情報処理学会論文誌　Vol. 49　No. 10　3657–3669 (Oct. 2008)

ラベルなしデータを用いた素性増強による日本語固有表現抽出方法

岩倉友哉[†1]　岡本青史[†1]

本論文では，日本語固有表現抽出における精度改善のための手法を提案する．日本語の固有表現抽出においては，単語の明確な境界がないために，単語認識を行った後に抽出を行う方法が幅広く用いられている．しかし，この方法では，単語の一部だけが固有表現を構成しうるという問題が生じる．この問題に対し，本論文では，単語単位の固有表現抽出の後に，文字単位の固有表現抽出を行う 2 段階の抽出手法を提案する．続いて，従来の固有表現抽出で幅広く利用されてきた，人手で作成された固有名詞辞書やシソーラスの代わりに，複数の固有表現抽出器でラベルなしデータを解析した結果から各単語がなりうる固有表現クラスや各単語と共起する固有表現クラスなどを獲得し，固有表現抽出の手がかりとして利用する方法を提案する．本手法を，IREX の固有表現抽出タスクにおいて評価を行った．その結果，単語単位と文字単位の抽出の組合せおよび，ラベルなしデータの利用が，日本語固有表現抽出の精度改善につながることが示せた．

Japanese Named Entity Extraction by Augmenting Features with Unlabeled Data

Tomoya Iwakura[†1] and Seishi Okamoto[†1]

This paper proposes two methods for improving the performance of Japanese Named Entity (NE) extraction. The first one is the combination of word-unit and character-unit extraction. Most Japanese NE extractors use words segmented by a Japanese morphological analyzer because Japanese language has no word boundary marker. However, word unit is not always consistent with NE unit. To solve this problem, we propose to combine word-unit and character-unit extraction. The other is feature argumentation techniques by using extraction results of NE from unlabeled data with NE extractors. Our method collects the candidate NE classes of each word and the NE classes of its surrounding words from unlabeled data. We use these collected information of each word as features. We apply our NE extraction methods to IREX Japanese NE extraction task. The results show that our methods contribute improved accuracy.

1. はじめに

固有表現（Named Entity=NE）抽出は，テキストに出現する人名や地名，組織名，日付，時間などを抽出する技術である．固有表現抽出は，情報抽出における要素技術の 1 つであり，構文解析などの他の要素技術の精度改善にもつながる技術である．

固有表現抽出の実現方法としては，人手作成の規則に基づく抽出および教師あり機械学習手法に基づく抽出の 2 種類の手法が幅広く用いられている．近年では，教師あり学習手法に基づく固有表現抽出が高い精度を残せることが，数多くの言語において報告されている[1),4),7),28),33)]．そこで，本論文でも，教師あり学習手法を用いて，固有表現抽出を実現する．

日本語では，英語やドイツ語のような言語と異なり，明確な単語境界が存在しない．そのため，形態素解析器を用いて単語を切り出した結果に対し，固有表現を抽出する方法が幅広く用いられている．しかし，切り出された単語の境界と固有表現の境界は必ずしも一致しないため，日本語固有表現抽出では，単語が固有表現の一部になるという問題が起きる[3),22),28),33)]．

本論文では，この問題に対して，単語単位と文字単位の固有表現抽出を組み合わせた抽出手法を提案する．本手法では，単語単位の固有表現抽出を行った後に，単語単位の固有表現抽出結果を手がかりとしながら，文字単位での固有表現抽出を行う．

また，日本語の固有表現抽出に限らず，教師あり学習手法に基づく手法においては，素性設計が高い抽出精度を得るために重要となる．固有表現抽出では，形態素解析結果から得られる品詞や単語，単語の文字種情報などに加え，人名・組織名・地名などの辞書やシソーラスなどの外部リソースから得られる情報を素性として用いるのが一般的である．

外部リソースを用いることで，固有表現抽出において精度改善につながる事例が数多く報告されている[3),4),22),27),28),33)]．しかし，これら外部リソース情報をタスクごとに作成するコストは無視できない．

本論文では，人手作成コストの問題に対し，ラベルなしデータを用いた素性増強手法について提案する．本手法では，ラベルなしデータから，固有表現抽出器を使って，各単語がな

†1 株式会社富士通研究所
Fujitsu Laboratories Ltd.

りうる固有表現クラス，各単語と共起する固有表現クラスの候補などの情報を収集し，素性として利用する．さらに，これらの情報を，漏れ少なくあるいは誤り少なく獲得するために，複数の固有表現抽出器の出力を用いた収集手法を提案する．

本論文の構成は以下のとおりである．2 章で，単語単位と文字単位の抽出を組み合せた日本語固有表現抽出手法を説明し，3 章で，ラベルなしデータを用いた素性増強方法について説明する．4 章では評価実験結果を述べ，5 章に従来手法との比較を述べる．6 章で今後の課題について述べ，7 章で本論文をまとめる．

2. 日本語固有表現抽出手法

本固有表現抽出では，単語単位での抽出を行った後に，文字単位の抽出を適用する．本章では，まず，複数の単語（単語の chunk）からなる固有表現を表現するための chunk 表現方法について説明する．続いて，単語単位の固有表現抽出手法，文字単位の抽出手法について説明し，本固有表現抽出で用いる機械学習アルゴリズムを説明する．

本論文では，IREX[9] の固有表現抽出タスクを用いて評価を行う．**表 1** は IREX で定義された固有表現のクラスとそれらの例である．IREX では，ARTIFACT（製品名，法律名などの固有物名），LOCATION（場所表現），ORGANIZATION（組織名），PERSON（人名），DATE（日付表現），MONEY（金額表現），PERCENT（割合表現），TIME（時間表現）の合計 8 種類の固有表現クラスが定義されている．

2.1 Chunk 表現方法

固有表現は 1 つの単語からだけでなく，複数の単語（単語の chunk）から構成される場合がある．そこで，固有表現を抽出するために，固有表現を構成する単語の chunk を判別する手法を用いる．単語の chunk を表現するための方法として，IOB1[15]，IOB2，IOE1，IOE2[21]，Start/End 法[22] の 5 種類の chunk 表現が提案されている．下記に，それぞれの chunk 表現を説明する．

- **IOB1**：I，O，B の 3 つのタグを用いる．I は chunk の中，O は chunk の外，B は chunk の先頭であることを意味する．IOB1 では，B については，他の chunk に連続して出現する chunk の先頭の単語にだけ付与される．

表 1　IREX で定義された固有表現クラス
Table 1　NE examples defined by IREX committee.

ARTIFACT	LOCATION	ORGANIZATION	PERSON	DATE	MONEY	PERCENT	TIME
ノーベル化学賞	日本	外務省	山田太郎	5 月	100 円	100%	今日

- **IOB2**：I，O，B の 3 つのタグを用いる．I，O に関しては，IOB1 と同じ意味である．しかし，B については，IOB1 と異なり，chunk の先頭に位置する単語につねに付与される．
- **IOE1**：I，O，E の 3 つのタグを用いる．I は chunk の中，O は chunk の外，E は chunk の最後であることを意味する．IOE1 では，E については，ある chunk の直前に連続して出現する chunk の最後の単語にだけ付与される．
- **IOE2**：I，O，E の 3 つのタグを用いる．I，O に関しては，IOE1 と同じ意味である．しかし，E については，IOE1 と異なり，chunk の最後に位置する単語につねに付与される．
- **Start/End (SE) 法**：S，B，I，E，O の 5 つのタグを用いる．S はある 1 つの単語が 1 つの chunk を構成する場合に用いる．B は 2 つ以上の単語で構成される chunk の先頭の単語，E は 2 つ以上の単語で構成される chunk の最後の単語に用いる．I は 3 つ以上の単語で構成される chunk の中の単語に用いる．O は上記以外の単語に用いる．

これらの 5 種類の表現方法を用いて，5 種類の固有表現タグセットを定義する．それぞれの固有表現タグは，固有表現のクラスと B，I，E，S の chunk 内での位置を示すタグの組合せで表現される．

たとえば，PERSON を SE 法で表現する場合は，PERSON の先頭である単語に用いられる B-PERSON，PERSON の中の単語に用いられる I-PERSON，PERSON の最後の単語に用いられる E-PERSON，単独で PERSON になる単語に用いられる S-PERSON の 4 種類のタグが作成される．また，O タグに関しては，固有表現ではない単語に対して用いる．

IREX では 8 種類の固有表現が定義されているので，IOB1，IOB2，IOE1，IOE2 に基づく固有表現タグセットは $(8 \times 2) + 1 = 17$ 種類から構成される．SE 法に基づく固有表現タグセットは，$(8 \times 4) + 1 = 33$ 種類から構成される．

本論文では，これら 5 種類の固有表現タグセットを基に複数の固有表現抽出器を作成し，ラベルなしデータからの素性増強のための単語情報収集に用いる．**図 1** に各固有表現タグセットによる例を載せる．

2.2 単語単位の固有表現抽出

単語単位での固有表現抽出では，各単語を，2.1 節で定義した固有表現タグセットのいずれかの固有表現タグに分類することで，固有表現を構成する単語の chunk とその chunk のクラス判別を行う．

単語単位の固有表現抽出においては，先行研究の手法を参考にし，着目している単語およ

Chunk 表現 / 文	田中 (Tanaka)	使節 (mission)	団 (party)	は (particle)	日 (Japan)	米 (U.S.A)	間 (between)
IOB1	I-ORG	I-ORG	I-ORG	O	I-LOC	B-LOC	O
IOB2	B-ORG	I-ORG	I-ORG	O	B-LOC	B-LOC	O
IOE1	I-ORG	I-ORG	I-ORG	O	E-LOC	I-LOC	O
IOE2	I-ORG	I-ORG	E-ORG	O	E-LOC	E-LOC	O
SE	B-ORG	I-ORG	E-ORG	O	S-LOC	S-LOC	O

図 1　固有表現タグの例．ORG と LOC は ORGANIZATION，LOCATION の略
Fig. 1　Examples of NE tags. ORG and LOC indicate ORGANIZATION and LOCATION.

表 2　素性に用いられる文字種および数値情報
Table 2　Character types and number types used as features.

文字種	平仮名，片仮名，漢字 大文字アルファベット，小文字アルファベット，その他
数値情報（N は数値）	$N \leq 12$, $25 \leq N$，$N \leq 100$，$N \leq 2000$，$2000 < N$

びその前後 2 単語から得られる次の情報を素性として用いる[22),28)]．以下に単語単位の抽出で利用する素性について述べる．以下の説明では，文が，m（$0 < m$）個の単語 $\{w_1,...,w_m\}$ から構成されるとする．

- **単語**：形態素解析器で切り出された単語の表記を素性として用いる．i（$1 \leq i \leq m$）番目の単語 w_i を分類する場合には，現在位置および前後 2 単語 $\{w_{i-2},...,w_{i+2}\}$ を素性として用いる．今回は，形態素解析器として ChaSen version 2.3.3，形態素解析の辞書として ipadic 2.6.3 を用いた[⋆1]．
- **品詞**：ChaSen によって各単語に付与される品詞情報を素性として用いる．$POS(w_i)$ を i 番目の単語 w_i の品詞とすると，本論文では，現在位置および前後 2 単語の品詞 $\{POS(w_{i-2}),...,POS(w_{i+2})\}$ を素性として用いる．
- **文字種**：単語を構成する文字の種類を素性として用いる．単語が 1 文字である場合は，**表 2** にある文字種をそのまま用いる．単語が 2 文字以上で構成される場合は，表 2 にある文字種の組合せで文字種を表現する．
  たとえば，「寝る」のように「漢字」1 文字と「平仮名」1 文字で構成される単語は「漢字-平仮名」という文字種とする．ここでの「-」は文字種の境界を意味する．もし，「食べる」の「べる」の箇所のように同じ文字種が 2 回以上連続する場合は，「漢字-平仮名+」と「+」をつけて表現する．もし，単語が数字であった場合は，表 2 にある数値の範囲に基づいて文字種素性を決定する．$CT(w_i)$ を単語 w_i の文字種とする．本論文では，現在単語および前後 2 単語の文字種 $\{CT(w_{i-2}),...,CT(w_{i+2})\}$ を素性として用いる．
- **前 2 単語に対して付与された固有表現タグ**：解析方向からみて 2 つ前の単語に付与された固有表現タグを素性として用いる．$NT(w_i)$ を固有表現抽出器によって単語 w_i に付与された固有表現タグとする．解析方向として，文頭から文末，文末から文頭という 2 種類を考える．文頭から文末の解析であれば，i 番目の単語の固有表現タグを判別する場合は，文頭側の 2 単語の $NT(w_{i-2})$，$NT(w_{i-1})$ を素性として用いる．文末から文頭の解析であれば，i 番目の単語の固有表現タグを判別する場合は，文末側の 2 単語の $NT(w_{i+1})$，$NT(w_{i+2})$ を素性として用いる．これらを動的素性と呼ぶことにする．また，これらの解析方向に加えて，解析方向において前 2 単語に対して付与された固有表現タグを素性として用いない場合（解析方向なし）の場合も考慮する．

実際には，現在位置および前後 2 単語内に同じ単語や品詞，文字種などが出現する可能性がある．そこで，各素性は現在位置の単語からの出現位置を添字として与えることで区別する．たとえば，i 番目の単語が現在位置である場合に，現在位置 i の単語から得られる素性は，“単語 0=w_i” のように，2 つ前の単語であれば，“単語-2=w_{i-2}” のように素性の種類と現在位置からの出現位置を添字として表現する．

今回は，5 種類の固有表現タグセットと，文頭から文末，文末から文頭の 2 種類の解析方向と，解析方向なしという 3 種類のパターンを組み合わせて $(5 \times 3) = 15$ 種類の単語単位の固有表現抽出器を作成する．

2.3　文字単位の固有表現抽出

日本語では，単語の一部が固有表現を構成する場合がある．たとえば，“田中使節団は訪米” という文が，“田中 / 使節 / 団 / は / 訪米” のように形態素解析されたとする．“/” は単語の区切りである．この文の中で，“田中 / 使節 / 団” という 3 つの単語で ORGANIZATION となる固有表現は，単語の境界と固有表現の境界と一致する．しかし，LOCATION である “米” は，“訪米” という単語に含まれており，単語の境界と固有表現の境界が一致しないため，単語単位の抽出手法では抽出ができない．

この問題に対して，本抽出手法では，文字単位の固有表現抽出[3),33)] を単語単位の抽出の後に適用する方法を用いる．本手法の文字単位での抽出では，現在着目している文字とその前後の文字情報に加えて，各文字を含んでいる単語の単語単位での抽出結果を Stacking[25)]

⋆1 http://chasen-legacy.sourceforge.jp/（参照 2008-7-10）

単語	品詞	文字種	固有表現タグ
田中	固有名詞-姓	漢字+	B-ORG
使節	名詞-一般	漢字+	I-ORG
団	名詞-接尾辞	漢字	E-ORG
は	助詞-係助詞	平仮名	O
訪米	名詞-サ変接続	漢字+	S-LOC

↓

文字	品詞	文字種	文字が出現する単語と位置	単語の固有表現タグ	単語文字種	固有表現タグ
田	B-固有名詞-姓	漢字	B-田中	B-B-ORG	漢字+	B-ORG
中	E-固有名詞-姓	漢字	E-田中	E-B-ORG	漢字+	I-ORG
使	B-名詞-一般	漢字	B-使節	B-I-ORG	漢字+	I-ORG
節	E-名詞-一般	漢字	E-使節	E-I-ORG	漢字+	I-ORG
団	S-名詞-接尾辞	漢字	S-団	S-E-ORG	漢字	I-ORG
は	S-助詞-係助詞	平仮名	S-は	S-O	平仮名	O
訪	B-名詞-サ変接続	漢字	B-訪米	B-S-LOC	漢字	O
米	E-名詞-サ変接続	漢字	E-訪米	E-S-LOC	漢字	B-LOC

図 2　素性の例：上側が単語単位抽出．下側が文字単位抽出．ORG と LOC は ORGANIZATION, LOCATION の略

Fig. 2　Feature expression for training: Top one is an example of word unit chunking one. Bottom one is character unit chunking one. ORG and LOC indicate ORGANIZATION and LOCATION.

の形で素性として利用する．**図 2** の上側に単語単位，下側に文字単位の抽出の例を示す．

文字単位での固有表現抽出における素性および固有表現タグセットは，先行研究[3),33)]を参考にした．今回の文字単位での固有表現抽出では，現在着目している文字から得られる情報に加えて，前後 2 文字から得られる次の 5 種類の情報を用いる．以下の説明では，文が，n（$0 < n$）個の文字 $\{c_1,...,c_n\}$ から構成されるとする．

- **文字**：文字そのものを素性として利用する．本論文では，i（$1 \leq i \leq n$）番目の文字 c_i を分類する際には，現在位置および前後 2 文字 $\{c_{i-2},...,c_{i+2}\}$ を素性として用いる．
- **文字種**：表 2 の分類に従い文字の文字種を決定し素性とする．$CT(c_i)$ を i 番目の文字 c_i の文字種とする．本論文では，現在位置および前後 2 文字の文字種 $\{CT(c_{i-2}),...,CT(c_{i+2})\}$ を素性として用いる．
- **対象の文字を含む単語の表記**：$W(c_i)$ を i 番目の文字 c_i を含む単語とする．また，$P(c_i)$ を i 番目の文字 c_i が単語 $W(c_i)$ において，出現する位置とする．出現する位置の表現としては，SE 法と同じく，先頭（B），中間（I），終わり（E），単独（S）を用いる．この組合せ「$P(c_i) - W(c_i)$」を素性として用いる．たとえば，“外務省は” という文字列が，“外務省 / は” と区切られた場合，「W(外) = 外務省」，「W(務) = 外務省」，「W(省) = 外務省」，「W(は)=は」となる．また，この例では，それぞれの文字が出現する位置は，「P(外) =B」，「P(務) = I」，「P(省) = E」，「P(は)=S」となる．本論文では，前後 2 文字の情報を用いるので，i（$1 \leq i \leq n$）番目の文字 c_i を分類する際には，$\{P(c_{i-2}) - W(c_{i-2}),...,P(c_{i+2}) - W(c_{i+2})\}$ を素性として用いる．
- **対象の文字を含む単語の品詞情報**：$POS(W(c_i))$ を i 番目の文字 c_i を含む単語の品詞とする．本論文では，単語 $W(c_i)$ における文字 c_i の出現位置 $P(c_i)$ と $POS(W(c_i))$ の組合せ「$P(c_i) - POS(W(c_i))$」を素性として用いる．本論文では，前後 2 文字の $\{P(c_{i-2}) - POS(W(c_{i-2})),...,P(c_{i+2}) - POS(W(c_{i+2}))\}$ を素性として用いる．
- **対象の文字を含む単語の文字種情報**：$CT(W(c_i))$ を i 番目の文字 c_i を含む単語の文字種とする．本論文では，前後 2 文字の $\{CT(W(c_{i-2})),...,CT(W(c_{i+2}))\}$ を用いる．
- **対象の文字を含む単語の固有表現タグ**：単語単位の固有表現抽出による分類結果を素性として用いる．$NT(W(c_i))$ を i 番目の文字 c_i を含む単語の固有表現タグとすると，本論文では，出現位置 $P(c_i)$ との組合せ「$P(c_i) - NT(W(c_i))$」を素性とする．本論文では，前後 2 文字の $\{P(c_{i-2}) - NT(W(c_{i-2})),...,P(c_{i+2}) - NT(W(c_{i+2}))\}$ を素性として用いる．今回は，単語の固有表現タグは SE 法で表現されるものとする．異なる chunk 表現に基づく単語単位の固有表現抽出器においては，各表現による単語単位の抽出結果を SE 法による表現に変換して適用する．
- **前 2 文字に対して付与された固有表現タグ**：解析方向からみて 2 つ前の文字に付与された固有表現タグを素性として用いる．解析方向として，先行研究の文字単位の抽出で良い結果を示している文末から文頭の方向を採用した[3),33)]．$NTC(c)$ を，文字単位の固有表現抽出器がある文字 c に付与した固有表現タグとする．この解析方向では，文末側の 2 文字の固有表現タグ $NTC(c_{i+1})$，$NTC(c_{i+2})$ を素性として用いる．

chunk 表現としては，先行研究の文字単位の固有表現抽出で用いられている IOB2 を採用した[3),33)]．各文字は，$(8 \times 2 + 1) = 17$ 種類のいずれかの固有表現タグへ分類される．

単語単位の抽出と同様，現在位置および前後 2 文字内に同じ素性が出現する可能性がある．そのため，i 番目の文字が現在位置である場合，2 つ前の文字であれば，“文字-2=c_{i-2}” のように素性の種類と現在位置からの出現位置を添字として表現する．

2.4　Support Vector Machines に基づく固有表現抽出

固有表現抽出のための機械学習アルゴリズムとして，日本語固有表現抽出において，高い性能を示している Support Vector Machines（SVM）[24)] を用いる．次に，SVM に基づく

固有表現抽出器の作成方法について説明する．入力を，M 個の素性ベクトルと正例・負例の 2 つのクラスからなる学習データ $(\mathbf{x}_1, y_1)$,...,$(\mathbf{x}_M, y_M)$ とする．$\mathbf{x}_i \in R^N$（$1 \leq i \leq M$）は，学習データ中の i 番目の素性ベクトルで，N 次元の素性ベクトル（$\mathbf{x}_i = [f_1,...,f_N]$）で表現される．$y_i \in \{+1, -1\}$ は i 番目のクラスを示す値であり，正例（+1），負例（−1）のいずれかとなる．

SVM は，これらの学習データから，クラスラベル $y_i \in \{+1, -1\}$ への識別関数 $f : R^N \rightarrow \{+1, -1\}$ を導出する．$f(\mathbf{x}) = +1$ の場合，$\mathbf{x}$ は正例に，$f(\mathbf{x}) = -1$ の場合は，$\mathbf{x}$ は負例に判別されたことを意味する．

SVM による分類器は，

$$f(\mathbf{x}) = sign(g(\mathbf{x}))$$

という形の識別関数となる．$g(\mathbf{x})$ は，

$$g(\mathbf{x}) = \sum\nolimits_{\mathbf{z_j} \in SV} \alpha_j y_j K(\mathbf{x}, \mathbf{z_j}) + b$$

の形となる関数である．ベクトル $\mathbf{z_j}$ はサポートベクトルと呼ばれるものであり，$y_j \in \{-1, +1\}$ は，$\mathbf{z_j}$ のラベル，$0 < \alpha_j$ は，$\mathbf{z_j}$ の重み，b はバイアス項と呼ばれるパラメータである．

$K(\mathbf{x}, \mathbf{z})$ はカーネル関数と呼ばれるものであり，素性ベクトルを高次な次元に拡張することができる．本実験では，$K(\mathbf{x}, \mathbf{z}) = (1 + \mathbf{x} \cdot \mathbf{z})^2$ で与えられる多項式カーネルの 2 次を用いた．多項式カーネルの 2 次は，SVM に基づく日本語固有表現抽出において高い性能を示している[3),28),33)]．また，ソフトマージンのパラメータは先行研究を参考に 1 を設定した[28),33)]．

SVM の入力は数値ベクトルである．そこで，素性集合を数値ベクトルに変換し SVM への入力とする方法を用いる．変換方法としては，各素性をベクトルの各次元に関連付け，素性が出現する場合は対応する次元の値を 1，それ以外は 0 とする方法を用いた．

また，SVM は二値分類器であるため，今回の固有表現抽出問題に適用するためには，多値分類を扱えるように拡張する必要がある．多値分類への拡張方法として，One-vs-Rest 法を用いた．この手法では，l 個のクラス c_1,...,c_l が対象である場合に，対象のクラスか対象のクラスではないかのいずれかに分類する l 個の二値分類器を用意する．

接続条件を満たした最適な固有表現タグ列を最終的な結果として得るために，ビタビアルゴリズムを用いる．今回は，SVM に基づく各分類器の出力 g_{c_1},...,g_{c_l} を $s(x) = 1/(1 + exp(-\beta x))$ で定義される sigmoid 関数を用いて，出力値を 0-1 の値にマッピングした[14)]．本実験では，β の値として 1 を用いた．

3. 素性増強のためのラベルなしデータからの単語情報獲得

本章では，3.1 節で，ラベルなしデータから収集する各単語の固有表現クラスの候補や共起する固有表現クラスの候補などの単語情報について述べ，3.2 節で，これらの収集方法について説明する．

本論文では，次の理由から，複数の固有表現抽出器を用いて単語情報を収集し，素性として利用する方法を用いる．まず，固有表現抽出においては，多くの場合が次の 2 つのパターンで抽出されていると予想される．

(1) 単語およびその品詞情報で固有表現と判別できる場合：一般的な姓名，県名・国名などの場所表現，会社名など

(2) 周辺の単語情報で固有表現と判別できる場合：人名を示唆する「さん」，場所を示唆する「出身」，組織名を示唆する「株式会社」など

このような特徴的なパターンにおいては，多くの固有表現抽出器において，高精度な抽出が行えると予想される．さらに，複数の固有表現抽出器による解析結果を比較した場合，ほとんどが一致すると予想される．そこで，複数の固有表現抽出器の出力を考慮することで，(1) のパターンからは固有表現を示唆する表現，(2) のパターンからは固有表現の候補，を漏れ少なくあるいは誤り少なく獲得できると期待される．また，これらの結果から，学習データには出現しない固有表現になりうる単語や固有表現を示唆する単語が数多く収集できると期待され，これらを素性と用いることで固有表現抽出の精度改善が期待される．

3.1 ラベルなしデータから獲得する単語情報

本論文では，固有表現抽出器によってラベルなしデータを解析した結果から，次の 3 種類の情報を獲得する．本論文では，これらの情報を各単語の単語情報と呼ぶことにする．今回は，各単語の固有表現タグの候補および共起する固有表現タグの候補に加えて，それらの頻度情報，順位も素性して利用する．

- **固有表現タグ情報**：各単語がなりうる固有表現タグおよび，各単語と共起する固有表現タグを固有表現タグ情報とする．今回は，固有表現タグとして，SE 法で表現される固有表現タグを収集する．固有表現タグ情報は，各単語の固有表現タグおよび，各単語と共起する前後 2 単語の固有表現タグから収集する．**表 3** は，ラベルなしデータの解析結果とそれらの結果から獲得される単語情報の例である．たとえば，“田中” の固有表現タグ候補として，B-ORGANIZATION と S-PERSON が獲得される．また，“田中” の後（文末側に 1 つ隣）出現する固有表現タグの候補として，E-ORGANIZATION と

表 3　抽出結果の例（上側）と，収集される単語情報の例（下側）
Table 3　Examples of extraction results (top) and examples of information of words collected from the extraction results (bottom).

田中 /B-ORGANIZATION	株式会社 /E-ORGANIZATION	上場 /O	。/O
田中 /S-PERSON	社長 /O	。/O	
田中 /S-PERSON	さん /O	。/O	

↓

単語	現在の単語からの位置	出現する固有表現タグの候補	頻度	順位
田中	現在位置	S-PERSON	2	1
		B-ORGANIZATION	1	2
	1 つ後	O	2	1
		E-ORGANIZATION	1	2
	2 つ後	O	3	1
さん	現在位置	O	1	1
	1 つ前	S-PERSON	1	1
	1 つ後	O	1	1
		:		

O が，2 つ後の位置に出現する固有表現タグの候補として O が収集される．

- **固有表現タグの頻度情報**：ラベルなしデータの解析結果から単語とその固有表現タグが共起する回数を基に，固有表現タグの頻度情報を決定する．たとえば，表 3 の例では，“田中” の固有表現タグとして，B-ORGANIZATION は 1 回，S-PERSON は 2 回収集される．また，“田中” の 1 つ後（文末側に 1 つ隣）出現する固有表現タグとして，E-ORGANIZATION は 1 回，O が 2 回，2 つ後に出現する固有表現タグとして，O が 3 回収集される．これらの頻度情報を二値の素性として表現するために，各単語と共起するそれぞれの固有表現タグ情報の頻度 f を，$f \leq 10$，$10 < f \leq 100$，$100 < f$，の範囲で区切って，あてはまる範囲を素性として用いる．
- **候補ラベルの順位**：各固有表現タグ候補の順位を，頻度情報を基に決定し，素性として用いる．たとえば，表 3 の例では，“田中” の固有表現タグ候補として，S-PERSON は 2 回，B-ORGANIZATION は 1 回獲得されるので，S-PERSON は 1 位，S-ORGANIZATION は 2 位と順位付けする．“田中” の後ろに出現する候補ラベルとして，O は 2 度獲得されるので 1 位，E-ORGANIZATION は 1 度獲得されるので 2 位，2 つ後の候補 O に対しては，1 位が付与される．

今回の実験では，各単語の単語情報として，その単語および前後 2 単語位置に出現する合計 5 単語分の上記 3 種類の情報を集める．さらに，今回の固有表現抽出では，33 種類の固有表現タグが存在するため，1 つの単語に対して最大 $(5 \times 33 \times 3) = 495$ 種類の素性が集められることになる．

学習および分類時では，現在着目している単語に加え，前後 2 単語の情報を素性として用いるので，最大 $(495 \times 5) = 2475$ 種類の素性が追加されることになる．

たとえば，表 3 の例から収集された情報を用いて「山田さん」という文の「山田」を解析するとする．この場合では，2.2 節で述べた前後 2 単語から得られる素性に加えて，1 つ後 (単語+1) の単語「さん」の単語情報として，「単語+1 の候補=O」，「単語+1 が O の頻度=≤ 10」，「単語+1 が O の順位=1」，「単語+1 の 1 つ前の候補=S-PERSON」，「単語+1 の 1 つ前 S-PERSON の頻度=≤ 10」，「単語+1 の 1 つ前 S-PERSON の順位=1」が素性として用いられる．

3.2　複数の固有表現抽出器を用いたラベルなしデータからの単語情報の獲得

3.1 節で紹介した素性増強のために用いる単語情報の獲得を次の手順で行う．

Step 1　2.1 節で示した 5 種類の chunk 表現に基づき，5 種類の学習データを作成する．2.2 節で紹介したように，“文頭から文末” および “文末から文頭” の 2 種類の解析方向に基づく動的素性の利用方法と，動的素性を利用しない場合の合計 3 種類を組み合わせることで，1 つの学習データから 15 (5×3) 種類の固有表現抽出器を作成する．

Step 2　Step1 で作成した 15 種類の固有表現抽出器を使って，ラベルなしデータから固有表現抽出を行う．

Step 3　Step 2 での抽出結果から，3.1 節で述べた単語情報を収集する．

Step 4　Step 1 の学習データに対し，単語情報を新たな素性として追加する．

Step 5　Step 4 で作成した単語情報を素性として追加した学習データから，新しい固有表現抽出器を作成する．

自動抽出結果から収集した情報を基に素性を拡張する場合は，収集の誤りと漏れの問題がある．これらの問題に対し，Step 3 における収集方法として，次の 3 つを考える．

- **収集手法 1**：素性増強のための単語情報を 1 つの固有表現抽出器の出力を使って収集する．今回は，15 種類の固有表現抽出器の中で，最も高い精度を示した固有表現抽出器を用いる．精度として，F-measure を用いた．最も高い F-measure を示す固有表現抽出器を使って単語情報収集することで，漏れの少なさおよび誤りの少なさにおいてバランス良く集めることができると期待される．
- **収集手法 2**：15 種類の固有表現抽出器の出力が一致した個所だけ収集対象とする．この手法で収集することで，収集の漏れは増えるが，誤りの少ない単語情報が得られると

期待される．

- **収集手法 3**：15 種類の固有表現抽出器のすべての出力から収集する．この手法では，仮に，ある文のある単語に対する出力結果がすべて異なっていても，それらのすべての結果を単語情報として収集する．この手法に基いて収集することで，誤りは増えるが，漏れの少ない単語情報が得られると期待される．この手法では，固有表現タグの頻度は 15 種類の固有表現抽出器の結果における平均とする．

単語情報を収集する際には，固有表現タグは SE 法の形式で収集した．SE 法以外の表現に基づく固有表現抽出器は，出力を SE 法の表現に変換してから収集を行った．

また，Step 4 における素性増強手法として次の 4 種類を用いる．

- **Best**：収集手法 1 で収集した単語情報を素性として用いる．
- **Intersection**：収集手法 2 で収集した単語情報を素性として用いる．
- **Union**：収集手法 3 で収集した単語情報を素性として用いる．
- **All**：収集手法 1〜3 で収集した単語情報を素性として用いる．

SVM に基づく固有表現抽出では処理速度が問題となる[12),28)]．そこで，処理速度の改善のために，Polynomial Kernel Expanded（PKE）法[12)] を用いた．PKE 法は Polynomial Kernel に基づく分類器を，すべての素性の組合せを展開することで，linear 分類器に変換する手法である．今回の実験では，SVM に基づく chunker である YamCha [12)] に実装されている PKE 法を使って，polynomial kernel による分類器を linear 分類器に変換した．

4. 実験・評価

本章では，提案固有表現抽出手法である，単語単位と文字単位の組合せと，素性増強手法の評価結果を述べる．まず，4.1 節で実験データについて述べる．続いて，4.2 節で実験設定について述べ，4.3 節で，実験結果を述べる．

4.1 実験データ

本実験では，次のデータセットを用いた．

- **学習データ**：CRL 固有表現データ（CRL データ）を学習データとして用いる．また，交差検定による評価にも用いる．
- **評価データ**：IREX-NE 予備試験トレーニングデータ（dry-run training data），IREX-NE 予備試験データ（dry-run data），IREX-NE 本試験限定課題トレーニングデータ（domain-specific training data），IREX-NE 本試験限定ドメインデータ（ARREST），IREX-NE 本試験一般ドメインデータ（GENERAL），の 5 種類からなる IREX [9)] のデータセットを評価に用いる．
- **ラベルなしデータ**：素性増強のための単語情報を獲得するために，毎日新聞，1991〜1993，1995〜1998，2000〜2002．合計 10 年分のニュース記事を用いる．IREX の評価データは，1994 年と 1999 年の記事に含まれている．そのため，今回の実験では，1994 年と 1999 年の記事は利用しないことにした．

表 4 に学習データ，評価データに付与されている固有表現タグ数を載せる．

表 4　日本語固有表現抽出のための学習データおよび評価データの内訳
Table 4　Training and test data for Japanese NE extraction.

	学習	評価					
NE / Data	CRL データ	本試験 一般課題	本試験 限定課題	本試験限定 トレーニング	予備試験 データ	予備試験 トレーニング	Total
ARTIFACT	871	49	13	11	42	67	182
DATE	3654	277	72	69	110	137	665
LOCATION	5660	416	106	165	192	255	1134
MONEY	390	15	8	19	33	32	107
ORGANIZATION	3813	389	74	80	214	270	1027
PERCENT	500	21	0	3	6	19	49
PERSON	3870	355	97	94	169	138	853
TIME	503	59	19	18	24	8	128
Total	19261	1581	389	459	790	926	4145

4.2 実験設定

4.1 節で述べたデータを用いて次の 2 種類の実験を行った．

- **実験 1**：CRL データにおける 5 分割交差検定：CRL データを記事単位で 5 分割し，交差検定を行い，評価を行った．5 分割の交差検定では，4/5 の CRL データを学習に，残りの 1/5 を評価に用いることを 5 回繰り返す．交差検定による評価では，交差検定の各ラウンドで，4/5 の学習データから作成される固有表現抽出器を使って，素性増強のための単語情報を収集した．交差検定における評価では，それぞれのラウンドで 15 種類，合計，$(5 \times 15) = 75$ 種類の固有表現抽出器を作成した[⋆1]．
- **実験 2**：5 種類のデータによる評価：CRL データを学習データとして作成した固有表現抽出器を，IREX-NE 予備試験トレーニングデータ，IREX-NE 予備試験データ，IREX-

⋆1 交差検定における素性増強のための単語情報収集では，各ラウンドで作成される 15 種類の固有表現抽出器で 10 年分の新聞記事を処理したので，合計で $(15 \times 5 \times 10) = 750$ 年分に相当する新聞記事を解析した．

表 5　固有表現抽出の結果（ラベルなしデータの利用なし）．数値は F-measure．“-F”，“-B”，“-N” は前向き解析，後向き解析，解析方向なしを指す．Av.FM は実験 2 での平均 F-measure．Av.rank は実験 2 での平均順位

Table 5　NE extraction results without feature augmentation ($F_{\beta=1}$). “-F”, “-B”, and “-N” indicate forward direction NE extraction, backward direction NE extraction, and NE extraction without using preceding NE labels, respectively. Av.FM and Av.rank are average F-measure and Average ranking on Exp.2.

	実験 1 (Exp.1)	実験 2 (Exp.2)						
Chunk Rep. / Data	CRL Cross	formal-run GENERAL	formal-run ARREST	domain-specific training	dry-run	dry-run training	Av. FM	Av. rank
IOB1-F	82.97	84.17	86.02	88.94	81.32	82.10	83.89	8.6
IOB1-B	83.85	83.88	84.69	88.84	81.85	**84.06**	84.18	8
IOB1-N	81.67	80.50	82.64	86.93	79.43	79.64	81.04	17.2
IOB2-F	85.11	82.92	85.56	87.95	80.27	81.49	82.92	14
IOB2-B	**86.07**	84.03	84.84	89.23	81.98	83.74	84.25	6.8
IOB2-N	81.60	83.53	84.44	88.72	81.11	82.65	83.55	13.4
IOE1-F	82.01	83.79	85.87	89.26	81.37	81.87	83.73	10.4
IOE1-B	82.64	83.80	84.65	88.69	81.74	83.77	84.04	10
IOE1-N	81.65	80.52	82.79	86.46	79.27	79.38	80.92	17.6
IOE2-F	82.70	84.65	85.90	89.94	82.68	81.78	84.36	4.8
IOE2-B	83.23	84.00	84.76	89.31	81.50	83.56	84.11	8.2
IOE2-N	81.60	84.11	**87.25**	89.35	82.77	82.09	84.30	4.8
SE-F	85.15	83.95	84.46	89.66	82.15	80.87	83.62	10
SE-B	85.90	83.67	85.52	89.46	81.85	83.11	84.03	8.4
SE-N	85.85	**85.49**	86.13	**90.47**	**83.27**	83.80	**85.83**	1.6

NE 本試験逮捕トレーニングデータ，IREX-NE 本試験限定ドメインデータ（ARREST），IREX-NE 本試験一般ドメインデータ（GENERAL），の 5 種類のデータに対する抽出精度で評価した．

固有表現抽出器の精度評価として，次に定義される F-measure を用いた．

Recall = NUM / (抽出すべき固有表現の数)

Precision = NUM / (固有表現抽出器が出力した固有表現の数)

F-measure = 2 × Recall × Precision / (Recall + Precision)

NUM は固有表現抽出器により正しく抽出された固有表現の数である．

4.3 実験結果

単語単位と文字単位抽出の組合せおよび，素性増強手法の評価結果について述べる．まず，**表 5** に，3.1 節に述べた素性増強を行っていない 15 種類の固有表現抽出器の評価結果を載せる．これらの抽出器のうち，SE 法を基にした動的素性を用いない固有表現抽出器（SE-N）が，実験 1 では 3 番目の結果，実験 2 の平均では最も良い結果を示した．本実験では，SE-N のモデルをベースラインとし，収集手法 1 用の単語情報獲得のための固有表現抽出器とする．

表 6　SE 法に基づく動的素性を用いない固有表現抽出結果．Rec.，Pre. は Recall および Precision の略

Table 6　Performance of the baseline model on IREX data. Rec. and Pre. indicate Recall and Precision.

	formal-run GENERAL		formal-run ARREST		domain-specific training		dry-run		dry-run training	
	Rec.	Pre.	Rec.	Pre.	Rec.	Pre.	Rec.	Pre.	Rec.	Pre.
単語単位の固有表現抽出										
ORGANIZATION	71.72	82.30	63.51	87.04	73.75	85.51	68.22	83.43	67.04	81.17
PERSON	88.45	89.46	85.57	89.25	97.87	94.85	91.72	91.72	90.58	88.03
LOCATION	76.44	83.25	77.36	77.36	75.76	75.76	84.38	85.26	85.10	84.11
ARTIFACT	36.73	48.65	53.85	43.75	45.45	62.50	9.52	23.53	28.36	70.37
DATE	93.86	95.94	95.83	93.24	98.55	97.14	82.73	85.85	90.51	96.88
TIME	94.92	98.25	94.74	100.00	94.44	100.00	87.50	91.30	87.50	100.00
MONEY	100.00	100.00	100.00	100.00	94.74	94.74	96.97	96.97	81.25	100.00
PERCENT	80.95	89.47	-	-	66.67	66.67	83.33	83.33	89.47	94.44
Total	80.77	86.81	80.72	85.09	84.10	86.16	77.97	85.67	77.32	86.37
F-measure	83.68		82.85		85.12		81.64		81.60	
単語単位+文字単位の固有表現抽出										
ORGANIZATION	73.78	84.41	66.22	90.74	77.50	95.38	70.09	85.23	71.11	85.71
PERSON	88.45	89.46	85.57	89.25	97.87	94.85	92.31	92.86	90.58	88.03
LOCATION	80.77	87.96	86.79	87.62	87.27	87.80	86.98	88.83	89.02	86.97
ARTIFACT	36.73	48.65	53.85	43.75	45.45	62.50	9.52	25.00	28.36	70.37
DATE	93.86	95.94	95.83	93.24	98.55	97.14	82.73	85.85	90.51	96.88
TIME	94.92	98.25	94.74	100.00	94.44	100.00	87.50	91.30	87.50	100.00
MONEY	100.00	100.00	100.00	100.00	94.74	94.74	96.97	96.97	81.25	100.00
PERCENT	90.48	100.00	-	-	66.67	66.67	100.00	100.00	89.47	94.44
Total	82.54	88.65	83.80	88.59	88.89	92.10	79.37	87.57	79.59	88.48
F-measure	85.49		86.13		90.47		83.27		83.80	

また，**表 6** に，ベースラインモデルの評価データにおける “単語単位の抽出” だけの結果および “単語単位および文字単位抽出の組合せ” による，固有表現クラス別の結果を載せる．表 6 から，単語単位と文字単位の固有表現抽出を組み合わせることですべての評価データにおいて，精度が改善されていることが分かる．これらの結果においては，単語単位だけの抽出と比較し，文字単位での抽出を加えた場合は，全体の F-measure の平均で 2.85 ポイントの改善が得られている．

続いて，Intersection，Union，All，Best の 4 種類の素性増強を行った場合の結果を**表 7** に載せる．表 7 から，素性増強を行った固有表現抽出器は，表 5 にある素性増強を行って

表 7　ベースラインおよび素性増強手法に基づく固有表現抽出器による結果．Av.FM は実験 2 の抽出結果の平均 F-measure
Table 7　Experimental results obtained with our proposed methods and SE-N of the base line (F-measure). Av.FM is for Exp.2.

	実験 1 (Exp.1)	実験 2 (Exp.2)					
	CRL Cross	formal-run GENERAL	formal-run ARREST	domain-specific training	dry-run	dry-run training	Av. FM
ベースライン	85.85	85.49	86.13	90.47	83.27	83.80	85.83
Best	87.54(+1.69)	**87.20**	90.31	92.05	84.37	85.68	87.92 (+2.09)
Intersection	88.04(+2.19)	86.06	89.84	92.51	84.95	85.47	87.77 (+1.94)
Union	88.26(+2.41)	86.41	**91.85**	**92.61**	**85.49**	85.94	**88.46** (+2.63)
All	**88.50**(+2.65)	87.09	90.20	91.78	**85.49**	**86.13**	88.14 (+2.31)

いないすべての抽出器より良い精度を示していることが分かる．これらの結果から，ラベルなしデータを解析した結果を用いた素性増強手法は精度改善に寄与することが分かる．

実験 1 においては，複数の固有表現抽出器を用いて素性増強を行う Intersection，Union および All が，1 つの固有表現抽出器を用いて素性増強を行う Best による結果より，F-measure で 0.5 ポイント〜0.96 ポイント高い結果を示している．また，実験 2 においては，Union と All が，Best より，F-measure で，0.54 ポイント，0.22 ポイント高い結果を残している．このことから，複数の固有表現抽出器を用いて単語情報を収集し素性として用いることで，さらなる精度改善につながったことが分かる．

表 8 は素性増強手法 All を用いた場合の “単語単位の抽出” だけの結果および “単語単位および文字単位抽出の組合せ” による各固有表現クラス別の結果である．この結果から，素性増強後においても，単語単位の抽出と文字単位の抽出を組み合わせることで精度改善が得られることが分かる．All に基づく固有表現抽出では，単語単位だけの抽出と比較し，文字単位での抽出を加えた場合は，全体の F-measure の平均で 2.94 ポイントの改善が得られている．また，表 6 の素性増強を行わない単語単位の抽出と比較し，表 8 の単語単位と文字単位の抽出と All による抽出との結果は，実験 2 の平均の F-measure で 5.16 ポイントの改善が得られている．

また，表 6，表 8 のベースラインおよび All に基づく固有表現抽出器の評価データ上での結果，**表 9** の CRL データにおける交差検定の結果から，本提案抽出手法により，8 種類の固有表現クラスにおいて，ほとんどの場合に Precsion を下げることなく，Recall が改善されていることが分かる．

表 10 に，素性増強に利用するラベルなしデータのサイズによる 5 種類の評価データの平均 F-measure，precision，recall の変化を示す．4 種類の素性増強手法，Best，Intersection，Union，All は，ラベルなしデータとして新聞記事 1 年分，5 年分，10 年分を用いた場合において，ベースラインより高い F-measure を残せていることが分かる．また，表 10 から，ラベルなしデータ 1 年分を使った All を除いて，ベースラインより高い，precision および recall を残せていることが分かる．

しかし，ラベルなしデータとして新聞記事 10 年分を用いた Best〜All の precision は，新聞記事 5 年分を用いた場合 precision と比較して，低い値となっている．この理由の 1 つとしては，ラベルなしデータのサイズを増やすことで，各単語の単語情報の種類が増大し，曖昧性解消が難しくなっているためと予想される．

この解決方法として，文節情報や[33] 係り受け情報から得られる新たな素性の利用などが

表 8　素性増強手法 All に基づく固有表現抽出結果．Rec.，Pre. は Recall および Precision の略
Table 8　Performance of the model based on All on IREX data. Rec. and Pre. indicate Recall and Precision.

	formal-run GENERAL		formal-run ARREST		domain-specific training		dry-run		dry-run training	
	Rec.	Pre.	Rec.	Pre.	Rec.	Pre.	Rec.	Pre.	Rec.	Pre.
単語単位の固有表現抽出										
ORGANIZATION	74.29	81.64	71.62	80.30	71.25	91.94	71.50	82.26	70.74	82.33
PERSON	91.27	90.76	93.81	91.00	98.94	94.90	97.04	93.18	92.75	90.14
LOCATION	79.33	84.40	76.42	81.82	77.58	78.05	87.50	83.58	88.24	87.89
ARTIFACT	30.61	55.56	61.54	72.73	36.36	57.14	16.67	36.84	38.81	72.22
DATE	95.67	95.67	97.22	95.89	100.00	98.57	90.91	85.47	93.43	95.52
TIME	96.61	98.28	100.00	100.00	94.44	100.00	95.83	88.46	100.00	88.89
MONEY	100.00	100.00	100.00	100.00	94.74	94.74	96.97	96.97	84.38	96.43
PERCENT	80.95	80.95	-	-	100.00	75.00	83.33	83.33	89.47	94.44
Total	82.99	87.47	84.83	87.77	84.75	88.21	82.53	85.34	80.99	87.72
F-measure	85.17		86.27		86.44		83.91		84.22	
単語単位+文字単位の固有表現抽出										
ORGANIZATION	76.35	83.90	74.32	83.33	73.75	95.16	73.36	84.41	73.70	86.52
PERSON	90.99	90.73	93.81	91.00	98.94	94.90	97.04	93.71	92.03	90.07
LOCATION	84.13	89.06	88.68	94.95	90.91	91.46	90.63	87.44	92.16	90.73
ARTIFACT	30.61	55.56	61.54	72.73	36.36	57.14	16.67	36.84	38.81	72.22
DATE	95.67	95.67	97.22	95.89	100.00	98.57	90.91	85.47	93.43	95.52
TIME	96.61	98.28	100.00	100.00	94.44	100.00	95.83	88.46	100.00	88.89
MONEY	100.00	100.00	100.00	100.00	94.74	94.74	96.97	96.97	84.38	96.43
PERCENT	95.24	95.24	-	-	100.00	75.00	100.00	100.00	89.47	94.44
Total	84.88	89.41	88.69	91.76	89.98	93.65	83.92	87.12	82.83	89.71
F-measure	87.09		90.20		91.78		85.49		86.13	

表 9　CRL データ上での 5 分割交差検定結果
Table 9　Experimental results on five hold cross-validation with CRL data.

	ベースライン（素性増強なし）		提案手法 (All)	
	Recall	Precision	Recall	Precision
ORGANIZATION	76.00	85.08	79.72	86.96
PERSON	85.67	89.02	90.88	91.89
LOCATION	86.46	87.46	88.81	90.46
ARTIFACT	32.63	66.62	39.67	66.70
DATE	89.99	94.65	92.88	94.35
TIME	86.98	93.21	89.82	90.41
MONEY	87.43	95.93	89.29	96.39
PERCENT	91.17	97.54	94.07	97.39
Total	82.98	88.92	86.48	90.62
F-measure	85.85		88.50	

表 10　単語情報獲得に用いるラベルなしデータのサイズによる実験 2 における F 値（FM），Precision（Pre.），Recall（Rec.）の平均の変化．data size はラベルなしデータとして利用する新聞記事の年数
Table 10　Average F-measure (FM), Precision (Prec.) and Recall (Rec.) on Exp.2 obtained with NE extractors using 1 year, 5 year and 10 year news articles for augmenting features. "data size" means the amount of news articles.

Baseline（ラベルなしデータの利用なし）

	Rec.	Prec.	FM
	82.83	89.07	85.83

ラベルなしデータの利用あり

data	Best			Intersection			Union			All		
size	Rec.	Prec.	FM	Rec.	Prec.	FM	Rec.	Prec.	FM	Rec.	Prec.	FM
1	85.18	89.27	87.17	84.52	89.27	86.81	85.20	89.60	87.33	84.73	88.97	86.79
5	86.17	90.22	88.13	85.81	90.27	87.97	85.99	90.78	88.31	86.29	90.55	88.36
10	86.05	89.89	87.92	85.82	89.81	87.77	86.80	90.20	88.46	86.06	90.33	88.14

考えられる．また，獲得されたすべての単語情報を用いるのではなく，上位数個の単語情報だけを用いるなどの方法が考えられる．

5. 先 行 研 究

5.1 日本語固有表現抽出における先行研究との比較

表 11 に IREX のデータを用いた先行研究の結果（F-measure）と本提案手法による結果を載せる．Uchimoto ら[22] は固有表現抽出器の作成のために最大エントロピー法（Maximum Entropy=ME）を用いている．トレーニングデータとしては，CRL データ，dryrun データ，dryrun トレーニングデータ，分野限定課題のトレーニングデータを用いている．さらに，人手で作成された固有表現辞書から得られる情報を素性として用いることで，GENERAL において 80.17，ARREST において 85.75 の精度（F-measure）を得ている．

竹本ら[32] は，人手で作成された規則に基づく固有表現抽出器によって GENERAL において 83.86 の F-measure を得ている．Utsuro ら[23] は，ME 法に基づく 3 種類の固有表現抽出器の結果を手がかりとする Decision List（DL）に基づく固有表現抽出器によって，GENERAL において，84.07 の F-measure を得ている．

山田ら[30] は，SVM に基づく固有表現抽出手法で，CRL データの交差検定で，83.2 という F-measure を得ている．

磯崎ら[28] は，CRL データを学習データとし，SVM を用いて固有表現抽出器を作成して

表 11　先行研究との比較．GE と AR は GENERAL と ARREST を指す
Table 11　Comparison with previous works. GE and AR indicate GENERAL and ARREST.

Method	GE	AR	CRL-DATA	抽出手法	外部資源
Uchimoto ら[22]	80.17	85.75	-	ME に基づく単語単位の抽出と書き換え規則	人手作成固有表現辞書
竹本ら[32]	83.86	-	-	人手作成抽出規則と複合語分割辞書	-
Utsuro ら[23]	84.07	-	-	3 種類の ME に基づく固有表現抽出器の出力を DL で Stacking	-
山田ら[30]	-	-	83.2	SVMs に基づく単語単位の抽出と学習データに出現した事例は分割	-
磯崎ら[28]	85.77	-	86.77	SVMs に基づく単語単位の抽出とテンプレートルールの利用	日本語語彙大系
Asahara ら[3]	-	-	87.21	SVMs をによる文字単位での抽出と形態素解析の N-best 解の利用	日本語語彙大系
中野ら[33]	-	-	89.03	SVMs による文字単位での抽出文節情報を利用	日本語語彙大系
山田[29]	-	-	88.33	SVMs に基づく Shift-Reduce 法による抽出文節情報を利用	-
本手法ベースライン	85.49	86.13	85.85	SVMs による単語単位と	-
Best	**87.20**	**90.31**	87.54	文字単位の抽出の組合せ	ラベルなしデータ
Intersection	**86.06**	**89.84**	88.04		（新聞記事 10 年分）
Union	**86.41**	**91.85**	88.26		
All	**87.09**	**90.20**	88.50		

いる．素性増強のために，人手で作成されたシソーラスである日本語語彙大系[31] を用いている．この結果，GENERAL においては，85.77，CRL データの交差検定で，86.77 という F-measure を得ている．

Asahara ら[3] は，形態素解析結果の上位複数解と日本語語彙大系を用いる文字単位の固有表現抽出器により CRL データの交差検定で，87.21 という精度を得ている．

本提案手法である SVM を用いた単語単位と文字単位の抽出による固有表現抽出手法は，CRL データを学習データとし，外部リソースを用いずに，GENERAL において 85.49，ARREST で 86.13 を得ている．また，CRL データの交差検定では，F-measure で 85.85 を得ている．

また，本提案手法基づくラベルなし新聞記事 10 年分を外部資源として用いた 4 種類の固有表現抽出器は，IREX の GENERAL，ARREST タスクにおいて，先行研究と比較し，高い精度を残せていることが分かる．素性増強手法 All においては，GENERAL で 87.09，ARREST で 90.20，CRL データの交差検定で，88.50 という F-measure を残している．

しかし，文節情報を素性として用いる手法と比較すると同等あるいは低い精度となっている．山田[29] は，文節から得られる情報を素性として用いた Shift-Reduce 法に基づく固有表現抽出手法において，CRL データの交差検定で 88.33 の F-measure を得ている．また，中野ら[33] は，文節から得られる情報と日本語語彙大系を利用する文字単位の固有表現抽出によって，CRL データの交差検定で 89.03 の F-measure を得ている．

中野ら[33] は，文字単位の抽出において文節から得られる素性を加えることで大幅な精度改善が得られることを報告している．このことから，本実験で文節から得られる素性を用いていないことが，中野らと比較し，低い精度となった 1 つの理由と考えられる．本提案手法であるラベルなしデータからの単語情報の収集による素性増強手法は，これら先行研究と組み合わせて利用することが可能である．今後の課題としては，さらなる精度改善のために，先行研究で提案されている素性情報を用いた抽出手法の開発があげられる．

5.2　ラベルなしデータの利用における先行研究との比較

ラベルなしデータを用いた素性の増強方法として，固有表現抽出や Word Sence Disambiguation のタスクにおいて，ラベルなしデータ中の単語をクラスタリング結果を素性として加えることで，精度改善することが報告されている[2),8),13)]．

他のラベルなしデータを用いる手法としては，検索クエリログからの固有表現に関する知識の獲得[17]，Wikipedia 情報を用いた素性の増強[10]，bootstrapping 法に基づく規則や辞書獲得手法[5),16),26)] なども研究されている．

また，Ando ら[1] は，自動作成されるラベル付きデータを用いた Alternating Structure Optimization（ASO）という半教師あり学習手法を提案している．ASO では，ラベルなしデータにおける前後の単語から現在位置の単語の推定，ラベル付きデータから得られる異なる素性集合から作成される分類器の出力予測など，目的タスクに関連するタスクの学習データを自動作成する．その後，自動作成された各関連タスクの学習データから，これら関連タスクで共有される structure parameter を獲得する[⋆1]．目的タスクの学習・分類時には，それら関連タスクから得られた structure parameter を用いることで，元々の入力ベクトルを，その入力ベクトルと「structure parameter と入力ベクトルの積」との組に変換し，それらに対し，学習・分類を行う．Ando らは，この ASO により，英語，ドイツ語における固有表現抽出で精度改善が得られることを報告している．

これらの先行研究と比較し，本提案手法は，以下の特徴を持つ．まず，本手法では，複数の固有表現抽出器の出力を考慮し，収集の振舞いの調節を行う．たとえば，漏れ少なく収集を行う場合には，すべての抽出器の出力結果から収集することで実現する．また，誤りの少ない収集を行う場合は，すべての抽出器の結果が一致した場合から収集することで実現する．さらに，異なる chunk 表現方法を用いることで，新たな学習データを作成することなく，1 つの学習データから複数の固有表現抽出器を作成できる．

また，クラスタリングに基づく素性増強，Wikipedia 情報を用いた素性増強と比較では，本手法では，固有表現抽出器の出力を用いるので，各単語の固有表現クラス の候補，各単語と共起する固有表現クラスの候補など直接目的のタスクに関連する単語情報を収集しているという点で異なる．

ASO との比較では，本手法では増強される素性の決定の方法が異なる．本手法では，複数の固有表現抽出器によるラベルなしデータからの抽出結果から，各単語がなりうる固有表現クラス，各単語が共起した固有表現クラスという情報を収集し，素性として用いる．

ASO では，入力の事例と structure parameter により決まる素性空間上での学習・分類となるので，追加される素性は入力事例の素性すべてを使って決まる．これに対し，本手法では，単語ごとに追加される素性をラベルなしデータからの抽出結果全体を見て決定しているので，事例に出現する単語にだけ依存して決まる．

⋆1 structure parameter 獲得は大きく次のような手順で行われる．(1) 関連タスクからの学習を，現在の structure parameter を考慮しながら行う．(2) すべての関連タスクの学習が終了後に，関連タスクからの学習結果を行列で表現し，singular value decomposition を用いて次元圧縮を行う．(3) その結果を，次の structure parameter とする．(4) 収束するまで (1)〜(3) を繰り返す．

6. 今後の課題

今後の課題として，学習時間の改善があげられる．本論文では，今回提案した素性増強手法により，精度改善が得られることを実験的に示した．しかし，素性が増えたために，より長い学習時間がかかるという問題がある．たとえば，SE 法の動的素性を用いない場合の学習における学習時間が約 7 時間であったのに対し，ラベルなしデータとして毎日新聞 10 年分を用いた素性増強手法 All を用いた場合は学習時間は約 105 時間であった[⋆1]．この問題の解決方法の 1 つとして，より計算量の少なくかつ高い精度を示せる online 学習アルゴリズム[6] の適用が考えられる．

また，より高精度な固有表現抽出器の作成のために，文節情報や[29),33)]，人手で作成されたリソース[31)]，先行研究で用いられているラベルなしデータを用いた素性増強[1),2),8),10),13)]など，先行研究で用いられている素性情報を組み合わせた固有表現抽出手法の開発があげられる．

さらなる本提案手法の有効性の評価のためには，スペイン語，オランダ語[18),⋆2]．英語，ドイツ語[20),⋆3]，中国語[⋆4]など，他の言語の固有表現抽出タスクでの評価，Noun Phrase Chunking[⋆5]，English Syntactic Chunking[19),⋆6]など固有表現抽出同様の枠組みが適用できるタスクでの評価があげられる．

7. ま と め

本論文では，日本語固有表現抽出における精度改善のための手法提案を行った．本手法では，日本語固有表現抽出における単語の一部が固有表現となる問題の解決のために，単語単位と文字単位の組合せによる抽出を行う．また，複数の固有表現抽出器を使ってラベルなしデータを解析した結果から，各単語がなりうる固有表現クラス，各単語と共起する固有表現クラスなどの情報を獲得し，素性として用いる．日本語固有表現抽出コンテスト IREX のデータセットおよびラベルなしデータとして毎日新聞 10 年分を用いて評価を行った．その結果，本提案手法である素性増強と文字単位の抽出を組み合わせることで，F-measure で 5 ポイント以上の改善が得られることが示せた．

⋆1 Intel(R) Xeon(TM) CPU 3.06 GHz，4 GB のメモリを搭載した Linux で計測.
⋆2 http://www.cnts.ua.ac.be/conll2002/ner/（参照 2008-7-10）
⋆3 http://www.cnts.ua.ac.be/conll2003/ner/（参照 2008-7-10）
⋆4 http://www.sighan.org/（参照 2008-7-10）
⋆5 http://www.cnts.ua.ac.be/conll99/npb/（参照 2008-7-10）
⋆6 http://www.cnts.ua.ac.be/conll2000/chunking/（参照 2008-7-10）

参 考 文 献

1) Ando, R. and Zhang, T.: A High-Performance Semi-Supervised Learning Method for Text Chunking, *Proc. ACL'05*, pp.1–9 (2005).
2) Ando, R.K.: Semantic Lexicon Construction: Learning from Unlabeled Data via Spectral Analysis, *Proc. CoNLL'04*, pp.9–16 (2004).
3) Asahara, M. and Matsumoto, Y.: Japanese Named Entity Extraction with Redundant Morphological Analysis, *Proc. HLT-NAACL'03*, pp.8–15 (2003).
4) Carreras, X., Màrques, L. and Padró, L.: Named Entity Extraction using AdaBoost, *Proc. CoNLL'02*, pp.167–170 (2002).
5) Collins, M. and Singer, Y.: Unsupervised models for named entity classification, *Proc. EMNLP and VLC'99*, pp.100–110 (1999).
6) Crammer, K., Dekel, O., Keshet, J., Shalev-Shwartz, S. and Singer, Y.: Online Passive-Aggressive Algorithms, *Journal of Machine Learning Research*, Vol.7, pp.551–585 (2006).
7) Florian, R., Ittycheriah, A., Jing, H. and Zhang, T.: Named Entity Recognition through Classifier Combination, *Proc. CoNLL'03*, pp.168–171 (2003).
8) Freitag, D.: Trained Named Entity Recognition using Distributional Clusters, *Proc. EMNLP'04*, pp.262–269 (2004).
9) IREX, C.: *Proc. IREX Workshop* (1999).
10) Kazama, J. and Torisawa, K.: Exploiting Wikipedia as External Knowledge for Named Entity Recognition, *Proc. EMNLP-CoNLL'07*, pp.698–707 (2007).
11) Kudo, T. and Matsumoto, Y.: Chunking with Support Vector Machines, *Proc. NAAC'01*, pp.192–199 (2001).
12) Kudo, T. and Matsumoto, Y.: Fast Methods for Kernel-Based Text Analysis, *Proc. ACL'03*, pp.24–31 (2003).
13) Miller, S., Guinness, J. and Zamanian, A.: Name Tagging with Word Clusters and Discriminative Training., *HLT-NAACL'04*, pp.337–342 (2004).
14) Platt, J.C.: *Probabilities for SV machines*, MIT Press (2000).
15) Ramshaw, L. and Marcus, M.: Text Chunking Using Transformation-Based Learning, *Proc. VLC'95*, pp.82–94 (1995).
16) Riloff, E. and Jones, R.: Learning Dictionaries for Information Extraction by Multi-Level Bootstrapping, *AAAI/IAAI'99*, pp.474–479 (1999).
17) Sekine, S. and Suzuki, H.: Acquiring ontological knowledge from query logs, *Proc. WWW'07*, pp.1223–1224 (2007).
18) Tjong Kim Sang, E.F.: Introduction to the CoNLL-2002 Shared Task: Language-Independent Named Entity Recognition, *Proc. CoNLL'02*, pp.155–158 (2002).

19) Tjong Kim Sang, E.F. and Buchholz, S.: Introduction to the CoNLL-2000 Shared Task: Chunking, *Proc. CoNLL and LLL'00*, pp.127–132 (2000).
20) Tjong Kim Sang, E.F. and De Meulder, F.: Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition, *Proc. CoNLL'03*, pp.142–147 (2003).
21) Tjong Kim Sang, E.F. and Veenstra, J.: Representing Text Chunks, *Proc. EACL'99*, pp.173–179 (1999).
22) Uchimoto, K., Ma, Q., Murata, M., Ozaku, H., Utiyama, M. and Isahara, H.: Named Entity Extraction Based on A Maximum Entropy Model and Transformation Rules, *Proc. ACL'00*, pp.326–335 (2000).
23) Utsuro, T., Sassano, M. and Uchimoto, K.: Combining Outputs of Multiple Japanese Named Entity Chunkers by Stacking, *Proc. EMNLP 2002*, pp.281–288 (2002).
24) Vapnik, V.: *Statistical Learning Theory*, John Wiley & Sons (1998).
25) Wolpert, D.H.: Stacked Generalization, *Neural Networks*, Vol.5, pp.241–259 (1992).
26) Yarowsky, D.: Unsupervised Word Sense Disambiguation Rivaling Supervised Methods, *Proc. ACL'95*, pp.189–196 (1995).
27) Zhang, T. and Johnson, D.: A Robust Risk Minimization based Named Entity Recognition System, *Proc. CoNLL'03*, pp.204–207 (2003).
28) 磯崎秀樹，賀沢秀人：SVM に基づく固有表現抽出の高速化，*IPSJ SIG notes NL-149-1*, pp.1–8 (2002).
29) 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，自然言語処理研究会報告 NL-179，pp.13–18 (2007).
30) 山田寛康，工藤　拓，松本裕治：Support Vector Machines 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).
31) 池原　悟，宮崎正弘，白井　諭，横尾昭男，中岩浩巳，小倉健太朗，大山芳史，林　良彦（編）：日本語語彙大系：CD-ROM 版，岩波書店 (1999).
32) 竹本義美，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).
33) 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).

(平成 20 年 1 月 4 日受付)
(平成 20 年 7 月 1 日採録)

岩倉　友哉（正会員）

2003 年 3 月九州工業大学大学院情報工学研究科博士前期課程修了．同年株式会社富士通研究所入社．自然言語処理技術の研究開発に従事．

岡本　青史

1991 年九州大学大学院総合理工学研究科情報システム学専攻修士課程修了．1998 年学位（理学（博士））取得．株式会社富士通研究所勤務．北陸先端科学技術大学院大学客員教授．推論・機械学習・情報検索・知識発見の研究開発に従事．人工知能学会会員．