Wikipedia を利用した上位下位関係の詳細化

山田 一郎[†]・橋本 力^{††}・呉 鍾勲^{††}・鳥澤健太郎^{††}・黒田 航^{†††,††††}・
Stijn De Saeger^{††}・土田 正明^{††††}・風間 淳一^{††}

単語の上位下位関係を自動獲得する研究はこれまで活発に行われてきたが、上位概念の詳細さに関する議論はほとんどなされてこなかった。自動獲得された上位下位関係の中には、例えば「作品 → 七人の侍」や「作品 → 1Q84」のように、より適切と考えられる上位概念「映画」や「小説」と比べて広範囲な概念をカバーする上位概念（「作品」）が含まれることがある。このような上位概念を検索や質問応答などのタスクにおいて利用すると、より詳細な上位概念を利用する手法と比較して有用でないことが多い。そこで本論文では、自動獲得した上位下位関係を、Wikipedia の情報を利用することでより詳細にする手法を提案する。例えば「作品 → 七人の侍」から、「作品 → 映画監督の作品 → 黒澤明の作品 → 七人の侍」のように、単語「七人の侍」の上位概念（かつ、単語「作品」の下位概念）として、2 種類の間ノード「黒澤明の作品」、「映画監督の作品」を生成することにより、元の上位下位関係を詳細化する。自動獲得した 1,925,676 ペアの上位下位関係を対象とした実験では、最も詳細な上位概念となる一つ目の中間ノード（「黒澤明の作品」など）を重み付き適合率 85.3% で 2,719,441 個、二つ目の中間ノード（「映画監督の作品」など）を重み付き適合率 78.6% で 6,347,472 個生成し、高精度に上位下位関係を詳細化できることを確認した。さらに、生成した上位下位関係が「対象 - 属性 - 属性値」として解釈できることについても報告する。

キーワード：上位下位関係獲得、「対象-属性-属性値」抽出、Wikipedia

Generating Information-Rich Taxonomy Using Wikipedia

ICHIRO YAMADA[†], CHIKARA HASHIMOTO^{††}, JONG-HOON OH^{††}, KENTARO TORISAWA^{††},
KOW KURODA^{†††,††††}, STIJN DE SAEGER^{††}, MASA-AKI TSUCHIDA^{††††} and JUN'ICHI KAZAMA^{††}

Hyponymy relation acquisition has been extensively studied. However, the informativeness of acquired hypernyms has not been sufficiently discussed. We found that the hypernyms in automatically acquired hyponymy relations are often too vague for their hyponyms. For instance, “work” is a vague hypernym for “work→Seven Samurai” and “work→1Q84”. These vague hypernyms sometimes cause the lower accuracy for NLP applications such as information retrieval or question answering. In this paper, we propose a method of making (vague) hypernyms more specific ex-

[†] NHK 放送技術研究所, Science & Technology Research Laboratory, Japan Broadcasting Corporation

^{††} 情報通信研究機構, National Institute of Information and Communications Technology

^{†††} 京都大学, Kyoto University

^{††††} 早稲田大学総合研究機構, Comprehensive Research Organization, Waseda University

^{†††††} 日本電気株式会社, NEC Corporation

ploting Wikipedia. For instance, our method generates two intermediate nodes “work by Akira Kurosawa” and “work by film director” for a original hyponymy relation “work→Seven Samurai”. We show that our method acquires 2,719,441 hyponymy relations with the first intermediate concepts (such as “work by Akira Kurosawa”) with 85.3% weighted precision and 6,347,472 hyponymy relations with the second intermediate concepts (such as “work by film director”) with 78.6% weighted precision. Furthermore, we confirm that hyponymy relations acquired by our method can be interpreted as “object – attribute – value”.

Key Words: *Hyponymy relation acquisition, Object-attribute-value acquisition, Wikipedia*

1 はじめに

上位下位関係は自然言語処理の様々なタスクにおいて最も重要な意味的關係の一つであり、それゆえ盛んに研究されてきた (Hearst 1992; Hovy, Kozareva, and Riloff 2009; Oh, Uchimoto, and Torisawa 2009; Ponzetto and Strube 2007; 隅田, 吉永, 鳥澤 2009; Suchanek, Kasneci, and Weikum 2007; Nastase and Strube 2008; Snow, Jurafsky, and Ng 2005). これらの過去の研究では、上位下位関係を、「A は B の一種あるいはインスタンスである A と B の関係」と定義している。本論文の上位下位関係もこの定義に従う。ただし、「概念」の詳細な表現を可能にするために、単一の語だけでなく、「黒澤明の映画作品」のような句や複合語も考慮する。このように制限を緩めることで、上位概念をより詳細に表現することが可能となる。

上記の定義によれば、次のペアはいずれも上位下位関係にあると考えられる¹。

- (1) 「黒澤明の映画作品 → 七人の侍」
- (2) 「映画作品 → 七人の侍」
- (3) 「作品 → 七人の侍」

質問応答等のアプリケーションを考えた場合、これらの上位下位関係の有用性は異なると考えられる。例えば、「“七人の侍”とは何ですか?」という質問に対して、上の3つの上位下位関係の上位概念のうち、答えとして適切なのは最も詳細な上位概念である(1)の「黒澤明の映画作品」と考えられる。一方、(3)の上位概念「作品」は、「何の作品であるか」という必要な情報が欠落しているため「黒澤明の映画作品」という答えに比べて適切ではない。

本論文では、以下の2つの条件を満たす場合に「下位概念 C に対して、A は B より詳細な上位概念」と呼ぶ。

- A と B は同じ下位概念 C を持つ
- B は A の上位概念である

¹ 本稿では上位下位関係を「A→B」のように表す。「A」が上位概念で「B」が下位概念である。

上記の例では、全ての上位概念が「七人の侍」という同じ下位概念を持ち、かつ、上位概念間には、それぞれ上位下位関係が成り立つ。「黒澤明の映画作品」の上位概念は「映画作品」、または「作品」、さらに「映画作品」の上位概念は「作品」と考えられる。従って、下位概念「七人の侍」に対して「黒澤明の映画作品」は「映画作品」や「作品」より詳細な上位概念であり、「映画作品」は「作品」より詳細な上位概念とすることができる。また、ある上位概念をより詳細な上位概念に置き換える処理を「上位概念の詳細化」と呼ぶ。

本研究では、自動獲得した上位下位関係の上位概念と下位概念の間に、より具体的な上位概念を中間ノードとして追加することで、元の上位下位関係を詳細化する。中間ノードとして追加されるより具体的な上位概念は、元の上位下位関係が記述されている Wikipedia 記事のタイトルと元の上位概念を「A の B」の形式で連結することで自動獲得する。

例として「作品 → 七人の侍」を挙げる。この上位下位関係は、タイトルが「黒澤明」の Wikipedia 記事の中に現れる。具体的には、当該記事の「作品」というセクションに「七人の侍」が記載されている。本手法では、この情報から、「七人の侍」は黒澤明の「作品」であると推測し、「黒澤明の作品 → 七人の侍」を新たに獲得する。さらに、「黒澤明」の上位概念が「映画監督」であることが獲得済みの上位下位関係から判明すれば、「映画監督の作品 → 七人の侍」も獲得できる。最終的に、元の「作品 → 七人の侍」から、「作品 → 映画監督の作品 → 黒澤明の作品 → 七人の侍」を得ることができる。

本稿ではさらに、本手法により獲得した上位下位関係（例えば「黒澤明の作品 → 七人の侍」）が「対象 - 属性 - 属性値」関係（例えば「黒澤明 - 作品 - 七人の侍」）として解釈できることについて議論する。この解釈では、Wikipedia 記事のタイトルが対象に、上位概念が属性に、下位概念が属性値に対応づけられる。実験で生成した上位下位関係 2,719,441 ペアは、94.0%の適合率で、「対象 - 属性 - 属性値」関係として解釈可能であることを確認した。

以下、2 節では、既存の手法で獲得された上位概念の問題点を例とともに述べる。3 節では、Wikipedia からの上位下位関係獲得手法 (隅田 他 2009) について説明する。4 節では、我々が開発した、Wikipedia を用いた詳細な上位下位関係の獲得手法について説明する。5 節では、提案手法の評価とエラー分析の結果について述べる。6 節では、提案手法により獲得した詳細な上位概念をより簡潔に言い換える試みと、詳細な上位下位関係の「対象 - 属性 - 属性値」関係としての解釈について議論する。7 節で関連研究について述べる。最後に 8 節で結論を述べる。

2 自動獲得された上位概念の問題

本節では、隅田ら (隅田 他 2009) の手法の出力を例に、自動獲得された上位概念に見られがちな問題点について述べる。

自動獲得された上位概念の中には、一般的なシソーラスにおいてルートノードの近くに位置

して広範囲な下位概念をカバーするものや、意味的に曖昧なものが存在するという問題が見られる。例えば「作品 → 七人の侍」における上位概念は「作品」だが、世の中には「作品」と呼べる物が数多く存在する。さらに極端な例として、上位概念が「物」や「事」になっている上位下位関係も、自動で獲得されてしまう可能性がある。このような上位概念を質問応答などの自然言語処理のアプリケーションで利用すると、より詳細な上位概念と比較してその有用性が低いことが多い。例えば1節の例で言及したように「“七人の侍” とは何ですか？」という質問に対しては、より詳細な上位概念である「黒澤明の映画作品」のほうが「作品」より適切な回答と考えられる。また、「黒澤明の作品には何がありますか？」といったリスト形式の回答を求めるとような質問に対して、上位下位関係を回答の知識源として使うことによって、上位概念「黒澤明の作品」の下位概念をリスト形式で回答できる²。一方、上位概念「作品」は他の映画作品や小説作品、音楽作品などの下位概念を持つため、「作品」が上位概念として含まれる上位下位関係のみを知識源として利用しても、このような質問に回答することは難しい。

表1に、隅田らの手法で獲得された上位下位関係で頻出した上位概念を挙げる。例えば、「アルバム」は、写真のアルバムなのか音楽が収録されているアルバムなのか分からず、曖昧である。一方、「出演者」は、これだけでは何に出演したのか分からない。この表から、自動獲得した上位下位関係の上位概念には、曖昧、または広範囲な下位概念をカバーする語が頻出していることがわかる。

このような問題点は、隅田らの手法に限らず発生すると考えられる。「A などの B」といった上位下位関係を明示する構文パターンから抽出する手法 (Hearst 1992) においても、例えば「七

表 1 隅田らの手法で獲得された上位下位関係中の上位概念 (出現頻度の降順上位 20 語)

頻度	上位概念	頻度	上位概念
250,914	出演作品	26,883	ゲーム
162,558	作品	26,764	スタッフ
129,487	登場人物	23,325	施設
117,542	キャスト	22,000	公立小学校
73,995	TV アニメ	20,742	シングル
54,145	TV ドラマ	18,831	小学校
53,591	出身者	18,481	アルバム
51,971	映画	17,443	部活動
47,399	収録曲	17,072	登場キャラクター
32,305	出演者	14,990	卒業生

² リスト形式の質問応答を行うタスクは、評価型ワークショップである TREC QA task (Dang, Lin, and Kelly 2006, 2007) で実施された。例えば「チューインガムの名前は？」といった質問に対して、そのインスタンスをすべて回答する。

人の侍などの作品」というフレーズからは、「七人の侍」の上位概念として「作品」が抽出される。つまり、他の多くの上位下位関係獲得手法についても当てはまる。

3 Wikipedia を用いた上位下位関係の獲得

本節では、隅田らが提案した Wikipedia を用いた上位下位関係の獲得手法 (隅田 他 2009) について述べる。この手法により獲得した上位下位関係が、4 節で説明する詳細な上位下位関係獲得の処理対象となる。

この手法では、Wikipedia 記事の階層的なレイアウト構造を利用して上位下位関係を獲得する。図 1 に、Wikipedia 記事の例として「アップル インコーポレイテッド」の記事を挙げる。この記事は「Apple ショップ」や「製品」という節があり、「Apple ショップ」の下位には「北海道地方」、「製品」の下位には「コンピュータ」、「iPod」、「iPhone」などの小節がある。さらに

図 1 Wikipedia 記事の例：アップル インコーポレイテッド

小節の中には、「Mac mini」や「MacBook」,「MacBook Air」といった項目が存在する。以後、これらの節見出し、小節タイトル、項目名を term と呼ぶことにする。

図1に示す Wikipedia 記事から上位下位関係候補を抽出する処理では、Wikipedia がデータベースのダンプデータとして提供している MediaWiki ソースコード (図2) を利用する。MediaWiki ソースコードでは、節見出し、小節タイトル、項目名を表現するために特殊な修飾記号が用いられる。例えば節見出しでは「==製品==」, 項目名では「*** Mac mini」などの記号が用いられ、その修飾記号の種類、繰り返し数により、レイアウト構造上の上下関係が決定する。

隅田らの手法では、まず、記事のレイアウト構造上の上下関係 (節タイトルは小節タイトルより上位にあり、小節タイトルは項目名より上位にある) を守りながら、2つの term から1つの上位下位関係候補を獲得する。例えば図1の場合、「製品 → コンピュータ」や「コンピュータ → Mac mini」, 「製品 → Mac mini」などが獲得される。次に、SVM (Vapnik 1995) を用いて、獲得された上位下位関係候補を正しそうなものとそうでないものに分類する。素性として、以下に示す特徴を上位概念候補、下位概念候補から抽出して利用する。

- 上位概念候補、下位概念候補の品詞。
- 上位概念候補、下位概念候補に含まれる形態素。
- 上位概念候補、下位概念候補の表層文字列。
- 上位概念候補、下位概念候補が属性語 X に一致するか否か。(属性語として、各記事の根ノード以外のノードに出現する単語を利用。)
- 上位概念候補、下位概念候補の修飾記号 (“=”, “*” など)。
- 上位概念候補と下位概念候補間のレイアウト構造上の距離。
- 上位概念候補が「主な～」, 「～のリスト」などの上位概念を表現する典型的なパターンに一致するか。
- 上位概念候補と下位概念候補の末尾の1文字が一致するか。

```
<title>アップル インコーポレイテッド</title>
== Appleショップ ==
=== 北海道地方 ===
* Apple Shop ビックカメラ 札幌店 - 2007年6月16日オープン
== 製品 ==
==== コンピュータ ====
*** [[Mac mini]] Intel Core 2 Duo 2.4GHz/2.66GHz を搭載
*** [[MacBook]] (13.3インチ) Intel Core 2 Duo 2.4GHzを搭載
*** [[MacBook Air]] (13.3インチ) Intel Core 2 Duo 1.6GHz/1.86GHz を搭載
==== iPod ====
* [[iPod nano]] - 2モデル (8GB, 16GB) 9色
* [[iPod touch]] - iPhoneを踏襲。3モデル (8GB, 32GB, 64GB) 1色、
==== iPhone ====
```

図2 Media Wiki ソースコードの例：アップル インコーポレイテッド

訓練データは, 隅田らが実験で用いたデータと同じものを使用した³. このデータは, Wikipedia から獲得した上位下位関係候補から 29,900 対を抽出し, 人手により上位下位関係か否かを判定することにより作成している.

この処理を 2009-09-27 版の Wikipedia に適用することにより, 1,925,676 ペアの上位下位関係を適合率 90% で獲得した. この上位下位関係をベース上位下位関係 (図 3(a)) と呼び, 4 節で説明する詳細な上位下位関係獲得の処理対象とする.

階層的なレイアウトを利用する手法とは別に, 隅田らは, Wikipedia 記事の定義文 (記事の第一文に該当) を用いた手法と, 記事下部にあるカテゴリ情報を用いた手法も提案している. これらの手法では記事タイトルが下位概念として使われるため, 我々が提案する記事タイトルによる上位下位関係の詳細化が適用できない. そこで, これら 2 つの手法により得られた上位下位関係はベース上位下位関係として用いず, G-上位概念の生成の際に用いる (4.2 節). この処理により, 2009-09-27 版の Wikipedia からは, 522,709 個の記事タイトルに対して 1,472,035 個の上位概念を適合率 90% で獲得した.

4 詳細な上位下位関係の獲得手法

2 節で述べた通り, ベース上位下位関係の上位概念の中には広範囲な下位概念をカバーするものや意味的に曖昧なものが存在する. そこで本節では, ベース上位下位関係を処理対象とした詳細な上位下位関係の獲得手法について述べる. 図 3 に, 提案手法の処理の流れの全体像を示

図 3 提案手法の処理の流れ

³ この訓練データにより学習されたモデルファイルと上位下位関係獲得ツールは <http://alaginrc.nict.go.jp/hyponymy/index.html> で公開されている.

す。まず、ベース上位下位関係の各上位概念を Wikipedia 記事のタイトルで詳細化し、詳細化された上位概念を元の上位概念と下位概念の間に中間ノードとして挿入する (4.1 節)。以降では、Wikipedia 記事のタイトルで詳細化された上位概念を T-上位概念と呼ぶ。また、T-上位概念を中間ノードとして挿入された上位下位関係を T-上位下位関係と呼ぶ (図 3(b))。次に、T-上位概念中の記事タイトル箇所をその上位概念で抽象化する事で、元の上位概念よりは詳細だが T-上位概念よりは抽象的な新たな上位概念を得る。以降では、この上位概念を G-上位概念と呼ぶ。G-上位概念は、T-上位下位関係の上から二番目、つまり元の上位概念の直下に挿入される (4.2 節)。T-上位概念に加え G-上位概念が挿入された上位下位関係を、これ以降、G-上位下位関係と呼ぶ (図 3(c))。なお本手法では、上位概念に関わらず、全てのベース上位下位関係を本提案手法により詳細化する。以下、各処理手順を詳しく説明する。

4.1 T-上位下位関係の獲得

Wikipedia の記事に出現する節タイトル、小節タイトル、項目名などは、その記事のタイトルによって情報を補足できると考えられる。ベース上位下位関係の上位概念は、Wikipedia の記事に出現する節タイトル、小節タイトル、項目名などに対応するため、T-上位下位関係の獲得処理では、ベース上位下位関係の上位概念を Wikipedia 記事タイトルで情報を補い、T-上位概念を生成する。上位概念を補う記事タイトルは、その上位概念と下位概念の抽出元の記事から取得する。

T-上位概念は、元の上位概念と Wikipedia 記事タイトルを、助詞「の」によって連結して生成する。例えば、上位概念「作品」と記事タイトル「黒澤明」は、助詞「の」によって連結されて「黒澤明の作品」という T-上位概念になる。

生成した T-上位概念は、元の上位概念と下位概念の間に挿入する。この結果、「作品 → 黒澤明の作品 → 七人の侍」のように、三階層の T-上位下位関係が生成される (図 3(b))。

4.2 G-上位下位関係の獲得

T-上位概念は、Wikipedia 記事タイトルとベース上位概念を「の」で連結して生成した。次に、T-上位概念の中の Wikipedia 記事タイトルの箇所を、その上位概念で置き換えることによって、さらなる上位概念となる G-上位概念を生成する。例えば「黒澤明の作品」という T-上位概念の場合、その Wikipedia 記事タイトルの箇所である「黒澤明」を上位概念である「映画監督」で置き換えて、「映画監督の作品」という G-上位概念を生成する。

G-上位概念の生成では、Wikipedia 記事タイトルの上位概念が必要になる。Wikipedia 記事タイトルの上位概念は、隅田らの手法のうち、3 節の最後で述べた、Wikipedia 記事の第一文を用いる手法と、記事下部のカテゴリ情報を用いる手法によって獲得する。例えば図 1 の場合、記事タイトルである「アップル インコーポレイテッド」の上位概念の候補がその第一文 (「アップ

ル社は、アメリカ合衆国... 製造する多国籍企業である.」)と記事下部にあるカテゴリ情報(カリフォルニアの企業, 多国籍企業, 携帯電話メーカー, ...)に記載されている. これらの上位概念候補は, 3節で述べたSVM分類器によって上位概念か否か判定される.

生成したG-上位概念を, T-上位下位関係の中の元の上位概念とT-上位概念の間に挿入し, G-上位下位関係を生成する. G-上位下位関係は, 例えば「作品 → 映画監督の作品 → 黒澤明の作品 → 七人の侍」のように, 四階層の上位下位関係となる.

5 評価実験

提案手法を評価するため, 2009-09-27 版の日本語 Wikipedia ダンプデータを対象として, 提案手法により G-上位下位関係を獲得した. 表2に, 獲得した G-上位下位関係の例を挙げる.

生成した G-上位下位関係から以下の三種類の上位下位関係ペアを抽出し, 各ペアが上位下位関係として妥当か評価を行った(図4).

ベース上位下位関係: 隅田らの手法により獲得した上位下位関係(例えば「作品 → 七人の侍」)

G-上位概念ペア: G-上位概念とベース上位下位関係の下位概念(例えば「映画監督の作品 → 七人の侍」)

T-上位概念ペア: T-上位概念とベース上位下位関係の下位概念(例えば「黒澤明の作品 → 七人の侍」)

Wikipedia ダンプデータを解析した結果, 1,925,676 個のベース上位下位関係, 6,347,472 個の G-上位概念ペア, そして, 2,719,441 個の T-上位概念ペアを獲得した. ベース上位下位関係は二記事以上に出現することがあるため, 出現した記事タイトルを補うことにより生成される T-上位概念ペアの数はベース上位下位関係の数より多くなる. また, 2,719,441 個の T-上位概念ペアのうち 2,113,040 ペアに対しては, 一つの Wikipedia 記事タイトルに2つ以上の上位概念が獲

表2 評価実験で獲得された G-上位下位関係の例

元の上位概念	G-上位概念	T-上位概念	元の下位概念
登場人物	SF 映画の登場人物	WALL-E の登場人物	M.O
製品	企業の製品	シリコングラフィックスの製品	IRIS Crimson
作品	アメリカの小説家の作品	J・D・サリンジャーの作品	A Boy in France
町	イングランドの州の町	イースト・サセックスの町	アックフィールド
監督	ミュージカル映画の監督	雨に唄えばの監督	スタンリー・ドーネン
卒業生	カリフォルニアの大学の卒業生	スタンフォード大学の卒業生	鳩山由紀夫
食材	カレーの食材	奥美濃カレーの食材	奥美濃ヘルシーポーク
イベント	放送局のイベント	フジテレビジョンのイベント	お台場どっと混む!

図 4 評価対象の三種類の上位下位関係ペア

得られたため、G-上位概念ペアの数はT-上位概念ペアの数より多い。一方、T-上位概念ペアのうち 342,884 ペアに対しては、上位概念（Wikipedia 記事タイトル）が獲得できなかったため、それらに対応する G-上位概念ペアが得られなかった。

獲得した G-上位下位関係から 200 サンプルを評価対象として抽出し、それら 200 サンプルからベース上位下位関係、T-上位概念ペア、G-上位概念ペアを取得した。サンプリングした G-上位下位関係の中で、22 個の T-上位概念に対する上位概念が自動獲得できなかったため、これらに対応する G-上位概念ペアが得られなかった。最終的に、ベース上位下位関係として 200 ペア、T-上位概念ペアとして 200 ペア、そして、G-上位概念ペアとして、サンプリングした G-上位下位関係から抽出可能な 178 ペアを評価した。

いずれも筆者ではない被験者三名により、これらのペアが上位下位関係として正しいかどうか評価を行った。被験者は次の三種類の評価ラベルを評価サンプルの各ペアに付与した。

Good: 上位下位関係として正しい。

Less good: 上位下位関係としては正しいが、「“下位概念” とは何？」といった質問の回答として相応しくない上位概念

Bad: 上位下位関係として間違っている。あるいは、上位概念または下位概念が意味不明である。

評価サンプルの各ペアに対して、被験者二名以上が選択したラベルを最終的な評価ラベルとした。被験者が三名とも異なる判断をした場合は、著者の一人によって最終的な評価ラベルを判断した⁴。被験者三名による評価アノテーションの Kappa 値は 0.58 であった。これは、本評価実験の評価アノテーションにまずまずの安定性があることを示している。

評価の指標として、次のように定義される重み付き適合率を用いた。

⁴ 578 ペアの評価サンプルのうち 9 ペアがこのケースに該当した。

表 3 上位下位関係の評価結果

	Good	Less good	Bad	重み付き適合率	Good ラベルのみを 正解とした適合率
ベース上位下位関係	0.500 (100/200)	0.460 (92/200)	0.040 (8/200)	0.730	0.500
G-上位概念ペア	0.702 (125/178)	0.169 (30/178)	0.129 (23/178)	0.786	0.702
T-上位概念ペア	0.850 (170/200)	0.005 (1/200)	0.145 (29/200)	0.853	0.850

$$\text{重み付き適合率} = \frac{\#Good \times 1 + \#Less\ good \times 0.5 + \#Bad \times 0}{\#Good + \#Less\ good + \#Bad} \quad (1)$$

ここで, $\#Good$, $\#Less\ good$, $\#Bad$ は, それぞれのラベル数を示す. 本評価実験における重み付き適合率の計算式では, Good ラベルを 1 つの正解サンプルとしてカウントし, Bad ラベルを正解サンプルとしてはカウントしない. この点は通常の適合率の計算と同じだが, Less good ラベル 1 つにつき, 0.5 を正解サンプル数に追加する点が通常と異なる. この重み付き適合率の計算方法は Pasca (Pasca 2007, 2009) も採用している. また, Good ラベルのみを正解とした適合率の計算も行った. 表 3 に評価結果を挙げる. この表の重み付き適合率のコラムを見ると, ベース上位下位関係, G-上位概念ペア, T-上位概念ペアと, 獲得される上位概念が詳細なほど重み付き適合率と Good ラベルのみを正解とした適合率が高くなっていることが読み取れる.

次に, 本実験における SVM 分類器の効果について考察する. 隅田らは, SVM によるフィルタリング処理を行わない場合, Wikipedia 記事の階層的なレイアウト構造から獲得した上位下位関係候補の適合率は 0.284 であると報告している (隅田 他 2009)⁵. つまり, SVM によるフィルタリング処理を行わない場合のベース上位下位関係では Good と Less Good を合わせても全体の 0.284 しか無く, 残りの候補が上位下位関係では無いと判断される. また, SVM によるフィルタリング処理を行わない場合, 定義文から獲得した上位下位関係候補は 0.894, カテゴリ情報からは 0.705 の適合率であると報告している. 本提案手法でも G-上位概念を生成する際に, 定義文とカテゴリ情報からフィルタリング処理を行い獲得した上位概念を使用している. この SVM によるフィルタリングの効果を明確にするため, 表 3 と同じ実験対象に対してフィルタリング処理を行わずに G-上位概念を生成し, 被験者三名による評価実験を行った. 結果を表 4 に示す. SVM によるフィルタリングを用いない結果 (表 4) は, フィルタリングを用いる結果 (表 3) と比較して, 重み付き適合率が 0.157 低く, SVM によるフィルタリングが効果的であることがわかる.

⁵ 本実験における上位下位関係候補の獲得処理は, 隅田らの手法と同じアルゴリズムを用いているため, この値が大きく変わることは無いと考えられる.

表 4 上位下位関係の評価結果 (SVM によりフィルタリング処理を行わない場合)

	Good	Less good	Bad	重み付き適合率	Good ラベルのみを 正解とした適合率
G-上位概念ペア	0.545 (97/178)	0.169 (30/178)	0.287 (51/178)	0.629	0.545

表 3 の結果では、Good と Less good, Bad のコラムからは、ベース上位下位関係、G-上位概念ペア、T-上位概念ペアと、獲得される上位概念が詳細なペアほど Less good と判定されるペアが減少し、Good あるいは Bad と判定されるペアが増加する傾向にあることが読み取れる。つまり、獲得される上位概念が詳細なペアほど、詳細で正しい上位概念だけでなく、詳細だが間違っている上位概念も増加する傾向にある。詳細で正しい上位概念が増加することは、本研究の当初の狙い通りのポジティブな側面であるが、間違っている上位概念が増加するのは予期しなかったネガティブな側面である。そこで、T-上位概念ペアに焦点を当てて、本提案手法による誤りの原因を次の三種類に分類した。

エラータイプ 1: ベース上位下位関係の誤りが原因となり、間違いと判定された。エラー全体の 27.6% を占める。

例) 「リンチバーグのヘリテージ高校 → ペリーモン小学校」

エラータイプ 2: 助詞「の」が元の上位概念と Wikipedia 記事タイトルを連結する表現として不適切であり、その結果生成された T-上位概念が意味不明なため、間違いと判定された。エラー全体の 3.4% を占める。

例) 「原山理一郎のアナウンサー → 小林豊」

エラータイプ 3: Wikipedia 記事タイトルによる詳細化によって上位概念が下位概念を包含する概念ではなくなったため (上位下位関係ではなくなったため)、間違いと判定された。エラー全体の 69.0% を占める。

例) 「大垣市の公共施設 → 図書館」

エラータイプ 1 の例「リンチバーグのヘリテージ高校 → ペリーモン小学校」におけるベース上位下位関係は「ヘリテージ高校 → ペリーモン小学校」であるが、これは上位下位関係として間違いである。3 節で述べた通り、ベース上位下位関係は隅田らの手法で獲得されるものであり、本提案手法は隅田らの手法の wrapper として機能するため、隅田らの手法のエラーはそのまま本提案手法に引き継がれる。つまり、エラータイプ 1 は提案手法を原因とはしていない。

エラータイプ 2 の例「原山理一郎のアナウンサー → 小林豊」におけるベース上位下位関係は「アナウンサー → 小林豊」であり、上位下位関係として正しい。しかし、本提案手法により、「アナウンサー → 小林豊」を獲得した Wikipedia 記事のタイトル「原山理一郎」を元の上位概念「アナウンサー」に助詞「の」によって連結したため、「原山理一郎のアナウンサー」という

意味不明な上位概念が生成された。この意味不明な上位概念が本来意味するところは「原山理一郎と同期入社のアナウンサー」である。つまり、元の上位概念と Wikipedia 記事タイトルを一樣に助詞「の」で連結するというナイーブな手法がこのタイプのエラーの原因となっている。

エラータイプ 3 の例「大垣市の公共施設 → 図書館」におけるベース上位下位関係は「公共施設 → 図書館」であり、これは上位下位関係として正しい。このベース上位下位関係を獲得した Wikipedia 記事のタイトルが「大垣市」である。そのため、本提案手法により「大垣市の公共施設 → 図書館」という T-上位下位関係が獲得された。しかし、「大垣市の公共施設」という概念は、元の上位概念である「公共施設」より詳細になってはいるが、「図書館」という概念を包含していない（大垣市の図書館以外にも図書館は存在する）ので、「大垣市の公共施設 → 図書館」は上位下位関係としては間違いとなる。

本提案手法の間違いの中でエラータイプ 3 に属するものが 69.0% と多数を占める。エラータイプ 3 に属する不適切な上位下位関係ペアの多くは、下位概念が普通名詞によって表されるものであり、正解と判定された上位下位関係ペアは、下位概念が固有名詞によって表されるものがほとんどであった。つまり、下位概念が普通名詞で表されている上位下位関係ペアを出力から除外することでエラータイプ 3 に属する間違いを減らすことができると考えられる。そこで、次の条件のいずれかに合致する term は普通名詞である可能性が高いと仮定し、下位概念が普通名詞である上位下位関係ペアを、評価サンプルのベース上位下位関係、T-上位概念ペア、G-上位概念ペアから除外した。

- Wikipedia 記事の節タイトル、あるいは小節タイトルとして使われている term
- 一定記事数（実験では 30 記事）以上に出現した term

表 3 と同じ処理対象に対して、下位概念が普通名詞と判断された上位下位関係を除外した場合の評価結果を表 5 に示す。下位概念が普通名詞と判断された上位下位関係を除外したため、ベース上位下位関係と T-上位概念では処理対象数が 200 ペアから 150 ペアに、G-上位概念では 178 ペアから 129 ペアに減少している。表 3 の結果と比べると、G-上位概念ペアの重み付き適

表 5 普通名詞で表される下位概念を持つ上位下位関係を除外した場合の評価結果

	Good	Less good	Bad	重み付き適合率	Good ラベルのみを 正解とした適合率
ベース上位下位関係	0.500 (75/150)	0.467 (70/150)	0.033 (5/150)	0.733	0.500
G-上位概念 ペア	0.767 (99/129)	0.171 (22/129)	0.062 (8/129)	0.853	0.767
T-上位概念 ペア	0.933 (140/150)	0.007 (1/150)	0.060 (9/150)	0.937	0.933

合率が 6.7%, T-上位概念ペアの重み付き適合率が 8.4%向上していることがわかる。しかし、全処理対象に対する獲得ペア数は、T-上位概念ペアが 2,719,441 ペアから 1,958,117 ペアへ、G-上位概念ペアが 6,347,472 ペアから 4,960,751 ペアへと減少した。獲得ペア数を保ちながら重み付き適合率を向上させる手法の開発は今後の課題とする。

6 応用

本節では、G-上位概念をより簡潔に言い換える手法と、T-上位概念ペアの「対象 - 属性 - 属性値」関係としての解釈について議論する。

6.1 G-上位概念のより簡潔な表現への言い換え

G-上位概念のいくつかはより簡潔な表現に言い換えることができる。この言い換え処理が自動化できれば、本提案手法で獲得した上位下位関係を既存のシソーラスと関連づけることが可能になる。例えば、G-上位概念として生成された「映画監督の作品」は「映画」に言い換えても問題ないと考えられる⁶。この言い換えにより、本提案手法で獲得した「映画監督の作品」の下位概念（映画のタイトルなどのインスタンスを含む）を既存のシソーラスの「映画」の位置に追加することができる。

そこで予備実験として、本提案手法で獲得した G-上位概念のうち最頻出の 20 概念に対して簡潔な言い換え表現を手作業で作成し、それらによって上位概念が言い換えられた G-上位下位関係の適合率を評価した。表 6 に、G-上位概念とその言い換え表現の例を挙げる。言い換え対象の 20 の G-上位概念を含む G-上位概念ペアは全部で 59,890 ペア、この G-上位概念に含まれる下位概念の異なり数は 54,981 個であった。その中から 200 ペアをサンプリングし、言い換え後の上位概念と下位概念のペアが上位下位関係であるか判定する実験を行った。実験では、筆者を含まない三名の被験者により判定を行い、二名以上が支持した結果を最終的な判定として使用した。三名の被験者の一致率を示す Kappa 値は 0.674 で、十分な一致率であると考えられ

表 6 G-上位概念の簡潔な言い換え表現の例

元の G-上位概念	言い換え後の G-上位概念
映画監督の作品	映画
放送局の番組	TV・ラジオ番組
小説家の作品	小説
ゲームクリエイターの作品	ゲーム
作曲家の作品	音楽

⁶ 「映画監督の作品」には「小説」などの可能性もあるが、ここでは主となる言い換え対象のみを扱う。

る。実験の結果、言い換え後の上位概念と下位概念のペアが上位下位関係として正しいと判定された適合率は 78.0%であった。

言い換え後の上位概念は既存のシソーラスに存在する単語を利用しているため、言い換え表現を 20 表現用意するだけで、異なり数 54,981 個の下位概念を適合率 78.0%で既存のシソーラスに追加できることがわかる。全ベース上位下位関係における下位概念異なり数は 1,199,826 個であり、わずか 20 個の言い換え表現で、下位概念全体の 4.6%をカバーしていることがわかる。今後、重複する下位概念などの情報を利用することによりこの言い換え表現を自動獲得し、カバー率を向上させることが課題となる。

6.2 T-上位概念ペアの「対象 – 属性 – 属性値」関係としての解釈

T-上位概念ペアは、Wikipedia 記事から獲得したベース上位下位関係と、その Wikipedia 記事のタイトルから構成される。この Wikipedia 記事のタイトルとベース上位下位関係の上位概念、下位概念は、対象とその属性、属性値という 3 つ組として解釈することができる。例として「黒澤明の作品 → 七人の侍」という T-上位概念ペアを挙げる。この T-上位概念ペアでは、「黒澤明」が Wikipedia 記事のタイトルで、「作品 → 七人の侍」がその記事から獲得された元のベース上位下位関係である。この場合、「作品」と「七人の侍」を「黒澤明」という対象の属性、属性値と解釈することができる。同様に、「シリコングラフィックスの製品 → IRIS Crimson」という T-上位概念ペアの場合も、「製品」と「IRIS Crimson」を「シリコングラフィックス」という対象の属性、属性値と解釈することができる。

5 節にある通り本提案手法による上位概念の詳細化は高い性能を示しているが、このことは、T-上位概念ペアが「対象 – 属性 – 属性値」関係として解釈可能であるという上記の観察結果によって、次のように説明できる。一般的に、属性は、それがどの対象の属性かを明示することで詳細化できると言える。本提案手法は、属性と上位概念の term、対象と Wikipedia 記事タイトルを対応づけた上でこの一般論に倣い、上位概念の term がどのタイトルの Wikipedia 記事から得られた term かを明示することで上位概念を詳細化している。従って、どの対象かを明示することで属性を詳細化できるという一般論が正しい限りにおいて、本提案手法は正しく上位概念を詳細化できる。

T-上位概念ペアが「対象 – 属性 – 属性値」関係として解釈できるという仮説が正しいかどうかを明らかにするために、T-上位概念ペアを「対象 – 属性 – 属性値」関係として評価した。まず、5 節の評価実験で使用した G-上位下位関係 200 サンプル（普通名詞で表される下位概念を持つ上位下位関係も含む）から、ベース上位下位関係に対応する元の上位概念と下位概念、T-上位概念の Wikipedia 記事タイトル箇所を取り出し、「Wikipedia 記事タイトル – 上位概念 – 下位概念」の 3 つ組を 200 個用意した。この評価データを「T-上位概念セット」と呼ぶ。これらとは別に、比較のため、隅田らの手法の処理途中で得られる上位下位関係候補（SVM で分

類される前のベース上位下位関係の候補. 3 節を参照) と, それらの出所である Wikipedia 記事のタイトルによって, 「Wikipedia 記事タイトル – 上位概念候補 – 下位概念候補」の 3 つ組を 200 個用意した. この評価データを「上位下位候補セット」と呼ぶ. 2 つの評価データの違いは, 上位下位候補セットには上位下位関係としては不適切な上位概念と下位概念がより多く含まれているという点にある.

次に, 3 名の被験者 (いずれも著者ではない) によって, これらの 3 つ組が「対象 – 属性 – 属性値」として正しいかを評価する実験を行った. 評価サンプルは, T-上位概念セットの 200 と上位下位候補セットの 200 の計 400 である. これら 400 サンプルはシャッフルした上で被験者に提示した. 評価の際は, 次の 3 種類の評価ラベルを使用した.

Vital: 「対象 – 属性 – 属性値」として適切.

Okay: 「対象 – 属性 – 属性値」として適切だが, その対象にとって当該の属性, 属性値は本質的なものとは言えない.

Wrong: 「対象 – 属性 – 属性値」として不適切.

5 節の評価実験と同様, 2 名以上の被験者が付与したラベルを各 3 つ組の最終的な評価ラベルとした. もし 3 名の被験者が皆異なる判断をした場合, 著者の一人が最終的な評価ラベルを決定した⁷. 被験者 3 名による評価ラベリングの Kappa 値は 0.51 であり, 本実験の評価ラベリングにまずまずの安定性があることを示している. 重み付き適合率は, 5 節の評価実験で使用した, 式 (1) と同様に, ラベルが Vital であるものを 1.0, Okay を 0.5, Wrong を 0 として正解サンプル数をカウントして算出した (Pasca 2007, 2009). 評価結果を表 7 に示す.

T-上位概念セットの「対象 – 属性 – 属性値」関係としての重み付き適合率が 94.0% であることから, T-上位概念ペアが「対象 – 属性 – 属性値」関係として解釈できるという仮説は正しいと考えられる. この重み付き適合率は, 表 3 における T-上位下位概念ペアの重み付き適合率より高い. これは, 5 節で述べたエラータイプ 3 のものが, 「対象 – 属性 – 属性値」関係としては, 正しい関係と判定されることに起因する. 例えば, エラータイプ 3 の例「大垣市 – 公共施設 – 図書館」は, 「対象 – 属性 – 属性値」関係としては正しい.

表 7 T-上位概念ペアの「対象 – 属性 – 属性値」としての評価結果

	Vital	Okay	Wrong	重み付き適合率
T-上位概念セット	0.940 (188/200)	0 (0/200)	0.060 (12/200)	0.940
上位下位候補セット	0.525 (105/200)	0.020 (4/200)	0.455 (91/200)	0.535

⁷ 著者の一人が評価ラベルを決めたケースは 400 サンプル中 9 サンプルであった.

一方, 上位下位候補セットの「対象 – 属性 – 属性値」関係としての適合率は 53.5% と低い. このことは, Wikipedia 記事タイトルとその記事から取り出した 2 つの term (節タイトル, 小節タイトル, 項目名) ならどんなものでも「対象 – 属性 – 属性値」関係として解釈できるわけではない, ということを示唆している. つまり, 2 つの term が上位下位関係として適切な場合にのみ, 「Wikipedia 記事タイトル – 上位概念の term – 下位概念の term」が「対象 – 属性 – 属性値」関係として解釈できる, ということを意味している.

7 関連研究

大量文書からの上位下位関係の獲得手法はこれまでに数多く提案されてきた. これらは言語表現パターンを用いるもの (Hearst 1992; Ando, Sekine, and Ishizaki 2004), クラスターリングに基づくもの (Pantel and Ravichandran 2004; Etzioni, Cafarella, Downey, Popescu, Shaked, Soderland, Weld, and Yates 2005), HTML 文書の構造を利用するもの (Shinzato and Torisawa 2004), Wikipedia の構造を利用するもの (隅田 他 2009; Oh et al. 2009; Yamada, Torisawa, Kazama, Kuroda, Murata, De Saeger, Bond, and Sumida 2009) に大きく分類することができる.

上位下位関係を構成する概念の詳細さの問題に取り組んだ研究は我々の知る限り Hovy らの研究 (Hovy et al. 2009) のみである. Hovy らは, Doubly-Anchored Pattern と呼ばれる語彙統語パターンを用いた bootstrap 手法によって, “people / Shakespeare” といった上位下位関係に中間語 writers を挿入する手法を提案した. しかし彼らの手法では, あらかじめ決めた “animals” と “people” という 2 種類のルートコンセプトのみを対象としている. 一方, 本提案手法では, 処理対象に制限はなく, あらゆる上位概念を扱うことができる.

本提案手法では Wikipedia を知識獲得源として利用しているが, Wikipedia からの知識獲得研究は近年活発化している (Kazama and Torisawa 2007; Ponzetto and Strube 2007; Suchanek et al. 2007; Nastase and Strube 2008; 隅田 他 2009; Oh et al. 2009; Yamada et al. 2009). Wikipedia からの知識獲得という文脈における本研究の新規性は, Wikipedia の百科事典としての性質を利用することで, 上位下位関係としてだけでなく, 「対象 – 属性 – 属性値」関係としても解釈可能な知識を獲得する手法を開発した点にある. 一般的に, 「対象 – 属性 – 属性値」関係における属性と属性値のペアは, 上位下位関係と解釈できないものも多数存在する. 提案手法により獲得できる「対象 – 属性 – 属性値」関係は, その属性と属性値が上位下位関係を持つものに限定しているが, 「対象 – 属性 – 属性値」関係を大量かつ高精度に獲得している.

8 おわりに

本稿では、自動獲得した上位下位関係の上位概念を、Wikipedia の情報を利用することで、より詳細にする手法を提案した。本手法により、2,719,441 個の T-上位概念ペアを重み付き適合率 85.3%で、6,347,472 個の G-上位概念ペアを重み付き適合率 78.6%で獲得することができた。さらに、下位概念が普通名詞である上位下位関係ペアを除く処理を行うことにより、1,958,117 個の T-上位概念ペアに対する重み付き適合率を 93.7%、4,960,751 個の G-上位概念ペアの重み付き適合率を 85.3%に向上できることを確認した。この結果は、ベースとしている上位下位関係獲得手法 (隅田 他 2009) における適合率 (1,925,676 ペアに対して 90.0%) と比較して十分な精度であると考えられる。また、G-上位概念をより簡潔に言い換える (例えば「映画監督の作品」を「映画」に言い換える) 実験を行い、わずか 20 個の G-上位概念の言い換え表現を作成することで、59,890 個の下位概念を適合率 78.0%で既存のシソーラスに追加できる可能性があることを明らかにした。最後に、本手法で獲得した上位下位関係が、「黒澤明 – 作品 – 七人の侍」などのように、「対象 – 属性 – 属性値」として解釈できることについて示した。

提案手法により生成した詳細な上位下位関係を使用することによって、質問応答におけるより適切な回答の生成や、「黒澤明の作品」の一覧といった「対象—属性」に対する属性値の検索が可能となる。この「対象—属性」に対する属性値の検索結果は、リスト形式の回答を求めるような質問応答のタスク (Dang et al. 2006, 2007) でも有用となる。さらに提案手法は、上位概念を詳細化して既存のシソーラスを拡張する手法としても利用可能と考えられる。

提案手法では T-上位概念を生成する際、元の上位概念と Wikipedia の記事タイトルを助詞「の」によって連結した。助詞「の」は多様な意味で用いることができるので、我々が実験した範囲では、この単純な方法がほとんどの場合に成功する。しかし、助詞「の」以外に、上位概念と Wikipedia 記事タイトルを結ぶより適切な表現が存在することもある。例えば「作品」と「黒澤明」の場合、「の」よりも「による」で連結した方が、日本語表現として適切な T-上位概念を生成できる。Torisawa (Torisawa 2001) は与えられた 2 つの名詞を連結する最も適切な表現を選択する手法を開発した。Torisawa の手法により我々の提案手法がさらに洗練されたものになる可能性が高いが、これは今後の課題とする。

参考文献

- Ando, M., Sekine, S., and Ishizaki, S. (2004). “Automatic Extraction of Hyponyms from Japanese Newspaper Using Lexico-syntactic Patterns.” In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC)*, pp. 387–390.
- Dang, H., Lin, J., and Kelly, D. (2006). “Overview of the TREC 2006 Question Answering

- Track.” In *Proceedings of the Fifteenth Text REtrieval Conference*.
- Dang, H., Lin, J., and Kelly, D. (2007). “Overview of the TREC 2007 Question Answering Track.” In *Proceedings of the Sixteenth Text REtrieval Conference*.
- 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.
- Hearst, M. A. (1992). “Automatic Acquisition of Hyponyms from Large Text Corpora.” In *Proceedings of the 14th conference on Computational Linguistics (COLING)*, pp. 539–545.
- Hovy, E., Kozareva, Z., and Riloff, E. (2009). “Toward Completeness in Concept Extraction and Classification.” In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 948–957.
- Kazama, J. and Torisawa, K. (2007). “Exploiting Wikipedia as External Knowledge for Named Entity Recognition.” In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pp. 698–707.
- Nastase, V. and Strube, M. (2008). “Decoding Wikipedia Categories for Knowledge Acquisition.” In *Proceedings of the 23rd AAAI Conference on Artificial Intelligence (AAAI)*, pp. 1219–1224.
- Oh, J.-H., Uchimoto, K., and Torisawa, K. (2009). “Bilingual Co-Training for Monolingual Hyponymy-Relation Acquisition.” In *Proceedings of ACL-09: IJCNLP*, pp. 432–440.
- Pantel, P. and Ravichandran, D. (2004). “Automatically Labeling Semantic Classes.” In *Proceedings of the Human Language Technology and North American Chapter of the Association for Computational Linguistics Conference (HLT-NAACL)*, pp. 321–328.
- Pasca, M. (2007). “Organizing and Searching the World Wide Web of Facts—Step Two: Harnessing the Wisdom of the Crowds.” In *Proceedings of the 16th World Wide Web Conference (WWW)*, pp. 101–110.
- Pasca, M. (2009). “Outclassing Wikipedia in Open-Domain Information Extraction: Weakly-Supervised Acquisition of Attributes over Conceptual Hierarchies.” In *Proceedings of the 12th Conference of European Chapter of the Association of Computational Linguistics (EACL)*, pp. 639–647.
- Ponzetto, S. P. and Strube, M. (2007). “Deriving a Large-Scale Taxonomy from Wikipedia.” In *Proceeding of the 22nd Conference on the Advancement of Artificial Intelligence (AAAI)*, pp. 1440–1445.
- Shinzato, K. and Torisawa, K. (2004). “Extracting Hyponyms of Prespecified Hypernyms from

- Itemizations and Headings in Web Documents.” In *Proceedings of the 20th Conference on Computational Linguistics (COLING)*, pp. 938–944.
- Snow, R., Jurafsky, D., and Ng, A. Y. (2005). “Learning Syntactic Patterns for Automatic Hypernym Discovery.” In *Proceedings of the Neural Information Processing Systems (NIPS)*.
- Suchanek, F. M., Kasneci, G., and Weikum, G. (2007). “Yago: A Core of Semantic Knowledge.” In *Proceedings of the 16th World Wide Web Conference (WWW)*, pp. 697–706.
- Torisawa, K. (2001). “An Unsupervised Method for Canonicalization of Japanese Postpositions.” In *Proceedings of the 6th Natural Language Processing Pacific Rim Symposium (NLPRS)*, pp. 211–218.
- Vapnik, V. N. (1995). *The Nature of Statistical Learning Theory*. Springer-Verlag New York, Inc., New York, USA.
- Yamada, I., Torisawa, K., Kazama, J., Kuroda, K., Murata, M., De Saeger, S., Bond, F., and Sumida, A. (2009). “Hypernym Discovery Based on Distributional Similarity and Hierarchical Structures.” In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 929–937. Association for Computational Linguistics.
- 隅田飛鳥, 吉永直樹, 鳥澤健太郎 (2009). Wikipedia の記事構造からの上位下位関係抽出. 自然言語処理, 16 (3), pp. 3–24.

略歴

- 山田 一郎：1993 年名古屋大学大学院修士課程修了。同年 NHK 入局。2008 から 2011 年（独）情報通信研究機構出向。現在 NHK 放送技術研究所主任研究員。博士（情報科学）
- 橋本 力：2005 年京都大学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。博士（言語科学、情報学）。
- 呉 鍾勲：2005 年 KAIST（韓国科学技術院）電子電算学科電算学専攻博士課程卒業。同年 KAIST 研究員を経て、（独）情報通信研究機構に専攻研究員として着任。博士（工学）。自然言語処理の研究に従事。
- 鳥澤健太郎：1995 年東京大学大学院博士課程中退。同年同大学院助手。北陸先端科学技術大学院大学助教授を経て、現在、（独）情報通信研究機構情報分析研究室室長。博士（理学）。日本学術振興会賞など受賞。
- 黒田 航：現京都大学・京都工芸繊維大学（非常勤講師）、早稲田大学情報教育研究所（招聘研究員）。元（独）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言語基盤グループ研究員。京都大学から人間・環境学博士を

取得. 言語学の認知科学と自然言語処理と言語教育を融合する研究に従事.

Stijn De Saeger: 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.

土田 正明: 2005 年東京理科大学大学院修士課程修了. 同年 4 月より NEC に入社. 2009 年 4 月から 2011 年 3 月まで (独) 情報通信研究機構に出向し, 現在は NEC に復帰. 2008 年人工知能学会大会優秀賞を受賞.

風間 淳一: 2004 年東京大学大学院情報理工学系研究科博士課程修了. 博士 (情報理工学). 同年北陸先端科学技術大学院大学助手. 2008 年より情報通信研究機構. 現在, 情報分析研究室主任研究員.

(2011 年 3 月 22 日 受付)

(2011 年 6 月 21 日 再受付)

(2011 年 8 月 3 日 採録)