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

福島　健一† 　鍜治　伸裕† 　喜連川　優†

† 東京大学生産技術研究所　〒 153-8505 東京都目黒区駒場 4–6–1
E-mail: †{ken,kaji,kitsure}@tkl.iis.u-tokyo.ac.jp

あらまし　本稿では、ウェブ上のテキストなどの大規模コーパスから効率的に情報を抽出し、それを固有表現抽出タスクで利用する手法を提案する。コーパスから固有名リストの形式で知識を抽出し、この知識を固有表現抽出の教師付き学習モデルに組み込む。50 億文規模の巨大なウェブコーパスを使った実験を行い、提案手法が教師データに出現しない固有表現（未知固有表現）の抽出精度を向上させることを確認した。
キーワード　固有表現抽出、ウェブコーパス

Use of Massive Amounts of Web Text in Japanese Named Entity Recognition

Ken'ichi FUKUSHIMA†, Nobuhiro KAJI†, and Masaru KITSUREGAWA†

† Institute of Industrial Science, University of Tokyo　Komaba 4–6–1, Meguro-ku, Tokyo, 153–8505 Japan
E-mail: †{ken,kaji,kitsure}@tkl.iis.u-tokyo.ac.jp

Abstract　In this paper we propose a method to efficiently extract information from large corpora such as text on the Web and use it in named entity recognition. The method extracts knowledge from corpora in the form of an *entity list* and incorporate the knowledge into feature design of existing supervised models. We conducted experiments with a Web corpus comprised of as many as 5 billion sentences and confirmed that our proposed method improves the recall of unknown named entities, which are not included in the training data.
Key words　Named Entity Recognition, Web Corpus

1. はじめに

テキスト中で人名、地名、時刻表現、金額表現などの固有表現が使われている箇所を同定し、その分類を行う固有表現抽出は、自然言語処理における基礎技術の一つであり、質問応答や情報抽出などの高度なテキスト処理アプリケーション構築のために必要不可欠だと考えられている。自然言語処理における他の様々なタスクと同様に、固有表現抽出では教師付き学習を用いることが一般的であり、そのアプローチは一定の成功を納めている。

教師付き学習を用いた固有表現抽出では、正解の抽出結果をテキストに人手で付与した教師データを用意し、それを学習モデルに与える。モデルは固有表現抽出規則を教師データから学習し、このモデルを未知のテキスト（テストデータ）に適用することによって固有表現の抽出を行う。モデルはデータの特徴（素性）を表現する素性関数から構成されており、モデルを学習させることは固有表現を認識する手がかりとなる素性を見出すことに等しい。例えば、「福田首相らと会談へ」という文の「福田」が人名であるという教師データからモデルが『「首相」に先行する文字列は人名である可能性が高い』という規則を学習すれば、「ブラウン首相の側近によると」という文の「ブラウン」が人名であることを認識できるようになる。この「首相」の例のように、教師データ中の事例と有効な素性を共有していれば、モデルはテストデータ中の固有表現を抽出できる。逆に教師データ中の事例と有効な素性を共有していない固有表現は原理的に認識不可能である。例えば教師データが先と同じでテストデータが「オバマの演説のうまさ」の場合、素性を共有していないので「オバマ」が人名であることを認識するのは不可能だと思われる。この例を認識するためには何らかの方法で素性を共有させてやる必要がある。

一方、近年のウェブの発展により我々が利用可能なテキストは爆発的に増えている。ウェブ上のテキスト量は教師データの典型的なサイズに比べて桁違いに大きく、非常に利用価値の高い言語資源として有望視されている。ただし教師付き学習を用いた従来手法においてそのままウェブテキストを利用することはできず、工夫が必要である。半教師付き学習や、コーパスを前処理して有用と思われる情報だけを従来モデルに素性として

組み込むなどのアプローチが提案されている。

我々はウェブテキストを用いて日本語固有表現抽出の精度を向上させる手法を提案する。具体的には、固有名リストという形式でテキストから知識を抽出し、この知識を従来モデルの固有表現抽出手法で素性として利用する。この新たな素性によって先の例のような素性の不共有を解消し、固有表現の認識制度を上げることが提案手法の狙いである。手法の直観的な解釈を、認識に失敗した先の「オバマ」の例で説明する。「福田」と「オバマ」がどちらも「政治家」というカテゴリに属す語句であることがコーパスから分かるならば、この知識を素性としてエンコードし教師データとテストデータの間で共有させてやることによって「オバマ」を認識できる可能性が生まれる。

固有表現抽出に関する従来研究は既知固有表現と未知固有表現を区別せずに抽出手法を評価していた。既知・未知固有表現とはそれぞれ教師データに出現する・しない固有表現である。既知固有表現の場合、固有表現の表層的な文字列そのものからまったく汎用性のない規則を学習してしまうことがありうる。例えば教師データ「オバマ氏が逆転」から『「オバマ」は人名である』という自明な規則を学習してしまえば、テストデータ「オバマの演説のうまさ」を正しく処理することは確かに可能である。しかし我々が教師付き学習に期待しているのはこのような自明で汎用性のない規則（まる覚え）ではなく、最初の例の「首相」のように未知固有表現をも抽出可能な汎用的な規則を学習することである。既知・未知固有表現を区別しない従来の評価方法では、まる覚えの事例が未知固有表現の抽出精度を覆い隠してしまう可能性があるので、本来これらは切り分けて評価するべきである。この切り分けを行って我々が固有表現抽出の精度を評価した結果が表 1 である。既存手法の最大公約数的なシンプルな手法で、日本語固有表現抽出の標準的なベンチマークである CRL データを使って実験を行った。固有表現の種類にもよるが、未知固有表現の再現率は既知固有表現に比べて 30% も低い。

表 1　既知固有表現と未知固有表現の再現率の違い。括弧内の数字は事例数を表す。

クラス	再現率–既知	再現率–未知
固有物名	89.05 (288)	31.66 (459)
地名	94.80 (4485)	63.85 (978)
組織名	93.29 (2560)	58.10 (1116)
人名	96.30 (2454)	82.07 (1386)
TOTAL	94.61 (9787)	64.81 (3939)

本論文の構成は以下の通りである。次の 2 章で固有表現の定義など基礎的なことがらと、固有表現抽出の標準的な定式化手法である系列ラベリング問題について説明する。3 章で我々の提案手法の詳細を述べ、それを実験によって評価した結果を 4 章で報告する。5 章で本研究と関連する既存研究について触れ、6 章でまとめを述べて本論文の結びとする。

2. 固有表現抽出

2.1 定　　義

人名・地名などの固有名詞と日付・金額などの数値表現を総称して固有表現 (named entity; NE) という。テキスト中の固有表現の出現箇所を同定しその種類（固有表現クラス）を判断することを固有表現抽出といい、質問応答、情報抽出などのより高次のテキスト処理のために必要不可欠な基礎技術の一つと認識されている。固有表現クラスの定義はテキストの分野や目的のアプリケーションに強く依存するが、日本語固有表現抽出手法の評価・比較の際には IREX ワークショップ [2] による定義が標準的に用いられる。IREX は表 2 に示す 8 つの固有表現クラスを定義している。その内訳は固有物名 (ART)、地名 (LOC)、組織名 (ORG)、人名 (PSN) の 4 つの固有名詞、日付 (DAT)、時刻 (TIM)、金額 (MNY)、割合 (PNT) の 4 つの数値表現となっている。注意が必要なのは ART で、これは LOC、ORG、PSN のいずれにも該当しない固有名詞をまとめた多様性の高いクラスである。固有表現抽出では歴史的な経緯から固有名詞と数値表現が同等に扱われているが、現実にはその性質はかなり異なる。本論文では原則的に固有名詞のみを念頭において議論を進める。

表 2　IREX による固有表現クラスの定義

クラス	例
ART 固有物名	ノーベル文学賞、我輩は猫である
LOC 地名	アメリカ、インド洋、お台場
ORG 組織名	自民党、トヨタ
PSN 人名	福田康夫、朝青龍、浜崎あゆみ
DAT 日付	二月三日、2007/08/12
TIM 時刻	午後三時、AM 10:30
MNY 金額	２０億円、35＄
PNT 割合	３５%、2 割

2.2 系列ラベリング問題としての定式化

現在用いられているほとんどの固有表現抽出手法では、タスクを系列ラベリング問題としてモデル化し、モデルのパラメータを教師付き学習の枠組みでデータから学習する。

系列ラベリング問題とは、与えられた入力トークン列 $\boldsymbol{x} = (x_1, \ldots, x_N)$ に対して、最適な出力ラベル列 $\boldsymbol{y} = (y_1, \ldots, y_N)$ を対応付ける問題である。日本語固有表現抽出では x_n は処理対象の文中の文字、y_n は固有表現の存在とクラスをエンコードするラベルである。ラベルを使った固有表現抽出を図 1 を使って説明する。図 1 は「…自民党総裁の福田首相は…」という文に対してラベル付けを行った様子である。この文には「自民党」(ORG)、「福田」(PSN) という 2 つの固有表現が出現している。B-XXX、I-XXX というラベルがこれらの文字列が固有表現であることを表現している。B-XXX は固有表現文字列の 1 文字目、I-XXX は 2 文字目以降を意味する。XXX 部分には ORG、PSN などの固有表現クラスが入る。固有表現でない部分には O というラベルが付与される。固有表現クラスは 8 種類あるので、y_n は $2 \times 8 + 1 = 17$ 通りの値をとる変数である。

$\boldsymbol{y}$	…	B-ORG	I-ORG	I-ORG	O	O	O	B-PSN	I-PSN	O	O	O	…
$\boldsymbol{x}$	…	自	民	党	総	裁	の	福	田	首	相	は	…

図 1　ラベル付けの例

系列ラベリングのモデルは入力文 $\boldsymbol{x}$ を与えられるとそのラベル列の推定値

$$\hat{\boldsymbol{y}} = \underset{\boldsymbol{y}\in \mathrm{GEN}(\boldsymbol{x})}{\operatorname{argmax}} \sum_{k=1}^{K}\sum_{n=1}^{N-1} \lambda_k f_k(\boldsymbol{x}, y_n, y_{n+1}, n)$$

を出力する。$\mathrm{GEN}(\boldsymbol{x})$ は $\boldsymbol{x}$ に対して可能なあらゆるラベル列の集合、$\sum_k \sum_n \lambda_k f_k(\boldsymbol{x}, y_n, y_{n+1}, n)$ は $\boldsymbol{x}$ と $\boldsymbol{y}$ の対応の良さを評価するスコア関数、λ_k はモデルのパラメータである。スコア関数は素性関数 f_k によって構成され、パラメータはその重みを与える形になっている。素性関数は入力文の様々な特徴とラベルの依存関係を記述する。素性関数はその引数の形さえ守れば任意の構造、任意の値域（実数）が許されるが、典型的には入力部分とラベル部分に分離した真偽値関数のみを使う。例えば、

$$f_k = \begin{cases} 1 & \text{if } y_n = \text{I-ORG} \wedge x_n = \text{“党”} \\ 0 & \text{otherwise} \end{cases}$$

という素性関数を定義すれば「“党” で終わる文字列は組織名の可能性が高い」という依存関係をモデルに取り込むことができる。また素性関数は入力 $\boldsymbol{x}$ の表層的な文字のみならず、外部のリソースを使って導出される情報をなんでも利用することができる。例えば形態素解析の結果を利用して、

$$f_k = \begin{cases} 1 & \text{if } y_n = \text{O} \wedge x_n\text{の品詞} = \text{“助詞”} \\ 0 & \text{otherwise} \end{cases}$$

という素性関数を定義すれば「助詞は固有表現の一部でない可能性が高い」という関係を表現できる。素性関数が記述するデータの特徴を素性という。日本語固有表現抽出における典型的な素性設計を図 2 に示す。n 番目のラベル（I-PSN）の推定

	位置	$n-2$	$n-1$	n	$n+1$	$n+2$
$\boldsymbol{y}$		O	B-PSN	I-PSN	O	O
$\boldsymbol{x}$	文字	の	福	田	首	相
	字種	HIRA	OTHER	OTHER	OTHER	OTHER
	単語	B-の	B-福田	I-福田	B-首相	I-首相
	品詞	B-助詞	B-固有名詞	I-固有名詞	B-一般名詞	I-一般名詞

図 2　固有表現抽出における素性設計の例

に使われる素性が図の $\boldsymbol{x}$ 部分に記されている。元の入力文に対して字種や形態素解析結果の単語・品詞などの情報を展開し、あらかじめ定められたウインドウ（図ではそのサイズは 5）内にある情報を利用してラベルを推定する。

モデルのパラメータ λ_k は教師付き学習の枠組みでデータから学習する。系列ラベリングのための学習アルゴリズムは Conditional Random Fields [4]、Structured Perceptron [1]、Max-Margin Markov Networks [8] などが存在し、固有表現抽出においても特別な工夫をせずにそれらを利用可能である。学習アルゴリズムの詳細については参考文献を参照されたい。

3. 提 案 手 法

本論文では、大規模なウェブコーパスから固有名リストという形式で知識を収集し、このリストの情報を素性として系列ラベリングのモデルに取り込むことによって固有表現抽出の精度を上げる手法を提案する。固有名リストは「カテゴリ名– 固有名」というペアの集合である。本節ではコーパスから固有名リストを収集する方法、固有名リストの情報の素性表現について説明する。

3.1　候補ペアの収集

まずカテゴリ名–固有名ペアの候補を生成する。テキストを形態素解析器によって単語に分割し、その単語列に次のパターン（カギ括弧パターン）を適用することによって候補を生成する。

$$noun「any+」$$

$noun$ は名詞（IPA 品詞体系 [13] における品詞細分類が “名詞–一般”、“名詞–サ変接続”、“名詞–接尾– 一般” のいずれかの単語）の単語、any は任意の品詞の単語とする。$noun$ 部分をカテゴリ名、any+ 部分を固有名として候補ペアが生成される。例えば “…地域政党「新党大地」を…” という句からは “政党–新党大地” という候補が、“…国際企業「トヨタ」の…” という句からは “企業–トヨタ” という候補が生成される（図 3）。

単語	…	地域	政党	「	新党大地	」	を	…
品詞		*	名詞–一般	*	*	*	*	
			カテゴリ名		固有名			

図 3　カギ括弧パターンの適用例

3.2　候補ペアのフィルタリング

カギ括弧パターンは確かに我々が意図したようなカテゴリ名と固有名を結びつける用法で使われるが、それ以外の用法も存在する（表 3）。典型的な用法は発話者とその発話内容の結びつけである（表 3 の iii–v）。さらに、単純に特定の語や句を強調するためだけにも用いられる（表 vi–vii）。これはカギ括弧パターンの $noun$ 部分、any+ 部分には原理的にほとんどあらゆる語句の組合せが入りうることを意味する。つまり、パターンの適用によって生成された候補ペア集合はカテゴリ名–固有名の関係にないペア（ノイズ）を多く含む。従ってこのペア集合を知識として固有表現抽出で利用するためには適切なフィルタリングを行って確かににカテゴリ名–固有名の関係にあるペアだけを選別する必要がある。ここでは候補ペア集合からカテゴリ名–固有名ペアとして信頼できる部分だけをとりだす、ノイズを取り除くという 2 つの観点に基づき、以下の 3 種類のフィルタリングを行う。

表 3　カギ括弧パターンの用法

タイプ		用例
固有表現	i	高速列車「ユーロスター」に乗って
	ii	松竹映画「男はつらいよ」シリーズ
発話	iii	ナレーション「とか言いつつひかりにあげるものを探すタケルだった。」
	iv	佐々木「それじゃ。夜遅く掛けてごめんね。教えてくれてありがとう」
	v	大泉「いえいえ」
強調	vi	かたちやスタイル「だけ」を気にしている
	vii	この作業の区切り「まで」は片付けよう

a）統計量に基づくカテゴリ名のフィルタリング

あるカテゴリ名候補について、それとペアをなす固有名候補の全体的な性質を見ることで、そのカテゴリ名候補がどの用法で用いられるのかを推定することができる。例えば固有名候補に句読点が多く含まれている、動詞が多く含まれている、文字数が多いなどの特徴があれば、そのカテゴリ名候補は表 3 の発話のパターンで使われる単語である可能性が高いと考えられる。この考察に基づき、カテゴリ名候補毎に以下の統計量を算出し、それが閾値を超える場合にはリストから除外する。

- 固有名の平均文字数
- 固有名に含まれる動詞の平均個数
- 固有名に含まれる句読点の平均個数
- 固有名の平均 n-gram 頻度(注1)

b）語彙統語パターンに基づくカテゴリ名のフィルタリング

本研究ではカギ括弧パターンを候補生成の手がかりとして利用するが、ある語句がカテゴリ名や固有名であることを示唆するパターンは他にも存在する。例えば「とある X が」というパターンの X 部分に、カテゴリ名である「企業」が入るのは自然だが、カテゴリ名ではない「発酵」が入ることは考えづらい。同様に「X ひとりが」というパターンでは、X に入る語句は人名を導くカテゴリ名である可能性が高いと考えられる。逆に、ある単語がカテゴリ名ならばその単語は必ず上記のパターンの中で使われているはずだと考えることができる。この考察に基づき、下記のパターンの X 部分に各カテゴリ名候補を代入し、コーパス中でのその使用頻度を数え、それが閾値を下回っている場合はその候補はカテゴリ名でないとしてリストから除外する。

- （ひとつ | 一つ | 1つ）の X（が | を）
- （ひとり | 一人 | 1人）の X（が | を）
- X（ひとり | 一人 | 1人）（が | で | を）
- どの X（が | を | に）
- とある X（が | を | に）
- 各 X（が | を）
- X（たち | 達）（が | を）

コーパス中でのパターンの使用頻度は、そのパターンの検索エンジン（Google）でのヒット件数で代用する。例えば X が「企業」の場合、「各 X（が | を）」というパターンは「“各企業が” OR “各企業を”」というクエリに展開され検索エンジンに投げられる。

(注1)：n-gram 頻度は固有名リスト収集と同じコーパスを使って数えた。

	…	は	新	党	大	地	に	…
⋮		⋮	⋮	⋮	⋮	⋮	⋮	
政党	…	ϕ	B-政党	I-政党	I-政党	E-政党	ϕ	…
⋮		⋮	⋮	⋮	⋮	⋮	⋮	
企業	…	ϕ	ϕ	ϕ	ϕ	ϕ	ϕ	…
⋮		⋮	⋮	⋮	⋮	⋮	⋮	

図 4　固有名リストの情報のカテゴリ素性による表現

c）n-gram 頻度に基づく固有名のフィルタリング

これは、コーパス中に頻出する語句が固有表現である可能性は低いだろうという考察に基づくフィルタリングである。あらかじめ大規模なコーパスを処理してそこに多く出現する文字 n-gram を列挙しておき(注1)、固有名候補がその中に存在する場合はその候補を固有名リストから除外する。原理的にはどんな文字列も固有名になりうるが、先に説明したようにカギ括弧パターンは単なる強調の目的などにも用いられるので、こうした文字列を残しておくよりも除外することによる固有名リストのノイズの低減の方が重要だと考える。

3.3　系列ラベリングにおける素性設計

固有名リストの情報は図 4 に示すような方法でエンコードされ、素性としてモデルに取り込まれる。本論文ではこの素性をカテゴリ素性と呼ぶ。入力文のある部分文字列がリストに固有名として含まれている場合、その範囲をその固有名とペアをなすすべてのカテゴリ名でマークする。固有名は一般に複数の文字からなるので 2.2 項で説明した方式にしたがって文字単位の情報に変換する。図 4 は「 … は新党大地に … 」という文にカテゴリ素性を展開した例である。カテゴリ素性はリスト中に固有名として存在する部分にのみ展開され、、それ以外の部分には展開されない。図中の ϕ は “ϕ” という素性ではなく、素性がないことを意味している。

3.4　素 性 選 択

固有表現抽出に固有名リストを用いる動機は、入力文中の部分文字列がリストに含まれているという事実が、その部分文字列が固有表現であることを判断する手がかりになるだろうという期待であった。これが目論見どおりに機能するためには、固有表現と完全一致するところでのみカテゴリ素性が展開されることが理想である。しかし 3.1 項、3.2 項の方法で収集した固有名リストを 3.3 項の方法でエンコードしたカテゴリ素性は、現実にはその理想からは遠く、一部のカテゴリ素性しか固有表現と一致せず、多くのカテゴリ素性が固有表現とは無関係のところで展開されてしまう。こうした望ましくないカテゴリ素性の存在はモデルの学習に悪影響を及ぼし、固有表現抽出の精度をむしろ下げてしまう。

カテゴリ素性が固有表現と無関係なところで展開されてしまうのは主に次の 2 つの理由による。1 つ目の理由は固有名リストのノイズである。カギ括弧パターンはカテゴリ名-固有名ペア

以外にも発話や強調など多様な用例を抽出してしまう。これらのノイズを除去するために各種のフィルタリングを行うが、しかしそれでもノイズをゼロにすることはできない。結果として固有表現とはまったく関係ないところにカテゴリ素性が展開されてしまうということになる。2 つ目の理由は、固有表現の多義性である。ある種の文字列は文脈に応じて異なるカテゴリの固有表現として使れたり、あるいは固有表現であったりなかったりする。前者の例として「富士」、後者の例として「さくら」などがあげられる。

こうした問題を根本的に解決するためには、固有名リストに含まれるノイズをさらに低減すること、固有表現の多義性を考慮した素性設計を行う必要がある。しかしここでは、固有表現抽出の精度をあげることを直接的に指向したヒューリスティクスを導入してこれらの問題を部分的に解決することを試みる。具体的には、展開されたカテゴリ素性の中から固有表現と一致する（可能性が高い）ものだけを選別する素性選択のステップを設ける。

先に述べたような望ましくないカテゴリ素性の事例を観察したところ、それらの事例に頻繁に現れるいくつかの特徴を見出した。その特徴とは、複合名詞の中の一部の名詞のみと固有名がマッチしている、固有名が名詞 1 単語である、などである。こうした観察に基づき、次のようなルールを定義し、そのいずれかを満たすカテゴリ素性は系列ラベリングのアルゴリズムを走らせる前にすべて削除することとした。

- 固有名の直前または直後の単語が名詞である。
- 固有名の直前または直後の単語が中点「・」である。
- 固有名が 1 形態素からなり、かつその品詞が次のいずれかである。
  - 名詞–一般
  - 名詞–サ変接続
  - 名詞–形容動詞語幹
  - 名詞–副詞可能
  - 動詞–自立
  - 形容詞–自立
  - 副詞–一般
- 固有名の末尾が数助詞でない。

4. 実　　験

4.1 固有名リストの構築

本研究で固有名リスト収集に用いるコーパスは、1999 年から 2007 年の間にクロールされた日本語ウェブページから作成されたものである。まずウェブページからテキスト部分だけを抜き出し、文に分割する。ミラーページの影響をできるだけ抑えるための工夫として、重複する文は一つだけを残して他はすべて取り除く。その他、いくつかのヒューリスティックなルールを適用して文とは認められないような文字列断片を取り除く。完成したコーパスのサイズは 325GB であり、54 億異なり文からなる。

このコーパスにカギ括弧パターンを適用して、2481 万種、述べ 4561 万個のペアからなる候補集合を得た。この集合は 4.5 万種のカテゴリ名候補、1785 万種の固有名候補を含む。これに 3.2 項で説明したフィルタリングを施して固有表現抽出で使うための固有名リストを作成する。各フィルターの閾値は、出来るだけ有用な情報を保持しつつノイズが除去されるように、予備実験を行いながら手作業で決定した。このようにして作られた最終的な固有名リストは、ペアを 638 万種、カテゴリ名を 378 種、固有名を 519 万種含む。

4.2 固有表現抽出実験

固有名リストの利用が固有表現抽出精度にどう寄与するのかを明らかにするために CRL データセットを使ってで固有表現抽出実験を行う。CRL データセットは日本語固有表現抽出手法を評価するための代表的なベンチマークであり、1995 年の毎日新聞 1174 記事 10718 文に 2.1 節で説明した IREX が定義する 8 種の固有表現 18677 個を人手でマークして作られている。

固有表現抽出の精度を評価する代表的な指標は、適合率 P、再現率 R、F 値 F の 3 つである。テストデータ中にマークされた固有表現の集合を A、モデルが固有表現だと推定して出力したものの集合を B とすると、これらの指標は次の式で定義される。

$$P = \frac{|A \cap B|}{|B|}$$

$$R = \frac{|A \cap B|}{|A|}$$

$$\frac{1}{F} = \frac{1}{2}\left(\frac{1}{P} + \frac{1}{R}\right)$$

つまり、適合率はモデルが固有表現だとして出力したものの正答率、再現率はモデルが全固有表現のうちいくつを見つけることができたかを意味する。そして適合率と再現率は一般にトレードオフの関係にあるので、これらのバランスを考慮した全体的な認識精度のよさを与える指標として F 値がある。

機械学習を使った手法の評価を行うためには教師データとテストデータを厳密に区別しなければならない。CRL データセットを使った過去の研究に倣い、本研究では 5 分割交差検定によって評価を行う。5 分割交差検定ではまずデータを 5 つのセットに分割する。このうち 4 つを教師データ、1 つをテストデータとして実験のラウンドを回す。5 つのセットそれぞれをテストデータとする計 5 回のラウンドを回し、各ラウンドでの精度の平均を全体の精度とする。

表 4 が固有表現抽出（ラベル付け）の精度である。固有名リストを使うことが未知固有表現の認識精度向上にどう寄与するのかを明らかにするために、既知/未知を区別した再現率を載せる。一方、適合率を既知/未知を区別して議論することにはあまり意味がないと考えるので、それらの数値は省略した。また括弧内の数値は、ベースライン（固有名リストを用いなかった場合）との差である。比較対象の数値は表 1 にあげたものと等しい。すべてのクラスで未知固有表現の再現率が上がり、提案手法の有効性を確認できた。元の再現率が低かった ART で向上幅が 7.00%と特に大きい。既知の固有表現に対してもベースラインの再現率を概ね保っており、カテゴリ素性を導入したことによる副作用は無視できる範囲である。

表 4　固有表現抽出精度。括弧内はベースラインとの差（固有名リストを用いなかった場合）との差。

クラス	再現率–既知	再現率–未知
ART	88.11 (−0.94)	38.66 (+7.00)
LOC	95.06 (+0.26)	65.30 (+1.45)
ORG	93.65 (+0.36)	60.58 (+2.48)
PSN	96.21 (−0.09)	82.16 (+0.09)
TOTAL	94.78 (+0.17)	66.69 (+1.88)

最後に、固有表現抽出手法の性能を端的に表す数値に、既知と未知、固有名詞と数値表現を区別せずに計算した F 値がある。CRL データを使った固有表現抽出実験では伝統的にこの数値によって抽出手法が評価されてきた。提案手法ではこの数値は 89.29%であり、ベースラインの 89.01%を 0.28%上回った。

5. 関 連 研 究

固有表現抽出は自然言語処理における基礎的な技術であり、日本語を対象とした研究も数多く存在する。かつてはどの教師付き学習モデルをどのように適用するかということを追求した研究が多かったが [9]〜[12], [14]、学習モデルの使用法に関する知見や経験が蓄積されてきた現在では、どのような素性をモデルに組み込むかということに興味の中心が移って来ている [5], [15]。

外部の言語リソースから自動獲得した知識を固有表現抽出で利用する手法として、英語を対象とした Talukdar ら [7] や風間ら [3] の研究がある。Talukdar [7] は、我々のカギ括弧パターンに当たる、固有表現を示唆するようなパターン自体を教師データから帰納的に獲得し、そのパターンをコーパスに適用して固有表現を収集した。風間ら [3] は Wikipedia から我々のカテゴリ名–固有名ペアと類似の知識を抽出しラベル付けに利用した。いずれの手法も固有表現抽出の精度向上に有効であると報告されている。ただし、これらの研究が対象とする英語には固有表現はキャピタライズされるという極めて強いヒューリスティクスが存在し、これらの手法をそのまま日本語に適用するのは難しいと思われる。

固有表現抽出タスクの精度向上を直接的に指向したものに限らなければ、ウェブなどの大規模コーパスから固有表現やそれに類する知識の収集を行った研究は多数存在する。隅田ら [6] は我々と同様にカギ括弧を手がかりにして固有名リストを収集し、さらにそれをシードにしてパターンのないところからも固有名の収集を行っている。隅田らの手法の目的は品質の高い固有名リストの作成である。また手法の適用対象が名詞列のみであり、収集できる固有名の種類が限られている。

6. お わ り に

本論文では、日本語固有表現抽出においてコーパスから自動獲得した知識を利用する手法を提案した。具体的には、コーパスから固有名リストの形式で知識を抽出し、この知識を従来の教師付き学習モデルに素性として組み込む。50 億文規模の大規模な実験を行い、提案手法が教師データに出現しない未知固有表現の再現率を向上させることを確認した。

文　　献

[1] Michael Collins. Discriminative Training Methods for Hidden Markov Models: Theory and Experiments with Perceptron Algorithms. In *Proc. of EMNLP*, pp. 1–8, 2002.

[2] IREX 実行委員会（編）. IREX ワークショップ予稿集, 1999.

[3] Jun'ichi Kazama and Kentaro Torisawa. Exploiting Wikipedia as External Knowledge for Named Entity Recognition. In *Proc. of EMNLP-CoNLL*, pp. 698–707, 2007.

[4] John Lafferty, Andrew McCallum, and Fernando Pereira. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proc. of ICML*, pp. 282–289, 2001.

[5] Ryohei Sasano and Sadao Kurohashi. Japanese Named Entity Recognition Using Structural Natural Language Processing. In *Proc. of IJCNLP*, 2008.

[6] Asuka Sumida, Kentaro Torisawa, and Keiji Shinzato. Concept-Instance Relation Extraction from Simple Noun Sequences Using a Full-Text Search Engine. In *Proc. of Workshop on Web Content Mining with Human Language Technologies*, 2006.

[7] Partha P. Talukdar, Thorsten Brants, Mark Liberman, and Fernando Pereira. A Context Pattern Induction Method for Named Entity Extraction. In *Proc. of CoNLL*, pp. 141–148, 2006.

[8] Ben Taskar, Carlos Guestrin, and Daphne Koller. Max-Margin Markov Networks. In *Proc. of NIPS*, 2004.

[9] 山田寛康. Shift-Reduce 法に基づく日本語固有表現抽出. 情報処理学会研究報告 NL-179, pp. 13–18, 2007.

[10] 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.

[11] 福岡健太. Semi-Markov Conditional Random Fields を用いた固有表現抽出に関する研究. Master's thesis, 奈良先端科学技術大学院大学 情報科学研究科, 2006.

[12] 磯崎秀樹, 賀沢秀人. 固有表現抽出のための SVM の高速化. 情報処理学会論文誌, Vol. 44, No. 3, pp. 970–979, 2003.

[13] 浅原正幸, 松本裕治. ipadic version 2.7.0 ユーザーズマニュアル, 2003.

[14] 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.

[15] 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.