Proceedings of the International Workshop
on Modern Science and Technology
Kitami, Japan/September 2010

Automatic Haiku Generation Using Web Search and Japanese Weblogs as Input

Takuya Emori, Rafal Rzepka, and Kenji ARAKI

Graduate School of Information Science and Technology, Hokkaido University,
{txearth,kabura,araki}@media.eng.hokudai.ac.jp

Abstract: *Haiku*[1] is a kind of Japanese poem. In case of computer generated *haiku*, the poems are generally constructed by words or phrases limited to a given database. However, the existing systems are not sufficient for user's demands because of the limited sources of examples. Our method solves this problem by using Japanese blogs as a default input and Web search. At first, the blog entry is parsed for retrieving nouns, which are used for Web search query. Our system sorts out a keyword from the entry according to Web search hit number. The system gains various words for creating a *haiku*. We evaluate the artificially created poems by Semantic Differential method. Finally, it is found that the generated *haiku* using proposed method compares favorably with previous one, which does not consider a relation degree between words in results of impression evaluation.

Key words: haiku generation, poetry, artificial art

1. Introduction

We describe an automatic *haiku* generation method using Web search and Japanese blogs. *Haiku* is a kind of Japanese poem with minimal length of seventeen syllables and several grammatical rules. These rules include elements of ancient Japanese and nature-related expressions *kigo*. *Kigo* is used for making people imagine a year season. *Haiku* was born about 700 years ago. Today men and women of all age groups are also familiar with *haiku*. As a related research, Yoshioka built a *kigo*-database for automatically specifying the *kigo* [1] and developed "*haiku* entry and appreciation system" for developing the database [2]. Tosa created "Hitch *Haiku*", which supports for creating a haiku [3]. Wu achieved an automatic *haiku* generation system [4].

However, the existing systems are not sufficient for user's demands because of depending on the limited databases. The databases restrict a literature expression. And quite difficult grammatical rules are an obstacle for users who want to create a *haiku* by themselves. Our method solves these problems by using Japanese blogs as the default input and Web search. We hold up two purposes in our research. One is removing the limited literature expression by no depending on the database. The other is providing an appropriate poem for *haiku* beginners for their entertainment.

In recent years, everyone can gain Web search results with free of charge. Such results are statistically trustworthy [5]. Web search also achieves high robustness for unknown words which do not exist in the database. A user writes a blog as a diary, and our system automatically depicts the article text as an original *haiku*. At first, the blog entry is parsed for retrieving nouns, which are used for Web search query. Our system sorts out a keyword from the entry according to the Web search hit number. The chosen keyword becomes the core element of the *haiku* and helps our system to gain related words. The system uses these words for creating a *haiku*. And they are rearranged according to the order of the relation degree. Then these words are inserted into *haiku* templates. The manually prepared are chosen to fit parts of speech of retrieved words. The system chooses the most appropriate *haiku* from the generated candidates by using Web search. The appropriate *haiku* means to be correct grammatically and make sense naturally.

Finally, we describe impression evaluation of the poems generated by our systems (baseline and enhanced) with using Semantic Differential method [6].

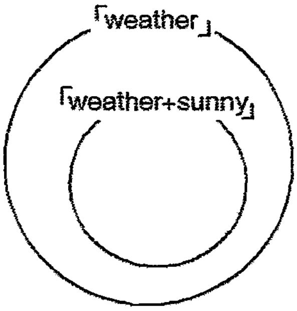

Fig. 1 . Co-Occurrence Rate Model

2. Methods

2.1 Retrieving related words

In this paper, we consider a related degree between words as Co-Occurrence rate, which shows the possibility of Co-Occurrence per a word. The greater the relation degree is, the higher the rate becomes. At first, the blog entry is parsed for retrieving nouns by using a morphological analyzer. Our system considers blog title as the summary, which is used for Web search query. The system acquires the Web search hit number and calculates the score by numerical formula (1) where N_1 is

[1] We use italic for Japanese language words and expressions.

substituted for a noun and N_2 for the title.

$$CoP(N_1,N_2)=\frac{Ret(N_1,N_2)}{Ret(N_1)} \quad (1)$$

A noun N_1 with the highest CoP (Cooccurrence Probability) score becomes the key word in the entry. Similarly, every noun in the entry is scored by using the keyword K by numerical formula (2) where N_1 is substituted for a noun and N_2 for title and N_3 for the keyword.

$$N(N_1,N_2,K)=\frac{CoP(N_1,K)}{CoP(N_1,N_2)} \quad (2)$$

(if N (Noun) score exceeds 1, it changes into reverse-ratio score.)

These nouns are sorted according to the order of N score. Our system also acquires an appropriate adjective for every noun by using the Web search engine snippets. A snippet is parsed and adjectives are extracted. The most appropriate adjective is sorted out by numerical formula (1) where N_1 is substituted for the keyword and N_2 for the adjective candidates. Similarly, our system sorts out the most appropriate *kigo* from *kigo*-database by numerical formula (1) where N_1 is substituted for the keyword and N_2 for the *kigo* candidates. The *kigo*-database has about 2,000 words, which are separated per a season.

2.2 Creating a haiku

Our system creates a *haiku* with retrieved words. These words are inserted into *haiku* templates to gain maximum score. The templates are automatically created from 314 *haikus* composed by famous *haiku* poets. The *haikus* are chosen from 20,000 *haikus* at random from a modern *haiku* database[2]. The system chooses the most grammatically correct *haiku* from the generated candidates by using Web search hit number which is acquired by using the candidates for search query for getting rid of grammatically incorrect *haiku*.

2.3 Evaluating the generated haiku

Subjects were 13 men and 10 women in their early twenties. Their average frequency of writing a blog is once a month and they are familiar with this type of poetry but have no habit of composing *haiku*. We used Semantic Differential method for evaluating the generated *haiku* and applied Varimax Normalized method with assuming no correlation between factors [7]. We prepared 15 pairs of adjectives. Subjects filled out a questionnaire by selecting one from eight degrees of an adjectives pair and evaluate one *haiku* for avoiding an order effect. One *haiku* is created by using baseline method which does not use the Web and words from the entry and a *kigo* from *kigo*-database are inserted at random. The season of the *kigo*-database is aligned with the blog entry. The other *haiku* is created by our proposed method.

Table 1. Adjective Pairs

unenjoyable (*tsumaranai*)	enjoyable (*omoshiroi*)
restless (*ochitukinonai*)	calm (*ochitukinoaru*)
tedious (*kurushii*)	fun (*tanoshii*)
ugly (*minikui*)	beautiful (*utsukushii*)
inferior (*ototteiru*)	superior (*sugureteiru*)
unarranged (*barabarana*)	arranged (*matomatta*)
unfamiliar (*shitashiminikui*)	familiar (*shitashimiyasui*)
shallow (*asai*)	deep (*fukai*)
bad (*warui*)	good (*yoi*)
complex (*fukuzatsuna*)	simple (*tanjyunna*)
unpleasant (*kimochinowarui*)	pleasant (*kimochinoyoi*)
dull (*nibui*)	sharp (*surudoi*)
poor (*mazushii*)	rich (*yutakana*)
old (*furui*)	new (*atarashii*)
stupid (*orokana*)	wise (*kashikoi*)

3. Results

Table 2. Results of Impression Evaluation

Minimum Score:0		Maximum Score:7	
	Baseline	Proposed	T-test
Average	3.69	4.51	*

*: significant difference ($p<0.05$)

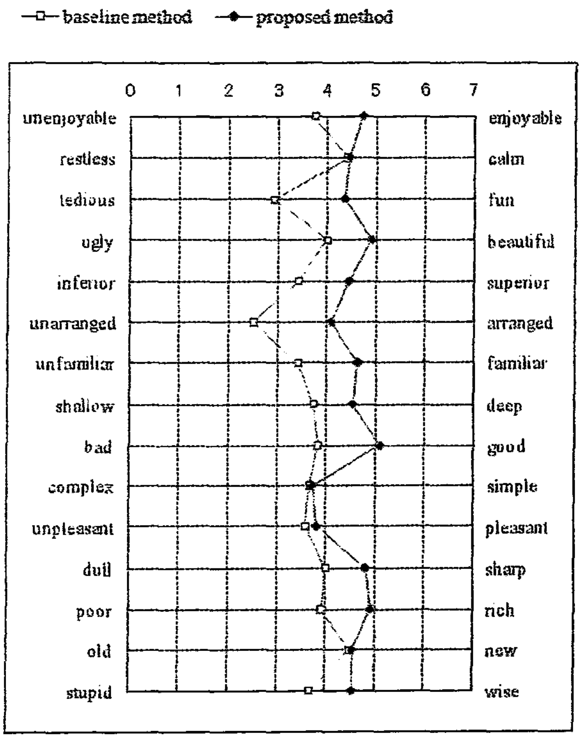

Fig. 2 . Impressions for Each Adjective Pair

[2] http://www.haiku-data.jp/

Table 3. Factor Loadings (Varimax Normalized)

	I	II	III
enjoyable	**0.634**	0.371	0.374
calm	0.157	**0.639**	-0.182
fun	**0.565**	**0.641**	-0.013
beautiful	0.312	**0.817**	0.141
superior	**0.606**	**0.561**	0.071
arranged	0.427	0.318	0.48
familiar	**0.875**	-0.018	-0.056
deep	0.196	**0.504**	0.191
good	**0.834**	0.346	0.134
simple	-0.152	-0.011	**0.563**
pleasant	0.188	-0.027	**0.859**
sharp	0.496	0.288	0.209
rich	**0.721**	0.441	-0.302
new	0.488	**-0.551**	-0.001
wise	**0.894**	0.246	0.039
proportion	0.317	0.204	0.111

Table 4. Averages of Factor Scores

	Satisfaction	Deepness	Articulateness
Baseline	-0.349 (σ=0.815)	-0.296 (σ=0.836)	-0.203 (σ=0.785)
Proposed	0.381 (σ=1.071)	0.322 (σ=1.059)	0.222 (σ=1.234)
T-test	+	–	–

+: significant tendency ($p<0.10$)

4. Discussion

It is found that proposed method's score is higher than the baseline in impression evaluation and there is a significant difference between both methods in the test result. It suggests that proposed method has achieved a positive impression. It has been shown that every pair of adjective's score in proposed method is higher than that in previous one. The maximum difference is 1.59 pts in arranged – unarranged and minimum difference is 0.04 pts in calm – restless, which suggests that the generated *haiku* with random words causes a negative impression. 15 pairs of adjectives are classified by Factor Loadings. It is found that adjectives of enjoyable, good, and rich are high score in factor I. The height of the score leads to name factor I "Satisfaction". "Satisfaction" is defined as the word which includes of the meanings of enjoyable, good, and rich. Similarly, we create factor II "Deepness" which is defined as the meanings of calm and deep and old, and factor III "Articulateness" which is by simple and pleasant. It can be observed that "Satisfaction" shows a significant difference between both methods. However, other factors show no significant difference. It suggests that proposed method especially have an improvement room in "Deepness" and "Articulateness".

5. Conclusion and Future Work

The paper describes *haiku* generation method using Web search and blog entry. The generated *haiku* with proposed method compares favorably with the baseline system in the result of impression evaluation. All the factors have shown their superiority in the score, but significant difference has not been shown. Therefore, our proposed method needs to improve especially in "Deepness" and "Articulateness". We consider that addition of rephrasing process to our method could lead to be more appropriate haiku in "Deepness" and "Articulateness".

References

[1] R. Yoshioka, To Build a *Kigo*-database and a Trial to specify the *Kigo* for the *Haiku* Automatically, IPSJ SIG Technical Report, 2000, pp. 57-64.

[2] R. Yoshioka, The Development of the *Kigo* –database and Outline of the *"Haiku Entry and Appreciation System"*, IPSJ SIG Technical Report, 71, 2006, pp.25-32.

[3] N. Tosa, H. Obara, M. Minoh, & S. Matsuoka, Hitch-Haiku, Japanese Haiku Poem Creation Support System by Computer, The Institute of Image Information and Television Engineers, 62, 2, 2008, 247-255.

[4] Xiaofeng Wu, N. Tosa, and R. Nakatsu, New Hitch Haiku: An Interactive Renku Poem Composition Supporting Tool Applied for Sightseeing Navigation System, Entertainment Computing, ICEC2009, 2009, 191-196.

[5] Page, L., Brin, S., Motwani, R., & Winograd, T, The Pagerank Citation Ranking: Bringing Order to the Web, Tech. Rep. Computer Systems Laboratory, Stanford University, Stanford, CA, 1998.

[6] M. Inoue, & T. Kobayashi, The Research Domain and Scale Construction of Adjective-pairs in a Semantic Differential Method in Japan, Japanese Journal of School Psychology, 33(3), 1985, 253-260..

[7] H. Kaiser, Computer Program for Varimax Rotation in Factor Analysis, Educational and Psychological Measurement, XIX, 1959, 413-420.