固有表現の経年変化と頑健な抽出

Analysis and Robust Extraction of Named Entity

遠藤 翔子 *,　　土屋 雅稔 †,　　中川 聖一 *

豊橋技術科学大学

* 情報工学系 / † 情報メディア基盤センター

1. はじめに

人名・組織名といった語句を同定する固有表現抽出タスクは，情報検索や情報抽出の基礎技術としてのみならず，自然言語処理における構文解析や意味解析などに大きな影響を及ぼすため，重要な問題である[1),2)]．

固有表現を抽出する方法は，大きく 2 つに分類することができる．第 1 は人手で作成した規則に基づく手法[3)] であり，第 2 は統計的機械学習に基づく手法である．統計的機械学習のアルゴリズムとしては，Maximum Entropy[4)]，決定リスト[5)–7)]，Support Vector Machine[8),9)] などの様々な手法が適用され，人手で作成した規則に基づく手法と比較して，カバー率と精度の両方の面で優れていることが示されている．

統計的機械学習に基づく手法を採る場合には，訓練データとして，対象となる固有表現に対して十分な量の固有表現タグ付きコーパスが必要である．しかし，固有表現は非常に種類数が多く，かつ，新たな固有表現が生まれ続けているため，全ての固有表現に対して十分な量の訓練データを用意することは事実上不可能である．さらに，訓練データのテキストが作成された日時と，テストデータのテキストが作成された日時が大きく異なると，それぞれのテキストで用いられている固有表現も大きく異なり，固有表現抽出器の性能が低下する恐れがある．英語固有表現抽出においては，訓練データのテキストの作成日時とテストデータのテキストの作成日時が大きく異なると，固有表現抽出器の性能に悪影響が現れることが報告されている[10)]．

本稿では，日本語固有表現の種類数などの経年変化について，1995 年 ∼1998 年および 2005 年の毎日新聞記事を対象として調査した結果を述べる．また，その経年変化に対して頑健な固有表現抽出手法について述べる．

2. 固有表現の出現傾向の経年変化

2.1 調査対象テキストの選定

筆者らの知る限り，公開されている唯一の日本語の固有表現タグ付きコーパスは，IREX ワークショップ実行委員会によって公開されている訓練用のコーパスである (以後，このコーパスを **IREX コーパス**と呼ぶ)[2)]．IREX コーパスは，1995 年 1 月 1 日から 1 月 10 日までの間に発刊された 1,174 件の毎日新聞記事からなり，その記事中の 18,677 個所の固有表現がタグ付けされている．

固有表現の出現傾向の経年変化を調査するには，作成日時以外の要因ができるだけ類似したテキストを用いるべきである．すなわち，IREX コーパスを調査対象テキストの一部とする場合は，IREX コーパスとは異なる日時に発刊された毎日新聞記事を対象とすることが適切と考えられる．そこで本研究では，1996 年，1997 年，1998 年および 2005 年の 6 月および 10 月の平日から，任意の 1 日を調査対象日として選択した．

調査対象日に掲載された全ての記事を人手で固有表現タグ付けすることは，作業時間の面で困難があるため，約 30%の記事のみを固有表現タグ付けの対象とする．新聞は社会面・スポーツ面・家庭面などの様々な紙面によって構成されており，その紙面によって，固有表現の出現傾向が異なる可能性がある．そこで，1991 年 ∼1999 年の毎日新聞の紙面毎の文字数を調査し，その比率と一致するように記事の選択を行った．

2.2 固有表現タグ付け作業

調査対象テキストとして選定した毎日新聞記事に対して，IREX ワークショップの定義に従い，人手により固有表現タグを付与した．人手により固有表現タグを付与する際，以下のような固有表現のあいまい性が問題となった．

組織名と地名

(1) ソ連 が崩壊し、ユーゴスラビアやチェコスロバキアも分解した。

(2) 東京都新宿区の 新宿署新宿駅西口交番 に、福生市熊川、無職、遠藤孝司容疑者が、「自宅で知人を殺した」と自首。

例文 (1) の下線部「ソ連」は，地名と組織名の 2 通りの判定が考えられる．本研究では，国名は地名とすることにした．例文 (2) の下線部「新宿署新宿駅西口交番」も同様に，地名と組織名の 2 通りの判定がが考えられる．この例文では，例文末尾の「自首」と言う表現に注目し，組織名と判定した．

組織名と固有物名

(1) 7 月 20 日付 毎日新聞 朝刊は…というワシントン・ポストの記事を紹介。

(2) 7 月 12 日の 朝日新聞 の記事は、…血友病患者が死亡したことを伝え、…

(3) 翌日の各紙は「日本は患者ゼロ」(毎日新聞) …などの記事を一斉に掲載。

表 1 調査対象テキスト

掲載日	1995 年	1996 年		1997 年		1998 年		2005 年	
	1/1∼1/10	6/5	10/15	6/10	10/7	6/8	10/21	6/23	10/12
記事数	1174	120	133	106	117	96	126	90	99
文字数	407881	60790	53625	46653	50362	51006	67744	49038	44344
固有表現種類数	6979	1446	1656	1276	1350	1190	1226	1230	1113
固有表現頻度	18677	2519	2652	2145	2403	2126	2052	1902	2007
1 文字あたり固有表現種類数	0.0171	0.0238	0.0309	0.0274	0.0268	0.0233	0.0181	0.0251	0.0251
1 文字あたり固有表現頻度	0.0458	0.0414	0.0495	0.0460	0.0477	0.0417	0.0303	0.0388	0.0453

表 2 掲載年が異なる記事中での固有表現出現率 (種類数)

掲載日	比較対象テキストの掲載年						
	1993 年	1994 年	1995 年	1996 年	1997 年	1998 年	1999 年
1995 年 1/1∼1/10	5106 (73.2%)	5484 (78.6%)	—	5210 (74.4%)	4537 (65.0%)	4491 (64.4%)	4418 (63.3%)
1996 年 6/5, 10/15	1953 (67.2%)	2082 (71.7%)	2098 (72.2%)	—	2244 (77.3%)	2209 (76.0%)	2182 (75.1%)
1997 年 6/10, 10/7	1730 (71.2%)	1783 (73.4%)	1808 (74.4%)	1909 (78.6%)	—	1963 (80.8%)	1908 (78.6%)
1998 年 6/8, 10/21	1639 (72.5%)	1685 (74.6%)	1723 (76.2%)	1800 (79.7%)	1870 (82.7%)	—	1896 (84.0%)
2005 年 6/23, 10/12	1355 (62.3%)	1394 (64.1%)	1452 (66.8%)	1494 (68.7%)	1547 (71.2%)	1585 (72.9%)	1605 (73.8%)

表 3 固有表現種類数の変化

	1995 年	1995 年 ∼1996 年	1995 年 ∼1997 年	1995 年 ∼1998 年	1995 年 ∼2005 年
固有表現種類数	6979(1.00)	9302(1.33)	11071(1.59)	12601(1.81)	14104(2.02)

(括弧内は，1995 年を 1 とする増加率)

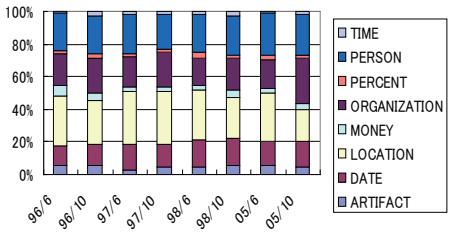

図 1 固有表現の種別毎の内訳比率

例文 (1) の下線部「毎日新聞」は，その日付の毎日新聞という物理的実体を指しているため，固有物名と考えられる．しかし，例文 (2) の下線部「朝日新聞」や例文 (3) の下線部「毎日新聞」のように，固有物名と組織名の 2 通りの判定が考えられる場合がある．本研究では，文脈からは判定が困難である場合は，組織名と判定することにした．

2.3 分 析 結 果

調査対象として選定したテキストの掲載日，記事数，文字数および固有表現の頻度と種類数を**表 1** に示す．表 1 において掲載日が 1995 年 1 月 1 日 ∼10 日となっているテキストは IREX コーパスであり，それ以外のテキストが本研究で固有表現タグ付けを行ったテキストである．表 1 より，固有表現の頻度および種類数は，調査対象テキストの掲載日によらずほぼ一定であることが分かる．次に，調査対象テキストに含まれる固有表現の種別毎の内訳比率を**図 1** に示す．図より，固有表現の種別毎の内訳比率も，調査対象テキストの掲載日によらずほぼ一定であることが分かる．

記事が執筆される時期が近いと，類似した話題 (人名や地名など) に関する記事が多くなることが予想される．そのため，ある日時に掲載された記事中で用いられた固有表現が，異なる年に掲載された記事中に出現する割合を調査した (**表 2**)．このような調査を完全に行うには，比較対象となる年における全記事に固有表現タグ付けされている必要があるが，現時点では，そのようなデータは利用できない．そのため，表 2 では，調査対象テキスト (例えば 1996 年 6 月 5 日および 10 月 15 日) に含まれる固有表現と同じ文字列が，比較対象となる年の新聞記事 (例えば 1995 年の毎日新聞全記事) に含まれている割合を求めた．表 2 より，ある年に用いられる固有表現の 70%から 80%は，既に前年の記事に出現していることが分かる．また逆に，ある年に用いられた固有表現の 70%から 80%だけが，翌年の記事に出現している．全ての固有表現が同等かつ独立であり，ある年に固有表現に出現した固有表現が翌年にも出現する確率が 80%であると仮定すると，ある年に出現した固有表現が 2 年後に出現する確率は 64%となる．しかし，表 2 によると，調査対象テキストの 2 年後 (または 2 年前) の比較対象テキストにおける出現率は，殆んどの場合で 70%を越えている．よって，掲載年に関わらず出現し易い固有表現と，ある掲載年のみに出現し易い固有表現とがあることが分かる．

掲載年の異なる調査対象テキストをひとまとまりのテキストとして考えた場合の固有表現種類数を**表 3** に示す．表 3 より，新規固有表現が追加されて，固有表

形態素素性 MF		類似形態素素性 SF		文字種素性	チャンク
表層形	品詞	表層形	品詞	CF	ラベル
今日	名詞–副詞可能	今日	名詞–副詞可能	$\langle 1,0,0,0,0,0\rangle$	O
の	助詞–連体化	の	助詞–連体化	$\langle 0,1,0,0,0,0\rangle$	O
石狩	名詞–固有名詞	**関東**	名詞–固有名詞	$\langle 1,0,0,0,0,0\rangle$	B-LOCATION
平野	名詞–一般	平野	名詞–一般	$\langle 1,0,0,0,0,0\rangle$	I-LOCATION
の	助詞–連体化	の	助詞–連体化	$\langle 0,1,0,0,0,0\rangle$	O
天気	名詞–一般	天気	名詞–一般	$\langle 1,0,0,0,0,0\rangle$	O
は	助詞–係助詞	は	助詞–係助詞	$\langle 0,1,0,0,0,0\rangle$	O
晴れ	名詞–一般	晴れ	名詞–一般	$\langle 1,1,0,0,0,0\rangle$	O

図 2 学習データの例

現の種類が増え続けていることが分かる．

3. 出現傾向の変化に対して頑健な固有表現抽出

増加し続けている固有表現を網羅したタグ付きコーパスを用意することは，非現実的である．筆者らは，固有表現タグは付与されていないものの，大量に利用可能なタグなしコーパス (例えば，新聞記事データ) を併用して，タグ付きコーパスに頻出しない (または，出現しない) 語を含む固有表現を頑健に抽出できる固有表現抽出法を提案している[11]．提案手法は 2 段階からなる．最初に，タグ付きコーパスに頻出しない語に対して，タグなしコーパスから求めた周辺ベクトルに基づいて，固有表現タグ付きコーパスに頻出し，かつ，その前後の文脈の出現が良く類似している語を対応付ける．次に，元々の語と，新たに対応付けた類似語の両方を素性として，従来からの機械学習手法を適用する．例えば，**図 2** の左端列のような文があり，この文に含まれる「石狩」という形態素が，タグ付きコーパスには頻出しないとする．この形態素「石狩」に対して，良く類似していると同時に，タグ付きコーパスに頻出する形態素「関東」を対応付ける．そして，元々の形態素「石狩」と，対応付けられた形態素「関東」の双方を素性として用いて機械学習を行う．以下では，提案手法について説明し，提案手法が固有表現の出現傾向の変化に対して頑健であることを示す．

3.1 類似形態素の対応付け

ある形態素に対して，その前後の文脈の出現が最も類似している形態素を求める方法を，以下に述べる．

ある形態素 m の周辺ベクトル V_m は，あらゆる可能な unigram, bigram を次元とし，その unigram, bigram が形態素 m の直前直後に出現した頻度を各次元の値とするベクトルである．形式的には，次式によって定義される．

$$V_m = \begin{pmatrix} f(m,m_0), & \cdots & f(m,m_N), \\ f(m,m_0,m_0), & \cdots & f(m,m_N,m_N), \\ f(m_0,m), & \cdots & f(m_N,m), \\ f(m_0,m_0,m), & \cdots & f(m_N,m_N,m) \end{pmatrix},$$

ここで，$M \equiv \{m_0, m_1, \ldots, m_N\}$ は，タグなしコーパスに出現する全ての形態素からなる集合である．また，$f(m_i, m_j)$ は，形態素 m_i と形態素 m_j がタグなしコーパスに連続して出現した頻度であり，$f(m_i, m_j, m_k)$ は，形態素 m_i, m_j, m_k がタグなしコーパスに連続して出現した頻度である．

タグ付きコーパスに頻出する形態素からなる集合を M_F とする．この時，ある非頻出形態素 $m_u \in M \cap \overline{M_F}$ に対して，周辺ベクトルの観点から最も類似した頻出形態素 $\hat{m_u}$ は，以下の式を解くことによって得られる．

$$\hat{m_u} = \underset{m \in M_F}{\operatorname{argmax}}\, sim(V_{m_u}, V_m), \qquad (1)$$

ベクトル間の類似度を求める関数 sim としては，様々なものが利用可能であるが，本稿では cosine 類似度を用いる．

3.2 素　　性

本稿では，i 番目の形態素 m_i に対する素性 F_i を，形態素素性 $MF(m_i)$，類似形態素素性 $SF(m_i)$，文字種素性 $CF(m_i)$ の 3 つ組として定義する．

$$F_i = \langle\, MF(m_i),\ SF(m_i),\ CF(m_i)\,\rangle$$

形態素素性 $MF(m_i)$ とは，その形態素 m_i の表層形と品詞の組である．類似形態素素性 $SF(m_i)$ は，形態素 m_i に対して最も類似した頻出形態素の形態素素性であり，次式のように定義される．

$$SF(m_i) = \begin{cases} MF(\hat{m_i}) & \text{if } m_i \in M \cap \overline{M_F} \\ MF(m_i) & \text{otherwise} \end{cases}, \qquad (2)$$

$\hat{m_i}$ は，形態素 m_i に対して周辺ベクトルの観点で比較して最も良く似ていると同時に頻出する形態素であり，式 (1) によって求められる．文字種素性 $CF(m_i)$ は，6 個の 2 値のフラグからなる．フラグはそれぞれ，形態素 m_i の表層形が，漢字・平仮名・片仮名・アルファベット・数字・その他の文字を含むか否かを表す．

i 番目の形態素 m_i に対するチャンクラベル c_i を決定するときには，下記のように周囲の 5 つの形態素に対する素性 $F_{i-2}, F_{i-1}, F_i, F_{i+1}, F_{i+2}$ と，先行する 2 つのチャンクラベル c_{i-2}, c_{i-1} を参照する．

			⟶ 解析方向 ⟶		
素性	F_{i-2}	F_{i-1}	F_i	F_{i+1}	F_{i+2}
チャンクラベル	c_{i-2}	c_{i-1}	$\boxed{c_i}$		

3.3 実 験 結 果

1996 年，1997 年，1998 年および 2005 年の調査対

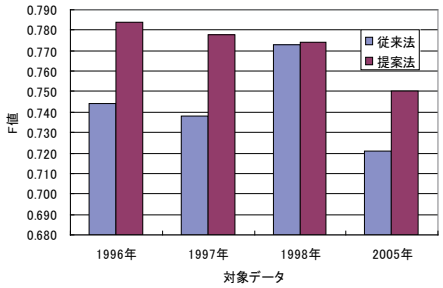

図 3 提案手法と従来手法の比較

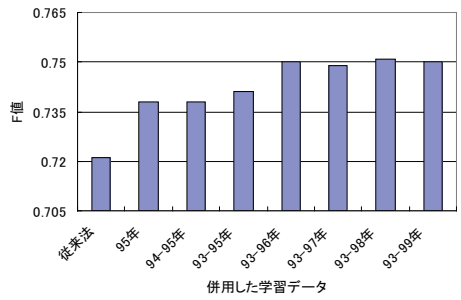

図 4 タグなしデータの分量と性能

象テキストに対する実験結果を**図 3** に示す．先に述べた通り，提案手法は，タグ付きデータとタグなしデータの 2 種類のデータを併用する．図 3 の実験では，タグ付きデータとして IREX コーパスを利用し，タグなしデータとして，1993 年から調査対象テキストの前年までの毎日新聞記事を利用した．例えば，1998 年に掲載された調査対象テキストをテストデータとする場合には，1993 年から 1997 年に掲載された毎日新聞記事をタグなしデータとして用いた．従来手法は，類似形態素素性 SF を用いず，形態素素性 MF および文字種素性 CF のみを用いる手法である．図 3 より，提案手法は，従来手法よりも良い性能を示している．また，従来手法の性能は，テストデータによって大きく変化しているのに対して，提案手法の性能の変化は小さく，提案手法が頑健であることが分かる．

2005 年の調査対象テキストに対して，タグなしデータの分量を変化させて提案手法を適用した実験結果を**図 4** に示す．図 4 より，タグなしデータを増やすと提案手法の性能は改善されることが分かる．

4. むすび

本稿では，日本語固有表現の出現傾向の経年変化について，1995 年 ∼1998 年および 2005 年の毎日新聞記事を対象として調査した結果について述べた．調査した範囲では，掲載日時によらず，固有表現の種類数および出現頻度には大きな差が見られないことを明らかにした．また，固有表現の種類数は年々増加し続けており，ある年の記事に出現した固有表現が翌年の記事にも出現する割合は約 70%から 80%であり，年月の経過につれて出現する割合が低下することを示した．

このような状況では，対象とする固有表現に対して十分な量の固有表現タグ付きコーパスを必要とする従来手法は安定した性能を発揮できない．それに対して，筆者らの提案手法は，そのような出現率の低下に対しても頑健であることを示した．

参 考 文 献

1) Grishman, R. and Sundheim, B.: Message Understanding Conference-6: a brief history, *Proc. of the 16th COLING*, pp.466–471 (1996).
2) Sekine, S. and Eriguchi, Y.: Japanese named entity extraction evaluation: analysis of results, *Proc. of the 18th COLING*, pp.1106–1110 (2000).
3) 桝井文人，鈴木伸哉，福本淳一：テキスト処理のための固有表現抽出ツール NExT の開発，言語処理学会第 8 回年次大会発表論文集，pp.176–179 (2002).
4) 内元清貴，馬　青，村田真樹，小作浩美，内山将夫，井佐原均：最大エントロピーモデルと書き換え規則に基づく固有表現抽出，自然言語処理，Vol.7, No.2, pp.63–90 (2000).
5) Sekine, S., Grishman, R. and Shinnou, H.: A Decision Tree Method for Finding and Classifying Names in Japanese Texts, *Proc. of VLC '98*, pp.171–178 (1998).
6) 宇津呂武仁，颯々野学，内元清貴：正誤判別規則学習を用いた複数の日本語固有表現抽出システムの出力の混合，自然言語処理， Vol.9, No.1, pp. 65–100 (2002).
7) Isozaki, H.: Japanese named entity recognition based on a simple rule generator and decision tree learning, *Proc. of ACL '01*, pp.314–321 (2001).
8) 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).
9) Isozaki, H. and Kazawa, H.: Efficient support vector classifiers for named entity recognition, *Proc. of the 19th COLING*, pp.1–7 (2002).
10) Mota, C. and Grishman, R.: Is this NE tagger getting old?, *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)* (2008).
11) 土屋雅稔，肥田新也，中川聖一：非頻出語に対して頑健な日本語固有表現の抽出，情報処理学会研究報告，Vol.2008–NL–46, pp.1–6 (2008).