情報処理学会論文誌 Vol. 52 No. 12 3435–3447 (Dec. 2011)

分布類似度とWikipediaから獲得した構造情報を利用した上位下位関係獲得

山田一郎[†1] 鳥澤健太郎[†2] 風間淳一[†2]
黒田航[†3,†4,†5] 村田真樹[†6] ステインデ・サーガ[†2]
フランシスボンド[†7] 隅田飛鳥[†8] 橋本力[†2]

質問応答などの自然言語処理アプリケーションが実用レベルに至るには，計算機で扱うことのできる，世界についての膨大な知識を構築する必要がある．本論文では，そのような知識の筆頭といえる，「サッカー選手/長友佑都」などの語句間の上位下位関係を自動獲得する手法を提案する．提案手法は，Wikipedia から獲得した上位下位関係と，Web テキストから獲得した語句間類似度情報を併用することで，網羅的かつ高精度に上位下位関係を獲得する．評価実験では，提案手法の適合率が，複数のベースライン手法の適合率に比べて，スコア上位 10,000 ペアでは 0.155 から 0.650 の差で，スコア上位 100,000 ペアでは 0.190 から 0.500 の差で上回ることを確認した．また，提案手法の獲得結果の中には，広く用いられている語彙統語パターンによる手法では獲得できない上位下位関係が多く含まれていることを確認した．

Hyponymy Relation Acquisition Based on Distributional Similarity and Hierarchical Structure of Wikipedia

Ichiro Yamada,[†1] Kentaro Torisawa,[†2]
Jun'ichi Kazama,[†2] Kow Kuroda,[†3,†4,†5]
Masaki Murata,[†6] Stijn De Saeger,[†2]
Francis Bond,[†7] Asuka Sumida[†8]
and Chikara Hashimoto[†2]

In order to make natural language processing (NLP) applications such as question answering accurate enough for practical use, it is essential to build a large-scale, computer-tractable semantic knowledge base. In this paper, we target hyponymy relation like "football player/Yuto Nagatomo," which is one of the most important semantic relations for NLP. We propose a new method of large scale hyponymy relation acquisition from Web texts that combines a hyponymy relation database constructed from Wikipedia and the distributional similarity between words calculated from Web texts. Experimental results showed that, in terms of precision, our method outperformed nontrivial baseline methods by 0.155 to 0.650 for the top 10,000 pairs and by 0.190 to 0.500 for the top 100,000 pairs. Furthermore, we confirmed that our method could acquire hyponymy relation pairs that widely-used lexico-syntactic pattern based approaches could not.

1. はじめに

質問応答などの自然言語処理アプリケーションが実用レベルに至るには，計算機で扱うことのできる，世界についての膨大な知識を構築する必要がある．「サッカー選手/長友佑都」のような語句間の上位下位関係はそのような知識の筆頭といえ，これまでに多くの研究がなされてきた[3),4),9),12)–14),17),19)–21),24)]．上位下位関係は，「長友佑都とは誰ですか」のような質問への答えとして直接的に用いられるほか，自然言語処理の基盤的意味解析の 1 つである語義曖昧性解消タスクにおいても利用される[18)]．上位下位関係に関する言語資源はこれまでにいくつか構築されているが，大規模なものであっても，固有名や新語を十分網羅しているとはいい難い[25)]．最新の固有名や新語を含む網羅的な上位下位関係言語資源を低コストで構築し，メンテナンスし続けるには，世界最大のテキストデータである Web を情報

†1 NHK 放送技術研究所
Science & Technology Research Laboratories, Japan Broadcasting Corporation
†2 情報通信研究機構
National Institute of Information and Communications Technology
†3 京都工芸繊維大学
Kyoto Institute of Technology
†4 京都大学
Kyoto University
†5 早稲田大学総合研究機構
Comprehensive Research Organization, Waseda University
†6 鳥取大学大学院工学部研究科
Graduate school of engineering, Tottori University
†7 南洋理工第学
Nanyang Technological University
†8 株式会社 KDDI 研究所
KDDI R&D Laboratories

源とする，高精度な上位下位関係自動獲得手法の開発が欠かせない．

本論文では，大量の Web テキストから自動獲得した大規模な語句間類似度情報と Wikipedia を情報源として，網羅的かつ高精度な上位下位関係を自動獲得する手法を提案する．Wikipedia からは信頼性の高い上位下位関係を自動獲得できるが，Wikipedia の記述内容に偏りがあるため，獲得結果の網羅性は高くない．一方，大量の Web テキストからは，網羅的な語句集合とそれらの語句間の信頼性の高い類似度情報を大量に自動獲得できる．両者を組み合わせることによって網羅的で高精度な上位下位関係を自動獲得できる，というのが本研究のポイントである．

2 章で詳述するとおり，従来の上位下位関係自動獲得のアプローチは，語彙統語パターンを用いるもの，クラスタリングを用いるもの，Wikipedia を用いるものの 3 つに大別できる．これらのアプローチには，同一文中で共起していない上位語句と下位語句からなる上位下位関係が獲得できない，1 つのクラスタ中の下位語句集合には同じ上位語句が一様に付与されてしまう，獲得結果の上位下位関係に偏りがある（網羅性に欠ける）などの欠点がある．一方提案手法は，高い網羅性と精度を示しつつ，これらの問題点を回避している．

提案手法は大きく分けて次の 3 つのステップからなる．

(1) Wikipedia からの上位下位関係自動獲得（獲得結果を以後，Wikipedia 上位下位関係データベースと呼ぶ）
(2) Web テキストからの語句集合の取得と語句間の分布類似度計算
(3) Wikipedia 上位下位関係データベースと語句間分布類似度を用いた対象語句の上位語句獲得

本研究では，ステップ (1) には Sumida ら[21] の手法を，ステップ (2) には風間ら[28] の手法を用いた．それぞれ 3 章と 4 章で述べる．ステップ (3) では，Wikipedia 上位下位関係データベースに出現しない語句を対象語句として，その上位語句を獲得する．この処理が，我々の主たる貢献部分となる．ステップ (3) は次の 3 つのサブステップからなる（**図 1**）．

(3-1) 対象語句との類似度が最も高い k 語句（以後，k-最類似語句と呼ぶ）を Wikipedia 上位下位関係データベースから選択する．
(3-2) k-最類似語句を利用して，Wikipedia 上位下位関係データベース中の上位語句に対して，対象語句に対する上位語句らしさを示すスコアを与える．その際，Wikipedia 上位下位関係データベースの親子関係や兄弟関係を表す木構造の情報を利用する．
(3-3) スコアが最大の上位語句を対象語句の上位語句として獲得する．

評価実験では，Wikipedia 上位下位関係データベースに存在しない対象語句約 670,000 語

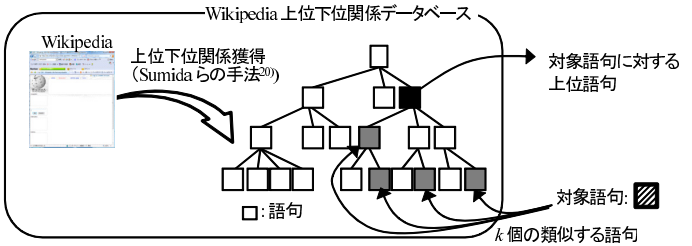


図 1　提案手法の全体像
Fig. 1　Overview of the proposed method.

句を下位語句と見なし，その上位語句を Wikipedia 上位下位関係データベースの中から自動獲得した．実験の結果，提案手法の適合率が，複数のベースライン手法の適合率に比べて，スコア上位 10,000 ペアでは 0.155 から 0.650 の差で，スコア上位 100,000 ペアでは 0.190 から 0.500 の差で上回ることを確認した．また，提案手法の獲得結果の中には，広く用いられている語彙統語パターンによる手法では獲得できない上位下位関係が多く含まれていることを確認した．

なお本論文では，次のいずれかが成り立つ A（下位語句）と B（上位語句）の関係を上位下位関係と定義する．

- A は B の一種である
- A は B の一例である

前者の条件は A と B がともに概念である場合，後者は B が概念で A がインスタンスである場合である．従来，上位下位関係は概念間の関係としてとらえられてきたが，本研究では固有名も対象に含めることが実用レベルの自然言語処理アプリケーションにとって重要であると考えるため，「A は B の一例である」が成立する A と B も上位下位関係に含める．また本論文では，A，B として，「サッカー選手」や「インテルに所属するサッカー選手」のような，複数形態素あるいは複数文節からなる表現も認めることにする．このように条件を緩和することで，より情報量の多い上位下位関係を獲得できる．その結果，たとえば質問応答では，「長友佑都とは誰ですか」のような質問に対し，単に「選手」と答えるだけでなく，「サッカー選手」や「インテルに所属するサッカー選手」のように，より詳細な答えを提供できる．

以下，2 章では上位下位関係自動獲得の先行研究について述べ，提案手法の優位性を主張する．3 章では，Wikipedia 上位下位関係データベースの獲得手法と獲得結果について述べ

る．4 章では語句間分布類似度の計算方法について述べる．5 章では，Web テキスト中の対象語句に対する上位語句獲得手法について説明する．6 章で評価実験の結果について報告し，7 章で結論と今後の課題について述べる．

2. 関 連 研 究

テキストから語句間の上位下位関係を自動獲得する研究は従来からさかんに研究されてきた．これまでに提案されてきたアプローチは，語彙統語パターンを用いるもの，クラスタリングに基づくもの，Wikipedia を用いるものの 3 つに大別できる．

語彙統語パターンによる手法として，たとえば Hearst[9] や Pantel ら[12]，安藤ら[24]，Snow ら[19),20)] のものがある．語彙統語パターンとはたとえば “B such as A” のようなパターンであり，このようなパターンに合致する A と B のペアは上位下位関係（B は A の上位語句）と見なせる．語彙統語パターンによる手法は Hearst によって最初に提案された．Hearst は少数の語彙統語パターンをシードとして，それにより新たな語彙統語パターンを獲得することによって上位下位関係を自動獲得した．Pantel らは，語彙統語パターン間の編集距離を利用して上位下位関係を表すパターンを自動獲得し，大量の上位下位関係を獲得する手法を提案した．安藤らは，語彙統語パターンを日本語の新聞記事に対して適用して，日本語の語句の上位下位関係獲得を試みた．Snow らは，語句間の係り受け関係を素性として機械学習により上位下位関係を学習する手法を提案した．語彙統語パターンによる手法の問題点は，同一文中で共起している上位語句と下位語句から構成される上位下位関係しか獲得できない点である．この制約は大規模な上位下位関係の獲得にとってボトルネックとなる（**図 2** 左）．一方，我々の提案手法は，上位語句が Wikipedia に，下位語句が Web テキストに出現していなくてはならないという比較的緩い条件が課せられてはいるが，上位語句と下位語句の同一文中での共起は前提としていないため，大規模な上位下位関係の獲得により適していると考えられる．

クラスタリングに基づく手法として，Caraballo[3] や Pantel ら[13]，Shinzato ら[17]，Etzioni ら[4] のものがあげられる．Caraballo，Pantel ら，Shinzato らは，クラスタリングによって得られたクラスタ中の語句群に対して共通の上位語句を付与する手法を提案した．Etzioni らはクラスタリングによる手法と語彙統語パターンによる手法を組み合わせた手法を提案した．クラスタリングに基づく手法の問題点は，1 クラスタ中の語句すべてに同一の上位語句を付与する点，いい換えれば，クラスタのメンバである語句それぞれに最適な上位語句を選択できない点にある（図 2 右）．一方提案手法では，1 章で述べたステップ (3) で，個々の対象語句に対して個別に上位語を推定する処理を行うため，クラスタリングに基づく手法で生じる問題は発生しない．

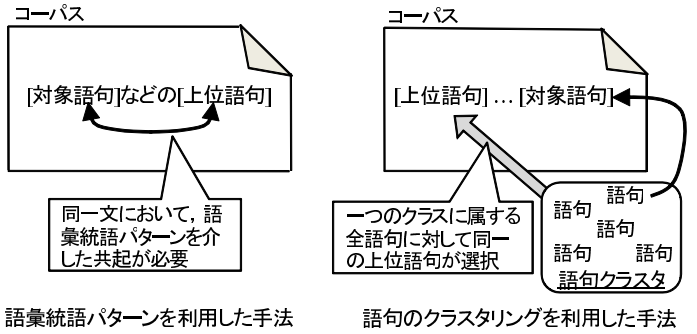


図 2　語彙統語パターンによる手法とクラスタリングによる手法の問題点
Fig. 2　Bottlenecks of lexico-syntactic pattern based methods and clustering baesd methods.

Ponzetto ら[14] は Wikipedia の category の情報を，Sumida ら[21] は Wikipedia の記事タイトル，節タイトル，小節タイトル，箇条書きなどの階層構造についての情報を用いて，Wikipedia から信頼性の高い上位下位関係を獲得する手法を提案した．しかし，Wikipedia で議論されているとおり[⋆1]，Wikipedia に出現する語句には偏りがあり，Web テキストには出現しているが Wikipedia にはない語句が多数ある．そのため，Wikipedia のみを情報源とする手法には網羅性の点で問題がある．一方提案手法は，Wikipedia から獲得した上位下位関係と，Web テキストから獲得した語句群と語句間類似度情報とを組み合わせることによって，信頼性と網羅性を高いレベルで両立させることができる．

Blohm ら[1] は，Web に対する検索結果と Wikipedia のタイトルとハイパーリンク情報を用いて，語句間の関係を獲得する手法を提案している．Wikipedia と Web テキストの 2 種類の情報源から獲得した学習データを利用した実験では，アルバムとそのアーティストなど 7 種類の関係を対象とし，Wikipedia のみを使用する手法の精度を維持しながら学習データ量を有意に減少できることを示している．また Wu ら[23] は，Wikipedia の Infobox 名と WordNet のノード（synset）を統合してより精錬されたオントロジを生成する KOG（Kylin Ontology Generator）を提案した．この手法では，Wikipedia の Infobox の情報と，Web に対する検索結果を学習処理における素性として利用している．これらの研究は，

⋆1 http://ja.wikipedia.org/wiki/Wikipedia: なぜウィキペディアは素晴らしくないのか.

目的とそのアプローチが提案手法と異なるが，単語間の特定の関係を抽出する処理において Wikipedia と Web テキストの両方を利用することの有効性を実証している．

3. Wikipedia 上位下位関係データベース

提案手法は，Wikipedia から獲得した上位下位関係（Wikipedia 上位下位関係データベース）と，Web テキストから獲得した語句群と語句間類似度情報とを併用することで，大量の上位下位関係を獲得する．本章では Wikipedia 上位下位関係データベースの獲得手法について述べる．

本研究では，Sumida らの手法をもとにして Wikipedia 上位下位関係データベースを構築した．Sumida らの手法は，Wikipedia 記事の階層構造に着目して上位下位関係を獲得する．Wikipedia 記事の階層構造とは，記事タイトル，節タイトル，小節タイトル，箇条書きなどで表現される記事の論理的な構造のことである．**図 3** に Wikipedia 記事「ペンギン」のソースファイルの一部を例としてあげる．この記事は「形態」「生態」「繁殖」「分類と種のリスト」などの節に分かれ，「分類と種のリスト」はさらに「コウテイペンギン属」「アデリーペンギン属」に分かれる．このようなタイトル–節関係あるいは節–小節関係などのうちのいくつか，たとえば記事タイトルと小説タイトルのペアである「ペンギン/キングペンギン」，「ペンギンの本/ペンギンになった不思議な鳥」などは正しい上位下位関係と見なすことができる．

```
ペンギンはペンギン目・ペンギン科に
属する鳥類の総称である．
== 形態 ==
== 生態==
== 繁殖 ==
== 分類と種のリスト ==
* コウテイペンギン属
**[[コウテイペンギン]]
** [[キングペンギン]]
* アデリーペンギン属
== ペンギンと文化 ==
== ペンギンの本 ==
* ペンギンになった不思議な鳥
* ペンギンハンドブック
== 関連項目 ==
* [[ペンギノン]]
* [[ペンギン（ミサイル）]]
[[Category:ペンギン目]]
```

図 3　Wikipedia 記事「ペンギン」ソースファイルの一部
Fig. 3　A part of data dump clipped from the article “Penguin” in Wikipedia.

Sumida らの手法は，Wikipedia からの上位下位関係候補の取得と，上位下位関係候補の正例/負例への分類の 2 ステップからなる．分類には SVM を用いており，上位語句と下位語句を構成する各語句の品詞，上位語句と下位語句それぞれに含まれる形態素，Wikipedia 記事における出現位置の階層上の差などを素性として利用している．この手法を 2007 年 3 月のバージョンの Wikipedia に適用したところ，適合率 0.90 [⋆1]で約 240 万ペアの上位下位関係が獲得できた．

Sumida らの手法で得られた上位下位関係は，これまでに構築されてきたシソーラス，たとえば日本語 WordNet [2]，日本語語彙大系[26]，分類語彙表[30] と比べて次のような特長を持つ．

- 語句数が多い
- 固有名を数多く含む

日本語 WordNet には約 93,000 語句，日本語語彙大系には約 300,000 語句，分類語彙表には約 96,000 語句が収録されている一方，Sumida らの手法で得た上位下位関係には約 1,200,000 語句[⋆2]が収録されている．また，これら既存のシソーラスには含まれていないが Sumida らの手法の出力には含まれている語句の多くは固有名である．実用に耐える自然言語処理アプリケーションによって固有名を網羅的に扱えることは必須であるが，既存のシソーラスに比べて Sumida らの手法の出力はその要件をより良く満たしているといえる．

しかし，関連研究で述べたとおり，Wikipedia に出現する語句には偏りがあり，Web テキストには出現しているが Wikipedia にはない語句が多数ある．たとえば，アニメやゲームに関する情報は詳細をきわめている一方で，日本の上代文学史に関する情報はわずかである．自然言語処理における最重要技術の 1 つである意味役割付与に至っては記事すら存在しない．つまり，Wikipedia のみを情報源とする手法には網羅性に欠けるという弱点がある．我々の提案手法は，Web テキストから得た語句群とそれら語句間の類似度情報を Sumida

⋆1 SVM のしきい値により調整．
⋆2 2007 年 3 月の Wikipedia を対象として得られた 240 万ペアの上位下位関係における，下位語句の異なり数である．

らの手法の出力に組み合わせることで，より網羅性の高いシソーラスを構築するものである．

Sumida らの手法の出力は上位語句と下位語句の 2 項関係の集合であるが，我々の提案手法では，上位語句あるいは下位語句の一致を手がかりにこれらを 1 つの木構造にまとめあげたうえで利用する．しかし，Sumida らの手法の出力から得た木構造はルートからリーフまでの距離の平均が 2.34 であり，比較的浅い．我々の提案手法では語句間の木構造上の距離を手がかりの 1 つのとして利用するため，階層構造からより多くの情報が得られることが望ましい．そこで，我々の提案手法を適用する前に，Sumida らの手法の出力の階層をより深いものにした．具体的には，上位語句が複合名詞である場合，その複合名詞の主辞を当該複合名詞の上位語句として新たに追加することによって階層を深くした．日本語は主辞後続型言語なので，複合名詞の右側の形態素列を主辞と見なすことができる．たとえば，上位語句が「大分麦焼酎」という複合名詞である場合，その上位語句として「麦焼酎」を，さらにその上位語句として「焼酎」を新たに追加できる．この処理では Kuroda らの手法[27] に従い，形態素解析器を用いて新たな上位語句候補を自動抽出し，この上位語句候補に対して人手チェックを行う．人手チェックにより上位語句として適切と判定されたものだけを Sumida らの手法の出力に追加した結果，ルートからリーフまでの距離の平均は 2.74 となった．この結果を Wikipedia 上位下位関係データベースとし，本提案手法において，対象語句の上位語句獲得時の手がかりとして利用する．

4. 分布類似度

提案手法では Web テキスト中に出現する語句の上位語句として適切なものを Wikipedia 上位下位関係データベース中から選択するが，その際に語句間の分布類似度を用いる．分布類似度とは，「類似する文脈に出現する語句は意味的に類似する」という分布仮説[7] に基づく類似度尺度である．分布類似度計算に用いる文脈と計算法にはいくつかの選択肢がある．提案手法では，文脈として，語句に付属する助詞と，「その語句＋助詞」の係り先の動詞を採用した．たとえば語句が「焼酎」なら，その文脈は「を 飲む」「を 買う」「で 仕込む」「が 楽しめる」などとなる．この文脈は Rooth ら[15] や Torisawa[22]，Hagiwara ら[6]，風間ら[28] でも採用され，良好な結果が報告されている．以下本章では，本研究で用いる，風間らの手法に基づいた分布類似度計算について述べる．まず，提案手法で使用する分布類似度計算法 CDS（Clustering based Distributional Similarity）について，その後，6 章で述べる提案手法の変種の 1 つで用いる分布類似度計算法 DDS（Dependency relation based Distributional Similarity）について述べる．

提案手法で用いる，2 つの語句 n_1，n_2 の間の分布類似度計算法 $CDS(n_1, n_2)$ は次のように定義される．

$$CDS(n_1, n_2) = 1 - D_{JS}(P(\cdot \mid n_1) \parallel P(\cdot \mid n_2)) \tag{1}$$

$CDS(n_1, n_2)$ は 0 から 1 の間の値をとる．n_1 と n_2 が類似しているほど値は 1 に近づき，逆にまったく類似していない場合は 0 になる．$D_{JS}(P(\cdot \mid n_1) \parallel P(\cdot \mid n_2))$ は n_1 と n_2 の確率分布間の距離である．確率分布間の距離を求める関数はいくつか提案されているが[11]，ここでは風間ら[28] の手法に従い次の Jensen-Shannon Divergence を用いる．

$$\begin{aligned} D_{JS}(P(\cdot \mid n_1) \parallel P(\cdot \mid n_2)) = \frac{1}{2} \Bigg(& D_{KL}(P(\cdot \mid n_1) \parallel \frac{P(\cdot \mid n_1) + P(\cdot \mid n_2)}{2}) \\ & + D_{KL}\left(P(\cdot \mid n_2) \parallel \frac{P(\cdot \mid n_1) + P(\cdot \mid n_2)}{2}\right) \Bigg) \end{aligned} \tag{2}$$

ここで，D_{KL} は式 (3) で定義される Kullback-Leibler divergence である．

$$D_{KL}(P(\cdot \mid n_1) \parallel P(\cdot \mid n_2)) = \sum P(\cdot \mid n_1) \log \frac{P(\cdot \mid n_1)}{P(\cdot \mid n_2)} \tag{3}$$

語句の確率分布は，後述する隠れクラス $a \in A$ への語句の所属確率 $P(a|n)$ から構成される．$P(a|n)$ は，次のように定義される．

$$P(a|n) = \frac{P(n|a)P(a)}{\sum_{a \in A} P(n|a)P(a)} \tag{4}$$

Torisawa は，「焼酎 を 飲む」のような係り受け関係を成す語句 n，助詞 rel，動詞 v の 3 項組 $< v, rel, n >$ が出現する確率 $P(< v, rel, n >)$ を次のように定義したうえで，与えられたコーパスの尤度が最大になるような $P(< v, rel > |a)$，$P(n|a)$，$P(a)$ の値を EM アルゴリズムにより計算した．

$$P(< v, rel, n >) =_{def} \sum_{a \in A} P(< v, rel > |a)P(n|a)P(a) \tag{5}$$

ここで a は隠れクラスを，$P(< v, rel > |a)$ はクラス a における $< v, rel >$ の生起確率を，$P(n|a)$ はクラス a における n の生起確率を，$P(a)$ は a の事前確率を表す．$P(< v, rel > |a)$，$P(n|a)$，$P(a)$ はコーパス中で観測できないため，直接算出することができない．そこで EM アルゴリズムを用い，与えられたコーパスにおける各確率値を計算する．E ステップでは，$P(a| < v, rel >)$ を計算し，M ステップでは，尤度を最大にするように $P(< v, rel > |a)$，$P(n|a)$，$P(a)$ を更新する．EM アルゴリズムの中で用いる各種頻度は log 関数によって補正したものを用いた．頻度を直接用いず log 関数によって補正する手法は，従来いくつかの

論文で提案されており，類似度の計算において良好な結果が報告されている[28),29)]．EM アルゴリズムの結果得られた隠れクラスごとの $P(n|a)$，$P(a)$ により，式 (4) を用いて $P(a|n)$ を計算することができる．

このようにして得られた $P(a|n)$ 群は語句クラスタリングの結果と見なすことができる．CDS の要点は，クラスタリングによって語句 n が出現する文脈に対して一種のスムージングを行うことで，データスパースネスの問題に対処している点である．

一方，提案手法の変種の 1 つで用いられる分布類似度計算法 DDS ではクラスタリングを用いない．CDS と DDS の違いは，語句の確率分布が，$P(a|n)$ ではなく，次のように定義される $P(< v, rel > |n)$ から構成される点である．

$$P(< v, rel > |n) = \frac{f(< v, rel, n >)}{f(n)} \tag{6}$$

ここで，$f(n)$ は語句 n の出現頻度，$f(< v, rel, n >)$ は 3 項組 $< v, rel, n >$ の出現頻度を表す．DDS においても頻度は log 関数によって補正したものを用いた．つまり，CDS と DDS の違いは，前者における語句の確率分布が $P(a|n)$ から構成されるのに対し，後者における語句の確率分布は $P(< v, rel > |n)$ から構成される，という点にある．

5. 対象語句の上位語句獲得

提案手法では，Web テキストには出現しているが Wikipedia 上位下位関係データベースには出現していない語句（対象語句）を対象に，その上位語句として適切なものを Wikipedia 上位下位関係データベース中の上位語句候補の中から，後述するスコアリング結果に基づいて獲得する．その際，対象語句と Wikipedia 上位下位関係データベース中の下位語句との分布類似度と，Wikipedia 上位下位関係データベースの階層構造の情報を手がかりとして利用する．Wikipedia 上位下位関係データベースについては 3 章で，本研究で使用する分布類似度については 4 章で述べた．本章では，対象語句に対する上位語句の獲得方法について述べる．

まず，対象語句と Wikipedia 上位下位関係データベース中の下位語句との間の分布類似度を前章で述べた計算結果から取得する．この分布類似度の値の降順に上位 k 個の下位語句（k-最類似語句）を抽出する．ただし，順位が k 以内でも，分布類似度の値が閾値 S_{min} より低いものは処理対象から除く．

次に，式 (7) に基づいて，Wikipedia 上位下位関係データベース中の各上位語句候補 n_{hyper} に対象語句 n_{trg} の上位語句らしさを表すスコア $score(n_{hyper}, n_{trg})$ を付与する．

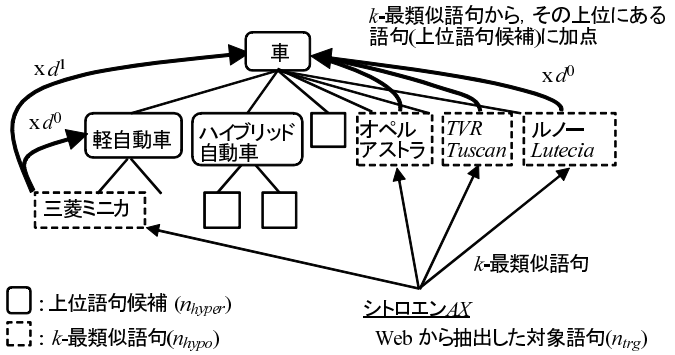


図 4　上位語句獲得のためのスコアリング処理の例
Fig. 4　Example of the scoring process for hypernym acquisition.

$$score\,(n_{hyper}, n_{trg}) = \sum_{n_{hypo} \in Desc(n_{hyper}) \cap ksimilar(n_{trg})} d^{r(n_{hyper}, n_{hypo})-1} \times sim(n_{trg}, n_{hypo}) \tag{7}$$

ここで，$Desc(n_{hyper})$ は n_{hyper} の下位語句集合，$ksimilar(n_{trg})$ は n_{trg} の k-最類似語句集合を表す．また，$r(n_{hyper}, n_{hypo})$ は n_{hyper} と下位語句 n_{hypo} の Wikipedia 上位下位関係データベースにおける階層の深さの差を表し，直接の親子関係にある場合に 1 を，祖父と孫の関係にある時は 2 をとる．つまり n_{hyper} と n_{hypo} の階層差が大きくなるにつれて値が増加する．d は階層差に対するペナルティ値を表し，0 から 1 までの値をとる．$sim(n_{trg}, n_{hypo})$ は n_{trg} と n_{hypo} の間の分布類似度を表し，提案手法では CDS が，提案手法の変種の 1 つでは DDS が用いられる．

最後に，スコアが最大の上位語句候補を対象語句の上位語句として獲得する．スコアは上位語句獲得結果の信頼性を示すための値としても利用される．つまり，スコアが大きいほど信頼できる獲得結果と見なす．

図 4 に，対象語句「シトロエン AX」に対する上位語句獲得処理を例としてあげる．まず，「シトロエン AX」の k-最類似語句として Wikipedia 上位下位関係データベースから「三菱ミニカ」，「オペルアストラ」，「TVR Tuscan」，「ルノー Lutecia」を抽出する．次に，各 k-最類似語句からその上位語句に対して式 (7) に基づきスコアを与える．「オペルアストラ」，「TVR Tuscan」，「ルノー Lutecia」はその上位語句である「車」に，「三菱ミニカ」はその上位語句である「軽自動車」と「車」にスコアリングする．最後に，スコアが最大の「車」

が「シトロエン AX」の上位語句として獲得される．

6. 評価実験

本章では，他の手法との比較を通して，提案手法の有効性を主張する．以下では，まず，実験で使用するデータと比較対象の手法，各手法のパラメータの推定方法について述べ，その後，実験結果を考察とともに報告する．最後に，広く利用されている語彙統語パターンを用いた手法では獲得できない上位下位関係が提案手法によって獲得できることを示す．

6.1 実験で使用するデータ

実験では，開放型検索エンジン基盤 TSUBAKI[16] において収集されたコーパス[⋆1]に出現している約 670,000 語句を対象語句とした．対象語句約 670,000 語句は，まず，TSUBAKI コーパスから，共起する動詞，助詞のペア $< v, rel >$ の異なり数が多い 1,000,000 語句を取得し，その中から，Wikipedia 上位下位関係データベースに出現している語句や，上位語句，下位語句としては不適切なものを除外することで得た．上位語句，下位語句としては不適切なものとは，数字や記号のみからなる文字列（たとえば「000」，「‥」，「＃」）や語句の断片と思われる文字列（たとえば「の中」，「ぁ」）などであり，不適切かどうかは人手により判定した．実験に先立ち，約 670,000 語句の間の分布類似度を CDS，DDS のそれぞれを用いて計算した．CDS の隠れクラス数は 2,000 とした．

実験で用いる Wikipedia 上位下位関係データベースは 3 章で述べたものである．上位下位関係の総数は約 2,400,000 ペアで，そのうち，上位語句の異なり数が約 95,000 語句，下位語句の異なり数が約 1,200,000 語句である．実験では，これらの上位語句約 95,000 語句のうち，上述した対象語句群に含まれる 28,015 語句を，いい換えれば，対象語句との分布類似度が計算済みのもののみを対象語句の上位語句候補とした．下位語句も同様に，対象語句群に含まれる 175,022 語句のみを上位語句獲得時の手がかりとして使用する．**図 5** に，対象語句 670,000 語句，上位語句候補 28,015 語句，上位語句獲得時の手がかりとして使用する下位語句 175,022 語句，Wikipedia 上位下位関係データベース全体の関係を示す．

以上で述べたデータから，評価データセットと，6.3 節で述べるパラメータ推定用の開発データセットを作成する．開発データセットは，対象語句約 670,000 語句からランダムサンプリングした 698 語句のそれぞれに対し，上位語句候補 28,015 語句の中から人手で選択した上位語句 1 語句以上を付与することで作成した．開発データセットの上位下位関係の総数は 7,016 ペア（1 語句に対して平均 10.1 個の上位語句を付与）である．評価データセットには，開発データセットとして使用する 698 語句以外の対象語句をすべて用いる．ただし，システム出力に対する人手による正解判定を行うのはその一部である．詳しくは 6.4 節で述べる．

⋆1 日本語 Web ページ約 1 億件，文数にして約 60 億文が収録されたバージョンを使用．

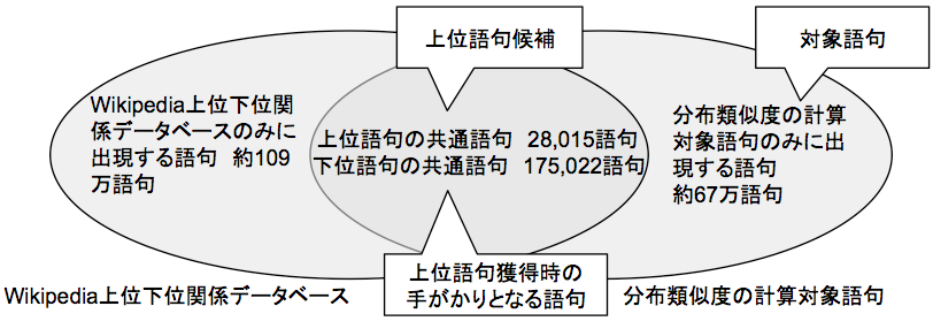


図 5　対象語句，上位語句候補，手がかりとして利用する下位語句の関係
Fig. 5　The relation between target words, hypernym candidates, and hyponym words that are used as cues.

6.2 比較対象の手法

比較対象の手法は，提案手法の変種の手法 2 つとベースライン手法 3 つの計 5 手法である．以下それぞれを，提案手法・変種 1，提案手法・変種 2，ベースライン手法 1，ベースライン手法 2，ベースライン手法 3 と呼ぶ．

提案手法・変種 1 は，使用する分布類似度計算法が DDS である点が提案手法と異なる．つまり提案手法・変種 1 は，提案手法と異なり，分布類似度計算の際にクラスタリングを用いない．これは，提案手法と提案手法・変種 1 との比較を通して，分布類似度を計算する際のクラスタリングによるスムージングの効果を検証することを目的としている．

提案手法・変種 2 は，上位語句候補のスコアの計算法 (7) において，階層差に対するペナルティを表すパラメータ d の値を 0 とする点が提案手法と異なる．$d = 0$ とすることで，k-最類似語句と直接の親子関係にある上位語句候補のみにスコアが付与される[⋆2]．つまり提案手法・変種 2 は，提案手法と異なり，Wikipedia 上位下位関係データベースの階層構造の

⋆2 式 (7) において，親子関係にある語句 h_{hyper}，h_{hypo} の階層の差は $r(h_{hyper}, h_{hypo}) = 1$ となり，$d^{r(h_{hyper}, h_{hypo})-1} = 0^0$ となるが，ここでは便宜上，$0^0 = 1$ として計算する．

利用が限定的である．これは，提案手法と提案手法・変種 2 との比較を通して，Wikipedia 上位下位関係データベースの階層構造全体を利用することの効果を検証することを目的としている．分布類似度計算法は提案手法と同じ CDS を用いる．

ベースライン手法 1 は，対象語句と最も類似する Wikipedia 上位下位関係データベース中の下位語句を抽出したうえで，それと直接の親子関係にある上位語句候補を対象語句の上位語句として選択する．このベースライン手法は，類似する語は同じ上位語句を持つ傾向がある，というヒューリスティックスに基づく．分布類似度計算法は提案手法と同じ CDS を用いる．対象語句と当該下位語句の分布類似度の値は，上位語句らしさの指標となる信頼性の値として利用する．この値が大きいほど信頼性が高い獲得結果と見なす．なおこの手法では，1 つの対象語句に対して複数の上位語句が選択される場合がある．これは Wikipedia 上位下位関係データベース中の下位語句が複数の上位語句を持つ可能性があることに起因する．この手法では，どれが最も尤もらしいかは判定せず，それらすべての上位語句候補を対象語句の上位語句として出力する．

ベースライン手法 2 は，Wikipedia 上位下位関係データベース中のすべての上位語句候補の中で，対象語句との分布類似度が最も大きいものを対象語句の上位語句として選択する．このベースライン手法は，上位語句は対象語句と類似する，というヒューリスティックスに基づく．分布類似度計算法は提案手法と同じ CDS を用いる．分布類似度が最も高い上位語句候補が対象語句の上位語句として選択される．分布類似度の値は，上位語句らしさを表す信頼性の値として利用する．

ベースライン手法 3 は，ベースライン手法 1 と同様，類似する語は同じ上位語句を持つ傾向がある，というヒューリスティックスに基づく．ベースライン手法 1 との違いは，対象語句に最も類似する下位語句を持つ上位語句を選択するのではなく，対象語句に最も類似する下位語句集合を持つ上位語句を選択する点にある．ベースライン手法 1 では，対象語句に最も類似する 1 つの下位語句から上位語句を選択したが，ベースライン手法 3 では，類似する複数の下位語句を利用するため，より頑健な処理が期待できる．具体的には，上位語句候補 n_{hyper} の直接の下位語句集合 $Ch(n_{hyper})$ が隠れクラス $a \in A$ に帰属する確率分布 $P(a|n_{hyper})$ を，下位語句集合に含まれる語句の確率分布の平均として次式により計算する．

$$P_{child}(a|Ch(n_{hyper})) = \frac{\sum_{n_{hypo} \in Ch(n_{hyper})} P(a|n_{hypo})P(n_{hypo})}{\sum_{n_{hypo} \in Ch(n_{hyper})} P(n_{hypo})} \quad (8)$$

対象語句 n_{trg} との確率分布間の距離を，式 (2) の Jensen-Shannon Divergence により計算し，この距離が最も小さい下位語句集合の上位語句を，対象語句の上位語として選択する．式 (1) の値は，上位語句らしさを表す信頼性の値として利用する．Wikipedia 上位下位関係データベースには不正確な上位下位関係が混入している場合があるため，下位語句が少数しかない上位語句候補に対しては信頼性の高い分布類似度は計算できない．そこで，この手法では閾値 min_{hypo} 語句以上の下位語句を持つ上位語句のみを上位語句候補とした．

6.3　パラメータの推定

開発データセットを用いて，提案手法と比較手法の各パラメータの最適な値を推定した．具体的には，使用する k-最類似語句の数 k，類似度の閾値 S_{min}，階層差に対するペナルティ値 d，ベースライン手法 3 の下位語句数の閾値 min_{hypo} をそれぞれ次のように変化させて上位語句獲得を行い，精度が最良となる値の組合せを選択した．

$k = \{20, 40, 60, 80, 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000\}$
$S_{min} = \{0.6, 0.65, 0.7, 0.75, 0.8, 0.85, 0.9, 0.95, 1.0\}$
$d = \{0.5, 0.55, 0.6, 0.65, 0.7, 0.75, 0.8, 0.85, 0.9, 0.95, 1.0\}$
$min_{hypo} = \{1, 3, 5, 10, 20, 30\}$

パラメータ推定実験の結果，次の値の組合せが選択された[⋆1]．

- 提案手法のパラメータ
  $k = 100$
  $S_{min} = 0.7$
  $d = 0.65$
- 提案手法・変種 1 のパラメータ
  $k = 60$
  $S_{min} = 0.95$
  $d = 0.6$
- 提案手法・変種 2 のパラメータ
  $k = 100$
  $S_{min} = 0.7$
  $d = 0$
- ベースライン手法 3 のパラメータ
  $min_{hypo} = 20$

⋆1 ベースライン手法 1 とベースライン手法 2 にはパラメータがないことに注意する．

提案手法と提案手法・変種 1 は k の値（使用する k-最類似語句の数）が大きく異なり，提案手法では $k = 100$ と大きい値である場合に最良の値となっている．一方，提案手法・変種 1 の $k > 60$ では，精度の低下が見られた．これは，提案手法において類似語のランクが低いものでも上位語句獲得に有効であることを示し，分布類似度計算時にクラスタリングを用いる手法（提案手法）はクラスタリングを用いない手法（提案手法・変種 1）と比較して，類似する語句を多く集めることができているためと考えられる．また，提案手法と提案手法・変種 1 では類似度の閾値 S_{min} も大きく異なるが，これは類似度の計算方法の違いから生じた差と考えられる．

6.4 実験結果

最適化したパラメータを用いて，各手法で上位下位関係獲得実験を行った．各手法の評価は次の手順で行った．まず，各手法の獲得結果である上位下位関係群を，その手法における信頼性の値によりソートする．次に，上位 10,000 ペア，上位 100,000 ペア，全ペア（約 670,000 ペア）から上位下位関係 200 ペアをランダムサンプリングする．ランダムサンプリングされた上位下位関係群は，出所の手法がどれか分からないようにシャッフルされたうえで，著者を含まない評価者 4 名により上位下位関係として正しいかどうかが判定された．評価者には，上位語句 B と下位語句 A が「A は B の一種である」か「A は B の一例である」のいずれかにあてはまれば正解と判定するよう指示した．本実験では，3 名以上の評価者が正解と判定したものだけを正解とした．正解/不正解の判定が 2 名ずつに分かれるケースはなかった．評価者による判定の一致度を表す Fleiss' Kappa 統計量[5] は 0.693 だった．この値は評価者間の判定の一致度が高いこと（substantial agreement [10]）を示している．

なお，ベースライン手法 1 は他の手法と異なり，1 つの対象語句に対して複数の上位語句を選択する場合がある．本実験では，ベースライン手法 1 に対しては，1 つの対象語句と 1 つ以上の上位語句の組を 1 つの出力と見なして評価した．より具体的には，1 つの対象語句に対して正しい上位語句が 1 つでも選択されていれば，その出力は正しいものと見なした．つまり，ベースライン手法 1 に対しては他の手法に比べて緩い基準で評価したことになる．後述するとおり，ベースライン手法 1 に対しては緩い基準で評価したにもかかわらず，提案手法は統計的に有意な差でベースライン手法 1 より高精度だった．

表 1 に提案手法，提案手法・変種 1，提案手法・変種 2 の上位 10,000 ペア，上位 100,000 ペア，全ペアにおける適合率を，**表 2** に 3 つのベースライン手法の適合率をあげる．適合率は次の式で計算される．

表 1　提案手法と提案手法の変種のランク別適合率
Table 1　Precision of the proposed method and its variants.

ランク	提案手法 (CDS)	提案手法・変種 1 (DDS)	提案手法・変種 2
10,000	**0.795**	0.715	0.735
100,000	**0.740**	**0.740**	0.725
全ペア	**0.425**	0.385	0.405

表 2　ベースライン手法のランク別適合率
Table 2　Precision of baseline methods.

ランク	ベースライン手法 1	ベースライン手法 2	ベースライン手法 3
10,000	0.640	0.145	0.310
100,000	0.550	0.245	0.240
全ペア	0.325	0.090	0.145

$$\text{適合率} = \frac{\text{3 名以上の評価者が正解と判定した数}}{\text{サンプリングした上位下位関係数 (200 ペア)}} \tag{9}$$

これらの結果から次の 3 点が明らかになった．

(1) 提案手法は，3 つのベースライン手法のいずれに対しても，統計的に有意な差で高い精度を示す．

(2) クラスタリングを用いる分布類似度である CDS はクラスタリングを用いない分布類似度である DDS より，上位下位関係獲得タスクにおいて優れている．

(3) Wikipedia 上位下位関係データベースの階層構造全体を利用する方が，直接の親子関係のみを利用するより，高い精度を示す．

1 点目に関して，提案手法の評価結果と 3 つのベースライン手法の評価結果を対象として Fisher の正確確率検定[8] を行ったところ，上位 10,000 ペア，上位 100,000 ペア，全ペアのすべてにおいて，提案手法と全ベースライン手法との差が統計的に有意であることを確認した（$p < 0.05$）[⋆1]．具体的には，提案手法の適合率が，複数のベースライン手法の適合率に比べて，スコア上位 10,000 ペアでは 0.155 から 0.650 の差で，スコア上位 100,000 では 0.190 から 0.500 の差で上回った．これは，分布類似度だけを手がかりとする手法に比べて，分布類似度の情報と Wikipedia 上位下位関係データベースにあるような階層構造の情報を

⋆1 提案手法の変種 1 に対しては「ベースライン手法 1 の全ペア」以外で，提案手法の変種 2 に対してはすべてにおいて，ベースライン手法との差が統計的に有意であった（$p < 0.05$）．

表 3　提案手法による獲得した上位下位関係の例（下線は上位語句が誤りであることを示す）
Table 3　Examples of hyponymy relations acquired by the proposed method, with acquisition errors underlined.

Score	対象語句	上位語句
58.6	INDIVI	ブランド
54.3	クレオメ	花
34.4	UOKR	ゲーム
21.7	置戸	町
20.5	スマートフォーツー	車
15.6	深川めし	料理
8.9	John Barry	作曲家
8.5	JVM	ソフトウェア
6.6	メタンガス	元素
5.4	メールセミナー	本
3.9	グルメット	商品
3.1	スプリングバック	現象

併用する提案手法が優れていることを意味する．

2 点目は，提案手法と提案手法・変種 1 の評価結果の違いから明らかである．つまり，上位 100,000 ペアでは同等の精度だが，上位 10,000 ペアと上位約 670,000 ペアでは，CDS を用いる提案手法が DDS を用いる提案手法・変種 1 より高い精度を示している．

3 点目は，提案手法と提案手法・変種 2 の評価結果の違いから明らかである．つまり，上位 10,000 ペア，上位 100,000 ペア，全ペアのすべてにおいて，Wikipedia 上位下位関係データベースの階層構造全体を利用する方が直接の親子関係のみを利用するより高い精度を示している．

表 3 に，提案手法で得られた上位下位関係の一部をそのスコアとともにあげる．たとえば，対象語句「INDIVI」の上位語句として「ブランド」が，対象語句「クレオメ」の上位語句として「花」が正しく選択されているのが分かる．一方，提案手法の誤りの多くは，たとえば表 3 の「本/メールセミナー」などのように，実際は「本」でも何らかの「作品」でもない対象語句に対して「本」や「作品」が上位語句として選択されるケースだった[⋆1]．この原因は次のように考えられる．世の中に存在する本や作品の中には，そのタイトルがその本あるいは作品自体を指すだけでなく，それ以外の事物を指すものがある．たとえば，「菩提樹」は歌曲のタイトルだがシナノキ科の落葉高木も指す．「高瀬川」は小説のタイトルだが京都の川も指す．その結果，たとえば，何らかの河川を意味するが本のタイトルではない対象語句が，「高瀬川」などの実在する川も本のタイトルも指す複数の下位語句と高い類似度を示した場合，当該対象語句に対して「本」が上位語句として誤って選択されうる．つまり，本や作品のタイトルが持ちうる曖昧性が提案手法の誤りの大きな原因となっている．この問題への対策として，下位語句の多くがこの種の曖昧性を持つような上位語句候補をあらかじめ特定し，その上位語句候補を所与の対象語句の上位語句として選択する際は何らかの条件を別途設ける，という方法が考えられる．何らかの条件としては，たとえば，上位下位関係を表す語彙統語パターンにおける対象語句と上位語句の共起があげられる．つまり，「本」や「作品」を上位語句として選択する際はより厳しい条件を設定する，ということである．この可能性の追求は今後の課題とする．

以下，ベースライン手法のエラー分析結果について述べる．ベースライン手法 2 の誤りの多くは，選択された上位語句が実際には上位語句ではなく類義語句である場合だった．たとえば「クリーニング工場」の上位語句として「セメント工場」が選択されたが，「セメント工場」は「クリーニング工場」の上位語句ではなく類義語句である．ベースライン手法 1 とベースライン手法 3 は，ともに，上位語句候補の下位語句を手がかりとする点で共通している．これらの手法の誤りの多くは，Wikipedia 上位下位関係データベースに存在する不適切な上位下位関係に起因するものだった．いい換えれば，これらの手法は Wikipedia 上位下位関係データベース構築時のエラーに影響されやすい．

6.5　獲得可能な上位下位関係に関する語彙統語パターンによる手法との比較

従来提案されてきた上位下位関係の獲得手法のうち，最も広く用いられているのは語彙統語パターンによる手法であろう．一方，我々の提案手法は語彙統語パターンを利用しない．もし提案手法で獲得可能な上位下位関係と語彙統語パターンによる手法で獲得可能なものが大きく異なるのであれば，両手法を組み合わせることで上位下位関係をより大規模に獲得できる可能性がある．

我々は次の手順により，提案手法で獲得した上位下位関係の中で，語彙統語パターンによる手法では獲得できないものがどのくらい含まれているのかを調査した．まず，上位下位関係を表す語彙統語パターンとして以下のもの（A が下位語句，B が上位語句）を用意した．

A などの B，A など B，A のほか B，A の他 B，A という B，A という B，Aっていう B，Aっていう B，A に似た B，A と呼ばれる B，Aって呼ばれる B

これらのパターンを安藤ら[24]の手法により TSUBAKI コーパスに適用することで上位

⋆1「メールセミナー」という本は少なくとも本実験で使用したデータの中には存在しない．

下位関係候補約 2,374 万ペアを獲得した[⋆1]．一方，提案手法で獲得した上位下位関係の中からは 300 ペアをランダムサンプリングした．この 300 ペアから，安藤らの手法で獲得した上位下位関係候補に含まれるペアを除外した結果，243 ペアが残った．つまり，提案手法で獲得した上位下位関係のうち 81%（243/300）が語彙統語パターンによる手法では獲得できないものだったということができる．これらのパターン以外にも上位下位関係を表す語彙統語パターンは考えられるが，使用したパターンは上位下位関係を表す代表的なものであり，かつ，約 60 億文が含まれる大規模な TSUBAKI コーパスを利用しているため，語彙統語パターンで得ることができる上位下位関係の大部分はこのパターンから得られた上位下位関係候補に含まれると考えられる．以上の結果から，我々は，提案手法と語彙統語パターンによる手法を組み合わせることで上位下位関係をより大規模に獲得できる可能性があると考える．

7. おわりに

本論文では，大量の Web テキストから自動獲得した大規模な語句間類似度情報と Wikipedia 上位下位関係データベースを情報源として，網羅的かつ高精度な上位下位関係を自動獲得する手法を提案した．Wikipedia からは信頼性の高い上位下位関係を自動獲得できるが，Wikipedia の記述内容に偏りがあるため，獲得結果の網羅性は高くない．一方，大量の Web テキストからは，網羅的な語句集合とそれらの語句間の信頼性の高い類似度情報を大量に自動獲得できる．本研究のポイントは，両者を組み合わせることによって網羅的で高精度な上位下位関係を自動獲得できる，という点にある．評価実験から，次の 4 点が明らかになった．

(1) 提案手法の適合率はベースライン手法の適合率と比較してスコア上位 10,000 ペアでは 0.155 から 0.650 の差で，上位 100,000 ペアでは 0.190 から 0.500 の差で上回る．
(2) クラスタリングを用いる分布類似度である CDS は，クラスタリングを用いない分布類似度 DDS より上位下位関係獲得タスクにおいて優れている．
(3) Wikipedia 上位下位関係データベースの階層構造全体を利用する手法は直接の親子関係のみの上位下位関係を利用するより高い精度を示す．
(4) 提案手法で獲得される上位下位関係の多くは，語彙統語パターンを利用する手法では獲得することができないものである．

今後は，提案手法と語彙統語パターンによる手法を組み合わせることで，信頼性の高い上位下位関係をより大規模に獲得することを目指す．

⋆1 ただしこの中には，上位下位関係として不適切なものも含まれている．

参考文献

1) Blohm, S. and Cimiano, P.: Using the Web to Reduce Data Sparseness in Pattern-Based Information Extraction, *Proc. 11th European Conference on Principles and Practice of Knowledge Discovery in Databases* (*PKDD2007*) (2007).
2) Bond, F., Isahara, H., Kanzaki, K. and Uchimoto, K.: Boot-strapping a WordNet using Multiple Existing WordNets, *Proc. 6th International Conference on Language Resources and Evaluation* (*LREC*) (2008).
3) Caraballo, S.A.: Automatic construction of a hypernym-labeled noun hierarchy from text, *Proc. 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics* (*ACL*), pp.120–126 (1999).
4) Etzioni, O., Cafarella, M., Downey, D., Popescu, A.-M., Shaked, T., Soderland, S., Weld, D.S. and Yates, A.: Unsupervised named-entity extraction from the web: An experimental study, *Artificial Intelligence*, Vol.165, No.1, pp.91–134 (2005).
5) Fleiss, J.L. and Cohen, J.: The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability, *Educational and Psychological Measurement*, Vol.33, pp.613–619 (1973).
6) Hagiwara, M., Ogawa, Y. and Toyama, K.: PLSI Utilization for Automatic Thesaurus Construction, *Proc. International Joint Conference on Natural Language Processing* (*IJCNLP*), pp.334–345 (2005).
7) Harris, Z.: Distributional Structure, *The Philosophy of Linguistics*, Katz, J.J. (Ed.), pp.26–47, Oxford University Press (1985).
8) Hays, W.L.: Statistics: Analyzing Qualitative Data, *Proc. 14th Conference on Computational Linguistics, Ch.18*, pp.769–783 (1988).
9) Hearst, M.A.: *Automatic acquisition of hyponyms from large text corpora*, pp.539–545, Rinehart and Winston, Inc. (1992).
10) Landis, J. and Koch, G.G.: The measurement of observer agreement for categorical data, *Biometrics*, Vol.33, pp.159–174 (1977).
11) Lee, L.: Measures of Distributional Similarity, *Proc. 37th Annual Meeting of the Association for Computational Linguistics* (*ACL*), pp.25–32 (1999).
12) Pantel, P., Ravich, D. and Hovy, E.: Towards terascale knowledge acquisition, *Proc. Conference on Computational Linguistics* (*COLING*), pp.771–777 (2004).
13) Pantel, P. and Ravichandran, D.: Automatically Labeling Semantic Classes, *Proc. Human Language Technology and North American Chapter of the Association for Computational Linguistics Conference* (*HLT-NAACL*), pp.321–328 (2004).
14) Ponzetto, S.P. and Strube, M.: Deriving a Large-Scale Taxonomy from Wikipedia, *Proc. 22nd Conference on the Advancement of Artificial Intelligence* (*AAAI*),

pp.1440–1445 (2007).
15) Rooth, M., Riezler, S., Prescher, D. and Beil, G.C.F.: Inducing a Semantically Annotated Lexicon via EM-based Clustering, *Proc. 37th Annual Meeting of the Association for Computational Linguistics* (*ACL*), pp.104–111 (1999).
16) Shinzato, K., Shibata, T., Kawahara, D., Hashimoto, C. and Kurohashi, S.: Tsubaki: An open search engine infrastructure for developing new information access, *Proc. 3rd International Joint Conference on Natural Language Processing* (*IJCNLP*), pp.189–196 (2008).
17) Shinzato, K. and Torisawa, K.: Extracting hyponyms of prespecified hypernyms from itemizations and headings in web documents, *Proc. 20th International Conference on Computational Linguistics* (*COLING*), pp.938–944 (2004).
18) Shirai, K. and Yagi, T.: Learning a Robust Word Sense Disambiguation Model using Hypernyms in Definition Sentences, *Proc. 20th International Conference on Computational Linguistics* (*COLING*), pp.917–923 (2004).
19) Snow, R., Jurafsky, D. and Ng, A.Y.: Learning Syntactic Patterns for Automatic Hypernym Discovery, *Proc. Neural Information Processing Systems* (*NIPS*) (2005).
20) Snow, R., Jurafsky, D. and Ng, A.Y.: Semantic taxonomy induction from heterogenous evidence, *Proc. 43rd Annual Meeting of the Association for Computational Linguistics* (*ACL*), pp.801–808 (2006).
21) Sumida, A., Yoshinaga, N. and Torisawa, K.: Boosting Precision and Recall of Hyponymy Relation Acquisition from Hierarchical Layouts in Wikipedia, *Proc. 6th International Conference on Language Resources and Evaluation* (*LREC*) (2008).
22) Torisawa, K.: An Unsupervised Method for Canonicalization of Japanese Postpositions, *Proc. 6th Natural Language Processing Pacific Rim Symposium* (*NLPRS*), pp.211–218 (2001).
23) Wu, F. and Weld, D.: Automatically Refining the Wikipedia Infobox Ontology, *Proc. 17th World Wide Web Conference* (*WWW-08*) (2008).
24) 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，Vol.2003, No.98, pp.77–82 (2003).
25) 山田一郎，呉　鍾勳，鳥澤健太郎，黒田　航，風間淳一，村田真樹：Wikipedia を利用した日本語 WordNet への用語追加の検討，言語処理学会第 16 回年次大会発表論文集，pp.948–951 (2009).
26) 池原　悟，宮崎正弘，白井　諭，横尾昭男，中岩浩巳，小倉健太郎，大山芳史，林良彦：日本語語彙大系，岩谷書店 (1997).
27) 黒田　航，李　在鎬，野澤　元，村田真樹，鳥澤健太郎：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 16 回年次大会発表論文集，pp.76–79 (2009).
28) 風間淳一，ステイン デ・サーガ，鳥澤健太郎，村田真樹：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 16 回年次大会発表論文集，pp.84–87 (2009).
29) 寺田　昭，吉田　稔，中川裕志：文脈情報による同義語辞書作成支援ツール，情報処理学会研究報告，Vol.2006, No.124, pp.87–94 (2006).
30) 中野　洋：分類語彙表増補改定版，国立国語研究所 (1996).

(平成 23 年 4 月 12 日受付)
(平成 23 年 7 月 11 日採録)

山田　一郎（正会員）

1991 年名古屋大学工学部卒業．1993 年同大学院修士課程修了．博士（情報科学）．同年 NHK 入局．1996 年より NHK 放送技術研究所にて自然言語処理を利用した情報抽出，メタデータ生成の研究に従事．2003〜2004 年スタンフォード大客員研究員．2008〜2011 年独立行政法人情報通信研究機構専門研究員．現在，NHK 放送技術研究所主任研究員．映像情報メディア学会，言語処理学会各会員．

鳥澤健太郎（正会員）

1992 年東京大学理学部卒業．1994 年同大学院修士課程修了．1995 年同大学院博士課程中退．同年同大学院助手．1998 年科学技術振興事業団さきがけ研究 21 研究員兼任（2002 年まで）．北陸先端科学技術大学院大学助教授を経て，2008 年より独立行政法人情報通信研究機構言語基盤グループ，グループリーダー．2011 年より同機構情報分析研究室室長，現在に至る．博士（理学）．自然言語処理の研究に従事．日本学術振興会賞等受賞．言語処理学会，人工知能学会，ACL 各会員．

風間　淳一（正会員）

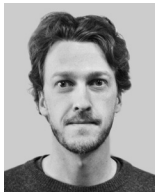
独立行政法人情報通信研究機構ユニバーサルコミュニケーション研究所情報分析研究室主任研究員．2004 年東京大学大学院情報理工学系研究科コンピュータ科学専攻博士課程修了．博士（情報理工学）．同年北陸先端科学技術大学院大学情報科学研究科助教．2008 年より情報通信研究機構．自然言語処理の研究に従事．

黒田　　航（正会員）

京都工芸繊維大学（非常勤講師），京都大学（非常勤講師），早稲田大学総合研究機構（招聘研究員）．元独立行政法人情報通信研究機構知識創成コミュニケーション研究センター MASTAR プロジェクト言語基盤グループ短時間研究員．京都大学人間・環境学博士．言語学と自然言語処理を融合する研究に従事．

村田　真樹（正会員）

1993 年京都大学工学部電気工学第二学科卒業．1997 年同大学院工学研究科電子通信工学専攻博士課程修了．博士（工学）．同年京都大学にて日本学術振興会リサーチ・アソシエイト．1998 年郵政省通信総合研究所入所．独立行政法人情報通信研究機構主任研究員を経て，現在，鳥取大学大学院工学研究科情報エレクトロニクス専攻教授．自然言語処理，情報抽出の研究に従事．2005 年 FIT2005 論文賞受賞．共著書に『事例で学ぶテキストマイニング』（共立出版）等がある．言語処理学会，人工知能学会，電子情報通信学会，計量国語学会，ACL 等の会員．



ステイン　デ・サーガ

2006 年北陸先端科学技術大学院大学知識科学研究科博士課程修了．博士（知識科学）．北陸先端科学技術大学院大学研究員を経て，2007 年に情報通信研究機構に入所．2008 年に NICT MASTAR プロジェクト言語基盤グループに専攻研究員として着任．自然言語処理を用いた知識獲得の研究に従事．

フランシス　ボンド

1988 年 Queensland 大学文学部日本語学科卒業．1990 年同大学工学部電気工学科卒業．1991 年日本電信電話株式会社入社．2001 年 Queensland 大学大学院博士課程修了．Ph.D.（英語学）．2006 年情報通信研究機構入所．2009 年南洋理工大学准教授．機械翻訳，辞書学，意味理解の研究に従事．IEEE，ACL，言語処理学会各会員．

隅田　飛鳥（正会員）

2005 年石川工業高等専門学校電子機械工学専攻卒業．2007 年北陸先端科学技術大学院大学情報科学研究科前期課程修了．2010 年同学後期課程修了．博士（情報科学）．同年株式会社 KDDI 研究所入所．自然言語処理の研究に従事．現在，株式会社 KDDI 研究所研究員．言語処理学会会員．

橋本　　力（正会員）

1999 年福島大学教育学部卒業．2001 年北陸先端科学技術大学院大学博士前期課程修了．2005 年神戸松蔭女子学院大学大学院博士後期課程修了．京都大学大学院情報学研究科産学官連携研究員を経て，2007 年山形大学大学院理工学研究科助教，2009 年より独立行政法人情報通信研究機構専攻研究員．2011 年京都大学大学院情報学研究科博士後期課程修了，現在に至る．自然言語処理の研究に従事．博士（言語科学，情報学）．言語処理学会，ACL 各会員．