Vol.2010-NL-197 No.5
2010/7/22

特徴的部分文字列と共起情報を用いた固有表現の詳細ラベル付与

小林　のぞみ[†1]　松尾　義博[†1]　菊井　玄一郎[†1]

本稿では，固有表現を従来の IREX 定義よりも詳細に分類するタスクについて検討する．与えられた固有表現をあらかじめ設定したクラスに分類する問題に焦点をおき，文章中で共起する語および語彙大系のカテゴリ情報と，あるクラスに特徴的な，固有表現を構成する部分文字列を素性とする手法を提案する．この手法を blog および Web 新聞中の固有表現を対象として評価し，F 値が 0.67 から 0.72 に向上することを確認した．また，クラス毎に素性の有効性を調査し，出現頻度の低い語を多く含むクラスで部分文字列素性が有効であり，固有表現の曖昧性が多いクラスで共起情報が有効であることを確認した．

Named Entity Classification Based on Substring Patterns and Co-occurrence Information

Nozomi Kobayashi,[†1] Yoshihiro Matsuo[†1]
and Gen'ichiro Kikui[†1]

This paper presents a method for classifying named entities into finer classes than those of the IREX definition, which is a standard for Japanese named entity recognition. This method uses substring patterns and co-occurrence information; words in the text and their categories derived from a thesaurus. The experimental results show that the proposed method improves F-measure to 0.72 from 0.67. Furthermore, our results show that the substring patterns are effective for the classes which have many low-frequent named entities, while the co-occurrence information is effective when the class includes many ambiguous named entities.

†1 日本電信電話株式会社 NTT サイバースペース研究所
NTT Cyber Space Laboratories, Nippon Telegraph and Telephone Corporation

1. はじめに

テキストから情報を抽出する上で，人名，組織名，場所名などの固有表現を認識，抽出することは重要な課題である．固有表現抽出は，これまで MUC[⋆1]，ACE[⋆2]などの評価型ワークショップを通して活発に研究がなされてきた．日本でも 1999 年に IREX[⋆3] が開催され，このときに定義された固有表現体系が日本語の固有表現抽出のスタンダードとして使用されている．

IREX では，組織名，人名，地名，固有物名，日付表現，時間表現，金額表現，割合表現の 8 種類の固有表現が定義されている．しかしながら，ブログに代表されるウェブテキストを対象にした情報抽出，および抽出された情報を使った評判検索などのサービスを考えると 8 種類では十分ではなく，より細かな分類体系が必要である．

IREX 定義に基づく固有表現手法のほとんどは，(1) テキストから固有表現を切り出し（チャンキング），(2) クラスに分類する という二つのタスクをシーケンシャルラベリングの問題として同時に解いている．学習に用いる素性は，現在着目している形態素（もしくは文字）の前後の形態素や品詞など，比較的狭い範囲の情報であることが多い．近年報告されているより詳細な固有表現体系に基づく手法も，これらと同様に周辺の情報を使用した手法となっている[3),10)]．しかしながら，より詳細な固有表現クラスを認識するタスクでは固有表現の曖昧性が増え，周辺文脈のみでは判定できない場合が多く存在する．例えば以下の 2 文に出現する**ソフトバンク**は，IREX 定義であればどちらも「組織名」であるが関根の拡張固有表現階層[⋆4]では前者は企業名，後者は競技組織名となる．

(1) **ソフトバンク**の孫さんの講演に行ってきました。
(2) **ソフトバンク**の内川さん、最近よく打ちますね。

この問題を解くためには，形態素の周辺情報より広い情報をみる必要があると考えられる．例えばこの例では，後者が出現した文章が野球について書かれていることがわかれば企業名よりもプロ競技組織名と判定できる可能性がある．

そこで我々は，広く文書全体の情報も使用して固有表現に詳細なクラス語を付与する手法について検討する．固有表現のチャンキングについては，周辺の情報を用いることで IREX

⋆1 http://www-nlpir.nist.gov/related_projects/muc/index.html
⋆2 http://www.itl.nist.gov/iad/mig/tests/ace/
⋆3 http://nlp.cs.nyu.edu/irex/index-j.html
⋆4 http://sites.google.com/site/extendednamedentityhierarchy/

定義の固有名的表現の人名，組織名、地名については 8 割から 9 割の高い精度で抽出できることが報告されているため（例えば文献[8),9)]）従来の手法で解くことを考え，本稿ではクラス分類の問題に絞って話を進める．

以下，2 節で手法全体の枠組みと導入する素性について述べ，3 節で導入した素性の各クラスでの有効性について調査した結果を報告する．4 節で関連研究について述べ，最後に 5 節でまとめる．

2. 固有表現の詳細ラベル付与

固有表現の詳細ラベル付与問題を,「固有表現が与えられたときに，あらかじめ与えられたクラスに分類する」タスクとして設定し，分類問題として解く．

ある固有表現が属するクラスを判断する際には，固有表現抽出手法でこれまで使用されてきた周辺文脈の情報以外に以下の情報が手がかりとなると考えられる．

特徴的な文字列 固有表現を構成する部分文字列は所属するクラスを判定する上で有用な手がかりである．例えば「ABC 航空」という固有表現があったとした場合，名称のみからおそらく会社であろうと推測することができる．

文章全体から得られる情報 ある文章が何について語られているかは，表記のみで決定できない場合に有用な手がかりである．例えば 1 節で挙げたソフトバンクの例において，(2) が野球について語られている記事であるとわかればこのソフトバンクは競技組織名であると推測できる．

以上の観察に基づき，それぞれの情報を素性として導入し，有用な素性の組み合わせを機械学習に基づいて求めることでこの問題にアプローチする．以下ではそれぞの情報の獲得方法および素性へのエンコード方法について述べる．

2.1 特徴的部分文字列

固有表現そのものはクラスを決める上で重要な手がかりである．しかしながら文章に出現する固有表現は種類が多く，詳細なクラス体系を考えた際に訓練コーパスから固有表現そのものを学習することは難しい．また，部分文字列をそのまま素性としても，訓練コーパスにその部分文字列が複数回出現しない場合は手がかりとして獲得できない可能性が高い．例えば「航空」という部分文字列は会社名であることの手がかりと考えられるが，実際に後に述べるコーパスでは「航空」を含む会社名の出現数は 10 回に満たず，手がかりとして獲得できていなかった．

そこで，あらかじめ大量のデータからあるクラスに特徴的な部分文字列を獲得しておき，入力された固有表現のクラスを推定する素性として使用することを考える．以下ではまず表現とクラスを表す語（以下，クラス語）の対を獲得する方法について述べ，次に表現とクラス語の対から特徴的な部分文字列を獲得する方法を述べる．ここで，クラスは複数のクラス語を取ると考え（例えば，クラス「企業名」について「会社」「事業者」など），区別のために「クラス語」という語を使用する．

2.1.1 表現とクラス語対の獲得

Web 上のオープンな百科事典である Wikipedia[⋆1]を使用して表現とクラス語の対を獲得する．Wikipedia はユーザが自由に項目を作成，編集できるため，通常の辞書や事典にはないような固有表現に関する記述が多数存在しており，最近話題になった固有表現に対する項目についても項目がすぐに作成される．これらのことを踏まえて我々の目的にかなう事典であるといえる．

Wikipedia の各項目にはユーザによって多くのカテゴリタグが付与されているが，その中にはクラスを表す語以外にその項目と関連する語も含まれている．例えば「日本電信電話株式会社」の場合,「特殊会社」「日本の電気通信事業者」「千代田区の企業」などに加え,「通信に関する制度」「日経平均株価」のように項目のクラスを表さない語もカテゴリとして付与されている．

そこでカテゴリを直接使うのではなく，風間ら[5)]と同様に Wikipedia の各記事の一文目からクラス語を抽出する手法をとる．Wikipedia の一文目は,「X は Y」のように定型文で書かれていることが多く，見出し語の定義文ととらえることができる．ここで，X が見出し語だとすると，Y はクラス語を表すと仮定できる．

実際に 2009 年 1 月 24 日の Wikipedia ダンプデータに対し，風間らの手法を参考に Wikipedia の各記事の一文目から,「一種」「こと」「もの」などを除いた最後の名詞もしくは未知語をクラス語として抽出した．その結果，Wikipedia 特有のテンプレートなどを除いた見出し語約 51 万の辞書項目から約 45 万項目でなんらかのクラス語を獲得した．以下に抽出された表現とクラス語の対をいくつか示す．

伊東スタジアム, 野球場
日本電信電話株式会社, 会社
NTT ファイナンス株式会社, 会社

⋆1 http://ja.Wikipedia.org/

2.1.2 特徴的な文字列の獲得

固有表現に特徴的な文字列は連続文字列として見られることが多い．例えば上の例では，「株式会社」「会社」の文字列が「会社」の特徴的文字列と考えられる．

この考えに基づき，獲得した表現とクラス語対から任意の n-gram を文字列として獲得する．これによりクラス語 C とある部分文字列 A の共起回数が求められ，A と C が共起した回数 a，A に対して C が出現しなかった回数 b，C に対して A が出現しなかった回数 c を求めることができ，PMI や dice などのさまざまな尺度を利用して共起の強さを求めることができる．今回の実験では，予備実験で精度がよかった重み付き dice 係数を共起の強さとして使用した．重み付き dice 係数は，上の a，b，c から以下の式で求められる．

$$\log_2 a \times \frac{2 \times a}{(a+b)+(a+c)}$$

実際に獲得できた文字列を確認すると,「県立」を特徴的な文字列として持つ dice 係数の高いクラス語として「高等学校」「公園」「博物館」「美術館」などが得られていた．

2.2 文章全体から得られる共起情報

固有表現クラスの粒度が細かくなることで，同じ表現が複数のクラスで現れるようになり曖昧性が増加する．実際に 3.1 で述べるコーパスに出現した固有表現のうち，多いクラスで約 3 割に曖昧性があり，IREX 定義の 2 倍であった．この問題に対処するには曖昧性を解消するための何らかの情報を導入する必要がある．

例として「イーグルス」という固有表現を考える.「イーグルス」は少なくともロックバンド（公演組織）とスポーツ組織の曖昧性がある．従来の方法では,「イーグルス」の前後数形態素に着目するが，下記のような場合に周辺文脈のみでクラスを判断することは難しい．

(3) 内訳は「イーグルス」が 32%でトップ。

しかし，この文の前に以下の文があった場合,「スポーツ」「チーム」などの語からスポーツ組織であると判断できる．

(4) 河北新報社が仙台のスポーツチームについて意識調査を行った結果が出ていました。

そこで文章内で共起している重要な語を抽出し，素性として使用することを考える．重要語の抽出には情報検索や自動要約などで一般的に用いられている tf-idf を使用する．tf-idf は以下の式で求める．

$$\text{tf-idf}(w) = tf(w) \times log\frac{N}{df(w)}$$

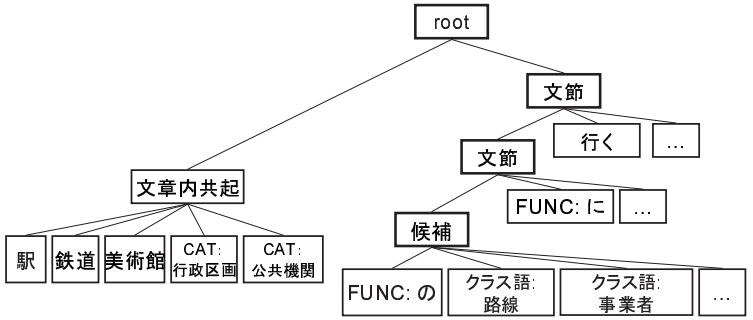

図 1 素性の例

ここで w は単語，$tf(w)$ はある文章での出現頻度，N は総文章数，$df(w)$ は単語 w を含む文章の数である．

また，単語のみではデータがスパースになることが予想されるため，日本語語彙大系[4] のカテゴリ番号も使用する．語彙大系のカテゴリ番号についても同様に tf-idf を計算し，tf-idf 値の高い上位の語およびカテゴリ番号を文書の共起情報として使用する．

2.3 素性へのエンコーディング

前節で述べた二つの情報に加え，文の各形態素の情報，文節の主辞の品詞，機能語などの統語的な情報も使用してモデルを作成する．これは，固有表現 X があるクラス C である場合に出現しやすい文構造があるという仮定に基づいている．

以下の文章を例にどういう木を構成するかを説明する．

(5) 南阿蘇鉄道の白水高原駅（略称）に行ってきました。松江の庭園美術館前駅（略称）を抜いて、日本で一番長い駅名になったそうです。

図 1 は，一文目に出現している固有表現「南阿蘇鉄道」に関する素性の例である．特徴的部分文字列を，2.1 で述べた方法で獲得した文字列を辞書引きして素性として使用する．具体的には，処理対象としている固有表現の部分文字列を n-gram 文字列として獲得し，これらを 2.1 で獲得した特徴的部分文字列で辞書引きする．結果，得られた重み付き dice の値が一定以上（実験では 0.1 を使用）だったクラス語を素性として抽出する．図 1 の「南阿蘇鉄道」の場合，部分文字列「南阿蘇」「鉄道」「阿蘇鉄道」などを獲得した特徴的部分文字列と照合し，値が 0.1 以上だった「路線」「事業者」をクラス語として使用する．

文章内の共起情報として，前節で述べた tf-idf の高い語およびカテゴリ番号を 5 つを使用

表 1 コーパス中のタグのうちわけ

地名		組織名		建物名	
クラス	数	クラス	数	クラス	数
GPE	6154	国際組織名	148	GOE	3669
その他場所	1030	公演組織名	787	路線名	163
地形名	593	競技組織名	2555	その他施設名	287
アドレス	714	法人名	2877		
		政治的組織名	1440		
		その他組織名	510		

する．idf はクロールした blog および Web 新聞約 335 万記事を形態素解析した結果から求めた．図 1 の例では，上位 5 位の「駅」,「鉄道」,「美術館」，語彙大系カテゴリの「行政区画」，および「公共機関」を素性として使用している．

文章内の共起情報で部分木を作成し，処理対象としている文の木をあわせて一つの木を構築し，学習器への入力とした．これらの素性に加え以下の素性を使用する．

- 固有表現を構成する形態素の品詞，日本語語彙大系でのカテゴリ番号
- 固有表現の文字種（平仮名，カタカナ，アルファベット，数字，その他）
- 固有表現直後の接尾辞
- 文節中の各形態素の品詞などの語彙情報
- 文節の主辞の語彙大系でのカテゴリ番号および品詞
- 文節の機能語

モデルの学習には構造情報を考慮した学習アルゴリズムである Bact[6] を用いる．Bact は入力された木構造のデータから部分木を素性とする decision stumps を抽出し，それを弱学習器として用いるブースティングアルゴリズムである．Bact は二値分類器のため，タスクにあわせて多クラスに拡張する必要がある．クラス数が多くなることを考慮し，分類器のスコアが最も高かったクラスを選択する one-versus-rest 法を採用する．

3. 実　　験

前節で導入した特徴的部分文字列と共起情報の二つの素性の効果を明らかにするため実験を行った．

3.1 コーパス

IREX で定義される組織名，地名を対象に，関根の拡張固有表現階層の組織名，地名，施設名以下のカテゴリを参考にクラスを付与した．関根の拡張固有表現階層のクラス数は組織名，地名，施設名のみでも 70 近く存在するが，今回は階層の深さ 2 に限定し，表 1 に示す 13 クラスとした．深さ 2 には 13 クラス以外に民族名，家系名などがあるが，IREX 定義で固有表現ではなかったものは除外した．また，場所名に関してはアプリケーションを想定し，住所を表す表現を含みかつ GPE でないものは全てアドレスとし，GPE，アドレス，地形名にあてはまらないものをその他場所とした．

このコーパスは，Web 上の新聞 1000 記事，blog 記事 2000 記事の計 3000 記事に対して作業者 1 名がタグを付与したものである．表 1 にタグを付与したクラス数とその数について示す．

表 2 分類結果

ベースライン	0.67	(14084/20927)	0.67	(14084/20927)
文字列あり	0.70	(14750/20927)	0.70	(14750/20927)
文字列+共起	**0.72**	(15068/20927)	**0.72**	(15068/20927)

3.2 評価・考察

3000 記事で 5 分割交差検定を行い，以下の式で求める精度，再現率で評価した．

$$\text{精度} = \frac{\text{システムが正しくクラス付与できた数}}{\text{システムが出力した数}},$$

$$\text{再現率} = \frac{\text{システムが正しくクラス付与できた数}}{\text{人手でクラスが付与された数}}$$

以下の 3 つのモデルを比較し，部分文字列と共起情報の効果について検証した．

- ベースライン: 固有表現を構成する形態素および品詞，文字種，固有表現の前後 2 形態素，品詞，文字種，および前 2 形態素に付与された固有表現のクラス情報
- 文字列あり: 2 節で述べた手法で，共起情報を使用しない
- 文字列+共起: 特徴的部分文字列および共起情報のいずれも使用

ベースラインは橋本ら[3] が使用している固有表現の周辺情報と固有表現抽出でよく使用される文字種（カタカナ，ひらがな，アルファベット，その他）を使用するモデルである．

結果を表 2 に示す．表 2 から，部分文字列を使用することで 3%精度が向上し，さらに共起情報を入れることで 2%精度が向上していることがわかる．また，マクネマー検定を行ったところ有意水準 1%でベースラインと文字列あり，文字列ありと文字列+共起，ベースラインと文字列+共起のいずれについても有意であることが確認できた．このことからいずれの素性も全体の精度向上に効果があったといえる．

次に，固有表現のクラスによって各素性の影響範囲が異なると考え，クラスごとの精度再

現率を調査した．その結果を表 3 に示す．また，有意水準 1%でマクネマー検定を行い，有意であった結果については F 値に'*' を付与している．

表 3 から，特徴的部分文字列は 13 クラス中 8 クラスで有意性が確認できたが，共起情報については競技組織名と GOE でのみ有意差が見られた．また，その他に属する固有表現については全体的に精度が低い傾向にあった．その他は他のクラスと比較してさまざまな固有表現が入っていることと，コーパス中の出現数がそれほど多くないことから十分に学習できていない可能性があり，今後学習曲線を見るなど詳しく分析する予定である．

その他場所については 1000 以上の事例がコーパス中に存在しているが，アドレスや GPE であると誤って判断した例が多かった．例えば「先斗町」は，地名としては存在しないため今回の定義ではその他場所になる．しかしながら「先斗町」が「町」というアドレスによく出現する特徴的文字列を含んでいたことから誤ってアドレスと判断されていた．このような問題についてはあらかじめ先斗町が地域であることを知らなければ人も判定できない例であると考えられるため，今後知識を構築することも検討する必要がある．

特徴的文字列素性の効果が見られたクラスの多くは，コーパス中に 1 回しか出現しなかった固有表現が多い傾向にあった．一例を挙げると，F 値が大幅に向上した路線名はコーパス中に出現した表現の 86%が頻度 1 であり，特徴的文字列に基づいたクラス語を素性とすることでうまく特徴をとらえられたと考えられる．実際に特徴的文字列から得られたクラス語が「路線」「高速道路」であれば路線名らしいという規則が獲得できていた．

一方，効果の見られなかった GPE，競技組織名は，頻出する固有表現の数がある程度限られていること，訓練コーパス中に出現した数が他のクラスより多かったことから表記の情報である程度カバーできていたと考えている．実際にベースラインモデルで学習された規則を確認すると，GPE では「東京都」などの表記,「市」「県」などの形態素が，競技組織名では「中日」「日本ハム」などの表記がそれぞれスコアの高い特徴として得られていた．

次に共起情報の効果について調査した．有意差が見られた競技組織名は，出現した固有表現に曖昧性が最も高かったクラスであり，共起情報を入れることによってベースラインでは GPE や企業名だと誤っていた事例を正しく競技組織名と判断できていた．結果，F 値が 0.69 から 0.71 に改善した．また，GOE について共起情報を導入することで解けた例を見ると，以下の「関内」のように駅や空港名で地名と曖昧性のある場合に語彙大系のカテゴリ情報（「乗り物」など）を使用して正しく GOE と判断できていた．

(6) 15 時過ぎの新幹線に乗り、まずは横浜へ。その後**関内**で降りて…

しかし，2 つのクラスでは有意であったものの他の 11 クラスでは有意差なしという結果となった．競技組織名の他に曖昧性が高かったクラスには，住所，GPE があったが，これらのクラスは，文書が何について書かれているかにあまり依存しないこと，表記の情報が強いということから効果が得られなかったと考えている．多かった誤りは，GPE が答えのときに住所やその他場所名，またその逆のケースであり，誤りの 5 割程度を占めていた．先ほど挙げた「先斗町」のほか，自治体がないため GPE ではない地名などを誤って判定していた．地名については外部知識がなければ解けない問題も多いため，今後の検討課題である．

4. 関 連 研 究

近年の固有表現抽出手法は，タグ付きコーパスから教師あり学習でモデルを学習するものが主流となっている．特に Support Vector Machines に基づく手法[11]，条件付確率場に基づく手法[7] などがよい精度を得ている．IREX に基づく日本語の固有表現抽出手法でも教師あり学習に基づき，対象とする語とその前後の形態素（文字）情報，字種などを素性とした手法が多数報告されている（文献[1),8),9)] など）．これまでに報告されている関根の拡張固有表現体系での固有表現抽出に取り組んだ研究も従来の IREX に基づく固有表現抽出と同様に，対象とする語とその前後の形態素情報，字種などに基づく手法である[3),10)]．本稿では，曖昧性の多いクラスについて，文章全体の情報が手がかりになると考え，文章分類問題などで使われている文章中で共起する語を素性として使用する手法を提案した．この手法を blog および Web ニュースで評価したところ，曖昧性の多かった競技組織名で有効に働き，全体の精度向上につながることを確認した．

今回我々が提案した特徴的な文字列による素性は，これまでの手法でも固有表現を抽出する際に形態素や文字列として学習されている素性である．最近では，Wikipedia や Web から獲得した固有表現を抽出時の素性として使用することによる効果が福島ら[2)] や風間ら[5)] によって報告されている．しかし，前者の方法では訓練データ中の固有表現についてのみしか学習されず，出現頻度が低い場合の特徴的な文字列をとらえることは難しい．後者の場合も表現そのものが固有表現として獲得されていない場合は難しいと考えられる．これに対し，我々は別途用意した大量の固有表現とクラス語の対から特徴的な文字列を獲得することで，訓練中もしくは辞書として獲得されていない固有表現に対しても，文字列特徴から固有表現のクラスを推定することを試みた．

5. お わ り に

本稿では，固有表現に従来の IREX 定義よりも詳細な分類クラスを付与するタスクにつ

表 3 クラスごとの結果

タイプ	ベースライン			文字列あり			文字列+共起		
	精度	再現率	F 値	精度	再現率	F 値	精度	再現率	F 値
GPE	0.78 (5705/7338)	0.93 (5705/6154)	0.85	0.83 (5538/6684)	0.90 (5538/6154)	0.86	0.83 (5555/6668)	0.90 (5555/6154)	0.87
その他場所	0.71 (358/505)	0.35 (358/1030)	0.47	**0.54 (456/852)**	**0.44 (456/1030)**	**0.48***	0.55 (454/818)	0.44 (454/1030)	0.49
地形名	0.70 (174/250)	0.29 (174/593)	0.59	**0.75 (359/481)**	**0.61 (359/593)**	**0.67***	0.77 (350/456)	0.59 (350/593)	0.67
アドレス	0.70 (367/523)	0.51 (367/714)	0.59	0.60 (390/652)	0.55 (390/714)	0.57	0.59 (401/679)	0.56 (401/714)	0.57
国際組織名	0.96 (91/95)	0.61 (91/148)	0.75	**0.67 (114/171)**	**0.77 (114/148)**	**0.71***	0.80 (107/134)	0.72 (107/148)	0.75
公演組織名	0.41 (235/568)	0.30 (235/787)	0.35	0.44 (278/632)	0.35 (278/787)	0.39	0.51 (294/576)	0.37 (294/787)	0.43
競技組織名	0.71 (1704/2408)	0.67 (1704/2555)	0.69	0.70 (1504/2144)	0.59 (1504/2555)	0.64	**0.77 (1680/2171)**	**0.66 (1680/2555)**	**0.71***
法人名	0.56 (1753/3138)	0.61 (1753/2877)	0.58	**0.63 (2060/3248)**	**0.72 (2060/2877)**	**0.67***	0.64 (2088/3264)	0.72 (2088/2877)	0.68
政治組織名	0.72 (1027/1429)	0.71 (1027/1440)	0.72	**0.85 (1161/1370)**	**0.80 (1161/1440)**	**0.82***	0.88 (1181/1344)	0.82 (1181/1440)	0.84
その他組織名	0.56 (186/331)	0.36 (186/510)	0.44	0.52 (148/281)	0.29 (148/510)	0.37	0.53 (153/290)	0.30 (153/510)	0.38
GOE	0.57 (2417/4213)	0.66 (2417/3669)	0.61	**0.63 (2616/4180)**	**0.71 (2616/3669)**	**0.67***	**0.62 (2676/4301)**	**0.72 (2676/3669)**	**0.67***
路線名	0.71 (48/68)	0.29 (48/163)	0.42	**0.74 (82/110)**	**0.50 (82/163)**	**0.60***	0.75 (79/105)	0.48 (79/163)	0.59
その施設名	0.31 (19/61)	0.07 (19/287)	0.11	**0.36 (44/122)**	**0.15 (44/287)**	**0.22***	0.41 (50/121)	0.18 (50/287)	0.24

いて，固有表現を構成するクラスに特徴的な文字列と，文章中で共起する重要な語および語彙大系のカテゴリ情報を使用した分類手法を検討した．IREX 定義の地名，組織名について 13 クラスに詳細化したデータで評価し，全体の精度向上に貢献したことを確認した．また，クラスごとにそれぞれの情報の効果を確認したところ，特徴的文字列素性については，特にコーパス中で出現頻度の低い固有表現が多かったクラスで効果が見られ，文章全体から得た共起情報については，固有表現の曖昧性が多かった競技組織で効果があったことを確認した．一方で地名間のクラス分類については知識として知らなければ解けない問題が多くあり，知識を構築する方向も考える必要があることが分かった．

今後の課題として，今回は抽出された固有表現に対してそのクラスを分類する手法に焦点を当てたが，実際に固有表現を従来のチャンキング手法で抽出した結果と組み合わせた固有表現全体の抽出精度について検証する必要がある．またその結果を，橋本ら[3] が報告しているチャンキングと分類を同時に解く手法による結果と比較し，拡張固有表現抽出を解く上で難しい問題について明らかにしたい．

参 考 文 献

1) 浅原正幸，松本裕治：日本語固有表現抽出におけるわかち書き問題の解決，情報処理学会論文誌， Vol.45, No.5, pp.1442–1450 (2004).
2) 福島健一，鍛治伸裕，喜連川優：日本語固有表現抽出における超大規模ウェブテキストの利用，第 19 回データ工学ワークショップ論文集 (DEWS2008) (2008).
3) 橋本泰一，中村俊一：拡張固有表現タグ付きコーパスの構築- 白書，書籍，Yahoo!知恵袋コアデータ-，言語処理学会第 16 回年次大会発表論文集，pp.916–919 (2010).
4) 池原 悟，宮崎正弘，白井 諭，横尾昭男，中岩浩巳，小倉健太郎，大山芳史，林良彦：日本語語彙大系，岩波書店 (1997).
5) 風間淳一，鳥澤健太郎：Web 上の資源から構築した複数の固有表現辞書を用いた日本語固有表現認識，言語処理学会第 14 回年次大会発表論文集，pp.813–816 (2008).
6) 工藤 拓，松本裕治：半構造化テキストの分類のためのブースティングアルゴリズム，情報処理学会論文誌， Vol.45, No.9, pp.2146–2156 (2004).
7) McCallum, A. and Li, W.: Early results for Named Entity Recognition with Conditional Random Fields, Feature Induction and Web-Enhanced Lexicons, *Proceedings of the Seventh Conference on Natural Language Learning*, pp.188–191 (2003).
8) 中野桂吾，平井有三：日本語固有表現抽出における文節情報の利用，情報処理学会論文誌， Vol.45, No.3, pp.934–941 (2004).
9) 笹野遼平，黒橋禎夫：大域的情報を用いた日本語固有表現認識，情報処理学会論文誌，Vol.49, No.11, pp.3765–3776 (2008).
10) 新納浩幸，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 回年次大会発表論文集，pp.105–108 (2006).
11) 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， Vol.43, No.1, pp.44–53 (2002).