Automatic Generation of *Iroha-Uta* Poetry

Takeshi Akatsuka, Masahide Sugiyama

The University of Aizu, Tsuruga, Ikki-machi,
Aizuwakamatsu City Fukushima, 965-8580 JAPAN
{m5051101, sugiyama}@u-aizu.ac.jp

Abstract. Study on word play using computer is both computational and artistic challenge. This study is an attempt on automatic generation of *Iroha-Uta* poetry using the computer. *Iroha-Uta* is poetry containing all *kanas* without duplication and has been composed by the intelligence of humans. Here, *kanas* are Japanese syllabic writing system. Generation of *Iroha-Uta* is a combinatorial problem of how to choose appropriate words from given vocabulary. This paper describes an algorithm based on the frequency distribution of *kanas* and an advanced algorithm using word bigram.

1 Introduction

Humans have used words and voice sounds in their languages as one of the instruments of play. Study on word play using computer is both computational and artistic challenge. *Iroha-Uta* poetry is one of the most artistic word plays in Japan and has been composed by sophisticated people. This study is an attempt to generate *Iroha-Uta* poetry automatically using the computer. *Iroha-Uta* is poetry satisfying the following two conditions.

1. *Iroha-Uta* must contain all *kanas*.
2. Duplication of *kana* is not permitted.

Here, *kanas* are Japanese syllabic writing system. One of the most famous *Iroha-Uta* poetry is as follows:

色は匂へど	散りぬるを	わが世誰ぞ	常ならむ
(i ro wa ni o e do)	(chi ri nu ru o)	(wa ga yo da re zo)	(tsu ne na ra mu)
有為の奥山	今日越えて	浅き夢見じ	酔ひもせず
(u i no o ku ya ma)	(ke fu ko e te)	(a sa ki yu me mi ji)	(e i mo se zu)

Generation of *Iroha-Uta* poetry is a combinatorial problem of how to choose appropriate words from given vocabulary. *Iroha-Uta* poetry generation using a neural computing has been proposed [5]. This study has formulated *Iroha-Uta* generation problem as a tree search and proposed the efficient searching algorithms based on the frequency distribution of *kanas* [4]. However, the generated *Iroha-Utas* had no meanings because these algorithms did not consider the order of words. This paper describes an algorithm based on the frequency distribution of *kanas* and an advanced algorithms using word bigram.

2 Algorithms of *Iroha-Uta* Poetry Generation

In this study, *Iroha-Uta* poetry is defined as the sequence of the words satisfying the two conditions. Partial *Iroha-Uta* is the sequence of words without duplication of *kana*, and complete *Iroha-Uta* is partial *Iroha-Uta* whose word length is equal to 46, which is the number of *kanas*.

The flag bit corresponding to *kana* is prepared for each word. The bits corresponding to *kanas* which appear in the word are set 1; the others are set 0. In order to check the duplication of *kana* between two words, w_1 and w_2, $F(w_1) \otimes F(w_2)$ is calculated. Here, $F(w_1)$ and $F(w_2)$ are the flag bit of w_1 and w_2. If the result of $F(w_1) \otimes F(w_2)$ is 0, there is no duplication of *kana* between two words. However, if the result is not 0, there is duplication of *kana* between two words. The flag bit of the word composed by connecting two words $w_1 + w_2$ is calculated as follows:

$$F(w_1 + w_2) = F(w_1) \oplus F(w_2). \tag{1}$$

The partial *Iroha-Uta* and n-th word in the vocabulary are denoted as W and w_n. N is the size of vocabulary used for generation and M is the number of *kanas* and is equal to 46.

2.1 Algorithm using Frequency Distribution of *Kanas*

The algorithm using frequency distribution of *kanas* is as follows:

Step (0) $X = \{w_1, w_2, \ldots, w_N\}$ and $W = \phi$ (empty). $h(X)$ is calculated.
Step (1) $k_{\min}$ is determined based on $h(X)$.
Step (2) $w_{\min}$ is determined based on $k_{\min}$ and is connected with W.
- If the word length of new W is equal to M, it is complete *Iroha-Uta* and stop.
- Otherwise, go to Step (3).

Step (3) All words containing the duplication of *kana* with $w_{\min}$ are deleted from X and go to Step (1).

where X, $k_{\min}$ and $w_{\min}$ are vocabulary, the *kana* with the lowest frequency and the word containing $k_{\min}$. The flag bit for each word is considered as the vector and the frequency distribution of *kanas* is calculated by the sum of the vectors of all words in the vocabulary and defined as follows:

$$h(X) = \sum_{w \in X} h(w), \tag{2}$$

When the minimum frequency of *kana* is 0, the search is back tracked because it is obvious that complete *Iroha-Uta* cannot be generated. After $w_{\min}$ is connected with W, all words containing the duplication of *kana* with w_{min} are deleted from

X. If there is duplication of *kana* between $w_{\min}$ and word w, w is excluded from X, and $h(X)$ is updated.

$$h(X) = h(X) - h(w), \tag{3}$$

where $h(w)$ is the vector of w.

2.2 Algorithm using Word Bigram

This algorithm uses word bigram to consider the connection of words and is formulated using Dynamic Programming (DP) technique. Eq. (4) defines the algorithm Fig. 1.

$$g_i^n = \max_{1 \le m \le M} \left\{ g_m^{n-1} + d_{m,i}^{n-1} - a \times h(w_i) \cdot h(w_m) \right\}, \tag{4}$$

where a, $h(w_m)$ and $h(w_i)$ are the weight, the frequency distribution of *kanas* of word w_m and w_i. In Fig. 1, n-th layer contains M nodes, and m-th node in $n-1$-th layer is connected to i-th node in n-th layer. Nodes and layers mean the words used for generation and the connections. The connection is characterized by distance, $d_{m,i}^{n-1}$, which is defined as Eq. (5) using word bigram.

$$d_{m,i}^{n-1} = P(w_i | w_m), \tag{5}$$

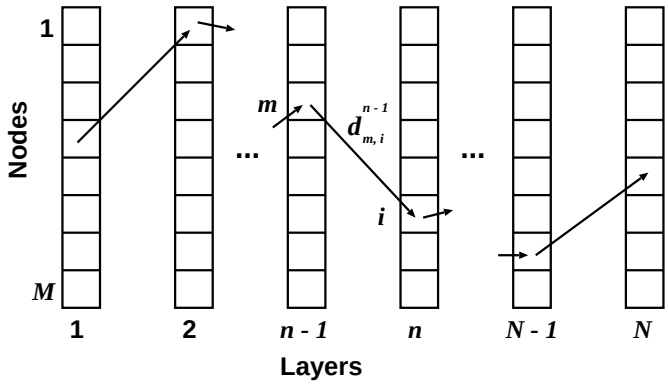

Fig. 1. Formulation of Optimal Word Connection

3 Experiment of *Iroha-Uta* Poetry Generation

In the experiment, the word sets composed of 500 to 700 words were used and 10 sets were prepared for each number of words. Fig. 2 shows the results for algorithm using frequency distribution of *kanas*. For 700 words, the average and

maximum execution time was about 2,000 seconds and 4,000 seconds, and the average and maximum number of generated complete *Iroha-Utas* was about 37 millions and 80 millions. The execution time was almost proportional to the number of generated complete *Iroha-Utas*.

The experimental results using word bigram will be shown in the final paper.

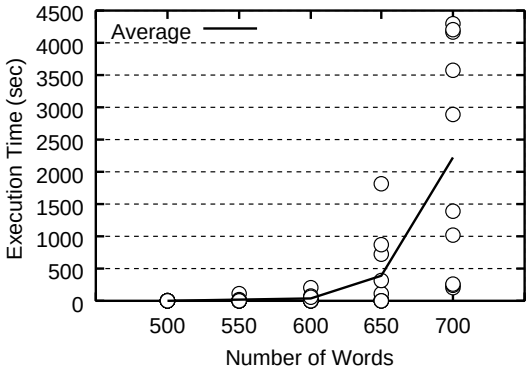

Fig. 2. Results for algorithm using frequency distribution of kanas

4 Conclusion

This paper describes two algorithms of *Iroha-Uta* poetry generation. The first algorithm, an algorithm using frequency distribution of *kanas*, could generate all *Iroha-Utas*, but the generated *Iroha-Utas* had no meanings. The second algorithm, an advanced algorithm using word bigram, could generated more meaningful *Iroha-Utas*.

In the final paper experimental results using word bigra will be shown.

References

1. T. Akatsuka and M. Sugiyama, "Study on Algorithm of *Iroha-Uta* Generation", *ECEI2000*, 2H-24 (Aug. 2000). (in Japanese)
2. T. Akatsuka, "Generation of *Iroha-Uta*", Under Graduation Thesis in the Univ. of Aizu, UGT2000-1051001 (Feb. 2001).
3. T. Akatsuka and M. Sugiyama, "Generation of *Iroha-Uta* using Frequency Distribution of *Kana*", *Proc. of IPSJ*, 1Q-6 (Mar. 2001). (in Japanese)
4. T. Akatsuka, M. Sugiyama, "Generation of Iroha-Uta using Frequency Distribution of Kana", *Proc. of CIT2001 (Journal of Shanghai University, English Edition, Vol.5, Suppl.)*, pp.152-157 (Sep. 2001).
5. N. Yoshiike, M. Kitabata, and Y. Takefuji, "A Neural Computing Approach for Composing *Iroha-Uta*," *Technical Report of IPSJ*, Vol. 2000, No. 85, pp. 61-64 (Sep. 2000). (in Japanese)