2007－NL－182 (8)
2007／11／20

文書からの組織名抽出における辞書利用

乾 孝司　村上 浩司　橋本 泰一　内海 和夫　石川 正道

概要：本稿では，固有表現クラスのひとつである組織名を題材にして，固有表現抽出課題において使用される固有表現辞書に関する幾つかの調査結果を報告する．まず，固有表現辞書の特性を定量的に記述する特性パラメータを定義し，特性パラメータと固有表現抽出精度の関係性を調査した．調査結果をもとに，辞書知識の自動整備の進め方について検討した．次に，効果的な固有表現辞書の利用法を提案した．提案手法は辞書エントリ集合から正規表現規則を自動獲得する．評価実験を通して，正規表現規則の情報が固有表現抽出精度の向上に有効に働くことを確かめ，既存手法と同程度の性能（$F_{\beta=1}$ で 84.58）を達成できることを確認した．

キーワード：固有表現抽出，固有表現辞書，組織名

A Study on Using Gazetteers for Organization Name Recognition

Takashi Inui　Koji Murakami　Taiichi Hashimoto
Kazuo Utsumi　Masamichi Ishikawa

Abstract: In this paper, we report some issues on the named entity recognition (NER) task which is performed with methods using gazetteer information, especially focusing on organization names as the target named entities. First, we defined a set of parameters, P, to describe the characteristics of gazetteers used to the NER, and investigated relationships between the P, and accuracy of the NER which is achieved by the method using gazetteers holing P. We also proposed a method of automatically acquiring a set of regular expression rules, which are applied to the NER task substitute for gazetteers. Experimental results showed that the rules acquired by our proposed method improved the NER performance, and achieved the 84.58 F-value for evaluation data.

Keywords: named entity recognition, gazetteers, organization name

1 はじめに

固有表現抽出（named entity recognition；NER）課題 [7, 9] とは，人名や地名，組織名等の固有な事物を指す表現をテキストから自動抽出する課題であり，情報抽出技術の必須の要素技術として，多くの技術開発がなされている．NER 課題では，技術開発の初期の頃より，抽出すべき固有表現の部分集合をリストアップした知識（固有表現辞書と呼ぶ）が利用されることがある．表 1 は辞書エントリの例であり，「国会」，「東京工業大学」が組織名を表す可能性があること，「クリントン」，「東京都」がそれぞれ人名，地名を表す可能性があることが示されている．先行研究 [2, 18] では，このような辞書知識が NER 課題に有効に働くと報告されている．

本稿では，固有表現辞書を利用して固有表現抽出課題を実施する状況を考える．そして，この状況において，これまで十分な議論がなされていないと考えられる以下の２つの項目について調査，検討をおこなう．

- 利用する固有表現辞書の特性と抽出精度の関係
- 固有表現辞書の効果的な利用法

先述したように，辞書知識が NER に有用であることは既に認識されているが，抽出すべき固有表現クラスが定まった時に，どのような辞書を利用するのが望ましいかについて，現在までまとまった知見は得られていない．そこで 4 節において，まず，辞書の特性を表す特性パラメータを定義する．次に，異なる特性パラメータをもつ複数の辞書を用意し，辞書の特性パラメータと NER の精度の関係を調査する．そして，その結果をもとに，固有表現辞書の自動整備の進め方について検討する．

近年，NER 課題は，入力テキストへの系列ラベリング問題として形式化されることが多く，Support Vector Machines（SVMs）を用いた手法 [20, 19] や，Conditional Random Fileds（CRFs）を用いた手法 [6, 12] 等が盛んに研究されている．また，辞書知識が NER に有用であることから，固有表現辞書の自動整備を狙った研究も幾つか報告されている [17, 14]．しかしながら，固有表現

表 1　固有表現辞書のエントリ例

エントリ	固有表現クラス
国会	ORGANIZATION, 組織名
東京工業大学	ORGANIZATION, 組織名
クリントン	PERSON, 人名
東京都	LOCATION, 地名

東京工業大学 統合研究院.
Tokyo Institute of Technology.
連絡先：inui@iri.titech.ac.jp

表 2　入出力系列の例

入力系列	出力系列
タイ	B–LOCATION
・	O
チュラロンコン	B–ORGANIZATION
大学	I–ORGANIZATION
の	O

辞書の知識の利用面についてはあまり議論されていない．5 節では，NER 課題における新しい固有表現辞書の利用法を提案する．具体的には，固有表現辞書の特性と抽出精度の関係調査から得られた知見をもとに，辞書知識から固有表現照合用の正規表現規則を自動獲得し，NER 課題に適用する方法を提案する．

上述した調査項目は，NER の手法や，抽出対象とする固有表現クラスに依存すると考えられる．本研究では，NER の手法として山田らの手法 [16]，また抽出対象として，IREX [9] で定義された固有表現クラスのひとつである組織名（ORGANIZATION）を選び，各調査項目を検討した．

以下本稿では，2 節と 3 節で，ベースとなる固有表現抽出手法，データ等，調査環境の諸々の設定について説明する．その後，4 節で，固有表現辞書の特性と抽出精度の関係について延べ，5 節 で，効果的な固有表現辞書の利用法について述べる．6 節で関連研究を紹介し，7 節で本稿をまとめる．

2　準備

2.1 節で，本研究で扱う固有表現クラスである組織名の定義，2.2 節で，本研究でベースとした固有表現抽出手法である山田らの手法の概要を説明する．また，2.3 節において，山田らの手法に辞書知識を取り込む単純な手法を導入する．

2.1　IREX における組織名の定義

組織名の定義は IREX（version 990214）に従う．IREX における組織名の定義の記述部分を以下に抜粋，引用する．

- （3.1.1 組織名）
  組織名とは、複数の人間で構成され、共通の目的を持った組織等の名称の事である。株式会社等の会社、固有の政府組織、学校、軍、スポーツチーム、国際組織、労働組合、工場、ホテル、空港、病院、教会やなんらかの目的を持ったグループ等もその対象が組織としての意味で使われている文脈においては組織名とする。

2.2　山田らの固有表現抽出手法

山田らの固有表現抽出手法 [16] の概要を説明する．4 節 以降で述べる内容に関し，解析の際，設定が必要なパラメータは山田らの評価実験の結果を参考にして最も性能の良い値に固定した．そのため，以降の説明でもパラメータの自由度がある場合には，ある固定した値に限定し説明をおこなう．手法の詳細説明は [16] に譲る．

近年，NER 課題は入力テキストへの系列ラベリング問題として扱うことが多く，山田らの手法でも系列ラベリングとして問題を扱っている．系列ラベリング問題は，入力系列が与えられた時に，それに対応する（何らかの意味で）最適な出力系列を求める問題である．NER 課題では，固有表現の情報が出力系列として符号化される．表 2 に，単語系列を入力とした場合の出力ラベル系列の例を示す．左列が入力，右列が出力であり，入力中の単語ごとにひとつの出力ラベルが与えられる．例では，単語「タイ」が地名，単語列「チュラロンコン大学」が組織名を表していることが示されている．出力ラベルは，チャンク情報と固有表現クラスの情報を組み合わせることで構成される．チャンク情報の符号化には幾つか種類があるが，本研究では IOB2 表現を採用した．IOB2 表現では，固有表現の開始位置の要素に「B」タグ，固有表現の中で開始位置以外の要素に「I」タグを付与することでチャンク情報を表す．また，固有表現でない箇所には「O」タグが付与される．上記の説明では，IREX の固有表現クラスセットを想定して述べたが，実際の調査では固有表現クラスとして組織名のみを扱うので，考慮すべき出力ラベルは「B–ORGANIZATION」，「I–ORGANIZATION」，「O」の 3 種類となる．

山田らの手法では，文頭あるいは文末から順に，各単語に対する出力ラベルを SVMs [13] によって決定的に推定していくことで出力ラベル系列を得る．解析順序は，文頭から文末，あるいは文末から文頭の 2 通りが考えられるが，本研究では文末から文頭に向かって解析する．SVMs は 2 クラスを対象とした 2 値分類器であるが，一般に，考慮すべき出力ラベル数は 2 クラス以上になる（本研究では 3 クラスになる）ため，実際には，SVMs を多値に拡張したものを用いる．代表的な多値への拡張方式には，pairwise 方式と one-versus-rest 方式がある．山田らは pairwise 方式しか検討していないが，予備実験として両方式を比較した結果，one-versus-rest 方式の方が良い性能を示したため，本研究では one-versus-rest 方式を採用する．

分類の素性には，対象単語および前後 2 単語の計 5 単語に対し，単語自身，品詞，単語を構成する文字種を用いる．単語境界情報，品詞情報は ChaSen [5] で解析をして得る．文字種は，「カタカナ」，「平仮名」，「漢字」，「記号」，「数字」，「アルファベット」の 6 種類で，単語に含まれる文字種すべてを素性情報として用いる．また，文末側の推定済みのラベル情報も素性情報として用いる．ただし，学習時には学習コーパスによって与えられた正しいラベル情報を用いる．表 3 に，入力系列「タイ・チュラロンコン大学の」中の単語「チュラロンコン」の出力ラベルを推定する際に使用する素性の例を示す．

実際には，以上の素性情報を 0/1 のベクトルに変換して SVMs の学習，分類に用いる．単語自身の情報が含まれるため，高次元ベクトルとなるが，値が 1 となる成分はわずかしかなく，各ベクトルは非常に疎である．

単語ベースで処理を行う場合，固有表現の境界と単語の境界で不整合が生じる（すなわち，単語内の一部の文字列として固有表現を含む）と，ラベル情報が表現できず適切に対処できない [20]．例えば，ChaSen では文字列「自公民」はひとつの単語として解析されるが，後述する学習データでは，「自」，「公」，「民」の各文字がそれぞれ組織名として判定されており，不整合が生じている．本研究では，上例のような不整合が生じる事例については学習データ中で固有表現クラス情報を無視した．つまり，上例の「自公民」等には「O」タグを付与した．ただし，評価データ中の該当箇所については評価対象に含めるため，該当箇所ではすべて誤ることになる．

表 3　素性情報の例

単語	品詞	文字種	出力ラベル
タイ	名詞-固有名詞-地域-国	カタカナ	(使用しない)
・	記号-一般	記号	(使用しない)
チュラロンコン	名詞-サ変接続	カタカナ	(推定箇所)
大学	名詞-一般	漢字	I-ORGANIZATION
の	助詞-連体化	平仮名	O

表 4　辞書素性

単語	Simple
タイ	O
・	O
チュラロンコン	I
大学	I
の	O

2.3　組織名辞書の利用法

山田らの論文 [16] では，固有表現辞書の情報は素性情報として利用していない．そこで，山田らの手法に対して，辞書知識を利用する単純な方法を導入する．

まず，入力単語系列から名詞列（品詞が名詞である単語列）を抽出する．そして，抽出された各名詞列について以下の操作を行い，素性情報を獲得する．以降，この素性情報獲得法を**単純照合法（Simple）**と呼ぶ．

単純照合法

- 名詞列と各辞書エントリとの文字列照合を試みる．名詞列が，あるエントリの文字列すべてを包含している場合を照合が成立したとみなし，名詞列の照合箇所に「I」タグを与える．照合しない残りの他の箇所には「O」タグを与える．ただし，名詞列と照合するエントリが複数ある場合は最長のものを選ぶ．また，前節末尾の議論と同様の理由で，名詞列に含まれる単語内の一部の文字列と照合する場合は，未照合として扱う．

辞書エントリとして「チュラロンコン大学」が存在する場合の辞書素性の例を表 4 に示す．

他の素性と同様，辞書素性も，対象単語および前後 2 単語の計 5 単語の値を使用する．

4 節では，単純照合法によって辞書知識を取り込むことで，辞書特性と抽出精度の関係を調査する．ここで紹介した単純照合法は先行研究から容易に類推できる方法であるが，続く 5 節では，単純照合法よりも効果的な素性情報の抽出を実現する手法を提案する．

3　利用するリソース

3.1　テキストコーパス

CRL 固有表現データ*を学習，評価用のテキストコーパスとして利用した．CRL 固有表現データは，毎日新聞 1995 年版 1,174 記事に対して IREX で定義された固有表現が人手でタグ付与されたものである．タグ付けが困難であったことを示す OPTIONAL タグを対象から除外すると，組織名に関するタグは 3676 箇所に付与されている．

3.2　組織名辞書

2 つの組織名辞書（**oracle 辞書**，**日外辞書**）を利用する．各辞書はそれぞれ以下の方法で用意した．

oracle 辞書は，CRL 固有表現データから組織名として認定されている異なり文字列を抽出し，「東京」，「社会」等†の，一般名詞との間で高い曖昧性をもつ文字列を人手で削除することで作成した．oracle 辞書は仮想的な辞書であり，CRL 固有表現データを実験データとして利用する際に，最も理想的な組織名辞書として機能することが期待される．辞書エントリ数は 1481 件である．oracle 辞書は 4 節で述べる，組織名辞書の特性と抽出精度の関係調査の際に主に使用する．

日外辞書は，日外アソシエーツ社が提供している機関名辞書 [21] と，ipadic（version 2.7.0）‡を混合して作成した．日外アソシエーツ社の機関名辞書は，上場企業や教育機関等を含む機関名のリストであり，略称（例えば，「東京工業大学」に対する「東工大」）も含まれる．リストは毎週更新されており，我々の手元にある版（2007 年 5 月頃の版）では，登録文字列の異なりで数えると，約 12 万 2 千件が登録されている．これに，ipadic のうち，品詞が「名詞 – 固有名詞 – 組織」となるエントリを加え，そこから地名を表しやすいエントリを取り除いたものを日外辞書として用いる．最終的なエントリ数は約 12 万 7 千件である．なお，ここでは，{ 市，区，町，村，都，道，府，県 } のいずれかの文字で終わるエントリを地名を表しやすいエントリとして除外した．先の oracle 辞書とはちがい，日外辞書は現実に存在するリソースを組み合わせることで作成されている．日外辞書は 5 節で使用する．

3.3　負例リスト

CRL 固有表現データから名詞列を抽出し，そのうち oracle 辞書のエントリと重複のない文字列，約 15000 件を負例リストとして用意した．本リストは，組織名辞書の特性と抽出精度の関係調査の際に使用する．

4　固有表現辞書の特性と抽出精度の関係

ここまで，調査の環境，設定について述べた．本節では，NER 課題で利用する固有表現辞書の特性と抽出精度の関係を調査する．4.1 節では，調査に先立ち，固有表現辞書の特性を定量的に記述する特性パラメータを導入する．4.2 節で特性パラメータを用いた調査の方法を説明し，4.3 節で調査結果を述べる．

4.1　辞書の特性パラメータ

辞書の特性パラメータとして，先行研究では量の観点からのパラメータを設定している．例えば，落谷 [22] や Shinzato et al. [11] は，エントリ数と抽出精度の関係を評価している．

ただ近年では，1 節でも述べたように，固有表現辞書の自動整備を狙った研究もある [17, 14]．そして，このような自動整備された辞書知識は，一般には，知識の内容面に関して量と質の間でトレードオフ関係をもつこ

---

* http://nlp.cs.nyu.edu/irex/index-j.html

† IREX の定義により，省略形であっても文脈から組織名であると判断できる場合はタグが付与されているため，例のような一般名詞との間で高い曖昧性をもつ文字列も含まれる．

‡ http://sourceforge.jp/projects/ipadic/

とが多い．このことから，特性パラメータを考える際には，（エントリ数等の）量の観点に加え，質の観点を含めるべきであると考えられる．

上記議論を踏まえ，本稿では，照合再現率（matching recall : MR），照合適合率（matching precision : MP）の2つの変数を定義し，この2変数で辞書特性を記述する．MR が辞書エントリの量を捉える変数であり，MP が質を捉える変数である．ここでは，質を評価するために，評価用のタグ付きコーパスの存在を仮定する．昨今の NER 課題は教師あり学習に基づく手法が採用されることが多く，このようなタグ付きコーパスの存在を仮定することは自然である．本研究では CRL 固有表現データを使う．

ある固有表現辞書（組織名辞書）D と，評価用コーパスがある時，コーパス中で組織名と認定されている事例の集合を A，コーパス中で，辞書 D に含まれるいずれかのエントリと文字列照合する事例の集合を B とする．この時，辞書 D の照合再現率（MR），照合適合率（MP）はそれぞれ次式で定義される：

$$\mathrm{MR} = |A \cap B|/|A|,$$
$$\mathrm{MP} = |A \cap B|/|B|.$$

ここで，$|\cdot|$ は集合の要素数を示す．また，状況に応じ，2変数の要約として，MR と MP の調和平均 MF $= (2 \times \mathrm{MR} \times \mathrm{MP})/(\mathrm{MR} + \mathrm{MP})$ を用いる．

照合適合率によって，先行研究では扱っていない質の観点を導入する．また，照合再現率は，エントリ数に比べて，より正確に機能すると考えられる．IREX の定義によると，同じ組織名という固有表現クラスの中にあっても，株式会社等の会社からスポーツチームまで，さまざまな組織が含まれることがわかる．この時，仮に会社名の抽出精度を向上させることが当面の目的であると仮定する．この状況において，スポーツチームのエントリ数を増補させると，辞書のエントリ数は増加するが，このエントリ数の増加は会社名の抽出精度の向上にはほとんど寄与しない．極端ではあるが，この例から，エントリ数と抽出精度の連動性はそれほど高くないと考えられる．一方で，上記の状況では照合再現率は変化せず，また，目的に合致したエントリが増補された時のみ適切に値が変化する．

さらに，辞書エントリを構成する単語について，固有表現辞書の特性を調査したところ，辞書エントリはそれを構成する単語の品詞に関して偏りをもつことがわかった．表 5 は，3 節で述べた 3 つの語彙的資源について，エントリを構成している単語の品詞の出現割合を調べた結果である．エントリを構成する単語の大多数は名詞である．ここでは，ipadic の品詞体系に従い，名詞の第2階層のカテゴリのうち，頻出上位4カテゴリの値を示している．左半分は各語彙的資源の全エントリを対象とした場合の結果であり，右半分は各語彙的資源のうち，構成単語数が2以上のエントリを対象とした場合の結果である．例えば，左上の数値「51.5」は，oracle 辞書の全エントリの中で，51.5% のエントリが，品詞が「名詞 – 一般」である単語を構成要素として含んでいることを表している．

表から，組織名辞書（oracle 辞書と日外辞書）と負例リストとの間を比較した際，「名詞 – 固有名詞」の値に顕著な差があることがわかる．負例リストでは品詞が「名詞 – 固有名詞」である単語は2割ほどしか含まないが，組織名辞書では7割以上のエントリで「名詞 – 固有名詞」である単語を含んでいる．組織名辞書の中に固有名詞が多く含まれることは自明であると思われるかも知れない．しかし，構成単語数が2以上のエントリに限定しても同様の傾向が見られる．つまり，固有名詞ひとつで組織名を表す場合だけでなく，複数単語から構成される組織名においても，その構成要素として固有名詞を含みやすいことがわかる．

表 5　品詞毎の含有率（単位は%）

	全エントリ			構成単語数が2以上		
名詞 - *	oracle	日外	負例	oracle	日外	負例
一般	51.5	67.3	60.4	75.8	83.8	61.8
サ変接続	18.7	23.9	29.3	28.8	29.8	32.7
固有名詞	**76.9**	**73.8**	**18.0**	**72.6**	**76.4**	**22.1**
接尾	29.4	26.3	33.8	45.7	32.8	51.7

以上を踏まえ，組織名辞書の全エントリの中で固有名詞を含むエントリの割合を固有表現含有率（proper noun percentage : PNP）と呼び，特性パラメータとして使用することとする．

4.2　調査方法

利用する固有表現辞書の特性と抽出精度の関係を調べるにあたり，固有表現辞書の特性を，先述した3つの特性パラメータの組 $(\mathrm{MR}, \mathrm{MP}, \mathrm{PNP})$ で記述する．

ある特性パラメータ $(\mathrm{MR}_i, \mathrm{MP}_j, \mathrm{PNP}_k)$ をもつ辞書 $D_{(i,j,k)}$ を用意し，$D_{(i,j,k)}$ の知識を辞書素性として取り込み組織名抽出を実施し，抽出精度を計測する．この手続きを異なる特性パラメータをもつ辞書ごとに行い，固有表現辞書の特性と抽出精度の関係を調査する．

組織名抽出手法は 2.2 節で述べた手法を用い，辞書知識は 2.3 節で述べた単純照合法で取り込む．データには CRL 固有表現データを用い，5分割交差検定の後，F 値（$\beta = 1$）で抽出精度を評価する．具体的な抽出手続きの実装には，YamCha[§] を利用した．また，SVMs のカーネル関数は2次の多項式カーネルを使用した．なお，データ，評価尺度等，組織名抽出の手続きにおける各種設定は以降の調査でも同様である．

複数の，異なる特性パラメータを持つ辞書は，oracle 辞書から作成する．oracle 辞書の特性パラメータは，$(0.947, 0.868, 0.769)$ である．

ある特定の特性パラメータ $(\mathrm{MR}_i, \mathrm{MP}_j, \mathrm{PNP}_k)$ をもつ辞書 $D_{(i,j,k)}$ の作成にあたり，まず，oracle 辞書の複製 D_c を作成する．次に，D_c に対し，以下で定義される，D_c の特性パラメータを変化させる2つの操作を繰り返し行うことで目的の $D_{(i,j,k)}$ を得る．ここで，照合再現率と照合適合率ともに，パラメータ値を下げることは比較的容易であるが，逆に上げることは困難である．そのため，以下の操作は，基本的には，各パラメータの数値を下げる操作として設計されている．

操作 R　D_c からエントリをひとつ選び，それを D_c から削除することで D_c を更新する（操作 P で追加されるエントリは選ばない）．

操作 P　負例リストからエントリをひとつ選び，それを D_c に追加することで D_c を更新する．

[§] http://www.chasen.org/~taku/software/yamcha/

表 6　特性パラメータと抽出精度の関係 (PNP=0.7 の場合)

		MR				
		.5	.6	.7	.8	.9
MP	.5	84.08	85.41_a	86.04	87.48	87.77_b
	.6	84.69_a	86.35	87.36_b	88.22	89.08
	.7	85.49	87.27_b	87.87	89.15_c	90.05
	.8	86.74	88.52	89.24_c	90.85	92.35
	.9	-	-	-	-	-

(値は $F_{\beta=1}$ の平均値)

表 7　特性パラメータと抽出精度の関係 (PNP=0.8 の場合)

		MR				
		.5	.6	.7	.8	.9
MP	.5	83.49	83.92_a	85.46	86.46	87.37_b
	.6	83.87_a	84.45	86.01_b	87.25	88.49
	.7	84.40	85.02_b	86.82	87.91_c	89.70
	.8	85.42	86.38	88.12_c	88.93	90.76
	.9	-	-	-	-	-

(値は $F_{\beta=1}$ の平均値)

操作 R によって照合再現率を制御し，操作 P によって照合適合率を制御する．また，上記操作の中で，操作 R においてエントリをひとつ選ぶ際，現在の PNP 値 (PNP_c) と目標の PNP 値 (PNP_g) を比べる．$PNP_c \geq PNP_g$ の関係であれば，固有名詞を含むエントリを選び，そうでなければ固有名詞を含まないエントリを選ぶ．同様に，操作 P において負例リストからエントリをひとつ選ぶ際，$PNP_c \geq PNP_g$ の関係であれば，固有名詞を含まないエントリを選び，そうでなければ固有名詞を含むエントリを選ぶ．この操作によって，照合再現率，照合適合率と同時に固有表現含有率を制御する．

以上の手続きに従い，本研究では，以下に示す各パラメータ値を組合せた計 40 種類の特性パラメータをもつ組織名辞書を作成した．

$$\begin{aligned} \mathrm{MR} &= \{0.5, 0.6, 0.7, 0.8, 0.9\} \\ \mathrm{MP} &= \{0.5, 0.6, 0.7, 0.8\} \\ \mathrm{PNP} &= \{0.7, 0.8\} \end{aligned}$$

操作 R，操作 P では，エントリを選ぶ際に任意性がある．そこで，同じ特性パラメータをもつ辞書を 5 つ作成することにした．結果として，40 種類 × 5 つの計 200 個の組織名辞書を作成し，調査に用いた．

4.3　調査結果

調査結果を表 6，表 7 に示す．各セルは，特定の特性パラメータをもつ辞書を利用して組織名を抽出することを辞書の数（5 回）だけ繰り返し，抽出精度の平均値を求めた値を示す．表 6 は PNP=0.7 である特性パラメータ群に関する結果であり，表 7 は PNP=0.8 である特性パラメータ群に関する結果である．視認性を高めるため，対角に位置するセルの値に下線を引いた．

表から以下のことがわかる．以下の観察結果は表 6，表 7 のいづれの表からも読み取れる．

観察 1：MR を固定して各 MP における抽出精度を比較した場合，各セルの値は異なる．MP を固定して各 MR における抽出精度を比較した場合も同様である．例えば，表 6 で MR=0.5 と固定した際の各 MP における抽出精度は，84.08，84.69，85.49，86.74 であり，それぞれ異なる．

観察 2：MR と抽出精度，MP と抽出精度の間には共に正の相関関係がある．つまり各セル値は，表の左側よりも右側の方が高く，上側よりも下側の方が高い．

観察 3：同じ（あるいはほぼ同じ）MF 値をもつセルの値を比べることを考える．MF が低い場合（表の左上に近い部分）は，MP よりも MR の値が高いセルの方が抽出精度が高い．逆に，MF が高い場合（表の右下に近い部分）は，MR よりも MP の値が高いセルの方が抽出精度が高い．例えば，添え字記号 a，b，c が含まれるセルに注目する．ここで，同じ記号をもつセルの MF 値は（ほぼ）等しい[¶]．同じ記号をもつセル値を比べた場合，記号 a，b では MP よりも MR の値が高いセルの方が抽出精度が高く，記号 c では逆に MR よりも MP の値が高いセルの方が抽出精度が高い．

4.4　まとめ

以上の観察から得られた知見をもとに固有表現辞書の自動整備の進め方について考える．まず，観察 1 から，辞書の特性と抽出精度の間の関係を検討するうえで，辞書を定量的に評価する際は，質と量を捉えるパラメータ (MR と MP) が少なくとも必要であることがわかる．観察 2 の相関関係から，MR と MP の値を向上させるよう辞書整備を進めることによって，結果として，最終的な固有表現の抽出精度も向上することが期待できる．さらに，観察 2 と比べて，観察 3 はより細かな情報を提供している．観察 3 から，辞書整備の段階において MR と MP の 2 つのパラメータに注目する際，MF 値が低い時点では MR を重視して整備を進めた方が MP を重視して整備を進めるよりも固有表現抽出の精度向上が計れると期待できる．また，表 6，表 7 中の添え字記号 b，c をもつセル周辺の値からの見積もりによると，MF=0.64 ～0.75 前後で重視すべきパラメータが MR から MP に移行することが示唆される．

5　正規表現に基づく固有表現辞書の利用法

本節では，日外辞書を題材にして，組織名抽出課題における辞書知識の効果的な利用法を提案，検証する．日外辞書の特性パラメータは (0.543, 0.648, 0.738) であり，MF=0.591 である．前節の結果から，この数値は MR の向上を重視した戦略がより有効な範囲に属する．そこで，MP の低下を抑えながら MR を向上させることを考える．ただし，ここでは，辞書エントリを増補するのではなく，2.3 節で述べた単純照合法を改良することで，MR の向上を実現する．

5.1　基本的な考え方

単純照合法では，辞書エントリが入力単語系列内に完全に包含されている時に限り文字列照合が成立し，エントリ情報が素性として反映される．提案手法では，辞書エントリが入力単語系列内に完全に包含されていない状況でも照合が成立するように改良する．具体的には，辞

[¶] 添え字記号 b をもつセルでは，MR=0.9，MP=0.5 の時，MF=0.643，MR=0.6，MP=0.7 (逆に，MR=0.7，MP=0.6) の時，MF=0.646 でありほぼ等しい．添え字 a および c に関しては同じ添え字をもつ各セルの MF は等しい．

表 8　固有表現の構成例

固有表現		固有表現クラス
インスタンス指定部	クラス指定部	
東京工業	大学	学校名
情報処理	学会	協会名
ＪＦケネディ	空港	空港名
仏	教	宗教名
ニュートンの	法則	理論名

書エントリの集合から正規表現規則を獲得し，この規則を照合対象とすることで MR の値を改善する．

例えば，「○○大学」，「△△大学」，「□□大学」の３つのエントリをもつ組織名辞書があると仮定する．また，この辞書エントリ集合から「**(.+) 大学**」という正規表現が獲得できたと仮定する．「(.+)」は，１文字以上の任意の文字列と照合する正規表現記号である．この時，入力系列「◇◇大学は...」内の名詞列「◇◇大学」は，単純照合法では辞書中のどのエントリとも文字列照合が成立せず，その結果，辞書素性は反映されない．一方で，「◇◇大学」は正規表現「**(.+) 大学**」と照合可能であるので，「◇◇大学」に対して何らかの辞書知識を素性として反映させることができ，また，MR の向上が見込める．

5.2　正規表現規則の獲得

まず，規則獲得の際に手がかりとなる情報について述べる．その後，具体的な規則獲得の手続きを説明する．

5.2.1　手がかりとなる情報について

組織名辞書のエントリ集合の中で高い頻度で出現する単語系列は，組織名を表す特徴的な文字列であると考えられる．そこで，高い頻度で出現する単語系列はそのまま残し，それ以外の箇所を正規表現記号「(.+)」で置き換えることを考える．

次に，固有表現の構成的な特徴について検討する．さまざまな種類の固有表現を観察することで，固有表現には次のような構成的な特徴があることがわかる．

固有表現の構成的な特徴　固有表現の多くは，固有表現クラスを指定する部分と，そのクラスにおける個々のインスタンスを指定する部分の２つの要素から構成される．また，固有表現クラスを指定する部分は固有表現の末尾に位置する．

具体例を表 8 に示す．表の固有表現クラス欄の値は，Sekine et al. [10] で定義された拡張固有表現階層でのクラスである．さまざまな種類の固有表現クラスが上記特徴を有することを示すために，[10] の定義を用いた．なお，(文脈にもよるが) IREX ではこのうち，上３件の例が組織名に該当する．

獲得したい正規表現規則は，固有表現のインスタンスではなくクラスを指定する情報を保持すべきである．ここから，辞書エントリの末尾側の文字列が頻出単語系列ではない場合，そのような実体は，正規表現規則として相応しくないと考えられる．

5.2.2　獲得の手続き

以上を踏まえた正規表現規則の獲得の手続きを以下に示す．また，図 1 に具体的なデータ例を示す．

ステップ１：フィルタリング．　先述した固有表現の構成的な特徴は，構成単語数がある程度長い事例ほど有

```
#　（辞書エントリ集合．スラッシュ記号は単語境界を示す．）
…
日本/航空機/エンジン/協会
AGF
日本/通訳/学会
栄町/商店/街/振興/組合
銘林
延岡/市/商店/街/連合/会
…
 ↓ステップ１：フィルタリング
…
日本/航空機/エンジン/協会
日本/通訳/学会
栄町/商店/街/振興/組合
延岡/市/商店/街/連合/会
…
 ↓ステップ２：前処理
…
4_START 3_日本 2_航空機 1_エンジン 0_協会
3_START 2_日本 1_通訳 0_学会
5_START 4_栄町 3_商店 2_街 1_振興 0_組合
6_START 5_延岡 4_市 3_商店 2_街 1_連合 0_会
…
 ↓ステップ３：頻出パタン・マイニング
#　（行先頭の数値は各パタンの頻度）
…
458 4_START 3_日本 0_協会
454 3_START 0_学会
451 2_街 0_組合
450 6_START 1_連合 0_会
…
 ↓ステップ４：正規表現規則の生成
…
 ^日本 (.+) 協会
 ^(.+) 学会
 街 (.+) 組合
 ^(.+) 連合会
…
```

図 1　規則獲得の処理の流れ

しやすいと考えられる．例えば，「銘林」や「AGF」といった事例は，クラス指定部とインスタンス指定部に分割できそうにない．そこで，日外辞書から構成単語数が２以下の辞書エントリは除外して，残ったエントリ集合に対し，以下のステップの処理を施す．

ステップ２：前処理．　辞書エントリを系列，エントリの構成単語をアイテムと見立てて頻出パタン・マイニング [1, 8] を適用することで，辞書エントリ集合から頻出単語系列を抽出する．そのために，各エントリに対して次の前処理を施す．

2-1 先頭に疑似単語「START」を加える．
2-2 単語境界ごとにスペースを加え，アイテムを明確化する．
2-3 末尾側を０番とし，各単語に位置インデックスを付与する．

ステップ３：頻出パタン・マイニング．　PrefixSpan [8] アルゴリズムを用いて，頻出単語系列を抽出する．抽出結果のうち，０番の位置インデックスをもつアイテムを含まない系列を削除する．

ステップ４：正規表現規則の生成．　頻出単語系列に対し，次の処理を施すことで正規表現規則を生成する．

表 9　実験結果

	再現率	適合率	$F_{\beta=1}$	p 値
Non	74.32	87.51	80.38	-
Simple	78.37	89.08	83.38	-
Simple + Regex				
$\xi = 5$	79.33	88.77	83.78	0.301
10	79.57	**89.23**	84.12	0.072
15	**79.60**	89.21	**84.13**	0.017
20	79.46	88.97	83.95	0.033

4-1 位置インデックスを参照し，系列内の単語が元エントリにおいて連続していない場合は，当該の単語間に「(.+)」を挿入する．
4-2 スペース，位置インデックスを削除する．
4-3 疑似単語「START」を，文字列先頭のみと照合する正規表現記号「^」に変更する．

5.3　正規表現規則の適用

正規表現規則を次のように使用して素性情報を獲得する．まず，入力単語系列から構成単語数が 2 以上の名詞列を抽出する．そして，抽出された各名詞列について以下の操作を行い，素性情報を獲得する．以降，この素性情報獲得法を**正規表現照合法（Regex）**と呼ぶ．

正規表現照合法

- 名詞列と各正規表現規則との照合を試みる．名詞列が，ある規則と照合する場合，名詞列の照合箇所に「I」タグを与える．照合しない残りの他の箇所には「O」タグを与える．ただし，名詞列と照合する規則が複数ある場合は最も単語を多く含む規則を選ぶ．また，名詞列に含まれる単語内の一部の文字列と照合する場合は，未照合として扱う．

他の素性と同様，正規表現照合法で得られた素性情報も，対象単語および前後 2 単語の計 5 単語の値を使用する．

節頭でも述べたように，ここでの方針は MR を上げることである．MR を上げることを考えた場合，正規表現照合法を単独使用するのではなく，単純照合法と正規表現照合法を併用した方がよい．そこで，単純照合法と正規表現照合法を併用して素性情報を獲得する手法を**提案手法**と呼び，評価実験では単純照合法と提案手法を比較する．

5.4　評価実験

評価実験を通して提案手法の有効性を検証する．

5.4.1　実験条件

組織名辞書には日外辞書，データには CRL 固有表現データを用い，5 分割交差検定の後，F 値（$\beta = 1$）で抽出精度を評価する．PrefixSpan では最小サポート値（ξ）以上の回数現れる系列を頻出パタンとみなして抽出する．実験では，$\xi = \{5, 10, 15, 20\}$ を試行した．

5.4.2　実験結果

実験結果を表 9 に示す．表の上段から，“Non” は辞書知識を利用しない場合の結果，“Simple” は単純照合法によって素性情報を獲得した場合の結果である．“Simple+Regex” は，提案手法（単純照合法と正規表現照合法の併用）で素性情報を獲得した場合の結果である．

表 10　実験結果（8 種類の固有表現クラスを考慮）

	再現率	適合率	$F_{\beta=1}$
Simple + Regex			
$\xi = 5$	80.82	**88.24**	84.37
10	80.90	88.12	84.36
15	81.18	88.15	84.52
20	**81.31**	88.12	**84.58**
山田 (SR 法)[15]	-	-	84.53
中野ら [19]			
意味素性なし	-	-	84.30
意味素性あり	-	-	84.69
oracle	93.42	96.62	94.99

表 11　構成単語数の違いによる性能の比較

構成単語数	$= 1$	$= 2$	$= 3$	≥ 4	all
Non	87.87	76.96	66.09	56.21	80.38
Simple	89.35	81.45	72.32	63.99	83.38
Simple + Regex ($\xi = 15$)					
	89.73	81.55	74.21	67.35	84.13

（値は $F_{\beta=1}$）

提案手法の中の最良値を太字で記す．表から，提案手法の有効性が確認できる．特に，再現率の向上が $F_{\beta=1}$ の値の向上に貢献していることがわかる．“Simple” を基準にして，各最小サポート値ごとの性能を単語単位の MAPSSWE 検定（Matched Pairs Sentence-Segment Word Error Test）[4, 3] で検定した．表の最右列に p 値を示す．$\xi = 15$ および 20 においては，有意水準 5 % で有意な差があることが確認できた．最適な最小サポート値の自動選択は今後の課題である．

表 11 に構成単語数の違いによる性能の比較の結果を示す．表から，構成単語数が多くなるに従って，提案手法の有効性が顕著になっていることがわかる．

次の例は，提案手法で素性情報を獲得することによって新たに抽出できるようになった組織名の例である．

- 香港インド商工会議所
- 日本かるた院本院
- 軍事休戦委員会
- 日本医療機器関係団体協議会

続いて，提案手法の性能を先行研究と比較する．ここまで，組織名に限定して議論を進めたが，対等な比較を行うために，IREX で定義された 8 種類の固有表現クラスを考慮して抽出精度を評価した．この場合は 17 種類のラベル分類問題となるが，それ以外の設定は先と同様である．実験結果を表 10 に示す．8 種類の固有表現クラスを考慮しているが，表の数値はあくまで組織名抽出の精度である．提案手法の性能が表 9 に比べて高くなっているが，これは，組織名以外の固有表現クラスの情報が組織名抽出に有効な素性情報を提供したことによると考えられる．表から，提案手法は先行研究とほぼ同等な性能を達成できることがわかる．

提案手法では，表 10 で挙げた先行研究 [15, 19] で有

効とされる文節情報を考慮していない．辞書知識は固有表現の内部的な素性情報を提供する．一方で文節情報は固有表現の外部（文脈）的な素性情報を提供し，両者の性質は互いに異なると考えられる．このことから，辞書知識と文節情報を両方考慮したモデルを構築すれば，より高い性能が達成できると考えられる．

表 10 の “oracle” は，単純照合法で oracle 辞書の知識を取り込んだ場合の結果である．この数値は，山田らの手法 [16] に対して辞書知識を取り込むような組織名抽出手法における性能の上限値を与える．ここから，CRFs に基づく手法等の複雑な抽出モデルを構築せず，辞書整備を進めるだけでも，高い性能を達成できる可能性があることがわかる．

6 関連研究

利用する固有表現辞書の特性と抽出精度の関係に注目した研究として，落谷 [22] や Shinzato et al. [11] は，抽出実験の一部として，辞書エントリ数と抽出精度の関係を評価している．本稿では，タグ付きコーパスの存在を仮定したうえで，特性パラメータを定義し，彼らの研究報告を発展させた．

塩入ら [17] や Watanabe et al. [14] 等は辞書エントリを獲得する研究を行っている．これらは，固有表現に関する語彙的知識を豊かにする方法であるという点で，辞書エントリ集合から正規表現規則を自動獲得することと類似している．しかし，彼らが Wikipedia や新聞記事等の言語資源を使用する一方で，正規表現規則の自動獲得では辞書知識を使用している点が異なる．また，我々の手法では明示的な辞書エントリは獲得されない点が異なる．

7 おわりに

本稿では，固有表現クラスとして組織名を選び，固有表現辞書を利用して固有表現抽出課題を実施する状況において，利用する固有表現辞書の特性と抽出精度の関係，効果的な固有表現辞書の利用法について述べた．

辞書特性を定量的に記述する特性パラメータを導入し，特性パラメータと抽出精度の関係性をある程度明らかにした．また，辞書エントリ集合から正規表現規則を自動獲得する手法を提案し，正規表現規則が固有表現の抽出精度の向上に有効に働くことを示した．

本稿の内容のうち，特性パラメータは固有表現クラスに依存しない概念である．また，5.2 節で述べた「固有表現の構成的な特徴」をもつ固有表現クラスであれば，正規表現規則の獲得手法を適用できる．上記以外の内容については，固有表現クラスに依存している可能性があり，今後，組織名以外のクラスに対しても同様の調査を実施したい．

謝辞

本研究は，文部科学省科学技術振興調整費「戦略的研究拠点育成プログラム」の支援の下に実施した．

参考文献

[1] R. Agrawal and R. Srikant. Mining sequential patterns. In Philip S. Yu and Arbee S. P. Chen, editors, *Eleventh International Conference on Data Engineering*, pp. 3–14. IEEE Computer Society Press, 1995.

[2] A. Borthwick, J. Sterling, E. Agichtein, and R. Grishman. Exploiting diverse knowledge sources via maximum entropy in named entity recognition. In *Proc. of the 6th Workshop on Very Large Corpora*, pp. 152–160, 1998.

[3] L. Gillick and S. Cox. Some statistical issues in the comparison of speech recognition algorithms. In *Proc. of ICASSP*, pp. 532–535, 1989.

[4] NIST Speech Grounp Significance Test Home. `http://www.nist.gov/speech/tests/sigtests/mapsswe.htm`.

[5] Y. Matsumoto, A. Kitauchi, T. Yamashita, Y. Hirano, H. Matsuda, and M. Asahara. *Japanese Morphological Analyzer ChaSen Users Manual version 2.0.* Technical Report NAIST-IS-TR990123, Nara Institute of Science and Technology Technical Report, 1999.

[6] A. McCallum and W. Li. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In *Proc. of the 7th Conference on Computational Natural Language Learning*, 2003.

[7] MUC6. The 6th message understanding conference, 1995.

[8] J. Pei, J. Han, B. Mortazavi-Asi, H. Pinto, Q. Chen, U. Dayal, and M. C. Hsu. PrefixSpan: mining sequential patterns efficiently by prefix projected pattern growth. Proc. of 1st. Conference of Data Enginnering (ICDE2001), pp. 215–226, 2001.

[9] S. Sekine and H. Isahara. IREX project overview. In *Proc. of the IREX Workshop*, 1999.

[10] S. Sekine and C. Nobata. Definition, dictionaries and tagger for extended named entity hierarchy. In *Proc. of LREC*, pp. 1977–1980, 2004.

[11] K. Shinzato, S. Sekine, N. Yoshinaga, and K. Torisawa. Constructing dictionaries for named entity recognition on specific domains from the web. In *Proc. of the Web Content Mining with Human Language Technologies workshop on the fifth International Semantic Web*, 2006.

[12] J. Suzuki, E. McDermott, and H. Isozaki. Training conditional random fields with multivariate evaluation measures. In *Proc. of the 44th Annual Meeting of the Association for Computational Linguistics*, pp. 217–224, 2006.

[13] V. N. Vapnik. *The Nature of Statistical Learning Theory.* Springer, 1995.

[14] Y. Watanabe, M. Asahara, and Y. Matsumoto. A graph-based approach to named entity categorization in wikipedia using conditional random fields. In *Proc. of the joint Conference on EMNLP-CoNLL*, pp. 649–657, 2004.

[15] 山田寛康. Shift-reduce 法に基づく日本語固有表現抽出. 情報処理学会自然言語処理研究会 (NL-179-3), pp. 13–18, 2007.

[16] 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, 2004.

[17] 塩入寛之, 関根聡, 梅村恭司. 拡張固有表現獲得の精度向上. 情報処理学会自然言語処理研究会 (NL-180-12), pp. 67–72, 2007.

[18] 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, Vol. 42, No. 6, pp. 1580–1591, 2001.

[19] 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, 2004.

[20] 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, 2004.

[21] 日外アソシエーツ. DCS-機関名辞書.

[22] 落谷亮. 組織名抽出のための知識収集. 言語処理学会第 5 回年次大会, 1999.