情報処理学会論文誌　Vol. 51　No. 10　1234–1242 (Oct. 2010)

意味カテゴリに基づく語義曖昧性解消における Web 資源の活用について

村本英明[†1]　鍛治伸裕[†2]
吉永直樹[†2]　喜連川優[†2]

従来の語義曖昧性解消の技術は，語句をシソーラスで定義された語義に分類していたため，シソーラスに未記載の語句や語義は扱うことができないよいう問題点があった．本稿では，上記の持つ問題点を解決するために，語句を意味カテゴリに分類することで語句の語義の曖昧性の解消を目指す．分類器の学習に用いる訓練データを，人手で作成するにはコストがかかるため，Wikipedia の記事間リンクを用いて，訓練データを半自動的に生成する手法を提案する．加えて，大規模な Web テキストから取得した文脈情報を利用して，分類対象語が取りうる意味カテゴリの絞り込みを行うことで，分類精度の向上を図る．

Class-Based Word Sense Disambiguation Using Web Resource

Hideaki Muramoto,[†1] Nobuhiro Kaji,[†2]
Naoki Yoshinaga[†2] and Masaru Kitsuregawa [†2]

英語のアブストを書く．

1. はじめに

近年のインターネットの爆発的な普及に伴い，人々はブログ等の CGM(Consumer Generated Media) を通じて自由に情報発信を行うことが可能となった．これにより，インターネット上には，人々の意見や感情が表出したテキストが多数流通することとなった．このようなテキストデータは社会分析やマーケティングなどの情報源として高い潜在的価値を有することから，これを解析するための自然言語処理技術に大きな期待が集まっている．

こうしたテキストを計算処理する際には，語義の曖昧性，すなわち，一つの語句が複数の語義を持ちうることが大きな問題となる．例えば「ライオン」という語は，少なくとも，会社と動物の二つの意味を持つ．そのため，Web テキストから会社の「ライオン」に関する言及を抽出したい場合には，会社の「ライオン」と動物の「ライオン」を何らかの方法で区別する必要がある．

1.1 意味カテゴリにもとづく語義曖昧性解消

語句の意味が持つ曖昧性を解消するためのテキスト処理技術は，語義曖昧性解消と呼ばれ，自然言語処理の分野において古くから研究が行われてきた．語義曖昧性解消の問題は，テキストに出現する語句に対して，WordNet[9] や国語辞典などの辞書資源において定義された語義を割り当てる分類問題として定式化されることが一般的である[10]．英語では WordNet，日本語では国語辞典が使われることが多い．

しかし，このような枠組みでは，新語や固有名など，辞書資源に未登録である語句を扱うことが難しい．さらに，辞書に登録されている語句であっても，新しい語義が作られることがあり，これに対応することも困難となっている[?),11]．これらはいずれも，Web テキストのような，多様な語句が出現するテキストを処理対象とした場合には無視することのできない問題となる．

このような問題意識にもとづき，本論文では，語義曖昧性解消の問題を，語句の意味カテゴリへの分類問題として解くことを試みる．... 節で詳しく述べるように，我々は 46 の意味カテゴリを定義し，テキストに出現する語句を，そのいずれかに分類することによって語義曖昧性の解消を行う．この枠組みでは，各語句に対する語義を列挙した辞書資源の存在を仮定しないため，既存の辞書資源に登録されていない新語や新語義に対しても柔軟に対応可能となる．(以下では，我々の扱うタスクのことを意味カテゴリ分類と呼ぶ．) 本タスクと語義曖昧性解消は全く異なる考えに基づいているわけではなく，両者の本質的な相違点は分類を行う語義の粒度であり，.. 節でより詳しく議論を行う．

1.2 Web 資源の活用

本論文では，語句を意味カテゴリに分類する分類器を構築するために，Web 資源を有効利用することによって，手軽にカテゴリ分類器を構築する．具体的には以下の二点について

†1 東京大学　情報理工学系研究科
Graduate School of Information Science and Technology, the University of Tokyo
†2 東京大学 生産技術研究所
Institute of Industrial Science, the University of Tokyo

© 2010 Information Processing Society of Japan

議論を行う．

- 高精度な分類器の構築のためには，大量のラベル付きデータが必要不可欠となるが，それを全て手作業によって作成することはコスト面から考えて望ましくない．そこで，Wikipedia の記事間リンクを利用することによって，人手で作成した少数の規則から大規模なラベル付きデータを構築する方法を提案する．
- 従来の語義曖昧性解消では，WordNet などの辞書資源を利用することによって，各単語が持ちうる語義が絞り込まれていた．例えば，SENSEVAL-2 日本語タスクにおいては，一つの語に対する平均語義数は 6.5 である．一方，我々のタスク設定において，分類先となるカテゴリ数は語句によらず 46 となっており，分類問題として難しい設定となっていると考えられる．そこで，語句の意味カテゴリを列挙した辞書を大規模 Web テキストから構築し，その辞書を利用して分類先となるカテゴリを絞り込む手法を提案する．

2. 関 連 研 究

2.1 語義曖昧性解消と固有表現認識

我々の導入した分類タスクは，語義曖昧性解消および固有表現認識と関連が深い．これら三つの問題における本質的な相違点は，語句の分類先となるカテゴリの粒度であり，

まず，語義曖昧性解消においては，分類先カテゴリは WordNet や国語辞典などの辞書資源において定義されている語義となる．一般的に，そのような語義は詳細な粒度で区別されているという特徴がある．しかし，そのような詳細な語義を網羅的に列挙した辞書作成は困難であり，辞書に登録漏れとなった語句や語義を扱うことができない．また，実用上の観点から見ても，例えば WordNet で定義されている語義集合が，機械翻訳や質問応答といったアプリケーションにおいて必要とされるよりも細かく設計されていることは，よく指摘されている事実であり，そのような詳細語義が必要であるかについては疑問が残る[6]．

一方，固有表現認識は，単語ごとに語義を列挙したのではなく，単語横断的な意味カテゴリを定義しておくという立場をとる．そのため，任意の単語を処理対象とすることが可能であり，語義曖昧性解消で問題となった未知語や未知語義の問題は発生しない．しかし，従来の固有表現認識において用いられているカテゴリ数は高々10 程度であり[⋆1]，語句の意味を十分に区別できるような問題設定になっているとは言い難い．例えば，IREX で定義されている固有表現カテゴリに基づくと，「洋服のワンピース」と「漫画のワンピース」は共に人工物となってしまい，これら二つの意味を区別することができない．我々の扱う問題は，任意の語句を処理対象としつつ，従来の固有表現認識よりも詳細な意味カテゴリを用いているものとして位置付けることができる．

固有表現認識よりも詳細なカテゴリに語句を分類するという点で，拡張固有表現認識[12]と共通している．実際に我々の研究でも拡張固有表現をもとに設計したカテゴリを分類先として用いている (3.1 節参照)．我々の研究は語義曖昧性解消を目的としているのに対して拡張固有表現認識は固有表現抽出を目的としており，タスク設定の目的が異なっている．

Ciaramita らの研究は，我々の研究と同様にカテゴリ分類にもとづく語義曖昧性解消に取り組んだでいる[4]．Ciaramita らの研究は，人手でタグ付けされたコーパスやシソーラスを用いているのに対して,．我々の研究は，訓練データや辞書は自動生成するため，Wikipedia と大規模な言語資源という人手での作業なしに入手できる言語資源しか前提としない点が Ciaramita らの研究と比較して優れている点である．また Ciaramita らの研究が一般表現が中心の意味カテゴリを用いているが，我々の研究は Web テキスト分析の対象になることが多い商品名や会社名などの固有表現を中心としたカテゴリを用いている点も異なっている．

2.2 語義曖昧性解消や固有表現認識における Web 資源の利用

どのような意味カテゴリを用いるかは，応用に強く依存することから，一般的に最適な意味カテゴリを定義することは事実上不可能であると考えている．そのため，応用に依存しにくいカテゴリ集合を定義することだけでなく，各研究者が意味カテゴリを自由に設計しやすい手法を構築することも重要であると我々は考えている．

人手で作成した訓練データやシソーラスを用いる手法では，自由に意味カテゴリを設計することが難しい．なぜなら，意味カテゴリを設計する度に，訓練データやシソーラスを人手で再作成しなといけないため，そのコストが大きくなるためである．そこで，我々は人手で作成したコーパスやシソーラスの代わりに，Wikipedia や Web テキストを用いて訓練データや辞書を自動構築する手法を提案する．以下，語義曖昧性解消や固有表現認識に Wikipedia や Web テキストを利用した研究について述べる．

語義の曖昧性解消に Wikipedia を利用した研究として Mihalcea や Bunescu らの研究がある[8)2]．Mihalcea や Bunescu らは Wikipedia の記事を一つの語義としてみなし，語句を Wikipedia の記事に分類することで，語義の曖昧性の解消を図っている．Mihalcea は記事間リンクに着目し，アンカーテキストをラベル付きデータとみなし，教師あり学習で分類器を構築している．また Bunescu らはを各々の記事と分類対象語の周辺の文脈との類似度に

⋆1 ここには時間表現や数値表現などのカテゴリなども含まれていることに注意されたい．

© 2010 Information Processing Society of Japan

基づき，分類対象語を記事に分類する手法を提案している．我々の研究は，Mihalcea の研究と同様に記事間リンクを手掛かりにラベル付きデータの自動生成を行うが，記事を意味カテゴリに対応付けることで，意味カテゴリ分類に必要なラベル付きデータを自動生成する点が異なっている．また，Mihalcea や Bunescu らの研究は記事に分類対象語を分類するため，記事が存在しない語句や語義については扱うことができない．一方，我々の手法は意味カテゴリに分類対象語を分類するため，Wikipedia の記事に記載されていない語句についても扱える点が優れている点である．

また，固有表現認識に Web テキストから訓練データを自動生成を行った研究として Whitelaw らの研究がある[13]．Whitelaw らは，Web テキスト分析を Whitelaw らの手法は分類器の学習の際に言語依存の性質を用いているため，汎用性が低いという問題がある．一方，我々の研究はこうした言語依存の性質を用いない．

3. Wikipedia を用いた意味カテゴリ分類器の学習

本節では Wikipedia から訓練データを自動生成し，得られた訓練データから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，語句を意味カテゴリに分類することによって，語句の意味の曖昧性を解消することにある．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，詳細な意味カテゴリを定義した拡張固有表現[12] 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．

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，12：地域，13：地形，14：天体，15：遺跡，16：GOE，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：罪，27：キャラクター，28：乗り物，29：食べ物，30：芸術作品，31：出版物，32：主義方式，33：規則，34：称号，35：言語，36：単位，37：催し物，38：事故事件，39：自然災害，40：元素，41：化合物，42：鉱物，43：生物，44：生物部位，45：動物病気，46：自然色，47：その他

3.2 Wikipedia を用いた訓練データの半自動生成手法

本節では，Wikipedia の記事間リンク[8)?] 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訓練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．

(1) 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，(1) の「トヨタ」に正解意味カテゴリが付与された訓練データを得ることができる．

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[⋆1]の上位語を Wikipedia の先頭文[⋆2]から自動抽出し[?]，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**

⋆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」.

⋆1 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．

⋆2 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．

© 2010 Information Processing Society of Japan

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，...

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，できるだけ多くの訓練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 節で得られた訓練データから，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．学習アルゴリズムには，提案手法により Wikipedia から自動生成した訓練データは大規模なため，高速に学習ができる平均化パーセプトロン[5] を用いる．

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[3]．

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．

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，全訓練データ中で 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．

5. 評 価 実 験

提案手法で Wikipedia から自動生成した訓練データが意味カテゴリ分類のタスクに用いることができるかどうかの確認と，Web テキストから自動構築した意味カテゴリ辞書による絞込手法が分類精度向上に寄与することの確認のために評価実験を行った．

5.1　実験の設定

評価実験では，Wikipedia は 2010 年 3 月 17 日のものを用いた．用いた Wikipedia は約 130 万記事，約 1500 万文からなる．また，意味カテゴリ辞書の自動構築には，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用いた．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．

分類精度の評価に用いるデータは，訓練データと同様に Wikipedia から作成する．評価実験では一義語に対する分類精度と多義語に対する分類精度の両方の評価を行う．具体的に

© 2010 Information Processing Society of Japan

表 3　評価に用いた語句 (多義)(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, 43)，エセックス (11, 23, 28)，オアシス (4, 12, 13, 47)，オセロ (4, 18, 30)，オレンジ (11, 29, 43, 46)，カサブランカ (11, 30)，クリーム (4, 29)，サラトガ (11, 23, 28)，サルサ (29, 30, 32, 47)，シカゴ (4, 11)，ジェネシス (4, 23, 28)，ジャングル (12, 13, 30, 32, 47)，スズキ (8, 43)，セオドア・ルーズベルト (1, 23, 28)，タンポポ (4, 43)，ハイヒール (4, 20)，ハンニバル (1, 30)，ホーネット (18, 23, 28)，ボール (18, 23, 47)，レベル (8, 36, 47)，レンジャー (23, 28, 34)，レーダー (1, 2, 18, 23, 27)，ヴァルナ (2, 11)，安全地帯 (4, 47)，少年 (31, 47)，忍者 (4, 34)，石 (36, 42, 47)，羞恥心 (4, 30, 47)，銀河 (14, 17, 23, 28)，長征 (28, 38)

は，下記のように Wikipedia から評価用データの作成を行った．

一義語データ　Wikipedia の記事間リンクにおいて，100 回以上リンクされている記事をサンプリングする．そして，得られた記事を人手で意味カテゴリと対応付ける．二つ以上の意味カテゴリと対応付けられる記事は削除する．対応付けた結果,「その他」の意味カテゴリに該当する記事は評価データとして用いないこととする．こうして得られた 30 の記事を分類対象語とし，こういった記事にリンクしているアンカーテキストを分類対象語とした評価用データを作成した．評価に用いた分類対象語は表 4 になった．

多義語データ　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う．人手での対応付けを行った結果，2 つ以上の意味カテゴリに属する記事を評価用の分類対象語として用いる，こうした記事 40 語をサンプリングし，こういった記事にリンクしているアンカーテキストを評価用データとして用いた．評価に用いた分類対象語は表**??**のようになった．

なお，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⋆1を利用し，形態素解析には MeCab⋆2，構文解析には J.DepP⋆3を利用した．

⋆1 http://alaginrc.nict.go.jp/hyponymy/index.html
⋆2 http://mecab.sourceforge.net/
⋆3 http://www.tkl.iis.u-tokyo.ac.jp/~ynaga/jdepp/

表 4　評価に用いた語句 (一義)(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)

つるの剛士 (1)，もののけ姫 (30)，アトランタ (11)，アルストム (8)，インド洋 (13)，ウラジオストク (11)，オペレッタ (30)，セルビア人 (1)，セーラー服 (20)，ソルボンヌ大学 (16)，ツァーリ (34)，テイチクエンタテインメント (8)，テニス (32)，ヒスパニック (6)，ボーイング (8)，ヤッターマン (27)，ロッテルダム (11)，京都 (11)，信越放送 (8)，名古屋市営地下鉄 (17)，国立劇場 (16)，執事 (34)，強制収容所 (16)，待合室 (16)，情熱大陸 (30)，戦闘機 (28)，最高経営責任者 (34)，朝日放送 (8)，漢 (6)，漫才師 (34)，特殊相対性理論 (32)，笑福亭鶴瓶 (1)，紀行番組 (30)，統一地方選挙 (37)，西条市 (11)，連合艦隊司令長官 (34)，重油 (18)，高射砲 (23)，高野山 (13)，魔法のプリンセスミンキーモモ (30)

5.2 実験結果

Wikipedia からの訓練データの自動生成には，600 個の上位語に対してルールを記述して上位語シードとして用いた．提案手法により，約 540 万の訓練事例が得られた．得られた訓練事例のカテゴリごとの事例数は，表**??**になる．

提案手法により得られた訓練データと，他の言語資源 (IREX[7]，拡張固有表現タグ付きコーパス[?)]) と比較を行ったのが表 6 である．この表より，提案手法により得られた訓練データは従来の言語資源と比較して大規模な訓練データであると言える．また，訓練データが大規模なため，出現頻度の少ないカテゴリに対しても多くの訓練データが得られていることが分かる．

次に，訓練データのサイズと分類精度の関係を調べるために，ランダムサンプリングで訓練データのサイズを変化させたときの，分類精度の変化を調べた (図 1)．図 1 の最も左の点が，本実験で得られた全ての訓練データを用いて学習した場合の結果を示している．一義語データ，多義語データともにこの図から訓練データ数が増加することで分類精度が上昇していることが見て取れる．また，得られた訓練データを全て用いても分類精度はまだ向上していることが分る．こうした点から，詳細意味カテゴリ分類のタスクにおいて，高精度で分類を行うには，大規模な訓練データが必要であるといえ，既存の言語資源に比べて大規模なデータを得ることができる提案手法は，分類精度の面でも有効であると言える．

一方，提案手法は Wikipedia の記事間リンクを訓練データとして用いているため，記事間リンクの総数以上の訓練データを得ることはできない．. 具体的には，評価実験では Wikipedia 中の記事間リンクの約 80 %を用いて訓練データの生成を行っているため，多くてもあと 25 %程度しか訓練データを増やすことはできない．そのため，Wikipedia の記事間リンク以外の手掛かりで方法で訓練データを生成する手法を今後検討していきたい．

© 2010 Information Processing Society of Japan

表 5　半自動生成されたラベル付きデータ数

意味カテゴリ	ラベル付きデータ数	意味カテゴリ	ラベル付きデータ数
人物	859,435	賞	18,134
神	14,454	勲章	2,679
国際組織	4,077	罪	7,203
公演組織	34,195	キャラクター	16,621
家系	29,665	乗り物	73,036
民族	13,047	食べ物	42,667
競技組織	76,437	芸術作品	382,701
法人	395,991	出版物	58,750
政治的組織	183,641	主義方式	196,006
温泉	4,286	規則	54,937
GPE	1,294,979	称号	103,700
地域	10,614	言語	84,948
地形	209,546	単位	35,871
天体	15,173	催し物	85,566
遺跡	2,160	事故事件	94,170
GOE	459,867	自然災害	5,842
路線	160,917	元素	16,781
製品その他	172,486	化合物	50,605
材料	2,283	鉱物	5,137
衣類	4,742	生物	92,794
貨幣	3,871	生物部位	10,946
医薬品	3,040	動物病気	16,434
武器	22,484	自然色	5,719
		合計	5,438,637

表 6　既存の言語資源との比較

	IREX	拡張固有表現タグ付きコーパス	提案手法
用いた言語資源	新聞	新聞，白書	Wikipedia
作成方法	人手	人手	半自動
訓練データ数	19,254	326,966	5,438,637

提案手法の辞書の絞込の効果を調べるために，辞書構築の際の閾値を 0〜1 まで変化させたときの評価データにおける分類精度の変化を実験した．閾値が 0 の場合は，全てのかカテゴリが辞書に加えられるため，絞込を行わない場合の分類精度に相当する．また，閾値が 1 の場合は上位語の拡張を行わない場合に相当する．

閾値を 0〜1 に大きくさせるにつれて，辞書に追加される意味カテゴリの数は次第に減少していくため，分類先の候補数は減少していく．分類先の候補数が減少すると，分類問題として容易になるため，分類精度が向上していくことが予想される．一方で辞書の意味カテゴリ数が減少するに伴って，正解の意味カテゴリが辞書に登録されなくなることが考えられる．そうすると，分類先の候補に正解意味カテゴリが含まれなくなるため，分類精度が低下することが想定される．辞書の絞込は上記の 2 つの要因により分類精度が増減すると考えられる．

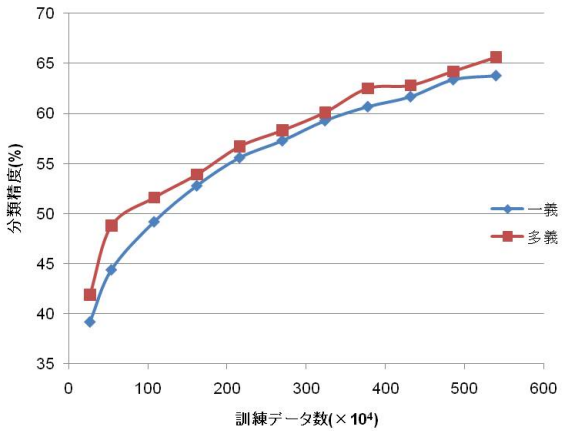


図 1　訓練データ数と分類精度

まず，多義語のデータについて考察する．図??から閾値が 0.0〜0.35 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．この結果から，意味カテゴリ辞書による分類先カテゴリの絞込は有効であるといえる．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．

一方，一義語のデータには辞書構築の際 の閾値が 0 から 1 に向上するにつれて，分類精度は，閾値が 0.05 の場合を除いて分類精度が向上していることが分かる．また，一義語のデータの場合は閾値が 1 の場合が最も精度が良いため，提案手法により上位語の拡張を行う必要がないことが分かる．これは一義語は属する意味カテゴリ数が多義語に比べて少ない

© 2010 Information Processing Society of Japan

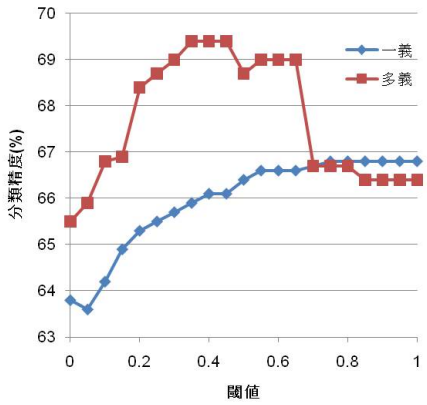


図 2　辞書構築の際の閾値と分類精度

ために，上位語の拡張を行う必要がないことが理由だと考えられる．

以下，辞書を行った場合の分類結果について考察していく．

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，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．

(2) 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，(2)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．

一方で，辞書による候補意味カテゴリの絞込が，辞書の誤りが原因で機能しない例が見られた．例えば，Web テキストから自動構築した意味カテゴリ辞書には，「オセロ」の取り得る意味カテゴリとして「人名」が列挙された．これは，例えば，下記のような文が Web テキスト中に含まれていたため，語彙統語パターンから「オセロ」の上位語が「芸人」として抽出され，「芸人」が意味カテゴリ「人名」の上位語と推測されたためである．

(3) a.　**芸人「オセロ」**が司会のテレビ番組．

また，(4)a. の例文の「オセロ」は，女性お笑いコンビの「オセロ」のため，正解は「公演組織」だが，周辺の文脈から，分類器は誤って意味カテゴリ「人名」に分類した．

(4) a.　松竹芸能は，1990 年代は森脇健児、近年では**オセロ**が大ブレイクした．

意味カテゴリ「人名」は分類先の候補に含まれるため，辞書による絞込を用いても正しく分類することができない．これは Web テキストから自動構築した辞書に誤って「人名」が登録されていることが誤りの原因であり，こうした誤りを防ぐために，今後，辞書構築の精度向上に取り組んでいきたい．

一方，文脈から正解カテゴリが明らかな場合でも分類器が誤る例が見られた．例えば，(5)a. の「オレンジ」は文脈から正解カテゴリが「自然色」であることは明らかだが，誤って「競技組織」に分類された．

(5) a.　大洋監督時代，日本プロ野球初となる**オレンジ**色のホームユニフォームを採用した．

(5)a. の例では，分類対象語「オレンジ」の後方 1-gram の「色」を手掛かりに，正解意味カテゴリが「自然色」と分る．しかし，分類器は (5)a. の bag of words(「監督」,「野球」,「ユニフォーム」など) を手掛かりに，誤って「オレンジ」を「競技組織」に分類した．これは，分類対象語が出現する文の全ての単語を等しい重みで特徴量として用いていることが原因であり，分類対象語の近傍の単語のみを特徴量として用いることで改善が期待できる．Villeneuve らの研究では，近接する特徴量を優先的に用いることで Semeval のタスクにおいて，分類精度が向上することを確認しており[1)]，こうした手法を用いることで精度向上を図ることは今後の課題である．

6. おわりに

本稿では，新語や新語義に対して適応可能な語句の意味が持つ曖昧性の解消の枠組みとして意味カテゴリ分類を導入した．意味カテゴリ分類において問題になる訓練データ作成のコストと，分類先のカテゴリ数が多い問題に対して，Wikipedia から訓練データを半自動生成する手法と，Web テキストから自動構築した辞書を用いて分類先のカテゴリを絞込を行う手法を提案した．そして，評価実験を通じて，上記の手法の有効性を確認した．

評価実験の考察を通じて，分類精度の向上には，提案手法よりも大規模な訓練データが必

© 2010 Information Processing Society of Japan

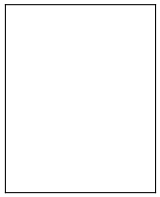
要であることと，高度な分類手法が必要であることを確認した．そのため，今後の課題として，提案手法よりも大規模な訓練データを自動生成するために，Wikipedia よりも大規模な Web テキストから訓練データを自動生成する手法を検討することと，分類精度向上のために分類手法の高度化を検討していきたい．

また，本稿では Web テキスト解析への利用を目的として，意味カテゴリ分類の枠組みを提案した．そのため，Web テキスト解析において，意味カテゴリ分類の枠組みが機能するかについても調査を進めていきたい．

参 考 文 献

1) Brosseau-Villeneuve, B., Nie, J.-Y. and Kando, N.: Towards an optimal weighting of context words based on distance, *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, Stroudsburg, PA, USA, Association for Computational Linguistics, pp.107–115 (online), available from ⟨http://portal.acm.org/citation.cfm?id=1873781.1873794⟩ (2010).
2) Bunescu, R. C. and Pasca, M.: Using Encyclopedic Knowledge for Named entity Disambiguation, *EACL*, The Association for Computer Linguistics, (online), available from ⟨http://acl.ldc.upenn.edu/E/E06/E06-1002.pdf⟩ (2006).
3) Chawla, N. V., Japkowicz, N. and Kotcz, A.: Editorial: special issue on learning from imbalanced data sets, *SIGKDD Explor. Newsl.*, Vol.6, pp.1–6 (online), DOI:http://doi.acm.org/10.1145/1007730.1007733 (2004).
4) Ciaramita, M. and Johnson, M.: Supersense tagging of unknown nouns in WordNet, *Proceedings of the 2003 conference on Empirical methods in natural language processing*, EMNLP '03, Stroudsburg, PA, USA, Association for Computational Linguistics, pp.168–175 (online), DOI:http://dx.doi.org/10.3115/1119355.1119377 (2003).
5) Collins, M.: Discriminative training methods for hidden Markov models: theory and experiments with perceptron algorithms, *Proceedings of the ACL-02 conference on Empirical methods in natural language processing - Volume 10*, EMNLP '02, Stroudsburg, PA, USA, Association for Computational Linguistics, pp.1–8 (online), DOI:http://dx.doi.org/10.3115/1118693.1118694 (2002).
6) Izquierdo, R., Suárez, A. and Rigau, G.: An empirical study on class-based word sense disambiguation, *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '09, Stroudsburg, PA, USA, Association for Computational Linguistics, pp.389–397 (online), available from ⟨http://portal.acm.org/citation.cfm?id=1609067.1609110⟩ (2009).
7) Kageura, K., Fukushima, T., Kando, N., Okumura, M., Sekine, S., Kuriyama, K., Takeuchi, K., Yoshioka, M., Koyama, T. and Isahara, H.: IR/IE/Summarisation Evaluation Projects in Japan.
8) Mihalcea, R.: Using Wikipedia for Automatic Word Sense Disambiguation, *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, Rochester, New York, Association for Computational Linguistics, pp.196–203 (online), available from ⟨http://www.aclweb.org/anthology/N/N07/N07-1025⟩ (2007).
9) Miller, G. A.: WordNet: a lexical database for English, *Commun. ACM*, Vol.38, pp.39–41 (online), DOI:http://doi.acm.org/10.1145/219717.219748 (1995).
10) Navigli, R.: Word sense disambiguation: A survey, *ACM Comput. Surv.*, Vol.41, pp.10:1–10:69 (online), DOI:http://doi.acm.org/10.1145/1459352.1459355 (2009).
11) Okumura, M., Shirai, K., Komiya, K. and Yokono, H.: SemEval-2010 task: Japanese WSD, *Proceedings of the 5th International Workshop on Semantic Evaluation*, SemEval '10, Stroudsburg, PA, USA, Association for Computational Linguistics, pp.69–74 (online), available from ⟨http://portal.acm.org/citation.cfm?id=1859664.1859676⟩ (2010).
12) Sekine, S., Sudo, K. and Nobata, C.: Extended named entity hierarchy, *Proceeding of the Third International Conference on Language Resources and Evaluation*, pp.1818–1824 (2002).
13) Whitelaw, C., Kehlenbeck, A., Petrovic, N. and Ungar, L.: Web-scale named entity recognition, *Proceeding of the 17th ACM conference on Information and knowledge management*, CIKM '08, New York, NY, USA, ACM, pp.123–132 (online), DOI:http://doi.acm.org/10.1145/1458082.1458102 (2008).

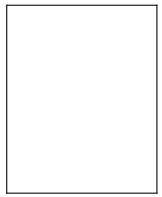
(平成 22 年 7 月 17 日受付)
(平成 22 年 9 月 17 日採録)



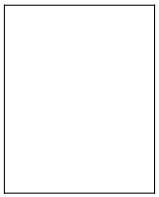
村本　英明（正会員）

昭和 62 年生．平成 21 年東京大学工学部電子情報工学科卒業．平成 23 年東京大学大学院情報理工学系研究科修了．

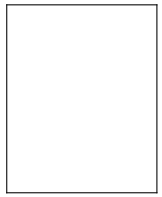
© 2010 Information Processing Society of Japan



ここに名前を入力する（正会員）



ここに名前を入力する（正会員）



ここに名前を入力する（正会員）

© 2010 Information Processing Society of Japan