社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2004－CH－61 (2)
2004／1／23

キーワードおよびその連想語による文例の検索に関する考察

中田 充† 葛 崎偉† 吉村 誠†

会話や文章作成の際に，頭に浮かんだイメージに対する適切な言語表現を導き出すことに苦慮したという経験は誰もが持っている．この問題を解決するために，我々は，頭に浮かんだイメージを断片的に表す複数のキーワードとそれらの連想語からイメージを的確に表現する言語表現を導き出す“思考イメージの言語表現支援システム”の実現を目指している．このシステムは，(A) 単語から連想語を導き出す，(B) 連想語から文例を導き出す，(C) 文例から言語表現を構成する，の3つから構成される．本論文では，システムの基本的な概念・設計方針を紹介すると共に，3つのプロセスのうちの(A)と(B)を対象として，キーワードおよびその連想語による文例の検索に関する考察を行う．

On Searching for Model Sentences via Keywords and Associated Words

Mitsuru NAKATA†, Qi-Wei GE† and Makoto YOSHIMURA†

All of us may have such experience once or more that how to express what appears in our minds by proper words during conversations or writings. To dissolve our such language expression problem, we try to realize a supporting system that can automatically lead our fragmented imagination to the proper language expression via some keywords and their associated words. This system is supposed to have the following three parts: (A) to deduce associated words from some keywords; (B) to deduce model sentences via the associated words; and finally (C) to generate proper language expression from the model sentences. In this paper, we introduce our basic concept and design policy of this system. And especially for the above (A) and (B), we discuss how to extract proper model sentences via keywords and the associated words from prepared model sentence database.

1 はじめに

我々が話したり文章を書いたりするときは，浮かんだイメージを頭にある「イメージと言葉との対応表」を参照しながら言語化するという作業を行っている．一般的にこの「対応表」では，一つのイメージに対して複数の言葉が対応すると同時に一つの言葉が異なるイメージに対応する．さらに，個人によって対応表の情報の量や質（語彙）に隔たりがあることが少なくない．そのため，ある一つのイメージを言語化する場合に，適切であろう言語表現を複数思い浮かべ，そのいずれが自分の思い浮かべたイメージを相手に最も的確に伝えるのか判断に苦慮し

† 山口大学 教育学部
† Faculty of Education, Yamaguchi University

たという経験や，適切であると思われる表現が全く考えつかないという経験は誰もが持っていることである．

例えば，“外で雨が降っている”というイメージを思い浮かべてそれを相手に伝える状況を考えてみる．その言語表現は単に「窓の外では雨が降っている」だけではく，なるべく臨場感あふれるように描こうとすると「なま暖かい風が吹いてきたかと思うと，土のにおいが漂うかのように大粒の雨が降り出してきた．」や「一天にわかにかき曇り，沛然として雨が降ってきた．」というようにその言語表現は様々なものが考えられる．

このような状況では，イメージを伝えようとする側（表現者）は，いずれの言語表現がより自分のイメージを適切に表現しているか，他にどのような言語表現が考えられるかなどを詳細に検討する必要がある．検討の過程において表現者は，自分が過去に触れたり使用した表現を参照したり，類義語辞書などの各種辞書や例文集を紐解いたり，既存の書籍や文学作品などに用いられている言語表現を検索することで，適切と思われる言語表現を選び出したり別の言語表現を構成する．しかし，この作業には多くの時間を必要とするため，表現者にとって大きな負担となっている．この傾向は，日常的に表現力豊な言語表現を求められることの多い作家や文学者，執筆業を生業とする者に代表されるいわゆる「文系の人間」よりも，技術者などの「理系の人間」の間で特に顕著であると考えられる．

この問題を解決することを目的として，筆者らは，頭に浮かんだイメージに対する的確な言語表現を見出す“思考イメージの言語表現支援システム”に関する研究を行っている．このシステムは，イメージを断片的に表す単語である複数のキーワードとそれらの連想語からイメージを的確に表現する言語表現を導き出すものであり，そのおおまかな処理のプロセスは，(A) 単語から連想語を導き出す，(B) 連想語から文例を導き出す，(C) 文例から言語表現を構成する，の3つから構成される．

本論文では，筆者らのシステムの基本的な概念・設計方針を紹介すると共に，上記3つのプロセスのうちの(A)と(B)を対象として，キーワードおよびその連想語による文例の検索に関する考察を行い，その課題を明らかにする．以降，2節で関連する研究に関して簡単にふれる．その後，3節で筆者らのシステムに関してその概要を述べ，4節でキーワードと連想語を用いた文例検索について述べる．

2 関連する研究

適切な言語表現を見出すために，従来は類義語辞典や国語辞典，文例集を参照していた．現在では，これらに代わって電子化辞書や全文検索システムが利用されて負担は軽減されているものの，適切な文例を容易に導き出す技術が望まれている．さらに，機械翻訳や文書管理システムにおいても，文例検索は必要とされている技術である．このような背景のもとで，類似文例検索に関する研究は従来から多くなされている．

参考文献[1]の研究では，単語を分類語彙表[2]に基づいてコード化し，その階層性を利用して類似例文を検索している．また，例文検索システムの応用例として，英文作成支援システムに関する研究[3][4]が挙げられる．参考文献[4]のシステムは，日本語の文章入力を受け付け，それに対する英語の例文を提示するというものであるが，日本語を英語に翻訳する際に，日本語の単語を概念ごとにグループ化し，グループを表現するキーワードを対応する英単語に変換するという方法を取っている．

また，概念の類似性判別のための枠組みの構築に関する研究もなされている．参考文献[5]の研究では連想実験により得られたデータから単語とそれに対する連想語の距離を定量化することで概念辞書を構築している．これに対して，

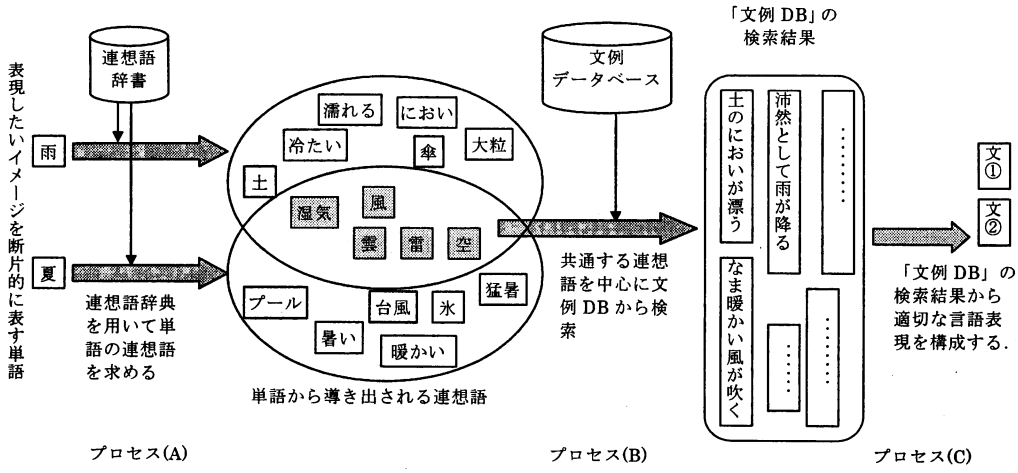

図1：システムの処理の流れ

参考文献[6]の研究では，データの収集の利便性を考慮して国語辞書や百科事典等を元にして概念ベースを構築し，それを様々な手法で精錬化している.

3 システムの概要

本研究の目的である思考イメージの言語表現支援システムの概要とその実現にあたり必要とされる 3 つの機能について述べる.

本システムは，イメージを断片的に表す単語である複数のキーワードとその連想語からそのイメージを的確に表現する言語表現を導き出すものであり，そのおおまかな処理のプロセスは以下のようになる（図 1）.

(A) 思い浮かんだイメージを断片的に表す複数のキーワードを入力として受け取り，それらから連想される連想語を導き出す.
(B) キーワードおよびその連想語を含む文例を検索する.
(C) 得られた文例をもとにして言語表現を構成して出力する.

これらの 3 つのプロセスを処理するために，(A)のプロセスでは，単語からその連想語を導き出す機能が必要とされる．そのために，まず，類義語辞書などの各種辞典，さらには古典・現代文学作品などを用いて，単語から連想される連想語を記した「連想語辞書」を構築しなければならない.

(B)のプロセスでは，(A)のプロセスで求められた連想語の集合を含む文例を検索するために，豊富なデータと高速な検索機能を持った文例データベースが必要とされる．その際，得られた文例をイメージとの適合度に応じて順序付ける機能も必要となる.

(C)のプロセスでは，(B)のプロセスで求めた文例から言語表現を構成する機能が求められる．これは，得られた文例の一部あるいは全部を用いて新たな文を構成する機能であるが本論文では対象外とする.

4 キーワードと連想語を用いた例文検索について

本節では，言語表現支援システムを実現するために必要な 3 つの機能のうち，(A)と(B)のプ

ロセスを処理する機能（以降，機能A，機能B）について考察し，その課題を明らかにする．

4.1 機能A：連想語を求める機能

機能Aは，キーワードからその連想語を導き出す機能であるが，前述のようにこれを実現するためには単語から連想される連想語を記した連想語辞書が必要とされる．連想とは「山から川を思い浮べるように，一つの観念につられてそれと関連のある他の観念が出現すること．（広辞苑第五版）」であることに従うと，連想語は「ある単語が表現する一つの観念に関連のある他の観念を表現する単語」と定義できる．しかし，豊な表現力を持つ言語表現を見出すための糸口としては，もとの単語が表現する観念と同一の観念を表現する同義語や類似の観念を表現する類義語も有用である．そこで，本論文ではこれらも含めて連想語と呼ぶ．

連想語辞書の構築にあたり，その第1段階として参考文献[7]-[10]などの類義語辞書の利用が考えられる．これらの辞書はいずれも類義語に関する情報を持つ辞書であるが，単語に対する類語を列挙し，それらの使い分けについて解説したもの[7]，単語を意味やカテゴリが近い語を集めてグループに分類したもの[8]，単語をカテゴリに分類し，そのカテゴリをさらに品詞の種類や意味の近さによってまとめた小分類に分けたもの[9][10]というように辞書の構造に違いがある．連想語辞書の作成の資料として用いるには，類義語が意味ごとにより細分化されていることが望ましい．

ここで，文献[10]の辞書の構造を用いた連想語辞書の構造を考える．文献[10]では，語句を1044のカテゴリに分け，さらに，カテゴリを意味の近さによってまとめた小語群に分けている．図2は文献[10]における“夏”の類義語を示している．“0276”は“夏”のカテゴリ番号であり，そのカテゴリには01～15までの番号を持つ小語群が存在する．例えば，小語群02は初夏の意味に近い語句が属する小語群であり，そこには“余春”などの類義語が属していることがわかる．なお，[夏]はその語句が夏の季語であることを表し，[気象]は小語群が気候に関するものであるという注意書きである．また，一般的に使われる語句はゴシック体で表し，各小語群中でひとまとまりとなる語句をセミコロンで区切っている．小語群13における「雷 [夏] 1001.4」は，“雷”がカテゴリ1001の小語群4に属する語句であることを示している．これにより，他のカテゴリに属する語句を“夏”からたどり着ける連想語として導き出すことが可能となる．

このような構造をそのまま連想語辞書に適用した場合，カテゴリ内での類義語の検索は可能であるが，カテゴリをまたいだ連想語の検索は，小語群中に他のカテゴリにも含まれる語句が存在する場合にのみ可能である．それ以外はたとえ類似の意味を持つカテゴリであったとしても，それらのカテゴリを対象とした検索は行えない．例えば，“荒天”，“風”，“曇・雲”，“霧・霞・靄”，“雨”などの類似した意味を持つカテゴリは，近接している存在している（カテゴリ番号が近い）ものの，これらをまたいだ連想語の検索は行えない．これは検索対象とする近接カテゴリの限定が困難であることによる．

この問題は，カテゴリをそれが持つ意味により階層化し，親子，兄弟などの関係にある限定されたカテゴリにのみを検索の対象とすることで解決可能である．

階層化されたカテゴリの例として，文献[11]の“一般名詞意味属性体系”が挙げられる．この文献では，カテゴリを意味属性と呼び，意味属性をそれが持つ意味により階層化（体系化）している．一般名詞意味属性体系とは，一般名詞の意味的用法を表す2710の意味属性の上位－下位関係，全体－部分関係を木構造で表したものである．

図3は意味属性体系の一部分を示したもので

0276 夏

01 夏
夏[夏]　夏季(なつき)　夏季(かき)　夏期　夏時　夏場[夏]
夏気 … … … …

02 初夏
初夏[夏]　初夏(はつなつ)[夏]　若夏[夏] ; 夏始め[夏]
夏の始め … 余春[夏] ; 立夏[夏]　芒種[夏] …
⋮

13 夏[気象]
夏の空[夏]　夏空[夏] … ; 雷[夏]1001.4 ; …
⋮

15 暑夏
暑夏(しょか)　炎夏[夏]　畏日 ; 暑し[夏]0987.02　蒸し暑し[夏]0987.04　酷暑0987.03

図2：文献[10]における“夏”の類義語

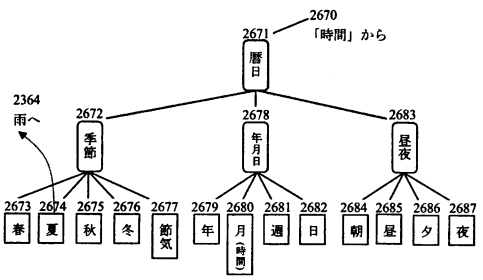

図3：一般名詞意味属性体系の一部

意味属性に属する単語の例（一部抜粋）
2671 暦日：　閏 年時
2672 季節：　折節 季 季候 季節 …
2674 夏 ：　夏期 夏季 梅雨 土用 つゆ明け …

単語体系の例（一部抜粋）
夏期：　2674夏 2695期間(自然・人間活動等)
夏季：　2674夏
梅雨：　2674夏 2695期間(自然・人間活動等)
2364雨
⋮

図4：意味体系に属する単語と単語体系の例

ある．図中の角の丸い矩形は下位の意味属性を持つものを表わす．図より，意味属性“季節”，“年月日”は意味属性“暦日”の下位意味属性であり，“夏”は“季節”の下位意味属性であることがわかる．意味属性には，図4のようにそれに属する単語が列挙されている．それぞれの意味属性に含まれる単語は，おのおのが一つ以上の意味属性に属するが，全ての単語について個々の単語が含まれる意味属性を示したものを単語体系と呼ぶ(図4)．単語体系を用いることで，“夏 →(梅雨)→ 雨”といった，上位－下位の階層関係に無い意味属性体系を横断する検索が可能となる．

我々は，基本的に一般名詞意味属性体系と単語体系の構造を用いて連想語辞書を構築する方針であるが，下記に示すような課題がある．

(1) より詳細な分類とその体系情報

意味属性体系では，その意味属性に属する単語を単に列挙しているが，連想語辞書として用いるには不都合がある．例えば，単語体系を用いて，“夏 → (梅雨) → 雨” とたどった場合，意味属性 “雨” に属する単語は，*“秋雨 雨脚 雨足 雨気 雨空 … 夕立 夕立ち 横降 横降り 雷雨”* など様々であり，全てが “夏” から連想される単語とは考えられない．そこで，文献[10]と同様に意味属性中の単語をその意味によってより詳細な小分類に分類し，その小分類に対する単語体系と同様な情報を構築する必要がある．

(2) 意味属性体系を横断する検索のための情報

より適切な連想語を得るために，単語体系以外に意味属性体系を横断する検索のための情報が必要である．このような情報として下記に挙げるものが考えられるが，ii)，iii)においては，n-gram[12][13]などの手法を用いて求めた単語間の結びつきの度合いを利用することが可能である．

i)　季語と季節の対応
ii)　慣用句・ことわざに現れる単語同士の対

応情報

iii) 古典・現代文学作品や現代文の文章中に頻繁に同時に使われている単語の対応情報

(3) 検索対象とする意味属性の絞込みの基準

連想語検索の際に，多くの意味属性を検索対象とすると，連想語とは思われない単語までが結果に含まれる．逆に対象とする意味属性が少なすぎると多様な連想語が得られない．階層化されている意味属性のどの範囲までを検索の対象とするかを決定する基準が必要である．例えば，意味属性体系の遷移をパスの長さで制限するといった方法が考えられる．

(4) 得られた連想語の順序付け

検索の結果，大量の連想語が得られることが考えられる．次のプロセスで連想語を用いて文例を検索することを考えると，これらの連想語を何らかの基準に応じて順序付けすることが望ましい．例えば，文献[10]に挙げられているような“一般的に使われる語句”は優先度を高くする，あるいは，始めに指定されたキーワードが属する意味属性からのパスの長さが小さい意味属性に属する単語の優先度を高くする，などが考えられる．

(5) 名詞以外の意味属性

名詞以外の動詞，形容詞などに関しては，単語体系に含まれるものの，それぞれの単語が含まれる意味属性が示されていない．現状では，名詞以外の連想語辞書の構築が非常に困難である．

なお，あるキーワードから連想される単語は個人によって異なるため，全ての人が連想する単語を辞書化することは不可能である．表現力豊な言語表現を見出すという本研究の目的を考えた場合，本研究における連想語辞書に求められるのは，多くの人が同じキーワードから同様に連想する単語を検索することである．したがって，本研究では連想における個人差は考慮しないものとする．

4.2 機能B：文例の検索機能

(B)のプロセスでは，(A)のプロセスで求められた連想語の集合を含む文例を検索するために，文例データベースとその検索機能が必要とされる．

文例データベース中の文例データのソースとしては，連想語辞書と同様に各種辞典の例文が考えられる．また，古典・現代文学作品，新聞記事テキストデータなどの文章やInternet上に存在する文章を分解して得られる文なども文例データとして利用可能である．これらのソースから得られた文例をデータベース中に格納し，キーワードと機能(A)によって得られた連想語を含む文例を検索するわけであるが，その実現には以下のような課題がある．

(1) 活用などへの対応

文例データベースから当該の文例を検索する際には，連想語と文例間における単なる文字列マッチングではなく，単語の活用や送り仮名の相違（“表す”と“表わす”など），さらには，異体字や歴史的仮名遣いなどを含んだ日本語の“ゆらぎ”を考慮する必要がある．異体字や歴史的仮名遣いは標準字体や現代仮名遣いに変換してからデータベース化することで解決できるが，活用や送り仮名の相違は，それらの対応情報が必要となる．

(2) 文例検索のための高速な検索機能

複数のキーワードとそれに関する連想語を含む文例を検索する際に，インデックスを使わない文字列検索を用いた場合には，様々な手法が提案されているものの，対象となる文例が大量

になると処理時間がかかるという問題がある[14]. 一般的な全文検索システムでは，インデックスの手法として，シグネチャファイル[14][15]，転置ファイル[14]などが広く用いられている．本システムにおいても何らかのインデックス手法を用いるが，いずれの手法を採用するかは今後の検討課題である．

(3) 検索結果の順序付けの基準の基準

文例の検索結果は膨大なものになることが考えられるため，検索結果を適合の程度が高い順に並べて提示する必要がある．そのために，得られた文例のキーワードに対する適合度を表す基準が必要である．例えば，多くのキーワードや連想語を含む文例の適合度を高くする，あるいは，キーワードそのものを含んでいる文例は，その連想語を含んでいる文例よりも適合度が高いと考えることが可能である．また，キーワードからより連想しやすい，つまり，概念的な距離が近い連想語をより多く含む文例が適合度の高い文例であるとするなどが考えられる．

5 おわりに

イメージを断片的に表す単語である複数のキーワードとそれらの連想語からイメージを的確に表現する言語表現を導き出す支援システムの実現を目的として，単語から連想語を導き出す機能（機能A），と連想語から文例を導き出す機能（機能B）の二つについて考察した．その結果，機能Aについては，(1) 基本とする既存の辞書よりも詳細な分類とその体系情報が必要，(2) 意味属性体系を横断する検索のための情報が必要，(3) 検索対象とする意味属性の絞込みのための基準が必要，(4) 結果として得られる連想語の順序付け基準が必要，(5) 名詞以外の意味属性がない，などの課題が明らかになった．また，機能Bについては，(1) 単語の活用などへの対応，(2) 文例検索のための高速な検索機能，(3) 検索結果の順序付けの基準，が必要である．今後は，これらの課題を解決しつつ，システムの詳細な設計と実装を行う予定である．

謝辞: 本研究の一部は平成15年度科学研究費補助金基盤研究(C)(課題番号:C15500071)による．

参考文献

[1] 兵藤安昭 河田実成 青山典生 浅井泰博 池田尚志，構文テキストベースの構築と意味分類コードを用いた類似例文検索への応用，情報処理学会研究報告. NL，自然言語処理, Vol. 100 Num. 3 pp.97-8104, (1994.3).

[2] 国立国語研究所編，分類語彙表，秀英出版，(1964.3).

[3] 寺濱幸徳 小澤邦昭 小嶋,弘行 絹川博之，英文作成支援システムにおける例文検索方式，情報処理学会全国大会講演論文集, Vol. 第 45 回平成 4 年後期 Num.3 pp.145-146 , (1992.).

[4] 武田明子 古郡廷治，例文をもとにした英文書作成支援システム，情報処理学会論文誌 Vol.35 Num.1 pp.53-61，(1994.1).

[5] 岡本潤 石崎俊，概念辞書の構築と概念空間の定量化：連想実験による概念空間の抽出，情報処理学会研究報告. NL，自然言語処理, Vol. 99 Num. 22 pp.81-88，(1999.3).

[6] 笠原要 松澤和光 石川勉，国語辞書を利用した日常語の類似性判別，情報処理学会論文誌, Vol. 38 Num. 7 pp.1272-1283, (1997.7).

[7] 田忠魁 泉原省二 金相順 編著，類義語使い分け辞典，研究社，1998.

[8] CD-ROM 日本語表現辞典, 岩波書店, 2002.

[9] 柴田武 山田進, 類語大辞典, 講談社, 2002.

[10] 山口翼 編，日本語大シソーラス，大修館書店，2003.

[11] NTT コミュニケーション科学基礎研究所監修，日本語語彙体系，岩波書店，1999.

[12] 長尾眞 森信介, 大規模日本語テキストのn

グラム統計の作り方と語句の自動抽出，情報処理学会研究報告．自然言語処理研究会報告，Vol.93 Num.61 pp.1-8，(1993.07).
[13] 近藤泰弘 近藤みゆき，N-gram の手法による言語テキストの分析方法，漢字文献情報処理研究 pp50-55 好文出版，(2001.10).
[14] 松井くにお 難波巧 井形信之，全文検索エンジン，情報の科学と技術，Vol.50 No.1 pp.9-13，(2000.1).
[15] 権藤夏男 金子邦彦 牧之内顕文，高速テキスト検索のためのパトリシアトライ構造化シグネチャファイル，電子情報通信学会技術研究報告．DE，データ工学，Vol.97 Num.161 pp.61-66 (1997.07).