和歌データベースにおける類似歌の発見

5K－10

山崎 真由美† 竹田 正幸† 福田 智子‡ 南里 一郎*

† 九州大学大学院システム情報科学研究科 ‡ 福岡女学院大学 * 純真女子短期大学

1. まえがき

和歌文学研究において，歌と歌の類似性の抽出は重要である．歌の類似性に着目することにより，過去や同時代の歌人による作品への影響を明らかにすることができ，また歌人の個性や時代による特徴を獲得することができる．本研究は，45万首もの和歌を収録した新編国歌大観のデータを対象に，計算機による和歌の類似性抽出を目指すものである．従来この類の研究は，任意の歌もしくは表現にまず注目し，次にその用例を収集するという方法で進められてきた．もし，大量の和歌のデータの中から類似したものを自動抽出することができれば，これまで見過ごされてきた新たな問題点をも，見出す端緒となりうるのである．

2. 和歌の類似性

和歌の類似性といっても，歌と歌との類似の仕方にはさまざまなものがあるが，ここでは，3種類の類似性を扱うことにする．

(1) 自立語が共通する．

(2) 付属語のなすパターンが共通する．

(3) (1),(2)と異なり，和歌を単なる文字の連鎖とみなした場合に共通した文字列を多く含む．

これまでは，(1)に着目した研究がなされてきた．しかし，例えば紀貫之が桜の歌を多く詠んだからといって，「貫之は桜の歌を好んだ」と結論するのは短絡である．なぜなら，当時は詠歌に際して題が与えられることが多く，歌人が歌語を任意に選択できない場合が少なくなかったと考えられるからである．そこで，著者らは，(2)の場合の類似性に着目した．例えば，「*ば*ざらましを*」などのパターンを扱うのである．これは反実仮想という表現技法に対応する．つまり，(2)の類似性は，表現技法上の類似性に対応している．そこで，最小記述長(MDL)原理に基づいた方法[1]を用いて，歌集からの付属語パターン自動抽出を試みた．得られたパターンの歌集ごとの相違は，歌人の個性や時代の好みを反映しているようであり，研究者に非常に興味深い視点を提供してくれている[2]．

大雑把にいえば，(1)は意味，(2)は構造の類似性に対応している．これに対して，(3)は単なる文字列としての類似性に対応すると考えられる．これは，同一の歌が改変されたものであるかもしれないし，また，本歌取りといって，古歌を踏まえて新たに歌を詠んだものであるかもしれない．いずれにせよ，(3)の類似歌が自動抽出できれば，きわめて重要な視点を与えてくれることは間違いない．

3. 和歌の類似度

Σ を文字の有限集合とする．$|\xi|$ で文字列 ξ の長さを表し，$\xi[i]$ は ξ の i 番目の要素を表す．ξ と τ を Σ 上の長さ1以上の文字列とする．$1 \leq i \leq |\xi|$ を満たす任意の i に対して，$\xi[i] = \tau[a_i]$ となるような整数の単調増加列 $a_1, ..., a_{|\xi|}$ が存在するとき，ξ を τ の部分列という．文字列 x と y に対し，ξ が両方の部分列となっているとき，ξ を x と y の共通部分列という．文字列 x,y に対する最長の共通部分列を**最長共通部分列** (longest common subsequence; LCS) という．例えば，$x = abcbdda, y = badbabd$ に対する LCS は $abbd$ であり，その長さは4である．

研究の第1段階として，LCS の長さを類似度の尺度として用いて類似歌を抽出することを考える．最も単純には，2つの和歌の LCS の長さを類似度とする方法が考えられる．しかし，本歌取りなどの場合には，対応する句の位置が変化することが一般的であるため問題が生じる．そこで，$5! = 120$ 通りの句の対応づけのうちで，対応する句の間の LCS の長さの総和の最大値

Discovering Similar Poems from Classical Japanese Poem Database

Mayumi Yamasaki†, Masayuki Takeda†, Tomoko Fukuda‡, and Ichiro Nanri*

† Department of Infomatics, Kyushu University, Fukuoka, 812-8581 Japan

‡ Fukuoka Jo Gakuin College, Ogori, 838-0141 Japan

* Junshin Women's Junior College, Fukuoka, 815-0036 Japan

表 1 各句に対する LCS（例 1）

1首目	2首目	LCS
はるかすみ	はるかすみ	5
たてるやいづこ	たてるはみやこ	5
みよしのの	さてもなほ	0
よしののやまに	やまのおくには	3
ゆきはふりつつ	ゆきやふるらむ	3

表 2 各句に対する LCS（例 2）

1首目	2首目	LCS
おもひいづる	あらぬものゆゑ	1
ときはのやまの	もみちのやまに	3
ほととぎす	ほととぎす	5
からくれなゐの	くれなゐの	5
ふりいててそなく	ふりててそなく	7

を考え，この値を類似度と呼ぶことにする．以下に示す例では，1 首目に対して 2 首目の各句がそのまま対応している．

（例 1） 春霞 たてるやいづこ みよしのの
よしのの山に 雪はふりつつ
春霞 たてるは都 さても猶
山のおくには 雪やふるらむ

各句に対する LCS の長さを表 1に示す．類似度は 16 となる．また，次に示す例では，1 首目に対して，2 首目が 1 句目から順にではなく，5-4-3-1-2 句と対応している．

（例 2） 思ひいづる ときはの山の 郭公
唐紅の ふりいでぞなく
くれなゐの ふり出てぞなく 郭公
もみぢの山に あらぬものゆゑ

各句に対する LCS の長さを表 2に示す．類似度は 21 となる．

4. 実験とその結果

この類似度の定義を検証するために以下のような実験を行った．

慈円の拾玉集の 3472 番から 3571 番までの百首は，古今集の歌を踏まえて詠まれたものであり，題詞にその本歌が示してある．そこで，この百首とその本歌を類似歌のサンプルとして用いることを考えた．ただし，古今歌を踏まえたとはいっても必ずしも文字列として似ているとは限らない．この百首について本歌との類似度を計算し，その分布を図 1に示した．また，本歌とされていない歌との類似度も比較のために計算し，その分布を図 2に示した．類似歌の 96% の類似度が 10 以上であった．また，非類似歌の 96% の類似度が 10 以下となった．閾値により類似歌を判定する場合には，閾値を 10 付近にとればよいようである．

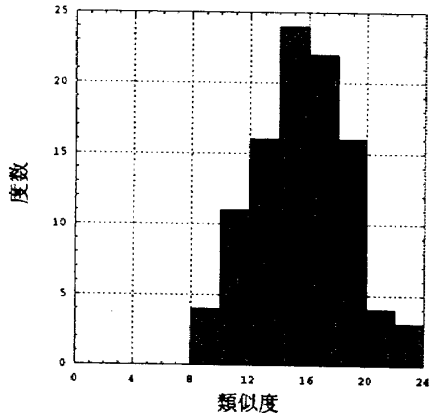

図 1 類似歌

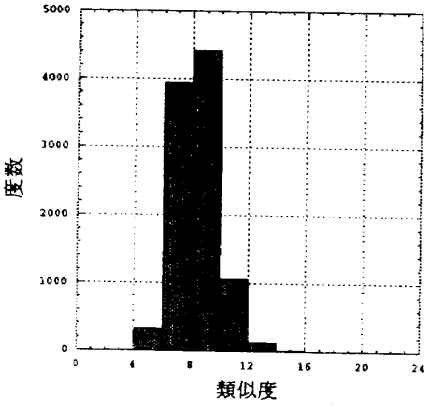

図 2 非類似歌

5. むすび

和歌のデータの中から類似したものを抽出するために，各句の LCS の長さの総和を類似度と定義した．また，類似歌，非類似歌のサンプルに対してこの類似度を計算し，その有効性を検証した．この定義では，単純に LCS の長さを類似度としており，LCS の各文字が連続している場合と，そうでない場合を区別していない．文字の連続性を考慮した類似度を定義することが，今後の課題である．

参考文献

1) A. Brazma, E. Ukkonen, and J. Vilo. Discovering unbounded unions of regular pattern languages from positive examples. In *ISACC96*, pp. 95–104, 1996.

2) M. Yamasaki, M. Takeda, T. Fukuda, and I. Nanri. Discovering characteristic patterns from collection of classical japanese poems. In *DS98*, 1998. to appear.