5ZF-5

概念階層を用いた万葉集和歌検索機能の考案

岡田　雄揮[†]　　中田　充[‡]　　葛　崎偉[‡]　　吉村　誠[‡]
山口大学大学院　教育学研究科[†]　　山口大学　教育学部[‡]

はじめに

国文学研究の分野では，日本最古の和歌集である万葉集の和歌を容易に参照したいという要求があり，筆者らも和歌検索システムを実現している[1]．このシステムは，文字列マッチングにより検索語を含む和歌を検索するが，より柔軟な和歌検索のためには，検索語のみならず，その同義語・類似語・関連語なども用いた，和歌の内容に基づいた検索機能が求められる．本稿では，このうち，検索語の同義語と類似語を用いた和歌検索機能を提案する．

提案手法では，単語の意味を表す概念の階層関係を用いて検索語の同義語と類似語を求めた上で，それらを含む和歌を検索する．なお，本研究で用いる概念に関する情報は，独立行政法人情報通信研究機構が提供しているEDR電子化辞書[2]を利用する．

定義

ここでは，提案手法における概念や単語などに関する定義を示す．

概念体系：ある概念について，より抽象的な意味を持つ概念と，より具体的な意味を持つ概念が存在する．これらそれぞれを上位概念(Broader concept)，下位概念(Narrower concept)と呼ぶ．このような概念間のつながりは，概念をノードとするDAG(Directed Acyclic Graph)として表現される．このグラフは，概念体系と呼ばれ，その唯一のソースノードはルート概念(c^R)と呼ばれる(図1)．隣接する上位概念を親概念(Immediate broader concept)，隣接する下位概念を子概念(Immediate narrower concept)と呼ぶ．ルート概念の子概念として，“事象”や“時”などの7つ概念があり，その他の概念はそれらの下位概念となっている．

概念と単語：本研究において，単語はその意味をあらわす概念に属し，同じ概念に属する単語は同じ意味を持つ．概念cは，$c = (cid, cs, cmt, IB_c, IN_c)$と定義される．ここで，$cid$は概念の識別子であり，$cs$は概念見出しと呼ばれる「その概念を代表する単語」の単語見出しである．また，cmtは，その概念が持つ意味を表す解説文であり，IB_cとIN_cは，それぞれ，概念cの親概念と子概念の集合である．単語wは，$w = (wid, ws, cid)$と定義される．widは単語の識別子，wsは単語見出し（単語を構成する文字列）である．

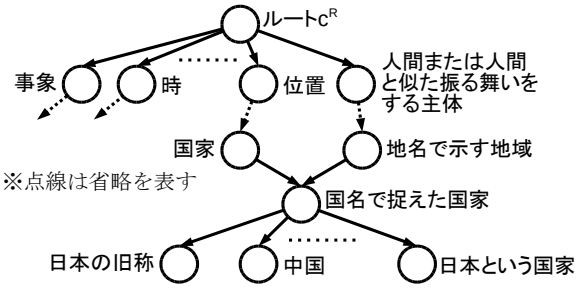

図 1：概念体系

概念
(3bcdec, “日本”, “日本という国家”, {444a40}, {})
(3bca94, “ジパング”, “日本の旧称”, {444a40}, {})
(444a40, “”, “国名で捉えた国家”, {30f772, 444a5f}, {3bcdec, 3bca94, …})
(30f772, “”, “国家”, {…}, {444a40, …})
(444a5f, “”, “地名で示す地域”, {}, {444a40, …})
単語
(JWD0373071,“日本”,3bcdec), (JWD0373072,“日本国”,3bcdec), (JWD0575082,“ジャパン”,3bcdec), (JWD0373060,“大和”,3bca94), (JWD0373051,“秋津島”,3bca94)

図 2：概念と単語の例

図2は，概念と単語の例である（それぞれ5つずつ）．図中の(3bcdec, “日本”, “日本という国家”, {444a40}, {})は，“日本という国家”を表す概念であり，その識別子は3bcdec，概念見出しは“日本”である．この概念は，“国名で捉えた国家”という概念の子概念であり，“日本の旧称”，“中国”という概念と兄弟の関係にある．

単語“日本”は，概念“日本という国家”に属し，同じ概念に属する単語“日本国”，“ジャパン”と同じ意味を持つ．なお，“国名で捉えた国家”のように，概念見出しを持たない概念も存在するが，これらは概念同士をつなぐ中間概念であり，いずれの単語もその概念に属さない．

同義語：同じ概念に属する単語をお互いの同義語と呼ぶ．例えば，単語“日本”，“ジャパン”，“日本国”は同義語である（図2）．

A proposal of search function for Japanese poems from Manyoshu by using lexical conceptual structure
[†]Yuki OKADA, [‡]Mitsuru NAKATA, [‡]Qi-Wei GE, [‡]Makoto YOSHIMURA
[†]Graduate school of Education, Yamaguchi University
[‡]Faculty of Education, Yamaguchi University

類似語：ある単語と全く同じ意味ではないが，類似した意味を持つ単語を類似語と呼ぶ．例えば，図2中の概念“日本の旧称”に属する単語“大和”と“秋津島”は，単語“日本”の類似語である．

単語xの類似語は，単語xが属する概念とその下位概念，ならびに，単語xが属する概念の類似概念とその下位概念に属する単語である．ここで，類似概念とは，類似した意味を持つ概念であり，ルート概念以外に共通の上位概念をもつ任意の概念c_1, c_2について，以下のいずれかが成立するとき，c_1とc_2は類似概念であるいう．

- 概念c_1に属する単語の単語見出しと同じ単語見出しを持つ単語を概念c_2が含む．
- 概念c_1に属する単語の単語見出しを概念c_2の概念説明（または概念見出し）が含む．

概念c_1とその類似概念c_2は，共通の上位概念と概念c_1との距離が短いほど意味が近い．また，単語xが属する概念c_xの下位概念NC_xは，c_xのより具体的な意味をもつ概念であるので，概念NC_xに属する単語は単語xの意味と極めて似た意味を持つ．

概念距離：単語xとその類似語yの意味の相違を表す尺度である．単語xが属する概念c_xと類似語yが属する概念c_yの関係に応じて，以下の2種類の概念距離を考える．

【概念距離1】$dis1(c_x, c_y)$：概念c_yが概念c_xの類似概念であり，c_xとc_yの最近傍の共通の上位概念を概念c_zとする．$c_x \neq c_z$かつ$c_y \neq c_z$のとき，概念c_xからみた概念c_yまでの概念距離1：$dis1(c_x, c_y)$は，概念c_xからc_zまでの距離（辺数）である．$c_x = c_z$または$c_y = c_z$のとき，$dis1(c_x, c_y) = dis1(c_y, c_x) = 0$である．

【概念距離2】$dis2(c_x, c_y)$：概念c_yが概念c_xの下位概念である（$c_x = c_z$）とき，概念c_xからみた概念c_yまでの概念距離2: $dis2(c_x, c_y)$は，概念c_xからc_yまでの距離である．

図3は概念体系の一部分を示している．概念aとbが類似概念であるとき，$dis1(a, b) = 3$，$dis1(b, a) = 2$である．また概念cがbの下位概念であるので，$dis1(b, c) = dis1(c, b) = 0$，$dis2(b, c) = 1$となる（同様に，$dis2(b, d) = 2$）．

同義語と類似語を用いた和歌検索機能

これまでに述べた検索語の同義語と類似語を用いた和歌検索の手順は以下の通りである．

手順1: 検索語kが属する概念c_kを求める．

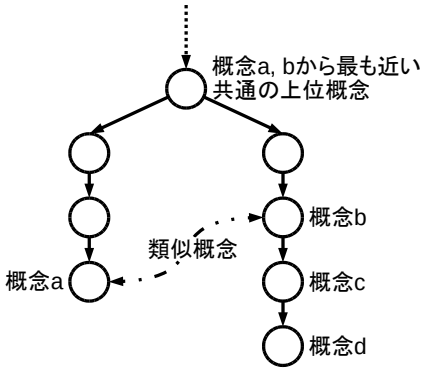

図 3：概念距離

手順2: 概念c_kの下位概念BC_kと類似概念SC_kを求める．

手順3: 類似概念SC_kの下位概念BC_{SC_k}を求める．

手順4: 概念BC_k，SC_k，BC_{SC_k}に属する単語をすべて求める．

手順5: 手順4で求めた単語を含む和歌を検索する．

手順6: 手順5で検索した和歌を，以下の基準に沿ってソーティングして，検索結果とする．

- 検索語kを含む和歌を最上位（順位1）とする．
- 検索語kの同義語を含む和歌を順位2とする．
- 検索語kの類似語のうち，概念距離1が0の概念に属する単語を含む和歌を順位3とする．但し，順位3の和歌が複数ある場合は，概念距離2の昇順とする．
- 検索語kの類似語のうち，概念距離1が0ではない概念に属する単語を含む和歌を順位4とする．但し，順位4の和歌が複数ある場合は，概念距離1の昇順とする．

さいごに

内容に基づいた和歌検索を実現するために，概念体系を用いた同義語，類似語を定義し，それらを含む和歌を検索する仕組みについて検討した．現在，これらの仕組みを実装中である．今後は，上位・下位以外の概念間のつながりも含めたより柔軟な検索機能についても検討する予定である．

謝辞：本研究は，一部，文部科学省科学研究費（挑戦的萌芽研究）（課題番号23650128）による．

文献

[1] 岡田，中田，葛，吉村：万葉集和歌検索システムの改良，平成24年度(第63回)電機・情報関連学会中国支部連合大会講演論文集，pp.452-453.
[2] 情報通信研究機構：EDR電子化辞書，http://www2.nict.go.jp/outpromotion/techtransfer/EDR/J_index.html