# 統語構造の特徴を利用した上位下位関係辞書の拡張

山田一郎[*]　鳥澤健太郎[*]　風間淳一[*]　黒田航[*]
村田真樹[*]　Francis Bond[*]　隅田飛鳥[†‡]

[*]情報通信研究機構　[†]奈良先端大学院大学　[‡]北陸先端科学技術大学院大学
E-mail: iyamada@nict.go.jp

## 1. はじめに

近年、Web テキストや大規模テキストコーパスなど膨大な量のテキストデータを計算機上で扱えるようになった。このようなデータには、人間が持つ常識や一般知識などが多く含まれる。そこで我々は、Web テキストを対象として、計算機が扱える知識の一つとなる用語の上位下位関係を自動獲得する研究を進めている。これまでに、隅田らにより Wikipedia から上位下位関係を自動獲得する手法が提案された[1]。現在、さらなる抽出精度の向上と抽出用語を増やすことによる上位下位関係辞書の拡張を目指している。辞書の拡張のためには、Wikipedia のみの情報源では限界があり、膨大な Web テキストに出現するあらゆる用語に対して上位概念となる用語を特定する必要がある。

従来、上位下位関係をテキストデータから抽出する研究は数多く行われてきた。安藤らは日本語の新聞記事を対象として、「A という B」といった上位下位を表す典型的な構文パターンを用いる手法を提案している[2]。この手法では、限られた構文パターンを用いているため、たとえ大量の Web データを対象としても 100 万語規模の上位下位関係を抽出することは難しい。また、新里らは HTML 文書を対象として、HTML 文書の構造と *df*、*idf* などの統計量、用語の係り受け関係などを用いる手法を提案している[3]。この手法では、HTML タグによる箇条書きなどの並列な構造にある用語のみを下位語候補としており、他の部分にある用語は処理対象としていない。

本稿では、Web テキストに出現し、かつ、上位下位関係を獲得できていない用語が、Wikipedia から自動獲得した上位下位関係辞書のどこに属するかを推定する手法を提案する。提案手法では、用語が、助詞を介して動詞を修飾する係り受け構造により特徴付けする。各用語の特徴として、用語のクラスタリングで利用する条件付き確率の分布を用いる。この確率分布を利用することにより、Web テキストに出現する用語の上位語として相応しい用語を、上位下位関係辞書の上位語から抽出し、上位下位概念辞書を拡張する。実験では、Web テキストに出現する用語からテストデータを作成し、このテストデータを対象とした上位語推定処理について報告する。

## 2. 上位下位関係辞書

隅田らの手法により抽出された上位下位関係辞書は、約 244 万対の上位下位関係と約 121 万種の下位語を含む。この手法では、Wikipedia の各記事に含まれる定義文、カテゴリー情報、階層構造を知識源として、SVM により上位下位関係か否かを判定することにより 90% の抽出精度が得られている。この上位下位関係辞書は、従来から広く利用されているシソーラス[4][5]などと比較して、以下の特徴を持つ。

- 大規模なデータである
- 固有名を数多く含む
- 木構造の木の深さが比較的浅い

Wikipedia データには膨大な量の用語が登録されており、各データ構造を解析することにより、大規模な上位下位関係が得られている。また、従来のシソーラスとの差分の多くは固有名であり、ニュース記事などの実データを解析する際に有用なデータと考えられる。

上位下位関係における下位語の多くが固有名であるため、さらなる下位語を持つことが少ない。上位下位関係辞書の木構造における木の深さは平均 1.89 であった。これは、6 階層の木構造で最下層のみに語が配置されている分類語彙表[4]と比較しても浅い構造となっていることがわかる。木構造の深さが浅いと、新たな用語の上位語推定処理では問題となる場合がある。そこで、自動生成された上位下位関係の最上位の上位語に対して、その形態素を語の先頭から順に削除する縮退処理を行い、上位語を生成する。例えば「いも焼酎」という語は、縮退した「焼酎」を「いも焼酎」の上位語として新たに上位下位関係に追加する。今回、縮退処理は人手により行い、その結果を利用した[6]。縮退した語を補間した結果、木の深さは平均 2.33 となった。この上位下位関係辞書を、Web に出現する用語の上位語推定の対象とする。

## 3. 用語の特徴抽出

Web に出現する用語と上位下位関係に出現する上位語、下位語を特徴付けるため、用語の係り受け関係を利用する。鳥澤は、用語が助詞を介して動詞を修飾する係り受け構造の類似性を利用して、助詞「は」「も」

の意味役割を推定する手法を提案した[7]。この手法では、用語 $n$ と、用語が助詞 $rel$ を介して動詞 $v$ に係る三項組<$v$, $rel$, $n$>の出現確率を以下の通り定義している。

$$P(<v, rel, n>) =_{def} \sum_{a \in A} P(<v, rel>|a)P(n|a)P(a)$$

ここで、$a$ は用語や動詞、助詞が属する隠れクラスを、$A$ は隠れクラスの集合を表す。$A$ に含まれる隠れクラス数はあらかじめ設定する。

三項組の出現確率 $P(<v, rel, n>)$を構成する $P(<v, rel>|a)$、$P(n|a)$、$P(a)$は、隠れクラス $a$ 自体を観測できないため、直接算出することができない。そこで、データ $L=\{<v_1,rel_1,n_1>, <v_2,rel_2,n_2>,\cdots, <v_m,rel_m,n_m>\}$を利用して EM アルゴリズムにより推定する。

EM アルゴリズムの結果得られた $P(n|a)$と $P(a)$から、用語 $n$ が隠れクラス $a$ に属する確率 $P(a|n)$を求めることができる。$P(a|n)$は用語のクラスタリングで利用可能な確率分布あり、最大の事後確率を持つ $\hat{a}$ は、用語 $n$ が属するクラスと考えることができる。同じ動詞と助詞の組み合わせ<$v$,$rel$>と頻繁に共起するような用語群は同じ隠れクラスへ分類される。

風間らは、鳥澤の手法を大規模な Web のテキストデータに応用して用語のクラスタリング処理を行い、その結果を利用して、固有名抽出処理の精度向上を実現した[8]。さらに[9]では、EM アルゴリズムで使う用語や動詞、助詞の出現頻度の代わりに、頻度の対数を使って計算し、良好な結果が得られている。本稿では、風間らが行った隠れクラス数 2000 個、用語数 100 万語に対する処理[9]により得られた確率分布 $P(a|n)$, $a=\{a_1, a_2,\cdots, a_{2000}\}$, $n=\{n_1,n_2, \cdots, n_{1000000}\}$を利用する。

## 4. 上位下位関係辞書への追加手法

用語クラスタリング処理対象の 100 万語は、用語が出現した構造（助詞と動詞の組み合わせ）の種類数の降順に選択されている。この 100 万語と、本稿で拡張対象としている Wikipedia から自動獲得した上位下位関係辞書の用語を比較した。結果を以下に記す。

| | |
|---|---|
| 上位語の共通用語数 | 27,705 語 |
| 下位語の共通用語数 | 156,710 語 |

用語クラスタリング処理対象の 100 万語中約 84 万語は上位下位関係の下位語として出現しておらず、上位下位関係辞書へ追加可能な拡張対象用語となる。

クラスタリング処理対象と上位下位関係辞書との共通用語を手がかりとして、拡張対象とする用語が上位下位関係辞書中のどの上位語の下位語となるかを推定する 4 つの手法を提案する。以下に、その詳細について説明する。

### 4.1 対象用語に最も類似する下位語の上位語(手法 1)

用語間の類似性は、各用語が隠れクラスに属する確率分布間の Jensen-Shannon divergence により評価できる[9]。これを拡張対象の用語と下位語に適用する。用語 $n_{trg}$ と下位語 $n_{hypor}$ が、隠れクラス $a$ に属する確率分布 $P(a|n_{trg})$と $P(a|n_{hypor})$の間の Jensen-Shannon divergence は、以下の式で定義される。

$$D_{JS}(P(a|n_{trg}) \| P(a|n_{hypor})) = \frac{1}{2}(D_{KL}(P(a|n_{trg}) \| \frac{P(a|n_{trg})+P(a|n_{hypor})}{2}) + D_{KL}(P(a|n_{hypor}) \| \frac{P(a|n_{trg})+P(a|n_{hypor})}{2}))$$

ここで $D_{KL}$ は Kullback-Leibler divergence を示す。Jensen-Shannon divergence は 0~1 の値を取り、類似する確率分布間は値が 0 に近づき、類似しない確率分布間は 1 に近づく。この値を利用し、拡張対象の用語 $n_{trg}$ が持つ各隠れクラスタへ属する確率分布に最も類似する確率分布を持つ用語を、下位語の共通用語から一つ抽出する。抽出された下位語と親子関係にある上位語を、対象用語の上位語とする。最も類似する下位語との Jensen-Shannon divergence が一定値 $\theta$ 以上のものは、上位語を抽出しない。次節以降の手法でも同様に、適切な上位語として抽出するための距離やスコアに対するしきい値 $\theta$ を設ける。

### 4.2 対象用語に類似する下位語群を持つ上位語(手法 2)

上位下位関係辞書に含まれる上位語を、その上位語と親子関係にある下位語により特徴付ける。上位語 $n_{hyper}$ が各隠れクラスに属する確率分布を以下の式により定義する。

$$P(a|n_{hyper}) = \frac{\sum_{n_{hypor}} P(a|n_{hypor})P(n_{hypor})}{\sum_{n_{hypor}} P(n_{hypor})}$$

ここで $n_{hypor}$ は、上位語 $n_{hyper}$ と直接、親子関係にある下位語を示す。Jensen-Shannon divergence により、拡張対象の用語 $n_{trg}$ の確率分布に最も類似する確率分布を持つ上位語 $n_{hyper}$ を一つ抽出し、対象用語の上位語とする。

### 4.3 対象用語に類似する下位語群を一部に持つ上位語(手法 3)

処理の対象としている上位下位関係辞書は Wikipedia から自動作成しているため、ノイズが 10%程度含まれる。手法 2 では、上位語を特徴付けるために、上位語と親子関係にある全下位語を利用しているが、ノイズが上位語の特徴付けの際に問題となる可能性がある。そこで、上位語を特徴付ける下位語を、対

象語に類似するもののみに制限する。上位語に属する下位語のうち、対象語に類似する上位 10 語により上位語の確率分布を決定する。拡張対象の用語 $n_{trg}$ と最も類似する上位語 $n_{hyper}$ を一つ抽出し、対象用語の上位語とする。この際、下位語の数が 10 語未満である上位語は、全ての下位語を使用する。

### 4.4 対象用語に類似する下位語の上位下位関係辞書における配置を利用した上位語推定(手法 4)

上位下位関係辞書から、拡張対象用語に類似する下位語を一定数だけ抽出し、その類似する下位語群と各下位語が隠れクラスに属する確率分布を利用して、対象用語の上位語を推定する。上位語に対して、以下の式により拡張対象用語の上位語らしさを評価する。

$$score(n_{hyper}) = \sum_{sn_{hypor}} d^{r} \times (1 - D_{JS}(P(a \mid n_{trg}) \,||\, P(a \mid sn_{hypor}))$$

ここで $sn_{hypor}$ は上位語 $n_{hyper}$ の下位に属し、かつ、拡張対象用語 $n_{trg}$ と類似する上位 $m$ 個に含まれる用語である。上位 $m$ 個を無条件で抽出すると類似しない用語まで含まれる可能性が出るため、加算対象を $D_{JS}$ が一定値 $D_{min}$ 以下のものに制限する。また、上位語の下位に属する用語を選択する際、親子関係にある下位語だけでなく、孫やひ孫など、間接的に上位下位の関係を持つ語も大きな情報源となる。そこで手法 4 では、このような間接的に下位に属する全ての語を $sn_{hypor}$ の候補として利用する。$r$ は上位語 $n_{hyper}$ と下位語 $sn_{hypor}$ の木構造中の階層の差を示し、親子関係にあるものを $r$=0 とする。この階層の差に対するペナルティの値を $d$ とする。この $score(n_{hyper})$ の値が最大の上位語 $n_{hyper}$ を一つ抽出し、対象用語 $n_{trg}$ の上位語とする。

## 5. 実験

前章で提案した手法を検証するために、上位語推定実験を行った。用語クラスタリング処理対象との共通用語である上位下位概念辞書の上位語を、拡張対象用語の上位語候補とし、共通する下位語を上位語抽出のための手掛かりとして利用した。

上位下位関係の下位語として出現していない用語クラスタリング処理対象の用語から無作為に 329 語を抽出し、手作業により上位語の正解データを与えた。1 つの用語に対して、平均 4.2 個の正解を与えている。抽出データのうち 100 語を各手法のパラメータ設定用データ、残りを評価用のテストデータとした。

最初に、パラメータ設定用データによる予備実験を行い、各手法の精度を最大とするパラメータを決定する。以下の時に予備実験における F 値が最大となった。

(手法 1) Jensen-Shannon divergence のしきい値 $\theta$ =0.60
(手法 2) Jensen-Shannon divergence のしきい値 $\theta$ =0.67
(手法 3) Jensen-Shannon divergence のしきい値 $\theta$ =0.67
(手法 4) 拡張対象用語と類似する上位個数 $m$=300
対象とする $D_{JS}$ の最低値 $D_{min}$=0.65
木構造中の階層の差に対するペナルティ値 $d$=0.6
$score(n_{hyper})$のしきい値 $\theta$ =0.7

予備実験で得られたパラメータを用い、テストデータを対象とした用語の上位語を抽出する実験を行った。処理結果の一部を表 1 に示す。表中の下線のある用語は、正解と判定した上位語を示している。例えば、拡張対象用語「AGP カード」は手法 1～3 により抽出された「プリント基板」と、手法 4 により抽出された「製品」の両方を正解としている。

表 2 に評価結果を示す。適合率は、しきい値 $\theta$ によって制限して抽出した中での正解の割合を示し、再現率は、全テストデータ 229 用語に対する正解の割合を示す。手法 1 では、最も類似する下位語に対する上位語が複数存在する可能性もあるが、そのうちの一つが正解に含まれていれば正解と判定した。手法 4 では、F 値が 0.609 と、他の手法に比べて良好な結果が得られた。手法 4 の処理で、しきい値を上げて適合率を 80%としたとき、再現率は 18.3%であった。クラスタリング処理と上位下位関係辞書との差分の 84 万語を全処理対象とすると、適合率 80%で、約 12 万語の適切な上位下位関係が新たに獲得可能と期待できる。

表 2 の結果では、手法 2 が最も低い F 値となっているが、これは上位下位関係辞書に含まれるノイズが問題となっていたためと考えられる。手法 1、手法 3 が手法 4 より精度が低い原因として、上位語が特定され過ぎていることが挙げられる。例えば表 1 において、「内分泌細胞」の上位語として「幹細胞」や「免疫細胞」が抽出されている。一方、手法 4 では上位下位関係の木構造を全て考慮しているため、対象用語に類似する下位語が複数の上位語の下位に出現している場合、それらの上位を抽出することが可能で、この場合でも、「幹細胞」や「免疫細胞」の上位にあたる「細胞」が抽出できている。

拡張対象用語がサ変名詞の時、手法 4 でも上位語を誤抽出することが多かった。自動構築した上位下位関係辞書には、サ変名詞が上位語として出現することが少なく、上位下位関係の情報不足が原因と考えられる。日本語 WordNet[5]などの既存のシソーラスを利用した上位下位関係辞書の補間を行うことにより、このような問題は解決できると考えられる。また、抽出された上位概念が同義語であるような誤りも見られた。例えば表 1 で、用語「ノートパソ」の上位語として、手法 1 と手法 4 では「ノートパソコン」が抽出されている。提案しているアルゴリズムでは、同義語の問題が考慮

表 1. 上位語抽出実験結果の一部（下線のある用語は正解と判定された上位語、括弧内の数値は各手法における評価指標の値を示す）

| 拡張対象用語 | 手法 1 | 手法 2 | 手法 3 | 手法 4 |
|---|---|---|---|---|
| AGP カード | プリント基板(0.36) | プリント基板(0.36) | プリント基板(0.36) | 製品(3.40) |
| EC 市場 | 情報サイト(0.30) | 自動車工業(0.53) | 企業(0.50) | 企業(4.02) |
| MrMrs スミス | 映画,作品(0.18) | 告白すること(0.30) | 映画(0.21) | 映画(165.64) |
| にぎり寿司 | キャラクター(0.26) | 沖縄料理(0.29) | 料理(0.30) | 料理(20.12) |
| アジャンター | メインレストラン,観光地,都市(0.46) | 遺跡など(0.53) | 遺跡など(0.53) | 世界遺産(2.01) |
| ノートパソ | ノートパソコン(0.33) | 製品シリーズ(0.39) | 製品シリーズ(0.39) | ノートパソコン(8.67) |
| 温総理 | 国会議員(0.27) | 会長(0.37) | 会長(0.37) | 政治家(5.34) |
| 結核性髄膜炎 | 疾患(0.189) | 疾患概念(0.23) | 疾患概念(0.23) | 疾患(45.65) |
| 言語治療 | テーマ(0.46) | 看護(0.48) | 看護(0.48) | 雑誌(9.86) |
| 児童ポルノ禁止法 | シンポジウム,議員立法(0.36) | 医事法(0.37) | 日本の法律(0.37) | 法律(54.82) |
| 食塩濃度 | 物質の性質(0.43) | 無次元数(0.44) | 定数(0.44) | 指標(4.60) |
| 銅地金 | 特産品,名産品(0.36) | 天然ガス(0.45) | 石油燃料(0.50) | 鉱石(2.57) |
| 内分泌細胞 | 幹細胞(0.32) | 抗原提示細胞(0.33) | 免疫細胞(0.33) | 細胞(5.98) |

表 2. 上位語抽出実験の評価結果

| | 適合率 | 再現率 | F 値 |
|---|---|---|---|
| 手法 1 | 0.347 (75/216) | 0.327 (75/229) | 0.337 |
| 手法 2 | 0.235 (50/213) | 0.218 (50/229) | 0.226 |
| 手法 3 | 0.359 (78/217) | 0.341 (78/229) | 0.350 |
| 手法 4 | 0.652 (131/201) | 0.572 (131/229) | 0.609 |

されていないため、同義語については別処理を行う必要がある。

今回は、語の多義性について考慮していない。拡張対象用語は辞書に登録されていないような語のため、固有名が多く、複数の語義を持つことが少ないと考えられるためである。今後、多義性のある語の処理についても検討を進める。

## 6. まとめ

本稿では、Web テキストに出現する用語の上位語として相応しい用語を、用語が Web テキスト中で出現する際の用語が助詞を介して動詞を修飾する係り受け構造を利用して、上位下位関係辞書の上位語から抽出する手法を提案した。実験により、対象用語に類似する下位語の上位下位関係辞書における配置を利用する手法で F 値 0.609 が得られ、最も効果的であることが分かった。今後、適合率を向上させた結果を利用して上位下位関係辞書を拡張し、ウェブ検索ディレクトリとしての言語資源として公開を予定している鳥式改[10]への応用を進める。

【 参考文献 】


[1] A. Sumida, N. Yoshinaga and K.Torisawa, "Boosting Precision and Recall of Hyponymy Relation Acquisition from Hierarchical Layouts in Wikipedia," In *Proceedings of the Sixth International Language Resources and Evaluation*, 2008

[2] 安藤, 関根, 石崎, "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, " 情処学研報, 2003-NL-157, pp.77-82(2003)

[3] K.Shinzato and K. Torisawa, "Acquiring Hyponymy Relations from Web Documents," In *Proceedings of HLT-NAACL*, pp73-80( 2004)

[4] 国立国語研究所, "分類語彙表 増補改定版" (2004)

[5] F. Bond, H. Isahara, K. Uchimoto, T. Kuribayashi, K. Kanzaki, "Extending the Japanese WordNet," 言語処理学会第 15 回年次大会発表論文集, C1-4(2009)

[6] 黒田, 李, 野澤, 村田, 鳥澤, "鳥式改の上位語データの人手クリーニング, " 言語処理学会第第 15 回年次大会発表論文集, C1-3(2009)

[7] K. Torisawa, "An unsupervised method for canonicalization of Japanese postpositions," In *Proceedings of the 6th Natural Language Processing Pacific Rim Symposium*(NLPRS), pp 211-218(2001)

[8] J. Kazama and K. Torisawa, "Inducing Gazetteers for Named Entity Recognition by Large-scale Clustering of Dependency Relations," In *Proceedings of ACL-08: HLT*, pp.407-415 (2008)

[9] 風間, De Saeger, Stijn, 鳥澤, 村田, "係り受けの確率的クラスタリングを用いた大規模類似語リストの作成," 言語処理学会第第 15 回年次大会発表論文集, C1-3(2009)

[10] 鳥澤, 隅田, 野口, 柿澤, 風間, Stijn De Saeger, 村田, 黒田, 山田, 塚脇, 太田, "ウェブ検索ディレクトリの自動構築とその改良 ---鳥式改---," 言語処理学会第 15 回年次大会発表論文集, P2-1(2009)