Wikipediaの記事構造からの上位下位関係抽出

隅田　飛鳥[†,††]・吉永　直樹[†††]・鳥澤健太郎[††††]

本稿では，Wikipediaの記事構造を知識源として，高精度で大量の上位下位関係を自動獲得する手法について述べる．上位下位関係は情報検索やWebディレクトリなど，膨大なWeb文書へのアクセスを容易にする様々な技術への応用が期待されており，これまでにも様々な上位下位関係の抽出手法が開発されてきた．本稿では，Wikipediaの記事構造に含まれる節や箇条書きの見出しから，大量の上位下位関係候補を抽出し，機械学習を用いてフィルタリングすることで高精度の上位下位関係を獲得する手法を開発した．実験では，2007年3月の日本語版Wikipedia 2.2 GBから，約77万語を含む約135万対の上位下位関係を精度90%で獲得することができた．

キーワード：上位下位関係，Wikipedia, SVM, 半構造情報，記事構造

Hyponymy Relation Acquisition from Hierarchical Layouts in Wikipedia

Asuka Sumida[†,††], Naoki Yoshinaga[†††] and Kentaro Torisawa[††††]

This paper describes a method of extracting a large set of hyponymy relations with a high precision from hierarchical layouts in Wikipedia articles. Hyponymy relation has been studied as one of the principal knowledge for information retrieval and web directory, which helps users to access the growing web. Various methods have been proposed to automatically acquire hyponymy relations. In this article, we first extract hyponymy relation candidates from sections and itemizations in hierarchical layouts of Wikipedia articles, and then filter out irrelevant candidates by using a machine learning technique. In experiments, we successfully extracted more than 1.35 million relations from the hierarchical layouts in the Japanese version of Wikipedia, with a precision of 90%.

Key Words: *hyponymy* (is-a) *relation, Wikipedia, SVM, semi-structured information, hierarchical layouts*

† 北陸先端科学技術大学院大学情報科学研究科, Japan Advanced Institute of Science and Technology School of Information Science
†† 奈良先端科学技術大学院大学情報科学研究科, Nara Institute of Science and Technology Graduate School of Information Science
††† 東京大学生産技術研究所, Institute of Industrial Science, University of Tokyo
†††† 独立行政法人情報通信研究機構, National of Information and Communications Technology

1　まえがき

本稿では，大量の上位下位関係を Wikipedia から効率的に自動獲得する手法を提案する．ここで「単語 A が単語 B の上位語である（または，単語 B が単語 A の下位語である）」とは，Miller の定義 (Fellbaum 1998) に従い，「A は B の一種，あるいは一つである (B is a (kind of) A)」とネイティブスピーカーがいえるときであると定義する．例えば，「邦画」は「映画」の，また「イチロー」は「野球選手」のそれぞれ下位語であるといえ，「映画／邦画」，「野球選手／イチロー」はそれぞれ一つの上位下位関係である．以降，「A ／ B」は A を上位語，B を下位語とする上位下位関係（候補）を示す．一般的に上位下位関係獲得タスクは，上位下位関係にある表現のペアをどちらが上位語でどちらが下位語かという区別も行った上で獲得するタスクであり，本稿でもそれに従う．本稿では概念―具体物関係（ex. 野球選手／イチロー）を概念間の上位下位関係（ex. スポーツ選手／野球選手）と区別せず，合わせて上位下位関係として獲得する．

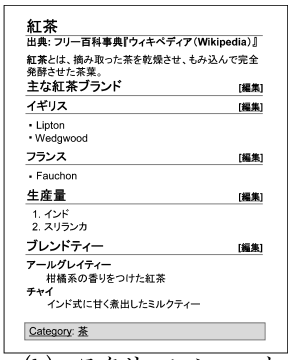
上位下位関係は様々な自然言語処理アプリケーションでより知的な処理を行うために利用されている (Fleischman, Hovy, and Echihabi 2003; 鳥澤, 隅田, 野口, 風間 2008)．例えば，Fleischman らは質問文中の語句の上位語を解答とするシステムを構築した (Fleischman et al. 2003)．また鳥澤らはキーワード想起支援を目的とした Web ディレクトリを上位下位関係をもとに構築した (鳥澤他 2008)．しかしながら，このような知的なアプリケーションを実現するためには，人手で書き尽くすことが困難な具体物を下位語とする上位下位関係を網羅的に収集することが重要になってくる．

そこで本稿では，Wikipedia の記事中の節や箇条書き表現の見出しをノードとするグラフ構造（以降，**記事構造**とよぶ）から大量の上位下位関係を効率的に獲得する手法を提案する．具体的には，まず記事構造上でノードを上位語候補，子孫関係にある全てのノードをそれぞれ下位語候補とみなし，上位下位関係候補を抽出する．例えば，図 1（b）の Wikipedia の記事からは 3 節で述べる手続きにより，図 1（c）のような記事構造が抽出できる．この記事構造上のノード「紅茶ブランド」には，その子孫ノードとして「Lipton」，「Wedgwood」，「Fauchon」，「イギリス」，「フランス」が列挙されている．提案手法をこの記事構造に適用すると，「紅茶ブランド」を上位語候補として，その子孫ノードを下位語候補群とする上位下位関係候補を獲得できる．しかしながら獲得した下位語候補には，「Wedgwood」，「Fauchon」のように下位語として適切な語が存在する一方，「イギリス」，「フランス」のような誤りも存在する．この例のように，記事構造は適切な上位下位関係を多く含む一方，誤りの関係も含むため，機械学習を用いて不適切な上位下位関係を取り除く．

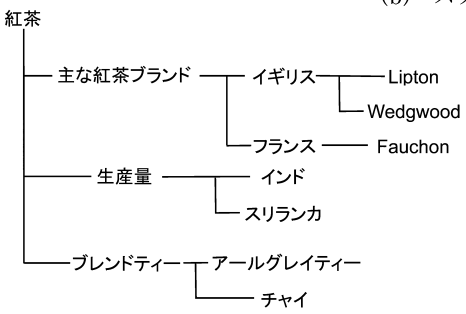
以下，2 節で関連研究と本研究とを比較する．3 節で提案手法で入力源とする Wikipedia の記事構造に触れ，4 節で提案手法について詳細に述べる．5 節では提案手法を日本語版 Wikpedia に適用し，獲得された上位下位関係の評価を行う．最後に 6 節で本稿のまとめと今後の展望に

```
1  紅茶とは、摘み取った茶を乾燥させ、もみ込んで完全
   発酵させた茶葉。
2  = 主な紅茶ブランド =
3  == イギリス ==
4  * Lipton
5  * Wedgwood
6  == フランス ==
7  * Fauchon
8  = 生産量 =
9  # インド
10 # スリランカ
11 = ブレンドティー =
12 ;アールグレイティー :柑橘系の香りをつけた紅茶
13 ;チャイ :インド式に甘く煮出したミルクティー
14 [[Category:茶]]
```

(a)　MediaWiki ソースコード



(b)　スクリーンショット



(c) 記事構造

図 1　「紅茶」に関する Wikipedia の記事の例

ついて述べる．

2　関連研究

本節では，既存の上位下位関係の自動獲得手法について説明する．上位下位関係の獲得は，1990 年代に Hearst が語彙統語パターンを用いて新聞記事から上位下位関係を獲得する手法を提案し (Hearst 1992)，以後各言語への応用がみられた (今角 2001; 安藤，関根，石崎 2003; Pantel and Pennacchiotti 2006; Sumida, Torisawa, and Shinzato 2006; 大石，伊藤，武田，藤井 2006)．その後 Web の発達に伴い，箇条書き表現などの Web 文書特有の手がかりを用いた獲得手法が提案されてきた (Shinzato and Torisawa 2004; Etzioni, Cafarella, Downey, Popescu, Shaked, Soderland, Weld, and Yates 2005) が，近年，具体物を含む概念間の知識を密に記述した Wikipedia に注目が集まっている (Ruiz-Casado, Alfonseca, and Castells 2005; Bunescu and Paşca 2006; Herbelot and Copestake 2006; Suchanek, Kasneci, and Weikum 2007; Kazama and Torisawa 2007)．以下では，まず新聞記事や Web 文書を対象とした上位下位関係の獲得手法について紹介し，その後 Wikipedia に特化した上位下位関係の獲得手法について述べる．以下，各手法について提案手

法との違いについて述べる．

2.1 新聞記事・Web 文書からの上位下位関係獲得

まず，語彙統語パターンを利用した研究として，(Hearst 1992; 今角 2001; 安藤他 2003; Sumida et al. 2006) があげられる．Hearst は英語の新聞記事を対象に，“〈上位語〉 such as 〈下位語〉”などのパターンを用いて上位下位関係を獲得した (Hearst 1992)．安藤らは Hearst に倣い，日本語の新聞記事コーパスを構文解析した結果から，“〈下位語 1〉（や〈下位語 2〉）*という〈上位語〉”などの同格・並列表現を含む語彙統語パターンを用い上位下位関係を獲得した (安藤他 2003)．また今角は日本語の新聞記事コーパスに対して“〈上位語〉「〈下位語〉」 のような括弧を用いたパターンを適用している (今角 2001)．この括弧を用いたパターンと名詞連続パターンを利用して，Sumida らは Web 文書から上位下位関係を獲得した (Sumida et al. 2006)．これらの手法では，信頼性の高いパターンを用いることで比較的高い精度で上位下位関係を獲得できるが，そのようなパターンで文書中に出現しない上位下位関係も数多く存在し，語彙統語パターンのみで大量の上位下位関係を獲得するのは本質的に難しい．

そこで，語彙統語パターンにマッチしない上位下位関係を獲得するため，Web 文書に頻出する箇条書き表現の文書構造を用いる手法が，Shinzato らや Etzioni らによって提案されている (Shinzato and Torisawa 2004; Etzioni et al. 2005)．Shinzato らは Web 文書中に繰り返し出現する HTML タグに囲まれた語の群を 1 つの単語クラスと見なし，この単語クラスに上位語を付与することで，上位下位関係を獲得する手法を提案した (Shinzato and Torisawa 2004)．また Etzioni らは語彙統語パターンを用いて抽出した上位下位関係をより広範な下位語に対応させるため，抽出した下位語を多く含むリスト構造を用いて，未知の下位語に上位語を割り当てる手法を提案した (Etzioni et al. 2005)．これらの手法でリソースとして用いている Web 文書の箇条書き表現は，上位下位関係の記述に限らず様々な用途に用いられるためノイズが多く，高い精度を保ったまま大量の上位下位関係を獲得することは難しい．これらの手法では箇条書き表現を基本的に下位語候補を収集するためにのみ用いており，上位語候補は別途獲得する必要があるが，我々の手法では上位語も含めて文書構造から獲得している点が異なる．

本研究と同様に分類器を用いて上位下位関係候補の正誤を判断する手法としては，Web からタグ構造を手がかりに収集した見出し語（用語）とその説明文（見出し語を含む段落）の組を入力として，見出し語間の上位下位関係を判定する手法を大石らが提案している (大石他 2006)．彼らが説明文中に含まれる単語を素性としているのに対し，我々は上位語候補／下位語候補自体に関する情報（例えば形態素）を主に素性として用いており，それぞれの手法の素性セットはほぼ独立である．また，彼らの手法の評価はコンピュータに関する用語のシソーラスを利用して人工的に作成したテストセットでの識別性能評価に止まっており，見出し語集合から生成した上位下位関係候補の分類精度は評価できていない．さらに，彼らの手法では上位語候補／

下位語候補は説明文が獲得できている用語に限定されるため，具体物を下位語とするような上位下位関係を大量に獲得することは難しいと考えられる．

また，以上の新聞記事・Web 文書を対象に上位下位関係を獲得する手法は，十分な量の関係を獲得するために，大量の文書が扱えるストレージやそれを処理するための高速な計算機などの大規模な計算機資源が必要となる．例えば，(Sumida et al. 2006) の手法を用いた場合，約 700 GB の HTML 文書を処理して獲得できる上位下位関係の数は，約 40 万対であるが，我々の手法ではわずか 2.2 GB の Wikipedia 文書から同程度の精度で約 135 万対の上位下位関係を獲得できている（詳しくは節 5.3 の実験結果を参照のこと）．

2.2 Wikipedia からの上位下位関係獲得

Wikipedia からの上位下位関係獲得についても新聞記事や Web 文書からの上位下位関係獲得のときと同様に語彙統語パターンを用いる手法が開発されている (Ruiz-Casado et al. 2005; Herbelot and Copestake 2006; Toral and Muñoz 2006; Kazama and Torisawa 2007)．これらの手法では，Wikipedia の記事に概念の定義を記述する定義文が多く含まれることに注目し定義文から上位下位関係を獲得している．図 1（b）では「紅茶とは，摘み取った茶を乾燥させ，もみ込んで完全発酵させた茶葉。」という定義文が含まれており，紅茶の上位語（の一つ）である茶葉を用いて紅茶が説明されている．この文に対し“とは*〈上位語〉。”というパターンを適用することで紅茶の上位語である茶葉を抽出することができる．Kazama らは，英語の固有表現抽出タスクのために，Wikipedia の記事の見出し語を下位語として記事の冒頭の一文を定義文とみなし，その定義文中の特定の語彙統語パターンにマッチする表現を上位語として獲得した (Kazama and Torisawa 2007)．また Herbelot らは，Wikipedia の記事の全文を意味解析し，定義文に対応する項構造を認識することで，約 88.5%の精度で上位下位関係を獲得している (Herbelot and Copestake 2006)．Ruiz-Casado らは WordNet (Fellbaum 1998) を利用して学習された上位下位関係からパターンを学習・適用することで，69%の精度で上位下位関係が獲得できたと報告している (Ruiz-Casado et al. 2005)．これらの手法は，Wikipedia に頻出する語彙統語パターンに着目した上位下位関係獲得手法であり，前節で述べた上位下位関係手法と同様に精度が高い一方で Wikipedia の記事数と同程度の数の下位語に関する上位下位関係しか獲得できないという問題がある．

一方，Suchanek らは Wikipedia の各記事の見出し語に対し，記事に付与されたカテゴリのラベルを上位語として上位下位関係を獲得する手法を提案している (Suchanek et al. 2007)．彼らは，英語特有の経験則を用いてカテゴリを選別し，外的知識として WordNet を利用することで，約 95%と高精度で上位下位関係を獲得している．提案手法では，WordNet などの外的な言語資源を用いることなく，機械学習のみで高精度の上位下位関係を大量に獲得することを目指す．また Kazama らや Suchanek らの手法のように，下位語候補が記事の見出し語に制限されない

ため，より網羅的な上位下位関係が獲得できると期待される．

また，本研究と同様に Wikipedia の記事構造を用いた研究として (渡邉，浅原，松本 2008) が存在する．渡邉らは Wikipedia の記事構造から Wikipedia のアンカーリンク間の関係を元に条件付確率場を学習し，そのモデルを適用することでアンカーリンクから固有表現を抽出した (渡邉他 2008)．本提案手法では記事構造から直接上位下位関係を獲得するのに対し，渡邉らの手法では記事構造をアンカー間の関係が同じカテゴリか，関連語か，部分全体関係かどうかの判定に用いており，異なる手法といえる．

3 Wikipedia の記事構造

本節では提案手法について述べる前に本研究で知識源として利用する Wikipedia の記事構造について述べる．Wikipedia は，様々な事物に関する常識的知識が密に記述されたフリーの多言語百科事典である．図 1（b）は見出し語「紅茶」に対する記事の例である．Wikipedia の記事は，明確な構造をもつ MediaWiki 構文により記述されており，多段の箇条書きを含む．この例のように，Wikipedia の記事には典型的なある概念（または具体物）の辞書的な定義に加えて，関連する概念（または具体物）の列挙を箇条書きとして含むことが多い．

本稿では Wikipedia の記事から上位下位関係候補を抽出するための媒体として，MediaWiki 構文で記事のレイアウト情報を扱う表 1 の修飾記号に注目し，記事から見出し（表 1 では *title* と標記）をノードとするグラフ構造（記事構造）を抽出する．具体的には，*title* に付与されている修飾記号の優先度が高く修飾記号の繰り返し数が少ないほど，グラフ構造上の高い位置にノードを配置する．このとき，修飾記号の優先度は記号の繰り返し数より優先される．例えば，「* リプトン」より「== イギリス ==」の修飾記号の優先度が高いので，グラフ構造上で「イギリス」が「リプトン」より高い位置に配置される．また，「== イギリス ==」は「= 主な紅茶ブランド =」と比較し修飾記号（この場合は “=”）の繰り返し数が多いので，「主な紅茶ブランド」よりグラフ構造上で低い位置に配置される．ただし，ルートノードは記事名とし，その修飾記号は繰り返し数 0 の「=」とする．図 1（b）の記事に対応する図 1（a）の MediaWiki コードをもとに，図 1（c）のような記事構造が抽出できる．

表 1 記事構造に関する修飾記号

優先度	修飾記号の種類	記述方法	例
1	節見出し	=+ *title* =+	== イギリス ==
2	定義の箇条書き	;*title*:*definition*	; チャイ: ミルクティー
3	番号付き箇条書き	#+ *title*	# インド
3	番号なし箇条書き	*+ *title*	* Lipton

注：*title* は見出しを，+は直前の記号が連続して出現しうることを示す．

4　提案手法

本節では，3 節の手続きで Wikipedia の各記事から構築した記事構造を知識源として，上位下位関係を獲得する手法を提案する．提案手法は以下の 2 ステップからなる．

Step1 Wikipedia の記事構造からの上位下位関係候補の抽出　3 節で説明した記事構造に含まれるノード間の先祖—子孫関係に注目して上位下位関係候補を抽出する．

Step2 機械学習によるフィルタリング　SVM(Vapnik 1998) を用いて，Step1 で抽出された上位下位関係候補から不適切な関係を取り除く．

以下，提案手法について詳しく述べる．

4.1　Step1: Wikipedia の記事構造からの上位下位関係候補の抽出

このステップでは，記事構造の各ノードを上位語候補，子孫関係にあるノードを下位語候補とする全ての組み合わせを上位下位関係候補として抽出する．例えば，図 1（c）の記事構造からは，「ブレンドティー／チャイ」や，「紅茶／リプトン」などの上位下位関係候補が抽出できる．

ここで，訓練データの記事構造から得られる上位語候補を調べたところ，階層構造中で上位語候補に対して箇条書きで下位語候補が列挙されるときには，上位語に箇条書き特有の修飾語が付くことが分かった．このような修飾語としては，主観で一部の下位語を選んで列挙していることを示す「主な〜」や「代表的な〜」などの接頭語，箇条書きが下位語の列挙であることを陽に示す「〜のリスト」や「〜の一覧」などの接尾語などがあり，基本的に上位語を箇条書きのタイトルとするために付けられたものであるため，適切な上位語を得るためには取り除く必要がある．

そこで我々は，抽出された上位語候補が図 2 のパターンをもつ場合，パターン中の X 以外の部分を取り除いた．パターン中の X は任意の文字列を示す．ただし，複数のパターンに一致した場合には，その中で，パターンの具体的な文字列部分（ex.「代表的な X」であれば「代表的な」）が最長一致するパターンを適用した．例えば，上位語「主な紅茶ブランド」はパターン「主な X」を適用することで，「紅茶ブランド」と置換される．

このようにして得られる上位下位関係候補には，明らかに誤りとみなせる上位下位関係候補

代表 X	代表的 X	代表的な X	著名 X	著名な X	主な X
基本 X	基本的 X	基本的な X	X 一覧	X の一覧	大きな X
主要 X	主要な X	おもな X	他の X	X リスト	X のリスト
X 詳細	X の詳細	一部の X	一部 X		

図 2　上位語候補の不要な修飾語を取り除くためのパターン

上位下位関係候補が以下の条件を満たすとき，その候補を削除する．

- 上位語候補と下位語候補が完全に一致
- 以下の記号を含む

  ”，’，↑，→，↑，⇔，←，⇒

上位語候補・下位語候補が以下の不要語を含むとき，その不要語を取り除き候補を訂正する

- HTML タグ，◆，◇，■，□，◎，●，○，△，▼，▲，▽，‡，文字化け
- ＊，※，…，\，空白（先頭あるいは末尾に含まれる場合のみ）
- ＃，＋（先頭に含まれる場合のみ）
- …（末尾に含まれる場合のみ）

図 3　上位下位関係候補の削除・訂正ルール

や，上位語または下位語に記号などの不要語を含む上位下位関係候補が含まれていたため，図 3 のルールに従って上位下位関係候補を削除，あるいは訂正した．

4.2　Step2: 機械学習によるフィルタリング

Step1 の手続きで得られた上位下位関係候補は多くの適切な関係を含む一方で，「生産地／インド」，「紅茶ブランド／イギリス」のような誤りも含む．Step2 では，Step1 で抽出した上位下位関係候補から教師あり機械学習を用い不適切な関係を取り除く．本稿では上位下位関係候補が適切な上位下位関係か否かを判定するため，Support Vector Machine (SVM)(Vapnik 1998) で学習された分類器を用いて上位下位関係候補を選別する．

SVM で各上位下位関係候補（上位語候補—下位語候補のペア）が適切な上位下位関係であるかどうかを判定するには，分類対象の上位下位関係候補を，素性ベクトルと呼ばれる分類対象の特徴（素性）を数値で表現したベクトルに変換する必要がある．この素性ベクトル（上位下位関係候補）に正解（適切な上位下位関係か否か）をつけたものを学習データとして，Step2 で用いる分類器 (SVM) を得る．

本研究では素性として，上位下位関係候補がある条件（特徴）を満たすかどうかを一つの素性として表現し，素性ごとに設定された条件を入力の上位下位関係候補が満たせば，対応する素性ベクトルの次元の値に 1 をセットし，満たさなければ 0 をセットする．実際に使用した素性をまとめたリストを表 2 に示す．表の各列は左から素性の種類，各素性に対応する素性ベクトルの次元の値を 1 にセットする条件，図 1 から抽出した上位下位関係候補「紅茶ブランド／Lipton」で実際に 1 にセットされる素性を表している． ただし，同じ表現の上位下位関係候補が異なる記事構造から抽出された場合，全ての抽出元の記事構造について生成した素性ベクトルの論理和を用いる．次に生成した素性ベクトルを SVM に入力し，その結果得られた SVM の

表 2　素性リスト

素性の種類	各素性に対応する素性ベクトルの次元の値を 1 にセットする条件	例)「紅茶ブランド／ Lipton」で 1 にセットされる素性
POS	上位語候補の末尾の形態素以外に付与された品詞が X	$\surd(X =$ ‘名詞—一般’)
	下位語候補の末尾の形態素以外に付与された品詞が X	
	上位語候補の末尾の形態素に付与された品詞が X	$\surd(X =$ ‘名詞—一般’)
	下位語候補の末尾の形態素に付与された品詞が X	$\surd(X =$ ‘名詞—固有名詞’)
MORPH	上位語候補の末尾以外の形態素が X	$\surd(X =$ ‘紅茶’)
	下位語候補の末尾以外の形態素が X	
	上位語候補の末尾の形態素が X	$\surd(X =$ ‘ブランド’)
	下位語候補の末尾の形態素が X	$\surd(X =$ ‘Lipton’)
EXP	上位語候補が X	$\surd(X =$ ‘紅茶ブランド’)
	下位語候補が X	$\surd(X =$ ‘Lipton’)
ATTR	上位語候補が属性 X に一致	
	下位語候補が属性 X に一致	
LAYER	上位語候補を修飾していた修飾記号の種類が X	$\surd(X =$ ‘=’)
	下位語候補を修飾していた修飾記号の種類が X	$\surd(X =$ ‘*’)
DIST	上位語候補と下位語候補間の距離が 2 以上である	$\surd$
	上位語候補と下位語候補間の距離が 1 である	
PAT	Step1 で上位語候補が図 2 のパターンのいずれかに一致	$\surd$
LCHAR	上位語候補と下位語候補の末尾の 1 文字が一致	

スコアが閾値以上の上位下位関係候補を正しい上位下位関係とみなす．以下で，各素性の設計方針について説明する．

POS　まず上位語候補・下位語候補の品詞は，誤りの判定に有効である．例えば，「木次線／管轄」のように上位語に固有名詞を含み，下位語に固有名詞を含まない場合，誤りの関係と推定できる．ここでは，品詞として IPA 辞書[1]の品詞細分類レベル（ex. 名詞—固有名詞など）まで考慮する．

また上位語候補・下位語候補に含まれる品詞のうち，主辞の品詞は語の意味的な特徴をよく捉えているため特に重要である．例えば，上位語候補の主辞の品詞が動詞であれば多くの場合その上位下位関係候補は誤りである．本稿では上位語候補・下位語候補の末尾の形態素を主辞とし，主辞の品詞を他の品詞と区別するように素性を設計した．

MORPH　品詞と同様に，上位語候補・下位語候補中の形態素の表層文字列は上位下位関係らしさの判定に有効である．例えば，「アメリカ映画／ウエスト・サイド物語」のように頻度が少ない，あるいは未知の上位語候補・下位語候補であっても，「映画」や「物語」などのより頻度が高い形態素に注目することで上位語らしさ・下位語らしさを判定するこ

[1] http://sourceforge.jp/projects/ipadic/

とが出来る．また品詞と同様に上位語候補・下位語候補の主辞の表層文字列は適切な上位下位関係であるかどうかの手がかりとなりやすいので，他の形態素と区別する．

EXP 上位語候補，下位語候補にはStep1の不要語処理ではカバーしきれない，「背景」や「あ行」などの不要語が多く存在する．これら不要語の特徴を捉えるため，上位語候補，下位語候補の表層文字列ごとに次元を割り当てるように素性を設計した．

ATTR 上位語候補，あるいは下位語候補が属性語である上位下位関係候補は誤りの関係となりやすい．ここで属性語とは，その単語についてユーザが知りたい観点を指す単語である (徳永，風間，鳥澤 2006)．例えば，「紅茶」の属性語としては「生産量」や「価格」があげられる．このような属性語を含む関係（例えば，「紅茶／生産量」や「生産量／1位インド」など）は多くの場合，属性語と概念（または具体物）間の関係となり上位下位関係となることは少ない．そこでこの素性ではあらかじめ抽出しておいた属性語リストの各語に固有の次元を割りあてるように設計した．

本研究では，属性語は以下のような手順で抽出した．まず各記事構造から根ノード以外のノードを抽出する．つぎに，抽出したノードのうち，Wikipedia中の複数の記事に出現するノードを属性とみなす．例えば「紅茶」と「タバコ」という記事の両方に「生産量」が見出しとして出現する場合，「生産量」を属性語とみなす．前述の上位語候補・下位語候補の表層文字列を素性とする素性EXPもこの素性と同じく不要語らしさを扱うことができるがこの素性では教師無しで構築された属性語リストを用いることで，より被覆率高く不要語を検出することが可能であることに注意されたい．

LAYER 記事構造の箇条書き表現から抽出された下位語候補をもつ上位下位関係は適切な関係になりやすい．例えば，図1（c）の記事構造の箇条書き表現には「Lipton」，「Wedgwood」などの固有名詞が列挙されており，これらは上位ノード「紅茶ブランド」の下位語として適切である．このような傾向を捉えるために，この素性では記事構造から抽出された上位語候補あるいは下位語候補のノードに付与されている修飾記号の種類（節見出し，定義の箇条書き，番号付き箇条書き，番号なし箇条書き）ごとに次元を割りあてた．

DIST 記事構造で上位語候補と下位語候補との間の距離が近ければ近いほど，正しい上位下位関係であることが多い．そこで，記事構造中における上位語候補・下位語候補間の距離を素性とすることで，この傾向を捉える．本稿では，上位語候補，下位語候補間の距離を記事構造中で上位語候補と下位語候補間に存在する辺の数とする．例えば，図1（c）の記事構造上で「Wedgwood」と「紅茶ブランド」間の距離は2である．素性DISTでは，上位語候補と下位語候補間の距離が2以上か否かという2つの状態にそれぞれ異なる次元を割りあてた．

PAT 上位語候補がStep1の時点で図2のパターンにマッチしていた場合，子孫ノードに適切な下位語が列挙されやすい傾向がある．例えば，図1（c）中の「主な紅茶ブランド」とい

うノードは下位階層に「Lipton」,「Wedgwood」などの適切な下位語が列挙されており，上位語が Step1 のパターンにマッチしていれば，その上位下位関係は適切だろうと推定できる．素性 PAT では，Step1 の時点で上位語候補がパターンにマッチしている場合この素性に対応する素性ベクトルの次元の値を 1 にセットするように設計した．

LCHAR　素性 MORPH では，形態素間の類似性を判断しているため,「高校」や「公立校」のように形態素の一部が一致する語の類似性はないと判断してしまう欠点が存在する．そこで上記のような事例を扱えるようにするため，素性 LCHAR では，上位語候補と下位語候補の末尾の 1 文字が共通する複合語に意味的に似た語が多い特徴を利用し，素性 MORPH の欠点を補う．具体的には，上位語候補と下位語候補の末尾が同じとき，この素性に対応する素性ベクトルの次元の値を 1 にセットするように設計した．

5　実験

5.1　実験設定

提案手法の有効性を評価するため，2007 年 3 月の日本語版 Wikipedia から Wikipedia 内部向けの記事を取り除いた 276,323 記事に対して，提案手法を適用した．Wikipedia 内部向けのページは，ユーザページ，特別ページ，テンプレート，リダイレクション，カテゴリ，曖昧さ回避ページを指すものとする．本稿では，形態素解析に MeCab[2] を利用しその辞書として IPA 辞書[3] を用いた．SVM には TinySVM[4] を利用した．SVM のカーネルには予備実験結果から 2 次の多項式カーネルを用いた．また Wikipedia から Step2 で必要となる属性語リストを抽出した結果，40,733 個の属性語が獲得でき，ランダムに取り出した 200 語を (徳永他 2006) の厳密な属性語を判定するための基準に従い評価したところ，精度は 73.0%だった．

まず Wikipedia の記事に Step1 を適用し，記事構造から重複を除いて 6,564,317 対の上位下位関係候補を獲得した．以降に示す全ての上位下位関係数は重複を除いた数を示す．次に，得られた上位下位関係候補からランダムに 1,000 対取り出してテストデータとした．続いて，テスト用データを除いた上位下位関係候補からランダムに 9,000 対，抽出元の記事構造中で上位語と下位語が直接の親子関係にあった候補から 9,000 対，図 2 のパターンにマッチしていた上位下位関係候補から 10,000 対，図 2 のパターンにマッチしなかった上位下位関係候補から 2,000 対をそれぞれランダムに取り出し，人手で正解をつけた．これらから重複を除いて得られた 29,900 対を訓練データとして用いた．訓練データのうち 19,476 対は，素性を決定するための予備実験に利用した．上位下位関係の正解付けは，Miller ら (Fellbaum 1998) の基準に従い 1 名で行っ

[2] http://mecab.sourceforge.net/
[3] http://chasen.naist.jp/chasen/distribution.html.ja
[4] http://chasen.org/~taku/software/TinySVM/

た．具体的には，各上位下位関係候補（上位語候補の表現 A と下位語候補の表現 B のペア）について，「B は A の一種あるいは 1 つである」という文が適切であるとき正解とした．

5.2　比較手法

提案手法の有効性を確認するため，2 節で説明した既存の語彙統語パターンに基づく上位下位関係獲得手法 (Sumida et al. 2006)，および既存の Wikipedia からの上位下位関係獲得手法 (Kazama and Torisawa 2007; Suchanek et al. 2007) と比較を行う．Wikipedia からの上位下位関係の獲得手法は，英語版 Wikipedia に特化したものであるため，以下で日本語版 Wikipedia に応用する際に変更した点を記載する．

Wikipedia の定義文からの上位下位関係の獲得

Kazama らの手法は英語版 Wikipedia のための手法であるため，ここでは国語辞書の語釈文から上位下位関係を獲得した Tsurumaru らの手法を参考に人手で図 4 のような語彙統語パターンを 1,334 パターン用意した (Tsurumaru, Hitaka, and Yoshida 1986; Kazama and Torisawa 2007)．

〈上位語〉 [〈数字〉+]	〈上位語〉ので、
〈上位語〉の{で|でも}あり、	〈上位語〉の事で、
〈上位語〉の事{で|でも}あり、	〈上位語〉の一つの{で|でも}あり、
〈上位語〉の一つの事で、	〈上位語〉の一つの事{で|でも}あり、
〈上位語〉のうち一つの{で|でも}あり、	〈上位語〉のうち一つの事で、
〈上位語〉のうち一つの事{で|でも}あり、	〈上位語〉の{で|でも}ある。
〈上位語〉の一つの{で|でも}ある。	〈上位語〉のうち一つの{で|でも}ある。
〈上位語〉のうち*もの{。|、}	〈上位語〉の代表的なもの{。|、}
〈上位語〉の代表格{。|、}	〈上位語〉を意味する{、|。}
〈上位語〉の意{、|。}	〈上位語〉と言う{、|。}
〈上位語〉に由来する{、|。}	〈上位語〉を意味する言葉{、|。}
〈上位語〉と言う言葉{、|。}	〈上位語〉に由来する言葉{、|。}
〈上位語〉を意味し、	〈上位語〉の意で、
〈上位語〉と呼び、	〈上位語〉に由来し、
〈上位語〉の一種の名称{、|。}	〈上位語〉の一種{、|。}
〈上位語〉の名称{、|。}	〈上位語〉の第〈漢数字〉番{、|。}
〈上位語〉第〈漢数字〉番{、|。}	〈上位語〉の第〈漢数字〉番目{、|。}
〈上位語〉第〈漢数字〉番目{、|。}	

注: +は直前の文字列が連続して出現しうることを，$\{A|B\}$ は A，B のいずれかが出現することを示す．

図 4　定義文に適用する語彙統語パターンの一例

図中の〈上位語〉は任意の名詞の連続，〈数字〉は 0〜9 までの数字の連続，〈漢数字〉は〇〜九などの漢数字の連続を示す．このパターンを定義文に適用することで見出し語を下位語，パターンで認識された〈上位語〉を上位語とする上位下位関係を獲得する．

Wikipedia のカテゴリからの上位下位関係の抽出

Suchanek らの手法に従い，各記事に付与されているカテゴリを上位語候補，記事の見出し語を下位語候補として上位下位関係候補を獲得する．例えば，図 1（b）の記事からは，「茶／紅茶」という上位下位関係候補が得られる．Wikipedia のカテゴリから獲得できる関係には上位下位関係以外に「喫茶文化／紅茶」などのように見出し語とその関連語間の関係も多く含まれる．Suchanek らの手法では，英語による経験則を用いて，さらに獲得した関係を選別しているが，日本語には適用できないため，ここではカテゴリから抽出できた全ての関係候補を上位下位関係とみなす．

5.3　実験結果

表 3 に提案手法と節 5.2 に述べた比較手法と (Sumida et al. 2006) の手法を比較した結果を示す．表 3 の各列は左から順に手法の種類，リソース，SVM の閾値，精度，SVM により選別された上位下位関係数，およびこれらより求めた期待される正しい上位下位関係の数を示す．ここでは以下のような評価尺度を用いた．

$$\text{精度 (Precision)} = \frac{\text{SVM により選別された上位下位関係のうち正解の関係の数}}{\text{SVM により選別された上位下位関係数}}$$

$$\text{再現率 (Recall)} = \frac{\text{SVM により選別された上位下位関係のうち正解の関係の数}}{\text{評価データ中に存在する正しい上位下位関係数}}$$

$$\text{正解率 (Accuracy)} = \frac{\text{SVM により正しく正例・負例を識別できた関係の数}}{\text{評価データ中の関係候補の数}}$$

$$\text{期待される正しい上位下位関係数} = \text{抽出できた上位下位関係数} \times \text{精度}$$

表 3　提案手法と比較手法により獲得した上位下位関係の比較

獲得手法		リソース（サイズ）	SVM の閾値	精度	上位下位関係数	期待される正しい上位下位関係数
(Sumida et al. 2006)		Web (700 GB)	—	89.0%	398,770	354,905
定義文		Wikipedia (2.2 GB)	—	89.4%	158,177	141,410
カテゴリ		Wikipedia (2.2 GB)	—	70.5%	596,463	420,506
提案手法	(Step1)	Wikipedia (2.2 GB)	—	28.4%	6,564,317	1,864,266
	(Step2)	Wikipedia (2.2 GB)	0.0	85.2%	1,738,500	1,481,400
	(Step2)	Wikipedia (2.2 GB)	0.36	**90.0%**	1,349,622	1,214,659

比較手法カテゴリ，定義文は 5.2 節で記述した手法を用い，提案手法で利用した Wikipedia と同じデータを利用し，評価サンプル数は 1,000 対である．また比較手法 (Sumida et al. 2006) は，Web から無作為に収集した約 700 GB（HTML タグ含む）の Web 文書に (Sumida et al. 2006) を適用した結果を示し，評価サンプル数は 200 対である．表より提案手法は Wikipedia を入力源とする手法と比較し，大量の上位下位関係を獲得することに成功した．また，提案手法と比較手法 (Sumida et al. 2006) を比べると，提案手法は小さなリソース（2.2 GB の XML 文書）から上位下位関係を抽出したにもかかわらず，より大量の上位下位関係（933,782 語を含む約 174 万対）を獲得できた．

獲得される上位下位関係の精度については，SVM の分類時の閾値を変更することであげることが可能である．精度と再現率とのトレードオフの関係を図 5 に示す．横軸は再現率，縦軸は精度を表す．このグラフより，SVM の閾値を大きくすることで，より信頼性の高い上位下位関係を獲得できることが確認できる．例えば閾値を 0.36 にすると，テストデータでの精度は 90%まで向上する（表 3）．この精度でも，他の比較手法より獲得できた上位下位関係は多く，またこの関係に含まれる語数は 774,135 語であった．

次に Step2 で利用した素性の効果を調べるために，各素性を除いたときの精度の比較を表 4 に示す．表 4 の各列は左から順に素性の種類，正解率，精度，再現率，F 値を表す．またこのときの精度と再現率とのトレードオフの関係を図 6 に示す．各素性は本稿で提案した全ての素性を含む素性セットを ALL，ALL から素性 X を除いた素性セットを ALL-X とした．また（）内は素性セット ALL-X の精度から素性セット ALL の精度を引いた結果であり，この値が低ければ低いほど，素性 X が提案手法の性能の向上に役立っていることを意味する．これらの結果より，全ての素性が Step2 のフィルタリング性能の向上に役立っていることが確認できた．また

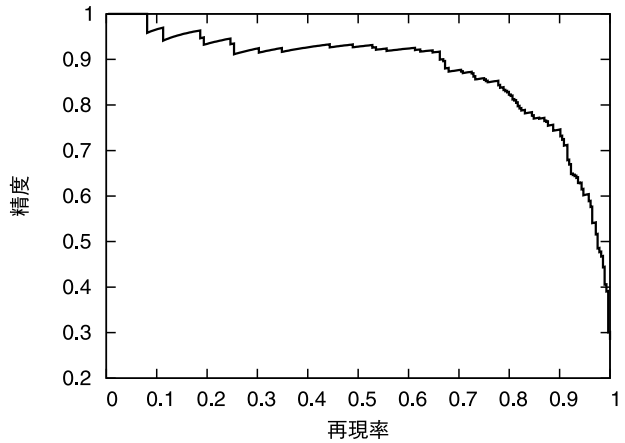


図 5　精度と再現率とのトレードオフ

表 4　各素性による効果

素性の種類	正解率		精度		再現率		F 値	
ALL-POS	89.0%	(−0.7)	83.7%	(−1.5)	76.1%	(−1.0)	79.7%	(−1.3)
ALL-MORPH	88.2%	(**−1.5**)	81.2%	(**−4.0**)	76.1%	(−1.0)	78.5%	(**−2.5**)
ALL-EXP	89.3%	(−0.4)	83.9%	(−1.3)	77.1%	(0.0)	80.4%	(−0.6)
ALL-ATTR	89.5%	(−0.2)	84.6%	(−0.6)	77.1%	(0.0)	80.7%	(−0.3)
ALL-LAYER	88.6%	(−1.1)	82.9%	(−2.3)	75.4%	(−1.7)	79.0%	(−2.0)
ALL-DIST	89.3%	(−0.4)	83.9%	(−1.3)	77.1%	(0.0)	80.4%	(−0.6)
ALL-PAT	89.5%	(−0.2)	83.8%	(−1.4)	78.2%	(1.1)	80.9%	(−0.1)
ALL-LCHAR	88.9%	(−0.8)	85.0%	(−0.2)	73.9%	(**−3.2**)	79.1%	(−1.9)
ALL	89.7%		85.2%		77.1%		81.0%	

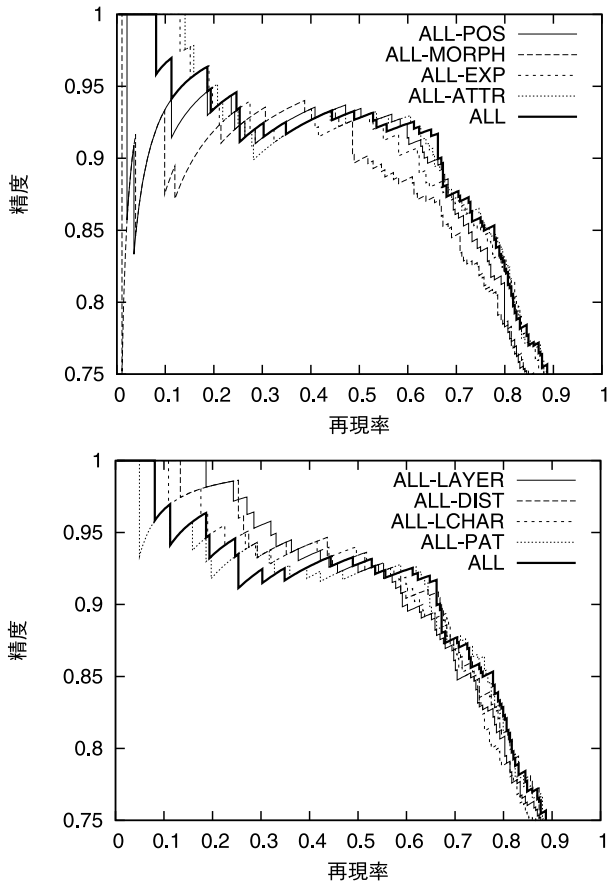
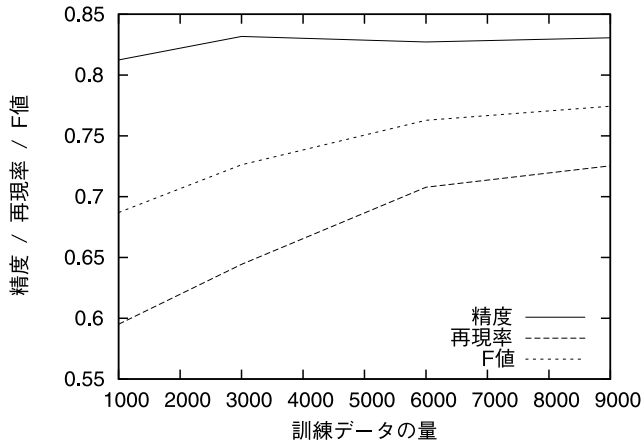


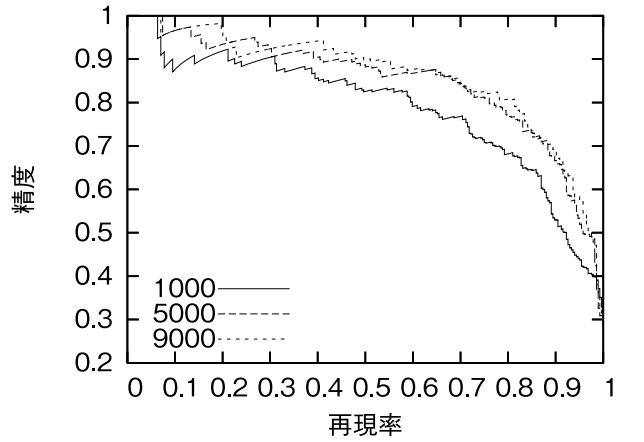
図 6　再現率—精度グラフによる素性の比較

表より全ての素性が精度の向上に寄与しており，特に素性 MORPH による効果が大きいことがわかった．一方，再現率の向上には素性 POS，MORPH，LAYER，LCHAR が寄与しており，特に素性 LCHAR が最も高い効果をもつことがわかった．

つづいて，訓練データの量を変化させたときの提案手法の性能の変化を調べた．訓練データは Step1 の結果からランダムに抽出した 9,000 対を利用し，1,000 対から 9,000 対まで 3,000 対ごとに評価を行った．その結果を図 7（a），（b）に示す．図 7（a）は SVM の分類時の閾値を 0 に固定したグラフで，横軸は訓練データの量，縦軸は精度，再現率，F 値を示す．また図 7（b）は SVM 分類時の閾値を変化させ精度と再現率のトレードオフを調べたグラフで，横軸は再現率，縦軸は精度を示す．図 7（a）より訓練データの量を増やすことで再現率の性能が向上する



（a） 訓練データの量の変化に伴う精度，再現率，F 値の変化



（b） 訓練データの量の変化に伴う再現率と精度のトレードオフの変化

図 7　訓練データの量による性能の変化

傾向がわかった．また図 7（b）より，SVM の閾値を変化させた場合でも，訓練データのサイズを増やすことで，性能が向上する傾向にあることが確認できた．この結果は訓練データをさらに増やせば提案手法の性能がさらに向上する可能性を示唆している．

5.4 考察

提案手法により得られた上位下位関係の例を表 5 に示す．ここでは，人手で選んだ 25 語についてランダムに 10 対の下位語を選択した．表中の * は上位下位関係が誤りの例を，# は小説や映画などのフィクション上でなりたつ架空の上位下位関係を示す． このような架空の上位語，あるいは下位語は，フィクション自体に関する記述（感想）や，比喩表現として日常的に用いられることも多いため，本稿ではそれ以外の上位下位関係と特に区別せず，有用な上位下位関係知識とみなし Miller ら (Fellbaum 1998) の基準で正解か誤りかを判断した．これらの架空の上位下位関係とそうでない上位下位関係を識別することは今後の課題の一つである． また表の各列は左から人手で選んだ上位語とその下位語の例を示している．この例より，ほとんどの上位下位関係は，上位語ごとに多少の精度の偏りがみられるものの正しく認識できていることが確認できる．

最後に，提案手法の性能を悪化させている原因を探るべく，SVM 分類器により誤りとされた上位下位関係候補を人手で分析した．テストデータ，訓練データ以外の上位下位関係候補からランダムに 1,000 対抽出し，人手で評価した．誤り分析用データに提案手法を適用した結果，その精度は 89.1%であり，この内訳は内訳は陽性が 233 対，陰性が 658 対，偽陽性が 22 対，偽陰性が 87 対であった．

表 6 に偽陽性の分類結果を示す．表の各列は，左から分類の種類，数，SVM スコアの平均，例を示す．この結果から，部分全体関係が最も頻出する誤りであるうえに，SVM スコアの平均から最も除去しにくい誤りであることがわかった．このような誤りを取り除くことは今後の課題である．また，精度を低下させる原因として，属性・属性値と facet を含む関係を上位下位関係と誤判定する問題が多いことも分かった．ここでいう facet とはインスタンスを分類するための属性の値である．例えば，図 1（c）の記事構造中の「主な紅茶ブランド」と「Wedgwood」との間に挿入されている「イギリス」は「Wedgwood」などのブランドを国別で分類するための facet であるといえる．提案手法では自動抽出した属性リストを用いてこのような誤りの除去を試みたが，表 6 より提案手法の対策だけでは不十分であり，新たに記事構造中の他のノードの情報を素性とするなど改善が必要であることがわかった．また「プロレス技／代表的な技」のように，素性 LCHAR が悪影響を及ぼしていると思われる例も存在した．

つづいて，陽性と偽陰性と判定された関係の上位語が訓練データ中に存在したか否かを調査した．陽性では 66.6%，偽陰性では 16.7%の上位語が訓練データ中に存在していることがわかった．未知の上位語であっても正しく判定できるようにするために，より上位の語や同義語の利

表 5 獲得した上位下位関係の例

上位語	下位語の例
湖	カリバ湖，ナセル湖，ツーク湖，ヌーシャテル湖，丹沢湖，シャスタ湖，ユタ湖，ダル湖／ダール湖，イシク湖，ウィンドメア湖
惑星	アスト IV#，アナサジ #，ドドー #，カタリナ #，天王星，ラロス #，パッサ #，ムトラル #，フリーザム #，ファルランド #
公園	中丸緑地，鹿島・扇平自然公園，元宮公園，南八幡宮児童遊園，諏訪ヶ原公園，香里ヶ丘西公園，堂山公園，牧野公園，かりん緑地，第八公園
公共施設	老人福祉センター，福祉施設，都立墨東病院，バグダード国際空港，仁保新町公園，稚内市総合文化センター，公立陶生病院，三島市民文化会館，泉崎村さつき公園，広島県警察学校
航空会社	ビーマン・バングラデシュ航空，シルク航空，タイ国際航空，ポリネシアン航空，エア・サイアム，エールリネール，新疆航空，琉球エアーコミューター，アシアナ航空，ノースウエスト航空
猟犬	前田犬，秋田犬，越路犬，赤城犬，琉球犬，レトリーバー，高安犬，ゴールデン・レトリーバー，ハウンド，薩摩犬
サクラ	エリザベス・サクラ・マツシタ *，オシドリザクラ，ヒウチダニキクザクラ，ウズザクラ，ニッコウザクラ，コトヒラ，ヤエノオオシマザクラ，シラタキザクラ，ショウドウザクラ，クシマザクラ
戦争映画作品	ホワイト・バッジ，ローレライ，ムルデカ，SHOAH ショアー，パール・ハーバー，マーフィの戦い，モスクワ大攻防戦，零戦燃ゆ，あゝ同期の桜，眼下の敵
民族楽器	クレタのリラ，アゴゴ，クラベス，ウード，高胡，二胡，馬頭琴，パンパイプ，ギロ，ボンゴ
文房具	のり，修正テープ，付箋，印章，輪ゴム，鉛筆，画鋲・虫ピン，綴じ具，画板，カッティングマット
工具	ロッキングプライヤ，ウォーターポンププライヤ，油圧工具，電動工具，ラチェットレンチ，研削工具，バイス，スナップリングプライヤ，振動・ハンマードリル，メタルソー
アジア系民族	マレー人，アイヌ民族，タイ人，ウズベク人，アラブ人，ニヴフ民族，漢民族，朝鮮民族，カザフ人，トルクメン人
彫刻家	オーギュスト・ロダン，鈴木実，平櫛田中，瀧口政満，イサム・ノグチ，高田博厚，佐藤忠良，ジャン・ティンゲリー，高芙蓉，雨宮敬子
学校行事	球技大会，卒業式，夏季休業中 *，学園祭，芸術鑑賞会，推薦・学業特待入学試験，クレメンティ校 *，学園祭，野外実習，応援合戦，学芸会，生徒会文化的行事
技	三角蹴り，炎戒 #，虎牙連斬 #，ネックブリーカー，月光 #，バリヤーガス #，リバースバイパー・ホールド，ラルフキック #，龍槌翔閃 #，エレクトリッガー #
スポーツ競技	混合競技，モーグル，フィギアスキー，トライアスロン，フットボール，ドラゴンボート，バスク・ペロタ，ボウリング，ライフル射撃，ワンダーフォーゲル
料理	紅白かまぼこ，黒臭豆腐，鶏肉ハム，コース料理，ボイシェル，カキご飯，オーストリア料理，ムリンチー，コスタリカ料理，チキン南蛮

表 6　偽陽性の分類結果

分類	数	SVM スコアの平均	例
部分全体関係	7	0.625	松下家／松下響子
概念—facet	5	0.214	私設応援団／浦和レッドダイヤモンズ
語—属性値	4	0.171	スタジオイースター／うた∽かた
文末一致	3	0.161	プロレス技／代表的な技
facet—語	1	0.095	ラトロア／ジェラルド・メイスン
その他	2	0.113	趣味・思考・特技／大塚に
計	22	0.315	

表 7　陰性の分類結果

分類	数
語—属性値	229
概念—属性	168
facet—語	64
部分全体関係	45
概念—facet	15
属性—属性値	15
属性—facet	2
その他	120
計	658

用を考えている．

最後に，表 7 に陰性を人手で分類した結果を示す．表の各列は左から誤りの分類，数を示す．ここでは，上位下位関係以外の何らかの概念間関係に分類できるかどうかに注目して分類した．この結果，約 80%については何らかの概念間関係になっていることが分かった．これらについては，正しく分類できれば語彙知識として有用である．本稿では上位下位関係に注目し，二値分類の分類器を用いたが，適切な関係に分類する多値分類を構築することで，Wikipedia の記事構造を余すことなく，語彙知識に変換することができそうである．

6　まとめ

本稿では，Wikipedia の記事構造を知識源とした上位下位関係獲得手法を提案した．提案手法は，「Wikipedia の記事構造中のノード間の関係は多くの上位下位関係を含む」という仮定と機械学習を併用することにより，約 135 万対の上位下位関係を精度 90%で獲得することに成功した．本稿では 2007 年 3 月時点での Wikipedia から上位下位関係を獲得したが，Wikipedia は現在も成長を続けており提案手法を最新の Wikipedia データに適用することでさらに多くの上

位下位関係を獲得することも可能であると考えられる．

実験結果より，Wikipedia の記事構造は上位下位関係だけでなく，属性—属性値の関係，部分全体関係などの記述にも頻繁に使われていることがわかった．今後の課題として，上位下位関係だけでなく部分全体関係や属性—属性値の関係を獲得したいと考えている．また 5 節で述べたように獲得した上位下位関係には，フィクションの世界でのみ成り立つ架空の上位下位関係が含まれている．これらの架空の世界でのみ成り立つ上位下位関係を識別することは今後の課題である． 更に，Wikipedia の記事には他の言語で記述された記事へのリンクが執筆者によって付与されており，これらのリンクを利用して様々な言語の上位下位関係を獲得することも考えている．

参考文献

Bunescu, R. C. and Paşca, M. (2006). "Using encyclopedic knowledge for named entity disambiguation." In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 9–16.

Etzioni, O., Cafarella, M., Downey, D., Popescu, A.-M., Shaked, T., Soderland, S., Weld, D. S., and Yates, A. (2005). "Unsupervised named-entity extraction from the web: An experimental study." *Artificial Intelligence*, **165** (1), pp. 91–134.

Fellbaum, C. (Ed.) (1998). *WordNet: An electronic lexical database*. MIT Press.

Fleischman, M., Hovy, E., and Echihabi, A. (2003). "Offline strategies for online question answering: Answering questions before they are asked." In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, pp. 1–7.

Hearst, M. A. (1992). "Automatic acquisition of hyponyms from large text corpora." In *Proceedings of the 14th International Conference on Computational Linguistics*, pp. 539–545.

Herbelot, A. and Copestake, A. (2006). "Acquiring ontological relationships from Wikipedia using RMRS." In *Proceedings of Web Content Mining with Human Language Technologies workshop on the fifth International Semantic Web Conference*.

Kazama, J. and Torisawa, K. (2007). "Exploiting Wikipedia as external knowledge for named entity recognition." In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 698–707.

Pantel, P. and Pennacchiotti, M. (2006). "Espresso: Leveraging generic patterns for automatically harvesting semantic relations." In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, pp. 113–120.

Ruiz-Casado, M., Alfonseca, E., and Castells, P. (2005). "Automatic extraction of semantic relationships for WordNet by means of pattern learning from Wikipedia." In *Proceedings of the 10th International Conference on Applications of Natural Language to Information Systems*, pp. 67–79.

Shinzato, K. and Torisawa, K. (2004). "Acquiring hyponymy relations from web documents." In *Proceedings of the 20th International Conference on Computational Linguistics*, pp. 73–80.

Suchanek, F. M., Kasneci, G., and Weikum, G. (2007). "YAGO: A core of semantic knowledge unifying WordNet and Wikipedia." In *Proceedings of the 16th International World Wide Web Conference*.

Sumida, A., Torisawa, K., and Shinzato, K. (2006). "Concept-instance relation extraction from simple noun sequences using a search engine on a web repository." In *Proceedings of the Web Content Mining with Human Language Technologies workshop on the fifth International Semantic Web Conference*.

Toral, A. and Muñoz, R. (2006). "A proposal to automatically build and maintain gazetteers for named entity recognition by using Wikipedia." In *Proceedings of Workshop on New Text held at the 11th Conference of the European Chapter of the Association for Computational Linguistics*.

Tsurumaru, H., Hitaka, T., and Yoshida, S. (1986). "An attempt to automatic thesaurus construction from an ordinary Japanese language dictionary." In *Proceedings of the 11th International Conference on Computational Linguistics*, pp. 445–447.

Vapnik, V. N. (1998). *Statistical Learning Theory*. Wiley-Interscience.

安藤まや，関根聡，石崎俊 (2003). "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.

今角恭祐 (2001). "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." Master's thesis, 九州工業大学.

大石康智，伊藤克亘，武田一哉，藤井敦 (2006). "単語の共起関係と構文情報を利用した単語階層関係の統計的自動識別." 情報処理学会研究報告 2006-SLP-61, pp. 25–30.

徳永耕亮，風間淳一，鳥澤健太郎 (2006). "属性語の Web 文書からの自動獲得と人手評価のための基準." 自然言語処理, **13** (4), pp. 49–67.

鳥澤健太郎，隅田飛鳥，野口大輔，風間淳一 (2008). "自動生成された検索ディレクトリ「鳥式」の現状." 言語処理学会第 14 回年次大会 発表論文集, pp. 729–732.

渡邉陽太郎，浅原正幸，松本裕治 (2008). "グラフ構造を持つ条件付確率場による Wikipedia 文書中の固有表現分類." 人工知能学会論文誌, **23** (4), pp. 245–254.

略歴

隅田　飛鳥：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終了．現在，同大学情報科学研究科後期課程在学中．修士（情報科学）．自然言語処理の研究に従事．

吉永　直樹：2005 年東京大学大学院情報理工学系研究科博士課程修了．2002 年より 2008 年まで日本学術振興会特別研究員 (DC1，PD)．2008 年 4 月より東京大学生産技術研究所特任助教．博士（情報理工学）．自然言語処理の研究に従事．

鳥澤健太郎：1995 年東京大学大学院理学系研究科情報科学専攻博士課程中退，同年同専攻助手．北陸先端科学技術大学院大学助教授を経て，2008 年より（独）情報通信研究機構・MASTAR プロジェクト・言語基盤グループ・グループリーダー．博士（理学）．自然言語処理の研究に従事．

（2008 年 10 月 16 日　受付）
（2008 年 12 月 28 日　再受付）
（2009 年 1 月 28 日　採録）