自動獲得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†‡
デサーガ ステイン*　土田正明*　風間淳一*
*情報通信研究機構 MASTAR プロジェクト 言語基盤グループ
†早稲田大学 総合研究機構 情報教育研究所
‡京都工芸繊維大学

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利用されている重要な意味的関係の一つであり，これまでに多くの研究が提案されている[1,2,3,4,5,6,7]．これらの過去の研究では，上位下位関係を「AはBの一種である」のようなAとBの関係と定義し[1]，本研究でもこの定義に従う．ただし，AとBには，単一の単語だけでなく，複数の形態素で構成される語句も許容する．この定義によれば，以下はいずれも上位下位関係にあると考えられる．

(1) 「選手／クリスティアーノ・ロナウド」[2]
(2) 「サッカー選手／クリスティアーノ・ロナウド」
(3) 「レアルマドリードのサッカー選手／クリスティアーノ・ロナウド」

質問応答などのアプリケーションを考えた場合，これらの有用性は異なる．例えば，「クリスティアーノ・ロナウドとは誰ですか?」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で，答えとして最も適切なのは(3)の「レアルマドリードのサッカー選手」であり，(1)の上位概念「選手」は，答えとしてはあまりに漠然としている．しかし，このような漠然とした上位概念は，自動獲得した上位下位関係において頻繁に現れる．

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手法を提案する．追加する中間ノードは，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助詞「の」で連結することで自動生成する．例として「作品／七人の侍」を挙げる．この上位下位関係は，記事タイトルが「黒澤明」のWikipedia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」を新たに生成する．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」も生成できる．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．

評価実験では，Wikipedia から自動獲得した 1,668,982 ペアの上位下位関係を対象として，Wikipedia のタイトルを補うことにより生成した上位下位関係（上記の例では「黒澤明の作品／七人の侍」）を適合率 93.7%で 1,958,117 ペア生成した．また，Wikipedia のタイトルの上位概念を補うことにより生成した上位下位関係（上記の例では「映画監督の作品／七人の侍」）を適合率 85.3%で 4,960,751 ペア生成した．この実験結果は，本手法によって，既存の手法で獲得した上位下位関係を高い精度で詳細化できることを示している．本稿ではさらに，生成した上位下位関係（例えば「黒澤明の作品／七人の侍」）が対象，属性名，属性値の関係（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．

2. 自動獲得された上位概念の問題

自動獲得した上位下位関係の上位概念には，漠然としている，あるいは意味的に曖昧なものが存在するという問題が見られる．表1に，隅田らの手法[5]により獲得した上位下位関係で，頻出した上位概念の上位20語を示す．例えば「アルバム」は，写真のアルバムなのか，もしくは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ず，漠然としている．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，または漠然としてい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，従来提案されている上位下位関係自動獲得手法の多くで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[1]においても，例えば「七人の侍などの作品」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，この上

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(出現頻度の上位20位)

出現頻度	上位概念	出現頻度	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	14,990	卒業生

[1] 「AはBのインスタンスである」のような関係も含む
[2] 本稿では上位下位関係を「上位概念／下位概念」と表記する

位概念は漠然としていると考えられる．つまり，他の多くの上位下位関係獲得手法においても共通する問題と考えられる．

3 Wikipedia を用いた詳細な上位下位関係の獲得手法

本節では，我々が提案する Wikipedia を用いた詳細な上位下位関係の獲得手法について述べる．図 1 に，詳細化の流れを示す．まず，隅田らの手法を用いて，詳細化対象のベースとなる上位下位関係を獲得する．この上位下位関係を「ベース上位下位関係」と呼ぶ．次に，ベース上位下位関係の上位概念を Wikipedia 記事のタイトルで詳細化し，ベース上位下位関係の中間ノードとして挿入する．これを「中間ノード A」と呼び，ベース上位下位関係に中間ノード A を挿入したものを「上位下位関係 A」と呼ぶ(図 1 中央)．最後に，中間ノード A を抽象化する事で，元の上位概念よりは詳細だが中間ノード A よりは抽象的な新たな中間ノードを得る．これを「中間ノード B」と呼ぶ．中間ノード B は，上位下位関係 A の上から二番目，つまり元の上位概念の直下に挿入する．中間ノード A に加え中間ノード B が挿入された上位下位関係を，「上位下位関係 B」と呼ぶ．以下に，各処理手順を詳しく説明する．

3.1 ベース上位下位関係の獲得

隅田らは，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手法を提案した[5]．例えば「アップル　インコーポレイテッド」の記事には，「Apple ショップ」や「製品」という節の見出しがあり，「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．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，「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 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．素性としては，term 中の形態素や品詞などの語彙的情報，Wikipedia 記事中での term 間の距離などの構造的情報を用いる．

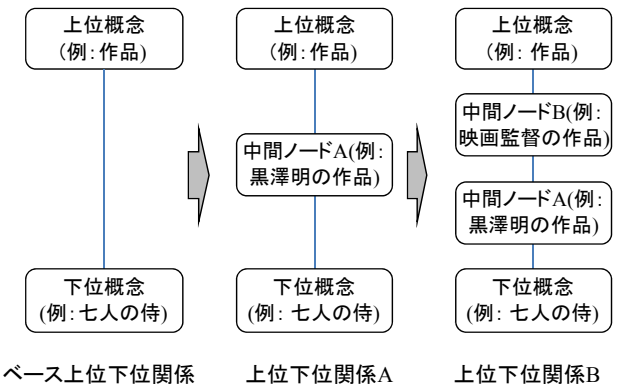

図 1．詳細化の流れ

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．例えば記事タイトル「アップル　インコーポレイテッド」の上位概念の候補は，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．」の「多国籍企業」）とカテゴリ情報（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，Wikipedia 記事の階層的なレイアウト構造を利用した手法と同様の SVM 分類器によって，上位概念か否かを判定される．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，中間ノード B の生成の際に必要とされる記事タイトルの上位語の情報として用いる．

3.2 上位下位関係 A の生成

Wikipedia の記事に出現する term 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 term に対応するため，上位下位関係 A の生成処理では，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，一つ目の詳細化したノードとなる中間ノード A 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．

中間ノード A 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」という中間ノード A になる．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している．

生成した中間ノード A は，元の上位概念と下位概念の中間に挿入する．この結果，「作品／黒澤明の作品／七人の侍」のように，三階層の上位下位関係 A が生成できる．

この処理において，生成した中間ノード A が詳細化されすぎ，カバーする意味範疇が下位概念より狭くなるような場合，中間ノード A と下位概念の上位下位関係が成立せずに問題となる．例えば，上位下位関係「公共施設／図書館」に対して，その記事タイトル「大垣市」を上位概念に補完して生成した中間ノード A「大垣市の公共施設」は，「図書館」の上位語として相応

しくない．このような事例は，下位概念が普通名詞(固有名詞以外の名詞)の場合に発生する．そこで，次の条件のいずれかに合致する場合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中間ノード生成対象のベース上位下位関係から除外して処理を行う．

- 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

3.3 上位下位関係 B の生成

中間ノード A は, Wikipedia 記事タイトルとベース上位概念を連結して生成した．次に，中間ノード A を抽象化した中間ノード B を生成する．この処理では，中間ノード A を構成する Wikipedia 記事タイトルの箇所を，その上位概念で置き換える．例えば「黒澤明の作品」という中間ノード A 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」という中間ノード B が生成できる．

中間ノード B 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, 3.1 節の最後で述べた, Wikipedia 記事の第一文を用いる手法と，カテゴリ情報を用いる手法によって獲得する．生成した中間ノード B を，上位下位関係 A の上位概念と中間ノード A の間に挿入し，上位下位関係 B を生成する． 上位下位関係 B は，「作品／映画監督の作品／黒澤明の作品／七人の侍」のような四階層の上位下位関係となる．

4. 評価実験

提案手法を評価するため，まず，3.1 節で説明した Wikipedia の階層的なレイアウト構造を利用する手法を 2009-09-27 版の Wikipedia に適用し，1,925,676 ペアの上位下位関係を適合率 90%で獲得した．また，Wikipedia 記事の定義文とカテゴリ情報を用いた手法により，522,709 個の記事タイトルに対して 1,472,035 個の上位語を適合率 90%で獲得した．

階層的なレイアウト構造を利用する手法で獲得した上位下位関係 1,925,676 ペアから，下位概念が普通名詞の場合のフィルタリング処理(3.2 節参照)を行い，1,668,982 ペアのベース上位下位関係を処理対象として生成した．このベース上位下位関係に対して，提案手法を適用することによって，1,958,117 個の中間ノード A ペアと，4,960,751 個の中間ノード B ペアを獲得した．ベース上位下位関係の中には，一つの関係が複数の記事に出現するものがあるため，出現した記事タイトルを補うことにより生成される中間ノード A ペアの数はベース上位下位関係の数より多くなる．また，1,958,117 個の中間ノード A ペアのうち 1,492,320 ペアに対しては， 1 つの Wikipedia 記事タイトルに複数の上位概念が獲得されたため，中間ノード B ペアの数は中間ノード A ペアの数より多い．一方，中間ノード A ペアのうち 273,603 ペアに対しては，Wikipedia 記事タイトルの上位概念が獲得できなかったため，それらに対応する中間ノード B ペアが得られなかった．

生成した上位下位関係 B から 150 サンプルを評価対象として抽出し，このサンプルからベース上位下位関係，中間ノード A と下位概念のペア（中間ノード A ペア），中間ノード B と下位概念のペア（中間ノード B ペア）を取得した．サンプリングした上位下位関係 B の中で，21 個の中間ノード A ペアに対する上位概念が自動獲得できなかったため，対応する中間ノード B 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 150 ペア，中間ノード A ペアとして 150 ペア，そして， 中間ノード B ペアとして，サンプリングした上位下位関係 B から抽出可能な 129 ペアを評価対象とした．

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，上位概念が漠然としているか曖昧である．
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．被験者三名による評価アノテーションの κ 値は 0.58 であった．これは，本評価実

表 2. 生成した上位下位関係 B の例

上位概念	中間ノードB	中間ノードA	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！

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 (1)式で定義した重み付き適合率を用いた．

$$適合率 = \frac{\#Good + \#Less\,good \times 0.5}{\#Good + \#Less\,good + \#Bad} \quad (1)$$

ここで，*#Good*，*#Less good*，*#Bad* は，それぞれのラベル数を示す．表 3 に評価結果を示す．この表の適合率を見ると，ベース上位下位関係，中間ノード B ペア，中間ノード A ペアと，上位概念が詳細なペアほど適合率が高くなっていることが読み取れ，元の上位下位関係の詳細化を実現していることが分かる．

表 3．詳細化した関係の評価結果

	Good	Less good	Bad	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
中間ノードA ペア	0.933 (140/150)	0.007 (1/150)	0.060 (9/150)	0.937
中間ノードB ペア	0.767 (99/129)	0.171 (22/129)	0.062 (8/129)	0.853

5．属性関係としての解釈

中間ノードAペアで使用したWikipedia記事のタイトルとベース上位下位関係の上位概念，そして，その下位概念は，対象と属性名，属性値という 3 つ組として解釈することができる．例えば「黒澤明の作品／七人の侍」という中間ノード A ペアでは，「黒澤明(Wikipedia の記事タイトル)」の「作品(上位概念)」が，「七人の侍(下位概念)」という対象，属性名，属性値と解釈することができる．

一般的に属性名は，それがどの対象の属性名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，属性名と上位概念の 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名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．

この仮説が正しいかどうかを明らかにするために，前節の評価実験で使用した中間ノード A ペア 150 サンプルから，「Wikipedia 記事タイトルー上位概念ー下位概念」の 3 つ組を抽出した．これらとは別に，Wikipedia から「Wikipedia 記事タイトルー記事の任意の節見出しの termー属性名とした term の構造上の下位に属する term」の 3 つ組をベースラインのデータとして抽出して，属性関係にあるか評価した．2 つの評価データの違いは，後者の「属性名ー属性値」には，上位下位関係としては不適切なものが，より多く含まれているという点にある．評価は前節と同様に，3 名の評価者によって，これら 2 つのデータが属性関係として適切であるかを，“Good”, “Less Good”, “Bad”の 3 種類の評価ラベルを使用して行った．評価結果を表 4 に示す．

表 4．属性関係の評価結果

	Good	Less good	Bad	適合率
中間ノードA の 3 つ組	0.967 (145/150)	0 (0/150)	0.033 (5/150)	0.967
ベースライン の 3 つ組	0.533 (80/150)	0.027 (4/150)	0.440 (66/150)	0.547

中間ノード A ペアから抽出した 3 つ組の適合率が 96.7%であることから，中間ノード A ペアが「対象ー属性名ー属性値」として解釈できるという仮説は正しいと考えられる．一方，ベースラインとして抽出したデータの適合率は 54.7%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属性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 記事タイトルー上位概念の termー下位概念の term」が属性関係として解釈できる，ということを意味している．

6．まとめ

本稿では，漠然とした上位概念を持つ上位下位関係を，Wikipedia の情報を利用することで，より詳細にする手法を提案した．本手法により，1,958,117 個の中間ノード A ペアを適合率 93.7%で，4,960,751 個の中間ノード B ペアを適合率 85.3%で生成することができた．さらに，詳細化した上位下位関係が，属性関係として解釈できることについて示した．

【 参考文献 】

[1] Hearst, Marti A., "Automatic acquisition of hyponyms from large text corpora," In *Proceedings of the 14th conference on Computational linguistics*, pp. 539-545 (1992).

[2] Hovy, Eduard and Kozareva, Zornitsa and Riloff, Ellen, "Toward completeness in concept extraction and classification," In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pp. 948-957 (2009).

[3] Oh, Jong-Hoon and Uchimoto, Kiyotaka and Torisawa, Kentaro, "Bilingual Co-Training for Monolingual Hyponymy-Relation Acquisition," In *Proceedings of ACL-09*, pp.432-440 (2009).

[4] Simone Paolo Ponzetto and Michael Strube, "Deriving a Large-Scale Taxonomy from Wikipedia," *In AAAI*, pp. 1440-1445 (2007).

[5] 隅田飛鳥, 吉永直樹, 鳥澤健太郎, "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(2009).

[6] 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. 情処学研報, 2003-NL-157, pp. 77–82 (2003).

[7] Yamada, Ichiro and Torisawa, Kentaro and Kazama, Jun'ichi and Kuroda, Kow and Murata, Masaki and De Saeger, Stijn and Bond, Francis and Sumida, Asuka, "Hypernym discovery based on distributional similarity and hierarchical structures," In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pp. 929-937 (2009).