Relationships between Flowers in a Word Embedding Space of Classic Japanese Poetry

Hilofumi Yamamoto
Tokyo Institute of Technology

Bor Hodošček
Osaka University

2017.5.8

Objectives

The present study, which is ongoing work, focuses on exploring the connection between the information content of words and their semantic role within classical poetic Japanese.

Word embedding methods such as Word2Vec (Mikolov et al. 2013) and more recent extensions, including multi-modal word distributions in (Chen et al. 2015, Athiwaratkun and Wilson 2017) and richer representations such as doc2vec in (Le and Mikolov 2014), have been shown effective in extracting semantic knowledge that performs well on analogy, semantic similarity, and entailment tasks. In this work, we will quantify the relationship between the information content of a word and its word embedding vector and examine the possibility of using word embedding spaces of classical poetic terms to explain the semantic relationships between terms.

First we will examine if word embeddings encode enough semantic information to be able to determine specific subordinate words via their superordinate concept, represented by a poetic term. For example, flowers frequently appear in classical poetry and tend to be used with specific words. Both *ka* (aroma) and *chiru* (fall) are used with flowers such as: "*hito wa isa, kokoro mo shirazu, furusato wa, hana zo mukashi no, ka ni nioikeru*" by Ki no Tsurayuki (the Kokinshū, No. 42; the state of human / hearts I cannot know and yet / the blossoms of this / familiar village still greet / me with the scent of years past / translated by Rodd and Henkenius (1984)). "*hisakata no, hikari nodokeki, haru no hi no, shizugokoro naku, hana no chiru ramu*" by Ki no Tomonori (the Kokinshū, No. 84; the air is still and / sun-warmed on this day of spring– / why then do cherry / blossoms cascade to the earth / with such restless changeful hearts/ translated by Rodd and Henkenius (1984) as well). If *ka* (aroma) and *chiru* (fall) are truly relating to specific flowers, it should be possible to acquire the specific name of a flower based on their word embeddings.

Second, we will examine that the residual of $\mathbb{A}$ minus a will be nothing if word

embeddings are strong enough to extract the specific name a from its superordinate concept $\mathbb{A}$ and when $a \in \mathbb{A}$ is established.

Methods

We use the Hachidaishū, classical Japanese poem anthologies compiled by the order of the Emperors (ca. 905–1205), comprised of approximately 9,500 poems. Each poem is tokenized by `kh` (Yamamoto 2007) which divides poem texts into tokens using a classical Japanese dictionary. In order to examine the notable relationships between 'ka' (fragrance), 'chiru' (fall), we look at the cosine similarity scores between terms in the word embedding space generated by Word2Vec. The Word2Vec model was generated using the Word2Vec implementation in gensim 2.0.0 (Řehůřek and Sojka 2010). A 50-dimensional skip-gram model with negative sampling was used with context window covering the whole poems.

Results

As a result of measuring the cosine distances between *ka* (fragrance) and other words, also between *chiru* (fall) and other words, the list of the former relationship indicates *ume* (plum) and the latter indicates *sakura* (cherry) as their corresponding flowers (Table 1 and 2).

As the flower of summer, we could obtain *tachibana* (the flower of orange), while after removing the vector of *tachibana*, we could not obtain any names of flowers (Table 3).

TABLE 1: Top five words similar to *ka* (fragrance) and the reverse examination of *ka* by using the term *ume* (plum). Each value indicates the cosine similarity between each word pair.

	ka (fragrance)		*ume* (plum)	
1	*ume* (plum)	0.96	*ka* (fragrance)	0.96
2	*niofu* (smell)	0.94	*niofu* (smell)	0.92
3	*nushi* (patron)	0.92	*kakine* (fence)	0.90
4	*chirasu* (make fall)	0.90	*nushi* (patron)	0.90
5	*moru* (raise)	0.89	*moru* (raise)	0.90

Discussion

The lists clearly shows that *ka* (fragrance) is related to *ume* (Mizutani 1983: 130). In terms of *chiru* (fall), the latter result replicates well-established knowledge in the literature that falling flowers denote sakura (cherry) and not ume (plum), and it is discernable that *sakura* (cherry) relates to *chiru* (fall), which indicates that people at

Table 2: Top five similar words to *chiru* (fall) and the reverse examination of *chiru* by using the term *sakura* (cherry). Each value indicates the cosine similarity between each word pair. pn. means 'place name'.

	chiru (fall)		*sakura* (cherry)	
1	*moru* (raise)	0.95	*yamazakura* (mountain cherry)	0.82
2	*sakurabana* (cherry blossoms)	0.94	*Yoshinoyama* (pn.)	0.82
3	*Yoshinoyama* (pn.)	0.93	*chirasu* (make fall)	0.80
4	*yahe* (eight fold)	0.93	*izure* (any)	0.80
5	*yamazakura* (mountain cherry)	0.92	*sakurabana* (cherry blossoms)	0.79

Table 3: Operations relating to *natsu* (summer), *hana* (flower), and *tachibana* (orange).

	flower + summer		flower + summer - orange	
1	*yadosu* (to dwell)	0.90	*yoru* (night)	0.69
2	*kaoru* (to smell)	0.90	*hikari* (light)	0.68
3	*tachibana* (orange)	0.89	*kohori* (ice)	0.67
4	*odoroku* (to surprize)	0.88	*tsura* (face)	0.66
5	*ushirometashi* (to feel guilty)	0.87	*harafu* (pay)	0.66
6	*katsura* (name of tree)	0.87	*fuyu* (winter)	0.66
7	*migaku* (polish)	0.87	*suzushi* (cool)	0.65
8	*issoshi* (more)	0.87	*akeshi* (to dawn)	0.65
9	*haku* (to sweep away)	0.87	*moru* (to leak)	0.65
10	*Musashino* (pn.)	0.86	*niwa* (garden)	0.65

the time lamented falling *sakura* (cherry) (Katagiri 1983: 84).

We next looked at whether the Word2Vec word embedding creates a vector space where geometric algebra is possible and vector distances in the space hold certain semantic meaning. Among summer flowers, the *tachibana* (the flower of orange) is very famous. We expect that if we subtract *tachibana* out from the summer vectors, the resulting space will be devoid of relationships between *natsu* (summer) and *hana* (flower). By calculating with the relational expressions $summer + flower$ and $summer + flower - tachibana$, the operations conducted by Word2Vec have been shown to reproduce our current understanding of the relationships between flowers and seasons as well as some emotions associated with them. As shown in Table 3, the $summer + flower$ operation indeed includes *tachibana* and it should be noted that the $summer + flower - tachibana$ operation did not include any remarkable values between $summer$ and $flower$.

Conclusion

We conducted the experiments using approximately 9,500 classical Japanese poem texts in order to examine the possibilities of extracting subordinate terms from superordinate concepts based on word embeddings. We found that the model also allows us to extract specific subordinate words based on the superordinate concept of classical terms such as: when the distance between two terms such as *tachibana* (orange) and *natsu* (summer) is close enough, the superordinate concept $\mathbb{A}$ indicates the subordinate concept a. We could therefore verify that it allows us to extract the concrete name from its superordinate concept.

References

Athiwaratkun, Ben and Andrew Gordon Wilson (2017) "Multimodal Word Distributions," *ArXiv e-prints*, April.

Chen, Xinchi, Xipeng Qiu, Jingxiang Jiang, and Xuanjing Huang (2015) "Gaussian Mixture Embeddings for Multiple Word Prototypes," *CoRR*, Vol. abs/1511.06246, URL: `http://arxiv.org/abs/1511.06246`.

Katagiri, Yoichi (1983) *Utamakura utakotoba jiten (Dictionary of poetic vocabulary)*, Vol. 35 of Kadokawa shojiten, Tokyo: Kadokawa Shoten.

Le, Quoc V. and Tomas Mikolov (2014) "Distributed Representations of Sentences and Documents," *CoRR*, Vol. abs/1405.4053, URL: `http://arxiv.org/abs/1405.4053`.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean (2013) "Efficient Estimation of Word Representations in Vector Space," *CoRR*.

Mizutani, Sizuo (1983) *Goi (Vocabulary)*, Vol. 2 of Asakura Nihogo Shin-Kōza, Tokyo, Japan: Asakura Shoten.

Řehůřek, Radim and Petr Sojka (2010) "Software Framework for Topic Modelling with Large Corpora," in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pp. 45–50, Valletta, Malta: ELRA, May, `http://is.muni.cz/publication/884893/en`.

Rodd, Laurel Rasplica and Mary Catherine Henkenius (1984) *Kokinshū - A Collection of Poems Ancient and Modern*, Boston MA USA: Cheng and Tsui Company.

Yamamoto, Hilofumi (2007) "Waka no tame no Hinshi tagu zuke shisutemu / POS tagger for Classical Japanese Poems," *Nihongo no Kenkyu / Studies in the Japanese Language*, Vol. 3, No. 3, pp. 33–39.