Wikipedia と Web テキストを利用した固有名の意味カテゴリの曖昧性解消

村本 英明 † 鍛治 伸裕 †† 吉永 直樹 †† 喜連川 優 ††
† 東京大学 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

1 はじめに

テキストからの情報抽出おいて人物や法人，製品等の固有物に関する言及が抽出の対象になることが多い．固有物に関する言及を抽出する際には，同じ名前が異なる固有物に対して与えられうることが問題になる．例えば，化学メーカーの「ライオン」についての言及を抽出したいとき，文字列との一致では以下の例のように法人名と生物名の両方の言及が抽出されてしまう．

(1) a. **ライオン**の接近を鳴き声で知らせる．
b. **ライオン**は漂白剤を 6 月に発売した．

言及抽出を行うには，(1)a.，(1)b. に出現する固有名「ライオン」がそれぞれどの固有物を参照しているかを明らかにする必要がある．

本研究では与えられた文中の固有名を**意味カテゴリ**(法人や生物など) に分類するタスクに取り組む．多くの場合，同一の意味カテゴリでは固有物の名前が重複することは少ないという直感から，固有名の意味カテゴリを明らかにできれば，その参照の曖昧性を解消することができると期待できる．

本論文ではこのタスクを教師あり学習を用いて解く方法を示す．教師あり学習にはラベル付きデータが必要となるが，人手でそれを作成するのにはコストがかかり困難である．そこで，我々は Wikipedia からラベル付きデータを半自動生成することでこの問題の解決を図る．また，各固有名が取り得る意味カテゴリを絞り込むために，大規模な Web テキストから自動構築した意味カテゴリ辞書を分類器と組み合わせて用いることで分類精度の向上を目指す．

2 関連研究

我々の研究と関連した研究として語義曖昧性解消 [2] がある．語義曖昧性解消では語句毎に人手で語義を定義し，その語義毎にラベル付けされたデータを人手で作成する必要があるため，分類器の構築にコストがかかる．[6][5] では Wikipedia の記事間リンクを用いることでこの問題の解決を図っているが，Wikipedia の記事に記載されていない語句に対しては分類器を構成できないという問題が残っている．これに対し本研究では分類対象語を語句横断的な意味カテゴリに分類するタスクを扱うので，予め与えられた意味カテゴリに対してラベル付きデータを用意すればよく，語義毎にラベル付きデータを作成する必要はない．

一方，テキスト中の語句に対して意味カテゴリを付与する技術として固有表現認識がある [3]．一般的な固有表現認識では「人名」や「組織」などの粗い意味カテゴリが用いられるが，これらの意味カテゴリでは固有名の曖昧性を扱うことのできないことがある．これに対して，我々は一般的な固有表現認識よりも詳細な粒度の意味カテゴリを用いる（3.1 節参照）．我々と同様に意味カテゴリを詳細化した研究として Whitelaw らの研究がある [1]．Whitelaw らは英語依存のヒューリスティクスを用いてラベル付きデータの生成を行っているが，我々の手法は Wikipedia の使用しか前提にしないため，英語以外の言語にも適用しやすい．

3 Wikipedia を用いた意味カテゴリ分類器の学習

本節では Wikipedia からラベル付きデータを自動構築し，意味カテゴリ分類器を構築する方法について説明する．

3.1 意味カテゴリ

本研究の目的は，固有名を意味カテゴリに分類することによって，固有名の曖昧性を解消することにある．

表 1: 拡張固有表現から構成した 46 の意味カテゴリ (ID; カテゴリ名で表記している)

1; 人物，2; 神，3; 国際組織，4; 公演組織，5; 家系，6; 民族，7; 競技組織，8; 法人，9; 政治的組織，10; 温泉，11;GPE[1]，12; 地域，13; 地形，14; 天体，15; 遺跡，16;GOE[2]，17; 路線，18; 製品その他，19; 材料，20; 衣類，21; 貨幣，22; 医薬品，23; 武器，24; 賞，25; 勲章，26; キャラクター，27; 乗り物，28; 食べ物，29; 芸術作品，30; 出版物，31; 主義方式，32; 規則，33; 称号，34; 言語，35; 単位，36; 催し物，37; 事故事件，38; 自然災害，39; 元素，40; 化合物，41; 鉱物，42; 生物，43; 生物部位，44; 動物病気，45; 自然色，46; その他

したがって，分類先の意味カテゴリは，同一意味カテゴリで名前の重複が少なくなるくらい詳細であることが望ましい．

そこで，本研究では，既存の固有表現認識よりも詳細な意味カテゴリを定義した拡張固有表現 [7] を元に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の第二層 (例: 人物，法人，材料，自然現象，生き物) を一つの意味カテゴリとみなした．加えて，拡張固有表現のどのカテゴリにも対応しない意味カテゴリ（例えば一般名詞）の分類先として,「その他」という意味カテゴリを準備した．その結果得られた意味カテゴリは計 46 個であった (表 1)．なお，提案手法は，拡張固有表現に依存した手法ではないため，任意の意味カテゴリ集合に適応できることに注意されたい．

3.2 Wikipedia を用いたラベル付きデータの半自動生成手法

本節では，Wikipedia の記事間リンク [6][5] と記事と意味カテゴリとの対応関係を用いて，意味カテゴリ分類器学習のためのラベル付きデータを半自動生成する手法について説明する．

Wikipedia は固有物を中心とする様々な事物に関する記事から構成され，対応する記事がある固有物については，以下の例のように文中で記事へのリンクが，“[リンク先記事タイトル | アンカーテキスト]” という形式で記述されている．

(2) a. プリウスは，[トヨタ自動車 | トヨタ] が発売したハイブリッドカーである．

ここで，リンク先の記事「トヨタ自動車」と意味カテゴリ「法人」を対応付けることができれば，(2) の「トヨタ」に正解意味カテゴリが付与されたラベル付きデータを得ることができる．

表 2: 上位語を介した記事と意味カテゴリの対応

意味カテゴリ	上位語	記事
法人	メーカー	ライオン (企業)，トヨタ自動車，ソニー，...
	銀行	三菱東京 UFJ 銀行，三井住友銀行，...
生物	哺乳類	ライオン (生物)，うさぎ，...
	魚類	ウナギ，アナゴ，...

各記事と意味カテゴリとの対応付けを人手で行うのはコストが大きい．そこで，記事[3]の上位語を Wikipedia の先頭文[4]から自動抽出し [9]，その上位語と意味カテゴリとの対応をとることで，上位語を介して間接的に記事と意味カテゴリを対応付ける．例えば上位語「メーカー」と意味カテゴリ「法人」を対応付ければ，上位語が「メーカー」の記事全て（「トヨタ自動車」,「ソニー」等）が意味カテゴリ「法人」と対応付けられたことになる (表 2)．本稿では，意味カテゴリと対応付けられた上位語のことを**シード上位語**，記事のことを**シード記事**と呼ぶ．

なお，人手で上位語と意味カテゴリの対応ルールを記述する際には，一つの対応ルールによって，できるだけ多くのラベル付きデータを得るために，Wikipedia 中でリンクされている頻度が高い記事に出現する上位語から順にルールを記述する．

3.3 分類器の学習方法

本節では 3.2 節で得られたラベル付きデータから，意味カテゴリ分類器を学習する手法について説明する．

意味カテゴリ分類器は一対他法で学習を行う．具体的には,「法人」の意味カテゴリについての学習を行う際には,「法人」カテゴリの訓練事例を正例とし,「法人」カテゴリ以外の全カテゴリの訓練事例は全て負例として 2 値分類器の学習を行う．我々が扱っている問題は分類先の意味カテゴリ数が 46 と多いため，一対他法の学習において，正例のデータ数が負例のデータ数に比べて少ないことが分類器の性能上の問題となる．そこで，正例のデータ数が負例のデータ数と同数になるようにサンプリングしたデータを用いて学習することで，この問題の解決を図る [4]．

[1]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」.

[2]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」.

[3]これは正確には記事タイトルのことであるが，文脈から明らかな場合は記事タイトルのことを単に記事と呼ぶこととする．

[4]例えば，記事「キリン（企業）」の先頭文は「ライオン株式会社は、洗剤、石鹸、歯磨きなどトイレタリー用品、医薬品、化学品を手がける日本の大手**メーカー**.」となっており，文末の名詞に着目することで上位語「メーカー」が抽出できる．

意味カテゴリ「その他」については，正例を準備するのが困難なため，全て分類器が負例と判断した場合にのみ「その他」に分類するとする．また，分類対象語に対して，複数の意味カテゴリに対する分類器が正例と判断した場合は，分類器の出力したスコアが最も高い意味カテゴリに分類する．

機械学習のアルゴリズムには平均化パーセプトロンを用いて学習を行う．素性は正解ラベルが付与された語句の出現する文の bag of words と，前後 n-gram(n = 1〜3)，及び，係り先の動詞の原型を用いる．ただし，ストップワード（「こと」,「とき」等）と，全ラベル付きデータ中で 5 回未満しか出現しない低頻度の素性は削除する．

4 意味カテゴリ辞書の自動構築

本研究で導入した意味カテゴリは 46 個であるが，各々の語句が取りうる意味カテゴリはそのうちの一部である．例えば，我々が調べた限り「ライオン」は「法人」,「生物」,「芸術作品」のカテゴリだけ考えれば十分である．こうしたタスクの性質を活用するために，語句が取り得る意味カテゴリを列挙した辞書 (意味カテゴリ辞書と呼ぶ) を構築し，分類対象語が取りうる意味カテゴリを絞り込むことで分類精度の向上を図る．

意味カテゴリ辞書は，Web テキストから語彙統語パターンにより上位語-下位語 (分類対象語) のペアを抽出し，抽出された上位語を意味カテゴリに対応付けることで構築する．

語彙統語パターンは,「w という h」,「w などの h」,「w 以外の h」,「h「w」」の 4 種類を用いる．パターン中の w は下位語 (分類対象語) で，h は上位語である．

上位語 h と意味カテゴリの対応付けにはシード記事とシード上位語 (3.2 節) を用いる．まず h が意味カテゴリ c のシード上位語に含まれる場合は，h を c に対応付ける．それ以外の場合は，h を c に対応付けたときに c に属する分類対象語が抽出できる確率 $p(c\,|\,h)$ を，シード記事を用いて以下の式で見積もり，それが閾値以上の場合は h と c を対応付ける．

$$p(c\,|\,h) = \frac{freq(c,h)}{\sum_{c'} freq(c',h)}$$

ここで，$freq(c,h)$ はカテゴリ c のシード記事に含まれる単語と h が語彙統語パターンで共起した回数を示している．

表 3: 評価に用いた語句 (カッコ内は評価データ中で付与されている意味カテゴリ ID)

イルカ (1,42)，エセックス (11,23,27)，オアシス (4,12,13,46)，オセロ (4,18,29)，オレンジ (11,28,42,45)，カサブランカ (11,29)，クリーム (4,28)，サラトガ (11,23,27)，サルサ (28,,31)，シカゴ (4,11)，ジェネシス (4,23,27)，ジャングル (12,13,31,46)，スズキ (8,42)，セオドア・ルーズベルト (1,23,27)，タンポポ (4,42)，ハイヒール (4,20)，ハンニバル (1,29)，ホーネット (18,23,27)，ボール (18,23,46)，レーダー (1,2,18,23,26)，レベル (8,35,46)，レンジャー (23,27,33)，ヴァルナ (2,11)，安全地帯 (4,46)，銀河 (14,17,23,27)，少年 (30,46)，石 (35,41,46)，長征 (27,37)，忍者 (4,33)，羞恥心 (4,29,46)

5 評価実験

5.1 実験の設定

ラベル付きデータの作成のために，600 の記事上位語に対して人手で意味カテゴリとの対応付けを行い，これを上位語シードとして用いた．得られた記事シード数は 281,188 であった．また，5,409,944 インスタンスのラベル付きデータが得られた．上記のデータを用いて分類器と意味カテゴリ辞書の構築を行った．

意味カテゴリ辞書の自動構築に用いた Web テキストは，我々の研究室で収集している 2006 年から 2009 年の 4 年分のブログ記事である．このテキストデータは，約 1 億 9 千万記事，20 億文からなる．Wikipedia の先頭文からの記事上位語の抽出には上位下位抽出ツール[5]を利用し，形態素解析には MeCab[6]，構文解析には J.DepP[7]を利用した．

評価に用いるデータは，ラベル付きデータと同様に Wikipedia から作成する．まず，Wikipedia から 10 回以上リンクしている記事が 2 つ以上ある語句を抽出する．次に，記事を意味カテゴリに対応付ける．この作業は半自動では行わず人手で行った．30 語ランダムにサンプリングして評価に用いた (表 3)．なお，評価に用いた語句に関するラベル付きデータは，分類器の訓練には用いないものとする．これは，新語，新用法に対しての頑健性を評価するためである．

5.2 実験結果

意味カテゴリ辞書を構築するときの閾値を変化させたときの分類精度の変化を評価した．閾値が 0 の場合は意味カテゴリ辞書による絞込みを行わない場合の分類精度に相当している．分類精度は分類対象語と正解

[5] http://alaginrc.nict.go.jp/hyponymy/index.html
[6] http://mecab.sourceforge.net/
[7] http://www.tkl.iis.u-tokyo.ac.jp/~ynaga/jdepp/

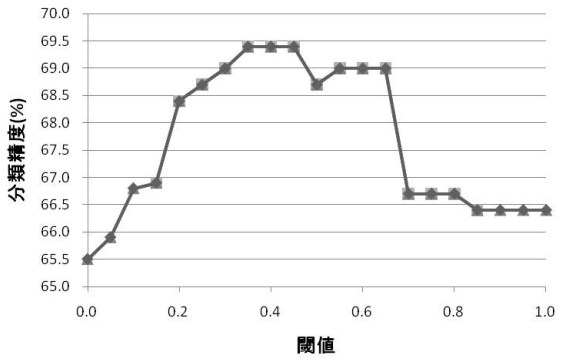

図 1: 分類精度

意味カテゴリのペアを 1 回の試行とした場合の正答率のマクロ平均を用いて計算した．

図 1 から閾値が 0.0〜0.2 では分類精度が向上していることが見て取れる．これは，閾値が大きくなるにつれて，意味カテゴリ辞書に追加される候補意味カテゴリの数が少なくなり，絞込みによる分類精度の向上の効果が大きくなるためである．閾値が 0.35〜0.45 の 5 ときの分類精度は 69.4 %で，意味カテゴリ辞書による絞込みを行はない場合に比べて 4 %分類精度が向上した．なお，閾値が 0.7 以上では分類精度が悪化しているのは，閾値を上げると意味カテゴリ辞書に含まれる候補意味カテゴリの数が減少し，正解意味カテゴリが意味カテゴリ辞書の候補意味カテゴリ中に含まれなくなるからである．以下，分類結果について考察する．

文脈から正解意味カテゴリを判断することが難しい場合でも，意味カテゴリ辞書を用いて候補意味カテゴリの絞り込みを行うことで，正しく分類することができるようになった例を以下に示す．

(3) a. **セオドア・ルーズベルト**大統領は全国レベルの行政改革を行った．

例えば，(3)a. の「セオドア・ルーズベルト」は国名 (例えば「アメリカ」) と置き換えても文脈上問題ないため,「国名」に分類するのか「人名」に分類するのか文脈情報から判断することは困難である．意味カテゴリ辞書を用いて分類対象語の候補意味カテゴリを絞ることで，上記の問題を部分的に解決することができた．

分類先の意味カテゴリが「その他」の語句 (例えば「羞恥心」) については，分類精度が他よりも低い傾向が見られた．この結果から,「その他」の意味カテゴリに対する我々の提案手法 (候補意味カテゴリに対する分類器が全て負例と判断した場合にのみ「その他」に分類する) が有効でないことが分かる．なお，正解意味カテゴリが「その他」のみの事例を評価から除くと分類精度の平均値は 73.7%まで上昇した (閾値 0.45)．

6 まとめ

本稿では，固有名の曖昧性解消を目的とした意味カテゴリ分類のタスクを導入し，Wikipedia から半自動で生成したラベル付きデータから教師あり学習により分類器を構築した．また，Web テキストから語彙統語パターンにより構築した意味カテゴリ辞書を分類器と合わせて用いる手法を提案し，評価実験により，分類精度が向上することを確認した．特に，意味カテゴリ辞書による絞込みは (3)a. の例のように，文脈情報からは正解意味カテゴリの判別が難しい場合に有効であることが分かった．

今後の課題としては，ラベル伝播 [8] 等のグラフ理論に基づくアルゴリズムを用いて，より精度の高い意味カテゴリ辞書の構築やすることや，より少ない人手での作業で分類器を構築する手法について検討することがあげられる．

参考文献

[1] C. Whitelaw, A. Kehlenbeck, N. Petrovic. Web-scale named entity recognition., In Proc. of CIKM, pp. 123–132. 2008.

[2] D. McCarthy. Word Sense Disambiguation:An Overview. Language and Linguistics Compass, 3(2), pp. 537–558. 2009.

[3] D. Nadeau, S. Sekine. A Survey of Named Entity Recognition and Classification, Journal of Linguisticae Investigationes30(1), pp. 3–26. 2007.

[4] N. V.Chawla, N. Japkowicz, A. Kolcz. Editorial: Special Issue on Learning from Imbalanced Data Sets. SIGKDD Explorations, 6(1), pp. 1–6. 2004.

[5] R. Bunnescu, M. Pasca. Using Encyclopedic Knowledge for Named Entity Disambiguation. In Proc of EACL, 2006.

[6] R. Mihalcea. Using Wikipedia for Automatic Word Sense Disambiguation. In Proc. of NAACL-HLT, pp. 196–203. 2007.

[7] S. Sekine, K. Sato, C. Nobata., Extended Named Entity Hierarchy. In Proc. of LREC. 2002.

[8] X. Zhu, Z. Ghahramani Learning from Labeled and Unlabeled Data with Label Propagation. Tech report CMU-CALD-02-107. 2002.

[9] 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.