人文科学とコンピュータ　37－5
（1998．1．31）

MDL原理を用いた和歌データからのパターン抽出

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 ‡ 福岡女学院大学文学部 * 純真女子短期大学

要旨． 本研究は，和歌の集合からその特徴を抽出することを目的とする．特徴として付属語のなすパターンを扱う．そこで，和歌の集合に対してその被覆となるパターンの集合を求める問題を考える．Brazmaらは，与えられた文字列の集合に対する最適な被覆をMDL原理に基づいて定義した．その定義は，パターンの記述長とパターンを用いて符号化された文字列の記述長の和を最小にする被覆を最適とするものである．本研究では，5つの歌集を対象にこの手法を用いてパターン抽出の実験を行ない，その有効性を検証した．

Finding Patterns from Classical Japanese Poems based on MDL Principle

Mayumi Yamasaki† Masayuki Takeda† Tomoko Fukuda‡ Ichiro Nanri*
† Department of Informatics, Kyushu University
‡ Fukuoka Jo Gakuin College * Junshin Women's Junior College

Abstract. In this paper we address a problem of finding features from a set of Wakas (classical Japanese poems). As the features we consider patterns consisting of auxiliary verbs or particles. The problem is to find a 'good' collection of patterns covering a set of Wakas. Brazma et. al defined a good collection of patterns for a given set of strings based on the minimum description length (MDL) principle. The best pattern set is the one that minimizes the sum of the length of the patterns and the length of the strings when encoded with the help of the patterns. We applied this method for finding patterns from five collections of Wakas, and obtained encouraging results.

1　まえがき

本稿では，和歌の集合からその特徴を抽出する問題を扱う．和歌の特徴といえば，これまでもっぱら歌語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花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．というのは，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．

それでは，和歌の特徴としてどのようなものを扱えばよいであろうか．以下に示す3首は，“三夕の歌”として知られる新古今集の名歌である．

さびしさは その色としも なかりけり
　　真木立つ山の 秋の夕暮
心なき 身にもあはれは しられけり
　　しぎ立つ沢の 秋の夕暮
見わたせば 花ももみぢも なかりけり
　　浦のとま屋の 秋の夕暮

この3首は，いずれも秋の夕の風情を詠んだ点が共通しており，表現世界が類似した歌であるといえる．しかし，この3首が三夕の歌と呼称されるゆえんは，表現技法の類似性にある．すなわち，この3首は，

けりの三句切れと結句の体言止め

が共通している．本稿では，和歌の骨格をなすこのような構造をふし(節)とよぶことにする．

和歌は，歌語という“素材”を，ふしという“器”に盛りつけたものと捉えることができる．素材(歌語)の選択が制限されていたとすれば，歌人たちの関心は，それをどのような器(ふし)に盛るかに集中したに違いない．したがって，ふしは和歌の表現技法に関する“特徴”と考えることができる．本研究は，和歌の集合に共通して現れるふしを抽出することにより，歌集ごとあるいは歌人ごとの特徴を獲得することを目的とする．

2 研究の方針

大雑把に言えば，歌語が自立語であるのに対し，ふしは付属語のつくるパターンと考えることができる．以下の3首には，

*れば*こそ*けれ*

というパターンが共通して現れている．

月見れば ちぢに物こそ かなしけれ
わが身ひとつの 秋にはあらねど

ゆふされば わが身のみこそ かなしけれ
いづれの方に 枕さだめむ

老いぬれば おなじ事こそ せられけれ
きみはちよませ きみはちよませ

ここで，“こそ”は助詞，“けれ”は助動詞でいずれも付属語であるが，“れば”は動詞の活用語尾＋助詞である．このように，本研究は，学校文法に則らず，自立語の活用語尾も付属語と同等に扱う立場をとる．

さて，和歌の集合からこのような付属語のなすパターンを抽出する問題を考えよう．このためには，次の二つが必要である．

(1) 和歌において付属語を決定すること．

(2) 和歌の集合の特徴となるような付属語のなすパターンを抽出すること．

(1)のためには，和歌に形態素解析を施せばよいが，その過程で曖昧さの問題が生じる．現代日本語文に対する形態素解析の研究は古くから行われているが，決定的な方法は未だに得られていない．まして対象は古文である．古文の形態素解析については少数の研究があるが，辞書等の未整備もあって，実用レベルにはほど遠い．そこで，本研究では，通常の形態素解析を放棄し，付属語と同形の文字列はすべて付属語とみなす．

(2)は，文字列の集合を正規パターン言語の有限和で覆う問題と考えることができる．Σを文字の有限集合とし，$*$をΣに含まれない文字とする．$(\Sigma \cup \{*\})^+$の元を**正規パターン**，または単にパターンと呼ぶ．パターンπに含まれる$*$をそれぞれ$(\Sigma \cup \{*\})^*$の元で置き換えてパターンπ'が得られるとき，$\pi' \preceq \pi$と書く．$L(\pi) = \{w \in \Sigma^* \mid w \preceq \pi\}$を$\pi$によって**定義される言語**という．文字列の有限集合Aに対してパターンの有限集合$\Pi = \{\pi_1, \ldots, \pi_k\}$が$A \subseteq \bigcup_{i=1}^{k} L(\pi_i)$を満たすとき，$\Pi$を集合$A$の**被覆**(cover)という．与えられた集合Aに対する被覆は一般に無限個存在するが，その中から和歌の特徴をよく表していて歌学の立場から“面白い”ものを得たい．そこで，被覆に“面白さ”の尺度を導入し，最適化問題を解くことを考える．

それでは，最適な被覆はどのように定義すべきであろうか．この問題は非常に難しい問題であるが，統計学的観点からは以下のような方法が考えられる．すなわち，仮説であるパターン集合Πの確率モデルとΠから文字列の集合Aを得る過程の確率モデルを仮定し，事後確率$P(\Pi|A)$を最大にするΠを最適な被覆とする方法である．また，これから最小記述長(minimum description length; MDL)原理を導くことができる[4]．Brazmaら[3]は，この方法に基づき，パターン集合Πの記述長と，集合AをΠを用いて符号化したときの記述長の和を最小にするΠを最適な被覆と定義した．また，この方法を遺伝子データに適用してその有効性を示した[2]．

一方，Arimuraら[1]は，被覆のパターン数の上限kがあらかじめわかっているという仮定のもとで，集合Aのk極小多重汎化(k-minimal multiple generalization; k-mmg)を定義し，これが正データからの帰納推論の観点から最適であることを示した．しかし，k-mmgは一意に定まらないこと，kの値をあらかじめ見積もる必要があること，kの値を少し大きくとると計算時間が現実的でなくなること等の問題がある．

本稿では，MDL原理に基づく最適な被覆の定義を採用する．和歌からふしを抽出する問題にこの方式が適しているかどうかはわからないが，研究の第1段階としてこの方式を採用し，その有効性を検証する．

3 付属語列とパターン

本研究では，付属語とは助詞，助動詞だけでなく，動詞や形容詞の活用語尾，接尾辞などを含むものとする．前節で示したパターン

*れば*こそ*けれ*

において，定数文字列は付属語もしくはその連接になっている．このような文字列を**付属語列**とよぶことにする．助動詞の連接を**助動詞列**，助詞の連接を**助詞列**とよぶ．助動詞列は動詞活用語尾または形容詞のカリ活用に接続し，助詞列は用言活用語尾または助動詞列に接続する．したがって，付属語列は，以下のいずれかの形となる．

- 動詞活用語尾 + 助動詞列
- 動詞活用語尾 + 助動詞列 + 助詞列
- 用言活用語尾 + 助詞列
- 助動詞列
- 助詞列
- 助動詞列 + 助詞列
- 用言活用語尾
- 接尾辞

付属語の連接においては前の語の活用形を考慮した．また，助動詞間および助詞間の連接に関しては，岩波古語辞典巻末の基本助動詞解説，基本助詞解説に従って，助動詞を5つに，助詞を6つにそれぞれ分類し，この分類に基づいた連接規則を用いた．

本研究では付属語列と同形の文字列はすべて付属語列とみなす．このため，付属語の誤認の問題が生じる．この問題をいくらかでも回避するため，次のようにした．

(a) 句末に生起する文字列に限定した．

(b) よみの同じ句を漢字表記のものに置き換えたデータを用いた．

ふしを構成する付属語列はほとんどが句末に生起するため，(a)のようにしても重要な付属語列を取りこぼすことはない．しかし，句末の文字列に限定しても，あまのがは(天河)の“は”や，あしひきの(足引の)の“きの”など，自立語の一部を付属語と誤認する場合が多い．このような誤認は，データが漢字表記されていれば避けることができる．そこで，よみによる句索引のデータを利用して，よみが同一の句を，漢字表記された句に統一して置き換えた．すなわち，

としのうちに 春はきにけり ひととせを
こぞとやいはむ ことしとやいはむ

という歌は

年の内に 春は来にけり 一年を
去年とやいはむ 今年とや言はむ

に置き換えた．

次に，新編国歌大観の和歌のうち350,197首を対象に，(a),(b)の方法によって上で定義した付属語列の生起頻度を調査した．生起した付属語列は，延べ3,010,248個，異なり29,203個であった．生起頻度順に並べた上位50の付属語列を表1に示す．ただし，生起頻度は重複を許した．すなわち，句末に“ものを”が生起した場合には，“を”としても計上した．

表 1: 高頻度付属語列

文字列	頻度	文字列	頻度
の	257536	つ	14068
に	156754	るらん	13791
る	125135	ける	13779
も	93023	ず	12841
て	85397	りけり	12538
は	85266	み	12536
を	64747	で	12112
ん	62650	なる	12043
り	52859	そ	11376
き	51769	とも	11208
し	47357	るかな	10592
な	47197	らぬ	10562
ば	44095	れて	10293
や	40642	らむ	10215
らん	39400	す	10062
かな	38936	しき	9976
く	38620	るる	9419
ぬ	35419	こそ	9383
と	32652	なり	9285
む	26794	ど	9155
けり	24627	つつ	8724
れば	19355	にも	8721
ふ	19205	より	8709
れ	17078	にけり	8681
ぞ	15340	べき	8350

$C \subseteq \Sigma^+$ を付属語列の有限集合とするとき，

$$\Phi(C) = \{*\beta_1 * \cdots * \beta_h * \mid h > 1 \wedge \beta_1, \ldots, \beta_h \in C\}$$

の要素を**付属語パターン**とよぶことにする．上の調査で 1 回以上生起した付属語列の集合を C としたときの $\Phi(C)$ のパターンについて，(a),(b) の方法を用いて，生起頻度を調査した．生起したパターンの延べ数は 46,900,640，異なり数は 22,719,627 であった．頻度順に並べた上位 20 個のパターンを表 2に示す．これらのパターンは，和歌の骨格をなす“ふし”とは認め難い．次節では，MDL 原理を用いて付属語パターンの中から“面白い”ものを抽出する方式について述べる．

表 2: 高頻度付属語パターン

パターン	頻度	パターン	頻度
＊の＊に＊	50112	＊の＊な＊	23917
＊の＊の＊	45248	＊の＊り＊	22079
＊の＊る＊	40725	＊も＊の＊	21428
＊に＊の＊	35162	＊の＊らん＊	21217
＊の＊ん＊	30160	＊の＊かな＊	21113
＊に＊る＊	28587	＊て＊の＊	20501
＊の＊も＊	27817	＊に＊て＊	20260
＊の＊は＊	26445	＊に＊ん＊	20091
＊る＊の＊	25877	＊は＊の＊	19721
＊の＊て＊	25551	＊の＊を＊	19671

4 MDL 原理に基づく被覆の抽出

4.1 被覆の最適性の定義

まず，Brazma ら [3] に沿って，MDL 原理に基づく最適な被覆の定義を与える．パターン

$$\pi = *\beta_1 * \cdots * \beta_h * \quad (\beta_1, \ldots, \beta_h \in \Sigma^+)$$

と文字列の集合 $B = \{\alpha_1, \ldots, \alpha_n\}$ について $B \subseteq L(\pi)$ であるとする．このとき，集合 B は，パターン π と以下の文字列群によって記述できる．

$$\begin{array}{cccc} \gamma_{1,0} & \gamma_{1,1} & \cdots & \gamma_{1,h} \\ \gamma_{2,0} & \gamma_{2,1} & \cdots & \gamma_{2,h} \\ & & \vdots & \\ \gamma_{n,0} & \gamma_{n,1} & \cdots & \gamma_{n,h} \end{array}$$

ここで，各 $i = 1, \ldots, n$ について，

$$\alpha_i = \gamma_{i,0}\beta_1\gamma_{i,1}\cdots\gamma_{i,h-1}\beta_h\gamma_{i,h}$$

である．集合 B のこのような記述法を，**パターン π を用いた符号化**とよぶことにする．

ある符号化を仮定し，そのもとでの文字列 α の記述長を $\|\alpha\|$ で表すことにする．簡単のため，文字列間の区切り文字などを無視して考えると，集合 B の記述長は，次のようになる．

$$\|\pi\| + \sum_{i=1}^{n}\sum_{j=0}^{h}\|\gamma_{i,j}\|$$

π 中の $*$ をすべて取り除いた文字列を $c(\pi)$ で表す．文字列の符号化において文字単位の符号化を仮定すると，上式は次のようになる．

$$\|\pi\| + \sum_{i=1}^{n}(\|\alpha_i\| - \|c(\pi)\|)$$
$$= \sum_{i=1}^{n}\|\alpha_i\| - \Big(\|c(\pi)\| \cdot |B| - \|\pi\|\Big)$$

文字列の有限集合 A に対して，パターンと A の部分集合の対の有限集合

$$\Omega = \{(\pi_1, B_1), \ldots, (\pi_k, B_k)\}$$

で以下を満たすものを A の被覆とよぶ．

- $B_i \subseteq L(\pi_i) \qquad (i = 1, \ldots, k)$.
- $A = B_1 \cup \cdots \cup B_k$.
- $B_1, \ldots, B_k$ は互いに素．

各 $i = 1, \ldots, k$ について，集合 B_i を π_i を用いて符号化するとき，集合 A の記述長は，

$$M(\Omega) = \sum_{i=1}^{n}\|\alpha_i\| - C(\Omega)$$

となる．ここで，$C(\Omega)$ は集合 A の記述長に関する利得であって，以下で与えられる．

$$C(\Omega) = \sum_{j=1}^{k}\Big(\|c(\pi_j)\| \cdot |B_j| - \|\pi_j\|\Big)$$

$M(\Omega)$ を最小にする Ω またはそのパターンの集合を，集合 A に対する最適な被覆と定義する．$M(\Omega)$ を最小化する問題は，$C(\Omega)$ を最大化する問題と同等である．

上で与えた最適な被覆の定義は，符号化の方法に依存する．そこで，どのような符号化を選ぶかという問題が新たに生じる．ここでは，以下に述べるような素朴な方法を選んだ．文字列 $w \in \Sigma^*$ を，w の長さと w の各文字を符号化して並べた符号列との対で表す．Σ 上の確率分布 P のもとでの最適符号による w の符号化の長さを $\ell_P(w)$ で表す．また，$|w| < M$ なる正数 M を仮定する．すると，$C(\Omega)$ は次のようになる．

$$C(\Omega) = \sum_{j=1}^{k}\Big(u(\pi_j) \cdot |B_j| - w(\pi_j)\Big)$$

ここで，

$$u(\pi) = \ell_P(c(\pi)) - n_*(\pi)\log_2 M + \log_2 M$$
$$w(\pi) = \ell_P(c(\pi)) + n_*(\pi)\log_2 M$$

であり，$n_*(\pi)$ はパターン π における $*$ の生起回数である．ほとんどの和歌は短歌形式であるため $M = 32$ とした．

4.2 近似アルゴリズム

与えられた文字列の集合に対して最適な被覆を求める問題は，最小被覆問題をその特別な場合として含むため，NP-困難である．そこで，Brazma ら [3] は，次のように問題を設定し，この問題に近似解を与える Greedy アルゴリズムを示した．

> 文字列の有限集合 A とパターンの有限集合 Δ が与えられたとき，Δ の元をパターンとするような A の被覆 Ω で，$M(\Omega)$ を最小にする Ω を求めよ．

Brazma らの近似アルゴリズムは，ループの各時点において，

$$u(\pi) - \frac{w(\pi)}{|L(\pi) \cap U|}$$

の値が最大となる π を選ぶものである．ここで，U は各時点においてそれまでに選んだパターンのいずれによっても覆われていない A の元の集合である．このアルゴリズムで得られる近似解に対する $M(\Omega)$ の値は最適解の場合の高々 $\log_2 |A|$ 倍であることが保証される．

5 実験とその結果

古今集，新古今集，壬二集，拾遺愚草，山家集の 5 つの歌集を用いてパターンの抽出実験を行なった．これら 5 つの歌集について，歌数と生起した付属語パターンの数を表 3に示した．ただし，長歌などを除いたため，歌数は実際より少なくなっている．歌集ごとに生起するパターン数は比較的大きいので，2 回以上生起するものに

表 3: 歌集とパターンの生起

歌集	歌数	生起パターン数	
古今集	1,106	165,676	(8,283)
新古今集	2,005	233,571	(12,456)
壬二集	3,200	187,100	(16,429)
拾遺愚草	2,983	215,325	(14,380)
山家集	1,551	280,999	(12,970)

表 4: 実験結果

歌集	歌数	被覆のパターン数
古今集	1,106	190
新古今集	2,003	270
壬二集	3,199	369
拾遺愚草	2,981	334
山家集	1,551	228

限定し，これを集合 Δ としてアルゴリズムを実行した．ただし，新古今集，壬二集，拾遺愚草については，Δ のいずれのパターンにも合致しない歌が含まれていたため，それを除いた和歌の集合に対して適用した．得られた被覆のパターン数を表 4に示す．例えば，古今集では 2 回以上生起するパターン 8,283 個のうちの 190 個が被覆として抽出された．抽出されたパターンを Greedy アルゴリズムが出力した順に 5 個ずつ示したものが表 5である．パターンの傾向は歌集ごとに異なっているようである．

6 むすび

和歌からふしとよぶパターンを抽出する問題に対して，MDL 原理に基づく方式を適用しパターン抽出の実験を行なった．抽出されたパターンは歌集ごとに傾向が異なっており，それぞれの特徴を反映しているのではないかと考えられる．

表 5: 抽出されたパターン

歌集	抽出パターン
古今集	*げれば*べらなり*
	*ぞ*しかりける*
	*こそ*りけれ*
	*りせば*らまし*
	*は*なりけり*
新古今集	*かりせば*まし*
	*の*にけるかな*
	*は*なりけり*
	*こそ*りけれ*
	*も*かりけり*
壬二集	*ばかり*るらん*
	*こそ*なりけれ*
	*や*なるらん*
	*は*なりけり*
	*の*なりけり*
拾遺愚集	*ばかり*らん*
	*の*なりけり*
	*らざりき*の*
	*や*るらん*
	*に*なるらん*
山家集	*まし*なりせば*
	*こそ*かりけれ*
	*ならば*らまし*
	*を*ふなりけり*
	*の*るなりけり*

参考文献

[1] Arimura, H., Shinohara, T., and Otsuki, S.: Finding minimal generalizations for unions of pattern languages and its application to inductive inference from positive data, *STACS'94*, pp. 649–660, 1994.

[2] Brazma, A., Jonassen, I., Ukkonen, E., and Vilo, J.: Discovering patterns and subfamilies in biosequences, *Proc. 4th International Conference on Intelligent Systems for Molecular Biology*, pp. 24–43, 1996.

[3] Brazma, A., Ukkonen, E., and Vilo, J.: Discovering unbounded unions of regular pattern languages from positive examples, *ISAAC'96*, pp. 95–104, 1996.

[4] Li, M. and Vitanyi, P.: *An introduction to Kolmogorov complexity and its applications*, Texts and monographs in computer science. Springer-Verlag, New York, 1993.