論文 / 著書情報
Article / Book Information

Title	Token-based semantic vector space model for classic poetic Japanese
Authors	Xudong Chen, Hilofumi Yamamoto, Bor Hodoscek
Citation	Proceedings of JADH conference, Vol. 2021, , pp. 77-81
Pub. date	2021, 9

Token-based semantic vector space model for classic poetic Japanese

Xudong Chen[1], Hilofumi Yamamoto[2], and Bor Hodošček[3]

1 Introduction

This paper explores the effectiveness of the token-based semantic vector space model (Heylen et al., 2012) for describing the classic poetic Japanese vocabulary.

The token-based semantic vector space model represents the semantics of each individual occur- rence of a word, while a type-based model aggregates over all occurrences of a word, giving a repre- sentation of a word's general semantics (Heylen et al., 2012, p. 17). In type-based models, context- or style-sensitive variation of semantics within word types is averaged and generalized in one vector and thus cannot be described in detail. In token-based models, the description of such variation is possible. Considering the variant referents, meanings, and stylistic usage of the Japanese poetic vocabulary, models on the token level are necessary.

Token-based solutions for the problem of contextualized meanings have been proposed from context- predicting deep learning approaches (e.g. Devlin et al., 2019). Historical data, however, is often too sparse to use the state-of-the-art machine learning methods (Kalouli et al., 2019, p. 109). Another method from a context-counting approach is proposed in Heylen et al. (2012), which does not use any machine learning techniques. Compared to context-predicting deep learning methods, this method is said to have greater transparency (De Pascale, 2019, p. 29). In the present paper, we, therefore, examine the effectiveness of the token-based model for Japanese poetic vocabulary.

2 Methods

2.1 Materials and preprocessing

Yamamoto and Hodošček (2021a) is used as an annotated corpus of Japanese poetry. Metacodes in Yamamoto and Hodošček (2021b) are used for annotating concept groups of word entries.

We select only poems that are within 41 kana in length. Choka, the long poems, for instance, are excluded. We also exclude any word annotated as a particle, auxiliary, auxiliary verb, prefix, suffix, adverb, interjection, and symbol.

2.2 Token-based vectorization

[1] Tokyo Institute of Technology
[2] Tokyo Institute of Technology
[3] Osaka University

Suppose that in a corpus with vocabulary size d, we obtain a token vector of target token t which occurs in the nth poem whose number of words is l. The token vector $\boldsymbol{v}_{t,n}$ can be calculated with Equation (1),

$$\boldsymbol{v}_{t,n} = \frac{1}{\sum_{i=1}^{l} w(t, c_i)} \sum_{i=1}^{l} w(t, c_i)\boldsymbol{c}_i \tag{1}$$

$$\boldsymbol{c}_i = \begin{pmatrix} w(c_i, \mathrm{word}_1) & w(c_i, \mathrm{word}_2) & \cdots & w(c_i, \mathrm{word}_d) \end{pmatrix}^{\mathsf{T}} \tag{2}$$

where c_i is the i^{th} context word of word t in the n^{th} poem. $w(t, c_i)$ is the mutual information between word t and c_i. $\boldsymbol{c}_i$ (Equation (2)) is a vector of context words c_i, which consists of the mutual information between c_i and all words in the corpus. For the mutual information, we use PPMI (Bullinaria & Levy, 2007) (Equation (3)),

$$w(a, b) = \mathrm{PPMI}\,(a, b) = \max\left(0, \log_2 \frac{p(a, b)}{p(a)p(b)}\right) \tag{3}$$

where $p(a)$, $p(b)$, and $p(a, b)$ indicate occurrence probabilities of word a, b, and probability of a and b occurring simultaneously, respectively.

2.3 Classification task with token vectors

In order to confirm the applicability of the token-based vector space model on Japanese poetic vocabulary, we perform classification tasks with token vectors generated by the method. In the classification tasks, we classify token vectors of two-word pairs and confirm whether 2 words in a pair can be correctly classified.

With metacodes in Yamamoto and Hodošček (2021b), we set the following four different types of classifications:

1. Word pairs matching at the concept group level,
   e.g., flower-flower pair *sakura* (cherry)-*mume* (plum);
2. Word pairs unmatched at the concept group level,
   e.g., flower-bird pair *mume* (plum)-*hototogisu* (cuckoo);
3. Noun pairs with the same lemmas (kana strings), but written with different surface forms,
   e.g., *sakura* (cherry) written as さくら/桜;
4. Verb pairs with the same lemmas, but written in different surface forms,
   e.g., *simu* written as 標む (mark as possession)/染む (dye).

We only include items whose document frequencies are within 20 to 90. Since there are large numbers of type 1 and 2 word pairs, we randomly sample 30 pairs of type 1

and 2 respectively. In type 3 and 4 pairs, if both surface forms of the type appear in the same poem, we exclude such cases. Finally, we obtained 29 pairs of type 3 and 7 pairs of type 4.

We use logistic regression as a classifier. We randomly sample 20 vectors of each word in each pair and use half of them as training data and the other half as test data.

We use a generalized linear mixed effects model with a binomial distribution to test how the above- mentioned types of word pairs predict the test accuracy (correct number out of 20 test pairs). In the analyses, we include one random effect, the individual differences of the pairs.

3 Results

Four examples from the results are shown in Figure 1. Dimensionality reduction of the vectors is conducted using multidimensional scaling (MDS). The values of the two dimensions in the current paper span a larger range than those reported in Heylen et al. (2012).

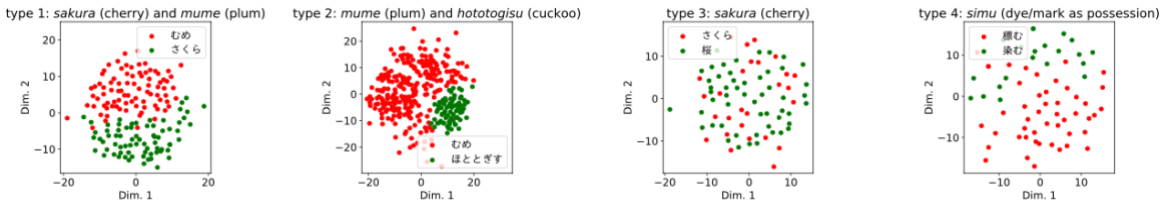

Figure 1: Two-dimensional visualization of token vectors: type 1, 2, and 4 pairs show clear boundaries; the example of type 3 pairs does not show a clear boundary.

The results of the classification task with token vectors is shown in Figure 2. Test accuracy differs among each type of pairs. Pairs whose lemmas are different (type 1 and 2) have the highest test accuracy, and type 2 does not differ from type 1. Noun pairs with the same lemmas written in different surface forms (type 3) have the lowest accuracy. Verb pairs with the same lemmas written in different surface forms (type 4) also have lower accuracy than those of type 1 and 2. Estimated accuracy of pairs with the same lemmas varies in a larger range than that of pairs in different lemmas.

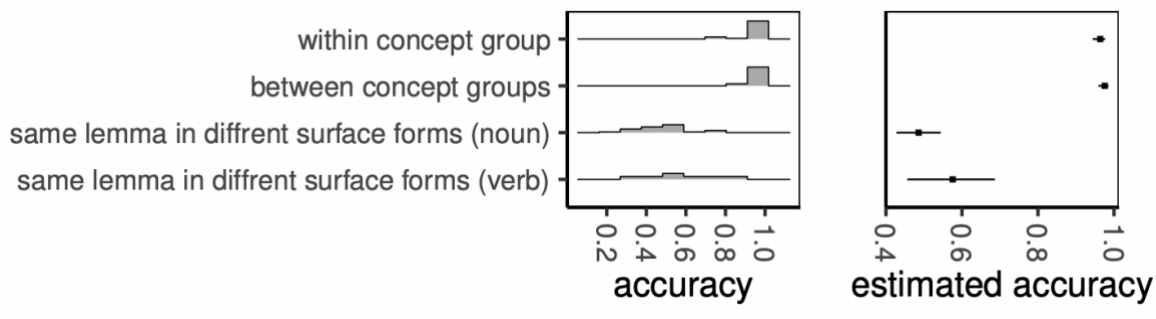

Figure 2: Distribution of test accuracy and estimated test accuracy in the classification task: the left shows the distribution of the test accuracy in each type of task; the right shows the accuracy in 95% CI predicted by generalized linear mixed effects model.

4 Discussion

The vector space generated by the model is sparse. But the token clouds in the 2-dimensional space can reflect the relations among the vocabulary. As shown in Figure 1, token clouds of a pair show clear boundaries when the pair differs more in meanings.

Compared to type 3 and 4, classification in type 1 and 2 pairs has high accuracy. This is because, in most of the cases, word meaning differs more in type 1 and 2 than in type 3 and 4. The accuracy of type 1 classification is slightly higher than that of type 2. This is because pairs in the same concept group share more similar word senses than pairs belonging to different concept groups.

Pairs with the same lemmas cannot be correctly classified. This indicates that information from word types can be important to the current method in a small scale corpus. Most type 3 pairs are often pairs having different surface forms that have no difference in meaning. Therefore the accuracy of type 3 classification is low. On the other hand, there also exist pairs in different surface forms with different meanings in the type 4 pairs (Table 1). The variance of accuracy in type 4 classification is, therefore, greater than that of other types.

Table 1: Examples of token vectors of type 4 pairs: pairs' surface forms with meaning change are well-classified; pairs' surface forms without meaning change are only correctly classified with a low test accuracy; these cases indicate the importance of contextual information to the pairs with the same word types in classification tasks.

Type	Surface forms		Test accuracy
	1	2	
okuru	送る (send; see off)	後る (be late)	0.722
koru	懲る (learn a lesson from failures)	樵る (chop down trees)	0.833
simu	染む (dye)	標む (mark as possession)	0.889
kikoyu	聞こゆ (be able to hear)	聞ゆ (be able to hear)	0.500

5 Conclusion

We conducted the experiments applying the token-based semantic vector space model (Heylen et al., 2015; Heylen et al., 2012) to Japanese classic poem texts in order to examine the possibilities of the model for small-scale corpora such as the Hachidaishu dataset. We found that although a small corpus generates a sparse vector space, it is possible to observe the differences between words at the token level with token clouds visualization generated by the model. The current method also allows us to relatively successfully classify senses between word pairs.

References

- Bullinaria, J. A., & Levy, J. P. (2007). Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior Research Methods*, 39(3), 510–526. https://doi. org/10/d8pmsm
- De Pascale, S. (2019). *Token-based vector space models as semantic control in lexical sociolectometry* (PhD Dissertation). KU Leuven. Leuven.
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirec- tional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. https://doi.org/10/ggbwf6
- Heylen, K., Wielfaert, T., Speelman, D., & Geeraerts, D. (2015). Monitoring polysemy: Word space models as a tool for large-scale lexical semantic analysis. *Lingua*, 157, 153–172. https://doi. org/10/gh58qv
- Heylen, K., Speelman, D., & Geeraerts, D. (2012). Looking at word meaning. An interactive visu- alization of Semantic Vector Spaces for Dutch synsets. *Proceedings of the EACL-2012 Joint Workshop of LINGVIS & UNCLH: Visualization of Language Patterns and Uncovering Lan- guage History from Multilingual Resources*, 16–24.
- Kalouli, A.-L., Kehlbeck, R., Sevastjanova, R., Kaiser, K., Kaiser, G. A., & Butt, M. (2019). ParHistVis: Visualization of Parallel Multilingual Historical Data. *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, 109–114. https:// doi.org/10/gh547n
- Yamamoto, H., & Hodošček, B. (2021a). Hachidaishu part of speech dataset. https://doi.org/10.5281/ zenodo.4835806
- Yamamoto, H., & Hodošček, B. (2021b). Hachidaishu vocabulary dataset. https://doi.org/10.5281/ zenodo.4744170