定型表現法と分布類似度を用いた特許データベースからの用語の上位、下位関係の抽出

Automatic hyponymy extraction from a patent database using pattern method and distributional similarity

広島市立大学大学院　情報科学研究科准教授
難波　英嗣

2001年北陸先端科学技術大学院大学情報科学研究科博士後期課程修了。博士(情報科学)。東京工業大学精密工学研究所助手等を経て、2010年より広島市立大学大学院情報科学研究科准教授。自然言語処理、テキストマイニングの研究に従事。

nanba@hiroshima-cu.ac.jp　TEL 082-830-1584

1　はじめに

オントロジーは複数の要素や概念の関係性を体系化したもので、文献を検索したり高度な言語処理を行ったりするための有用な情報源となっている。しかしながら、オントロジーの人手での構築は、一般に非常にコストがかかるため、テキストデータベースからオントロジーを自動的に構築する様々な手法が提案されてきた。

そのひとつに、「A や B などの C」や“A such as B, C”などの定型表現に着目して、用語の上位、下位関係を自動的に抽出するものがある [Hearst 1992]。この手法を、本稿では定型表現法と呼ぶことにする。定型表現法は、数多くの上位、下位関係の用語対を収集できる一方で、上位、下位関係にない用語対も誤って抽出してしまうという問題があった。この問題に対し、本研究では、定型表現法で収集された名詞句対の中から上位、下位関係にあるものを選定する手法を提案する。

2　用語の上位、下位関係の抽出

本稿では、1章で述べた問題について、分布類似度 [Lin 1998, Lee 1999] と呼ばれる尺度を用いて改善する手法を提案する。2.1節では、まず、定型表現法で上位、下位関係の用語対を収集しようとすると、どのような問題が生じるのかについて述べる。次に、2.2節で分布類似度を用いた関連語抽出手法について、2.3節では、分布類似度を用いた上位、下位関係抽出の改善手法について、それぞれ述べる。

2.1　定型表現法の問題点

今、以下の2文について考える。

（文1）「パソコンなどのOA機器」
（文2）「パソコンなどのキーボード」

この2文に定型表現法を適用すると、文1から「OA機器」―「パソコン」が、文2から「キーボード」―「パソコン」が、それぞれ上位、下位関係として抽出される。ここで、「OA機器」―「パソコン」は上位-下位関係として正しいが、「キーボード」―「パソコン」は正しくない。後者は、一般に部分、全体関係と呼ばれるが、「自動車等のエンジン」や「飛行機等の翼」など、定型表現法を用いると、部分、全体関係を上位、下位関係と誤認識するケースが少なからず存在する。本節では、このような誤認識を軽減する手法を提案する。

2.2　分布類似度を用いた関連語抽出

分布類似度とは、「用語Aと用語Bの意味が似ていれば、AとBの特定範囲に出現する共起語（特定のある語とともに頻出する語）も類似する」という仮定に基づき、共起語のベクトルで各用語を特徴付け、共起語ベクトルどうしの類似度で語の類似性を測る尺度である。

共起語ベクトルの作成には、単純に同一文内で共起する語を用いる方法と、文内で意味の上で結びついた係り受け関係にある語のみを用いる方法の2通り存在する。相澤は、新聞記事を対象にした調査で、後者の方が前者よりも精度良く関連語を収集できることを示している

[相澤：2008]。

ここでは、相澤と同様の手法により関連語を収集した結果について述べる。図 1 は、公開公報（1993～2016 年）を用いて作成した「パソコン」の共起語ベクトルである。このベクトルは、まず、日本語係り受け解析器 CaboCha を用いて、公開公報の明細書に含まれる 985,556,899 文を構文解析し、次に、各文中で直接係り受け関係にある「名詞-表層格と動詞」の対を抽出し、名詞ごとに tf*idf 法を用いて表層格と動詞を重み付けて作成されている。

重み	係り受け関係にある動詞
0.00520775792586484	を_使い慣れる
0.00469443553518948	に_取り込める
0.00436481024477683	を_買い換える
0.00434722420136261	を_持ち歩く
0.00429461761609426	に_とりこむ
0.00424172826128001	を_使いこなす
0.0041798969458932	を_携える
0.00415226314635009	を_使える
0.00398313521124071	で_扱える
0.00397020018583619	に_取りこむ
0.00392873472668023	を_持ち込む

図 1　「パソコン」の共起語ベクトルの一部

この共起語ベクトルと類似度の高い用語を収集した結果が図 2 である。図から、「パソコン」と類似する様々な機器の名前が収集できていることがわかる。このように、共起語ベクトルでは、共通の性質を持つ関連語が収集される傾向にある。

類似度	関連語
14.9394897027747	パソコン 1
14.4046855836949	パソコン 2
14.1539360812989	パソコン 10
14.1497330630177	コンピュータ 2
13.9830894793089	パソコン上
13.9462757374554	ワープロ
13.9094509648393	ワードプロセッサ
13.8846924586199	PC20
13.8822168194158	コンピュータ機器
13.750711445519	ノート型パソコン
13.7010977506601	計算機 1
…	
12.3596639985012	OA 機器

図 2　分布類似度を用いて収集された「パソコン」の関連語

2.3　分布類似度を用いた上位、下位関係抽出の改善手法

2.1 節で述べた 2 文を例に、分布類似度を用いて上位、下位関係にあるものとそうでないものを分離する手法について述べる。

「パソコン」の関連語を、分布類似度を用いて収集すると、「OA 機器」は収集結果の上位に出現する一方で、「キーボード」は出現しないことが期待できる。なぜならば、図 1 に示す「パソコン」のリストに出現する共起語の多くは「OA 機器」の共起語リストに含まれていても不自然ではないが、「キーボード」に含まれるのは不自然だからである。実際に、「キーボード」の共起語ベクトルの一部を図 3 に示す。

重み	係り受け関係にある動詞
0.00520410336117679	を_使い慣れる
0.00500417739891946	を_たたく
0.00476489998914416	を_立掛ける
0.00438619246937393	に_慣れ親しむ
0.00423463445265096	を_立てかける
0.00404156203197133	を_立て掛ける
0.00394834591570089	に_慣れる
0.00380550467639943	を_打てる
0.00379658953262475	で_打ち込む
0.00366171120248995	を_折りたたむ
0.00345465888178231	に_こぼす

図 3　「キーボード」の共起語ベクトルの一部

図 3 と図 1 を比べると、共起語の一つ目の「を_使い慣れる」は共通するものの、他に一致する共起語はない。このように、上位、下位関係にある用語対は、何らかの共通の性質を持つはずであり、その一側面が、係り受け関係にある動詞という形で表れていると考えることができる。

3　実験

2.3 節で述べた手法の有効性を確認するため、実験を行った。1993～2016 年の公開公報の明細書に含まれる 985,556,899 文のうち、「等の」または「などの」を含む文をすべて抽出し、これらの文を、CaboCha を用いて係り受け解析した後、上位、下位関係の候補となる名詞句対を抽出した。これらの候補に対し、2.2 節で提案した分布類似度を用いた手法を用い、上位、下位関

係にあるものを選定した。

実験の結果、提案手法は、計 14,167,501 件の候補のうち、6,418,750 件を上位、下位関係と判定した。上位、下位関係と判定したもの、上位、下位関係でないと判定したものの一部を図 4 および図 5 にそれぞれ示す。なお、網掛けになっている個所は、筆者が上位、下位関係ではないと判断したものである。

頻度	上位	下位
70391	ネットワーク	インターネット
39957	車両	自動車
34033	金属	アルミニウム
27927	影響	ノイズ
27216	画像形成装置	プリンタ
23909	発生	停電
22268	記憶装置	ハードディスク
20956	弾性体	ゴム
20221	液体	水

図 4　提案手法により上位、下位関係と判定された名詞句対の例

頻度	上位	下位
18986	ポインティングデバイス	マウス
16260	不活性ガス	窒素
13112	不活性ガス	アルゴン
11900	不活性ガス	窒素ガス
11664	理由	こと
11290	情報処理装置	パーソナルコンピュータ
10840	値	I/O
9072	問題	こと
8003	利点	こと

図 5　提案手法により上位、下位関係でないと判定された名詞句対の例

図 4、5 の結果より、本来上位、下位関係と判断されるべきものが上位、下位関係でないと判断されたケースが存在しているが、上位、下位関係であると判断されたものの大部分は実際に上位、下位関係にあったため、提案手法は上位、下位関係の判別に一定の効果があると考えられる。

4 おわりに

本稿では、定型表現法と分布類似度を用いた用語の上位、下位関係の抽出手法を提案した。

参考文献

[Hearst 1992] Hearst, M.A., “Automatic Acquisition of Hyponyms from Large Text Corpora,” Proceedings of COLING, pp.539-545, 1992.

[Lee 1999] Lee, L., “Measures of Distributional Similarity,” Proceedings of ACL, pp.25-32, 1999.

[Lin 1998] Lin, D., “Automatic Retrieval and Clustering of Similar Words,” Proceedings of COLING/ACL, pp. 768-774, 1998.

[相澤 2008] 相澤彰子、“大規模テキストコーパスを用いた語の類似度計算に関する考察,” 情報処理学会論文誌 , Vol.49, No.3, pp. 1426-1436, 2008.