F-020

テキストコーパスからの上下関係抽出

Extraction of Hyponymy from Text Corpus

中渡瀬 秀一† 相澤 彰子††
Hidekazu Nakawatase Akiko Aizawa

1．まえがき

本稿ではコーパス中に含まれる上下関係にある名詞を抽出するための計算方法を提案する．名詞間の上下関係は類義関係のうちでも重要なもののひとつであり，また高度な自然言語処理には不可欠であるシソーラスの構成要素でもある．これまでにも上下関係を自動抽出する手法がいつくか提案されているがそれらは，1）右主要部規則に基づいた方法，2）特定の文型や文の表層的パターンに基づいた方法，3）単語間の関係に基づいた方法，の3種類の方法に大別される．文には語形成レベル，表層構造レベルや深層構造レベルなどといった異なるレベルの構造が混在していることから，それらのどこに注目するかによって上記のような異なる系統の手法が導かれていた．また実用化に向けてはこれらの各手法を組み合わせることによって抽出精度を向上させる試みもある[4]．

本研究は前者の立場をとり，その中でも3番目の単語間の関係に注目した方法について考える．この系統の手法における先行研究では名詞の共出現関係と頻度を用いるものや名詞と形容詞の係り受け関係と相互情報量的尺度を用いるもの[5]などが存在していた．一方，我々は動詞とそれに対して特定の表層格（ヲ格）の関係にある名詞との関係から上下関係を抽出する方法を試みたのでその結果を報告する．

2．上下関係の抽出方法

上下関係を判定する基準を導くために次の仮定を置く．

- 上位語のすべての性質は下位語に継承されるがその逆は成立しない．
- 特定の格関係で名詞に結合する動詞はその名詞の性質を表している．

ここである名詞を A，また A とある格関係（ヲ格など）で結合する動詞の集合を V(A)とすると，上記の仮定により上位語 A とその下位語 B に関して

$$\text{A が B の上位語である} \Leftrightarrow V(A) \subset V(B) \quad (1)$$

が成立する．次に(1)を用いてコーパスから上下関係を抽出する手順を以下に説明する．

ステップ1：コーパスに含まれる任意の名詞 A に対して A とヲ格の関係で接続する動詞の集合 V(A)を生成する．

ステップ2：名詞 A に対して $|V(A) \cap V(Bi)| > a$(定数)となる名詞 Bi を選択しそれらの集合{Bi}を生成する．

ステップ3：Bi のそれぞれに対して

$$Ri = |V(A) \cap V(Bi)| / |V(Bi)| \quad (2)$$

を計算しこの Ri の降順で{Bi}をソートして名詞 A の下位語候補のランキングとする．

それぞれのステップについて説明する．まずステップ1ではコーパスを形態素解析，係り受け解析した結果をもとにして V(A)を生成する．今回，格関係はヲ格を用いた．ところが使用する実際のコーパスには十分な用例が含まれていないことが多いのでこのとき生成される V(A)は不完全な動詞集合となる．次にステップ2ではステップ1で得られる名詞 A のうち出現頻度が低いため十分な V(A)が構成できていないものを除外している．最後のステップ3では有限のコーパスを用いることによる V(A)，V(B)の不完全さを考慮して，上下関係の判定には(1)式に近い形式の(2)式による尺度を用いる．この(2)式は確率論的には条件付確率，またデータベースから連想規則を得るためによく用いられる信頼度とも同じ形式である．

3．実験

本手法を用いた名詞の上下関係の抽出について2種類の実験を行ったのでその方法と結果について述べる．

3．1 実験1

（1） 実験方法

実験1では与えられた名詞に対してその下位語候補を本手法によって計算し，そのランキングの上位 10 件を抽出した．以下に実験で使用したコーパス，ツールを示す．

コーパス：日本経済新聞を用い，名詞は 1996 年 1 月の 1 ヶ月分の記事の中から抽出し，動詞は 1996 年-2000 年の 5 年分の記事の中に含まれるものを用いた．

形態素解析器：chasen[1]を用いて分かち書き，品詞タグ付けを行った．

係り受け解析器：CaboCha[2]を用いて形態素解析結果から動詞とヲ格で接続する名詞を抽出した．

（2） 実験結果

ここでは例として名詞「商品」「言葉」「活動」に対する下位語候補のランキング上位 10 件を表 1 に示す．

順位	名詞：**商品**	名詞：**言葉**	名詞：**活動**
1	名産	故事	緑化
2	文具	感触	預託
3	中古	便り	遍歴
4	蓄音機	敬語	変転
5	単品	苦言	採掘
6	税目	知らせ	合成
7	詰め合わせ	言い訳	巡回
8	オーブン	くだり	節電
9	アフターケア	面白み	散布
10	ふろしき	逃げ場	検索

表 1:「商品」「言葉」「活動」に対する下位語候補

†総合研究大学院大学
††国立情報学研究所

ここでは3つの名詞を取り上げている．これらは次の観点で選択した．名詞の大きな分類に「物」と「事」がある．前者に属する代表例として「商品」，後者の例として「活動」，また双方に属さない例として「言葉」を選んだ．このとき大分類項目が異なる「商品」と「活動」とではその下位語にも共通部分するものがないはずである．実験によって得られたその下位語候補を見ると「商品」の下位語候補は「税目」を除いてすべて商品の一種であり下位概念となっている．同様に「活動」の下位語候補もすべて活動の一種である動詞由来の名詞となっていることがわかる．また活動の下位語候補のどれも「商品」の下位語になっていないことも分かる．そしてその逆もまた同様である．「言葉」の下位語候補に関しては，面白み，逃げ場のような不適切な候補も含まれるが概ね期待した効果が認められた．

3．2　実験2

実験2は既存の代表的シソーラスである分類語彙表[3]の体言の部（名詞）から上位語と下位語の組をサンプリングし，この語の組が本手法でも抽出できるかどうか確認するための実験である．

（1）　実験方法

分類語彙表は4階層からなるシソーラスである．その中の名詞部分に相当する体言の部ではそれぞれの階層に含まれる見出し語の異なり数が5語，43語，887語，54756語となっている．さらにこの中でコーパス中にも含まれる名詞数は7766語であった．この中から上位語と下位語の組を無作為にサンプリングし，第2階層中の名詞と第4階層中の名詞の組，第3階層中の名詞と第4階層中の名詞の組，それぞれ10組ずつを抽出した．次にこれらのサンプル中の上位語に対して本手法で下位語候補を上位200語まで抽出し，それらの候補にサンプル中の下位語が含まれているかを検証した．なおコーパスや形態素解析器などの実験条件は実験1と同様である．

（2）　実験結果

実験結果を表2に示す．本手法による分類語彙表中の上下関係サンプルにおける下位語の抽出は全20サンプルの中，19サンプルにおいて成功していることが確認できた．このことは分類語彙表の上下関係と本手法が抽出する上下関係が結果的に概ね適合していることを示している．

第2階層中の語と第4階層中の語	抽出結果	第3階層中の語と第4階層中の語	抽出結果
関係、きずな	○	力、実力	○
停止、打ち切り	○	時間、納期	○
場所、住所	○	形、形式	○
全体、全社	○	社会、世の中	○
団体、農協	○	心、胸中	○
解決、決着	○	言語、日本語	○
会議、討論	○	生活、暮らし	○
集会、催し	○	経済、損益	×
売買、転売	○	資材、燃料	○
料理、弁当	○	道具、おもちゃ	○

表2:分類語彙表に含まれる上下関係のサンプルとその抽出結果（○：抽出、×：抽出できず）

しかし分類語彙表における上下関係はその中で明確に定義されていないので本手法の意図する上下関係抽出との違いを陽に比較することはできない．中には本手法の仮定に適合しないような例も分類語彙表中には見られる．例えば「物質→天気」という上下関係がそれに該当する．この場合，「物質」から「天気」に継承されないような動詞の用例が存在する（“物質を分解する”と“*天気を分解する”など）．この点は分類語彙表以外のシソーラスとの比較を踏まえたうえで，手法の評価に関する課題として今後検討していきたい．

4．まとめ

本稿ではコーパスから上下関係にある名詞の組を自動的に抽出する方法を提案した．また本手法による上下関係抽出と分類語彙表に含まれる上下関係との比較実験も行いサンプル評価で95%の適合度を確認した．しかし分類語彙表中には本手法の仮定とは異なる上下関係が含まれていることも確認した．今後，評価方法の検討，その他の上下関係抽出手法との比較やその他のシソーラスとの比較実験など行う予定である．

参考文献

[1]　松本裕治，北内啓，山下達雄，平野善隆，今一修，今村友明: “日本語形態素解析システム「茶筌」使用説明書 Version1.5.1”, (1997)

[2]　工藤拓，松本裕治: “チャンキングの段階適用による係り受け解析”, 情報処理学会論文誌, Vol. 43, No. 6, (2002).

[3]　国立国語研究所編: “分類語彙表”，大日本図書, (2004).

[4]　新里圭司，鳥澤健太郎: “HTML 文書からの単語間の上位下位関係の自動獲得”，自然言語処理, Vol. 12, No. 1, (2005)

[5]　山本英子，神崎享子，井佐原均: “共起語の包含関係に基づく語彙の階層化への頻度情報の影響”，言語処理学会第11回年次大会予稿集, (2005)