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

Hilofumi Yamamoto, Tokyo Institute of Technology
yamagen@ila.titech.ac.jp

Bor Hodošček, Osaka University
bor@lang.osaka-u.ac.jp

Introduction

- Word embedding methods such as Word2Vec (Mikolov et al., 2013; Le and Mikolov, 2014) have been shown effective in extracting semantic knowledge from large corpora.
- Quantify the relationship between the content of a word and its word embedding vector.
- Examine the possibility of word embedding spaces to explain the semantic relationships between classical Japanese poetic terms.

Problem

- Can word embeddings trained on the Hachidaishu encode enough semantic information to find subordinate words via their superordinate concept?

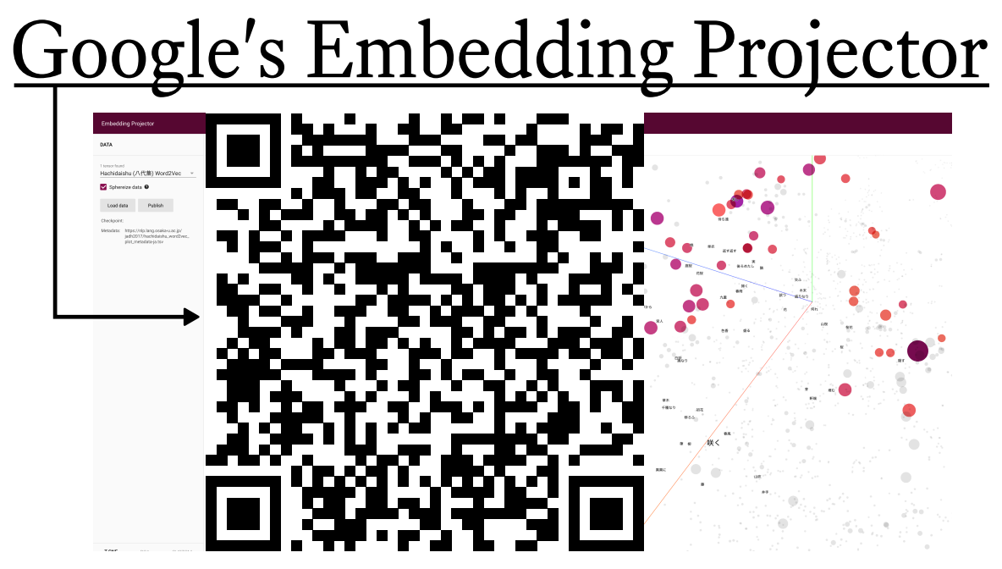
Materials

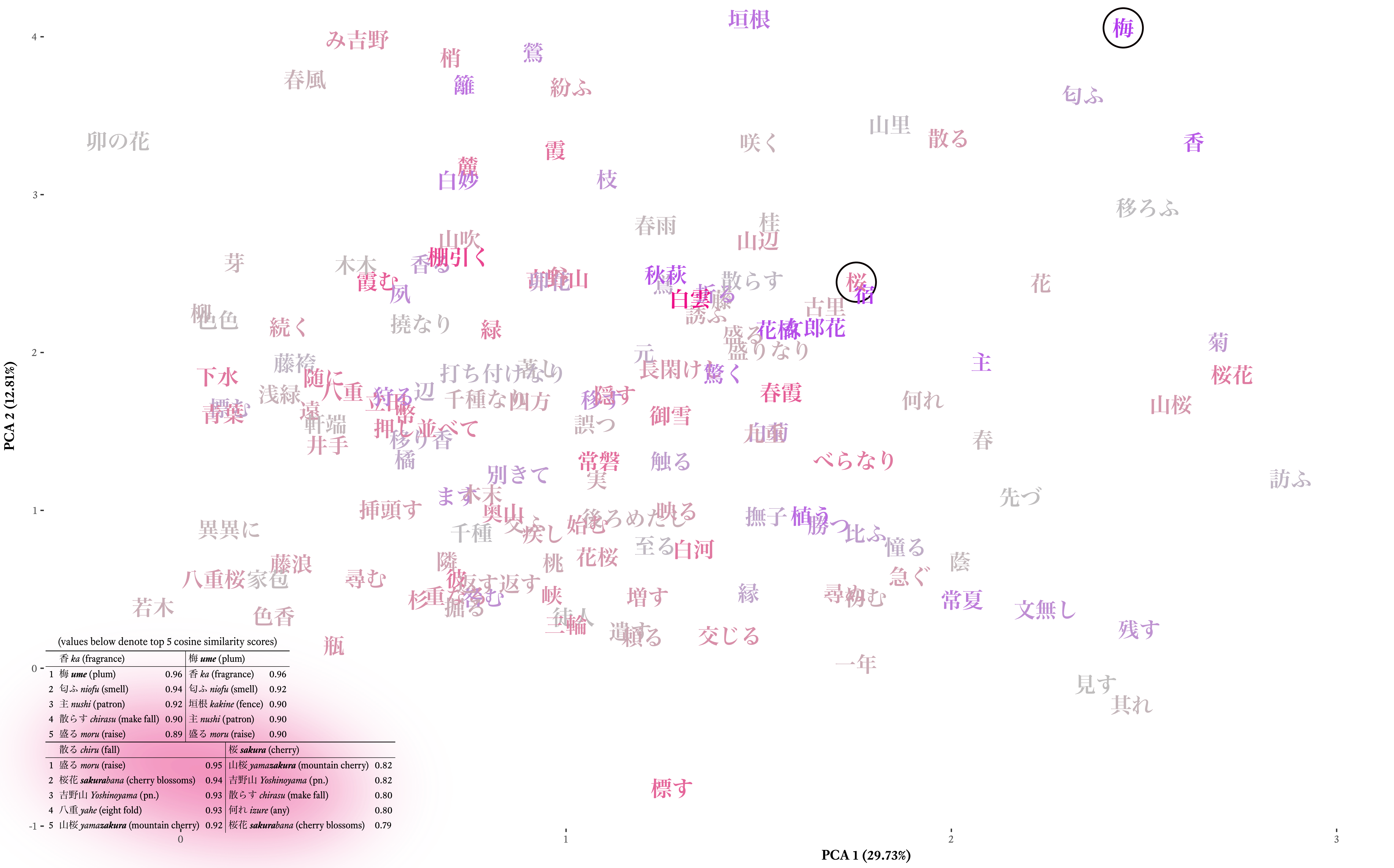
- *Hachidaishū*: classical Japanese poem anthologies compiled under decree by Emperors (ca., 905—1205), comprising approximately 9,500 poems and 159,183 tokens (Source: *KokkakaitanNijūichidaishū* database published by NIJIL).
- Each poem is tokenized into lemma forms by **kh** (Yamamoto, 2007) which divides poem texts into tokens using a classical Japanese dictionary.

Methods

- 50-dimensional skip-gram model with negative sampling, context window covering the whole poem using Gensim 2.3.0 (Řehůřek & Sojka, 2010).
- 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.

Access the dataset online using Google's Embedding Projector





(values below denote top 5 cosine similarity scores)

	香 *ka* (fragrance)		梅 ***ume*** (plum)	
1	梅 ***ume*** (plum)	0.96	香 *ka* (fragrance)	0.96
2	匂ふ *niofu* (smell)	0.94	匂ふ *niofu* (smell)	0.92
3	主 *nushi* (patron)	0.92	垣根 *kakine* (fence)	0.90
4	散らす *chirasu* (make fall)	0.90	主 *nushi* (patron)	0.90
5	盛る *moru* (raise)	0.89	盛る *moru* (raise)	0.90

	散る *chiru* (fall)		桜 ***sakura*** (cherry)	
1	盛る *moru* (raise)	0.95	山桜 *yamazakura* (mountain cherry)	0.82
2	桜花 ***sakurabana*** (cherry blossoms)	0.94	吉野山 *Yoshinoyama* (pn.)	0.82
3	吉野山 *Yoshinoyama* (pn.)	0.93	散らす *chirasu* (make fall)	0.80
4	八重 *yahe* (eight fold)	0.93	何れ *izure* (any)	0.80
5	山桜 *yamazakura* (mountain cherry)	0.92	桜花 ***sakurabana*** (cherry blossoms)	0.79

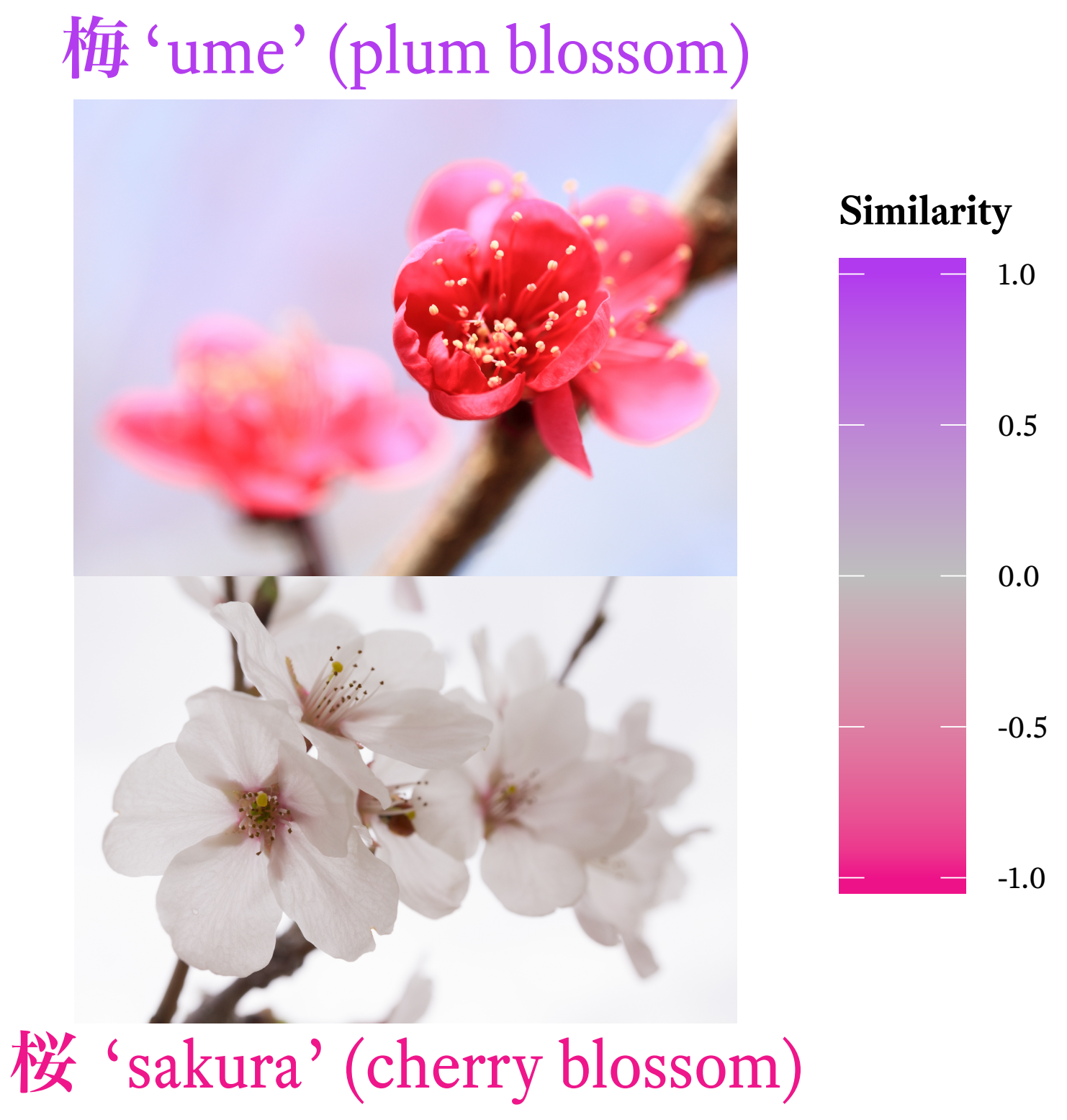


Figure 1: PCA of word embedding space (4157 words × 50 dimensions) filtered to include only top 100 similar words for each of ume and sakura (150 total). Similarity is represented by the difference in similarity scores between ume and sakura, scaled to [-1, 1].

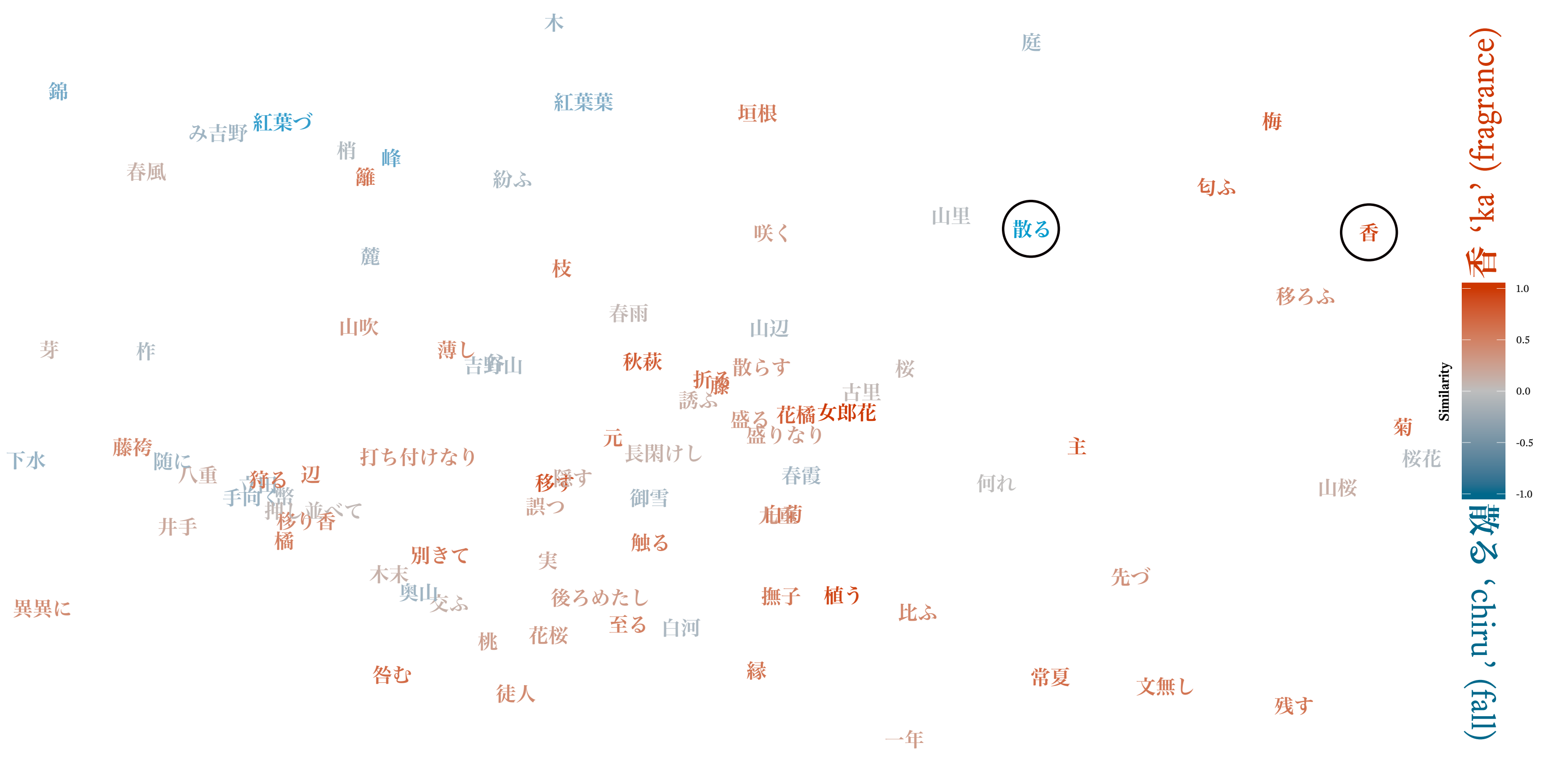


Figure 2: PCA of words similar to 香 'ka' (fragrance) and 散る 'chiru' (fall).

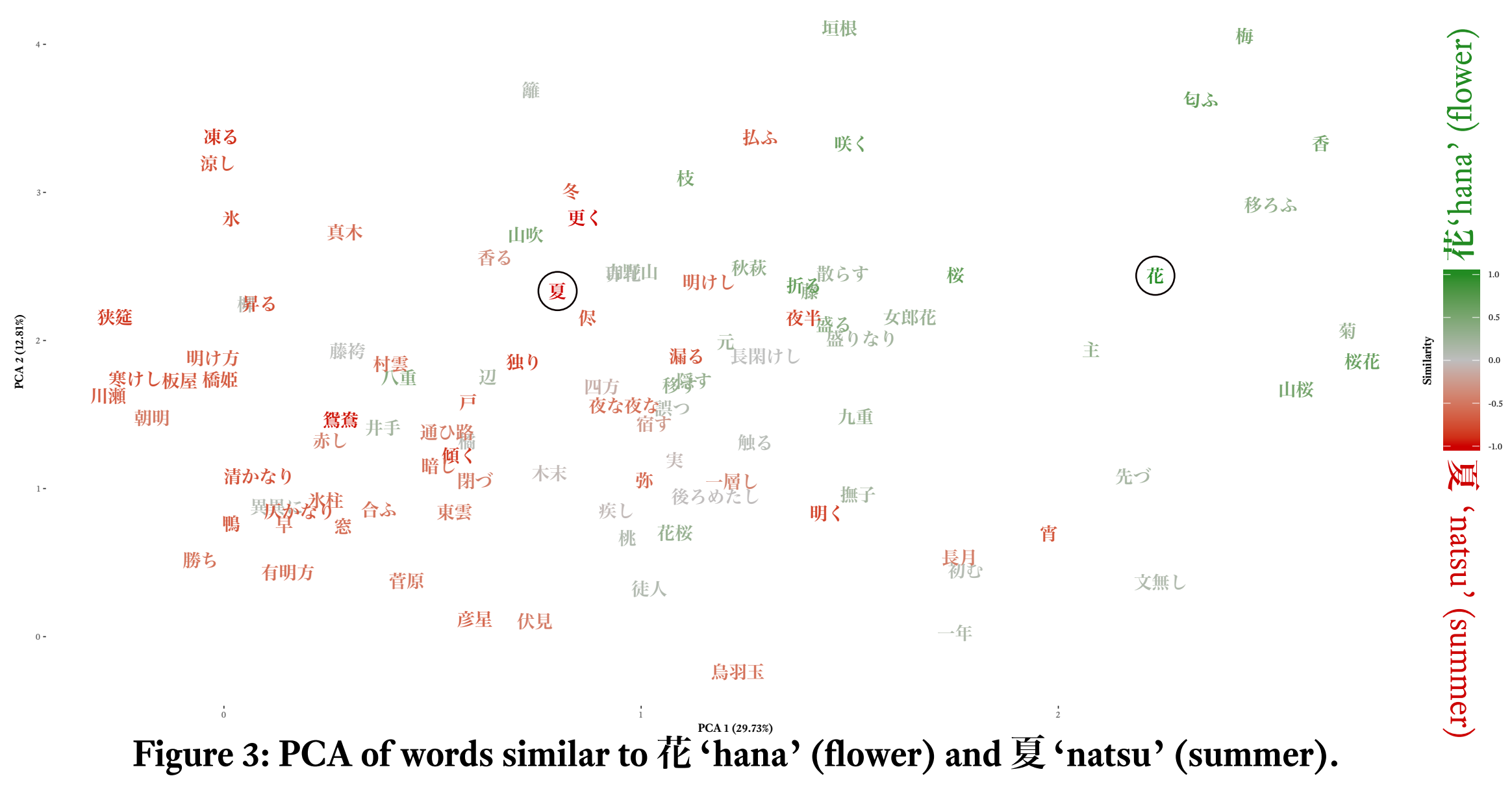


Figure 3: PCA of words similar to 花 'hana' (flower) and 夏 'natsu' (summer).

Among summer flowers, the tachibana (the flower of orange) is very famous.

	flower + summer (花＋夏)		flower + summer - orange (花＋夏－橘)	
1	宿す *yadosu* (to dwell)	0.90	夜 *yoru* (night)	0.69
2	香る *kaoru* (to smell)	0.90	光 *hikari* (light)	0.68
3	橘 ***tachibana*** (orange)	0.89	氷 *kohori* (ice)	0.67
4	驚く *odoroku* (to surprize)	0.88	面 *tsura* (face)	0.66
5	後ろめたし *ushirometashi* (to feel guilty)	0.87	払ふ *harafu* (pay)	0.66
6	桂 *katsura* (name of tree)	0.87	冬 *fuyu* (winter)	0.66
7	磨く *migaku* (polish)	0.87	涼し *suzushi* (cool)	0.65
8	一層し *issoshi* (more)	0.87	明けし *akeshi* (to dawn)	0.65
9	掃く *haku* (to sweep away)	0.87	漏る *moru* (to leak)	0.65
10	武蔵野 *Musashino* (pn.)	0.86	庭 *niwa* (garden)	0.65

Results

- 'ka' (fragrance) is related to 'ume' (plum) (replicating Mizutani, 1983).
- Falling flowers denote 'sakura' (cherry) and not 'ume' (plum); 'sakura' (cherry) relates to chiru (fall), which indicates that people at the time lamented falling sakura (falling cherry blossom petals) (replicating p. 84 in Katagiri, 1983).
- Subtracting tachibana out from the summer vectors reveals a vector space devoid of relationships between natsu (summer) and hana (flower). These relational expressions (summer + flower; summer + flower - tachibana) reproduce our current understanding of the relationships between flowers and seasons as well as some emotions associated with them in the word embedding space.

Conclusion

- Word embeddings allowed us to extract specific subordinate words based on the superordinate concept of classical terms → when the distance between two terms such as 'tachibana' (orange) and 'natsu' (summer) is close enough, the superordinate concept A indicates the subordinate concept *a*.
- We could therefore verify that it allows us to extract the concrete name from its superordinate concept.

References

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,* URL: http://arxiv.org/abs/1301.3781.

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.

JADH 2017: "Creating Data through Collaboration" @ Doshisha University, September 11-12, 2017