社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2008－NL－185 (1)
2008－SLP－ 71 (1)
2008／5／22

非頻出語に対して頑健な日本語固有表現の抽出

土屋雅稔[†1] 肥田新也[†2] 中川聖一[†2]

統計的固有表現抽出のためには，固有表現がタグ付けされた十分な量の学習コーパスが必要である．しかし，新規の固有表現が増加し続けていることを考慮すると，あらゆる固有表現に対応した学習コーパスを用意することは非現実的である．本稿では，この問題に対処するために，固有表現がタグ付けされたコーパスとタグ付けされていないコーパスを併用して，タグ付けされたコーパスに頻出しない語 (非頻出語) を含む固有表現を抽出する手法を提案する．提案手法は 2 段階からなる．最初に，タグ付けされていない大量のコーパスを用いて，入力テキストに含まれている非頻出語を，その非頻出語と良く似た頻出語に対応付ける．次に，元々の語から得られる素性と頻出語から得られる素性の両方を組み合わせて学習した統計的固有表現抽出器によって，固有表現を抽出する．IREX コーパスと NHK コーパスを用いた実験により，提案手法は，非頻出語からなる固有表現の抽出において効果的であることを示す．

Robust Extraction of Japanese Named Entity Including Infrequent Word

Masatoshi Tsuchiya,[†1] Shinya Hida[†2] and Seiichi Nakagawa[†2]

This paper proposes a novel method to extract named entities including infrequent words which do not occur or occur few times in a training corpus using a large unannotated corpus. The proposed method consists of two steps. The first step is to assign the most similar and frequent word to each infrequent word based on their context vectors calculated from a large unannotated corpus. After that, traditional machine learning approaches are employed as the second step. The experiments of extracting Japanese named entities from IREX corpus and NHK corpus show the effectiveness of the proposed method.

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．そのため，英語を対象とする固有表現抽出タスクについては MUC6[1] で，日本語を対象とする固有表現抽出タスクについては IREX[2] で，コンテストにおける課題の 1 つとしてとりあげられ，集中的な研究が行われた．

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．そのため，訓練データに出現しない固有表現に対しても適切に抽出できるような手法が必要である．

そのような手法として，固有表現がタグ付けされていない一般のコーパス (以後，**タグなしコーパス**と呼ぶ) は，タグ付きコーパスに比べて大量に利用できることに注目し，タグなしコーパスとタグ付きコーパスを併用する半教師あり学習が注目を集めている．実装が容易な半教師あり学習手法としては，タグなしコーパスから求めた素性をタグ付きコーパスに追加し，追加されたコーパスを用いて教師あり学習を行うという手法がある．Miller ら[10] は，タグなしコーパスからクラス言語モデルを作成し，タグ付きコーパスにクラスタリング結果を素性として追加して学習を行う方法

†1 豊橋技術科学大学情報メディア基盤センター
Information and Media Center, Toyohashi University of Technology
†2 豊橋技術科学大学情報工学系
Department of Information and Computer Sciences, Toyohashi University of Technology

を提案している．Andoら[11]は，タグなしコーパスから単語の予測問題を学習し，その予測問題を解いた予測値を素性として追加して学習を行う方法を提案している．これらに対して，鈴木ら[12]は，Conditional Random Field のポテンシャル関数と学習方法を拡張し，タグ付きコーパスとタグなしコーパスを直接に併用する手法を提案している．

本稿では，日本語固有表現抽出を対象として，タグなしコーパスから求めた類似語を素性として追加して学習を行う半教師あり学習手法を提案する．提案手法は 2 段階からなる．最初に，入力テキストに含まれている非頻出語を，大量のタグなしコーパスから求めた周辺ベクトルに基づいて，その非頻出語と良く似ていると同時に，タグ付きコーパスに頻出する語に対応付ける．次に，元々の語から得られる素性と対応付けた頻出語から得られる素性の両方を組み合わせて学習した統計的固有表現抽出器によって，固有表現を抽出する．IREX コーパスおよび NHK コーパスを対象として実験を行ったところ，提案手法は，特に非頻出語を含む固有表現の抽出において効果的であることが分かった．

2. 日本語固有表現抽出

2.1 IREX 日本語固有表現抽出タスク

IREX ワークショップにおける固有表現抽出タスクは，**表 1** に示す 8 種類の固有表現を認識するというタスクである[2]．

IREX ワークショップ実行委員会は，訓練用のコーパスを公開している (以後，このコーパスを **IREX コーパス**と呼ぶ)．IREX コーパスは，1995 年 1 月 1 日から 1 月 10 日までの間に発行された 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．筆者らの知る限り，公開されている日本語の固有表現タグ付きコーパスは，他には存在しない[⋆1]．しかも，現実に存在する固有表現の種類数を考えると，IREX コーパスの分量はまだ不十分だと考えられる．例えば，ある日本語組織名辞書[13]には，既に 171,708 個の項目が登録されており，なお増加中である．この数は，IREX コーパスの文数・固有表現数よりもかなり大きい．

2.2 チャンキングとしての固有表現抽出

日本語の固有表現抽出においては，

(1) 新規の固有表現 (例:新設された会社) が頻繁に出現するため，全ての固有表現を網羅した辞書を作成することは現実的ではない．
(2) 多くの固有表現は，既知の一般語の連続である (例:東京新聞)．

という 2 点の理由から，固有表現抽出と形態素解析とは独立に実行可能であるという仮定を置き，固有表現抽出を，形態素列に対するチャンキング問題として定式化することが一般的である．すなわち，**図 1** のように，形態素列を入力とし，形態素それぞれに対して，どのような固有表現の一部であるのか，または，固有表現の一部ではないのか，を表すラベルを付与していくという問題として定式化する．このように定式化すると，十分な量の固有表現タグ付きコーパスを訓練データとして機械学習的手法を適用することによって，適当な固有表現抽出器を得ることは容易である．

チャンクを表現するためのラベルの形式には，Start/End 形式や IOB 形式など様々な方法があるが，大きな性能差はないことが知られている[8),14),15)]．そこで本稿では，最も一般的な IOB2 形式を用いる．この形式では，以下のような 3 つのラベルを用いる．

B チャンクの先頭の形態素であることを表す．
I チャンクの中間の形態素であることを表す．
O チャンクに含まれない形態素であることを表す．

実際には，**B** ラベルおよび **I** ラベルと，固有表現のタイプ (8 種類) を組み合わせた 17 種類のラベル (例: **B–LOCATION**) を用いて，固有表現の種類毎に区別してチャンキングを行う．

日本語固有表現の抽出を，形態素列を対象とするチャンキングとして定式化すると，広く知られている通り，固有表現境界と形態素境界の不整合という問題が発生する．例えば，IREX ワークショップにおける固有表現の定義に従うと，「訪米」という形態素の部分文字列「米」を，アメリカを意味する固有表現として抽出しなければならないが，形態素を単位とする単純なチャンキングでは，このような抽出は不可能である．この問題に対処するため，内元ら[4]は，不整合を引き起こす形態素を書き換える規則を適用する手法を提案している．また，固有表現抽出を，形態素列を対象とするチャンキングとして定式化する代わりに，文字列を対象とするチャンキングとして定式化する手法も提案されている[16),17)]．にも関わらず，本稿では，固有表現抽出を，形態素列を対象とするチャンキングとし

表 1 IREX で使用する固有表現の種類
Table 1 Categories of Japanese Named Entities in IREX

固有表現の種類		例
ARTIFACT	固有物名	ノーベル文学賞
DATE	日付表現	五月五日
LOCATION	地名	日本，韓国
MONEY	金額表現	２０００万ドル
ORGANIZATION	組織名	社会党
PERCENT	割合表現	三割，二〇%
PERSON	人名	村山富市
TIME	時間表現	午前五時

⋆1 IREX ワークショップ実行委員会は，IREX ワークショップの参加者に限定して，テスト用固有表現タグ付きコーパスを提供していた．しかし，筆者らは参加者ではないため，このテスト用コーパスは利用できない．

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 1 学習データの例
Fig. 1 Example of Training Instance for Proposed Method

て単純に定式化する．第 1 に，このような不整合を引き起こす固有表現は少なく，IREX コーパスに含まれる 18,677 個所の固有表現の内，1,022 個所 (5.5%) に過ぎないことが報告されている[18]．第 2 に，このような単純な定式化であっても，従来手法と提案手法とを比較するには十分であるからである．

3. 非頻出語に対して頑健な固有表現抽出

先に述べた通り，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．そこで本稿では，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する．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 1 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，それぞれについて詳細を述べる．

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

表 2 IREX コーパスに含まれる固有表現
Table 2 Statistics of NE Types of IREX Corpus

種類	頻度 (%)
ARTIFACT	747 (4.0)
DATE	3567 (19.1)
LOCATION	5463 (29.2)
MONEY	390 (2.1)
ORGANIZATION	3676 (19.7)
PERCENT	492 (2.6)
PERSON	3840 (20.6)
TIME	502 (2.7)
計	18677

表 3 IREX コーパスに対する固有表現抽出
Table 3 NE Extraction Performance of IREX Corpus

	提案手法		ベースライン		NExT
	CRF	SVM	CRF	SVM	
ARTIFACT	0.487	0.518	0.458	0.457	-
DATE	0.921	0.909	0.916	0.916	0.682
LOCATION	0.866	0.863	0.847	0.846	0.696
MONEY	0.951	0.610	0.937	0.937	0.895
ORGANIZATION	0.774	0.766	0.744	0.742	0.506
PERCENT	0.936	0.863	0.928	0.928	0.821
PERSON	0.825	0.842	0.788	0.787	0.672
TIME	0.901	0.903	0.902	0.901	0.800
全体	0.842	0.834	0.821	0.820	0.732

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，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}	$\boxed{c_i}$		

素性の例を，図 1 に示す．学習時には，最初に，タグ付きコーパスを形態素解析し，形態素素性および文字種素性を得る．次に，タグなしコーパスから求めた周辺ベクトルに基づいて類似形態素素性を割り当てて，図 1 のような学習データ全体を得る．最後に，この学習データから，先に述べた範囲を素性として用いて，統計的固有表現抽出器を学習する．テスト時には，最初に入力テキストを形態素解析し，形態素素性および文字種素性を得る．次に，タグなしコーパスから求めた周辺ベクトルに基づいて類似形態素素性を割り当てる．このようにして得られたデータに対して，先に学習した統計的固有表現抽出器を適用し，固有表現を抽出する．なお，本提案手法では，タグなしコーパスにも出現しない未知語には適切な類似形態素素性を割り当てることができない．この場合は，式 (2) に定義した通り，入力テキスト中の元々の形態素をそのまま類似形態素として用いる．

4. 実　　験

4.1 実験条件

本稿では，固有表現タグ付きコーパスとしては，IREX コーパスを，周辺ベクトルを求めるためのタグなしコーパスとしては，毎日新聞データ (1993 年 ~1995 年) を用いる．ただし，IREX コーパスと毎日新聞データは，10 日間の記事が重複しているので，それらの重複している記事は，タグなしコーパスとしては使わないように取り除いた．結局，本稿で用いたタグなしコーパスの大きさは，3.5M 文・140M 形態素である．また，頻出語集合 M_F を，IREX コーパスに 5 回以上出現した全ての形態素の集合と定義する．実験全体を通して，形態素解析器としては MeCab[19] を用いた．統計的固有表現抽出器を学習する方法としては，Conditional Random Fields(CRF)[20] と，Support Vector Machine(SVM)[21] の 2 つを試みた．実際の実験には，CRF++[*1] および YamCha[*2] をそれぞれ用いた．

4.2 IREX コーパスに対する実験

IREX コーパスに含まれる固有表現の数を**表 2** に，IREX コーパスを対象として固有表現抽出を行った実験結果を**表 3** に示す．ここで，提案手法は，類似形態素素性を用いて学習とテストを行った場合の $F_{\beta=1}$ 値であり，5 分割交差検定を行って求めている．ベースラインは，類似形態素素性を用いずに学習とテストを行った場合の $F_{\beta=1}$ 値であり，5 分割交差検定を行って求めている．また，比較のために，人手規則に基づく手法として NExT[3] を用いた場合の結果も示した．なお，統計的機械学習を用いた場合とは異なり，NExT を用いた場合の結果のみは 5 分割交差検定を行っていない．

類似形態素素性を用いた CRF は，類似形態素素性を用いない CRF に対して，$F_{\beta=1}$ 値で 0.02 の改善を示している．また，類似形態素素性を用いた SVM は，類似形態素素性を用いない SVM に対して，0.01 の改善を示している．従来手法として用いている素性は，山田ら[8] が用いた素性とほぼ同じであり，ベースラインの SVM の結果は，山田らの結果と一致している．よって，類似形態素素性は，少なくとも従来手法に悪影響は与えないことが分かる．また，人手規則に基づく方法は，一貫して，機械学習に基づく方法より劣っている．これは，人手規則に基づく方法は，カバー率・精度の両面で劣っているという，先行研究の報告と一致している．

4.3 NHK コーパスを用いた実験

NHK コーパスは，日本放送協会 (NHK) によって 1996 年 6 月 1 日から 12 日にかけて放映された「お

*1 http://crfpp.sourceforge.net/
*2 http://chasen.org/~taku/software/YamCha/

表 4 NHK コーパスに含まれる固有表現
Table 4 Statistics of NE Types of NHK Corpus

種類	頻度 (%)
DATE	755 (19%)
LOCATION	1465 (36%)
MONEY	124 (3%)
ORGANIZATION	1056 (26%)
PERCENT	55 (1%)
PERSON	516 (13%)
TIME	101 (2%)
計	4072

表 5 NHK コーパスに対する固有表現抽出
Table 5 NE Extraction Performance of NHK Corpus

	提案手法		ベースライン		NExT
	CRF	SVM	CRF	SVM	
DATE	0.630	0.595	0.571	0.569	0.523
LOCATION	0.837	0.825	0.797	0.811	0.741
MONEY	0.988	0.660	0.971	0.623	0.996
ORGANIZATION	0.662	0.636	0.601	0.598	0.612
PERCENT	0.538	0.430	0.539	0.435	0.254
PERSON	0.794	0.813	0.752	0.787	0.622
TIME	0.250	0.224	0.200	0.247	0.260
全体	0.746	0.719	0.702	0.697	0.615

表 6 頻出語・非頻出語の比較
Table 6 Extraction of Frequent/Infrequent NEs

		頻出語のみを含む固有表現	非頻出語のみを含む固有表現	双方を含む固有表現
CRF	提案手法	0.789	0.654	0.621
	ベースライン	0.758	0.557	0.617
出現頻度		2,117 (52.0%)	1,390 (34.1%)	566 (13.9%)

はよう日本」などのニュース番組の発話内容を人手で書き起こしたコーパス[*1]であり，IREX コーパスおよび毎日新聞データとは，表現形式がかなり異なっている．また，IREX コーパスを構成する記事の発行時期は，NHK コーパスを構成するニュース番組の放送時期よりも過去であるから，NHK コーパスには，現実の入力テキストと同じように非頻出語を含む固有表現が出現すると考えられる．この 2 点から，NHK コーパスは，より現実に近い状況での固有表現抽出器の性能を知るために適している．NHK コーパスに対して，ARTIFACT 以外の種類の固有表現について，大学院生 1 名が IREX コーパスと同一の基準で固有表現タグ付けを行った結果を**表 4** に示す．

IREX コーパス全体を学習コーパスとして用いた固有表現抽出器を用いて，NHK コーパスに含まれる固有表現を抽出した場合の性能を**表 5** に示す．類似形態素素性を用いた提案手法は，類似形態素素性を用いないベースラインに対して，良い性能を示している．しかも，表 3 と比較すると，機械学習手法として CRF, SVM のどちらを用いた場合であっても，類似形態素素性を用いることによる改善が大きくなっている．例えば，CRF を用いた場合，表 3 では類似形態素素性を用いることによる改善は 0.02 であったのに対して，表 5 では 0.04 となっている．

提案手法とベースライン手法を用いて NHK コーパスを対象として固有表現抽出を行ったとき，IREX コーパスにおける非頻出語を含む固有表現に対してどのように振る舞うかを調べた結果を**表 6** に示す．IREX コーパスにおける非頻出語からなる固有表現は，1,390 個所 (34.1%) に出現している．よって，現実の入力テキストにおいては，訓練データにおける非頻出語が問題となることが強く示唆される．これらの固有表現に対して，ベースラインの CRF は $F_{\beta=1}$ 値で 0.557 と非常に低い性能を示しているのに対して，提案手法は $F_{\beta=1}$ 値で 0.654 と大きく改善している．また，IREX コーパスには全く出現しない語 (未知語) のみからなる固有表現は，NHK コーパス中の 869 個所に出現していた．これらの表現についても，提案手法は $F_{\beta=1}$ 値で 0.657 という精度で抽出することができた．

5. おわりに

本稿では，固有表現タグ付きコーパスに頻出しない (または，出現しない) 形態素を含む固有表現に対して頑健な固有表現抽出手法を提案した．提案手法は 2 段階からなる．最初に，タグ付けされていないコーパスを大量に用いて，入力文に含まれている非頻出語を，その非頻出語と良く似た頻出語に対応付ける．次に，元々の語から得られる素性と頻出語から得られる素性の両方を組み合わせて統計的固有表現抽出器を学習する．提案手法を用いた結果，IREX コーパスにおいては $F_{\beta=1}$ 値で 0.02 の改善を得た．また，完全にオープンなデータである NHK コーパスにおいては $F_{\beta=1}$ 値で 0.04 の改善を得た．特に，非頻出語からなる固有表現に限って見れば，$F_{\beta=1}$ 値で 0.1 という大きな改善があり，提案手法は非頻出語からなる固有表現の抽出に効果的であることを示した．

日本語の固有表現抽出を，形態素を単位とするチャンキングとして定式化すると，固有表現境界と形態

*1 NHK 技研によって作成されたコーパスであり，一般には公開されていない．

素境界の不整合の問題が生じるため，文字を単位とするチャンキングとして定式化する方法がある．浅原ら[16]は，文字単位でチャンキングし，形態素解析結果の N-best 解の情報を組み合わせることによって，$F_{\beta=1}$ 値で 0.87 を達成している．中野ら[17]は，文字単位でチャンキングし，隣接する文節の情報を組み合わせることによって，$F_{\beta=1}$ 値で 0.89 を達成している．これらは，本提案手法とは，まったく異なる素性を用いて得られた結果であるから，本稿で提案する素性と組み合わせることによって，更に性能の改善が期待される．よって，今後は文字を単位とする手法との統合および，周辺ベクトルの類似度など他の素性の組み込みなどを検討していく予定である．また，他の半教師あり学習手法との比較を進めていく予定である．

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) Miller, S., Guinness, J. and Zamanian, A.: Name Tagging with Word Clusters and Discriminative Training, *Proc. of HLT-NAACL 2004*, pp.337–342 (2004).
11) Ando, R. K. and Zhang, T.: A High-Performance Semi-Supervised Learning Method for Text Chunking, *Proc. of ACL '05*, pp.1–9 (2005).
12) 鈴木　潤，磯崎秀樹：大規模ラベルなしデータを利用した言語解析器の性能検証，言語処理学会第 14 年次大会発表論文集，pp.388–391 (2008).
13) 日外アソシエーツ（編）：DCS 機関名辞書，日外アソシエーツ (2007).
14) Tjong Kim Sang, E.: Representing Text Chunks, *Proc. of the 9th EACL*, pp.173–179 (1999).
15) 宇津呂武仁，颯々野学：ブートストラップによる低人手コスト日本語固有表現抽出，情報処理学会研究報告，Vol.2000–NL–139, pp.9–16 (2000).
16) 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).
17) 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).
18) Utsuro, T. and Sassano, M.: Minimally Supervised Japanese Named Entity Recognition: Resources and Evaluation, *Proceedings of the 2nd International Conference on Language Resources and Evaluation*, pp.1229–1236 (2000).
19) 工藤　拓，山本　薫，松本裕治：Conditional Random Fields を用いた日本語形態素解析，情報処理学会研究報告，Vol.2004-NL-161, pp.89–96 (2004).
20) Lafferty, J., McCallum, A. and Pereira, F.: Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data, *Proceedings of ICML*, pp.282–289 (2001).
21) Cristianini, N. and Shawe-Taylor, J.: *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*, Cambridge University Press (2000).