ウェブからの関連語・下位語の収集手法の検討と検索システムへの応用

芳鐘　冬樹 (E-Mail: fuyuki@niad.ac.jp), **野澤　孝之** (E-Mail: nozawa@niad.ac.jp)
（大学評価・学位授与機構　評価研究部）
辻　慶太 (E-Mail: keita@nii.ac.jp), **影浦　峡** (E-Mail: kyo@nii.ac.jp)
（国立情報学研究所　人間・社会情報研究系）

Abstract

本研究では，複合語を対象として，ある用語と関連する概念を持つ関連語や，より限定されたスペシフィックな概念を持つ下位語などの関係用語を，テキストから自動的に収集する手法の検討を行う。本研究が提案する関係用語の収集手法は，テキストからの言い換え表現（用語異形）の抽出に基づく。また，実際に関連語・下位語，同義表現（を含むウェブページ）の検索を行うシステムを作成し，そのパフォーマンスについて報告する。

1　はじめに

1 つ 1 つの用語は，基本的に，それぞれ独自の概念を表すものと考えられるが，全ての用語が各々全く独立した概念を表すわけではなく，多くの用語には，何らかの関係がある用語（同じような概念を表す同義語や，より広い/狭い概念を表す上位語/下位語，あるいはその他の関連語）が存在している。そのような異なる用語間の関係についての知識は，用語が集まってできている文同士，さらには文書同士の関係の理解に役立てることが可能である。例えば，文書検索において，検索語に関係する用語と関係のタイプについての知識が利用できれば，検索の再現率の向上だけでなく，出現する用語の関係のタイプから文書間の関係を整理して，検索結果を提示することで，ユーザの利便性の向上も期待できる。そこで，本研究では，ある用語と関係する用語を，関係のタイプごとに，テキストから収集する手法の検討を行う。さらに，その手法を応用したウェブページ検索システムについての報告も行う。

関係用語の自動収集手法に関しては，文書内の共起情報に基づくものなど，様々な研究が存在する (Chen ら, 1995; 永松 & 田中, 1996)。また，ウェブページを収集源とした研究も既に存在する (小原ら, 2004)。しかしながら，それらの多くは，関係用語の収集にとどまり，関係のタイプの区別までは行っていない。区別を行っているものも，収集源が辞書の語義文などに限定されていたり，区別の範囲が上位語や下位語に限定されており (鶴丸ら, 1992; 佐藤 & 佐々木, 2003)，必ずしも十分とは言えない。本研究の手法は，複合語の形態的/統語的な言い換えに着目し，元の用語を言い換えた異形（を加工したもの）を，元の用語に関係する用語として抽出する。複合語の形態的/統語的な異形に限定されるが，言い換えの分類に基づいて，関係のタイプ（同義表現，下位語，関連語）を区別しつつ，関係用語を収集することができる。

以下では，収集対象とする関係用語のタイプ，そして収集の手法について具体的に述べ，次いで，ウェブからの関係用語収集の実験を通して，提案手法の有用性を示す。

2　関係用語のタイプと収集対象

用語間の関係のタイプには，同義関係，階層関係，関連関係がある。さらに，階層関係は類種関係（包含関係），全体部分関係，例示関係に，関連関係は同一カテゴリに属する語の関係と，異なるカテゴリに属する語の関係に分けられる (日本工業規格, 1991)。本研究では，これらのうち，同一カテゴリに属する関連語，類種関係の下位語，そして，それらの関連語・下位語および元の用語の同義表現を収集の対象とする。

このような対象の制限は，形態的/統語的な言い換えの認識のみに基づく手法の限界によるものだが，逆に，これらを他と区別して収集できるという利点があるとも言える。同一カテゴリに属する関連語だけを収集し，共通するカテゴリを上位語として系統的に整理できれば，単に関連語を集めるよりも精緻な知識が得られるものと考えられる。下位語については，分野によるだろうが，階層関係の中で類種関係が占める割合は少なくないという報告もあり[1]，ある程度のカバレジは得られると考えられる。

また，関係用語収集の出発点とする入力用語は複合語に限定している。これも言い換え認識上の制限である。本研究の手法は，一般語，専門用語を問わず適用可能だが，専門用語は複合語が大部分を占めるため (Ishii, 1987)，専門用語に対して特に効果がある。例えば，人工知能分野の『人工知能大辞典』(Shapiro & Eckroth, 1987) では，3,869 語中 3,245 語（約 84%）が 2 要素以上から成る複合語である。表 1 に，複合語 3,245 語の品詞パターンの内訳をまとめた。ただし，N は名詞，TPN は名詞性接尾辞，S は助詞を表す。2 名詞から成る複合名詞（N N）

[1] 原田ら (1988) によれば，電気工学分野では，JICST シソーラス中の階層関係のうち約 45%が類種関係である。

が全体の約 40%を占めること，そして（N N）は他の多くのパターンの基礎になっていることが，表から確認できる。本研究は，関係用語収集の最初のステップとして，この（N N）を入力用語として想定した手法の検討を行う。

N N	1297
N N N	419
N TPN N	213
N TPN	197
N N TPN	112
N S N	91
N TPN N N	71
N N N N	68
N N S N	43
N N TPN N	41
その他	693
計	3245

表 1: 複合語の品詞パターン

3 収集手法

本研究が提案する関係用語の収集手法は，テキストからの言い換え表現（用語異形）の抽出に基づく。用いる言い換え規則は，Yoshikane ら (2003), Kageura ら (2004) の規則を拡充したものであり，それらを言い換え認識システム Fastr (Jacquemin, 1994) にインプリメントして，言い換え表現の抽出を行う。作成した 211 規則は，(i) 機能語の付加/削除，元の構成要素の品詞変化，主要部の入替のみが生じるもの，(ii) 修飾語として内容語が加わるもの，(iii) 元の構成要素と等位関係を持つ形で内容語が加わるもの，の何れかに分類できる。表 2 に，(i)(ii)(iii) それぞれの言い換え規則を例示した[2] 。

- (i) は主要部入替などを含み，言い換え後の異形は同義語と呼べるものではないが，元の内容語の削除や新たな内容語の付加はなく，表層的な構造は変わっても，およその概念は保たれている。これらの規則が適用される異形を，元の用語の同義表現として抽出する。(例:「生体計測」の同義表現として「生体の計測」を抽出）
- (ii) は，修飾語の付加によって，より限定された概念を表す異形に言い換えるものである。これらの規則が適用される異形を，元の用語の類種関係にある下位語として抽出する[3] 。（例 :「生体計測」の下位語として「生体物理計測」を抽出）
- (iii) は，元の用語と共通の構成要素を含み並列関係を持つ用語を，元の用語と組み合わせるものである。これらの規則が適用される異形から,「並列関係を持つ用語」を分離して，元の用語の関連語として抽出する。(「生体計測」の関連語として,「生体・環境計測」から分離した「環境計測」を抽出）

Jacquemin (1996) は，仏語を対象にして同様の手法で用語の収集を行っている。本研究は，そのアイデアを日本語に適用し，さらに Jacquemin (1996) では用いられていない外部からの修飾を含む言い換え（(ii-2) など）も加えることで，より包括的な関係用語の収集を目指す。

図 1: 用語間の関係

テキストから抽出した関係用語は，図 1 のような 2 系列のツリー形式に整理する。1 つめは，入力用語（元の複合名詞）の主要部（後項）をルートとするツリー，2 つめは修飾部（前項）をルートとするツリーである。入力用語を [a] の位置に，入力語の同義表現を [a'] の位置に，下位語（およびその同義表現）を [c] の位置に出力する。以上は，2 つのツリーに共通している。[b1][b2] の位置には，それぞれ，入力用語の主要部を共通のカテゴリとする関連語（およびその同義表現），主要部の対象・目的などを表す修飾部を共通のカテゴリとする関連語（およびその同義表現）を出力する[4] 。2 つめのツリーは，ルートと入力用語の主要部が一致しておらず，厳密には階層関係を表していないが，関連語との関係を示すために便宜的にツリー形式で表現することとした。

4 ウェブからの関係用語収集

前節で述べた収集手法は，収集源のテキストの種類を限定するものではないが，最新のデータを含む膨

[2] 表中の NS はサ変名詞，NA は形容動詞語幹になる名詞，S は助詞，SC は名詞接続助詞，MD は助動詞，L は読点・区切り記号を表す。

[3] 内容語が語中に挿入される場合，挿入後の結合パターンによっては，元の用語の下位語と見なせないケースもある。例えば,「[自動制御理論] 学習」は「自動学習」の下位語ではない。しかし，これも分野による違いはあろうが，情報処理分野の用語を調査した原田ら (1988) の報告によれば，そのようなケースは 1 割にも満たない。

[4] 2 名詞から成る複合名詞の構成要素間の関係には，並列，同格，修飾，格修飾があるが (石崎, 1987)，ツリーを 2 系列出力するにあたって，並列関係の用語（「研究開発」など）は想定していない。そのような用語の場合，(iii) の言い換え規則は適用されず，関連語が抽出されないため，2 つのツリーは同じものになる。

	規則例	言い換え例
(i-1)	N1 N2 → N1 S1 N2	生体計測 → 生体の計測
(i-2)	NA1 N1 → NA1 MD1 N1	曖昧情報 → 曖昧な情報
(i-3)	NS1 N1 → N1 NS1	共有メモリ → メモリ共有
(ii-1)	N1 N2 → N1 N3 N2	生体計測 → 生体物理計測
(ii-2)	N1 N2 → N3 N1 N2	情報検索 → オンライン情報検索
(iii-1)	N1 N2 → N1 SC1 N3 S1 N2	生体計測 → 生体と環境の計測
(iii-2)	N1 N2 → N1 N2 L1 N3	生体計測 → 生体計測・診断

表 2: 言い換え規則の例

大な量のデータが存在すること，検索エンジンを利用できることなど，ウェブページを収集源とする利点は多い。そこで，ユーザが入力した用語の関係用語をウェブから収集するシステムを作成した。関係用語の収集が第 1 の目的だが，ウェブページ検索システムとしても利用可能である。出現する関係用語のタイプ（同義表現，下位語，関連語）に従って，ウェブページ検索の結果を整理することで，ユーザが，検索要求に適ったページを探しやすくなるとともに，提示される関連語などは，次に検索語を修正する際の参考になり，ウェブページ検索の利便性の向上が期待できると考えられる。

4.1 収集の流れ

以下の流れで，ウェブから関係用語の収集を行う。

- 検索画面（図 2）で入力された検索語を形態素に分割し，Google (http://www.google.co.jp) でそれらの AND 検索を行う（日本語のページを検索）。
- Googleが検索結果として返すウェブページのサマリから，Fastr を用いて検索語の言い換え表現を抽出する。
- 抽出した言い換え表現を整理し，用語間の関係を表すツリー形式で収集結果を表示する（図 3）。その際，それぞれの用語を含むウェブページへのリンクを張る。

形態素への分割と品詞の判定には茶筌 (松本ら, 2000) を用いている。

4.2 収集実験

評価実験として，前述の『人工知能大辞典』に記載された用語の関係用語をウェブから収集する実験を行った。入力用語は，2 名詞から成る複合名詞を無作為に 50 語選び，Google の最大ヒットページ数は 100 とした。表 3 に，同義表現，関連語，下位語それぞれの抽出数と抽出精度を示した。再現率については，今回は評価を行わなかった。再現率の評価は，ウェブページのカバレジ，検索エンジンの性能，異形抽出の性能といった，いくつかの段階に分けて行う必要があるが，これは今後の課題にしたい。

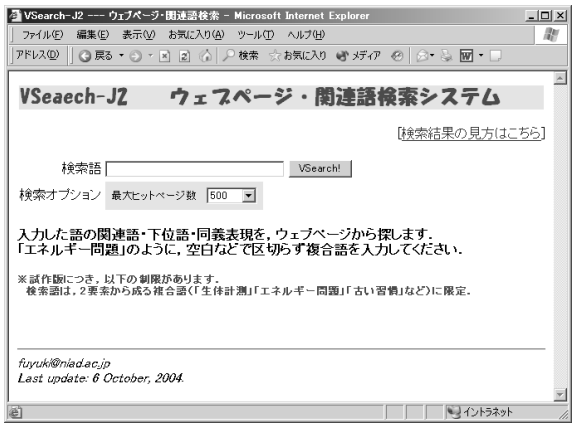

図 2: 検索画面

表 3 に示したように，約 77%の精度で関係用語を収集することができた。精度の面では，Yoshikane ら (2003) の，人工知能分野の論文抄録を抽出源にした異形抽出の結果（精度 76.8%，再現率 78.6%）とほぼ等しい。正しく収集できた関係用語数は，1 語あたり，同義表現 1.6 語，関連語約 1 語，下位語約 11 語であり，下位語の割合が非常に大きい。

4.3 誤り診断

関係用語収集の誤りは，以下の 4 つに大別できる。

(1) 形態素解析の誤りに由来するもの。

(2) 係り受け構造の抽出の誤りに由来するもの。
知識獲得 → ×知識の道具
「[知識の獲得]・[道具と人間の関わり]」から誤って「知識の獲得・道具」を抽出。

(3) 構成要素間の関係が元の用語と異なるものを収集したケース。
言語処理 → ×処理言語（〜を処理するための言語）

(4) 関係用語のタイプを誤って収集したケース。
啓蒙運動 → ×啓発啓蒙運動（下位語として）
正しくは，関連語として「啓発運動」を抽出すべき。

同義表現			関連語 [b1]			関連語 [b2]			下位語			全体
正	誤	精度	正	誤	精度	正	誤	精度	正	誤	精度	精度
80	13	0.86	24	2	0.92	27	10	0.73	534	169	0.76	0.77

表 3: 関係用語の収集結果

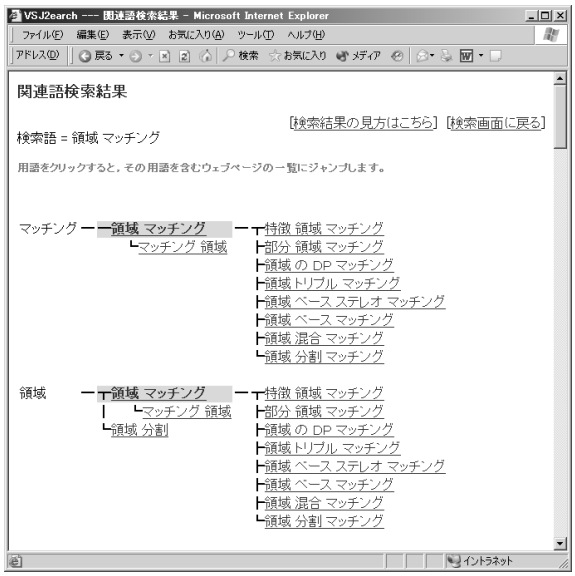

図 3: 検索結果画面

最初の 3 つは，論文抄録からの異形抽出を行った Yoshikane ら (2003) の指摘と共通している。一方，(4) は，元の用語を言い換えた異形としては正しいが，関係のタイプの区別を誤っている。形態的/統語的言い換え規則だけでは，完全な区別は難しい。

また，ウェブを収集源としたことによる問題点として，論文データと比べて誤字が多い，文の途中で改行が入っていることが多い，誤って中国語のページが混ざることがあるといった，テキストの質の低さが挙げられる。それらは (1) の誤りにつながる。

5 おわりに

本稿では，テキストからの言い換え表現の抽出に基づく，関係用語の収集手法の提案と評価を行った。下位語と比較して関連語の抽出数が非常に少ないという結果であったが，検索エンジンの最大ヒットページ数を増やすことで，下位語だけでなく関連語の抽出数もある程度は増やすことができる。収集源の分量を増やさなくても，異形抽出の前処理として構文解析を行い，並列構造を認識できれば，より多くの関連語が収集できると考えられる。また，得られた関係用語を入力用語にして，さらに，それらの関係用語を収集するというプロセスを繰り返すことにより，用語間の関係マップを広げていくことが可能になる。そのような方針で，今後研究を進めたい。

謝辞

本研究の一部は「科学研究費補助金若手研究 (B)」によるものであり，ここに謝意を表します。また，本研究を進めるにあたり，Fastr システムを提供してくださった CNRS-LIMSI の Christian Jacquemin 博士に感謝します。

References

Chen, H., Yim, T. and Fye, D. (1995) "Automatic thesaurus generation for an electronic community system," *Journal of the American Society for Information Science*, Vol.46, No.3, p. 175–193.

原田隆史, 細野公男, 田村俊作, 高柳敏子, 後藤智範, 岸田和明, 坂田亮子 (1988) "複合語の解析による語の上位－下位関係の自動抽出についての基礎的研究," 三田図書館・情報学会研究大会, p. 49–52.

Ishii, M. (1987) "Economy in Japanese scientific terminology," *Terminology and Knowledge Engineering '87*, p. 123–136.

石崎雅人 (1987) "日本語複合名詞の解析," 情報処理学会第 35 回全国大会, p. 1315–1316.

Jacquemin, C. (1994) "Fastr: a unification-based front-end to automatic indexing," *Proceedings, Intelligent Multimedia Information Retrieval Systems and Management (RIAO'94)*, p. 34–47.

Jacquemin, C. (1996) "A symbolic and surgical acquisition of terms through variation." In: Wermter, S. Riloff, E. and Scheler, G. (eds.) *Connectionist, Statistical and Symbolic Approaches to Learning for Natural Language Processing*. Heidelberg: Springer. p. 425–438.

Kageura, K., Yoshikane, F. and Nozawa, T. (2004) "Parallel bilingual paraphrase rule for noun compounds: concepts and rules for exploring web language resources," *The 4th Workshop on Asian Language Resources (ALR-04)*, p. 54–61.

小原恭介, 山田剛一, 絹川博之, 中川裕志 (2004) "ウェブを利用した関連用語収集," FIT2004 (第 3 回情報科学技術フォーラム), p. 183–184.

松本裕治, 北内啓, 山下達雄, 平野善隆, 松田寛, 高岡一馬, 浅原正幸 (2000)『日本語形態素解析システム『茶筌』version 2.2.1 使用説明書』, 奈良先端科学技術大学院大学松本研究室: 生駒, 21p.

永松健司, 田中英彦 (1996) "コーパスから抽出した係り受け共起情報に基づく類似度と文書検索における評価," 情報処理学会研究報告, NL-116, p. 73–78.

日本工業規格 (1991)『シソーラスの構成及びその作成方法：JIS X 0901-1991』, 日本規格協会: 東京, 24p.

佐藤理史, 佐々木靖弘 (2003) "ウェブを利用した関連用語の自動収集," 情報処理学会研究報告, NL-153-8, p. 57–64.

Shapiro, S. C. and Eckroth, D.; 大須賀節雄監訳 (1991)『人工知能大辞典』, 丸善: 東京, 1316p.

鶴丸弘昭, 井上淳, 日高達, 吉田将 (1992) "語義文からの階層関係の自動抽出," 電子情報通信学会技術研究報告, NLC92-2, p. 9–16.

Yoshikane, F., Tsuji, K., Kageura, K. and Jacquemin, C. (2003) "Morpho-syntactic rules for detecting Japanese term variation: establishment and evaluation," *Journal of Natural Language Processing*, Vol. 10, No. 4, p. 3–32.