論文 / 著書情報
Article / Book Information

Title	A study on the distribution of cooccurrence weight patterns of classical Japanese poetic vocabulary
Authors	Hilofumi Yamamoto, Bor Hodoscek
Citation	JADH2018 Proceedings of the 8th Conference of Japanese Association for Digital Humanities ‘Leveraging Open Data", Volume 2018, , pp. 179-182
Pub. date	2018, 9

JADH 2018

"Leveraging Open Data"

September 9-11, 2018
Hitotsubashi Hall, Tokyo

https://conf2018.jadh.org

Proceedings of the 8th Conference of Japanese Association for Digital Humanities

Organized by:
Japanese Association for Digital Humanities

Hosted by:
Center for Open Data in the Humanities, Joint Support-Center for Data Science Research, Research Organization of Information and Systems

Co-organized by:
JSPS Grant-in-Aid Project (S) "Construction of a New Knowledge Base for Buddhist Studies" (15H05725)
International Institute for Digital Humanities

Sponsored by:

Supported by:
Japan Art Documentation Society
IPSJ SIG Computers and the Humanities
Japan Society for Digital Archives
Japan Association for English Corpus Studies
Japan Association for East Asian Text Processing
The Mathematical Linguistic Society of Japan
Japan Society for Information and Media Studies
Japan Society of Information and Knowledge

Proceedings of the 8th Conference of Japanese Association for Digital Humanities

Edited by Chikahiko Suzuki

Published by:
Center for Open Data in the Humanities, Joint Support-Center for Data Science Research, Research Organization of Information and Systems
2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo
http://codh.rois.ac.jp

Online edition: ISSN 2432-3144 Print edition: ISSN 2432-3187

A study on the distribution of cooccurrence weight patterns of classical Japanese poetic vocabulary

Hilofumi Yamamoto[1], Bor Hodoscek[2]

Introduction

The present study focuses on ongoing work exploring the threshold values dividing words in classical Japanese text into three groups: content, functional, and in-between. Content or semantic based analyses usually employ some techniques of data cleansing, such as eliminations of tags, punctuations, or symbols, as a preprocessing step. Stop words are also a type of token to be eliminated since they contain comparatively less meaning for content analysis. In general, it can be said that the most frequent words will be common words such as 'the' or 'and,' which help build ideas but do not carry any significance themselves (Rajaraman and Ullman, 2012: 8). Lists of stop words are commonly used, but have some problems: 1) it is necessary to compile them in advance; 2) they necessarily change depending on the domains of analyses; and 3) it is not clear which words should be included when analyzing classical texts.

Our previous study grouped modern Japanese words into low-, mid-, and high-range groups according to their information content given by their term frequency-inverse document frequency (*tf-idf*) and found that low-range words corresponded to infrequent and highly topical words, and high-range words corresponded to functional words expressing the grammatical relations between words. The study did not find an automatic method capable of classifying tokens into low-, mid-, and high-range. Furthermore, we found that previous research almost exclusively ignored the properties of the mid-range (Hodoscek and Yamamoto, 2013).

One of the methods used in Hodoscek and Yamamoto (2013) exploited the occurrence not of individual words but of pairwise or cooccurrence patterns such as 'fragrance–flower' relationships and revealed that the distribution of cooccurrence weights in modern Japanese texts approximately fitted a Gaussian curve. In this study, we will attempt to expand this analysis to classical texts by utilizing the characteristics of the Gaussian distribution to automatically group words into three clusters of cooccurrence patterns.

Methods

We use the *Hachidaishu* as the material of the present study, which comprises the eight anthologies compiled under order of the Emperors (ca. 905–1205) and contains about 9,500 poems. We developed the corpus and a method of cooccurrence weighting similar to the *tf-idf* method, *cw* (Yamamoto, 2006), which calculates the weight of patterns of any two words occurring in a poem sentence (Spärck Jones, 1972; Robertson, 2004; Manning and Schutze, 1999; Rajaraman and Ullman, 2012).

$$
\begin{aligned}
w(t,d) &= (1+\log\ tf(t,d)) \cdot idf(t) \\
cw(t_1,t_2,d) &= (1+\log\ ctf(t_1,t_2,d)) \cdot cidf(t_1,t_2) \\
cidf(t_1,t_2) &= \sqrt{idf(t_1) \cdot idf(t_2)} \\
idf(t) &= \log \frac{N}{df(t)}
\end{aligned}
$$

Where *w* is a weight, *t* a token, and *N* the number of tokens. The function *idf* is called the "inverse document frequency" (Spärck Jones, 1972; Robertson 2004; Manning

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

and Schutze, 1999). The function *cw* is called the "cooccurrence weight," which allows us to examine the patterns of poetic word constructions through mathematical modeling.

As in Figure 1, there is a concept (Losee, 2001: 1019) of terms located in each layer being effective query terms. Luhn (1968) cuts the top and bottom words of the frequency and uses the mid-range vocabulary for the development of an automatic outline generation system (Figure 1). Nagao (1983: 28) also mentioned mid-range vocabulary to be effective in generating automatic abstracts. Nagao's viewpoint is slightly different from Luhn (1968) in that it allocates the distribution of word lengths around the Gaussian curve. The positions of both the upper cutoff and the lower cutoff are, however, assumed to be empirical; it is not discussed where to cut them off.

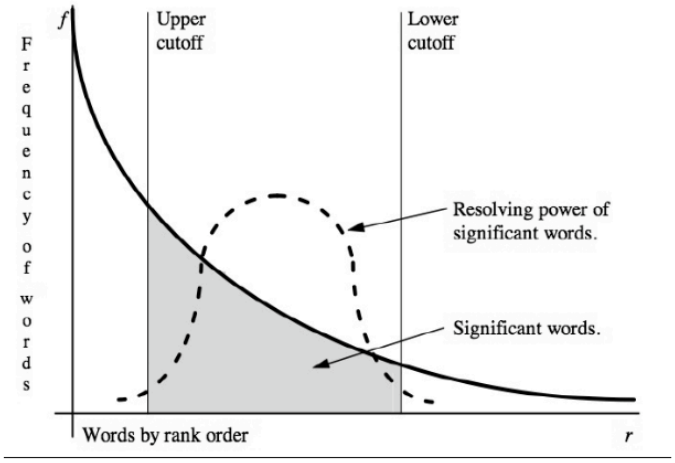

Fig. 1: Hyperbolic curve relating occurrence frequency with rank order; adapted from (Luhn 1968: 120)

Results

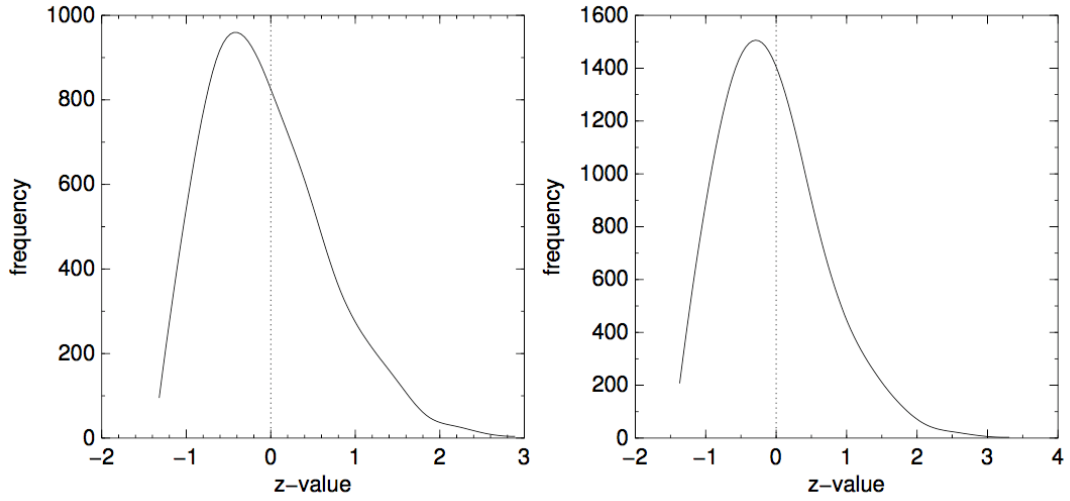

Fig. 2: The distribution of cw values *ume* (plum; left) and *sakura* (cherry; right) in Hachidaishū; The statistics of *ume* (plum): N=7016, min=-1.370, mean=0.138, max=3.700, SD=0.740, SE=0.009, CV=534.012%, Reliable interval low - upper = 0.116 - 0.161 (95%), skew=0.737, kurtosis=3.567, and that of *sakura* (cherry): N=4734, min=-1.320, mean=0.132, max=3.240, SD=0.716, SE=0.010, CV=544.116%, Reliable interval low - upper = 0.104 - 0.159 (95%), skew=0.740, kurtosis=3.345 indicate both approximately fit a Gaussian curve

The distribution of *cw* values is taken from the network model of both *ume* (plum) and *sakura* (cherry) and their curves belong to Gaussian curve as well as in classical texts (Figure 2). Therefore we will attempt to divide this shape into three layers by inflection points.

The cooccurrence patterns of *sakura* (cherry) under -0.9 (near -1) *cw* value are adjacent patterns comprising function words, and over 1 *cw* value are patterns with

content words as we expected (Table 1 and 2). As for the upper cutoff, we used an under -0.9 (near -1) σ value of *cw*, which could extract patterns of functional tokens: almost all patterns included functional words, while as lower cutoff, we used over 1 σ values, which could extract patterns of content tokens: almost all patterns included content words. Both under -1 and over 1 σ are regarded as inflection points which have mathematically interesting properties.

Discussion

Inflection points are defined as the points on the curve where the curvature changes its sign while a tangent exists (Bronshtein et al., 2004: 231). We consider the threshold values that separate upper cutoff, mid-range, and lower cutoff not as coincidental but as evidential points. It is, however, necessary to conduct further experiments and continue to discuss the mathematical traits behind the distributions of cooccurrence weights.

In terms of removing the low-range (upper cutoff) and extracting the high-range (lower cutoff) from poetic texts, we found that we do not need to use any filters to eliminate terms, since *cw* values returned semantically cooccurring patterns. Apart from low-range and high-range, the characteristics of the mid-range lexical layer are still unknown.

Table 1: Upper cutoff patterns of *ame* (sakura): *cw* = co-occurrence weight; *z* = z-value (normalized value of frequency). word annotations: ari(be), ba(cond.), ha(topic.), hana(flower), hito(human), keri(past.), ki(past.), koso(emphatic.), miru(see), mo (also), nasi(no exist), nu(neg.), o(obj.), omou(think), ramu(aux.will), su(do), te(p.), to(and), ware(we), zo(emphatic.), zu(neg.)

	cw	*z*	pattern		*cw*	*z*	pattern		*cw*	*z*	pattern
1	0.62	-0.91	mo–keri	11	0.59	-0.96	nasi–ha	21	0.52	-1.05	nu–o
2	0.62	-0.92	hana–o	12	0.57	-0.98	o–ramu	22	0.52	-1.05	o–zo
3	0.62	-0.92	o–koso	13	0.57	-0.98	mo–ramu	23	0.52	-1.05	miru–o
4	0.60	-0.94	zu–keri	14	0.57	-0.98	ha–ki	24	0.48	-1.09	ba–mo
5	0.60	-0.94	su–ha	15	0.56	-1.00	zu–mo	25	0.48	-1.09	o–keri
6	0.60	-0.94	to–ba	16	0.56	-1.00	o–te	26	0.43	-1.16	zu–ha
7	0.59	-0.96	ari–ha	17	0.55	-1.01	hito–mo	27	0.43	-1.16	to–o
8	0.59	-0.96	ari–mo	18	0.54	-1.02	zu–te	28	0.43	-1.16	te–ha
9	0.59	-0.96	ware–mo	19	0.52	-1.05	zo–ha	29	0.34	-1.27	o–ha
10	0.59	-0.96	nasi–o	20	0.52	-1.05	omou–o	30	0.34	-1.27	o–mo

Table 2: Lower cutoff patterns of *ame* (sakura) in Kokinshū: 30 out of 164 patterns extracted; *cw* = co-occurrence weight; *z* = z-value (normalized value of frequency) word annotations: ba(cond.), bakari(only), besi(should be), chiru(fall), fukakusa(deepgreen), hana(flower), isa(already), kakusu(hide), katu(win), koku(pull), komoru(go deep inside), magiru(mix), makasu(entrust), maku(wind up), manimani(as it is), masi(as), mazu(mix), me(eye), minami(south), miyako(city), mono(thing), nagara(even if), sakura(cherry), si(emphasic.), sumi(black ink), tatu(start,stand), tazumu(being around), tu(past.), uturou(change), watasu(give), yamakaze(mountain wind), yamu(stop), yanagi(willow), yononaka(world)

	cw	*z*	pattern		*cw*	*z*	pattern
1	3.86	3.18	yamu–manimani	106	2.38	1.31	si–fukakusa
2	3.75	3.04	minami–magiru	107	2.38	1.31	sakura–hana
3	3.67	2.93	minami–maku	108	2.38	1.31	sakura–isa
4	3.61	2.86	maku–magiru	109	2.38	1.31	sakura–ba
5	3.42	2.62	yanagi–koku	110	2.38	1.30	sakura–me
6	3.38	2.57	yamu–makasu	—			
7	3.38	2.56	mazu–koku	155	2.17	1.04	chiru–katu
8	3.27	2.43	yanagi–mazu	156	2.17	1.04	bakari–sumi
9	3.26	2.42	sakura–yamu	157	2.16	1.03	maku–besi
10	3.25	2.40	minami–yamakaze	158	2.16	1.03	tatu–maku
–				159	2.16	1.03	tatu–tazumu
101	2.40	1.33	uturou–komoru	160	2.16	1.03	tazumu–tu
102	2.40	1.33	sakura–watasu	161	2.16	1.03	miyako–sakura
103	2.40	1.33	katu–nagara	162	2.16	1.02	kakusu–si
104	2.39	1.32	sakura–masi	163	2.14	1.00	yononaka–sakura
105	2.39	1.31	sakura–makasu	164	2.14	1.00	mono–sakura

Conclusion

Using the distribution characteristics of cooccurrence weights, we were able to classify cooccurrence patterns into three layers of cooccurrence patterns: high-, mid-, and low-range patterns.

We found that 1) the distribution of classical texts fits a Gaussian curve as well as in modern texts; 2) the *cw* value can separate patterns into three layers (low-, mid-, and high-range) using inflection points (-1σ and 1σ); 3) of the three layers, the high-range could be extracted without a list of stop words; 4) the mid-range lexical layer might include mathematical traits not yet revealed in the present study.

References

Bronshtein, I.N., Semendyayev, K.A., Musiol, G., and Muehlig, H. (2004). Handbook of Mathematics: Springer-Verlag, 4th edition.

Hodoscek, B. and Hilofumi Y. (2013). "Analysis and Application of Midrange Terms of Modern Japanese", in Computer and Humanities 2013 Symposium Proceedings, No. 4, pp. 21–26.

Losee, Robert M. (2001). "Term dependence: A basis for Luhn and Zipf models", Journal of the American Society for Information Science and Technology, Vol. 52, No. 12, pp. 1019–1025.

Luhn, H. P. (1968). HP Luhn: Pioneer of Information Science: Selected Works: Spartan Books.

Manning, C.D. and Schutze, H. (1999). Foundations of statistical natural language processing, Cambridge, Massachusetts: The MIT press.

Nagao, M. (1983). Gengo kogaku (Language Engineering), Jinkochino sirizu 2 (Series of Artificial Intelligence): Shokodo.

Rajaraman, A. and Ullman, J.D. (2012). Mining of massive datasets, Cambridge: Cambridge University Press.

Robertson, S. (2004). "Understanding inverse document frequency: on theoretical arguments for IDF", Journal of Documentation, Vol. 60, pp. 503–520.

Spärk Jones, K. (1972). "A Statistical Interpretation of Term Specificity and Its Application in Retrieval", Journal of Documentation, Vol. 28, pp. 11–21.

Yamamoto, H. (2006). "Konpyuuta niyoru utamakura no bunseki / A Computer Analysis of Place Names in Classical Japanese Poetry", in Atti del Terzo Convegno di Linguistica e Didattica Della Lingua Giapponese, Roma 2005: Associazione Italiana Didattica Lingua Giapponese (AIDLG), pp. 373–382.