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

2003－FI－ 72　(11)
2003－NL－157　(11)
2003／9／30

定型表現を利用した新聞記事からの下位概念単語の自動抽出

安藤 まや†　関根 聡‡　石崎 俊†
† 慶應義塾大学 政策・メディア研究科
‡ ニューヨーク大学 コンピュータサイエンス学科
E-mail: †{maya, ishizaki}@sfc.keio.ac.jp, ‡sekine@cs.nyu.edu

高度な自然言語処理を行なう際には、構文情報のみならずさまざまな語と語の関連情報が重要となってくる。我々は「トマトなどの野菜」といった定型表現を用いて、新聞記事から、名詞の下位概念を自動的に抽出する手法を提案する。７種の定型表現を作成し、６年分の新聞記事をコーパスとして下位概念を抽出した。その結果、ほぼ6割以上の正解率で下位概念が得られた。また、抽出した下位概念と、人間が連想した下位概念との比較をおこない、２人以上の被験者が連想した下位概念のうち、平均８５%の下位概念をコーパスから自動抽出することができた。

Automatic Extraction of Hyponyms from Newspaper Using Lexicosyntactic Patterns

Maya ANDO†　Satoshi SEKINE‡　Shun ISHIZAKI†
† Graduate School of Media and Governance, Keio University
‡ Computer Science Department, New York University
E-mail: †{maya, ishizaki}@sfc.keio.ac.jp, ‡sekine@cs.nyu.edu

Not only syntactic information but also semantic relationships between words are important in advanced natural language processing. We describe a method to automatically extract hyponyms from newspaper. First, we discover patterns which can extract hyponyms of a noun, such as "A nado-no B (B such as A)", then we apply the patterns to the newspaper corpus to extract instances. The precision is 60-90 percent depending on the patterns. We compare the extracted hyponyms and those associated by human. 85 percent of the words associated by more than 1 person are extracted automatically.

１．はじめに

自然言語をコンピュータで処理するには、構文情報、意味情報などに加え、人間の持つ知識が必要である。このような知識は、高度な自然言語処理に重要な情報となる。現在、名詞の階層関係を記述した辞書にはＥＤＲ概念辞書、日本電信電話(株)の日本語語彙体系などがある。これらは大規模な辞書であるが、ドメインによらない辞書を目指しているため、ドメイン固有の知識に応じたものでない。また、規模は大きいものの、自然言語処理を施す際に重要な頻度情報などの重要度の情報がない。本研究では、「下位概念（トマト）などの野菜」といった定型表現を用いて、大規模コーパスから下位概念を自動的に抽出する手法を提案する。コーパスから自動的に下位概念が抽出できれば、頻度情報を自動的に取得することができると考えられる。また、提案する手法はコーパスを変更することも可能で、ドメインの特徴を捉えた辞書を構築するには有効な手段であると考える。

文中に現れる定型表現を用いて、上位・下位関係や部分関係など、単語の関係する語を抽出する研究には以下のようなものがある。
Marti A. Hearst は、上位・下位関係を構成する定型表現（ex. A such as B）を発見し、半年分の「The New York Times」に対してその定型表現を適用した。評価は such as パターンに限り 200例について人手で行なった。その結果、６３％が適切であると評価された[1]。同様に Mattew Berlandr らは”A of B”などの定型表現を用いて部分関係の抽出をはかり、人間による評価を行なった結果、５５％の正解率が得られたとしている[3]。国内においては山田がニュース用語集を自動作成する目的で、ニュース文の中の定型表現（ex. A というB）を用いて上位概念を抽出している[4]。
下位概念抽出の研究では、岡本らが構築する連想概念辞書中に下位概念が含まれている。連想概念辞書とは、連想実験の結果を電子化して構築したもので、実験は被験者に刺激語である名詞と連想の制約を与えるための課題（上位概念、下位概念など７種類）を提示して行なわれる。登録されている刺激語数は約 1000 語で被験者数が 10 人である。現在被験者数を増加中で、660 語の刺激語に対して 50 人の被験者での実験が終了している[5]。
まず、本稿では定型表現を用いて下位概念を自動的に抽出し、連想概念辞書との比較、評価を行なう。作成した７種の定型表現を用いて、６年分の新聞記事から下位概念を抽出した結果、ほぼ６割以上の正解率で下位概念が得られた。また、抽出した下位概念と、先に紹介した連想概念辞書の下位概念とを比較した結果、２人以上の被験者が連想した下位概念のうち、平均８５％の下位概念を新聞記事から抽出することができた。

２．実験対象語の選択

本研究では、まず、下位概念が現れる定型表現を作成し、次にその定型表現を使用しコーパスから下位概念にあたる単語を抽出した。そして、評価としてコーパスから抽出された下位概念と連想概念辞書との比較を行なった。
一連の作業の前段階として、まず、抽出する単語の上位概念にあたる対象語を規定した。今回考察した対象語は基本単語で日常性の高い約 60 語で、その内 17 語について、「にんじん」に対する「ニンジン」「人参」といったような約 20 種類の異表記についても調べた。対象語は、まず、対象語自身が連想概念辞書に含まれており、且つその対象語の上位概念と下位概念にあたる単語も連想概念辞書に刺激語として登録されているものを中心に選択した。次に、その対象語の上位概念、下位概念にあたる単語で、連想概念辞書に刺激語として登録されているものを選択した。今回調べた対象語例を次に示す。

- 階層関係の中で比較的上位に位置すると思われる対象語：生き物　食物　道具など
- 家具：椅子 たんす ソファ テーブルなど
- 乗り物：飛行機 自動車 自転車など
- 楽器：クラリネット ピアノ ギターなど
- 動物：ほ乳類 霊長類 アザラシ 犬など
- 果物：ブドウ グレープフルーツなど
- 野菜：ほうれん草　トマト 緑黄色野菜など

また、抽出対象となるコーパスは、ＪUMAN とＫＮＰで解析した６年分の毎日新聞記事（1994－1999 年）となっている。これらの解析結果にはエラーも含まれる。形態素解析結果については、今回取り扱っているデータが基本名詞であること、定型表現の解析結果に誤りはなかったことから、そのまま採用している。構文解析結果については、特に「名詞＋助詞」といった文節が続く場合の係りうけ関係にエラーが見受けられたが、これは構文解析以外の処理で今後解決していきたいと考えている。

３．定型表現の作成

新聞記事から下位概念を自動抽出するために、まず下位概念と共に現れる確率の高い定型表現を発見する。次に、得られた定型表現を用いて、コーパスから自動的に下位概念を抽出する。
はじめに、連想概念辞書の刺激語となっている名詞とその下位関係にある名詞の双方を含む文をコーパスから抽出した。刺激語と下位概念が含まれる文から、「下位概念を含む文節」が「刺激語を含む文節」に係っている場合、または「刺

激語を含む文節」が「下位概念を含む文節」に係っている場合を中心に調査した。その結果、約３０種の定型表現候補が見つかった。その中から、用例の少ないものなどを除き、以下７種の定型表現を決定した。

a1. 下位概念＋など＋対象語
a2. 下位概念＋などの＋対象語
b1. 下位概念＋に似た＋対象語
b2. 下位概念＋のような＋対象語
c1. 下位概念＋以外の＋対象語
d1. 下位概念＋という＋対象語
d2. 下位概念＋と呼ばれる＋対象語

今回採用しなかった定型表現には以下のようなものがある。「下位概念＋を除く＋対象語」「下位概念＋をはじめ＋対象語」「下位概念＋を含め＋対象語」などはコーパス出現頻度が少なかったために外している。

４．下位概念の自動抽出

前章にて作成した定型表現を新聞記事に適用し、下位概念を抽出する。まず、対象語が含まれている文を抽出する。その対象語に、定型表現を含む文節が係っている場合には、定型表現の直前の名詞を下位概念として抽出する。以下に対象語が「楽器」の場合の例をあげる。

ビオラやチューバなどの楽器を失った。

上記の例の場合、対象語「楽器に」定型表現 a1「など」を含む文節が係っている。そのため、定型表現の直前にある、「チューバ」が下位概念として抽出される。また、a1、a2、b2、c1 については、並列して名詞が列挙されている場合には、それも下位概念として抽出するようにしているため、「ビオラ」も下位概念として抽出される。今回抽出対象とした並列表現は、助詞「ト」、助詞「ヤ」、句点「、」の三種類である。助詞の場合は、例えば「チューバなど楽器」という表現があった場合、チューバの直前にある文節が「名詞＋と」もしくは「名詞＋や」となっていた場合に抽出する。さらにその直前の文節に対しても同様の処理を加え、「バイオリンやビオラやチューバなど楽器」といった場合には「バイオリン」も抽出できるようになっている。句点の場合は、「、＋名詞」という表現が続く限り下位概念として抽出するよう設定している。「名詞＋、」で抽出すると「高校時代、ピアノ、ドラムなどの楽器をこなした」といった表現で「楽器」の下位概念として「高校時代」を抽出してしまう恐れがあったためである。

このような定型表現抽出結果に対し、定型表現ごとに評価を行なった。評価は、「AはBである」という文のAに下位概念、Bに対象語を代入したときに、文として成立する場合を正解として、手作業で行なった。表１に結果を示す。

表１：定型表現ごとの正解率

	総出現回数	正解率
a1.**など**	803	0.59
a2.**などの**	549	0.85
b1.**に似た**	23	0.70
b.**のような**	167	0.69
c1.**以外の**	86	0.74
d1.**という**	347	0.61
d.**と呼ばれる**	19	0.68

定型表現 a2 から抽出される下位語の正解率が高く、比較的多くの下位概念を抽出することができた。正解率の低い b1、b2、d1 は「鬼のような人間」「議会という生き物」といったような比喩表現や「飼い主に似た犬」といった下位概念を表さない表現が多くみられ、正解率が下がる傾向にある。a1 は、「パンクなど自転車の簡単な修理」といった係り受け関係の自動判定が困難な表現が出現したため、正解率が下がっている。このような表現は対象語ごとに傾向が異なっている。表１においてコーパス出現回数が 100 以上だった a1、a2、b2、d1 について、いくつかの対象語を例に、対象語ごとの正解率を表２に示す。

表２：延べ語数で計算した対象語ごとの正解率

	a1.など		a2.などの		b1.のような		d1.という	
対象語	述べ語数（異なり）	正解率	述べ語数（異なり）	正解率	述べ語数（異なり）	正解率	述べ語数（異なり）	正解率
野菜	67(38)	0.85	111(50)	0.95	2(2)	0.5	8(7)	0.63
緑黄色野菜	12(8)	1.0	9(6)	0.78	2(2)	1.0	0(0)	0
トマト	3(3)	0	0(0)	0	1(1)	0	0(0)	0
乗り物	14(14)	1.0	14(13)	0.86	1(1)	1.0	5(5)	0.6
自動車	36(32)	0.14	13(12)	0	4(4)	0.25	4(4)	0.5
トラック	2(2)	0	0(0)	0	1(1)	0	0(0)	0
動物	174(120)	0.85	128(82)	0.94	22(15)	0.95	18(10)	0.83
霊長類	8(5)	0.88	0(0)	0	0(0)	0	0(0)	0
人間	136(124)	0.04	13(12)	0.08	72(44)	0.86	138(104)	0.54
猿	1(1)	0	0(0)	0	0(0)	0	2(2)	0.5

a1 では、「自動車」「人間」の場合のみ、コーパス内に多く出現したにもかかわらず先に書いたような係り受け関係の自動判定が困難な表現が高い頻度で現れたため、極めて低い正解率となった。「自動車」の場合、自動車そのものについての話題よりも自動車産業の話題が多く見受けられるためか、「自動車」の性能・機能などの話題が多く、正解率が下がったものと考えられる。「人間」も同様に、病気や性質など人間にまつわる話題が多く見受けられ、「など＋人間」「など＋自動車」といった表現ではそれぞれの対象語の下位概念にあたる単語は登場しなかった。この２語はコーパス内の出現頻度も比較的高かったため、この定型表現全体の正解率にも影響を及ぼしている。a1 とよく似た定型表現である a2 においても同様の傾向が見られたが、a1 に比べ a2 では、対象語「自動車」「人間」の出現頻度が低かったため、正解率への影響が少なくなっている。このような極めて正解率の悪い対象語に対して、何らかの処置を施すのか、また抽出された下位概念そのものを選別する手法を考えるのか、今後の検討していきたい。

全体的な考察としては、対象語約６０語のうち、a1、a2、b1、d1 にて下位概念が抽出された対象語数はそれぞれ 42 語、26 語、32 語、30 語だった。予測どおり、「トマト」「猿」といったような階層関係の中で比較的下位に位置すると思われる対象語は、コーパスに出現する頻度が非常に低く、下位概念が抽出しづらい対象語となっている。

正解率を概観すると Hearst の正解率が６割程度なのに対し、本稿の結果は多少高くなっている。Hearst の場合は、定型表現「such as」などを利用して、上位概念と下位概念のセットを抽出することを目標としている。評価は「such as」パターンに限り 200 例について行なっている。本稿とは目標が異なるがこの結果とあえて比較すると、「such as」パターンには、抽象名詞などが含まれる可能性もあり、正解率の低下につながると考えられる。試験的に、抽象名詞である「意味」「移動」について下位概念の抽出を試みた。対象語「意味」は連想概念辞書に下位概念が登録されていない。コーパスから抽出された文は「一局長の私的懇談会の方針など意味はない」といったようなものが多く、下位概念に該当する名詞は抽出されなかった。一方、対象語「移動」は、連想概念辞書に「瞬間移動」「引越し」といった下位概念が登録されている。しかしながら、抽出された文は、「歩道橋や駅の階段など移動を阻む『物理的な障壁』」といった文が多く、下位概念抽出には至らなかった。本稿では、具体物を対象語とし定型表現を適用しているため、正解率があがっているものと考えられる。

５．連想概念辞書との比較

コーパスから抽出された下位概念と第一章で紹介した連想概念辞書との比較を行なう。

まず、連想概念辞書に登録されている下位概念とどのくらい一致したのかを示し、次に、連想概念辞書の下位概念の頻度情報と、コーパスから抽出された下位概念との相関を示す。

連想概念辞書と、コーパスから抽出された下位概念を比較する際、表記のゆれが見受けられた。そのため、両者の表記のゆれを一致させた上で、比較を行なっている。

５．１ 連想概念辞書の下位概念との一致度

連想概念辞書の項目「下位概念」に構築されている単語がコーパスからどのくらい抽出されたのかを調べた。約６０概念中、一語も連想概念辞書の下位概念を抽出できなかった対象語が３２語見受けられた。それらはおおよそ二つのグループに分けられる。一つは、連想概念辞書においても、下位概念として連想されている数が少なくまた連想されていても連想した被験者の数が少ないもの、つまり、人間が考えても下位概念がなかなか思い浮かばないような対象語である。「アザラシ」「グレープフルーツ」「かぼちゃ」「クラリネット」などがそれにあたる。これらの対象語は階層関係の中で下位に位置する傾向があると考えられる。もう一つは、コーパス内には連想概念辞書の下位概念が存在しているが、本稿の定型表現では抽出できなかった場合である。「トラック」「椅子」などがそれにあたる。これらはそもそも新聞記事にはあまり登場しない単語であるが、それに加え、対象語「トラック」の場合、下位概念として「４トントラック」「大型トラック」などがあげられるが、「４トントラックなどのトラック」というような表現は、「トラック」が両単語に含まれるため現れる可能性は低い。このような単語の抽出が難しいと考えられる。

表３に、対象語ごとに連想概念辞書の下位概念との一致度（一致した数／連想下位概念の総数）、コーパスから抽出された連想下位概念、抽出されなかった連想下位概念の例をいくつかの対象語について示す。抽出語はコーパス頻度が高かったものから上位３語、未抽出語はコーパス頻度がゼロだった単語で、連想概念辞書において頻度が高い順に３語をあげる。全体的に、連想概念辞書の下位概念の連想語数が多い対象語が目立っている。連想概念辞書にありながら抽出されなかった下位概念には、「ほ乳類」「穀物」「緑黄色野菜」などグループを表すことばが多く見受けられた。

５．２連想概念辞書の頻度との相関

本セクションでは、コーパスから抽出した下位概念と、連想概念辞書の下位概念を連想した人の割合との間に関係が認められるのかを考察する。今回考察に使用した連想概念辞書の被験者数は 50 人もしくは 100 人であるが、一人しか連想しなかった連想語については、データの信頼性を高めるために使用しない。

表３：連想下位概念との一致度

対象語	家具	果物	楽器	乗り物	動物	野菜	食べ物
一致度	0.47 (7/15)	0.5 (17/34)	0.43 (20/46)	0.18 (8/44)	0.61 (23/38)	0.56 (24/43)	0.19 (11/59)
抽出語	ソファ テーブル いす	リンゴ ミカン メロン	ピアノ バイオリン ギター	飛行機 自転車 ジェット コースター	人間 猫 猿	トマト ニンジン キャベツ	果物 パン 米
未抽出語	本棚 棚 ちゃぶ台	柿 パパイヤ サクランボ	太鼓 ドラム サックス	船 バイク 自動車	ほ乳類 爬虫類 ペット	緑黄色野菜 ブロッコリー レンコン	穀物 ラーメン カレーライス

コーパスから抽出した下位概念と、連想下位概念の頻度情報との相関をとった。連想下位概念を連想した被験者の割合が 20%以上、10%以上 20%未満、２%以上 10%未満の３グループにわけ、そのグループに属する連想下位概念のうち、どのくらいがコーパスから抽出されたかという割合を出した。その例を図１に示す。

多くの被験者に連想された下位概念はコーパスから抽出されやすい傾向が強いことがわかる。連想した被験者の割合が 20%以上のグループの一致度は、最も低い値で 0.63、平均は 0.85 となっている。10%以上 20%未満のグループでは、対象語ごとにかなりのばらつきが見られた。しかし、全体的には、被験者が連想した割合が減るにしたがって、コーパスから抽出される下位概念の割合が減る傾向があった。つまり、連想概念辞書に近いかたちでコーパスから下位概念が抽出される傾向があることがわかった。

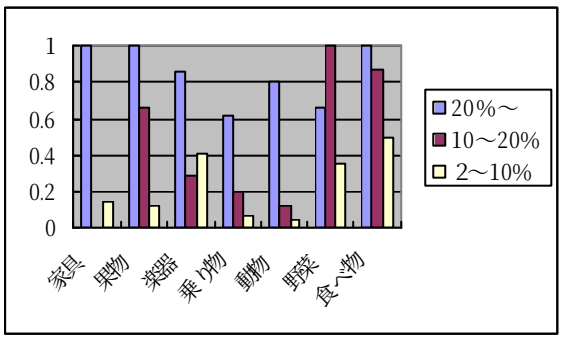

図１：連想下位概念の頻度との相関

６．今後の課題

定型表現を用いて、コーパスから下位概念を自動的に抽出し、連想概念辞書の下位概念との相関があることを示した。つまり、人間の連想結果を構築した連想概念辞書が、コーパスを使用した自動的な方法で置き換えられる可能性を示した。

今後は、現在使用している定型表現の正解率をあげる研究を行なう。Scot Cederberg はＬＳＡ（Latent Semantic Analysis）を用いてテキストから自動抽出した階層関係の精練を行ない、エラーを３０%減らすことに成功している[6]。このような手法を使用すれば、より曖昧な表現からも下位概念抽出が可能になり、定型表現を増やすことができる。筆者が発見した正解率の低い定型表現でも、このような手法を使えば精度をあげることができると考えている。

また、本稿では下位概念のみの抽出を扱っているが、上位概念、部分関係、動詞、形容詞など他の関連語についても抽出をおこなっていきたいと考えている。対象となるコーパスも新聞記事３８年分に増やすことや、ウェッブのデータの使用を予定しており、抽出語の増加をはかっていきたいと考えている。

謝辞

連想概念辞書のデータを提供してくださった慶應義塾大学の岡本潤氏に感謝いたします。

参考文献

[1] Marti A. Hearst,WordNet:An Electronic Lexical Database, Chapter 5, Automated Discovery of WordNet Relations, The MIT Press, Cambridge, Massachusetts, 1998.

[2] Marti A. Hearst. Automatic Acquisition of Hyponyms from Large Text Corpora. In Proc. Of the Fourteenth International Conference on Computational Linguistic COLING'92, 1992.

[3] Matthew Berland and Eugene Charniak,"Finding Parts in Very Large Corpora",In Proceedings of the ACL 1999, ACL. New Brunswick NJ, 1999.

[4] 山田一郎・住吉英樹・柴田正啓,ニュース記事に出現する用語と説明文の意味関係自動獲得，情報処理学会研究報告 NL152-21,pp145-152,2002

[5] 岡本潤・石崎俊, “概念間距離の定式化と電子化辞書との比較”, 自然言語処理, Vol.8 No.4, Oct. 2001.

[6] Scott Cederberg and Dominic Widdows, "Using LSA and Noun Coordination Information to Improve the Precision and Recall of Automatic Hyponymy Extraction", In Proceedings of the Seventh CoNLL conference held at HLT-NAACL 2003, Edmonton, 2003.