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[††]

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に与えられず，解析誤りが起こりやすくなる．そこで，本論文では文節区切りを行い，文節内外の情報を素性としてチャンカに与える手法を提案する．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．

Japanese Named Entity Extraction with Bunsetsu Features

Keigo Nakano[†] and Yuzo Hirai[††]

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 performed using contextual information such as lexical or part-of-speech features obtained through a fixed-size window. For longer entities, however, the chunker may not be given sufficient information because of the shorter and fixed-size window. To cope with this problem, we propose a method which makes use of Japanese base phrases, called *bunsetsu*, as features for chunking. Since our feature extraction method depends on the length of each *bunsetsu*, our chunker can take advantage of more flexible contextual information than conventional 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 effectiveness of our method.

1. はじめに

情報化が浸透した現在，個人が扱わなければならないコンピュータ上の情報量は，人の処理能力の限界を超えている．特に，新聞記事やホームページなどの大量なテキストの中から，人名や地名など必要な情報を素早く見つけることは至難の業である．このような人名，地名，組織名などを固有表現（named entity）という．固有表現抽出は，質問応答システムや重要文抽出などの高次の言語処理のための情報抽出においてきわめて重要なタスクとなっている[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．トークンの単位としては，単語や文字が考えられるが，Asahara ら[2)] は，文字を用いた手法が単語を用いた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．

そこで本論文では，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も行い，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[3)] や Asahara ら[2)] が採用している Support Vector Machines（以下 SVM）に基づくチャンカ yamcha を用いた．実験により先行研究と比較した結果，提案手法はこれまでで最も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とおりである．2 章では固有表現抽出の解説と問題の定式化を行う．3 章で提案手

† 筑波大学システム情報工学研究科
Doctoral Program Systems and Information Engineering, University of Tsukuba
†† 筑波大学電子・情報工学系
Institute of Information Sciences and Electronics, University of Tsukuba

表 1 IREX による固有表現の種類と例
Table 1 Types and examples of named entities defined by IREX.

固有表現の種類		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

文字	固有表現タグ
小	B-PERSON
泉	I-PERSON
首	O
相	O
は	O
日	B-LOCATION
米	B-LOCATION
首	O
脳	O

図 1 固有表現タグの例
Fig. 1 Examples of named entity tags.

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．4 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．5 章で実験によって提案手法の有効性を検証する．最後に 6 章でまとめと今後の課題について述べる．

2. 固有表現抽出タスク

情報抽出と情報検索を対象とした IREX（Information Retrieval and Extraction Exercise）[1] というワークショップでは，抽出対象として**表 1** に示す 8 種類の固有表現を定義している．また IREX では，各固有表現は重なり合うことはなく，入れ子になることもないとしている．

機械学習に基づく固有表現抽出においては，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判別するという手法が一般的である．各トークンのまとめあげ状態（chunk state）を表すために様々な手法が提案されている．本研究では SVM を用いた固有表現抽出[3] において最も精度が良いと報告されている Inside-Outside 法（以下 IO 法）[4] のバリエーションの 1 つである，IOB2 [5] と呼ばれるチャンクタグ集合を用いる．**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．ここで固有表現タグとは，チャンクタグと固有表現の種類をハイフンで結んだもののことをいう．IOB2 では固有表現の先頭トークンに B タグを付与し，それ以降のトークンに I タグを付与する．O タグは固有表現以外のトークンに付与される．

これらの表現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定式化することが可能になる．

3. SVM を用いた固有表現抽出

本章では提案手法のベースとなる Asahara ら[2] の文字をトークンの単位とした素性展開について解説する．

3.1 文字単位の素性展開

先行研究の多くはまとめあげの単位として形態素を用いている．しかし，この手法では形態素の境界が固有表現の境界と一致しないと抽出することができないという問題がある．たとえば「茨城県内で」を Chasen を用いて形態素解析を行うと「茨城/県内/で」と分割されてしまうために「茨城県」を地名として抽出することはできない．この問題に対処するために，書き換え規則を用いて分かち書きを修正するなどの処理が必要となる[6),7]．一方で Asahara ら[2] は，まとめあげの単位として文字を用いる手法を提案している．文字をまとめあげの単位として用いることによって，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．また，文字そのものを素性として使用するため単語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．

形態素を単位とする場合と異なり，文字には直接品詞情報を付加することはできない．そこで各文字が属する単語と品詞に，その文字の単語中の位置に応じて Start-End 法（以下 SE 法）[6] に基づくチャンクタグを付与したものを素性として用いる．**図 2** に例を示す．SE 法では形態素の先頭文字に対し B タグ，末尾に E タグ，内部に I タグ，1 文字からなる形態素に対しては S タグが付与される．固有表現抽出においては形態素の開始・終了位置が重要な情報となる．そのため単語・品詞情報の表現法としては，開始・終了位置のどちらにも固有のタグを付与する SE 法が有効である．

3.2 SVM の固有表現抽出への適用

Supprot Vector Machines [8] 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

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

図 2 チャンキングのための素性展開
Fig. 2 Feature expansion for named entity chunking.

いる．詳細は文献 3) 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 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．

チャンカは 3.1 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実線で囲まれた窓内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）で精度が異なる．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（図 2 は左向き解析の例）[2),3)]．

4. 文節情報を用いた固有表現抽出

4.1 従来法の問題点

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 2 つの例文を示した．どちらの例文においても同じ素性（太線で囲まれた前後 2 文字の窓内）がチャンカ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窓内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．また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

同 委員会にお いて
日本オリンピック 委員会にお いて
←解析方向

図 3 従来法の問題点の例
Fig. 3 An example of problem in conventional approach.

処するためには，窓外の情報を文脈に応じて適切に利用する必要がある．

4.2 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，形態素解析およ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文字単位のチャンキングを行う，

というステップからなる．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．

4.3 文節素性の展開

本論文で用いる文節の定義は京都大学コーパス ver3.0 [9)] の基準に従うものとする．また，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．

固有表現を抽出する方法には，文頭から解析していく方法と，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，本論文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．窓外の情報を利用する必要がある一方で，何を文節素

文字	品詞	文節内素性	固有表現タグ
日本	名詞-固有名詞-地域-国	*	B-ORGANIZATION
オリンピック	名詞-一般	名詞-固有名詞-地域-国	I-ORGANIZATION
委員	名詞-一般	名詞-固有名詞-地域-国	I-ORGANIZATION
会	名詞-接尾-一般	名詞-固有名詞-地域-国	I-ORGANIZATION

図 4 文節内素性の例
Fig. 4 Examples of *intra-bunsetsu features*.

文字	品詞	隣接文節素性	固有表現タグ
加工	名詞-一般	*	O
会社	名詞-一般	*	O
「	記号-括弧開	*	O
松原	名詞-固有名詞-地域-一般	会社	B-ORGANIZATION
シャーリング	名詞-固有名詞-一般	会社	I-ORGANIZATION
」	記号-括弧閉	*	O
の	助詞-連体化	*	O

図 5 隣接文節素性の例
Fig. 5 Examples of *neighboring-bunsetsu features*.

性として用いるかによっては過学習を起こす可能性がある．また，解析に用いる素性の数を増やすと，素性ベクトルの次元が大きくなり，モデルの学習時間が増大する．最適な素性の組合せはタスクによって異なるため，実験的に決定しなければならない．本論文では，日本語固有表現抽出に有効であると考えられる文節素性として以下の 3 種類を定義した．

- **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なるが，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．図中の二重線は文節の区切り位置を示している．この例では,「シャーリング」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（文節末から見て最初の自立語）を素性とする．たとえば「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．

☆ 解析時には各文字に文節素性が付与される．

表 3 固有表現の種類ごとの精度の比較
Table 3 Performance of the proposed methods for each category of NEs.

	頻度	base model	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	88.50	**88.78**	88.72

表 2 使用する文字種
Table 2 Character types for chunking.

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

5. 実　　験

5.1 実験データおよび実験設定

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付与されている．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，適合率と再現率の調和平均である F 値（$\beta = 1$）で各モデルの比較を行った．以下，断りのない限り F 値を性能尺度として用いる．

形態素解析器には Chasen [10] を使用し，磯崎ら [7] の手法と同様に未知語を「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☆を用いた．Cabocha は京都大学コーパス ver3.0 [9] から文節区切りのモデルを学習している．チャンキングには SVM に基づくチャンカ yamcha☆☆を使用し，チャンキングの解析方向はすべ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．

5.2 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．

(1) **base model**：文字，単語，品詞，文字種，前固有表現タグ（図 2 と同じ構成）
(2) **model A**：base model の素性＋文節内素性
(3) **model B**：base model の素性＋文節内素性＋隣接文節素性
(4) **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．実験用マシンとして OS は Linux，CPU は Xeon 2.0 GHz，メモリは 4 GB の計算機を用いた．base model の学習には 9 時間程度，model A，B，C の学習にはそれぞれ 11 時間，13 時間，15 時間程度必要とした．

5.2.1 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．表中の太字は固有表現の種類ごとの最高精度を示している．頻度は CRL データ中の出現頻度である．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．

5.2.2 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変化を，**表 5** に抽出数および正解数の変化を示す．表中の n は固有表現

☆ http://cl.aist-nara.ac.jp/~taku-ku/software/cabocha/
☆☆ 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
Table 4 Difference in performance according to the number of morphemes constituting NEs.

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 値	87.09	88.81	89.51	76.51	74.31
	再現率	85.44	**89.40**	87.21	69.92	66.11
	適合率	88.81	88.23	91.94	84.48	84.83
model C	F 値	**88.72**	**89.86**	**91.17**	**80.58**	**79.00**
	再現率	**86.65**	89.07	**88.18**	**75.67**	**73.41**
	適合率	**91.02**	**90.66**	**94.37**	**86.18**	**85.49**

表 5 固有表現の構成要素数による抽出数と正解数
Table 5 Difference in the number of extracted entity according to the number of morphemes constituting NEs.

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	抽出数	17968	10486	5102	1398	982
	正解数	15957	9252	4691	1181	833
model C	抽出数	17759	10168	5026	1483	1082
	正解数	16164	9218	4743	1278	925

を構成する形態素数を示している☆.

表 4 より，F 値で比較すると，n が多くなるに従って 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．

抽出数および正解数での比較では，n が小さいときには提案手法は正解数自体は従来法と大差はないが，抽出数は減少していることが分かる．一方 n が大きいときには，抽出数・正解数ともに従来法に比べて増加している．以上のことから，提案手法は従来法と比較して過抽出を抑えつつ，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分かる．

従来法では,「消防署」や「研究所」など単独では固有表現とならない文字列を固有表現として誤って抽出してしまう．また従来法では，構成要素数の多い固有表現が抽出できなかった場合，抽出すべき対象の一部を固有表現として過抽出してしまうことがあり，適合率の低下を招く原因となる．たとえば,「日本オリンピック委員会」という組織名に対し「日本」のみを地名として抽出してしまうことなどがその例である．提案手法は，文節素性の導入によって構成要素数の多い固有表現であっても正確に抽出できるため，従来法のような問題は起きない．このような過抽出の減少が構成要素数の少ない場合の適合率の差となって現れたと考えられる．

☆ 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形態素数とは若干異なる．

表 6 意味素性の利用による精度の変化
Table 6 The thesaurus and extraction performance.

	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**

5.3 意味素性の利用

先行研究[2),7)] 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が報告されている．そこで model C において，素性として用いている単語（各文字が属する単語および文節素性として用いる単語）を日本語語彙大系[11)] 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 6** 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すべて併記している．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．

5.4 考　　察

表 7 に先行研究との比較結果を示す．CRL DATA は CRL データの 5 分割交差検定による精度．IREX GENERAL は CRL DATA を学習データ，IREX 本

表 7 先行研究との比較
Table 7 Comparison with previous works.

	CRL DATA	IREX GENERAL	学習モデル	文脈長	分かち書き問題への対処	シソーラス
内元[6]		80.17	ME	±2	書き換え規則	無
山田[3]	83.2		SVM	±2	無	無
竹本[12]		83.86	辞書＋規則		複合語分割辞書	無
宇津呂[13]		84.07*	ME+決定リスト	一部可変長	無	無
磯崎[7]	86.77	85.10	SVM ＋ sigmoid	±2	書き換え規則	無
Asahara[2]	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

試験データを評価データとして得られた精度である．表より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ことが分かる．

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[13]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あげられる．そのため，2 つ以上の文節からな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．

6. おわりに

本論文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．

謝辞 本研究は，文部科学省の平成 14 年度科学技術振興調整費による「文脈主導型，認識・判断・行動機能実現のための動的記憶システムの研究」の一環として行われたものであり，その支援に感謝する．また，本学電子・情報工学系山本幹雄助教授には貴重なコメントをいただいた．合わせて感謝する．

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) Asahara, M. and Matsumoto, Y.: Japanese Named Entity Extraction with Redundant Morphological Analysis, *Proc. HLT-NAACL 2003* (2003).
3) 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，Vol.43, No.1, pp.44–53 (2002).
4) Ramshaw, L.A. and Marcus, M.P.: Text chunking using transformation-based learning, *Proc. 3rd Workshop on Very Large Corpora* (WVLC-95), pp.82–94 (1995).
5) Sang, E.F.T.K.: Noun phrase recognition by system combination, *Proc. NAACL00*, pp.50–55 (2000).
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日本語固有表現抽出，自然言語処理，Vol.7, No.2, pp.63–90 (2000).
7) 磯崎秀樹，賀沢秀人：固有表現抽出のための SVM の高速化，情報処理学会研究報告 NL149-1, pp.1–8 (2002).
8) Vapnik, V.N.: *Statistical Learning Theory*, A Wiley-Interscience Publication (1998).
9) 黒橋禎夫，長尾 眞：京都大学テキストコーパス・プロジェクト，言語処理学会第 3 回年次大会予稿集，pp.115–118 (1997).
10) 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).
11) 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 (1997).
12) 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).
13) 宇津呂武仁，颯々野学，内元清貴：正誤判別規則を用いた複数の固有表現抽出システムの混合，自然言語処理，Vol.9, No.1, pp.65–100 (2002).

(平成 15 年 9 月 26 日受付)
(平成 15 年 12 月 2 日採録)

中野 桂吾

2002 年筑波大学第三学群情報学類卒業．現在，同大学システム情報工学研究科 2 年在学中．

平井 有三（正会員）

1948 年生．1970 年慶應義塾大学工学部電気工学科卒業．1972 年慶應義塾大学大学院工学研究科修士課程電気工学専攻修了．1975 年慶應義塾大学大学院工学研究科博士課程電気工学専攻修了．同年富士通株式会社．1978 年筑波大学電子・情報工学系助手．1992 年より筑波大学電子・情報工学系教授，現在に至る．1993 年〜1994 年オックスフォード大学客員研究員．専門はニューラルネットワーク．電子情報通信学会，人工知能学会，日本神経回路学会，IEEE，INNS 各会員．