}$, where h is the size of context window, γ is the initial learning rate, n_{min} is the minimum length of character n-gram and n_{max} is the maximum length of character n-gram. For pv-dbow, pv-dm and sent2vec, we performed a grid search over $(h, \gamma) \in \{5, 10\} \times \{0.01, 0.05, 0.1, 0.2, 0.5\}$, where h is the size of context window and γ is the initial learning rate. For sembei and scne, we used the initial learning rate 0.01 and $n_{\min} = 1$. The maximum length of n-gram to consider $n_{\max}$ is grid searched over $\{4, 6, 8\}$ in the word and sentence similarity tasks. In the noun category prediction task, we used $n_{\max} = 8$ for sembei and the $n_{\max}$ of scne is grid searched over $\{4, 6, 8\}$. For sentiment analysis task, we tested both $n_{\max} = 8$ and $n_{\max} = 16$ for sembei and scne to see the effect of large $n_{\max}$. After carefully monitoring the loss curve and the performance in the word and sentence similarity tasks, we set the number of epochs 10 for all methods. In preliminary experiments, we also tested the number of epochs 20 for the word-segmentation-dependent baselines but there were no significant differences. In the two supervised downstream tasks, the learned vector representations are combined with the logistic regression classifier. The parameter C, which is the inverse of regularization strength of the classifier, is adjusted via a grid search over $C \in \{0.1, 0.5, 1, 5, 10\}$. Again, as explained in the main paper, the hyperparamters are grid searched on the determined validation set for all experiments.

A.2 Implementations

Here we provide the list of implementations of baselines which are used in our experiments. For skipgram[1], sisg[2], sembei[3], and sent2vec[4], we use the official implementations provided by the authors. Meanwhile, as for pv-dbow and pv-dm, we employ a widely-used implementation of Gensim library[5].

A.3 Word segmenters and word dictionaries for unsegmented languages

Below we list the word segmentation tools and word dictionaries which are used in our experiments. We employed a widely-used word segmentation tool for each language.

For Chinese language, we used jieba[6] with its default dictionary[7] or with an extended dictionary[8], which fully supports both traditional and simplified Chinese characters.

For Japanese, we used MeCab[9] with its default dictionary called IPADIC[9] along with specially designed neologisms-extended dictionary called mecab-ipadic-NEologd[10]. Note that, because this extended dictionary *mecab-ipadic-NEologd* is specially designed to include many neologisms, there is a significant word coverage improvement

[1] https://code.google.com/archive/p/word2vec/
[2] https://github.com/facebookresearch/fastText
[3] https://github.com/oshikiri/w2v-sembei
[4] https://github.com/epfml/sent2vec
[5] https://radimrehurek.com/gensim/models/doc2vec.html
[6] https://github.com/fxsjy/jieba
[7] https://github.com/fxsjy/jieba/blob/master/jieba/dict.txt
[8] https://github.com/fxsjy/jieba/blob/master/extra_dict/dict.txt.big
[9] http://taku910.github.io/mecab/
[10] https://github.com/neologd/mecab-ipadic-neologd

by using this word dictionary as it can be seen in the Japanese noun category prediction task in the main paper.

For Korean, we used mecab-ko[11] with its default dictionary called mecab-ko-dic[12] along with another extended dictionary called NIADic[13].

A.4 Training corpora

We prepared Wikipedia corpora and SNS corpora for Chinese, Japanese and Korean for our experiments. For the Wikipedia corpora, we used the first 10, 50, 100, 200 and 300MB of texts from the publicly available Wikipedia dumps[14]. The texts are extracted by using WikiExtractor tool[15]. For Chinese SNS corpus, we used 100MB of Leiden Weibo Corpus (van Esch, 2012) from the head. For Japanese and Korean SNS corpora, we collected Japanese and Korean tweets using Twitter Streaming API. We removed usernames and URLs from the SNS corpora. There were many informal words, emoticons and misspellings in the SNS corpora. We preserved them without preprocessing to see the effect of the noisiness of training corpora in our experiments.

A.5 Preprocess of Wikidata

For the noun category prediction task, we extracted noun words and their semantic categories from Wikidata (Vrandečić and Krötzsch, 2014) following Oshikiri (2017). We determined the semantic category set used in our experiments as follows: First, we collected Wikidata objects that have Chinese, Japanese, Korean and English labels. Next, we sorted the categories by the number of noun words, and removed categories (e.g., *Wikimedia category* or *Wikimedia template*) that do not represent any semantic category. We also removed out several categories that contain too many noun words (e.g., *human*) or too few noun words (e.g., *academic discipline*). Since there were several duplicated labels for different Wikidata objects, the number of nouns for each language is slightly different. Each category has at least 0.1k words and no more than 5k words. The numbers of extracted noun words that are used in our experiments were 22,468, 22,396 and 22,298 for Chinese, Japanese and Korean, respectively.

A.6 Movie review datasets

In the main paper, three movie review datasets are used to evaluate the quality of sentence embeddings. We used 101,114, 55,837 and 200,000 movie reviews and their rating scores from Yahoo奇摩電影[16], Yahoo!映画[17] and Naver Movies[18] for Chinese, Japanese and Korean, respectively.

B Additional Experiment on Japanese

In this section, we show the results of Japanese word similarity experiments. We use the datasets of Sakaizawa and Komachi (2018). It contains 4427 pairs of words with human similarity scores. We omit sentence similarity task since there is no public widely-used benchmark dataset for Japanese yet. Following the main paper, given a set of word pairs and their human annotated similarity scores, we calculated Spearman's rank correlation between the cosine similarities of the embeddings and the human scores. We use 2-fold cross validation for hyperparameters tuning. The same grid search is performed as explained in Section A.1. To see the effect of the noisiness of training corpora, we use two Japanese corpora, 100MB of Wikipedia corpus and 100MB of noisy SNS corpus (Twitter), which are also used in the Japanese noun category prediction task in the main paper. As seen in Table 4, the experiment results for Japanese are similar to those of Chinese in the main paper.

Table 4: Spearman rank correlations of the word similarity task on two different Japanese corpora.

	$\texttt{skipgram}_{rich}$	$\texttt{sisg}_{rich}$	sembei	sembei-sum	**scne**
Wiki.	8.3	15.4	4.0	9.3	**24.1**
SNS	5.3	12.7	2.8	9.3	**23.0**
Diff.	-3.0	-2.7	-1.2	-0.0	-1.1

References

Daan van Esch. 2012. Leidon weibo corpus.

[11] https://bitbucket.org/eunjeon/mecab-ko

[12] https://bitbucket.org/eunjeon/mecab-ko-dic

[13] https://github.com/haven-jeon/NIADic

[14] https://dumps.wikimedia.org/

[15] https://github.com/attardi/wikiextractor

[16] https://github.com/fychao/ChineseMovieReviews

[17] https://github.com/dennybritz/sentiment-analysis/tree/master/data

[18] https://github.com/e9t/nsmc

Takamasa Oshikiri. 2017. Segmentation-free word embedding for unsegmented languages. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 767–772. Association for Computational Linguistics.

Yuya Sakaizawa and Mamoru Komachi. 2018. Construction of a Japanese Word Similarity Dataset. In *Proceedings of the 11th Language Resources and Evaluation Conference*, Miyazaki, Japan. European Language Resource Association.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: A Free Collaborative Knowledge Base. *Communications of the ACM*, 57:78–85.