Wikipedia 記事を利用した曖昧性のある表現の固有表現クラス分類*

藤井裕也　　飯田 龍　　徳永健伸
東京工業大学 大学院情報理工学研究科
{yfujii,ryu-i,take}@cl.cs.titech.ac.jp

1 はじめに

近年 Web の規模はますます拡大し，膨大な量の文書が Web 上に溢れるようになってきている．そのため，膨大な量の文書から必要な情報を取り出す必要性が高まってきており，情報抽出の技術に注目が集まっている．文書から情報を抽出する際は，文書中に出現する固有表現を認識することが重要な要素技術となる．固有表現クラスの認識には，人手で作成した規則に基づく方法 [1] と機械学習による方法の 2 つの方法が存在する．人手で作成された規則に基づく方法では規則が人間にとって理解しやすいという利点があるが，新出の固有表現に対し新たに規則を記述，また既に作成されている規則との整合性をとるなど，保守・管理が困難であるという問題がある．一方，機械学習に基づく方法では固有表現タグが付与されたコーパスがあれば分類にどのような手がかりが役立つかを自動で学習できるため，現在ではこちらの手法が一般的に用いられている．機械学習による手法において，ある文脈上に現れる表現に対してその固有表現クラスを分類する際，その表現自身を構成する語の情報や前後文脈の単語の特徴などを用いて学習する手法が一般的であるが [2]，この手法では対象の表現がどの固有表現クラスになり得るかを事前に制約しないため，誤ったクラスに分類されることがある．例えば，例 (1) では「ライオン」という表現は企業名や生物名など限られた固有表現クラスにしかなり得ないが，単純に文脈情報をもとに構築された多値分類器を利用して固有表現クラスを決定する場合，文脈によってはその表現が取り得ないクラスを割り当てる可能性がある．

(1) ライオンが株の 15%を保有している．

そこで本研究では対象表現が取り得る固有表現クラスを Wikipedia 記事の情報を利用してあらかじめ推定することによって固有表現クラス候補集合を列挙した上で固有表現クラス分類を行う手法を提案する．また，Wikipedia 記事から抽出できる情報を用いて対象表現がどの語義に近いかを推定し，その情報も併用することで分類精度の向上を図る．

本稿では 2 節で関連研究について記述し，3 節で提案手法の解析手順を説明する．次に 4 節で評価実験を行い，最後に 5 節でまとめる．

*Named entity classification using Wikipedia articles
Yuya Fujii, Ryu Iida, and Takenobu Tokunaga
Tokyo Institute of Technology

2 関連研究

Wikipedia は誰でも閲覧，編集ができるインターネット上の百科事典である．Wikipedia の各ページ内には見出し語（ページのタイトル）とその説明が記述されており，必要に応じて関連する他の Wikipedia ページへのリンクを含んでいる．また，それぞれのページは編集者によって作成されたカテゴリという情報で分類されている．Wikipedia では新出用語などのページが日々生成，更新されるため，言語処理の研究でもこのページ集合を対象とした研究が進められている．固有表現抽出に関しては，例えば渡邊ら [3] は Wikipedia ページ内のリンクのグラフ構造に着目し，条件付確率場を用いてページの固有表現クラスの分類を行う手法を提案している．また，杉原ら [4] は Wikipedia が持つカテゴリの階層構造の情報を素性に用いることによってページを固有表現クラスへ分類する手法を提案している．ただし，どちらの手法も Wikipedia のページ，つまりページのタイトル名の固有表現クラスを分類しており，各見出し語がどの固有表現クラスに分類されるかという資源は蓄積されるが，実際の文脈中で固有表現クラスが曖昧な表現については，これらの手法で構築した資源はそのまま適用することができない．また，Wikipedia のページ，つまりページのタイトルを語義とみなすことで語義曖昧性解消の問題を考えることもできる．例えば，Mihalcea[5] は Wikipedia ページ内に出現するパイプリンクに着目し，語義曖昧性解消の問題に取り組んでいる．このパイプリンクとは，Wikipedia ページ編集者がページでの読みと実際のリンク先の両方の情報を編集するための形式で，例えば，例 (2) において，実際にこの文をブラウザで見る場合には「ライオン」という文字列が表示されるが，そのリンク先は Wikipedia ページ集合の中の「ライオン (企業)」というページとなる．

(2) 現在も [[ライオン (企業)| ライオン]] が株の 15%を保有している．

Mihalcea はこのパイプリンクの右側の表現を語義曖昧性の対象となる単語，左側の表現を語義とみなし，このパイプリンクとその近傍文脈を収集することで語義曖昧性解消の問題集合を人手作成のコストをかけることなく大量に収集可能としている．本研究でもこのパイプリンクを固有表現の曖昧性解消の手がかりとして用いる．

3 Wikipedia 記事を用いた固有表現分類の洗練

2 節で概観したように Wikipedia の情報は，固有表現抽出もしくは語義曖昧性解消の重要な手がかりとして利用できると考えられる．そこで，本研究では Wikipedia のパイプリンクとページのカテゴリ情報を利用して固有表現分類の精度向上を目指す．提案手法の具体的な処理の手順は以下の通りであり，以後の各項でそれぞれの詳細を述べる．

1. パイプリンクを利用し，分類対象となる単語が取りうる固有表現クラスを得る
2. ページのカテゴリ情報をもとに対象単語に関連した語義分類器を作成する
3. (1)，(2) の手がかりを既存の固有表現認識の手法で利用する

3.1 Wikipedia 記事の固有表現分類

2 節で述べたように Wikipedia の記事には，その記事内での読みとリンク先のページタイトルを表すパイプリンクが多数存在する．これを抽出することにより，単語とその語義に相当するページタイトルを容易に収集することができる．例えば，単語「ライオン」については「ライオン（企業）」や「ライオン[1]」,「ライオン（巡洋戦艦）」などのタイトル名を収集可能である．ここで収集したタイトル名を語義とみなすと,「ライオン（企業）」や「ライオン（巡洋戦艦）」は,「ライオン」という表現が文章中に出現した場合に取りうる語義の一覧とみなすことができる．ここでは，この語義の一覧を利用して，固有表現分類の際にあらかじめ取り得る固有表現クラスを限定することを考える．今ここで語義に相当するのは Wikipedia の各ページであり，各ページに対してはすでに先行研究によって固有表現のクラスを割り当てる手法が提案されている．つまり，これらを組み合わせることによって，ある表現が取り得る固有表現クラスの一覧を把握することができ，あらかじめ分類される可能性のない固有表現クラスを取り除くことにより，文脈中に出現する表現の固有表現分類の際の解析精度の向上が期待できる．

この手法ではどの固有表現のクラスラベルを使うかに制約はないが，本稿では後述するように関根の拡張固有表現階層（ver.7.1.0）[2]を利用した結果について報告するため，ここでは杉原ら [4] の研究で利用されている情報を参考に分類モデルを作成し，それを用いて各ページの固有表現クラスを決定する．具体的には，記事中の説明文に出現する形態素とページのカテゴリ情報を利用して学習を行い，one vs rest 法で分類対象となるページの固有表現クラスを一意に決定する．ここでカテゴリ情報として，杉原らの手法に従って Wikipedia のカテゴリ階層構造の最上位のカテゴリである「主要カテゴリ」ページから対象ページまでの最短パス上にあるカテゴリ名を素性として用いる．

[1]生物名であるライオンについてはページタイトルが「ライオン」の項目に記載されている

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

3.2 語義分類モデル

次に各ページに記載されているカテゴリの情報を分類に利用する方法について考える．Wikipedia の各ページには編集者が記述したカテゴリ集合があり，例えば,「ライオン（企業）」のページには「日本の化学工業メーカー」や「日本の医薬品メーカー」といったカテゴリが記述されている．一般的には類似するページには同一のカテゴリ名が付与されている可能性が高いため，同一カテゴリ名を持つページを参照することによって分類対象となるページからは得ることができない情報を分類に利用できる可能性がある．ここではその一例として，与えられた 2 つの異なる語義に相当するページのカテゴリ集合のうち対象表現が出現する文脈の近傍語の集合がどちらのカテゴリ集合に近いかを分類する分類器を作成し，その情報を固有表現クラス分類に利用する．以後，この分類器のことを語義分類器と呼ぶ．

ここでは文脈中に出現する「ライオン」という表現を例にどのような語義分類器を作成するかを説明する．前述のように単語「ライオン」は「ライオン（企業）」や生物名としての「ライオン」など複数の語義をとり，それぞれが Wikipedia のページに対応付けられている．この状況で「ライオン (企業)」のページが「日本の化学工業メーカー」と「日本の医薬品メーカー」の 2 つのカテゴリから成るカテゴリ集合，生物名「ライオン」のページが「ネコ科」と「特定動物」の 2 つカテゴリから成るカテゴリ集合を持つとすると,「ライオン」という表現が出現した文脈の近傍語の集合がどちらのカテゴリ集合に近いかという情報が最終的な固有表現クラスの分類に寄与すると考えられる．そこで，分類対象の表現に対する各語義候補のページに出現するカテゴリを抽出し，各語義候補のページの任意の組み合わせについて語義分類器を作成する．例えば,「ライオン」という表現が「ライオン (企業)」,「ライオン」,「ライオン (巡洋戦艦)」の 3 つの語義を持つ場合には,「ライオン (企業)」と「ライオン」,「ライオン (企業)」と「ライオン (巡洋戦艦)」,「ライオン」と「ライオン (巡洋戦艦)」のそれぞれの組み合わせについてそれらが持つカテゴリ集合間の語義分類器を作成し，それぞれの出力値を固有表現分類の素性として利用する．

各語義分類器を作成する際は，訓練事例として同一カテゴリに属する他のページを収集し，各ページを 1 事例として扱う．ただし，あるカテゴリに属するページが大量に存在する場合には，語義となるページに出現していたカテゴリを多く共有するページを優先して利用し，各訓練事例の上限を 1000 事例とした．学習に利用する素性にはページの説明文に出現している形態素を用いる．

3.3 固有表現分類モデル

最後に 3.2 の語義分類モデルが出力する値を素性として用いる固有表現分類モデルを作成する．このモデルは機械学習を用いた固有表現分類に関する既存の研究 [2] を参考に作成する．学習には以下の素性を利用した．

- 分類対象となる表現 w の品詞
- w の係り先の文節内の品詞と形態素の見出し語
- w の係り元の文節内の品詞と形態素の見出し語

- w の前後 2 文節内に出現する機能語の文字列
- w の文脈中に出現する形態素の見出し語

ただし，文脈中に出現する形態素に関しては w に近いほど重要であると考えられるため，w と該当文節の距離を d とした場合，その文節内に出現する形態素の素性値を $10/(5+d)$ とする．また w と異なる文内に出現する形態素に対しては素性値を 0.4 とする．上記の素性と 3.2 の語義分類モデルが出力する値を素性として学習した分類モデルを用い，one vs rest 法で評価の際に固有表現クラスを決定する．ただし分類対象となる表現が取り得る固有表現クラスは 3.1 であらかじめ決定したクラス候補内のいずれかとする．

表 1: 関根の拡張固有表現階層 ver.7.1.0 (第 2 層目まで)

名前	人名
	神名
	組織名
	地名
	施設名
	製品名
	イベント名
	自然物名
	病気名
	色名
	名前_その他
時間表現	略
数値表現	略

表 2: 評価用の単語の例とその語が取り得る固有表現クラス

単語	クラス	パイプリンクにより得られた語義候補
ハブ	施設名	ハブ空港 [16]
	製品名	ハブ (ネットワーク機器)[28]
	自然物名	ホンハブ [6], ハブ (動物)[51]
核	製品名	核兵器 [67], 核 (数学)[12], 原子力 [3], 核爆弾 [3], 原子力推進 [1], 零空間 [1]
	自然物名	細胞核 [83], 灰白質 [5], 原子核 [3]
ニュートン	人名	アイザック・ニュートン [97]
	製品名	ニュートン (雑誌)[11], アップル・ニュートン [4], ニュートン [1]
	地名	ニュートン (マサチューセッツ州)[15]
レインボー	施設名	レインボーロード [2]
	製品名	日野・レインボー [21], レインボー (テレビ番組)[9]
	組織名	レインボー (バンド)[76], 高田薬局 [1]
	自然物名	虹 [1]

括弧内の数値はその語義に対する文脈数を表す.

表 3: テスト単語集合に対する正解クラスの割合

クラス	文脈数	割合
人名	397	5.5%
施設名	226	3.1%
製品名	3885	53.9%
地名	573	8.0%
組織名	1108	15.4%
自然物名	778	10.8%
イベント名	3	0.0%
病気名	5	0.1%
色名	11	0.2%
神名	13	0.2%
時間表現	204	2.8%
合計	7203	100%

4 評価実験

4.1 実験設定

本研究では，関根の拡張固有表現階層 (ver.7.1.0) を固有表現クラスの定義として用いる．この拡張固有表現階層は 200 クラスの固有表現クラスが階層構造として定義されており，使用する階層の深度によって様々な粒度の解析を行える．本研究では，表 1 に示したクラスの内，名前クラスの第 2 階層目の「名前_その他」を除いた 10 クラスに「時間表現」クラスの合計 11 クラスの分類問題を考える.「名前_その他」と「数値表現」クラスはそれぞれ Wikipedia 記事から十分な量のデータが得られなかったため除外した．

また，評価に使用する単語集合は以下の基準を満たすものを選択した．

- 2 種類以上の固有表現クラスに分類される
- パイプリンクにより得られる文脈数が 100 以上である
- 形態素 1 つで構成されており，品詞が名詞-一般である[3]

これらの条件を満たす 29 単語を評価用単語として選択した．評価用単語の例とそのクラス語が取り得る固有表現クラスを表 2 に示す．これら評価用単語それぞれの文脈情報はパイプリンクをもとに抽出する．あらかじめ評価用単語 29 語がパイプリンクで記述されたページを抽出し，それぞれに正解クラスを付与することで評価用データを作成した．テスト単語集合に対する正解クラスは表 3 のような構成となった．

文脈から素性を抽出するために必要な形態素・係り受け解析についてはそれぞれ茶筌[4]と南瓜[5]を用い，分類器の学習には Support Vector Machine (SVM)[6]を使用した．なお，SVM のカーネルには線形カーネルを使用し，パラメタ c の値は 1.0 とした．

4.2 実験結果

まず 3.1 で説明した Wikipedia ページの固有表現分類モデルを 5 分割交差検定によって評価した．学習と評価には，NAIST Japanese ENE Dictionary on Wikipedia[7]，及び現代日本語書き言葉均衡コーパス[8]に関根の拡張固有表現階層の固有表現クラスがタグ付けされたデータ [6] を用いた．NAIST Japanese ENE Dictionary on Wikipedia は Wikipedia の見出し語に対して関根の拡張固有表現階層の定義に基づく固有表現クラスが付与されたデータであり，そのまま学習・評価用データに用いることができる．一方，現代日本語書き言葉均衡コーパスは白書や書籍の文章に出現する文字列に固有表現クラスのタグを付与したものであるため，コーパス中で固有表現タグが付いている表現のうち，Wikipedia ページの見出し語と一致するものを抽出することで学習・評価用データを作成した．なお，素性に用いる形態素は茶筌によって動詞，名詞 (「名詞-数」を除く)，形容詞，未知語と判定されたもののみを利用した．各クラスに対する再現率と精度を表 (4) に示す．精度，再現率ともに高い値を得ており，これらの分類器を用いて対象表現の固有表現クラス候補集合を十分に推定できると考えられる．

[3]品詞は naist-jdic (http://sourceforge.jp/projects/naist-jdic/) に従った

[4]http://chasen-legacy.sourceforge.jp/
[5]http://chasen.org/ taku/software/cabocha/
[6]http://chasen.org/ taku/software/TinySVM/
[7]http://birch.naist.jp/ masayu-a/p/NAIST-jene.html
[8]http://www.tokuteicorpus.jp/g_ghq/item-103.html

表 4: Wikipedia 記事に対する固有表現クラス分類の結果

クラス	精度	再現率
人名	97.03%	95.71%
施設名	96.47%	90.87%
製品名	82.99%	92.32%
地名	91.58%	92.97%
組織名	80.00%	75.32%
自然物名	93.73%	90.08%
イベント名	78.80%	56.89%
病気名	85.55%	73.60%
色名	84.17%	80.00%
神名	76.24%	52.86%
時間表現	99.14%	92.62%

表 5: 学習に用いた各クラスに対する文脈数

クラス	文脈数	割合
人名	2533	11.4%
施設名	957	4.3%
製品名	4670	21.1%
地名	3478	15.7%
組織名	3108	14.0%
自然物名	2139	9.6%
イベント名	608	2.7%
病気名	563	2.5%
色名	1041	4.7%
神名	644	2.9%
時間表現	2439	11.0%
合計	22180	100%

次に 29 単語の固有表現分類問題の評価実験を行った．3.3 の固有表現分類モデルの学習データには，3.1 の分類器を学習する際に用いた表現とその固有表現クラスの対に対して，Wikipedia 記事中でその表現へ直接リンクされた箇所の文脈を収集することで，文脈情報を収集した．この結果得られた学習データの数と各クラスの割合を表 5 に示す．

対象表現のクラス候補を制限せず、語義分類モデルの出力結果を素性として用いない場合をベースラインとし，クラス候補を制限した場合と語義分類モデルによる素性を用いた場合それぞれと比較した結果を表 6 に示す．結果を比較すると，クラスを制約した方がしない場合よりも全体の正解率が約 5%上昇している．また，語義分類モデルによる素性を加えた場合と加えない場合では加えた場合の方が正解率が 0.5%程度上昇した．最も正解率の高い (d) の各クラスに対する再現率と精度を表 7 に示す．この結果から全体的に製品名に分類される割合が高く，他のクラスの分類精度が良くないことが分かる．これは製品名クラスの定義の範囲が広いため，他のクラスとの境界が明確でないことが原因として考えられる．また，地名，人名クラスは一般的な固有表現認識の問題と比較して表層が一般名詞である表現のみを対象に問題を解いているため特に分類精度が低くなっている．そこで固有表現分類モデルの素性として使用していた「分類対象となる表現 w の品詞」の素性を除いて再度実験を行った結果，(a)〜(d) のいずれにおいても製品名クラスの再現率は低下するものの人名，地名クラスの再現率は大幅に改善された．具体的には，最も精度の良かった (d) の結果において，製品名クラスの再現率が 83.86%から 72.54%へ減少する代わりに，人名，地名クラスの再現率はそれぞれ 6.05%から 52.64%，5.93%から 26.88%へと増加し，全体の正解率も 60.77%から 61.32%に増加した．このことから対象表現の表層が一般名詞である場合は表層の品詞を素性に用いない方が精度が良いということが分かる．一般的な固有表現分類の問題に本手法を適用する際には，対象表現の表層が一般名詞である場合とそうでない場合で分類器を使い分ける必要があると考えられる．

表 6: 各方法による精度の比較

	マイクロ平均	マクロ平均
a) ベースライン	55.50%	51.83%
b) a)+クラス制約	60.25%	55.86%
c) a)+語義分類モデル	56.07%	52.65%
d) a)+クラス制約+語義分類モデル	60.77%	56.34%

マイクロ平均は全体の正解率を，マクロ平均は単語毎の正解率の平均を表す．

表 7: 方法 (d) の各クラスに対する結果

クラス	精度		再現率	
人名	42.86%	(24/56)	6.05%	(24/397)
施設名	54.59%	(113/207)	50.00%	(113/226)
製品名	62.58%	(3258/5206)	83.86%	(3258/3885)
地名	59.65%	(34/57)	5.93%	(34/573)
組織名	57.93%	(369/637)	33.30%	(369/1108)
自然物名	56.62%	(560/989)	71.98%	(560/778)
イベント名	—	(0/0)	0.00%	(0/3)
病気名	66.67%	(2/3)	40.00%	(2/5)
色名	58.82%	(10/17)	90.91%	(10/11)
神名	20.00%	(6/30)	46.15%	(6/13)
時間表現	100.0%	(1/1)	0.49%	(1/204)

5 おわりに

本研究では文脈中で固有表現クラスに曖昧性がある表現に対し，あらかじめ対象表現が取り得るクラス候補を制限することで固有表現分類の精度を向上させることができることを示した．また，Wikipedia のカテゴリから対象表現の各語義に相当する類似ページを集め，そこから作成した分類器の出力値を素性に用いることでも若干精度が向上することを確認した．

本研究では対象表現を固有表現クラスの粒度で分類する問題を扱ったが，単語の語義のような別の粒度で分類する問題に適用することも考えられる．分類する粒度としては Wikipedia のカテゴリが挙げられるが，Wikipedia のカテゴリは上位下位関係の粒度が統一されていないという問題がある．そこで，Wikipedia と既存のオントロジを統合する研究 [7] の成果を併用し，分類の粒度そのものを定義することで評価を行いたい．

参考文献

[1] 竹元義美, 福島俊一, 山田洋志. 辞書およびパターンマッチルールの増強と品質強化に基づく日本語固有表現抽出. 情報処理学会論文誌 42(6), pp. 1580–1591, 2001.

[2] 山田寛康, 工藤拓, 松本裕治. Support Vector Machine を用いた日本語固有表現抽出. 情報処理学会論文誌 (NL-43-1), pp. 44–53, 2002.

[3] 渡邉陽太郎, 浅原正幸, 松本裕治. グラフ構造を持つ条件付確率場による Wikipedia 文書中の固有表現分類. 人工知能学会論文誌 23(4), pp. 245–254, 2008.

[4] 杉原大悟, 増市博, 梅本宏. Wikipedia カテゴリ階層構造の固有名詞分類実験における効果. 情報処理学会研究報告. 情報学基礎研究会報告 (NL-189-9), pp. 57–64, 2009.

[5] Rada Mihalcea. Using Wikipedia for Automatic Word Sense Disambiguation. In *Proceedings of the North American Chapter of the Association for Computational Linguistics*, pp. 196–203, 2007.

[6] 橋本泰一, 乾孝司, 村上浩司. 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 (NL-188-17), pp. 113–120, 2008.

[7] 小林暁雄, 増山繁, 関根聡. 日本語語彙大系と日本語ウィキペディアにおける知識の自動結合による汎用オントロジー構築手法. 情報処理学会研究報告. 自然言語処理研究会報告 (NL-187-2), pp. 7–14, 2008.