HTMLタグの繰り返しパターンに注目した知識の自動獲得

新里 圭司 (210044)

北陸先端科学技術大学院大学 情報科学研究科

2003年2月13日

キーワード: 知識の自動獲得，統計的自然言語処理，上位語，下位語，World Wide Web.

近年，膨大な量の文書が計算機で扱えるようになり，多種多様な自然言語処理技術が利用されるようになってきた．しかし，より知的で高度な処理を行うためには，単語間の上位下位関係（*hyponymy relation*），類似関係（*synonymy relation*），包含関係（*part-whole relation*）などの知識がまだまだ不足しており，このような知識の獲得は今後ますます重要なものになるといえる．そこで本稿では，WWW上に大量に存在するHTML文書から広範な単語間の上位下位関係を自動的に獲得する手法について提案する．WordNetに代表されるような大規模なシソーラスを自動生成するという目的のもと，従来より単語間の意味的関係の自動獲得に関する研究は盛んに行われてきた[3, 4, 1, 7, 5, 2, 6]．しかし，そのほとんどはHearst[3]が用いた“ *such as* パターン ”に代表される，構文パターン（*lexico–syntactic pattern*）のマッチングによりコーパス中から獲得するものであった．しかし，(a）単語間の意味的な関係を表す構文パターンがコーパス中に頻繁に現れることは稀であり，また（b）たとえ大量のテキストを持ってきたとしても，構文パターンに現れない単語や句が大量に存在するため，従来手法では大量かつ幅広い単語間の上位下位関係を獲得することが難しいという問題があった．そのため，本研究では構文パターン以外の上位下位関係の特性を捕らえる手がかりを用いることで獲得を試みる．具体的には，(1）HTMLタグにより与えられる文書の構造，(2）情報検索などの分野で用いられるdfやidfなどの統計量，(3）大量の新聞記事から収集した名詞と動詞の係り受け関係，(4）予備実験により得られたヒューリスティクなルール，の4つの異なる要素を組み合わせることで上位下位関係の獲得を試みる．

本研究では，単語間の上位下位関係を獲得するにあたり，以下に示す3つ仮説をたてる．

仮説1 HTML文書中で同じパスを持つ表現同士は意味的に類似しており，共通の上位語を持ちやすい

仮説2 共通の上位語を持つような下位語の集合が与えられた時，それらに共通な上位語は各下位語を（少なくとも1つ）含む文書に現れやすく，それ以外の文書には現れにくい

仮説 3 上位語と下位語は意味的に類似しており，その類似性は上位語と下位語の持つ係り受け関係によって捕らえることができる

本研究で提案する上位下位関係の獲得方法は，上に挙げた仮説を考慮した次の 4 つのステップからなる．まずステップ 1 では，仮説 1 に基づき，WWW より大量に収集した HTML 文書中から同じパスを持つ表現同士を獲得する．以下では，このステップ 1 で獲得された同じパスを持つ表現の集合のことを**下位語候補集合**と呼ぶ．

続いてステップ 2 では，ステップ 1 で獲得された下位語候補集合の各要素に共通する上位語の獲得を試みる．上位語獲得に伴い，2 つ文書集合を準備する．1 つ目の文書集合は，ステップ 1 で得られた下位語候補集合の各要素（下位語候補）を検索語として検索エンジンより収集した HTML 文書集合からなるもので，これを**局所的文書集合**と呼ぶ．2 つ目の文書集合は，WWW より収集した HTML 文書 100 万件からなる文書集合で，これを**大域的文書集合**と呼ぶ．ステップ 2 では，局所的文書集合中に含まれる名詞のスコアを，仮説 2 に基づき，局所的文書集合中での文書頻度と，大域的文書集合中での文書頻度の両方を用いて計算し，スコアの最も高かった名詞を与えられた下位語候補集合に対する上位語として獲得する．

次にステップ 3 では，仮説 3 に従い，ステップ 1 で獲得された下位語候補集合とステップ 2 で獲得された上位語の組を，両者の類似度に基づきソートする．その結果，上位 N 組を後述するステップ 4 を適用後に出力とすることで，より尤もらしい上位語と下位語候補集合の組だけを獲得することが期待できる．類似度を計算するため，下位語候補集合全体の係り受け関係を局所的文書集合から，上位語の係り受け関係を大量の新聞記事よりそれぞれ求めた．そして，両者の係り受け関係をベクトルで表現することで，その類似度をコサイン尺度を用いて計算した．コサイン尺度とは，文書検索において文書間の類似度を求める際によく利用されている尺度である．

そして最後にステップ 4 として，予備実験の結果得られた 3 つのヒューリスティックなルールを適用することで，獲得された上位語と下位語候補集合の組を精錬することを行う．

実際に WWW より収集してきた約 87 万件の HTML 文書から，下位語の集合（の候補）を約 9 万個獲得することができた．そして，その中からランダムに抽出した集合 2,000 個について評価したところ，その中に含まれる約 14,000 個の順序付けられた上位下位関係のうち，全体の約 3.6 %にあたる上位 501 個については 85 %，全体の約 5 %にあたる上位 700 個の上位下位関係については 75 %，約 10 %にあたる 1,400 個については 60 %程度の精度で正しい上位下位関係を獲得することができた．また，従来手法と比較実験を行い，文書量が少なすぎて従来手法では獲得することができないような上位下位関係を，提案手法では同量の文書から獲得できることを確認した．これにより，少量の文書集合を対象に上位下位関係の自動獲得を試みる場合，本研究で提案する手法が有効であることがわかった．

参考文献

[1] Sharon A. Caraballo. Automatic construction of a hypernym-labeled noun hierarchy from text. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pp. 120–126, 1999.

[2] Michael Fleischman, Eduard Hovy, and Abdessamad Echihabi. Offline strategies for online question answering: Answering questions before they are asked. In *Proceedings of the 41th Annual Meeting of the Association for Computational Linguistics*, pp. 1–7, 2003.

[3] Marti A. Hearst. Automatic acquistition of hyponyms from large text corpora. Technical Report S2K–92–09, 1992.

[4] Marti A. Hearst. Automated discovery of wordnet relations. In Christiane Fellbaum, editor, *WordNet: an electronic lexical database*, chapter 5, pp. 131–151. MIT Press, 1998.

[5] Emmanuel Morin and Christian Jacquemin. Automatic acquisition and expansion of hypernym links. In *Computer and the Humanities 2003*, 2003. forthcoming.

[6] 安藤まや, 関根聡, 石崎俊. 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. 情報処理学会 研究報告 2003-NL-157, pp. 77–82, 2003.

[7] 今角恭祐. 並列名詞句と同格表現に着目した上位下位関係の自動獲得. 九州工業大学修士論文, 2001.