語の共起頻度とヒューリスティックスを用いたWebからの上位語の獲得

新里圭司　鳥澤健太郎
北陸先端科学技術大学院大学 情報科学研究科
{skeiji, torisawa}@jaist.ac.jp

1　はじめに

従来より提案されてきた単語間の上位下位関係の獲得手法は，名詞句の並置など文の表層的なパターンを利用するものであり，大量の上位下位関係の獲得が難しいという問題があった．我々は，既に従来手法とは全く異なるアプローチによる獲得手法として，HTML文書中のタグ情報，*df·idf* による名詞のスコア，名詞と動詞の係り受け関係の3種の情報を利用する手法を提案したが [5]，幾つかのヒューリスティックスを新たに導入することで，精度のさらなる向上が図れることがわかった．本稿では，それらのルールに焦点をあてるとともに，従来手法との比較実験についても報告する．

2　関連研究

Hearst[1] はコーパス中から単語間の上位下位関係を自動獲得する手法を提案している．Hearst は，“*NP* {,*NP*}*{,} *or other NP*” のような単語間の上位下位関係を明示する構文パターンを幾つか発見した．そして，“*or other* パターン” を対象に評価実験を行った結果，約52％の精度で妥当な上位下位関係が獲得されたと報告している．このような構文パターンを日本語の新聞記事に対して適用し，単語間の上位下位関係の獲得を試みた研究として今角 [4]，安藤ら [3] がある．今角は構文解析の結果より得られる同格・並列表現を含む文に対し，3種類の構文パターンを適用し上位下位関係の獲得を行っている．新聞記事4年分に対し実験を行ったところ約15,000件の上位語下位語対が獲得でき，そのうちの600件について人手で評価を行い，精度は77.2％であったと報告している．また安藤らは，大規模なシソーラスを自動的に生成するための準備として，連想概念辞書に登録されている日常性の高い約60語について，新聞記事からその下位語の獲得を行っている．構文解析済みの新聞記事6年分に対し，7種類の構文パターンを適用することで，いずれのパターンについても60％から85％程度の精度で上位下位関係が獲得できたと報告している．

3　提案手法

本研究では，以下に示す3つの仮説をたて，単語間の上位下位関係の獲得に用いている．

仮説 1 HTML文書中に現れる箇条書きやリストボックス，テーブルのセルなどの要素は，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下位語の集合が与えられた時，各下位語に共通する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比較的現れにくい

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捉えることができる

そして，上の仮説に基づいた以下に示す4つのステップを経ることで単語間の上位下位関係の自動獲得を行う．ここに挙げたステップ1，2，3は上の仮説1，2，3とそれぞれ対応している．

ステップ 1 HTML文書中のタグ情報に基づいた下位語候補集合の獲得

ステップ 2 *df·idf* 等の統計量に基づく上位語候補獲得

ステップ 3 上位語候補と下位語候補間の意味的類似度に基づく上位語候補と下位語候補集合の並べ替え

ステップ 4 ヒューリスティックなルールを用いた上位語候補と下位語候補集合の組の取捨選択

ここでステップ4は，上位下位関係獲得の精度を改善するために新しく導入したステップであり，ステップ3までで獲得された上位下位関係をヒューリスティックなルールを用いて修正，または削除するステップである．以下，各ステップについて述べる．

3.1　下位語候補集合の獲得 (ステップ 1)

ステップ1では，仮説1に従いHTML文書中に現れる各表現の持つ**パス**に注目することで，下位語候補集合の獲得を試みる．ここでパスとは，HTML文書中の表現を囲んでいるタグをそのネストの順序に従って，リスト形式で表したものである．例えば，

```
<LI>今月のお買得！</LI> <UL><LI>DVD-RW</LI>
<LI>ハードディスク</LI> <LI>プリンタ</LI></UL>
```

のようなHTML文書中の各表現は以下のようなパスを持っている．

(LI，今月のお買得！)
(UL，LI，DVD–RW)
(UL，LI，ハードディスク)
(UL，LI，プリンタ)

ステップ1では，HTML文書中に現れる同じパスを持つ表現同士をまとめることで，意味的に類似した共通の上位語を持つであろう表現の集合を獲得する．先程例に挙げたHTML文書からは，

{DVD–RW，ハードディスク，プリンタ }
といった集合が獲得される．本研究では，この獲得された集合のことを**下位語候補集合**と呼ぶ．

3.2 *df*，*idf*に基づく上位語候補の獲得 (ステップ 2)

ステップ 2 では，ステップ 1 より獲得された下位語候補集合に対する上位語候補の獲得を行う．これに伴い，ステップ 2 ではまず 2 つの文書集合を準備する．1 つ目は，大量の HTML 文書集合の中から無作為に選んだ HTML 文書からなるもので，これを**大域的文書集合**と呼ぶ．この文書集合は一般的な文脈においての単語の文書頻度を求める際に使用する．2 つ目は，ステップ 1 で獲得された下位語候補集合の各要素を 1 つでも含む文書を，既存のサーチエンジンより収集し作成するもので，**局所的文書集合**と呼ぶ．この文書集合は与えられた下位語候補集合の各要素と，ステップ 2 で獲得する上位語候補の関連の強さを測る際に用いる．

以下では，下位語候補集合を C，大域的文書集合を G，C について生成した局所的文書集合を $LD(C)$ と記述する．また，$LD(C)$ に含まれる名詞の中から，普通名詞，サ変名詞，地名を表す名詞を抽出して得られる名詞の集合を N とする[1]．ステップ 2 では，仮説 2 に基づき上位語候補 $h(C)$ を以下の式により求める．

$$h(C) = argmax_{n \in N}\{df(n, LD(C)) \cdot idf(n, G)\}$$

$$idf(n, G) = \log \frac{|G|}{df(n, G)}$$

ここで $df(n, D)$ は，文書集合 D 中で名詞 n を含む文書数を返す関数であり，$|G|$ は文書集合 G に含まれる文書数を表す．上式より，局所的文書集合中の多くの文書に現れ，かつ大域的文書集合中の文書には相対的にあまり現れない名詞が上位語候補として獲得される．

3.3 意味的類似度に基づく上位語候補と下位語候補集合の並べ替え (ステップ 3)

ステップ 3 では，仮説 3 に基づき上位語候補と下位語候補の持つ係り受け関係から，両者の意味的類似度を計算し，その類似度に基づいてステップ 2 により獲得された上位語候補と下位語候補集合の組をソートする．これにより，後述するステップ 4 を適用後，その上位 N 組を最終的な出力とすることで，相対的に高い精度で上位下位関係を獲得することが可能になる．下位語候補集合 C の要素のいずれかが助詞 p を介して動詞 v に係る頻度を $f_{hypo}(C, p, v)$，全ての助詞を $\{p_1, \cdots, p_l\}$，全ての動詞を $\{v_1, \cdots, v_m\}$ で表したとき，C の係り受け関係を表したベクトル（以降，**係り受けベクトル**と呼ぶ）を次のように定義する．

$$hypov(C) = \langle f_{hypo}(C, p_1, v_1), f_{hypo}(C, p_2, v_1), \\ \ldots, f_{hypo}(C, p_{l-1}, v_m), f_{hypo}(C, p_l, v_m)\rangle$$

同様に上位語候補 $h(C)$ の係り受けベクトルを，

$$hyperv(h(C)) = \langle f(h(C), p_1, v_1), f(h(C), p_2, v_1), \\ \ldots, f(h(C), p_{l-1}, v_m), f(h(C), p_l, v_m)\rangle$$

のように定義する．ここで $f(h(C), p, v)$ は，新聞記事 33 年分[2]より求めた，上位語候補 $h(C)$ が助詞 p を介して動詞 v に係る頻度を表している．下位語候補集合 C 中の全ての要素と上位語候補の意味的な類似度は，コサイン尺度を用いて次のように計算する．

$$sim(h(C), C) = \frac{hypov(C) \cdot hyperv(h(C))}{|hypov(C)| \times |hyperv(h(C))|}$$

この時，新聞記事に 500 回以上現れない語に関しては，正しい係り受け関係が得られていないと判断し，そのような語が上位語候補として獲得された組に関しては，類似度を 0 とした．ステップ 3 では，上位語候補と下位語候補集合の各組を以下の値に基づきソートする．

$$sim(h(C), C) \cdot df(h(C), LD(C)) \cdot idf(h(C), G)$$

3.4 ヒューリスティックスを用いた上位語候補と下位語候補集合の組の取捨選択 (ステップ 4)

ステップ 4 では，ステップ 3 までで得られた結果に対し，以下の 3 つのヒューリスティックなルールを適用して上位下位関係獲得の精度向上を図る．これらのルールは予備実験で得られた知見を基に作成されている．

ルール 1 獲得された上位語候補を検索語として検索エンジンに問い合わせ，その結果得られるヒット件数が，各下位語候補を検索語として得られたヒット件数の総和よりも少ない場合，その上位語候補と下位語候補集合の組を削除する

ルール 2 獲得された上位語候補が，下位語候補集合のいずれかの要素の部分文字列として現れていた場合，以下の条件に当てはまるような上位語候補と下位語候補集合の組は削除する

- 上位語候補が下位語候補の末尾以外の場所で部分文字列として現れている
- 下位語候補集合の半分以上の要素について上位語候補が末尾に現れていない

ルール 3 獲得された上位語候補が地名を表す語である場合，上位語候補を「地名」に変更する

ルール 1 では，誤って獲得された上位語候補を持つ組を削除し，精度の改善を図る．一般に，上位語は下位語と比べより広い文脈で使われており，下位語候補より上位語候補を含む文書の方が WWW 上により多く存在するはずと考えられる．このような上位語が持つと考えられる特性を利用したのがルール 1 である．

次いでルール 2 では，誤った上位語候補が獲得されている組及び，意味的類似性の見られない下位語候補集合を持つ組を削除することで精度の改善を図る．しかし，このルールは獲得された上位語候補が下位語候補の部分文字列として現れない場合には適用されない．日本語では，複合名詞の主辞は主として末尾に現れる

[1] 実際は，予め用意した 116 個の不要語を取り除いたものを N としている．

[2] 読売新聞 1987–2001，毎日新聞 1991–1999，日経新聞 1990–1998; 計 3.01GB

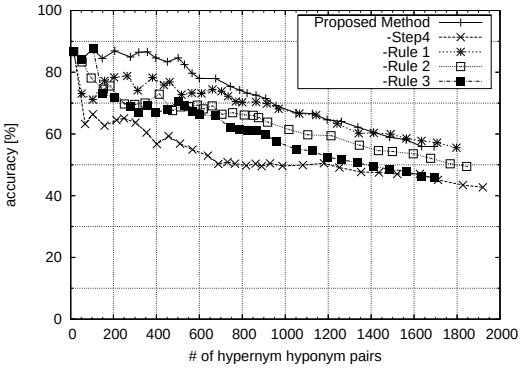

図 1: 新たに導入したルールの効果

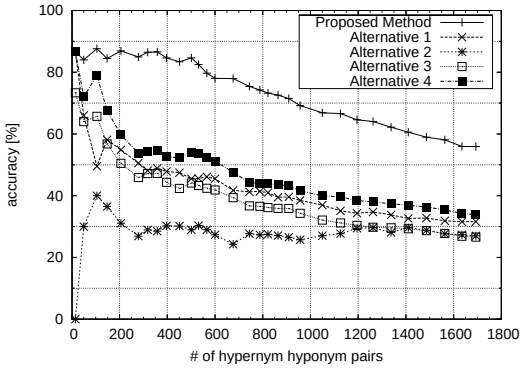

図 2: 提案手法と他の手法の比較

名詞であるため，下位語候補集合の多くの要素で共通の語が末尾に現れている場合，その語は妥当な上位語である可能性が高いと考えられる．それに対し，獲得された上位語候補が下位語候補の末尾以外の場所に現れる場合，その上位語候補が妥当である可能性は低いと考えられる．また，一部の下位語候補の末尾にだけ上位語候補が現れる場合，その下位語候補集合には意味的な共通性が見られ難く，共通の上位語も持ち難いと考えられる．そのため，このような条件に該当する上位語候補や下位語候補集合を持つ組は削除する．

最後にルール 3 では，獲得された誤った上位語候補を正しい上位語に置換することで精度の改善を図る．予備実験において，下位語候補集合の要素が地名の場合，それら地名を含む地域を指す地名が，上位語候補として獲得される場合が頻繁に見られた．例えば，下位語候補集合が「石川」,「富山」,「福井」という表現からなっていた場合，妥当な上位語としては「地域」などが考えられるが，ステップ 3 までで述べた方法で上位語候補を求めると,「日本」という結果が得られる．実際に獲得された上位語候補「日本」は，本研究でたてた仮説を満足するが,「石川」,「富山」,「福井」に対しては包含関係を表す語であり，上位語ではない．そこで，地名からなる下位語候補集合に対しても，正しい上位語を獲得できるようにするため，獲得された上位語候補が地名を表す語であった場合は「地名」に置換するようにした．

4 提案手法の評価実験

本研究では，約 4.66×10^6 件の HTML 文書（重複無し）を WWW より収集した．そして，その中から 1.00×10^6 件の HTML 文書 (約 1.26GB，タグ無し) を無作為に選び，大域的文書集合を作成した．大域的文書集合から単語の一般的な文書頻度を求めるため，タグを除去し，JUMAN を用いて形態素解析を行った．次に，先程の 4.66×10^6 件の文書集合より，約 8.71×10^5 件の HTML 文書 (10.4GB，タグ有り) を選び，3.1 節で述べた方法により，9.02×10^4 個の下位語候補集合（重複有り，全部で 6.01×10^5 個の下位語候補を含んでいる）を獲得した．続いて，この 9.02×10^4 個の下位語候補集合の中から重複を除き，無作為に選択した 2,000 個を評価実験用のデータとした．この 2,000 個の下位語候補集合には，13,790 個の下位語候補が含まれている．最後に，個々の下位語候補を検索語として検索エンジン goo より検索し，その結果得られた上位 100 件を収集して局所的文書集合を作成した．作成した局所的文書集合は，タグを除去し，JUMAN を用いて形態素解析を行い，さらに係り受けベクトルを生成するため，既存の構文解析器を用いて係り受け解析を行った．

実験の結果，各ステップはそれぞれ性能の向上に貢献していることが確認されている．各ステップの貢献に関するより詳細な実験結果及び，*df·idf* スコアの妥当性については，論文 [2] を参照されたい．また，本研究では評価用に用いた 2,000 組の上位語候補と下位語候補集合のうち，全体の 1 割にあたる上位 200 組を最終的に獲得された上位下位関係として評価対象にしている．残りの 1,800 組については，間違った上位語候補が獲得されやすいという観点から評価対象から外した．

新たに導入したステップ 4 の各ルールがどの程度精度の向上に貢献しているのかを確認するため，それぞれのルールを抜いた時の上位下位関係獲得の性能を評価した．その実験結果を図 1 に示す．図中の各グラフは次の式に従って描画されている．

$$\langle \sum_{k=1}^{j} |C_k|, \frac{\sum_{k=1}^{j} correct(C_k, h(C_k))}{\sum_{k=1}^{j} |C_k|} \rangle$$

ここで j は $1 \leq j \leq 200$ であり，$|C_k|$ は下位語候補集合 C_k の要素数である．さらに，$correct(C_k, h(C_k))$ は下位語候補集合 C_k 中で獲得された上位語候補 $h(C_k)$ が正しいとされる下位語候補の数を表している．図 1 において “– Rule X” となっているものは，今回提案した手法から “ルール X” を抜いた時の精度を，“–Step 4” は全ルールを抜いた時の精度をそれぞれ表している．図 1 より，どのルールを抜いた場合についても，上位下位関係獲得の精度が低下しており，各ルールが精度の向上に有効に働いていることがわかる．中でもルール 3 を抜いた時が最も精度が落ちていることから，提案手法により獲得された上位下位関係の中には地名に関するものが多く含まれていると考えられる．実際，提案手法により獲得された正しい上位下位関係数は 947 個であるのに対し，ルール 3 を抜いた場合では 779 個に減少していることから，提案手法により獲得された正しい上位下位関係のうち，およそ 17.7％は地名に関する上位下位関係で占められていることがわかる．表 1 に実際に獲得された上位下位関係の例を示す．

表 1: 提案手法により獲得された上位下位関係の例

下位語候補集合	上位語
殺人*, 放火*, 強姦*, 侵入盗*, 侵入強盗*, 非侵入盗*, 非侵入強盗*	犯罪
将軍*, 宮本武蔵*, 七人の侍*, ミッドウェイ*, 無法松の一生*, 太平洋の地獄*, 羅生門* 価値ある男*, 用心棒』『赤ひげ, 大統領の堕ちた日*, 武士道ブレード*	映画
モスクワ*, キエフ*, タシケント*,, ミンスク*, トビリシ*, ドゥシャンベ*, ビシュケク*, アスタナ*, キシニョフ*, アシハバード*, エレバン*, バクー*	地名
踊る大捜査線*, プロジェクトX, 世紀を越えて, 彼女たちの時代*	ドラマ
桑田真澄*, 上原浩治*, ワズディン*, 武田一浩*, 木村龍治*, 真田裕貴*, 鄭ミン台*, 趙成ミン*	投手

“*” が後についている下位語候補は，提案手法により妥当な上位語が獲得されたものを示す．

5 他の手法との比較実験

本研究では，以下に示す 4 種類の手法を対象に比較実験を行った．

手法 1 複数の下位語候補の末尾で共有される最長の語を上位語として獲得する．

手法 2 下位語候補集合を獲得した HTML 文書中の箇条書きや表データの直上，またはさらにその上のキャプションから，上位語を人手により獲得する．

手法 3 先行研究で用いている構文パターンを基に作成した正規表現パターンを用いて上位語を獲得する．

手法 4 手法 1，2，3 を組み合せて上位語を獲得する．

手法 3 では，図 3 に示した正規表現パターンを用いて上位下位関係の獲得を行っているが，これらのパターンは構文解析結果ではなく，文の表層的な情報だけを利用しており，先行研究で用いている構文パターンとは若干異なる．しかし図 3 に示した正規表現パターンは，先行研究で用いられている構文パターンで獲得される上位下位関係を漏れなく獲得できるため，手法 3 により得られる上位下位関係獲得の精度が，構文パターンを用いて上位下位関係の獲得を試みる手法の上限であると言える．また，正規表現パターンを適用する文書として，ステップ 2 で上位語候補を獲得する際に用いた局所的文書集合中に含まれている文書を利用した．手法 3 では，提案手法により予め獲得された正しい上位下位関係を正規表現パターンに与え，そのパターンに適合する文があるかどうかを評価の対象としている．つまり，提案手法により獲得された正しい上位下位関係を，正規表現パターンを用いて獲得できるかどうかだけしか確認していない．

また手法 4 の評価は，提案手法により獲得した正しい上位下位関係のうちどれだけの関係を手法 4 で獲得できるか，という観点で行った．手法 4 により正しい上位下位関係が獲得できたかどうかの判定は，提案手法で獲得した正しい上位下位関係を，手法 1，2，3 のいずれかで獲得できれば，正しい上位下位関係が獲得できたとした．比較実験により得られる提案手法と手法 4 の精度の差は，提案手法で獲得できて，手法 1，2，3 では獲得できない正しい上位下位関係数の差を表す．

上位語「**下位語**」，**下位語** .* 以外の .* **上位語**
下位語 .* に似た .* **上位語**
下位語 .* と呼ばれる .* **上位語**
下位語 .* のような .* **上位語**
下位語 .* と (い | 言) う .* **上位語**
下位語 .* など (、 | の)? .* **上位語**
下位語 .* (ら | たち) .* **上位語**

上位語，下位語は「」や “” で囲まれていても構わない．

図 3: 比較実験に用いた正規表現パターン

図 2 に提案手法と手法 1，2，3，4 の上位下位関係の獲得精度を示す．図 2 では，各グラフとも提案手法と同じ方法で下位語候補集合をソートしており（つまり各手法とも，$sim(h(C), C) \cdot df(h(C), LD(C)) \cdot idf(h(C), G)$ の値を用いてソートしている），上位語を獲得する方法だけが異なっている．

この図から，少なくとも 1 つの下位語候補あたり最大で 100 件の文書を収集し作成した文書集合から上位下位関係を獲得する場合においては，手法 1，2，3，4 では獲得できないような関係を提案手法はかなりの数獲得できているということがわかる．

6 まとめと今後の課題

本稿では，我々が提案した構文パターンを用いずに上位下位関係を獲得する手法 [5] に対し，3 つのヒューリスティクスを新たに導入することで精度の改善が図れることを実験により示した．

今後の課題としては，獲得精度のさらなる向上が挙げられる．今回提案した手法は，既存の手法と組み合わせることが可能であるため，これにより精度の向上が見込めるのではないかと考えている．次に，複数の語からなる上位語を獲得できるようにしたいと考えている．提案手法では，DVD–RW，やハードディスクなどのパソコン周辺機器からなる下位語候補集合に対して「機器」としか上位語を求めることができないが,「周辺機器」,「パソコン周辺機器」という上位語が獲得できれば，他の自然言語アプリケーションにとって，より有用な情報となるのではないかと考えている．

参考文献

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

[2] Keiji Shinzato and Kentaro Torisawa. Acquiring hyponymy relations from web documents. In *Proceedings of HLT/NAACL'04*, 2004. to appear.

[3] 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.

[4] 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.

[5] 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. 情報処理学会 研究報告 2003–NL–158, pp. 95–102, 2003.