日本語固有表現抽出におけるわかち書き問題の解決

浅原正幸[†] 松本裕治[†]

一般的に日本語固有表現抽出で提案されている手法は形態素解析とチャンキングの組合せによる．形態素解析出力結果をそのままチャンカの入力にすると，形態素解析結果より小さい単位の固有表現を抽出することは困難である．そこで，文字単位でチャンキングを行う手法を提案する．まず，統計的形態素解析器で入力文を冗長的に解析を行う．次に，入力文を文字単位に分割し，文字，字種および形態素解析結果の n 次解までの品詞情報などを各文字に付与する．最後に，これらを素性として，サポートベクトルマシンに基づいたチャンカにより決定的に固有表現となる語の語境界を推定する．CRL 固有表現データを用いて評価実験（交差検定 5-fold）を行った結果，F 値 0.87 という高精度の結果が得られた．

A Word Unit Problem in Japanese Named Entity Extraction

Masayuki Asahara[†] and Yuji Matsumoto[†]

Named Entity (NE) extraction is a task in which proper nouns and numerical information are extracted from texts. A method of cascading morphological analysis and chunking is usually used for NE extraction in Japanese. However, such a method cannot extract smaller NE units than morphological analyzer outputs. To cope with the unit problem, we propose a character-based chunking method. Firstly, input sentences are redundantly analyzed by a statistical analyzer. Secondly, the input sentences are segmented into characters. The characters are annotated with the character types and POS tags of the top n-best answers that are given by the statistical morphological analyzer. Finally, we do chunking deterministically based on support vector machines. We apply our method to IREX NE task using CRL Named Entities data. The cross validation result of the F-value being 0.87 shows the effectiveness of the method.

1. はじめに

固有表現抽出は，地名・人名・組織名などの固有名詞や日時・時間・通貨などの数値表現をテキスト中から切り出し分類する技術である．情報抽出や質問応答システムなどの基礎技術だけでなく，形態素解析や構文解析などにも影響を及ぼすために，重要な問題の 1 つである．また，英語における Message Understanding Conference（MUC-7）や日本語における Information Retrieval and Extraction Exercise（IREX）などで共通のデータセットが公開されて，多くの研究者が様々なモデルを提案し，この問題に取り組んできた．

一般的に固有表現抽出は，形態素解析をまず行い，前後 2 単語程度の品詞情報などを用いることにより，形態素解析結果の語の単位を基にしてまとめあげるという作業が行われる．しかし，この手法のままでは，形態素解析結果より短い単位の固有表現を抽出することが困難である．たとえば「小泉首相が 9 月に訪朝」という文について形態素解析を行うと「小泉/首相/が/9 月/に/訪朝」のように分割される．人名である「小泉」および日時である「9 月」は，このわかち書き単位から抽出が可能であるが，国名を表す「朝」（＝北朝鮮・朝鮮民主主義人民共和国の略称）は形態素解析のわかち書き単位より小さいために抽出が不可能である．本稿では，このような固有表現抽出における語のわかち書きの単位の問題にも対処する．

このわかち書きの問題に対し，先行研究は様々な前処理あるいは後処理により対処している．内元ら[4]は，このわかち書きの問題に対して，書き換え規則を導入し，わかち書きをしなおす手法をとっている．山田ら[10]は，学習データ中に出現したものについては，分割した単位で抽出している．これらの別処理による対処方法に対して，我々はより直接的な対処方法を提案する．提案手法ではテキストを文字単位に分割し，文字単位でチャンキングを行う．各文字により豊かな品詞情報

† 奈良先端科学技術大学院大学情報科学研究科
Graduate School of Information Science, Nara Institute of Science and Technology

	小	泉	首	相	は	日	朝	間	…
IOB1	I-PERSON	I-PERSON	O	O	O	I-LOCATION	B-LOCATION	O	
IOB2	B-PERSON	I-PERSON	O	O	O	B-LOCATION	B-LOCATION	O	
IOE1	I-PERSON	I-PERSON	O	O	O	E-LOCATION	I-LOCATION	O	
IOE2	I-PERSON	E-PERSON	O	O	O	E-LOCATION	E-LOCATION	O	
SE	B-PERSON	E-PERSON	O	O	O	S-LOCATION	S-LOCATION	O	

図 1 固有表現タグの例
Fig. 1 Named entity tag.

表 1 IREX で定義されている固有表現の種類と例
Table 1 Examples of named entity in IREX.

固有表現の種類		例
ARTIFACT	固有物名	ノーベル化学賞
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	二〇%，三割
PERSON	人名	村山
TIME	時間表現	午前五時

を与えるために，冗長的な形態素解析を導入し，n 次解までの品詞情報をチャンカの素性として利用する．チャンキングには山田ら[10] が採用している Support Vector Machine に基づくチャンカ yamcha[6] を利用し，既存の手法を上回る解析精度を達成した．

以下，次章では IREX 日本語固有表現抽出タスクについて述べる．3 章では，今回提案する手法の詳細について説明する．4 章で，抽出実験と考察を報告し，最後にまとめと今後の課題について述べる．

2. IREX 日本語固有表現抽出タスク

IREX 日本語固有表現抽出タスク[1] では，**表 1** に示す 8 種類の固有表現を定義し，それぞれの固有表現は重ならないとしている．固有表現抽出は，入力文中の単語列が固有表現か否かを識別するチャンク同定問題と見なすことができ，チャンク同定問題では 1 つ以上の要素列からなるチャンクを導出するために，トークン列にチャンクの開始位置や終了位置を示すチャンクタグを付与することによって行われる．

チャンクタグ集合として IOB1，IOB2，IOE1，IOE2 および SE と呼ばれる 5 種類が提案されている．チャンクタグ I はチャンクの内部，チャンクタグ B はチャンク開始位置，チャンクタグ E はチャンク終了位置，チャンクタグ O はチャンク外を表す．IOB1 ではチャンクタグ I，O，B を用いるが，チャンクタグ B はチャンクが連続する際のチャンク境界におけるチャンク開始位置にのみ付与する．IOB2 ではチャンクタグ I，O，B を用いるが，チャンクタグ B はすべてのチャンク開始位置に付与する．同様に，IOE1 ではチャンクタグ I，O，E を用いるが，チャンクタグ E はチャンクが連続する際のチャンク境界におけるチャンク終了位置にのみ付与する．IOE2 ではチャンクタグ I，O，E を用いるが，チャンクタグ E はすべてのチャンク終了位置に付与する．最後に，SE では，チャンクタグ I，O，B，E と 1 トークンでチャンクになる S を用いる．なお，SE では，1 トークンでチャンクでならない場合，すべてのチャンク開始位置に B を，すべてのチャンク終了位置に E を付与する．

トークンの単位としては形態素解析で切り出された単語を用いる場合が多いが，我々はこれを文字単位で用いる．**図 1** に文字単位でタグ付けした「小泉首相は日朝間における…」の文中の固有表現の例を示す．この文中では「小泉」：人名（PERSON），「日」「朝」：地名（LOCATION）が固有表現であるが，図中では IOB1，IOB2，IOE1，IOE2 および SE のチャンクタグ集合に基づいた固有表現タグを付与している．なお，ここで固有表現タグとは “I-PERSON” のように固有表現名とチャンクタグを “-”（ハイフン）で結んだものを指す．

3. 提 案 手 法

本章では提案手法について述べる．提案手法は以下の 3 ステップによる．

(1) 冗長的に形態素解析を行う．
(2) 文字単位に分割し，各文字が属する形態素の情報と，その形態素中における文字の位置情報を付与する．
(3) 文字に付与された情報を手がかりに，文字単位にまとめあげを行う．

以下，各ステップについて説明する．

3.1 冗長的な形態素解析

本手法で用いる日本語形態素解析はマルコフモデルに基づく．形態素解析は入力文 S の単語列 W に対する品詞タグ列 T を決定することと定義できる．目標は次の確率値を最大にするような品詞タグ列 T を発見することである．日本語や中国語の場合には，入力が文字列となるため，可能な単語列をすべて展開したうえで品詞列同定と単語列同定を同時に行うことに

位置	文字	字種	品詞情報（1 次解）	品詞情報（2 次解）	品詞情報（3 次解）	固有表現タグ
$i-2$	小	OTHER	名詞-固有名詞-人名-姓-B	接頭詞-名詞接続-S	名詞-一般-S	B-PERSON
$i-1$	泉	OTHER	名詞-固有名詞-人名-姓-E	名詞-固有名詞-地域-一般-E	名詞-固有名詞-一般-E	I-PERSON
i	首	OTHER	名詞-一般-B	名詞-一般-S	名詞-接尾-助数詞-S	O
$i+1$	相	OTHER	名詞-一般-E	名詞-接尾-一般-S	*	
$i+2$	が	HIRAG	助詞-格助詞-一般-S	*	*	

図 2 導入する素性
Fig. 2 An example of extracted feature.

表 2 字種の分類
Table 2 Character types.

字種タグ	説明
ZSPACE	空白
ZDIGIT	アラビア数字
ZLLET	英字小文字
ZULET	英字大文字
HIRAG	ひらがな
KATAK	カタカナ
OTHER	その他

なる．

$$T = arg\max_{T} P(T|W).$$

ベイズの定理を利用して，$P(W,T)$ は品詞タグ列の生起確率と単語列の生起確率として展開することができる．

$$arg\max_{T} P(T|W) = arg\max_{T} P(W|T)P(T).$$

単語生起確率はその品詞タグからのみに，品詞タグ生起確率は bi-gram モデルのみに制限して近似をする．これらの確率値はタグ付きコーパスの頻度から最尤推定される．推定されたパラメータを利用して，動的計画法の一種である Viterbi algorithm により，単語列 W に対して出現確率最大の品詞タグ列 T を決定する．実際の計算には確率の対数をとり，コストに変換して，可能な単語/品詞列からコスト和が最小になるようなものを選ぶことにより解析を行う．

本手法で用いる冗長解析は，最適解から設定したコスト幅の差以内の n 次解を出力することによる．各文字位置において，その文字を含む文頭からのコスト和が小さい順に n 次解として形態素を出力する．なお，未知語が出現した場合，不要な短い n 次解が多く出現しチャンク解析誤りを引き起こすことが考えられる．そこで，コスト和が設定したコスト幅を越えて異なる場合には，その解を出力しない．本手法ではコスト幅として，確率モデルを推定する際，既知語のみが出現した場合に最低確率となる事象に割り当てられるコストを用いた．

実際の解析には，形態素解析器『茶筌』および ipadic-2.4.4 を用いた．上記冗長解析は，茶筌解析実行時にオプション `-v -w 4000` を付けた結果を文字単位に分割することにより行うことができる．なお，辞書に対して，外部の固有表現辞書を追加するというようなことは行っていない．

3.2 チャンカのための素性展開

冗長的な形態素解析により認定された形態素の情報を文字単位に分割したものをチャンカの素性として導入する．品詞情報とともに，各形態素における，当該文字の位置について，チャンクタグ集合 SE に基づいたタグを付与する．

これらの品詞情報のほかに，字種，文字などを手がかりとしてチャンキングを行う．字種は**表 2** に示す 7 種類を導入する．

図 2 に前後 2 文字を用いた場合の導入する素性を示す．本手法では，異なる文字位置の事象は異なる素性とする．さらに，異なる n 次解（1 次解，2 次解，3 次解）で生起する事象は異なる素性とした．これは予備実験において，訓練データ量が少ない場合には異なる n 次解で生起する事象をグループ化した方が高い精度が得られたが，今回利用するデータ量の場合には異なる n 次解を異なる素性とした方が高い精度が得られたことによる．

3.3 サポートベクトルマシンを用いたチャンキング

チャンキングにはサポートベクトルマシンを基にしたチャンカ yamcha [6] を利用した．以下にサポートベクトルマシンを用いたチャンキングについて述べる．詳細は文献 10) を参照のこと．

サポートベクトルマシンは，素性ベクトル $\mathbf{x}_t$ と正・負の二値ラベル y_t の二つ組 $(\mathbf{x}_t, y_t)$ で表現される訓練事例に対して，正・負のラベルを正しく分離するような超平面 $\mathbf{W}\cdot\mathbf{\Phi}(\mathbf{x})+b$（ただし $\mathbf{W}$，$\mathbf{\Phi}(\mathbf{x}) \in R^n$）を求める二値線形分類器[2] である．正・負例を正しく分類する数多くの超平面の中から，分離超平面とそれに最も近い事例間との距離（マージン）が最大となるようなものを求めることによりモデルを作成する．未知の事例 $\mathbf{x}$ に対する正・負例の分類は，求められた超平面からの位置によって決定される．

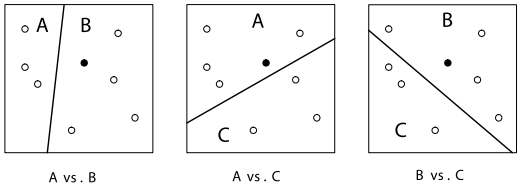

図 3 Pairwise 法
Fig. 3 Pairwise method.

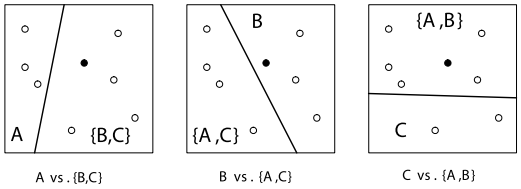

図 4 One vs. Rest 法
Fig. 4 One vs. Rest method.

$$f(\mathbf{x}) = sign(\mathbf{W} \cdot \mathbf{\Phi}(\mathbf{x}) + b)$$
$$= sign(\sum_{\mathbf{x_i} \in SV} \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}) + b)$$

ここで $K(\mathbf{x}_i, \mathbf{x})$ を Kernel 関数と呼ぶ．Kernel 関数を導入することにより，素性ベクトルをより高次元の空間に写像することができ，線形分離不可能な場合も扱うことができるようになる．本手法では d 次の多項式関数 $K(\mathbf{x}_i, \mathbf{x}) = (\mathbf{x}_i \cdot \mathbf{x} + 1)^d$ を Kernel 関数として利用した．これにより 2 個までの素性の組合せを考慮した学習が可能になる．

サポートベクトルマシンは正例・負例を分類する二値分類器であり，チャンキング抽出規則を学習するために 17 クラス☆に分類する多値分類に拡張する．代表的な手法として Pairwise 法と One vs. Rest 法がある．Pairwise 法は k 個のクラスから任意の 2 つのクラスに関する二値分類器を ${}_kC_2$ 個構築する手法である．これに対し，One vs. Rest 法は，あるクラスかそれ以外かという二値分類器をクラスの数だけ用意する手法である．以下，A，B，C の 3 クラスの問題に対し，例を用いて詳しく説明する．Pairwise 法の場合，**図 3** のように，A vs. B，A vs. C，B vs. C の 3 つの二値分類器を作成し，これらの結果の多数決により決定する．図中の黒い点は，B に分類される．One vs. Rest 法の場合，**図 4** のように，A vs. {B，C}，B vs. {A，C}，C vs. {B，C} の 3 つの二値分類器を作成し，これらの結果により決定する．図中の黒い点は，B に分類される．2 つ以上のクラスが選ばれた場合，分離超平面から最も離れた分類器の結果を用いる．Pairwise 法の方が学習データを効率的に利用できる特性がある一方，One vs. Rest 法では付与するタグごとに与える素性を変更できる特性がある．

☆ {I, B} もしくは {I, E} × 固有表現 8 種+ {O} の 17 クラス．

文字	1 次解	2 次解	固有表現
日	名詞-一般	名詞-固有名詞-地域-国	LOCATION
本			
人		名詞-接尾-一般	

図 5 n 次解の有効な例（1）
Fig. 5 Effect of n-best answers (1).

チャンキングは 3.2 節で示した素性をサポートベクトルマシンに与え，その出力クラスを基に文頭もしくは文末から 1 方向に決定的に行われる．すでに決定されたチャンクタグも素性に用いるために，チャンキングの解析方向が順方向（左→右）の場合と逆方向（左←右）の場合とで解析結果が異なる．図 2 に前後 2 文字文脈，字種，文字および冗長形態素解析結果 3 次解までの品詞を用いた場合に利用される素性を示す．この例では，位置 i における固有表現タグ O を推定するために，実線の内部にあるものすべてを独立した素性として利用する．

3.4 n 次解を用いることによる効果

文字単位にチャンキングを行うことにより，わかち書きの問題を対処することができる．さらに，n 次解を用いることにより日本語形態素解析器の振舞いをチャンキングに有効利用することが可能である．わかち書きの問題は形態素解析器の辞書に長い複合語が含まれている場合に起こる．長い複合語の品詞が固有表現抽出に有効でなくても，複合語を構成するより短い単位の語の品詞が固有表現抽出に有効な場合がある．1 次解のみを利用する場合この短い単位の語の素性が捨てられてしまうが，n 次解を用いることによりこの素性を救えることができる．**図 5** に例を示す．この例では，2 次解に付与された品詞情報から「日本」が LOCATION であると抽出することができる．

また，n 次解を利用することにより未知語の問題をも解決できると考える．未知語が出現すると形態素解析器の際に未知語出現位置でマルコフモデルに用いる文脈素性が途切れてしまう．よって未知語出現位置前後で品詞推定誤りすることが多い．しかしながら，n 次解には正しい品詞が出現する．**図 6** に未知語問題を解決できた例を示す．この例では，1 文字目の 1 次解の品詞「名詞-固有名詞-人名-姓」によって，人名の開始点を認識し，5 文字目の 2 次解の「名詞-接尾-一般」によって，4 文字目が人名の終了点を認識することができたと考える．

表 3 文脈長の違いによる精度の比較
Table 3 The length and contextual feature and the extraction accuracy.

Pairwise 法						
文脈長	左 1 右 1		左 2 右 2		左 3 右 3	
解析方向	順	逆	順	逆	順	逆
ARTIFACT	29.74	46.36	42.17	**48.30**	43.90	46.36
DATE	84.98	90.33	91.16	**94.14**	92.47	93.72
LOCATION	80.16	86.17	84.07	**87.62**	85.75	87.18
MONEY	43.46	94.00	59.88	**95.82**	72.53	94.34
ORGANIZATION	66.06	74.73	72.63	78.79	75.55	**79.48**
PERCENT	67.66	**96.37**	83.77	96.31	85.26	94.14
PERSON	83.44	85.60	85.35	**87.31**	86.31	87.24
TIME	88.21	87.55	**89.82**	87.47	89.54	87.49
全体	76.60	83.72	81.91	**86.19**	83.82	86.02
One vs. Rest 法						
文脈長	左 1 右 1		左 2 右 2		左 3 右 3	
解析方向	順	逆	順	逆	順	逆
ARTIFACT	29.79	45.59	39.84	**49.58**	42.35	47.82
DATE	85.15	90.22	91.21	**93.97**	92.42	93.41
LOCATION	80.22	86.62	84.31	**87.75**	86.06	87.61
MONEY	43.43	93.30	61.85	**93.85**	75.01	93.60
ORGANIZATION	65.69	74.80	72.74	78.33	75.95	**79.95**
PERCENT	69.12	95.96	85.66	**96.06**	88.56	94.16
PERSON	83.63	84.98	85.51	87.19	86.57	**87.65**
TIME	88.42	87.54	**90.38**	88.33	89.85	88.08
全体	76.65	83.71	82.12	86.11	84.16	**86.33**

冗長解析結果は 3 次解まで利用，素性（品詞，文字，字種，前固有表現タグ）．
Kernel 関数は 2 次の多項式関数．文脈長は文字数．

文字	1 次解	2 次解	固有表現
池	名詞-固有名詞-人名-姓	名詞-一般	PERSON
坊			
専	未知語		
永	名詞-固有名詞-人名-姓	形容詞-自立	
家	名詞-一般	名詞-接尾-一般	

図 6 n 次解の有効な例（2）
Fig. 6 Effect of n-best answers (2).

4. 評価実験

4.1 データ

実験には CRL（通信総合研究所）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タグ付けされている．このデータ中の固有表現の総数は 19,262 個であった．評価は CRL 固有表現データを 5 等分に分割し，訓練 4，テスト 1 の比率で交差検定を行い，それらの F 値 ($\beta = 1$) の平均を精度比較に利用する．

以下の実験において解析に利用するチャンクタグは，予備実験で最も精度が高かった IOB2 モデルに固定した．

4.2 文脈長の違いによる精度の比較

まず，文脈長の違いによる精度比較を行う．**表 3** に文脈長を変化させた際の実験結果を示す．精度は 8 つのタグすべての解析精度（F 値）による．チャンキングの方向が順方向（左→右）の場合と逆方向（左←右）の場合を比べると逆方向の方が精度が高いといえる．特に接尾辞の情報が有効な TIME 以外の数値表現で顕著に逆方向が有効であった．TIME では「午前」「午後」などの接頭辞を過不足なく切り出すために順方向解析が必要であることが分かった．ORGANIZATION を除く固有表現では前後 2 文字を見ることにより精度が高いことが分かる．ORGANIZATION でより長い文脈が有効に働いているのは,「運輸省」に対する「運輸省鉄道局技術企画課」のような長い表現の部分文字列も固有表現になるようなものが多いからであると考える．

4.3 冗長解析結果の深さの違いによる精度の比較

次に素性として利用する冗長解析結果の深さを変化させた場合の精度比較を行う．この実験では，素性として利用する文脈を左 2 文字右 2 文字に，利用する素性を品詞，文字，字種，前固有表現タグの 4 種類に固定して行った．**表 4** に結果を示す．3.4 節で示したような事例が少ないために，こちらが狙いとする n 次解の導入による精度の向上は数値として現れなかった．

4.4 素性の違いによる精度の比較

必要な素性について検討する．**表 5** に文脈を左右 2 文字に固定したうえで素性を変化させた結果を示す．

表 4 冗長解析結果の深さの違いによる精度の比較
Table 4 The depth of redundant analysis and the extraction accuracy.

Pairwise 法								
冗長解析	1 次解のみ		2 次解まで		3 次解まで		4 次解まで	
解析方向	順	逆	順	逆	順	逆	順	逆
ARTIFACT	44.37	**49.76**	43.57	48.84	42.17	48.30	42.10	49.04
DATE	90.53	93.81	91.22	**94.23**	91.16	94.14	91.00	93.71
LOCATION	84.35	**87.67**	84.20	**87.67**	84.07	87.62	83.92	87.60
MONEY	59.45	93.89	60.36	94.28	59.88	95.82	60.94	**95.96**
ORGANIZATION	73.83	79.12	73.71	**79.34**	72.63	78.79	72.46	78.39
PERCENT	84.44	**97.20**	84.87	96.76	83.77	96.31	83.51	96.81
PERSON	86.23	87.32	85.65	87.13	85.35	87.31	85.22	**87.46**
TIME	**90.22**	88.22	89.45	87.72	89.32	87.47	89.86	87.77
全体	82.37	86.25	82.31	**86.30**	81.91	86.19	81.74	86.08
One vs. Rest 法								
冗長解析	1 次解のみ		2 次解まで		3 次解まで		4 次解まで	
解析方向	順	逆	順	逆	順	逆	順	逆
ARTIFACT	43.11	48.96	41.12	**50.06**	39.84	49.58	38.65	48.45
DATE	90.79	**94.18**	91.19	**94.18**	91.21	93.97	90.96	93.83
LOCATION	84.72	87.65	84.67	87.61	84.31	87.75	84.15	**87.77**
MONEY	63.46	93.79	61.62	93.67	61.85	93.85	62.13	**95.47**
ORGANIZATION	74.37	78.96	73.70	**79.27**	72.74	78.33	72.73	78.12
PERCENT	86.07	**97.09**	86.23	96.02	85.66	96.06	85.51	96.28
PERSON	85.92	**87.69**	86.03	87.40	85.51	87.19	85.41	87.16
TIME	**90.98**	89.04	90.54	88.07	90.38	88.33	89.90	88.32
全体	82.72	**86.40**	82.58	86.35	82.12	86.11	81.95	86.07

左 2 文字右 2 文字文脈，素性（品詞，文字，字種，前固有表現タグ）．
Kernel 関数は 2 次の多項式関数．

表 5 素性の違いによる精度の比較
Table 5 The extraction accuracy for each feature set.

Pairwise 法								
素性	すべて		−文字		−字種		−品詞細分類	
解析方向	順	逆	順	逆	順	逆	順	逆
ARTIFACT	42.17	**48.30**	23.64	25.04	41.36	46.31	41.45	45.77
DATE	91.16	**94.14**	76.26	80.41	91.08	94.04	90.07	93.33
LOCATION	84.07	**87.62**	77.29	79.15	83.87	87.27	76.37	70.99
MONEY	59.88	**95.82**	47.09	87.48	58.44	95.81	57.84	90.91
ORGANIZATION	72.63	**78.79**	60.81	62.06	72.15	78.62	66.10	73.41
PERCENT	83.77	**96.31**	68.78	83.05	84.10	95.98	82.59	94.58
PERSON	85.35	**87.31**	81.46	83.05	84.59	86.29	73.55	78.42
TIME	**89.82**	87.47	83.33	81.56	89.53	87.57	89.68	86.26
全体	81.91	**86.19**	72.14	75.13	81.54	85.78	75.58	77.94
One vs. Rest 法								
素性	すべて		−文字		−字種		−品詞細分類	
解析方向	順	逆	順	逆	順	逆	順	逆
ARTIFACT	39.84	**49.58**	22.97	23.94	39.98	47.82	39.69	47.42
DATE	91.21	93.97	75.80	80.57	91.25	**94.09**	90.17	93.34
LOCATION	84.31	**87.75**	75.87	79.38	84.50	87.63	76.99	82.68
MONEY	61.35	93.85	45.19	85.19	60.33	**94.86**	59.62	89.89
ORGANIZATION	72.74	**78.33**	58.85	61.95	72.77	78.31	66.60	73.64
PERCENT	85.66	96.06	66.86	79.61	86.21	**96.09**	83.76	94.81
PERSON	85.51	**87.19**	80.43	82.33	84.87	86.59	73.92	79.07
TIME	**90.38**	88.33	80.44	77.31	90.36	88.27	88.96	86.59
全体	82.12	**86.11**	70.73	74.92	82.07	85.96	76.02	81.72

左 2 文字右 2 文字文脈，冗長解析結果は 3 次解まで利用．
Kernel 関数は 2 次の多項式関数．

表 6 多項式 Kernel 関数の次数の違いによる精度の比較
Table 6 The degree of polynomial Kernel function and the extraction accuracy.

Pairwise 法						
素性	1 次		2 次		3 次	
解析方向	順	逆	順	逆	順	逆
ARTIFACT	36.81	47.87	42.17	**48.30**	38.86	43.93
DATE	90.21	92.78	91.16	**94.14**	91.25	93.70
LOCATION	83.79	85.55	84.07	**87.62**	83.74	86.73
MONEY	55.22	95.42	59.88	**95.82**	59.63	93.88
ORGANIZATION	71.62	75.25	72.63	**78.79**	72.60	78.22
PERCENT	84.13	**97.04**	83.77	96.31	80.14	93.47
PERSON	83.25	85.15	85.35	**87.31**	85.13	86.48
TIME	89.09	88.42	89.82	87.47	**89.99**	85.80
全体	80.66	84.10	81.91	**86.19**	81.66	85.36
One vs. Rest 法						
素性	1 次		2 次		3 次	
解析方向	順	逆	順	逆	順	逆
ARTIFACT	32.62	45.26	39.84	**49.58**	38.32	44.25
DATE	90.11	93.02	91.21	**93.97**	91.45	93.63
LOCATION	83.57	85.88	84.31	**87.75**	84.36	87.26
MONEY	55.36	**94.21**	61.85	93.85	63.55	93.91
ORGANIZATION	71.22	75.61	72.74	**78.33**	72.76	78.13
PERCENT	81.86	95.35	85.66	**96.06**	83.10	94.18
PERSON	82.71	85.05	85.51	**87.19**	85.54	86.90
TIME	85.26	88.06	**90.38**	88.33	89.86	87.25
全体	80.36	84.23	82.12	**86.11**	82.17	85.65

左 2 文字右 2 文字文脈，冗長解析結果は 3 次解まで利用.
素性（品詞，文字，字種，前固有表現タグ）.

導入した素性は「文字」「字種」「品詞」「固有表現タグ」の 4 種類で，これらすべてを用いたもの,「文字」を用いなかったもの,「字種」を用いなかったもの,「品詞細分類」を用いなかったものについて，各固有表現の精度および全体の精度を示す.

一般に「文字」の情報を除くと精度が急激に下がってしまう．これは品詞表現だけでは粗いために各固有表現を開始位置または終了位置を捕捉するだけの情報が得られないことが理由であると考えられる.

また，本実験では導入しなかったが，予備実験で活用の情報を入れると精度が若干下がった．これは一般に固有表現は活用語を含まないためであると考えられる.

4.5 多項式 Kernel 関数の次数の違いによる精度の比較

適用する多項式 Kernel 関数の次数 d を 1 から 3 に変化させ，素性の組合せを考慮した学習がどれだけ重要であるかを調査した．**表 6** に結果を示す．先行研究[10] と同様にほとんどの表現について 2 次の Kernel 関数を用いるものが最も精度が高かった.

4.6 シソーラスの効果

上までの実験では，文献 10) で用いられている素性を基に比較を行ってきた．文献 3) では，シソーラス（NTT 語彙大系[11]）の情報も導入している．**表 7** に本手法にシソーラスの情報を導入した際の結果を示す．なお，シソーラスの情報は，各冗長解析結果である形態素が NTT 語彙大系のどの葉ノードに含まれているかのみについて，品詞と同様に SE タグとともに付与した．ARTIFACT と TIME を除く表現で，シソーラスが有効に働いていることが分かる．ARTIFACT については，シソーラスを用いても未知の表現に対する有用な素性を付与できなかったことが精度悪化の原因

表 7 シソーラスの有無による精度の比較
Table 7 The thesaurus and the extraction accuracy.

素性	シソーラスなし		シソーラスあり	
解析方向	順	逆	順	逆
ARTIFACT	41.12	**50.06**	43.28	49.15
DATE	91.19	94.18	91.78	**94.80**
LOCATION	84.67	87.61	85.78	**88.59**
MONEY	61.62	93.67	64.58	**95.34**
ORGANIZATION	73.70	79.27	75.69	**80.37**
PERCENT	86.23	96.02	86.64	**96.11**
PERSON	86.03	87.40	86.21	**87.73**
TIME	**90.54**	88.07	90.19	88.92
全体	82.58	86.35	83.58	**87.12**

左 2 文字右 2 文字文脈，冗長解析結果は 2 次解まで利用.
素性（品詞，文字，字種，前固有表現タグ）.
One vs. Rest 法による.

表 8 先行研究との比較
Table 8 Comparison with related work.

	CRL 公開データ	IREX GENERAL	学習モデル	文脈長	わかち書きの問題への対処	シソーラス
内元 2000 [4]		80.17	ME	±2	書き換え規則	無
颯々野 2000 [7]		82.8	ME	±2		無
山田 2002 [10]	83.7		SVM	±2	学習データにあるものは分割	無
竹元 2001 [8]		83.86	辞書+規則		複合語分割辞書	無
宇津呂 2002 [5]		84.07	ME +決定リスト	一部可変長		無
磯崎 2003 [3]	86.77	85.77	SVM + sigmoid	±2	書き換え規則	無
本手法	87.12		SVM	±2	解析単位を文字	有
中野 2003 [9]	89.03		SVM	±2，文節素性	解析単位を文字	有

注）「宇津呂 2002」の結果は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.

であると考える．TIME に関しては利用される語彙および文字が限られているため，シソーラスの情報までは不要であるのだろうと推察される．

4.7 考　　察

表 7 における「シソーラスあり・解析方向 逆」の結果が最も精度が高かった☆.

本手法はまとめあげ手法に文献 10) とまったく同じものを採用しているが，この文献では，わかち書きの問題が起きないような環境での評価実験で 85.9（F 値）であることを報告している．この値と今回の評価実験と比較すると，我々の提案手法がわかち書きの問題を解決している以上の大幅な精度改善を達成していることが分かる．

表 8 に先行研究との比較を示す．本手法の提案は，素性の展開手法であるために，過去に提案されている他の学習モデルのものにも適用することが可能であると考える．文献 9) は，本手法で導入した文字単位での展開だと，長い単位の固有表現がとれないという問題点を，文節素性を導入することにより，解決している．

5. まとめと今後の課題

本稿では日本語固有表現抽出タスクに対し，冗長的な形態素解析結果の利用する手法を提案し，その有効性を示した．形態素解析の n 次解を利用し，形態素解析器の性能を十分に引き出すことにより，高い精度を得ることができた．また，文字単位にまとめ上げを行うことにより，わかち書きの問題を解決し，未知語に対しても頑健なモデルを構成することができた．

本提案手法でも ARTIFACT の抽出は先行研究と同様あまり高い精度が得られなかった．今後，ARTIFACT 抽出に有効な素性および言語資源を考慮していきたい．また今回はシソーラスの木構造のノードについては考慮しなかったが，今後シソーラスのノードの上下関係を素性にうまく取り込むことにより，精度を向上させることができると考えている．

謝辞 有用な議論をしていただいた北陸先端科学技術大学院大学の山田寛康氏と筑波大学の中野桂吾氏に感謝の意を表します．また，yamcha を公開している奈良先端科学技術大学院大学の工藤拓氏に感謝の意を表します．

☆ 左 2 文字右 2 文字文脈，冗長解析結果は 2 次解まで利用，素性（品詞，文字，字種，前固有表現タグ），One vs. Rest 法による．

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) Vapnik, V.: *Statistical Learning Theory*, A Wiley-Interscience Publication (1998).
3) 磯崎秀樹，賀沢秀人：SVM に基づく固有表現抽出の高速化，情報処理学会論文誌，Vol.44, No.3, pp.970–979 (2003).
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) 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).
6) 工藤　拓，松本裕治：Support Vector Machine を用いた Chunk 同定，自然言語処理，Vol.9, No.5, pp.3–23 (2002).
7) 颯々野学，宇津呂武仁：統計的日本語固有表現抽出における固有表現まとめ上げ手法とその評価，情報処理学会研究会報告（自然言語処理研究会），No.2000-NL-139-1, pp.1–8 (2000).
8) 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).
9) 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，情報処理学会研究会報告（自然言語処理研究会），No.2003-NL-156-2, pp.7–14 (2003).
10) 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).
11) NTT コミュニケーション科学研究所：日本語語彙大系，岩波書店 (1997).

(平成 15 年 3 月 31 日受付)
(平成 16 年 3 月 5 日採録)

浅原　正幸（正会員）

1975 年生．1998 年京都大学総合人間学部基礎科学科卒業．同年奈良先端科学技術大学院大学情報科学研究科博士前期課程入学．2001 年同大学博士後期課程進学．同年より日本学術振興会特別研究員．2003 年同大学博士後期課程修了．2004 年より奈良先端科学技術大学院大学助手，現在に至る．自然言語処理の研究に従事．言語処理学会学生会員．

松本　裕治（正会員）

1955 年生．1977 年京都大学工学部情報工学科卒業．1979 年同大学大学院工学研究科修士課程情報工学専攻修了．同年電子技術総合研究所入所．1984 年〜1985 年英国インペリアルカレッジ客員研究員．1985 年〜1987 年（財）新世代コンピュータ技術開発機構に出向．京都大学助教授を経て，1993 年より奈良先端科学技術大学院大学教授，現在に至る．京都大学工学博士．専門は自然言語処理．人工知能学会，日本ソフトウェア科学会，言語処理学会，認知科学会，AAAI，ACL，ACM 各会員．