2ZF-5

単語の意味情報ならびに共起情報を用いた万葉集和歌検索手法の提案

岡田　雄揮†　　鍵村　好美‡　　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考にしたいという要望があり，筆者らも和歌を検索するシステムを研究している[1]．このシステムでは，和歌の本文と検索語との単純な文字列マッチングに加えて，概念体系を用いた検索語の同義語や類似語，単語の共起を用いた連想語による柔軟な検索を提案している．本稿では，柔軟かつ正確な和歌検索を実現するために，より精度の高い同義語や類似語，連想語を求める手法について考察する．万葉集に含まれる単語を抽出し，それらを意味情報と共に概念体系に組み入れることで，検索語の同義語と類似語の充実を図る．また，共起行列にも意味情報を反映することで連想語を求める精度向上を目指す．

従来のシステムについて

従来のシステムでは，検索語の同義語や類似語，連想語を求め，それらを含む和歌を検索可能とすることで，柔軟な和歌検索を実現している．検索語の同義語や類似語，連想語とは，それぞれ以下のような単語である．

同義語：検索語と全く同じ意味を持つ単語
類似語：検索語と全く同じ意味ではないが，近い意味を持つ単語
連想語：検索語と類似した意味を持つとは限らないが，検索語から連想できる単語

検索語の同義語と類似語は，EDR 電子化辞書[2]の概念体系に基づいて求める[1]．概念体系とは，概念同士のつながりを表した DAG（Directed Acyclic Graph）である．概念体系において，ある概念は一つの意味を表し，その意味を持つ単語が一つ以上属する．概念同士のつながりは親子関係で表される．ある概念の子概念以下の概念を下位概念と呼ぶ．下位概念は，より具体的な意味を持つ．逆に，ある概念の親概念以上の概念を上位概念と呼ぶ．上位概念は，より抽象的な意味を持つ．図 1 に概念体系のイメージ図を示す．

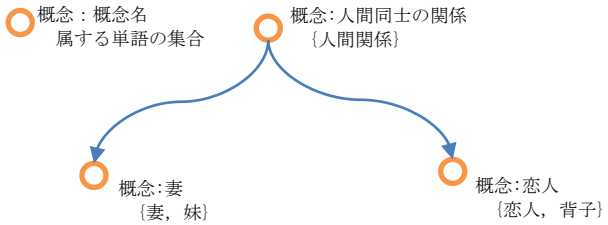

図 1:概念体系イメージ図

表 1：共起行列（一部抜粋）

	雲	草枕	大君	大和	天皇	都	宮処	国
雲		0	1	1	0	0	0	0
草枕	0		2	0	0	1	0	0
大君	1	2		1	0	2	1	3
大和	1	0	1		1	0	0	4
天皇	0	0	0	1		0	0	1
都	0	1	2	0	0		0	1
宮処	0	0	1	0	0	0		1
国	0	0	3	4	1	1	1	

従来のシステムにおいて，検索語の同義語とは，その検索語が属する概念に属する他の単語であり，図 1 の例では，「恋人」と「背子」は同義語である．また，検索語の類似語とは，検索語が属する概念と似た意味を持つ概念に属する単語であり，概念体系をたどることで求められる．図 1 の例では，「恋人」と「妹」は類似語である．

検索語の連想語は，万葉集から抽出した単語同士の共起情報を用いて求める[3]．同じ和歌に含まれる単語同士を共起しているといい，単語同士の共起回数をまとめた表を共起行列という．表 1 は万葉集の巻一（84 首）に含まれる単語同士の共起行列を一部抜粋したものである．共起行列の行は万葉集から抽出した単語(本文語)であり，列は本文語と共起している単語(共起語)である(本文語の集合と共起語の集合は等しい)．共起行列から連想関係にある単語の共起パターンを導くことで検索語の連想語が求められる．

従来のシステムでは，上記のような方法で，検索語の同義語や類似語，連想語を求めている

A proposal of search functions for Japanese poems of Manyoshu by using semantic and co-occurrence information
†Yuki Okada, ‡Yoshimi Kagimura, ‡Mitsuru Nakata, ‡Qi-Wei GE, ‡Makoto Yoshimura
†Graduate school of Education, Yamaguchi University
‡Faculty of Education, Yamaguchi University

が，以下の二つの問題がある．本稿では，これらの問題を解決するための手法を提案する．

問題1. 万葉集にのみ含まれる単語（古語など）や現代とは意味が異なる単語などが検索語の同義語や類似語として得られない．

問題2. 検索語との共起回数が低い連想語が存在する．

提案手法

上記の二つの問題を解決するために，以下の手法を提案する．

【手法1：概念体系の拡張】

EDR 電子化辞書に収録されている単語や概念は，現代日本語が中心であり，万葉集には含まれるが現代は使われなくなっている日本語（古語）などはあまり含まれていない．また，万葉集の単語には，現代日本語と同じ表記であるが，意味が異なるものも少なくない．そこで，万葉集に含まれる単語を抽出し，万葉集における意味に相応しい概念に属させるように概念体系に組み入れることで，問題 1 の解決を図る．例えば，「君が代も我が代も知るや岩代の岡の草根をいざ結びてな」という和歌から，単語「代」を抽出する．この和歌において，「代」は「年齢」という意味である．したがって，「年齢」という概念に属させるように「代」という単語を概念体系に組み入れる．

【手法2：共起行列の列の統合及び分割】

共起行列の列には，同じ意味を持つ単語が複数存在する．表 1 の例では，共起語「天皇」は「大君」などの他の共起語と同じ意味をもつ単語である．つまり，本文語「国」は，天皇という意味の単語と四回共起しているが，共起回数が複数の列（大君と天皇）に分散されている．本来，このような同じ意味を持つ共起語の共起回数は合算する必要がある．そのために，共起語の意味属性を考慮した共起行列を作成する．この共起行列では，列は一つの意味を表す共起概念となる．これにより，意味が同じ共起語は一つの列に統合され，共起語が複数の意味を持つ場合は複数の列が作成される．

実行例

上記の提案手法に則り，万葉集巻一(84 首)の和歌の訓読文を対象に単語（活用のない単語 509 語）を抽出し，手法 1,2 を実行した．図 2 は抽出した単語の一部である．手法 1 を実行した結果，これまでより多くの同義語や類似語に基づく検索が可能となった．例えば，検索語が「秋」のとき，類似語として「黄葉」を求めることができ，秋に関する和歌を幅広く検索できるようになった．

(単語の見出し，属する概念)

(黄葉，秋に色づいた木の葉)
(代，年齢)
(宮処，繁華で中心的な地域)
(皇子，天皇の息子)
(海神，地球の表面にある陸地以外の塩水をたたえている場所)
(蜻蛉島，日本国および大和国の古名で日本や大和にかかる枕詞)

図 2:抽出単語一部

表 2:共起語の意味情報を元に作成した共起行列（一部）

	雲	草枕	天皇	大和	都	昔の行政区分	国家	故郷
雲	-	0	1	1	0	0	0	0
草枕	0	-	2	0	1	0	0	0
大君	1	2	-	1	3	2	2	0
大和	1	0	2	-	0	1	2	1
天皇	0	0	-	1	0	1	0	0
都	0	1	2	0	-	1	0	0
宮処	0	0	1	0	-	1	0	0
国	0	0	4	4	2	-	-	-

※「-」は，共起概念に本文語が属していることを表す．

また，手法 2 により，表 2 のような共起行列が作成された．表 1 と比べ，同じ意味を持つ共起語「天皇」，「大君」の列が統合され「天皇」という共起概念になり，本文語「国」との共起回数が高くなる．これにより，問題 2 を解決している．また，「国」は複数の意味を持つ単語であるため，表 2 では，「昔の行政区分」，「国家」，「故郷」という三つの共起概念を新たに作成した．

まとめ

検索語の同義語や類似語を充実させることで，柔軟な万葉集和歌検索が可能となることが分かった．また，共起行列の同じ意味をもつ共起語を統合することにより，連想関係にある単語同士の共起回数が高くなり，連想語を求めやすくなった．

今後は，巻二以降に含まれる単語など，より多くの単語を概念体系に組み入れると共に，共起回数から連想語を求めるための閾値を求めるアルゴリズムについて検討していく．

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] 岡田雄揮，中田充，葛崎偉，吉村誠：「概念階層を用いた万葉集和歌検索機能の考察」，第75回情報処理学会全国大会講演論文集，pp837-838(2013年).
[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
[3] 岡田雄揮，鍵村好美，中田充，葛崎偉，吉村誠：「単語の共起情報に基づいた万葉集和歌検索機能の考察」，第64回電気・情報関連学会中国支部連合大会講演論文集，pp183-184(2013年).