日本語固有表現抽出における文節情報の利用

†中野桂吾　　‡平井有三

†筑波大学大学院　システム情報工学研究科

‡筑波大学　電子・情報工学系

E-mail: †hyper@viplab.is.tsukuba.ac.jp　‡hirai@is.tsukuba.ac.jp

日本語固有表現抽出処理は，形態素解析などを用いて入力文を適当な解析単位（トークン）に分割し，対象となるトークンの前後 2 トークン程度の文脈長の品詞情報などを用いて固有表現部分のチャンキングを行うことが一般的である．しかしながら文脈長を固定してしまうと，固有表現の構成要素数が多い場合には十分な素性がチャンカーに与えられず，解析誤りが起こりやすくなる．そこで，本稿では文節区切りを行い，文節内の情報を素性としてチャンカーに与えることを提案する．提案手法では各文節の長さに応じて素性展開を行うため，文脈長を固定したモデルでは用いることのできなかった情報をチャンキングに利用することができる．CRL 固有表現データを用いた評価実験（5-fold-cross-validation）の結果，F 値約 0.89 という結果が得られ，提案手法の有効性を確認できた．

キーワード：固有表現抽出，Support Vector Machines，文節区切り，チャンキング

Japanese Named Entity Extraction Using Busetsu Feature

†Keigo Nakano　　‡Yuzo Hirai

†Doctoral Program Systems & Information Engineering, University of Tsukuba

‡Institute of Information Science and Electronics, University of Tsukuba

E-mail: †hyper@viplab.is.tsukuba.ac.jp　‡hirai@is.tsukuba.ac.jp

In Japanese Named Entity (NE) extraction, the first step is to segment a sentence into a sequence of appropriate analytical tokens. Then NE chunking is applied using contextual information such as lexical or POS features obtained through a fixed-size window. For longer entities, however, the chunker may not be given sufficient information because of shorter and fixed-size window. To cope with this problem, we propose a method which makes use of Japanese base phrase, called *bunsetsu* being extracted by *bunsetsu* segmentation, as features for chunking. Since a *bunsetsu* feature is effective for whole length of each phrase, the proposed method will provide more flexible chunking information to the chunker than the fixed-size window methods. We evaluated our method on CRL named entity data and obtained 0.89 F-value for 5-fold cross validation test, which shows the effectiveness of our method.

Keywords：named entity extraction, Support Vector Machines, bunsetsu segmentation, Chunking

１　はじめに

テキスト中から組織名や人名を抽出する固有表現抽出（Named Entity Recognition）は，情報抽出や情報検索における重要な要素技術である．英語においては Message Understanding Conference (MUC-7)，日本語においては Information Retrieval and Extraction Exercise (IREX) などが開催され，共通のデータを用いて様々な手法が比較されてきた[1].

固有表現抽出は入力文を適当な解析単位（トークン）に分割し，その単位に基づき固有表現部分をまとめあげるという手法が一般的である．Asahara ら[3]は，トークンの単位として単語ではなく，文字を用いることによって既存の手法よりも高い抽出精度が得られることを示した．しかし，彼らの手法では，該当文字の前後 2 文字程度の品詞情報などを用いてまとめあげを行うため，固有表現を構成する単語の数が多くなるにつれて正確に抽出するのが困難になるという問題がある．

そこで本稿では，固有表現の構成要素数が多い場合においても有効な素性を利用できる手法を提案する．提案手法では，形態素解析だけでなく文節区切りまで行い，文節内の情報を固有表現抽出のための素性として利用する．チャンキングには山田ら[10]や Asahara ら[3]が採用している Support Vector Machines（以下 SVM）に基づくチャンカー yamcha を用いた．先行研究と比較した結果，高い精度を達成することができた．

本論文の構成は以下の通りである．2 章で提案手法のベースとなる従来の SVM に基づく固有表現抽出について解説する．3 章では従来法の問題点を指摘し，その解決法について述べる．4 章で実験によって提案手法の有効性を検証する．最後に 5 章でまとめと今後の課題について述べる．

２．SVM に基づく固有表現抽出

本節では，提案手法のベースとなる Asahara ら[3]の提案したトークンの単位を文字とした固有表現抽出手法について解説する．

２．１　基本的なアプローチ

固有表現抽出は入力文を適当な解析単位（トークン）に分割し，固有表現を構成する 1 つもしくは複数のトークンをまとめあげると同時に，まとめあげられたトークン列がどの種類の固有表現なのかを同定するチャンキングと呼ばれるタスクとみなすことができる．各トークンの状態を表すために IOB1, IOB2, IOE1, IOE2 および S/E と呼ばれるタグ集合を用いる手法が提案されている．図 1 に「小泉首相は日米首脳会談において…」という文章に対し文字単位に固有表現タグを付与した例を示す．ここで B は固有表現の開始位置，E は終了位置，I は固有表現内，S は一つのトークンからなる固有表現，O タグは固有表現以外を示すタグである．これらの手法を用いて固有表現タグを定義することにより，固有表現のまとめあげの規則の学習は，入力文中の各トークンに固有表現タグを付与する分類規則の学習として扱うことが可能になる．なお固有表現タグには先行研究[3]で用いられている IOB2 モデルのみを用いた．

２．２　文字単位の素性展開

入力文を文字単位に分割し，各文字に対し，形態素解析によって得られる単語情報や品詞情報と，その文字の位置に応じた S/E タグを付与する．図 2 に例を示す．このように文字単位でチャンキングすることの利点は（1）形態素解析による単語の境界と固有表現の境界の不一致の問題を解消できる（2）文字そのものを素性として使うため単語そのものを用いるよりも粒度の細かい情報を用いることができる，という点が挙げられる．

２．３　SVM を用いたチャンキング

SVM [2] は多くの自然言語のタスクで利用され，その有用性が示されている．固有表現抽出においても SVM が適用され高い精度を示している．詳細は文献[10]を参照のこと．

SVM は正負を分離する超平面をマージンが最も大きくなるように求める 2 値分類器である．固有表現抽出のようにクラス数が 3 以上ある場合には多値分類に拡張する必要があり，拡張法としては

	小	泉	首	相	は	日	米	首	脳 …
IOB1	I-PERSON	I-PERSON	O	O	O	I-LOCATION	B-LOCATION	O	O
IOB2	B-PERSON	I-PERSON	O	O	O	B-LOCATION	B-LOCATION	O	O
IOE1	I-PERSON	I-PERSON	O	O	O	E-LOCATION	E-LOCATION	O	O
IOE2	I-PERSON	E-PEROSN	O	O	O	E-LOCATION	E-LOCATION	O	O
S/E	B-PERSON	E-PERSON	O	O	O	S-LOCATION	S-LOCATOIN	O	O

図 1：固有表現タグの例

位置	文字	単語	品詞	文字種	固有表現タグ
i−2	茨	茨城-B	名詞-固有名詞-地域-一般-B	OTHER	B-LOCATION
i−1	城	茨城-E	名詞-固有名詞-地域-一般-E	OTHER	I-LOCATION
i	県	県内-B	名詞-一般-B	OTHER	I-LOCATION
i+1	内	県内-E	名詞-一般-E	OTHER	
i+2	の	の-S	助詞-格助詞-一般-S	HIRA	

図 2：チャンキングのための素性展開

one versus rest 方式と pairwise 方式が代表的である．one versus rest 方式では k 個のクラスに対し，あるクラスかそうでないかを分類する二値分類器を k 個作成する手法である．一方 pairwise 方式は任意の 2 つのクラスに関する二値分類器を ${}_kC_2$ 個作成する手法である．本研究では先行研究において良い精度を示している one versus rest 方式を用いた．

チャンカーは 2.2 節で示した素性を SVM の入力とし，その位置における固有表現タグを推定する．図 2 では i 番目の固有表現タグを推定するために太線で囲まれた部分の素性を用いている．ここで i – 1 番目，i – 2 番目の固有表現タグは解析時には未知であるため，各位置で推定した固有表現タグを用いて決定論的に解析する．文頭から文末へ解析するか（右向き解析），文末から文頭へ解析するか（左向き解析）で精度が異なり，日本語固有表現抽出においては接尾辞が重要な役割を果たすために左向き解析が有効であることが知られている [3][10].

3　提案手法

3．1　従来手法の問題点

先行研究の多くは固定長の文脈情報を用いているが，この手法では推定に必要な情報がチャンカーに与えられない場合がある．図 3 に「会」が固有表現であるか否かを判定する二つの例文を示した．どちらの例文においても同じ素性（太線で囲まれた部分）がチャンカーに与えられてしまうので，同じクラスに判断されてしまうが，実際には「同委員会」に含まれる「会」は固有表現ではない．この例においては太線内のウインドウにある単語や文字に関する情報をいくら付与しても正しく固有表現を抽出することはできない．抽出に必要な文脈長は固有表現ごとに異なり，文脈長を必要以上に長くすると過学習が起きやすくなるため，結果として抽出精度が下がってしまう．この問題に対処するためには，ウインドウ外の情報を文脈に応じて適切に利用する必要がある．

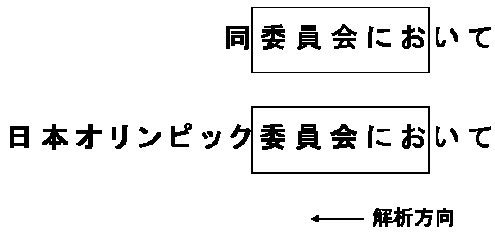

図 3：従来法の問題点

3．2　提案手法の概要

仮に前もって固有表現の長さが分かっていれば，固有表現の長さに応じた情報を用いることによって抽出精度の向上が期待できる．しかしながら抽出時には現在位置のトークンがどのような長さの

文字	品詞	文節内素性	固有表現タグ
日	名詞-固有名詞-地域-国	*	B-ORGANIZATION
本			
オ	名詞-一般	名詞-固有名詞-地域-国	I-ORGANIZATION
リ			
ン			
ピ			
ッ			
ク			
委	名詞-一般	名詞-固有名詞-地域-国	I-ORGANIZATION
員			
会	名詞-接尾-一般	名詞-固有名詞-地域-国	I-ORGANIZATION

図 4：文節内素性が有効な例

文字	品詞	隣接文節素性	固有表現タグ
加	名詞-一般	*	O
工			
会	名詞-一般	*	O
社			
「	記号－括弧開	*	O
松	名詞-固有名詞-地域－一般	会社	B-ORGANIZATION
原			
シ	名詞－固有名詞－一般	会社	I-ORGANIZATION
ヤ			
ー			
リ			
ン			
グ			
」	記号－括弧閉	*	O
の	助詞－連体化	*	O

図 5：隣接文節素性の例

固有表現を構成するか，あるいは固有表現を構成しないのかは未知である．そこで提案手法では文節区切りを用いることによって，各文節の長さに応じた素性展開を行う．以下，文節を考慮した素性を文節素性と呼ぶ．

提案手法は，

（1）入力文に対し，形態素解析及び文節区切りを適用

（2）各文字が属する単語や品詞情報に加え文節素性を展開

（3）従来法と同様に固定長文脈でチャンキングを行う

というステップから成る．提案手法では文節の長さに応じて文脈長を変えるのではなく，文節素性を展開することによって仮想的に可変長モデルを扱うことができる．

3．3　文節素性の展開

ほとんどの固有表現は名詞から構成されるので，文節素性を付与するのは名詞のみに限った．以下に実験で用いた文節素性について解説する．

文節素性を抽出する方法に，文頭から解析して行く方法と，文末から解析する方法がある．解析方向が異なれば文節素性の展開の仕方も変わるが，本稿では文末から解析する左向き解析の場合のみを考慮した．本稿で用いた文節素性は以下の 3 種類である．

- **文節内素性**：文節内で解析方向に向かって固有名詞が存在すれば，最も近い固有名詞の品詞細分類を，固有名詞がなければ文節の先頭の単語を素性として用いる．構成要素数の多い固有表現は，固有表現の内部に地名や組織名などを含むことが多い．図 4 に示した例のように，文節内の固有名詞「日本」以下の名詞に対して，「固有名詞-地名 - 国名」が文節内素性として付与される．文節内素性を用いることによって，従来法ではできなかった「日本オリンピック委員会において」と「同委員会において」の区別ができるようになる．

- **隣接文節素性**：解析方向に隣接する文節の末尾が名詞である場合に，その単語を素性として用いる．一般的に文節は自立語と付属語から成るが，文節が名詞で区切られている場合には何らかの重要な情報が含まれていると考えられる．隣接文節素性が有効である例を図 5 に示す．図中の太線は文節の区切り位置を示している．この例では，「シヤーリング」が組織名の一部であることを，隣接文節素性「会社」によって推定することができる．

- **主辞素性**：各文節の主辞を素性とする．例えば「アントノフ/ソ連/軍/参謀/総長」（/は形態素の区切り位置）において，主辞素性「総長」を用いることによって「アントノフ」を人名として正しく抽出できるといったことが期待できる．

表 3：固有表現ごとの精度の比較（F 値）

	頻度	base	model A	model B	model C
ORGANIZATION	3676	80.46	84.12	**84.31**	84.30
PERSON	3840	87.88	88.88	89.08	**89.16**
LOCATION	5463	88.34	89.85	**90.13**	89.91
ARTIFACT	747	50.66	51.25	**52.90**	52.20
DATE	3567	94.55	94.71	**94.73**	94.68
TIME	502	89.29	89.49	**91.45**	91.24
MONEY	390	93.59	94.06	93.50	**94.43**
PERCENT	492	96.87	96.79	97.08	**97.16**
TOTAL	18677	87.07	88.5	**88.78**	88.72

表 1：IREX による固有表現のクラスと例

固有表現のクラス		例
ORGANIZATION	組織名	共和党
PERSON	人名	小泉純一郎
LOCATION	地名	日本、アメリカ
ARTIFACT	固有物名	ノーベル賞
DATE	日付表現	４月２７日
TIME	時間表現	午後五時
MONEY	金額表現	５００万円
PERCENT	割合表現	二十%、三割

表 2：使用する文字種

文字種	説明
HIRA	ひらがな
KATA	カタカナ
NUM	アラビア数字
ALP_L	アルファベット小文字
ALP_U	アルファベット大文字
OTHER	その他

４． 実験

４．１ 実験データ及び実験設定

実験には CRL 固有表現データを使用した．CRL 固有表現データは毎日新聞９５年度版 1,174 記事，約 11,000 文に対して IREX で定義された固有表現が付与されている．表 1 に IREX で定義された固有表現の種類を示す．評価には CRL 固有表現データを記事単位に 5 等分し，訓練データ４，評価データ 1 の比率で交差検定を行い，それらの平均の適合率，再現率，F 値（$\beta = 1$）で各モデルの比較を行った．

形態素解析器には Chasen[9]を使用し，磯崎らの手法[5]と同様に未知語を「固有名詞-一般」として出力されるようにした．文節区切りには構文解析器 Cabocha[1]を，チャンキングには SVM に基づくチャンカー yamcha [2]を使用した．チャンキングの解析方向は全て文末から文頭方向へ解析する左向き解析で行い，文脈長はすべて対象文字の前後 2 文字に固定した． SVM のカーネル関数には 2 次の多項式カーネルを使用した．

４．２ 文節素性の効果

文節素性が抽出精度に与える影響を調べるために，以下に示す 4 種類の場合について比較した．

base model：文字，単語，品詞，文字種，前固有表現タグ（表 2 と同じ構成）

model A：base model の素性＋文節内素性

model B：base model の素性＋文節内素性＋隣接文節素性

model C：base model の素性＋文節内素性＋隣接文節素性＋主辞素性

ここで文字種は表 2 に示す 6 種類を用いた．

４．２．１ 固有表現の種類による精度の比較

表 3 に固有表現の種類ごとの精度および全体の精度を示す．表より，文節素性を用いることによって精度が向上していることが分かる．文節内素

[1] http://cl.aist-nara.ac.jp/~taku-ku/software/cabocha/

[2] http://cl.aist-nara.ac.jp/~taku-ku/software/yamcha/

表 4：固有表現の構成要素数による精度の比較

構成する形態素数		n ≧ 1	n = 1	n =2	n =3	n ≧ 4
頻度		18677	10349	5379	1689	1260
base model	F-value	87.09	88.81	89.51	76.51	74.31
	recall	85.44	**89.40**	87.21	69.92	66.11
	precision	88.81	88.23	91.94	84.48	84.83
model C	F-value	**88.72**	**89.86**	**91.17**	**80.58**	**79.00**
	recall	**86.65**	89.07	**88.18**	**75.67**	**73.41**
	precision	**91.02**	**90.66**	**94.37**	**86.18**	**85.49**

性が最も精度向上に貢献し，主辞素性は精度にほとんど影響を及ぼさなかった．

固有表現のクラス別に見ると，固有名詞的表現，特に組織名において精度が向上していることが分かる．組織名は接尾辞が長い（「研究所」，「委員会」など）ので，文節素性が有効に働いたと考えられる．一方数値的表現に対しては文節素性の効果は低かった．数値的表現は接尾辞（～円，～ドル，～時など）が抽出に対し決定的な役割を果たすので，文節素性の影響が小さかったと考えられる．

４．２．２　構成要素数の違いによる精度の比較

次に固有表現の構成要素数によって抽出精度がどのように変化するかを比較した．表 4 に固有表現を構成する形態素数による抽出精度の増減を示す．表中のｎは固有表現を構成する形態素数を示している．[3]

F 値で比較すると，ｎが多くなるにしたがって base model との精度の差が大きくなっていることから，提案手法が構成要素数の多い固有表現に対して有効であることが分かる．適合率と再現率での比較では，適合率に関してはｎが大きくなるにつれて差は小さくなり，逆に再現率に関してはｎが大きくなるにつれて差が大きくなっている．

長い固有表現を抽出できなかった場合，そのうちの一部を固有表現として過抽出してしまう（例：「日本オリンピック委員会」というフレーズに対し「日本」のみを地名として抽出しまう）．文節素性によって構成要素数の多い固有表現を抽出できるようになったことが，構成要素数の少ない場合において適合率が向上した理由であると考えられる．

４．３　意味素性の利用

先行研究[３]では意味素性を用いることによって精度が向上することを報告している．そこで model C において，文節素性で用いる単語（各文字が属する単語および文節素性として用いる単語）を日本語語彙体系[4]における意味素性に置き換えてモデルを学習した．結果を表 5 に示す．表より意味素性を用いることによって精度が若干向上したことが分かる．今回の実験においては，複数の意味素性を持つ単語に対しては意味素性を全て併記している．そのため意味素性の曖昧性を解消することによってさらなる精度向上が期待できる．

表 5：意味素性の利用と精度（F 値）

	model C	意味素性
ORGANIZATION	84.30	**84.69**
PERSON	89.16	**89.27**
LOCATION	89.91	**90.37**
ARTIFACT	52.20	**53.85**
DATE	94.68	**94.80**
TIME	91.23	**91.85**
MONEY	94.43	**95.48**
PRECENT	**97.16**	97.07
TOTAL	88.72	**89.03**

[3] 固有表現そのものを形態素解析しているので，実際に文章中で用いられている場合の構成要素数とは若干異なる．

表 6：先行研究との比較（F 値）

	ＣＲＬ 公開データ	IREX GENERAL	学習モデル	文脈長	わかち書き問題への対処	シソーラスの利用
内元 00 [6]		80.17	ME	±2	書き換え規則	無
山田 02 [10]	83.2		SVM	±2		無
竹本 01 [8]		83.86	辞書＋規則		複合語分割辞書	無
宇津呂 02 [7]		84.07*	ME+決定リスト	一部可変長		無
磯崎 02 [5]	86.77	85.10	SVM＋sigmoid	±2	書き換え規則	無
Asahara 03 [3]	87.21		SVM	±2	解析単位を文字	有
提案手法	**89.03**		SVM	±2，文節素性	解析単位を文字	有

*分かち書きと固有表現の境界が一致しない場合を除いた精度

４．４　考察

表 6 に先行研究との比較結果を示す．提案手法は我々の知る限り最も高い精度を達成している．

機械学習に基づく先行研究のほとんどは文脈長を固定したモデルを用いている．宇津呂ら[7]は現在位置のトークンがいくつのトークンから構成される固有表現の一部であるかを考慮して学習する可変長モデルを提案している．しかしながら宇津呂らの手法は，モデルの適用時には固定長モデルとして適用するので，学習時と適用時で考慮する素性集合が異なり，単独のモデルでは固定長モデルに比べて精度が高くないことを報告している．それに対し提案手法は，固定長のモデルを用いてはいるが，文節素性によって文脈に応じた素性展開を行う．また，単一のモデルでも高い精度を示すことができた．

文節素性を導入してもうまく抽出できなかったものとしては，「子どもの権利条約」など固有表現中に付属語や活用語を含むものが挙げられる．そのため，2 つ以上の文節から成る固有表現に対しても有効な手法の提案が必要であると考えられる．

５　おわりに

本稿では日本語固有表現抽出タスクに対し，文節素性を用いることを提案し，その有効性を示した．提案手法は各文節の長さに応じて適切な素性展開を行い，従来法では扱うことができなかった素性をチャンカーに与えることができる．実験の結果，SVM に基づく固有表現抽出システムについて今まで報告されているものの中で，最高の精度が得られた．

提案手法は専門用語抽出などの自然言語処理の他のタスクにも応用可能であると考えられる．今後の課題としては他の言語や他のタスクに提案手法を適用し，その有効性を検証していきたいと考えている．

参考文献

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

[2] V.N. Vapnic. *Statistical Learning Theory*. A Wiley-Interscience Publication, 1998.

[3] Masayuki Asahara and Yuji Matsumoto, Japanese Named Entity Extraction with Redundant Morphological Analysis. In *Proceedings of HLT-NAACL 2003.*

[4] NTT コミュニケーション科学基礎研究所　監修．日本語語彙大系．岩波書店，1997.

[5] 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, Vol44, No.3, pp.970-979.

[6] 内元清貴，馬青，村田真樹，小作浩美，内山将夫，井佐原均．最大エントロピー法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.

[7] 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，pp65-100，2002.
[8] 武本義美，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.
[9] 松本裕治，北内啓，山下達雄，平野善隆，松田寛，高岡一馬，浅原正幸．形態素解析システム『茶筌』 version 2.2.9 使用説明書．奈良先端科学技術大学院大学，2002.
[10] 山田寛泰, 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，Vol43，No.1，pp.44－43，2002.