

# Discovering Poetic Allusion in Anthologies of Classical Japanese Poems

Kouichi Tamari<sup>1</sup>, Mayumi Yamasaki<sup>1</sup>, Takuya Kida<sup>1</sup>, Masayuki Takeda<sup>1</sup>,  
Tomoko Fukuda<sup>2</sup>, and Ichirō Nanri<sup>3</sup>

<sup>1</sup> Department of Informatics, Kyushu University 33, Fukuoka 812-8581, Japan

<sup>2</sup> Fukuoka Jo Gakuin College, Ogōri 838-0141, Japan

<sup>3</sup> Junshin Women's Junior College, Fukuoka 815-0036, Japan

{tamari, yamasaki, kida, takeda}@i.kyushu-u.ac.jp

{tomoko-f@muc, nanri-i@msj}.biglobe.ne.jp

**Abstract.** WAKA is a form of traditional Japanese poetry with a 1300-year history. In this paper we attempt to semi-automatically discover instances of poetic allusion, or more generally, to find similar poems in anthologies of WAKA poems. The key to success is how to define the similarity measure on poems. We first examine the existing similarity measures on strings, and then give a unifying framework that captures the essences of the measures. This framework makes it easy to design new measures appropriate to finding similar poems. Using the measures, we report successful results in finding poetic allusion between two anthologies KOKINSHŪ and SHINKOKINSHŪ. Most interestingly, we have found an instance of poetic allusion that has never been pointed out in the long history of WAKA research.

## 1 Introduction

WAKA is a form of traditional Japanese poetry with a 1300-year history. A WAKA poem is in the form of TANKA, namely, it has five lines and thirty-one syllables, arranged thus: 5-7-5-7-7.<sup>1</sup> Since one syllable is represented by one KANA character in Japanese, a WAKA poem consists of thirty-one KANA characters.

WAKA poetry has been central to the history of Japanese literature, and has been studied extensively by many scholars. Most interestingly, FUJIWARA NO TEIKA (1162–1241), one of the greatest WAKA poets, is also known as a great scholar who established a theory about rhetorical devices in WAKA poetry.

One important device is HONKADORI (allusive-variations), a technique based on specific allusion to earlier famous poems, subtly changing a few words to relate it to the new circumstances. It was much admired when skilfully handled. This device was first consciously used as a sophisticated technique by TEIKA's father

---

<sup>1</sup> The term WAKA originally meant Japanese poetry as opposed to Chinese poetry, but it is frequently used as a synonym of TANKA (short poem), which is the clearly dominant form of Japanese poetry, although NAGAUTA (long poem) and SEDŌKA (head-repeated poem) are included in the term WAKA, as well.

---

<i>Poem alluded to.</i> (KOKINSHŪ #147)	Anonymous. <i>Topic unknown</i>
HO-TO-TO-KI-SU	<i>Oh, Cuckoo, you sing</i>
NA-KA-NA-KU-SA-TO-NO	<i>Now here, now there, all about,</i>
A-MA-TA-A-RE-HA	<i>In a hundred villages.</i>
NA-HO-U-TO-MA-RE-NU	<i>So I feel you are estranged to me,</i>
O-MO-FU-MO-NO-KA-RA.	<i>Though I think you are dear to me.</i>
<i>Allusive-variation.</i> (SHINKOKINSHŪ #216)	SAIONJI KINTSUNE.
<i>At the 1500 game poetry contest</i>	
HO-TO-TO-KI-SU	<i>Oh, Cuckoo, you must be singing</i>
NA-HO-U-TO-MA-RE-NU	<i>In some other villages now.</i>
KO-KO-RO-KA-NA	<i>But in this twilight,</i>
NA-KA-NA-KU-SA-TO-NO	<i>I cannot feel you are estranged to me,</i>
YO-SO-NO-YU-FU-KU-RE.	<i>Though you are not here with me.</i>

---

**Fig. 1.** An example of poetic allusion. The hyphens ‘-’ are inserted between syllables, each of which was written as one KANA character although romanized here. English translation due to Dr. Kei Nijibayashi, Kyushu Institute of Technology.

FUJIWARA NO SHUNZEI (1114–1204) and then established both theoretically and practically by TEIKA himself, although its use had begun in earlier times. Figure 1 shows an example of poetic allusion.

For interpretations of poems utilizing this device, one must know what poems they allude to. Although the poems alluded to might be obvious and well-known at the time of writing, they are not so for present-day researchers. The task of finding instances of poetic allusion has been carried out, up till now, exclusively by human efforts. Although the size of each anthology is not so large (a few thousand poems at most), the number of combinations between two anthologies is on the order of millions.

Recently, an accumulation of about 450,000 WAKA poems became available in a machine-readable form, and it is expected that computers will be able to help researchers in finding poetic allusion.

In this paper we attempt to semi-automatically detect instances of poetic allusion, or more generally, similar poems. One reasonable approach is to arrange all possible pairs of poems in decreasing order of *similarity* values, and to examine by human efforts only the first 100 pairs, for example. A reliable similarity measure on WAKA poems plays a key role in such an approach.

How to define *similarity* is one of the most difficult problems in AI. Since a WAKA is a natural language text, it seems to require both syntactic and semantic analyses. However such analyses are very difficult, especially when the text is a poem. In this paper we choose a different approach. That is, we take a poem simply as a string, i.e. a chain of characters, and define the similarity measure in such a way that no natural language processing technique is required.

We first re-examine the already existing similarity (dissimilarity) measures for strings, and give a unifying framework which captures the essences of the measures. This framework describes a measure in terms of the *set of common*

*patterns* and the *pattern scoring function*. Under the framework, we design new similarity measures which are appropriate for the problem of finding similar poems. Using these measures, we report successful results in finding similar poems between KOKINSHŪ (1,111 poems) and SHINKOKINSHŪ (2,005 poems), which are known as the best two of the twenty-one imperial anthologies, and have been studied most extensively. The results are summarized as follows.

- Of the most similar 15 of the over 2,000,000 combinations, all but two pairs were in fact instances of poetic allusion.
- The 55th similar pair was an instance of poetic allusion that has never been pointed out in the long history of such research.

Thus the proposed method was shown to be effective in finding poetic allusion. There have been very few studies on poetic allusion in the other anthologies, especially in private anthologies, until now. We hope that applying our method to such anthologies will enable us to build up a new body of theory about this rhetorical device.

In a previous work [4], we studied the problem of finding characteristic patterns, consisting of auxiliary verbs and postpositional particles, from anthologies, and reported successful results. The goal of this paper is not to find patterns, although we do use patterns in the definition of similarity measures. It may be relevant to mention that this work is a multidisciplinary study involving researchers in both literature and computer science. In fact, the fifth and sixth authors are, respectively, a WAKA researcher and a linguist in Japanese language.

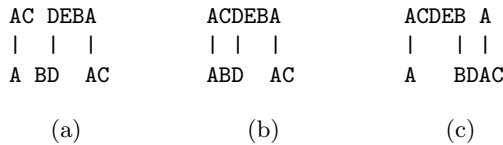
## 2 Similarity and Dissimilarity on Strings

In this section we first give an overview of already existing similarity and dissimilarity measures, and then show a unifying framework which captures the essences of the measures and makes it easy to design new measures that will be appropriate for problems in various application domains.

### 2.1 An Overview of Existing Measures

One simple similarity measure is the length of the longest common subsequence (LCS). For example, the strings ACDEBA and ABDAC have two LCSs ADA and ABA, and therefore the similarity value is 3. The alignment of the strings in Fig. 2 (a) illustrates the fact that the string ADA is an LCS of them. The pairs of symbols involving the LCS are written one above the other with a vertical bar, and the symbols do not involve the LCS are written opposing a blank symbol. Thus the LCS length is the number of the aligned pairs with a vertical bar.

On the other hand, one simple dissimilarity measure is the so-called *Levenshtein distance* between two strings. The Levenshtein distance is often referred to as the *edit distance*, and is defined to be the minimum number of editing operations needed for converting one string into the other. The editing operations

**Fig. 2.** Alignments

here are *insertion*, *deletion*, and *substitution* of one symbol. The Levenstein distance between the strings ACDEBA and ABDAC is 4, and the alignment shown in Fig. 2 (b) illustrates the situation. Note that the second symbols C and B of the two strings are aligned without vertical bar. Such a pair corresponds to the substitution whereas the unaligned symbols opposed to a blank symbol correspond to the insertion or the deletion. One can observe that the alignment in Fig. 2 (c) gives the LCS ABA with the same length as ADA, but does not give the Levenstein distance since it requires five editing operations.

The similarity and the dissimilarity are dual notions. For example, the LCS length measure is closely related to the edit distance in which only the insertion and the deletion operations are allowed. In fact the edit distance of this kind is equal to the total length of the two strings subtracted by the twice of their LCS length. This is not true if the substitution operation is also allowed.

Sequence comparison for the nucleotide or the amino acid sequences in molecular biology requires a slightly more complicated measure. A *scoring function*  $\delta$  is defined so that  $\delta(a, b)$  specifies the cost of substituting symbol  $a$  for symbol  $b$ , and  $\delta(a, \varepsilon)$  and  $\delta(\varepsilon, b)$  specify the costs of deleting symbol  $a$  and inserting symbol  $b$ , respectively. The distance between two strings is then defined to be the minimum cost of conversion of one string into the other. This measure is referred to as the *generalized Levenstein distance*.

In molecular biology, other editing operations are often used. For example, a deletion and an insertion of consecutive symbols as a unit are allowed. The cost of such an operation is called a *gap penalty*, and it is given as a function of the length of a gap. Usually, an affine or a concave function is used as the gap function. Other editing operations such as *swap*, *translocation*, and *reversal* are also used (see [1], for example).

## 2.2 A Unifying Scheme for Existing Measures

For many existing measures, similarity (dissimilarity) of two strings can be viewed as the maximum (minimum) ‘score’ of ‘pattern’ that matches both of them. In this view, the differences among the measures are the choices of

- (1) the pattern set  $\Pi$  to which common patterns belong, and
- (2) the pattern scoring function  $\Phi$  which assigns a score to each pattern in  $\Pi$ .

A pattern is generally a description that defines a language over an alphabet, but from practical viewpoints of time complexity, we here restrict ourselves to a string consisting of symbols and a kind of wildcards such as  $*$ . For example, if

we use the regular pattern set and define the score of a pattern to be the number of symbols in it, then we obtain the LCS length measure. In fact, the strings ACDEBA and ABDAC has a common pattern  $\mathbf{A*D*A*}$  which contains three symbols.

More formally, let  $\Sigma$  be the alphabet. A *wildcard* is an expression that matches one or more strings in  $\Sigma^*$ . The set of strings that a wildcard  $\gamma$  matches is denoted by  $L(\gamma)$ . Clearly,  $\emptyset \subset L(\gamma) \subseteq \Sigma^*$ . Let  $\Delta$  be a set of wildcards. A string over  $\Sigma \cup \Delta$  is called a *pattern*. Let  $L(a) = \{a\}$  for all symbols  $a$  in  $\Sigma$ . The *language*  $L(\pi)$  of a pattern  $\pi = \gamma_1 \cdots \gamma_m$  is then defined to be the concatenation of the languages  $L(\gamma_1), \dots, L(\gamma_m)$ . A pattern  $\pi$  is said to *match* a string  $w$  in  $\Sigma^*$  if  $w \in L(\pi)$ . A pattern  $\pi$  is said to be a *common* pattern of two strings  $x$  and  $y$  in  $\Sigma^*$  if it matches both of them, namely,  $x, y \in L(\pi)$ . We are now ready to define similarity and dissimilarity measures on strings. A *similarity (dissimilarity) measure* on strings over  $\Sigma$  is a pair  $\langle \Pi, \Phi \rangle$  such that

- $\Pi = (\Sigma \cup \Delta)^*$  is a set of patterns, and
- $\Phi$  is a function from  $\Pi$  to  $\mathbf{R}$ , which we call *pattern scoring function*,

where  $\Delta$  is a set of wildcards and  $\mathbf{R}$  denotes the set of real numbers. The *similarity (dissimilarity)* of two strings  $x$  and  $y$  is then defined to be the maximum (minimum) value of  $\Phi(\pi)$  among the common patterns  $\pi \in \Pi$  of  $x$  and  $y$ .

We introduce some notations to describe typical wildcards as follows.

- $*$  : a wildcard that matches any string in  $\Sigma^*$ ;
- $\phi$  : a wildcard that matches any symbol in  $\Sigma$ ;
- $\phi^{(n)}$  : a wildcard that matches any string in  $\Sigma^*$  of length  $n \geq 1$ ; and
- $\phi(u_1 | \cdots | u_k)$  : a wildcard that matches any of the strings  $u_1, \dots, u_k$  in  $\Sigma^*$ .

In addition, we use a pair of brackets, [ and ], to imply ‘optionality’. For example,  $[a]$  is a wildcard that matches both the empty string  $\varepsilon$  and the symbol  $a \in \Sigma$ . Similarly, the wildcard  $[\phi^{(n)}]$  matches the empty string  $\varepsilon$  and any string of length  $n$ .

Using these notations, we show that most of the existing similarity and dissimilarity measures can be described according to this scheme. Let

$$\begin{aligned} \Delta_1 &= \{*\}, \\ \Delta_2 &= \{\phi\}, \\ \Delta_3 &= \{\phi, [\phi]\}, \\ \Delta_4 &= \{[a] \mid a \in \Sigma\} \cup \{\phi(a|b) \mid a, b \in \Sigma \text{ and } a \neq b\}, \text{ and} \\ \Delta_5 &= \Delta_4 \cup \{[\phi^{(n)}] \mid n \geq 1\}. \end{aligned}$$

Let  $\Pi_k = (\Sigma \cup \Delta_k)^*$  for  $k = 1, \dots, 5$ .

*Example 1.* The pair  $\langle \Pi_1, \Phi_1 \rangle$  such that  $\Phi_1(\pi)$  is the number of occurrences of symbols within pattern  $\pi \in \Pi_1$  gives the similarity measure of the LCS length.

*Example 2.* The pair  $\langle \Pi_2, \Phi_2 \rangle$  such that  $\Phi_2(\pi)$  is the number of occurrences of  $\phi$  within pattern  $\pi \in \Pi_2$  gives the Hamming distance.

*Example 3.* The pair  $\langle \Pi_3, \Phi_3 \rangle$  such that  $\Phi_3(\pi)$  is the number of occurrences of  $\phi$  and  $[\phi]$  within pattern  $\pi \in \Pi_3$  gives the Levenstein distance.

*Example 4.* The pair  $\langle \Pi_4, \Phi_4 \rangle$  such that  $\Phi_4 : \Pi_4 \rightarrow \mathbf{R}$  is a homomorphism defined by  $\Phi_4([a]) = \delta(a, \varepsilon)$ ,  $\Phi_4(\phi(a|b)) = \delta(a, b)$ , and  $\Phi_4(a) = 0$  for  $a, b \in \Sigma$ , gives the generalized Levenstein distance mentioned in Section 2.1.

*Example 5.* The pair  $\langle \Pi_5, \Phi_5 \rangle$  such that  $\Phi_5 : \Pi_5 \rightarrow \mathbf{R}$  is a homomorphism defined in the same way as  $\Phi_4$  in Example 4 except that  $\Phi_5([\phi^{(n)}])$  is the gap penalty for a gap of length  $n$ , gives the generalized Levenstein distance with gap penalties mentioned in Section 2.1.

Although all the functions  $\Phi_i$  in the above examples are homomorphisms from  $\Pi_i$  to  $\mathbf{R}$ , a pattern scoring function  $\Phi$  does not have to be a homomorphism in general. In the next section, we give a new similarity measure under the framework which is suited for application to the problem of discovering the poetic allusion from anthologies of the classical Japanese poems.

### 3 Similarity Measures on WAKA Poems

We have two goals in finding similar poems in anthologies of WAKA poems. One goal is to identify poems that were originally identical, and where only a small portion has been changed accidentally or intentionally while being copied by hand. Pursuing reasons for such differences will provide clues on how the poems have been received and handed down historically.

The other goal is to find instances of poetic allusion. In this case, by analyzing how poems alluded to earlier ones, it will be possible to generate and test new theories about the rhetorical device.

In this section we consider how to define similarity between WAKA poems.

#### 3.1 Changes of Line Order

In poetic allusion, a large proportion of the expressions from an earlier poem were used in a new one. A poet must therefore take care to prevent his poem from merely being an imitation. FUJIWARA NO TEIKA gave the following rules in his writings for beginners “KINDAI SHŪKA” (Modern excellent poems; 1209) and “EIGA NO TAIGAI” (A basis of composing poems; ca 1221).

- The use of expressions from the poem alluded to should be limited to at most two lines and an additional three or four characters.
- The expressions from the poem alluded to should be located differently.
- The topic should be changed.

The second item forces us to consider all the possible correspondences between the lines of two poems. Since a poem consists of five lines, there are  $5! = 120$  different correspondences. We shall compute a best correspondence which maximizes the sum of similarities between paired lines, and define the similarity between the two poems to be the maximum sum.

Consider the two poems of Fig. 1. If we use the LCS lengths as the similarity values between the paired lines, the permutation (1, 4, 5, 2, 3) yields the correspondence that gives the maximum value 21 (see Table 1).

**Table 1.** Best correspondence

KOKINSHŪ #147	SHINKOKINSHŪ #216	sim.
1: HO-TO-TO-KI-SU	1: HO-TO-TO-KI-SU	5
2: NA-KA-NA-KU-SA-TO-NO	4: NA-KA-NA-KU-SA-TO-NO	7
3: A-MA-TA-A-RE-HA	5: YO-SO-NO-YU-FU-KU-RE	1
4: NA-HO-U-TO-MA-RE-NU	2: NA-HO-U-TO-MA-RE-NU	7
5: O-MO-FU-MO-NO-KA-RA	3: KO-KO-RO-KA-NA	1

### 3.2 Evaluation of Similarity Measure

To estimate the ‘goodness’ of a similarity measure, we need a sufficient number of examples of poetic allusion. The data we used here is from SHŪGYOKUSHŪ, a private anthology by the priest Jien (1155–1225). The 100 poems numbered from 3,472 to 3,571 of this anthology were composed as allusive variations of KOKINSHU poems, and the poems alluded to were identified by annotations. We used these 100 pairs of allusive variations as positive examples, and the other 9,900 combinations between the two sets of 100 poems as negative examples.

Using these examples, we estimated the performance of the measure based on the LCS length between the paired lines. It was found that 96% of the positive examples had similarity values greater than or equal to 10, and 96% of the negative examples had similarity values less than or equal to 10. This implies that an appropriate threshold value can classify the positive and the negative examples at high precision.

Let us denote by  $Succ_P(t)$  the number of positive examples with a similarity greater than or equal to  $t$  divided by the number of all positive examples, and by  $Succ_N(t)$  the number of negative examples with a similarity less than  $t$  divided by the number of all negative examples. The best threshold  $t$  is then defined to be the one maximizing the geometric mean  $\sqrt{Succ_P(t) \times Succ_N(t)}$ . In the above case we obtained a threshold  $t = 11$  which gives the maximum value  $\sqrt{0.9200 \times 0.9568} = 0.9382$ .

### 3.3 New Similarity Measure

Now we discuss how to improve the similarity measure. See the following poems.

*Poem alluded to.* (KOKINSHU #315)

YA-MA-SA-TO-HA / FU-YU-SO-SA-HI-SHI-SA / MA-SA-RI-KE-RU  
HI-TO-ME-MO-KU-SA-MO / KA-RE-NU-TO-O-MO-HE-HA

*Allusive-variation.* (SHUGYOKUSHU #3528)

YA-TO-SA-HI-TE / HI-TO-ME-MO-KU-SA-MO / KA-RE-NU-RE-HA  
SO-TE-NI-SO-NO-KO-RU / A-KI-NO-SHI-RA-TSU-YU

The best correspondence in this case is given by the permutation (1, 5, 4, 2, 3). One can observe that the pairs (YA-MA-SA-TO-HA, YA-TO-SA-HI-TE), (FU-YU-SO-SA-HI-SHI-SA, A-KI-NO-SHI-RA-TSU-YU), and (MA-SA-RI-KE-RU, SO-TE-NI-SO-NO-KO-RU) have scores of 2, 1, and 1, respectively, although these pairs seem completely dissimilar. That is, these scores should be decreased. On the other hand, the pair (KA-RE-NU-TO-O-MO-HE-HA, KA-RE-NU-RE-HA) has a score of 4, and it is relatively similar.

These observations tell us that the *continuity* of the symbols in a common pattern is an important factor. Compare the common pattern YA\*TO\* of YA-MA-SA-TO-HA and YA-TO-SA-HI-TE, and the common pattern KARENU\*HA of KA-RE-NU-TO-O-MO-HE-HA and KA-RE-NU-RE-HA. Thus we will define a pattern scoring function  $\Phi$  so that  $\Phi(*a*b*) < \Phi(*ab*)$ .

Let us focus on the length of clusters of symbols in patterns. For example, the clusters in a pattern  $*a*bc*d*$  are  $a$ ,  $bc$ , and  $d$  from the left, and their lengths are 1, 2, and 1, respectively. Suppose we are given a mapping  $f$  from the set of positive integers into the set of real numbers, and let the score  $\Phi(\pi)$  of the pattern  $\pi = *a*bc*d*$  be  $f(1) + f(2) + f(1)$ . For our purpose, the mapping  $f$  must satisfy the conditions  $f(n) > 0$  and  $f(n+m) > f(n) + f(m)$ , for any positive integers  $n$  and  $m$ . There are infinitely many mappings satisfying the conditions. Here, we restrict  $f$  to the form  $f(n) = n - s$  ( $0 < s \leq 1$ ).

For a parameter  $s$  varied from 0 through 1, we computed the threshold  $t$  that maximizes the previously mentioned geometric mean. The maximum value of the geometric mean was obtained for a parameter  $s = 0.8 \sim 0.9$  and for a threshold  $t = 8.9$ , and the value was  $\sqrt{0.9600 \times 0.9680} = 0.9604$ .

## 4 Experimental Results

In our experiment, we used two anthologies KOKINSHŪ (compiled in 922; 1,111 poems) and SHINKOKINSHŪ (compiled in 1205; 2,005 poems), which are known as the best two of the twenty-one imperial anthologies, and have been studied most extensively. We computed the similarity values for each of the over 2,000,000 combinations in order to find the SHINKOKINSHŪ poems that allude to KOKINSHŪ poems (not forgetting that many SHINKOKINSHŪ poems allude to poems in anthologies other than KOKINSHŪ). The similarity measures we used are as follows:

**Table 2.** Frequency distributions of similarity values

(a) Measure A			(b) Measure B		
sim.	freq.	cumulat. freq.	sim.	freq.	cumulat. freq.
23	1	1	16-17	2	2
22	0	1	15-16	1	3
21	3	4	14-15	4	7
20	4	8	13-14	8	15
19	5	13	12-13	26	41
18	26	39	11-12	32	73
17	52	91	10-11	77	150
16	114	205	9-10	137	287
15	268	473	8-9	332	619
14	916	1389	7-8	1066	1685
13	3311	4700	6-7	3160	4845
12	13047	17747	5-6	10089	14934
11	50284	68031	4-5	35407	50341
10	162910	230941	3-4	134145	184486
9	394504	625445	2-3	433573	618059
8	632954	1258399	1-2	873904	1491963
7	588882	1847281	0-1	717547	2209510
6	288190	2135471			
5	66873	2202344			
4	6843	2209187			
3	318	2209505			
2	5	2209510			
1	0	2209510			
0	0	2209510			

**Measure A.** The maximum sum of similarities computed line-by-line using the LCS length measure.

**Measure B.** The maximum sum of similarities computed line-by-line using the measure presented in Section 3.3.

Table 2 shows the frequency distributions of similarity values.

We first verified that changing the measure from A to B improves the results in the sense that most of the pairs which are not so similar as poems but had relatively high similarity, now have relatively low similarity.

Next we examined the first 73 pairs in decreasing order of Measure B that have a similarity value greater than or equal to 11. It was found that 43 of the 73 pairs were indicated as poetic allusion in [2,3], the standard editions of SHINKOKINSHŪ with annotations. The other 30 pairs were generally not considered to be poetic allusions, although some of them seem to be possible instances. Note that such judgements are to some extent subjective.

All but three of the first 15 pairs were identified as poetic allusion in [2,3]. One of the three exceptions seems actually to be poetic allusion, while this does not seem to be the case for the remaining two. The two had long expressions in common, HA-RU-KA-SU-MI TA-NA-HI-KU-YA-MA-NO and \*NO-MI-TO-RI-SO I-RO-MA-SA-RI-KE-RU, respectively. However, both of these expressions are frequent in WAKA poems, so cannot be considered specific allusions. By considering the frequencies of the expressions, the similarity values of such pairs can be decreased.

It should be emphasized that the following pair, ranked 55th in Measure B, was apparently an instance of poetic allusion of which we can find no indication in [2,3].

*Poem alluded to.* (KOKINSHŪ #826)

A-FU-KO-TO-WO / NA-KA-RA-NO-HA-SHI-NO / NA-KA-RA-HE-TE  
 KO-HI-WA-TA-RU-MA-NI / TO-SHI-SO-HE-NI-KE-RU

*Allusive-variation.* (SHINKOKINSHŪ #1636)

NA-KA-RA-HE-TE / NA-HO-KI-MI-KA-YO-WO / MA-TSU-YA-MA-NO  
 MA-TSU-TO-SE-SHI-MA-NI / TO-SHI-SO-HE-NI-KE-RU

It is considered that this pair has been overlooked in the long research history of WAKA poetry. Note that the rank of this pair in Measure A was 92 ~ 205.

The experimental results imply that the proposed measure is effective in finding similar poems. We cannot give a precise evaluation of the results since we have no complete list of instances of poetic allusion between the anthologies.

## 5 Discussion and Future Work

One can observe in the following two poems that the strings YOSHINO and YAMA, originally in the second line, appear separately in the first and the second lines of the new poem.

*Poem alluded to.* (KOKINSHŪ #321)

FU-RU-SA-TO-HA / YO-SHI-NO-NO-YA-MA-SHI / CHI-KA-KE-RE-HA  
 HI-TO-HI-MO-MI-YU-KI / FU-RA-NU-HI-HA-NA-SHI

*Allusive-variation.* (SHINKOKINSHŪ #1)

MI-YO-SHI-NO-HA / YA-MA-MO-KA-SU-MI-TE / SHI-RA-YU-KI-NO  
 FU-RI-NI-SHI-SA-TO-NI / HA-RU-HA-KI-NI-KE-RI

Our similarity measure is not suited for such situations since we considered only line-to-line correspondences. The following improvement will be appropriate. For strings  $u_1, \dots, u_n \in \Sigma^+$  ( $n \geq 1$ ), let  $\pi(u_1, \dots, u_n)$  be an *extended pattern* that matches any string in the language  $L(*u_{\sigma(1)}* \cdots *u_{\sigma(n)}*)$  for every permutation  $\sigma$  of  $\{1, \dots, n\}$ , and define the score to be the sum  $\sum_{i=1}^n f(|u_i|)$  for some function  $f$ .

As a preliminary result, we have shown that the function  $f$  defined by  $f(1) = 0$  and  $f(n) = n$  for  $n > 1$  gives the geometric mean  $\sqrt{0.9900 \times 0.9507} = 0.9702$ , which is better than those for Measures A and B.

As mentioned in the previous section, a highly frequent expression could not allude to a particular poem. Hence, a new idea is to assign a smaller score to a pattern if it is not *rare*, i.e., it appears frequently in other strings (poems). The *rarity* of common patterns can be formalized in terms of machine learning as follows. Suppose we are given a finite subset  $S$  of  $\Sigma^+$ , and we have only to consider the similarity on the strings in  $S$ . Let  $x, y \in S$  with  $x \neq y$ . Let us regard  $Pos = \{x, y\}$  as positive examples and  $Neg = S - Pos$  as negative

examples. The *rarest common pattern*  $\pi$  of  $x$  and  $y$  with respect to  $S$  is the one satisfying  $Pos \subseteq L(\pi)$  and minimizing the one side error  $|Neg \cap L(\pi)|$ , equivalently minimizing  $|S \cap L(\pi)|$ , the frequency of  $\pi$  in  $S$ .

Significance of common patterns depends upon both their scores and their rarity. Our future work will investigate ways of unifying the two criteria.

## References

1. S. Hannenhalli and P. A. Pevzner. Transforming men into mice (polynomial algorithm for genomic distance problem). In *Proc. 36th Ann. Sympo. Found. Comp. Sci.*, pages 581–592. Springer-Verlag, 1995.
2. J. Kubota. *Shinkokinwakashū*. Nihon koten shūsei. Shincho-sha, Tokyo, 1979.
3. Y. Tanaka and S. Akase. *Shinkokinwakashū*. Shin nihon koten bungaku taikei. Iwanami-shoten, 1992.
4. M. Yamasaki, M. Takeda, T. Fukuda, and I. Nanri. Discovering characteristic patterns from collections of classical Japanese poems. In *Proc. 1st Int. Conf. Discovery Science*, volume 1532 of *Lecture Notes in Artificial Intelligence*, pages 129–140. Springer-Verlag, 1998.