ラベルなしデータからの意味カテゴリタガーの学習

村本 英明† 鍜治 伸裕‡ 吉永 直樹‡ 喜連川 優‡
†東京大学大学院 情報理工学系研究科
‡東京大学 生産技術研究所
{muramoto, kaji, ynaga, kitsure}@tkl.iis.u-tokyo.ac.jp

要旨 固有表現認識など，テキスト中の語句に意味カテゴリを付与する技術は，高度なテキスト処理の実現には必要不可欠である．しかし，従来の意味カテゴリタグ付与手法は，人手で作成したラベル付きデータという，高価な言語資源に強く依存しており，そのため，意味カテゴリの拡張や異なるドメインへの適応が困難となっていた．そこで本論文では，カテゴリ辞書およびラベルなしデータという，廉価に入手可能な言語資源から意味カテゴリタガーを構築する方法を検討する．現在のところ，ウェブドメインを対象として，45 の意味カテゴリを付与するタガーの構築を進めており，本稿ではその経過報告，および今後の課題に関する議論を行う．

1. はじめに

物事を何らかのカテゴリに分類することによって，その意味を認識，識別することは，我々人間が有する基本的な認知能力の一つであり，言語理解とも深い関わりを持つ．

(1) a. イギリス人科学者のニュートンとフック．
b. 今月発売のニュートンの特集記事．

例えば我々は，(1a) において，ニュートンとフックが「人物」という共通の意味を持つことを認識できる．また，例文 (1b) におけるニュートンは「出版物」であり，例文 (1a) のニュートンとは意味が異なることも認識可能である．
これと同様の認識処理の機械化，すなわち，テキスト中の語句に対して適切な意味カテゴリを付与する言語処理技術の実現は，工学的に大きな意義を有すると考えられ，従来から盛んに研究が行われてきた．例えば，固有表現認識はその代表例であり，情報抽出や質問応答における要素技術として，これまで膨大な数の研究が行われてきた [1]．また，ACE[1]における言及検出（mention detection）タスクも，意味カテゴリ付与との関連が深いと言える．
従来，固有表現認識をはじめとする意味カテゴリのタグ付与は，主として新聞記事テキストを対象として議論が行われてきた．そのため，タグ付与の対象は，人物や組織といった，新聞記事に頻出する少数のカテゴリに限定されている．さらに，タガーの構築に利用するラベル付きデータは，新聞記事テキストから作成されたものが大半である．
このことは，例えばウェブのような，新聞記事以外のテキストを処理するときに大きな問題となる．従来の意味カテゴリタガーは，食品や出版物など，ウェブドメインにおける重要なカテゴリの多くを取り扱うことができない．また，新聞記事テキストから構築したタガーを，ウェブという異なるドメインのテキストに適用した場合，精度が大きく劣化することも懸念される．
上記の問題の根本的な原因は，意味カテゴリタガーの構築方法が，手作業で作成したラベル付きデータという高価な言語資源に依存している点にある．そうしたアプローチにおいては，取り扱う意味カテゴリを拡張したり，タグ付け対象のドメインを変更したりするためには，ラベル付きデータを再構築する必要がある．しかしながら，それには多大な作業コストが必要になるため，現実問題としては実現困難であると言わざるを得ない．
一方，近年では，ウェブの急速な発展に伴い，不特定多数による協調的な辞典編纂が実現され，ウィキペディアなどの巨大辞典が入手可能になった．そのため，これを利用すれば，大規模で高品質な**意味カテゴリ辞書**（名詞句とそれが取りうる意味カテゴリを記述した辞書）を容易に構築することができる．また，大量テキストからの言語知識獲得に関する研究の進展に伴い，意味カテゴリ辞書をテキストから自動構築することも技術的に可能になりつつある．
このような背景を踏まえ，我々は，意味カテゴリ辞書とラベルなしデータから，意味カテゴリタガーを構築するための方法の検討を行っている．このことが実現されれば，意味カテゴリ辞書という，ラベル付きデータよりも遥かに廉価な言語資源からのタガー構築が可能になる．その結果として，意味カテゴリの拡張や，対象ドメインの変更が格段に容易になることが期待される．現在のところ，ウェブドメインを対象として，45 の意味カテゴリタグを付与するタガーの構築を進めており，本稿ではその経過報告，および今後の課題に関する議論を行う．

2. 関連研究

半教師あり学習およびドメイン適応は，いずれも，大量のラベル付きデータを人手で作成するコストの削減を目的とした学習枠組みであり，固有表現認識においても多くの適用事例が報告されている[2,3,4,5,6]．これらの研究は，本研究と少なからず問題意識を共有していると考えられるが，カテゴリの拡張に関する議論は見られず，我々とは研究の目的が異なる．また我々は，人手で作成したラベル付きデータの代わりに，意味カテゴリ辞書を利用してタガーを構築するという問題設定を議論しており，この点においても上記の研究とは大きく異なっている．
固有表現認識で扱うカテゴリの拡張に関する試みとして，例えば Sekine らは拡張固有表現と呼ばれる約 200 のカテ

[1] http://www.itl.nist.gov/iad/mig//tests/ace/2008/

ゴリ集合を提案している[7]．しかしながら，ウェブテキストを対象として，そうした大規模な意味カテゴリタグを付与するタガーを構築することは，依然として技術的に困難である．本研究は，これに対する解決策を探究するものであると位置づけることができる．現在に至るまで，意味カテゴリ辞書とラベルなしデータからのタガー構築に関する研究事例は極めて少ない[9,10]．しかし，これらのような大規模に入手可能な言語資源の効果的な活用法を明らかにすることは，自然言語処理における重要な研究の方向性であると考えられる．

一方，テキストコーパスからの語句の意味カテゴリ獲得に関しては，Hearst の先駆的研究[11]を始めとして，これまでに数多くの研究事例が報告されている[12]．近年では，単なるテキストコーパスだけでなく，検索ログなどの活用も議論されており興味深い[13]．これらの研究が意味カテゴリ辞書の自動構築を目指しているのに対して，本研究は，それを用いてタガーを構築することを目的としている．

3. 手法の概要

本節では，任意のドメインのテキストを入力として，所与の意味カテゴリをテキスト中の各名詞(句) に付与する意味カテゴリタガーを，廉価で構築する手法を提案する．
我々はまず，所与の意味カテゴリ集合に対して，既存の上位下位関係の獲得手法[14,15,16]を利用して，ウェブからオンデマンドで大規模な意味カテゴリ辞書を構築する．次に，得られた意味カテゴリ辞書を利用して，多クラス分類器学習のためのラベル付きデータを獲得し，所与の名詞(句) を適切な意味カテゴリに割り当てる分類器を学習する．

このようにして得られた意味カテゴリ辞書と分類器を用いて，タガーは以下の手順で所与のテキスト中の名詞(句) に対して意味カテゴリを付与する．

タグ付け対象文字列の認識
テキスト中で意味カテゴリ辞書の辞書項目と一致する文字列をタグ付けの対象語句として認識する．

辞書と分類器による意味カテゴリの識別
認識された文字列について，意味カテゴリ辞書を利用して絞り込まれた候補カテゴリから，多クラス分類器を用いて適切な意味カテゴリ(タグ) を選択し付与する．

提案手法は，特定の意味カテゴリ集合を前提としないが，説明のため，本節では実際に実験で用いた意味カテゴリ集合を例に議論を進める．我々は，関根らが設計した拡張固有表現を元に，具体物を対象とした意味カテゴリ集合を構成した．具体的には，時間表現・数値表現を除いた固有名に関する拡張固有表現階層において，第二層の各クラス(例: 組織名，地域名，材料名，自然現象名) を一つの意味カテゴリとみなした[2]．表 1 が本研究で用いた 45 の意味カテゴリである．なお，意味カテゴリの導出に用いた拡張固有表現階層の詳細については，関根の拡張固有表現階層-7.1.0-[3]を参照されたい．
以下で，意味カテゴリ辞書の構築手法と，得られた意味カテゴリ辞書を用いたラベル付きデータの獲得手法について順に説明する．

表 1 拡張固有表現から構成した 45 の意味カテゴリ

人，神，国際組織，公演組織，家系，民族，競技組織，法人，政治的組織，温泉，GPE[4]，地域，地形，天体，遺跡，GOE[5]，路線，材料，衣類，貨幣，医薬品，武器，賞，勲章，罪，キャラクター，乗り物，食べ物，芸術作品，出版物，主義方式，規則，称号，言語，単位，催し物，事故事件，自然災害，元素，化合物，鉱物，生物，生物部位，動物病気，自然色

3.1. 意味カテゴリ辞書の構築

本節では，名詞(句) と意味カテゴリの対の形でウェブから意味カテゴリ辞書を自動構築する手法について述べる．本節で獲得する意味カテゴリ辞書は，意味カテゴリの多クラス分類器学習のためのラベル付きデータの獲得で使うほか，意味カテゴリ分類時の候補カテゴリの絞り込みに利用するため，高精度かつ網羅的であることが必要となる．
所与の意味カテゴリに属する名詞(句) を網羅的に獲得するために，我々は近年広く研究されているウェブを知識源とした上位下位関係の獲得手法[15,16]に着目した．基本的なアイデアとしては，各意味カテゴリと，上位下位関係の獲得手法で得られた上位語とを対応づけることで，下位語集合と意味カテゴリの間を間接的に対応づけ，意味カテゴリ辞書を構成する．

以下でまず，ウィキペディアの記事に付与された記事カテゴリを手がかりに，意味カテゴリ辞書を獲得する手法について述べる．その後，語彙統語パターン[14,15]を利用して，意味カテゴリ辞書を獲得する手法について述べる．

3.1.1. ウィキペディアの記事カテゴリを手がかりとした意味カテゴリ辞書の構築

ウィキペディアは，具体物を中心とした様々な事物に関する常識的知識を記述した百科事典である．ウィキペディア中の各記事は，記事を分類するための**記事カテゴリ**が付与されている．これらの記事カテゴリは，見出し語(事物) の上位語を獲得する手がかりとして利用されている[16]．例

[2] ただし，第三層が定義されていないクラスについては，第二層のクラス名を意味カテゴリとして採用し，具体物に関する網羅性を保つように留意した．

[3] http://sites.google.com/site/extendednamedentityhierarchy/

[4] GPE (Geological and Political Entity) 地名にも政治的組織名にもなり得るエンティティのこと．例えば，「日本」．

[5] GOE(Geological and Organizational Entity) 地名、組織名にもなりえるような施設の名前．例えば，「東京大学」．

えば，特定の人について記述した記事には，「1987 年生」のような人物の生年による記事カテゴリが与えられているし，企業について記述した記事であれば，「大阪府の企業」のように直接的に上位語を含む記事カテゴリが付与されている．従って，各記事に付与された記事カテゴリと，意味カテゴリとの間の対応を取ることで，見出し語(事物) と意味カテゴリの対を得ることができる．

表 2 意味カテゴリと対応する記事カテゴリパターン

意味カテゴリ	パターン	ウィキペディア の記事カテゴリ
人	**/¥d+ 年生/**	1987 年生，1988 年生，. . .
法人	**/.+ 企業/**	大阪府の企業，千代田区の企業

本研究では，各意味カテゴリごとに，記事カテゴリのパターン(**記事カテゴリパターン**) を人手で記述し，パターンにマッチする記事カテゴリを付与された記事の見出し(事物) と意味カテゴリの対を，意味カテゴリ辞書に追加する．表 2 に記事カテゴリパターン(正規表現)の例を示す．

3.1.2. 語彙統語パターンに基づくウェブテキストからの意味カテゴリ辞書の構築

前節で述べた手法は，信頼性の高い記事カテゴリパターンのみを用いることで辞書の精度を高く保つことができる反面，辞書項目がウィキペディアの見出し語に限定されるため，辞書の網羅性の点で限界がある．

そこで我々は，膨大なウェブテキストを知識源として，上位関係獲得のための既存の語彙統語パターン[14,15]を利用して，意味カテゴリ辞書を構築する．具体的には，まず，各意味カテゴリに少数のクラス語を人手で列挙し(例:人→「首相」,「歌手」)，既存の上位下位関係の獲得パターン(例: A という B) の上位語(B) を各クラス語で具体化したパターン(例: A という**歌手**) を得る．このパターンを大量のウェブテキストに適用し，得られた各下位語(A) と意味カテゴリの対を辞書に追加する．本研究で利用した上位下位関係獲得のための語彙統語パターンのテンプレート(A: 下位語; B: 上位語) は以下の三つである．

P1: A という B + 「の」以外の助詞
P2: A などの B + 「の」以外の助詞
P3: B A

パターン P1, P2 については，名詞連続またはカギ括弧で囲まれた文字列を下位語(A) として獲得した．一方，パターン P3 については，カギ括弧で囲まれた文字列のみを下位語(A) として獲得した．

表 3 語彙統語パターンを用いた意味カテゴリ辞書の獲得

文	辞書項目
宇多田ヒカルという歌手が. . .	人－宇多田ヒカル
労働基準法などの法律を. . .	規則－労働基準法
映画「**ローマの休日**」を観た.	芸術作品－ローマの休日

カギ括弧を意味カテゴリ獲得の対象の語句を認識する手がかりとして用いることで,「それでもボクはやってない」や「ツァラトゥストラはかく語りき」といった，名詞(句) とみなせないような語句に対しても意味カテゴリ辞書を獲得することができることに注意されたい. 例として，表 3 左の各文から，右の辞書項目を獲得できる．

3.2. 多クラス分類器のためのラベル付きデータの獲得

本節では，前節で得た意味カテゴリ辞書を利用して，与えられた名詞(句) を適切な意味カテゴリに割り当てる分類器の学習ためのラベル付きデータを自動的に獲得する手法について検討する．

意味カテゴリ辞書を利用して，ウェブテキストからラベル付きデータをかくとくする最も単純な方法としては，意味カテゴリ辞書中で一つの意味カテゴリしか持たない名詞(句) を手がかりに, そのウェブでの出現全てを(その意味カテゴリのラベル付きデータとして) 収集する方法がある．しかしながら，この方法には幾つか根本的な問題がある．まず, 名詞(句)が所与の意味カテゴリ集合に含まれない意味を持つ場合があることに注意されたい. 例えば,「ジャンプ」は出版物(雑誌) であると同時に動作の種類を意味する．他の難しい例として，部分マッチの問題がある．例えば,「食堂」は施設であるが,「かもめ食堂」は芸術作品である．

そこで，本稿では，なるべくノイズの無いラベル付きデータを得るための工夫として，節 3.1.1 と節 3.1.2 の手法で共通して獲得された信頼性の高い辞書項目(名詞(句) －意味カテゴリ対; 例: 芸術作品－チャイナタウン) に着目し，節 3.1.2 でその意味カテゴリに対して具体化した語彙統語パターンでその名詞(句) を含んでいた文をウェブから収集する．例えば，

チャイナタウンという映画を観た.

という文では，チャイナタウンという名詞(句) の意味カテゴリが，映画という上位語を通じて，芸術作品に限局している．従って，この語彙統語パターンにマッチした部分文字列全体(「チャイナタウンという映画」) を，芸術作品という意味カテゴリに属する名詞(句) の例とみなし，ラベル付けする．

表 4 意味カテゴリ辞書とラベル付きデータ

意味カテゴリ	意味カテゴリ辞書				ラベル付けされた名詞(句)数
	エントリ数		適合率		
	wiki	blog	wiki	blog	
人	196,812	23,918	1.00	0.75	31,199
神	3,031	4,301	1.00	0.50	1,537
国際組織	148	159	1.00	0.95	21
公演組織	726	0	1.00	NA	0
家系	33	169	0.80	0.10	0
民族	0	776	1.00	0.10	0
競技組織	4,875	9,474	1.00	0.90	3,400
法人	37,192	32,767	1.00	0.80	43,426
政治的組織	3,021	1,668	1.00	0.75	1,198
温泉	1,310	3,454	1.00	0.85	1,090
ＧＰＥ	5,153	26,798	1.00	0.70	2,169
地域	158	7,506	1.00	0.85	16
地形	9,700	10,009	1.00	0.95	7,583
天体	355	793	1.00	0.80	1,023
遺跡	1,071	817	1.00	0.85	119
GOE	38,732	26,064	1.00	0.85	11,505
路線	10,245	4,366	1.00	0.55	2,028
材料	741	3,392	1.00	0.90	278
衣類	267	190	1.00	0.95	26
貨幣	287	116	1.00	0.70	37
医薬品	360	3,628	1.00	0.95	930
武器	3,610	3,392	1.00	0.40	433
賞	1,272	1,380	1.00	0.85	733
勲章	216	339	1.00	0.45	37
罪	369	1,279	1.00	0.80	882
キャラクター	1,421	11,611	1.00	0.85	2,334
乗り物	10,314	9,596	1.00	0.80	13,258
食べ物	3,190	21,112	1.00	0.95	4,527
芸術作品	36,918	265,111	1.00	0.95	739,210
出版物	2,751	23,692	1.00	1.00	51,229
主義方式	14,326	13,332	1.00	0.60	10,160
規則	2,415	3,702	1.00	0.60	3,898
称号	440	8,630	1.00	0.90	2,841
言語	1,746	1,121	1.00	0.80	1,005
単位	1,410	2,984	1.00	0.30	2,182
催し物	5,084	7,645	1.00	0.90	1,418
事故事件	915	4,743	1.00	0.50	81
自然災害	879	813	1.00	0.80	150
元素	510	119	1.00	0.85	181
化合物	2,340	209	1.00	1.00	115
鉱物	505	309	1.00	0.90	251
生物	3,897	10,147	1.00	0.90	3,333
生物部位	78	0	1.00	NA	0
動物病気	332	4,863	1.00	0.95	1,434
自然色	375	5,365	1.00	0.80	5,127
合計	561,859	409,530			952,404

4. 予備実験

4.1. 実験の設定

意味カテゴリ辞書の構築，及び，ラベル付きデータの自動生成の実験を行った．実験には，我々の研究室で収集している 2006 年から 2009 年の 4 年分のブログ記事をコーパスとして用いた．このコーパスは，約 1 億 9 千万記事，20 億文，510 億語からなるコーパスである．また，ウィキペディアは 2010 年 3 月 28 日時点のものを用いた．

ウィキペディアからの意味カテゴリ辞書の構築には，45 カテゴリに対して，181 個のパターンを記述した．また，ブログテキストからの意味カテゴリ辞書の構築に用いる上位語は，195 個用意した(表 5)．

4.2. 意味カテゴリ辞書の自動構築手法の評価

我々の手法により得られた意味カテゴリ辞書のエントリ数，及び適合率を評価した結果は表 4 のようになった．ここでエントリ数とは，意味カテゴリと名詞（句）の対の数のことを指す． 表中で blog は，ブログ記事から語彙統語パターンにより得られた意味カテゴリ辞書，wiki はウィキペディアの記事カテゴリから得られた意味カテゴリ辞書のエントリ数をそれぞれ示している．また，適合率については，各意味カテゴリごとに 20 個ずつエントリをランダムサンプリングし，人手で名詞（句）と意味カテゴリのペアが正しいかどうかを判断した．

この結果から，意味カテゴリごとにウィキペディアとブログ記事から抽出できた名詞（句）の数に偏りがあることが分かる．例えば，ブログ記事から得られた辞書では芸術作品のエントリが全体の約 47%を占めている．これは，ブログ記事で映画や楽曲などが話題に登りやすいことに加え，それらが P1 ~ P3 のパターンで出現する頻度が高いことが原因だと考えられる（典型的には映画「ローマの休日」といった表現が頻出する）．一方，ウィキペディアから得られた辞書では人のエントリが約 46%を占める．これは，ウィキペディアには人に関する記事が多く含まれていることが原因だと考えられる．

また、ウィキペディアから得られた辞書とブログ記事から得られた辞書に共通して含まれる名詞（句）の数を調べ

たところ，各々の約 6%， 9%と，少ないこと分かった．これは，ウィキペディアとブログ記事から得られる名詞(句)の性質が異なっていることを示している．ブログ記事から取得できた例として，「八海山」（食べ物），「中村」（人，駅）などがある．

適合率の評価結果から，ウィキペディアから得られた辞書の適合率は非常に高いことが分かる．これは，ウィキペディアの記事カテゴリが人手で十分整理されているため，精度の高い意味カテゴリ辞書を取得できたためだと考えられる．一方で，ブログ記事から得られた辞書の適合率は低い．誤り例としては次のよう文が挙げられる．

(2) a. 打撃が持ち味という選手.
b.「環境を守ることが大切だ」という政党.
c. ロックの神様「エルビス・プレスリー」
d. ストレートという武器

a，b の例は，「A という B」のパターンが，同格関係ではなく，修飾関係を表現するために用いられている．これはパターンが持つ曖昧性により誤りが生じている例である．また，c，d は神様や武器という単語が比喩的に用いられているため，正しいカテゴリが獲得できていない．このように比喩的に用いられる上位語に基づく語彙パターンでは，得られる辞書の精度が下がる傾向にあることが見て取れた．

ブログ記事のサイズの増減に伴う，得られる辞書のエントリ数の変化を示したグラフが図 1 である．このグラフから，コーパスのサイズに対して，ほぼ線形に辞書のエントリが増加していることが分かる．この結果は，今後のブログ記事の増加に伴って，意味カテゴリ辞書のエントリ数も更に増加すると期待できることを示している．

4.3. ラベル付きデータの自動生成手法の評価

ウィキペディアから得られた辞書によりラベル付けできたラベル付きの名詞(句)の数は表 4 のようになった．

全体としては, 952, 445 と大規模な訓練事例が得られた．これは，拡張固有表現に対して人手でラベル付きデータを作成した[17] (約 21 万ラベル)の 5 倍程度の数である．

また，意味カテゴリごとに，得られた訓練事例数に差があることが見て取れる．例えば，訓練事例数のうち芸術作品が約 78%程度を占める．これは，意味カテゴリごとに語彙統語パターンの出現頻度に差があることが原因と考えられる．

また，意味カテゴリがラベル付けされた文字列を，各カテゴリにつき 20 個ずつ，計 45×20 = 900 個のランダムダムサンプリングをして，人手で正解かどうかを確認した結果，900 個全てに正しいラベルが付いていた．

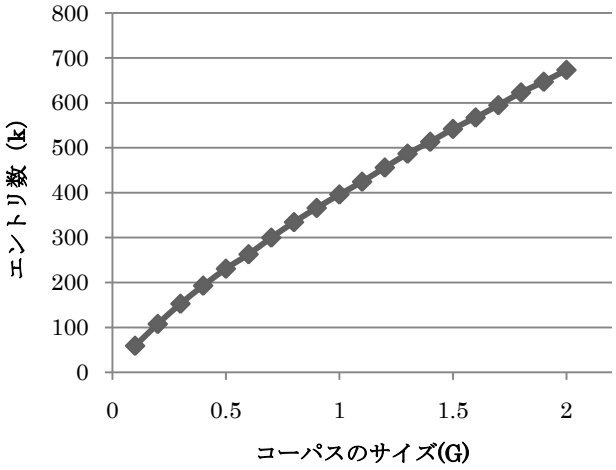

図 1 コーパスサイズの増加に伴う意味カテゴリ辞書のエントリ数の増加

5. まとめと今後の課題

本稿では，意味カテゴリタガーのラベルなしデータからの構築に向けて，意味カテゴリ辞書，及びラベル付きデータの自動生成手法の提案し，その評価を行った．評価実験の結果，高精度なラベル付きデータが意味カテゴリ辞書から自動生成できることが分かった．今後，得られたラベル付きデータを用いて，タガーの構築を行いたい．

今回，意味カテゴリ辞書の獲得には，ウィキペディアからの獲得に，181 個のパターン，ブログ記事からの獲得に 195 個の上位語を用意した．これは，人手でラベル付きデータを作成することに比較すると，手間がかからないが，所望の辞書が得られるまで，ある程度，試行錯誤でパターンを記述する必要がある．柔軟な意味カテゴリの設定を可能にするために，更に，廉価にラベル付きデータが生成できる手法の兼用を進めていきたい．

自動生成したラベル付きデータのラベル数は，意味カテゴリ間で偏りがあることが分かった．これは，意味カテゴリによって，語彙統語パターンでの文脈での出現頻度に偏りがあることが原因だと考えられる．そのため，今回，準備した 3 つの語彙統語パターンでの文脈での出現頻度が低い意味カテゴリについては，ラベル付きデータの生成が難しい．この問題点に対処するため，特定の語彙統語パターンの出現頻度に依存しない，より汎用的な手法の検討を進めていきたい．

参考文献

[1] David Nadeau and Satoshi Sekine, "A Survey of Named Entity Recognition and Classification," Journal of Linguisticae Investigationes, 30(1), pp.3-26, 2007.

[2] Scott Miller, Jethran Guinness, and Alex Zamanian, "Name Tagging with Word Clusters and Discriminative Training," In Proceedings of NAACL, pp.337-342, 2004.

[3] Hal Daume III, "Frustratigly Easy Domain Adaptation," In Proceedings of ACL, pp. 256-263, 2007.

[4] Jun'ichi Kazama and Kentaro Torisawa, "Inducing Gazetteers for Named Entity Recognition by Large-Scale Clustering of Dependency Relations," In Proceedings of ACL, pp. 407-415, 2008.

[5] Jun Suzuki and Hideki Isozaki, "Semi-Supervised Sequential Labeling and Segmentation Using Giga-Word Scale Unlabeled Data," In Proceedings of ACL, pp. 665-673, 2008.

[6] Dekang Lin and Xiaoyun Wu, "Phrase Clustering for Discriminative Learning," In Proceedings of ACL , pp. 1030-1038, 2009.

[7] Satoshi Sekine, Kiyoshi Sato, and Chikashi Nobata, "Extended Named Entity Hierarchy," 3rd international conference on Language resournce and evaluation(LREC-2002), 2002.

[8] Casey Whitelaw, Alex Kehlenbeck, and Nemanja Petrovic, *Web-scale named entity recognition*.: In Proceedings of CIKM., 2008.

[9] Andrew Carlson, Scott Gaffney, and Vasile Flavian, "Learning a named entity tagger from gazetteers with the partial perceptron," In AAAI Spring Symposium on Learning, 2009.

[10] Marti Hearst, "Automatic Acquisition of Hyponyms from Large Text Corpora," In Proceedings of COLING, pp. 539–545, 1992.

[11] Patrick Pantel, Eric Crestan, Arkady Borkovsky, Ana-Maria Popescu, and Vishnu Vyas, "Web-Scale Distributional Similarity and Entity Set Expansion," In Proceedings of EMNLP, pp. 938-947, 2009.

[12] 小町 守, 牧本 慎平, 内海 慶, and 颯々野 学, "ラプラシアンラベル伝播による検索クリックスルーログからの意味カテゴリの獲得," 人工知能学会論文誌, 25 巻, 1 号,C, pp. 196-204, 2010.

[13] 安藤 まや, 関根 聡, and 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 2003.

[14] Asuka Sumida, Kentaro Torisawa, and Keiji Shinzato, "Concept-instance relation extraction from simple noun sequences using a search engine on a web repository," In Proc. ISWC workshop on Web Content Mining with Human Language Technologies, 2006.

[15] Fabian M., Suchanek, Gjergji Kasneci, and Gerhard Weikum, "YAGO: A core of semantic knowledge unifying WordNet and Wikipedia," In Proc.WWW, 2007.

[16] 橋本 泰一, 乾 孝司, and 浩司 村上, *拡張固有表現タグ付きコーパスの構築*: 情報処理学会研究報告 自然言語処理研究会報告, 2008.

[17] 新納 宏幸 and 関根 聡, "拡張固有表現タガーの作成とその問題点の考察," 言語処理学会第 12 回年次大会, 2006.

表 5　意味カテゴリとそれに対応する記事カテゴリパターンと上位語

大分類	意味カテゴリ	記事カテゴリパターン	上位語
人	人	.*年生	選手，芸人，歌手，俳優，女優，アイドル...
神	神	神，.*の神	神，神様
組織	国際組織	国際機関，国際専門機関	国際組織，国際機関
	公演組織	.*のオーケストラ	公演組織
	家系	.*の士族，.*の家族	家系
	民族	NA	民族，州続，国籍
	競技組織	.*のサッカークラブ，.*野球チーム...	競技組織，チーム，サッカークラブ，リーグ
	法人	.*法人，.*の企業，.*の企業グループ	法人，社団法人，企業，会社，株式会社...
	政治的組織	.*の官公庁，.*の政党，.*の政治団体...	派閥，政府，政党，内閣，軍隊
地名	温泉	.*の温泉	温泉
	GPE	.*の町・字，.*の群，.*の都道府県，.*の州...	市，区，町，村，都市，都，道，府，県，州...
	地域	大陸，.*の地方	地域，地区，大陸，地域
	地形	.*の山地，.*の山，.*の島，.*の河川...	山，島，川，河，湖，海，湾
	天体	恒星，.*の惑星，88 星座	天体，恒星，惑星，星座
施設	遺跡	.*の考古遺跡，.*の古墳	遺跡，古墳
	GOE	.*の大学，.*の高等学校，.*の小学校...	施設，公共機関，学校，研究機関，研究所...
	路線	.*の鉄道路線，.*の道路，運河，.*の運河...	路線，道路，運河，航路，トンネル，橋
製品	材料	材料，.*材料	燃料，物質，染料，原材料
	衣類	装飾具，スポーツウェア，履物...	衣類
	貨幣	硬貨，金貨，銀貨，.*の硬貨	貨幣
	医薬品	一般用医薬品，漢方薬，抗生物質，抗炎症薬	薬，医薬品
	武器	.*兵器，.*ミサイル	武器
	賞	.*の賞	賞
	勲章	.*の勲章	勲章
	罪	犯罪	罪
	キャラクター	.*キャラクター	キャラクター
	乗り物	.*の車種，.*の列車，.*の航空機...	乗り物，車，自動車，列車，電車，飛行機...
	食べ物	ソフトドリンク，カクテル，果実酒，蒸留酒...	ソフトドリンク，カクテル，果実酒，蒸留酒...
	芸術作品	.*の美術館，絵画作品，.*テレビ番組...	芸術作品，作品，絵画，絵，番組，映画...
	出版物	.*の新聞，.*年創刊の雑誌	出版物，新聞，雑誌
	主義方式	.*の文化，.*の宗教，.*科学，流派，.*流派...	主義，文化，宗教，学問，流派，競技...
	規則	条約，.*の法律	規則，条約，法令，法律
	称号	職業	称号，地位，職業
	言語	.*の方言，語族，.*の言語	言語，言語
	単位	.*単位	単位，通貨
イベント	催し物	.*の祭，.*大会，議会，.*の議会	催し物，祭り，祭，競技会，大会，会議
	事故事件	.*の戦い，戦争	事件，事故，戦争
	自然災害	.*の災害，.*の地震	災害，地震
自然物	元素	元素	元素
	化合物	.*の化合物，化学物質	化合物
	鉱物	鉱物	鉱物
	生物	菌類，節足動物，イカ，タコ，貝類，二枚貝...	生物，菌類，軟体動物，節足動物，昆虫，魚...
	生物部位	器官	NA
病気	動物病気	病気	病気
色	自然色	色名	色