情報処理学会論文誌 Vol. 49 No. 11 3765–3776 (Nov. 2008)

大域的情報を用いた日本語固有表現認識

笹野遼平[†1,†2] 黒橋禎夫[†3]

本稿では大域的情報を用いた日本語固有表現認識手法を提案する．提案する手法では，SVM を用いた固有表現認識手法を基とし，構造的な解析などから得られる大域的な情報として，先行文における同一形態素の解析結果，共参照関係にある表現の解析結果，係り先から得られる情報，固有表現情報を付与した格フレームを用いた格解析から得られる情報の 4 つの情報を新たに導入する．CRL 固有表現データ（5 分割交差検定），IREX テストセット，および，ウェブテキストに固有表現を付与したデータを用いた評価実験の結果，従来手法より高い精度が得られ，手法の有効性が確認された．

Japanese Named Entity Recognition Using Non-local Information

Ryohei Sasano[†1,†2] and Sadao Kurohashi[†3]

This paper presents an approach that uses non-local information for Japanese named entity recognition (NER). Our NER system is based on Support Vector Machine (SVM), and utilizes four types of non-local information: cache features, coreference relations, syntactic features and case-frame features, which are obtained from structural analyses. We evaluated our approach on CRL NE data and obtained a higher F-measure than existing approaches that do not use non-local information. We also conducted experiments on IREX NE data and an NE-annotated web corpus and confirmed that non-local information improves the performance of NER.

†1 東京大学大学院情報理工学系研究科
Graduate School of Information Science and Technology, the University of Tokyo
†2 日本学術振興会特別研究員 DC
Research Fellow of the Japan Society for the Promotion of Science
†3 京都大学大学院情報学研究科
Graduate School of Infomatics, Kyoto University

1. はじめに

固有表現認識（Named Entity Recognition）とは，情報検索，情報抽出の基礎技術として，テキスト中から組織名，人名，地名などといった固有表現の自動的な認識を行う処理であり，高度な言語処理に向けて必要不可欠な技術であるといえる．このため，英語に関しては MUC-6 [11]，日本語に関しては IREX [4] においてコンテストのタスクとして取り上げられるなど，固有表現認識に関する研究は多く行われており，なかでも Support Vector Machine（SVM）[14] や Conditional Random Fields（CRF）[9] を用いた機械学習に基づく手法で高い精度が報告されている．

機械学習に基づく固有表現認識では，認識の手がかりとしてどのような素性を使用するかが問題となる．従来の日本語固有表現認識手法の多くは，固有表現であるかどうかを判定すべき表現と表層的に近くに出現するいくつかの表現を手がかりにして，その表現が属する固有表現を推定している．しかしながら，自然言語テキスト中には，表層的には離れていても，係り受け関係にある表現など，構造的には関係の深い表現が存在しており，このような表現も認識の手がかりとして使用できると考えられる．また，自然言語テキストは，前から読んでいくことを前提に記述されているため，後方の文の固有表現認識を行う際には，前方の文に出現した情報が手がかりとなる場合も多いと考えられる．

そこで本研究では，日本語固有表現認識において，従来から使用されてきた表層的に近くに出現する表現から得られる情報に加えて，構造的な解析などから得られる大域的な情報を手がかりとして用いる手法を提案する．提案する手法では，大域的な情報を用いるため，固有表現認識に先行して構文解析，格解析，共参照解析を行い，これらの解析結果から得られる情報，および，先行する文における固有表現認識の結果を後続する文における固有表現認識の手がかりとして利用する．

2. 関連研究

固有表現認識の手法は大きく分けて，人手によって記述されたルールに基づく手法と，正しいタグが付与されたデータから機械学習を用いてルールを獲得する手法が存在するが，近年，高い精度を実現しているものは機械学習を用いた手法である．

機械学習を用いた固有表現認識においては，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ていく手法が一般的である．この際，解析単位としてどのような単位を用いるかが問題となり，形態素区切りに基

づく手法と，文字単位で解析する手法が提案されているが，精度の差はほとんど確認できない[1),22)]．

使用する機械学習器としては，最大エントロピー法を用いた手法[24)]や，Support Vector Machine（SVM）を用いた手法[17)]，Conditional Random Fields（CRF）を用いた手法[20)]などが提案されている．日本語固有表現認識においては現在のところ SVM を用いた手法が最も高い精度を達成しており，文末または文頭から決定的に固有表現タグを決定していく手法[17)]や，SVM が出力した値にシグモイド関数を適用することにより確率類似量を計算し Viterbi アルゴリズムにより文全体で固有表現タグを決定する手法[22)]，Shift-Reduce 法に基づく手法[16)]などが提案されている．

近年の固有表現認識に関する研究は，タグ付けされていないコーパスから有用なデータの抽出・活用を目指した研究と，大域的な情報を用いるなどタグ付きコーパスのより効果的な利用を目指した研究に分けられる．

前者の例として，福島ら[19)]は大規模な Web テキストから「政党-新党大地」のように「カテゴリ名-固有名」というペアの集合である固有名リストを収集し，固有表現認識に利用している．また，Kazama ら[5)]は Wikipedia から抽出，および，Web 文書中での動詞と名詞の係り受け関係を大規模にクラスタリングすることにより大規模な固有名詞に関する辞書を構築し，それを CRF を用いた固有表現認識に用いている．土屋ら[15)]は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よく似た頻出語と対応付け，機械学習を行う際に，元々の語から得られる情報と新たに対応付けた類似語から得られる情報の両方を組み合わせて素性として使用している．これらの情報は本稿で提案する手法では使用していない情報であり，本稿で提案する手法においてもこれらの情報を新たに使用することにより固有表現認識精度が向上する可能性が考えられる．

本稿で提案する手法は基本的に後者に属する．後者に分類される日本語固有表現認識に関する研究はあまり行われていない．中野ら[18)]は文字を解析単位とした機械学習手法における学習の素性として，一般的に使用される前後 2 文字，計 5 文字が属する形態素の出現形，文字種，品詞，品詞細分類に加えて，文節内の解析方向に存在する固有名詞に関する素性（文節内素性），隣接する文節の末尾に関する素性（隣接文節素性），文節の主辞に関する素性（主辞素性）という 3 つの文節素性を導入している．しかしながら，使用している情報は文節内，および，直前の文節から得られる情報にとどまっている．佐竹ら[23)]は照応関係を考慮した日本語固有表現認識手法を提案している．ただし，佐竹らが設定したタスクでは，固有表現を参照している普通名詞も認識すべき固有表現と定義しており，一般的な固有表現定義に対してどの程度の効果があるかは不明である．

日本語以外を対象とした研究に目を向けると，英語を対象とした研究では，共参照関係を利用した手法[2)]や，解析対象としている表現が文章中の別の箇所に出現している場合にその情報も最大エントロピー法に基づく固有表現認識の素性として利用する手法[3)]，構文解析を使用し係り先に関する情報を利用する手法[10)]，固有表現認識を局所的な情報のみを考慮した 1 段階目と大域的な情報も考慮した 2 段階目に分けて行うことにより大域的な情報を考慮する手法[8)]など，多くの大域的情報を活用した固有表現認識手法が提案されている．本稿ではこのような大域的情報が日本語固有表現認識においても有用であることを示す．

3. SVM に基づく固有表現認識

本章では，大域的な情報を用いた固有表現認識を行うにあたり，ベースとする手法について説明する．

3.1 認識対象とする固有表現

固有表現認識とは，与えられた文書から人名，地名，組織名といった，あらかじめ決められたタイプの表現を認識する問題である．認識する表現のタイプとしては人名，地名，組織名などいわゆる固有名詞的表現のほかにも時間や割合などの時間表現や数値表現を対象とする場合もあり，また，入れ子構造をどのように扱うかなど様々な設定を考えることができる．このため，同一の文章に対してであっても，どのようなタイプの表現の認識を行うかにより求めるべき固有表現は異なってくる．

本研究では，多くの日本語固有表現認識に関する先行研究と同様に IREX[4)]で定義された固有表現を認識することを目的とする．IREX では，**表 1** に示す 8 つの固有表現が定義されている．また，固有表現が重なっている場合は，原則的に長い単位の表現を認識するこ

表 1　IREX における固有表現の種類と例
Table 1　The definition of Japanese named entity in IREX and it's examples.

	固有表現の種類		例
固有名詞的表現	組織名	ORGANIZATION	NHK 交響楽団，ICAO
	人名	PERSON	福田康夫，川崎憲次郎
	地名	LOCATION	アメリカ，新義州
	人工物名	ARTIFACT	ノーベル賞，ひかり 123 号
時間表現	日付	DATE	6 月 17 日，今年
	時刻	TIME	午後五時，正午
数値表現	金額	MONEY	500 円，五・七新ペソ
	割合	PERCENT	90%，三分の一

とを目的としている．

3.2 SVM を用いた固有表現認識

本研究ではベースラインとして，磯崎らの手法[22)] を基にした手法を使用する．磯崎らの手法の特徴としては，以下の 3 点があげられる．

- 解析の単位として基本的に形態素を使用
- チャンクタグとして Start/End 法[24)] を使用
- 固有表現タグを決定する際は Viterbi アルゴリズムを用いて文全体で決定

解析の単位として形態素を用いた場合，固有表現の区切りが形態素中にある場合にどのように対処するかという問題が存在するが，文字単位よりも形態素単位を使った方が大きな文脈を考慮するのに適していると考えられることから，本研究でも解析の単位として基本的に形態素を使用する．固有表現の区切りが形態素中にある場合の対応は 3.2.1 項で述べる．

先行研究では，チャンクタグとして Inside/Outside 法[12)] のバリエーションの 1 つである IOB2 を用い，文末から固有表現タグを決定していくものが多いが，解析単位として形態素を使用した場合は Start/End 法と Viterbi アルゴリズムを組み合わせた手法の方が高い精度を達成している[17),22)] こと，チャンクタグの分類が細かい Start/End 法の方が，先行文章中での解析結果を後続文における固有表現認識の手がかりとして使用する（詳細は 4 章）のに適していると考えられることから，本研究でも Start/End 法と Viterbi アルゴリズムを組み合わせて使用する．

3.2.1 形態素解析

形態素解析器は，主として JUMAN 6.0[26)] を使用する．多くの先行研究で用いられている ChaSen[28)] や MeCab[25)] と比較した場合，JUMAN 6.0 の特徴としては，品詞や活用形などといった基本的な情報に加えてカテゴリなどの意味情報（詳細は次項）が付与されること，連続する数字が 1 つの形態素となること，使用している辞書の固有名詞の登録数が少ないことなどがあげられる[⋆1]．

これらの特徴のうち，意味情報が付与されることは固有表現認識において有利に働くと考えられるが，固有名詞の登録数が少ないことは固有表現認識において不利であるといえる．そこで本研究では，JUMAN 6.0 による解析に加えて，MeCab[⋆2]での解析も行い，MeCab による解析で得られた固有表現に関する情報を Start/End 法を用いて JUMAN 6.0 の解析結果に付与することにより，固有名詞の登録数が少ないという問題に対処する．

表 2　JUMAN，MeCab による形態素区切り
Table 2　Morphological segmentation by JUMAN/MeCab.

JUMAN

日中	友好	協会	代表	団	が	訪	中	，	江沢	民	国家	主席	と	会見	．
時相名詞	普通名詞	普通名詞	サ変名詞	普通名詞	格助詞	普通名詞	地名	読点	人名	普通名詞	普通名詞	普通名詞	格助詞	サ変名詞	句点

MeCab

日	中	友好	協会	代表	団	が	訪中	，	江沢民	国家	主席	と	会見	．
地名	地名	普通名詞	サ変名詞	普通名詞	普通名詞	格助詞	サ変名詞	読点	人名	普通名詞	普通名詞	格助詞	サ変名詞	句点

また，固有表現認識の解析単位として形態素を使用する場合，固有表現区切りが形態素の途中にある場合への対処が問題となるが，本研究では，以下の 2 つの形態素分割ルールを適用することにより対処する．

- 学習コーパスにおいて，固有表現区切りが形態素途中にある形態素を抜き出し，複数回出現したパターンについてはルールを作成し，分割する[⋆3]．
- JUMAN で 1 形態素であると判断された表現中に，MeCab による解析で固有名詞が関係する形態素区切りがあった場合，そこで分割する．

分割された形態素は，分割された断片 1 つずつを入力とし，JUMAN で解析し直す．例として，次の文を考える．

(i)　日中友好協会代表団が訪中，江沢民国家主席と会見．

この文を JUMAN で解析すると**表 2** に示すように 16 形態素に分割される．この解析結果を，同じく表 2 に示した MeCab による解析結果と比較すると，大きく異なるのは次の 3 点である．

- MeCab では 2 形態素と解析される「日中」が，JUMAN では 1 つの時相名詞であると解析される．
- MeCab では 1 形態素と解析される「訪中」が，JUMAN では 2 形態素と解析される．
- MeCab では 1 形態素の人名と解析される「江沢民」が．JUMAN では「江沢」と「民」の 2 形態素と解析される．

この場合，JUMAN による「日中」，「江沢民」の解析は誤っており，MeCab による解析

⋆1 ChaSen や MeCab では複数の辞書が使用可能であるが，一般的に使用されている IPA 辞書を使用した場合を想定している．
⋆2 MeCab のバージョンは 0.93，辞書は IPA 辞書（mecab-ipadic-2.7.0-20060707）を使用した．
⋆3 具体的には，「自社」，「公民」，「社民」などの形態素や，「倍」，「市」，「県」で始まる形態素，「憲」，「長」で終わる形態素，「半」を含む形態素などが分割される．

の方が優れているといえる．しかしながら，「日中」の解析は「日中は暑かった.」という文でも同様の解析結果，すなわち，JUMAN では 1 つの時相名詞，MeCab では 2 つの地名となり，この場合は JUMAN による解析の方が正しい解析となる．また，「江沢民」の解析は単純に辞書に登録されているかどうかの差であり，一概に MeCab の解析の方が優れているとはいえない．本研究では，3.2.3 項で述べるように構文解析以降の処理に JUMAN による解析結果を入力とする KNP [27] を使用することから，基本的に JUMAN をベースとした形態素解析を行う．

例文 (i) の処理に話を戻すと，MeCab では「日中」が 2 つの固有名詞に分割されていることから，「日中」は「日」と「中」に分割され，それぞれ単独で出現したものとして JUMAN で再解析した結果，それぞれ「時相名詞」，「普通名詞」であると解析される．また，「日中」の「日」，「中」は MeCab による解析で “名詞-固有名詞-地域-国” であると解析されることから，単独（**S**ingle）で国を表すという意味で “S-国” が MeCab による解析で得られた固有名詞情報としてそれぞれ付与される．同様に，「江沢民」は MeCab による解析の結果 “名詞-固有名詞-人名-一般” と解析されることから「江沢」には人名の先頭（**B**egin）あるという意味で “B-人名”，「民」には人名の末尾（**E**nd）であるという意味で “E-人名” が MeCab による解析で得られた固有名詞情報として付与される．これらの処理を行った後の最終的に使用する形態素解析結果を**表 3** に示す．

3.2.2　学習に用いる素性

ベースラインモデルにおける分類の際の素性として，分類対象の形態素とその前後 2 つずつの計 5 形態素に対して，出現形，JUMAN による品詞，品詞細分類，文字種，意味情報，文節主辞の出現形，MeCab による固有名詞情報を素性として使用する．この際，JUMAN の解析に曖昧性がある場合は，すべての品詞，品詞細分類，意味情報を使用する．

JUMAN による意味情報としては，JUMAN によって与えられるカテゴリ情報，および，人名の後にあることが多い語に付与される「人名後」，組織名の末尾になることが多い語に付与される「組織名末尾」という情報を使用する．カテゴリ情報とは，基本的にすべての普通名詞，サ変名詞，時相名詞に付与される⋆1情報で，事前に定義された 22 種類のカテゴリから 1 つ以上のカテゴリが付与される．

中野ら [18] は 3 種の文節素性を用いることにより固有表現認識の精度が向上することを報告しているが，予備実験の結果，主辞素性を除く 2 つの文節素性を素性に追加しても精度に変化は見られなかったため，本研究ではこれらの文節素性は使用しない．文節素性を用いても精度が向上しない原因としては，本研究では解析の単位として形態素を用いており，また，Viterbi アルゴリズムにより前後の表現を考慮しているため，文字単位で文末から決定していく手法を用いる場合には文節素性を用いないと考慮されないような文節内情報・隣接文節情報の多くがすでに考慮されているためであると推測される．

⋆1 普通名詞，サ変名詞，時相名詞であっても連語として登録されているものには付与されない．

表 3　形態素解析結果の例
Table 3　Example of morphological analysis.

出現形	品詞	品詞細分類	JUMAN 意味情報	MeCab 固有名詞情報
日	名詞	時相名詞	<時間>	S-国
中	名詞	普通名詞	<場所-機能>	S-国
友好	名詞	普通名詞	<抽象物>	
協会	名詞	普通名詞	組織名末尾，<組織・団体>	
代表	名詞	サ変名詞	人名後，<人>, <抽象物>	
団	名詞	普通名詞	<組織・団体>	
が	助詞	格助詞		
訪	名詞	普通名詞	連語	
中	名詞	地名	連語	
,	特殊	読点		
江沢	名詞	人名		B-人名
民	名詞	普通名詞	<人>	E-人名
国家	名詞	普通名詞	組織名末尾，<組織・団体>	
主席	名詞	普通名詞	人名後，<抽象物>	
と	助詞	格助詞		
会見	名詞	サ変名詞	<抽象物>	
.	特殊	句点		

∗JUMAN 意味情報において “<>” で囲まれたものはカテゴリを表す．

例として，表 3 に示した形態素解析結果における 11 番目の形態素「江沢」の解析を考えると，分類対象の形態素自身に関する素性としては「出現形:江沢」，「品詞:名詞」，「品詞細分類:人名」，「文字種:漢字」，「文節主辞の出現形:主席」，「MeCab 固有名詞情報:B-人名」，1 つ後の形態素に関する素性としては「出現形:民」，「品詞:名詞」，「品詞細分類:普通名詞」，「文字種:漢字」，「カテゴリ:人」，「文節主辞の出現形:主席」，「MeCab 固有名詞情報:E-人名」などの素性が使用される．

3.2.3　解析の流れ

本稿で提案する固有表現認識の基本的な流れは次のとおりである．

(1)　入力文を JUMAN，MeCab を用いて形態素解析する．
(2)　KNP [27] を用いて構文・格解析を行う．

(3)　Sasano らの手法[13] を用いて共参照解析を行う．ただし，以下の 2 つの条件を満たした場合も，平仮名列 H は漢字列 C の読み方を表していると考え，同一の対象を指しているものとして扱う．

- 平仮名列 H で始まる括弧表現が存在し，その直前文節が漢字列 C で始まる．
- JUMAN によって与えられる漢字列 C の読みの先頭，末尾が，平仮名列 H の先頭，末尾と一致し，かつ，平仮名列 H の半分以上が漢字列 C の読みに含まれる．

(ex. 梅木崇 さん（うめき・たかし = 駒沢大教授，行政法専攻）···)

(4)　以下の 2 つの文分割ルールに基づき入力文を分割する．

- 句点，または，カギ括弧が出現した場合はそこで分割．
- 読点が出現し，その直前が数詞でない場合はそこで分割．

(5)　文頭から順に，分割された部分ごとに固有表現の認識を行う．

ベースライン手法では基本的に形態素解析結果から得られる情報しか使用しないため (2)〜(4) の処理は必要ないが，4 章で説明する大域的情報を導入するために，固有表現認識に先行してこれらの処理を行う．

本研究ではチャンクタグとして Start/End 法を使用するので，各形態素を IREX 固有表現の 8 種類に対して先頭（B），途中（I），末尾（E），単独（S）の 4 種類に分け，固有表現以外（O）を加えた合計 4×8+1=33 種類の固有表現タグを考える．学習時にはこれらのタグを正解として学習していく．この際，33 クラスへの分類が必要となるが，本稿では磯崎らにならい各クラスごとに分類器を用意する one-versus-rest 法を採用する．解析時にはさらに「組織名の先頭」に続くのは「組織名の途中」か「組織名の末尾」のみであるなどという制約を適用し，これらの制約を満たすすべての組合せの中でベストになる組合せを Viterbi アルゴリズムを用いて選ぶ．

磯崎らは文全体でベストになる組合せを決定しているが，本研究では 4 章で導入するキャッシュ素性をより細かい単位で使用できるようにするため，先述した文分割ルールに基づき 1 文を分割し，文頭から順に，分割された部分ごとにベストになる組合せを決定していく．固有表現の途中で文を分割した場合，その固有表現を認識することはできなくなるが，このような固有表現は非常に稀であり[⋆1]，また，もともと認識が難しい固有表現であるといえることから，全体的な固有表現認識の精度にはほとんど影響はないと考えられる．

⋆1 CRL 固有表現データに含まれる 18,677 個の固有表現のうち，途中で分割されるものは “「構造主義」認識論” のようにカギ括弧を含むものが 5 個，“長野，山梨両県警” のように読点を含むものが 13 個の合計 18 個のみである．

Viterbi アルゴリズムを使用する際に必要となるスコアについては，磯崎らの手法と同様に SVM の出力に式 (1) のようなシグモイド関数を適用することにより，確率類似量を計算し，スコアとして用いる[⋆2]．

$$s(x) = \frac{1}{1+\exp(-\beta x)} \tag{1}$$

4. 固有表現認識における大域的情報の利用

従来の固有表現認識手法では基本的に，表面的に近くに出現する文脈しか固有表現認識の手がかりとして使用していなかった．しかしながら，表面的には近くに出現していない表現であっても，係り受け関係にある表現や，同一の対象を指す表現は大きな手がかりになると考えられる．たとえば，以下のような 2 つの文があった場合，それぞれ 2 文目だけを読んでも，「川崎」が前者では組織名（ORGANIZATION），後者では人名（PERSON）であることを認識するのは非常に難しく，これらの固有表現クラスを正しく認識するためには 1 文目から得られる情報も使う必要があると考えられる．

(ii)　a.　第 10 節は**川崎**フロンターレと対戦．川崎 は現在 4 勝 3 分 2 敗で 6 位．
　　b.　今年の沢村賞に**川崎**憲次郎投手が決定．今季の 川崎 は，17 勝 10 敗．

そこで，3 章で導入した局所的な素性に加えて，構造的解析などから得られる大域的な情報として，キャッシュ素性，共参照関係，係り先素性，格フレーム素性という 4 つの大域的情報を固有表現認識の手がかりとして使用する．以下では，これら 4 つの情報について，これらの情報が有用であると考えられる理由，および，どのように使用するかについて説明する．

4.1 キャッシュ素性

同じ文章中に同じ形態素が出現した場合，同じ固有表現タグが付与される場合が多いなど，これらの形態素に付与される固有表現タグには何らかの相関があると考えられる．また，文章は基本的に前から読むことを前提に書かれているため，同じ対象を指す表現があった場合は前方に出現した表現の方が，その表現がどのようなものを表しているかに関する情報が多いと考えられる．このため，文章の前方に同じ形態素が出現している場合は，その固有表現解析結果を解析対象としている形態素の素性に加えることにより，固有表現解析の精度を向上させることができると考えられる．以下ではこの素性をキャッシュ素性と呼ぶ．

⋆2 実験では $\beta = 10$ としている．

たとえば，先述した例における「川崎」を解析する際，1 文目の解析で「川崎フロンターレ」と「川崎憲次郎」がそれぞれ組織名（ORGANIZATION），人名（PERSON）であると解析できていれば，2 文目の「川崎」を解析する際に，それぞれが前文で組織名の先頭（B-ORGANIZATION），人名の先頭（B-PERSON）であったという情報を素性として用いることにより，それぞれを正しく解析できるようになることが期待できる．

解析時に用いられるキャッシュ素性は，自動解析結果から作成するしかないため，学習する際も自動解析結果から作成された情報を素性として使用するべきであるが，本研究では正解として付与されている情報を使用するものとする．すなわち，「川崎フロンターレ」に続く，2 文目の「川崎」を解析する際に使用されるキャッシュ素性は，学習時にはつねに正しい “B-ORGANIZATION” という情報がキャッシュ素性として与えられるが，解析時は，1 文目の「川崎」の解析結果に依存し，“B-ORGANIZATION” というキャッシュ素性が与えられるとは限らない．

4.2　共参照関係

前節でも述べたように，同じ対象を指す表現があった場合は，前方に出現した表現の方が，その表現がどのようなものを表しているかに関する情報が多いと考えられる．このような情報は，キャッシュ素性を導入することにより，ある程度，活用することができると考えられる．

しかしながら，同義表現を用いていい換えられている場合や，ある表現に対してその読み方が記されている場合は，同一の形態素が出現しないため，キャッシュ素性は有効に働かない．たとえば，以下のような文があった場合，「NHK 交響楽団」と「N 響」，「梅木崇」と「うめき・たかし」はそれぞれ同じ対象を指しているものの，同じ形態素は含んでいないため，キャッシュ素性には反映されない．このため，このような表現に同一の固有表現を与えるためには別の手法が必要となる．

(iii) **NHK 交響楽団**客演指揮者として迎え入れられながら，団員のボイコットで ···．
その当時の団員がひとりもいなくなった N 響 と，···

(iv) **梅木崇**さん（うめき・たかし = 駒沢大教授，行政法専攻）···

そこで本研究では，共参照解析の結果，同一の対象を指していると解析された 2 表現のうち一方のみが固有表現と解析されている場合，もう一方にも同じ固有表現タグを付与するというルールを導入する．ただし，共参照情報の使用は基本的にキャッシュ素性が有効に働かないような異なる表記間の共参照関係を考慮するのが目的であること，および，誤った固有表現タグが引き継がれてしまう可能性があることから，このルールを用いるのは以下の 2 つの共参照関係に該当する表現に限定する．

(1) 大量のコーパス中に出現する括弧表現から自動的に獲得した同義表現[13] 間の，共参照関係である場合（ex. NHK 交響楽団 = N 響）

(2) 漢字で書かれた人名と，その読み方であると解析された場合（ex. 梅木崇 = うめき・たかし）

表 4　形態素解析結果
Table 4　Morphological analysis.

見出し	品詞	品詞細分類
新	接頭辞	名詞接頭辞
義	名詞	普通名詞
州	名詞	普通名詞

4.3　係り先素性

以下のような文の形態素解析を行うと，「新義州」は**表 4** のように 3 つの形態素に分割されるため，前後 2 形態素から得られる素性のみから，これらが地名（LOCATION）であることを認識することは非常に難しい．

(v) 新義州 から国境の鴨緑江を**渡り**···

しかしながら，構文解析を行うと，これらの形態素を含む文節は格助詞 “から” をともなって「渡り」という表現に係っていることが分かり，この情報を活用することにより「新義州」が地名（LOCATION）であると推測することは可能になると考えられる．

そこで本研究では，事前に構文解析を行ったのちに固有表現認識を行うことにより，解析対象の形態素が属している文節の格，および文節の係り先の文節の主辞の形態素が何であるかという情報を獲得し，獲得した情報を学習の素性として使用する．たとえば，先頭の「新」という形態素の解析を行う際には，前後 2 形態素，すなわち「義」，「州」に関する素性に加えて，「カラ格」として，「渡り」という表現に係っているという情報，具体的には，「所属文節の格:カラ」，「所属文節の係り先文節の主辞:渡り」などといった素性を，素性列に追加して使用する．

4.4　格フレーム素性

固有表現認識を行うにあたり，解析対象としている表現がどの用言のどんな格要素となっているかという情報は非常に有用な情報であると考えられる．たとえば，「派遣する」という用言のガ格は組織名（ORGANIZATION）に，ニ格は地名（LOCATION）になりやすいと考えられ，事前にこのような知識を獲得し，使用することで，固有表現認識の精度を向上させることができると考えられる．

そこで本研究では，事前に固有表現に関する情報を付与した格フレームを構築し，格解析結果とその格フレームから得られる固有表現に関する情報を，格フレーム素性として固有表現認識に利用する．

4.4.1　係り先素性との違い

格解析結果から得られる情報は，前項で述べた係り先から得られる情報と類似している．たとえば，(vi) のような文において「ICAO」の固有表現タグを推定する場合，ともに係り先である「派遣し」から得られる情報を利用することになる．

(vi) ICAO はこの結果を基に，航空関連のコンピュータソフトの改修が進んでいないなど，対応が遅れている国に専門家を**派遣し**，改善指導に乗り出す方針だ．

しかしながら，係り先素性は対象とする用言-格ペアが学習データ中に出現している場合のみ有効であるのに対して，格フレーム素性は対象とする用言-格ペアが学習データ中に 1 度も出現していなくても格フレームさえ構築されていれば使用できる点で異なり，格フレーム素性は係り先素性よりも多くの文に対して適用できる素性であるといえる．

たとえば，「派遣する」という表現が 1 度も学習データ中に出現していない場合における上記の文の「ICAO」の解析を考えると，係り先素性は有効に働かないのに対し，格フレーム素性は「派遣する」の格フレームが構築されてさえいれば利用できる．

4.4.2　固有表現情報付き格フレームの構築

格フレーム素性を使用するため，事前に大量のタグなしコーパスから固有表現情報付き格フレームの自動構築を行う．ここで，格フレームとは用言とそれがとる格要素の関係を記述した辞書であり，基本的に Kawahara らの手法[7] に基づいて構築する．

以下に Kawahara らの手法の概要を示す．

(1)　KNP[27] を用いてコーパスを構文解析し，構文的曖昧性のない述語項構造を抽出する．

(2)　抽出した述語項構造を用言とその直前の格要素のペアごとにまとめ，5 回以上出現したペアを用例パターンとして収集する．

表 5　「派遣」の格フレーム
Table 5　Case frame of "*haken* (dispatch)".

格	用例（素性）
ガ格	日本:23, 会:13, 国:12, 政府:7, 企業:6, 区:6, 団:5, 会社:5, 協会:5, 国連:4, アメリカ:4, 韓国:4, 隊:4, 国々:4, . . . （ORGANIZATION, LOCATION）
ヲ格	隊:1,249, 者:1,017, 員:932, 職員:906, 企業:214, 講師:823, 家:799, ヘルパー:694, スタッフ:398, 軍隊:347, . . .
ニ格	イラク:700, 現地:576, 海外:335, 家庭:172, 日本:171, 現場:145, インド洋:142, 中国:125, 市:119, . . . （LOCATION）
カラ格	日本:61, 会:11, 社:9, 会社:9, 県:8, 当社:8, 市:7, センター:7, 省:6, 内外:6, 町:6, 団体:6, 局:6, 学部:5, . . . （LOCATION, ORGANIZATION）
ヘ格	イラク:219, 基:108, 海外:93, 市:88, 企業:68, 現地:66, 日本:47, 地域:43, 中国:34, 州:33, 大学:28, 会:24, 元:24, . . . （LOCATION）
	⋮

(3)　用言ごとに，用例パターンのクラスタリングを行う．最終的に出来上がった各クラスタが格フレームである．

本研究では，固有表現に関する情報を付与した格フレームを構築するため，事前に基本的な素性のみを用いて構築した固有表現認識器により，格フレーム構築に使用するコーパスに固有表現タグを付与し，固有表現タグが付与されたコーパスから格フレームを構築する．そのうえで，最終的に収集された格要素の用例に各固有表現に分類された表現がどの程度含まれているかを調べ，それが 1 割を超えている場合はその格はその固有表現をとりやすいという格であるいう情報を付与する．

構築される格フレームの例を**表 5** に示す．この格フレームを用いることにより，「派遣」という用言は，「ガ格」，「ヲ格」，「ニ格」などをとることができ，そのうち「ガ格」は組織名（ORGANIZATION），または，地名（LOCATION），「ニ格」は地名（LOCATION）となる場合が多いことが分かるようになる．

4.4.3　固有表現情報付き格フレームの利用

本項では，固有表現情報付き格フレームをどのようにして固有表現認識において使用するかについて説明する．

固有表現情報付き格フレームを用いた固有表現認識を行う場合，まず，格解析を行い，文の格フレーム構造を決定する．そのうえで，固有表現情報が付与された格フレームに対応付けられた表現があった場合は，その表現の解析を行う際に，格フレームに記された固有表現情報を格フレーム素性として使用する．ここで，格解析とは，用言の格要素を同定する解析

であり，格が明示されている場合に加えて，「は」，「も」などの副助詞で表現されたり，連体節で表現されたりした場合にその格の推定を行う．本研究では Kawahara らの手法[6] を用いて格解析を行う．

格フレーム素性の利用の例として，4.4.1 項の例における「ICAO」の解析がどのように行われるかについて説明する．まず，格解析の結果，「ICAO」は格フレーム「派遣する」のガ格に対応付けられる．一方，「派遣する」の格フレームを調べるとガ格は組織名（ORGANIZATION），または，地名（LOCATION）となる場合が多いことが分かる．このような場合，「ICAO」を解析する際の素性として，“組織名（ORGANIZATION）が付与された格に対応付けられた”，“地名（ORGANIZATION）が付与された格に対応付けられた” という情報を追加して解析を行う．

5. 実　　験

5.1 使用するデータ

実験には，IREX の定義に基づいて作成された CRL 固有表現データ，IREX の GENERAL（一般課題）という公式テストデータ（IREX テストデータ），Web から獲得したコーパスに IREX の定義に基づいて固有表現タグを付与したデータ（WEB コーパス）を使用した．CRL 固有表現データは，10,718 文に対して，18,677 個の固有表現タグが付与されている．また，IREX テストデータは，981 文に対して，1,510 個の固有表現タグが，WEB コーパスは，2,399 文に対して 1,686 個の固有表現タグが付与されている．

CRL 固有表現データをテストデータとして用いる場合は，記事単位に 5 等分し，5 分割交差検定を行い，IREX テストデータ，WEB コーパスをテストデータとして用いる場合は，いずれも CRL 固有表現データを学習データとして使用して実験を行った．

5.2 大域的情報の有無による精度比較

大域的情報を使用することによる効果を確認するため，各テストデータに対して以下の 7 つの条件で固有表現認識実験を行った．

(1)　大域的情報を用いない（ベースライン）
(2)　ベースラインにキャッシュ素性を追加
(3)　ベースラインに加えて共参照関係を使用
(4)　ベースラインに係り先素性を追加
(5)　ベースラインに格フレーム素性を追加
(6)　すべての大域的情報を使用
(7)　すべての大域的情報に加えてシソーラスを使用

先行研究においてシソーラスから得られる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[1),18)] ことから，構造的情報に加えて，シソーラスの情報を導入した実験も行った．シソーラスとしては分類語彙表[21)] を使用した．

格フレーム素性を得るのに必要となる格フレームとしては，Web から獲得した約 5 億文のコーパスから構築した格フレームを使用した．この格フレームを構築する際，ベースライン手法で使用している素性のみを用いて CRL 固有表現データから学習した固有表現認識器を使用しているため，格フレーム素性を使用した実験は CRL 固有表現データに対してオープンな条件であるとはいえない．オープンな条件で格フレーム素性を使用した実験を行うためには，5 分割交差検定における学習データとテストデータの組合せごとに 5 種類の格フレームを構築する必要があるが，5 分割した CRL 固有表現データすべてを学習データとして使用した固有表現解析器と 4 つを使用した固有表現解析器の解析の差異は限定的であり，これらの差異が格フレームの固有表現情報に反映されることは少ないと考えられること，および，CRL 固有表現データすべてを使用して構築した格フレームを用いた場合でも CRL 固有表現データに対する格フレーム素性の効果はほとんど確認できないことから，学習データごとの格フレーム構築は行わない．このため，CRL 固有表現データに対して格フレーム素性を使用した実験を行う場合は，格フレーム素性を除いた条件での実験も行った．

また，それぞれの情報を使用することの有効性を確認するため，ベースラインとそれ以外の条件の間で，次のようなマクネマー検定に類似した方法で，精度の差に有意性があるかどうかの検定を行った[⋆1]．

(1)　ベースラインと出力結果が異なるものを収集する．
(2)　ベースラインの方が正しい出力，ベースラインでない方が正しい出力の数を数える．
(3)　上記の 2 つ出力の分布が 2 項分布であると仮定し 2 項検定を行う．

有意水準としては，0.1，0.01，0.001 の 3 つの値を使用した．実験結果を**表 6** に示す．検定結果は表 6 中において “*” を用いて示しており，“*”，“**”，“***” はそれぞれベースラインと比較した場合，有意水準 0.1，0.01，0.001 で精度の差が有意であることを表している．

いずれのテストデータに対しても，大域的情報を用いることにより固有表現認識の精度が

⋆1 マクネマー検定は正解の数とシステムの出力数が一致する場合に使用されるが，固有表現認識では一致しないためマクネマー検定をそのまま使用することはできない．

表 6　固有表現認識実験の結果
Table 6　Experimental results (F-measure).

	CRL 交差検定		IREX テストデータ		WEB コーパス	
ベースライン	88.63		85.47		68.98	
＋ キャッシュ素性	88.81	+0.18*	85.94	+0.47	69.67	+0.69*
＋ 共参照関係	88.68	+0.05	86.52	+1.05***	69.17	+0.19
＋ 係り先素性	88.80	+0.17*	85.77	+0.30	70.25	+1.27**
＋ 格フレーム素性	88.57	−0.06	85.51	+0.04	70.12	+1.14*
大域的情報をすべて使用	89.21	+0.58***	86.98	+1.51**	**71.27**	+2.29***
（格フレーム素性なし）	(89.09	+0.46***)				
＋ シソーラス	**89.40**	+0.77***	**87.72**	+2.25***	71.03	+2.05***
（格フレーム素性なし）	(**89.43**	+0.80***)				

*，**，***はそれぞれベースラインとの差が有意水準 .1，.01，.001 で有意であることを表す．

向上していることが確認できる．使用した情報別に見ると，キャッシュ素性，および，係り先素性はいずれのテストデータに対しても解析精度を向上させており一般的に有効な素性であるといえる．また，キャッシュ素性については，キャッシュ素性をより細かい単位で使用できるようにするため 3.2.3 項で述べたように 1 文をいくつかの部分に分割し固有表現を決定していくという手法を使用しているが，この効果を確認するため，文を分割しないで学習・解析した実験も行い比較を行った．その結果，文を分割した場合としなかった場合で出力に変化があった箇所のうち，分割した場合の方が正しかったものが 95 カ所であったのに対し，誤っていたものは 64 カ所であり，分割した部分ごとに固有表現を決定していく手法が有効であることが確認できた．

共参照関係の利用は，IREX テストデータに対しては大きく精度を向上させているものの，CRL 固有表現データ，WEB コーパスに対してはあまり精度に影響を与えていない．IREX テストデータにおいてのみ精度が大きく向上する理由は，CRL 固有表現データ，および，WEB コーパスで事前に削除されている括弧表現が IREX テストデータでは削除されていないため，IREX テストデータには他のデータよりも表記の異なる共参照関係にある表現が多く含まれているためだと考えられる．

格フレーム素性は，WEB コーパスに対しては大きく解析精度を改善させているものの，CRL 固有表現データ，IREX テストデータに対しては効果が確認できない．WEB コーパスにおいて精度が向上したことは，格フレーム素性が文章のドメインにあまり依存しない情報であることを表しており，格フレーム素性は新聞記事と Web の文章のように学習データとテストデータのタイプが大きく異なる場合に特に有効な素性であると考えられる．逆に，学習コーパスとテストコーパスのタイプが似ている場合は，認識対象の固有表現やその係り先用言などが学習コーパス中にそのまま出現している場合が多いと考えられ，このような具体的な手がかりが多くある場合には，ほとんど効果のない素性であると考えられる[⋆1]．

また，シソーラスを使わなかった場合と使った場合の精度を比較すると，新聞記事を用いた実験では，先行研究で報告されているとおりシソーラスの使用により精度が向上した．しかしながら，WEB コーパスに対する実験では有意な差はないものの精度が下がっており，やや新聞記事に特化した傾向を学習していると考えらえる．

次に，大域的情報を用いなかった場合（ベースライン），および，すべての大域的情報を使用した場合の固有表現のタイプごとの再現率と適合率を**表 7**，**表 8** に示す．大域的情報を使用することにより精度が向上したのは主に組織名，人名などいわゆる固有名詞的表現であることが分かる．

テストデータごとの傾向を見ると，テストデータとして CRL 固有表現データを用いた場合は，再現率，適合率がバランス良く向上しているのに対し，IREX テストデータを用いた場合は再現率が，WEB コーパスでは適合率がそれぞれ大きく向上している．IREX テストデータにおいては共参照関係が，WEB コーパスにおいては係り先素性と格フレーム素性が最も精度向上に貢献していることから，共参照関係の利用は主に再現率の向上に，係り先素性と格フレーム素性は主に適合率の向上に効果があると考えられる．

5.3　先行研究との比較

表 9 に先行研究との比較結果を示す．CRL 固有表現データに対して格フレーム素性を用いた場合はオープンな条件であるとはいえないことから，CRL 固有表現データのスコアは格フレーム素性を使用しなかった場合のスコアを使用している．また，本研究のために作成した WEB コーパスには対応する先行研究がないため比較を行っていない．

CRL 交差検定，および IREX テストデータを用いた実験において，それぞれこれまでで最も良い精度を得ている福島ら[19]，磯崎ら[22] と比較してみても，同等以上の精度が得られており，提案手法の有効性が確認できる．提案手法において高い精度が得られた要因としては，大域的情報を用いることによる精度向上に加えて，ベースライン手法の精度が比較的高いことがあげられる．特に CRL 固有表現データに対する実験では，本研究のベースライン手法の基にした磯崎らの精度と比べ F 値で 1.8 以上高い精度となっている．

ベースライン手法の精度が磯崎らの手法による精度と比べて高い要因を考えると，主に以

⋆1 新聞記事を対象にした実験で格フレーム素性の効果が確認できなかった要因としては，実験で使用した格フレームが Web テキストから構築されたものであったことも考えられるが，新聞記事から構築した格フレームを使用して実験を行ってもその効果は確認できなかった．

表 7　ベースラインにおける固有表現のタイプごとの精度
Table 7　Experimental results of each NE types when using baseline.

	CRL 交差検定		IREX テストデータ		WEB コーパス	
	再現率	適合率	再現率	適合率	再現率	適合率
組織名	81.07	85.90	73.13	82.24	43.82	57.07
人名	88.02	91.03	88.46	90.88	69.30	69.72
地名	89.64	91.07	86.68	87.32	73.28	76.40
人工物名	43.51	68.86	35.42	34.00	21.92	55.17
日付	94.11	94.09	93.46	95.29	94.01	89.00
時刻	90.24	92.07	94.44	92.73	57.50	76.67
金額	92.82	97.31	100.00	100.00	89.74	89.74
割合	97.36	97.96	100.00	95.45	80.00	84.21
合計	86.91	90.42	83.07	87.03	64.62	73.97
F 値	88.63		85.47		68.98	

表 8　すべての情報を使用した場合の固有表現のタイプごとの精度
Table 8　Experimental results of each NE types when using all information.

	CRL 交差検定		IREX テストデータ		WEB コーパス	
	再現率	適合率	再現率	適合率	再現率	適合率
組織名	82.40	86.89	79.22	81.25	44.57	68.79
人名	88.72	92.86	93.49	92.94	68.39	77.05
地名	90.15	91.56	86.44	89.03	75.65	76.97
人工物名	44.18	69.18	35.42	34.00	23.29	69.86
日付	93.61	94.19	92.31	94.49	93.31	90.14
時刻	89.64	92.40	94.44	98.08	52.50	77.78
金額	92.82	97.84	100.00	100.00	89.74	92.11
割合	95.73	97.31	100.00	95.45	80.00	66.67
合計	87.34	91.15	86.29	87.69	65.16	78.65
F 値	89.21		86.98		71.27	

下の 3 点が考えられる．

(1)　JUMAN により与えられる意味情報を使用している．
(2)　文節主辞の出現形を使用している．
(3)　複数の形態素解析器から得られる品詞情報を使用している．

これらの情報の利用がどのくらいベースライン手法の精度に貢献しているかを確認するため，CRL 固有表現データを対象として，ベースライン手法を基に以下の 4 つの条件で固有表現認識実験を行った．

表 9　先行研究との比較
Table 9　Comparison with previous work.

	CRL 交差検定	IREX データ	解析モデル	解析単位	特徴
磯崎 2003 [22]	86.77	85.10	SVM + sigmoid	形態素	Start/End 法，Viterbi
Asahara 2003 [1]	87.21		SVM	文字	形態素冗長解析，シソーラス使用
中野 2004 [18]	89.03		SVM	文字	文節素性，シソーラス使用
福岡 2006 [20]	87.71		Semi-Markov CRF	文字	CRF を使用
山田 2007 [16]	88.33		SVM+Shift-Reduce	形態素	文節境界素性使用
Kazama 2008 [5]	88.93		CRF	文字	Web から構築した辞書使用
福島 2008 [19]	89.29		SVM	文字	文節素性，固有名リスト使用
提案手法	**89.43**	**87.72**	SVM + sigmoid	形態素	大域的情報，シソーラス使用

表 10　ベースライン手法における各素性の貢献
Table 10　Contribution of each feature to the baseline model.

実験条件	CRL 交差検定	
ベースライン	88.63	
−JUMAN の意味情報	88.08	−0.55
− 文節主辞の出現形	88.48	−0.15
−MeCab による固有名詞情報	87.34	−1.29
上記の素性すべて不使用	86.65	−1.98

- JUMAN により与えられる意味情報を使用しない．
- 文節主辞の出現形を使用しない．
- MeCab による固有名詞情報を使用しない．
- 上記 3 つの情報をすべて使用しない．

結果を**表 10** に示す．JUMAN の意味情報，および，MeCab による固有名詞情報の利用がベースラインの精度に大きく貢献していることが確認できる．

また，先行研究において，Kazama ら[5]，福島ら[19] らは大規模な Web 文書から有用な情報を抽出し使用することにより固有表現認識精度が向上することを示しており，これらの情報と大域的情報を組み合わせて使用することで，さらに高精度な固有表現解析器を実現できる可能性があると考えられる．

6. おわりに

本稿では，大域的情報を用いた日本語固有表現認識手法を提案した．3 つのテストデータを用いた実験の結果，先行文における同一形態素の解析結果から得られる情報，共参照関係

にある表現から得られる情報，係り先の形態素から得られる情報，固有表現情報を付与した格フレームを用いた格解析から得られる情報を固有表現認識の手がかりとして用いることにより固有表現認識精度が向上することを確認した．今後の課題としては，Web テキストなどから得られる固有名に関する知識と，本稿で提案した大域的情報とを組み合わせて使用した固有表現認識システムの構築などが考えられる．

参 考 文 献

1) Asahara, M. and Matsumoto, Y.: Japanese Named Entity Extraction with Redundant Morphological Analysis, *Proc. 2003 Conference of the North American Chapter of the Assosication for Computational Linguisticsthe on Human Language Technology*, pp.8–15 (2003).
2) Charniak, E.: Unsupervised Learning of Name Structure From Coreference Data, *2nd Meeting of the North American Chapter of the Association for Computational Linguistics* (*NACL-01*), pp.48–54 (2001).
3) Chieu, H.L. and Ng, H.T.: Named Entity Recognition: A Maximum Entropy Approach Using Global Information, *Proc. 19th International Conference on Computational Linguistics*, pp.190–196 (2002).
4) IREX 実行委員会（編）：IREX ワークショップ予稿集 (1999).
5) Kazama, J. and Torisawa, K.: Inducing Gazetteers for Named Entity Recognition by Large-Scale Clustering of Dependency Relations, *Proc. ACL-08: HLT*, pp.407–415 (2008).
6) Kawahara, D. and Kurohashi, S.: Fertilization of Case Frame Dictionary for Robust Japanese Case Analysis, *Proc. 19th International Conference on Computational Linguistics*, pp.425–431 (2002).
7) Kawahara, D. and Kurohashi, S.: Case frame compilation from the web using high-performance computing, *Proc. 5th International Conference on Language Resources and Evaluation*, pp.1344–1347 (2006).
8) Krishnan, V. and Manning, C.D.: An Effective Two-Stage Model for Exploiting Non-Local Dependencies in Named Entity Recognition, *Proc. 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pp.1121–1128 (2006).
9) Lafferty, J., McCallun, A. and Pereira, F.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data, *Proc. 18th International Conference* (*ICML'01*), pp.282–289 (2001).
10) Mohit, B. and Hwa, R.: Syntax-based Semi-Supervised Named Entity Tagging, *Proc. ACL Interactive Poster and Demonstration Sessoins*, pp.57–60 (2005).
11) MUC-6 (Ed.): *Proc. 6th Message Understanding Conference*, Morgan Kaufmann Publishers, INC. (1995).
12) Ramshaw, L. and Marcus, M.: Text Chunking using Transformation-Based Learning, *Proc. 3rd Workshop on Very Large Corpora*, pp.82–94 (1995).
13) Sasano, R., Kawahara, D. and Kurohashi, S.: Improving coreference resolution using bridging reference resolution and automatically acquired synonyms, *Anaphora: Analysis, Alogorithms and Applications, 6th Discourse Anaphora and Anaphor Resolution Colloquium* (*DAARC2007*) (2007).
14) Vapnik, V.: *The Nature of Statistical Learning Theory*, Springer (1995).
15) 土屋雅稔，肥田新也，中川聖一：非頻出語に対して頑健な日本語固有表現の抽出，情報処理学会自然言語処理研究会 2008-NL-185-1，pp.1–6 (2008).
16) 山田寛康：Shift-Reduce 法に基づく日本語固有表現抽出，情報処理学会自然言語処理研究会 2007-NL-179-3，pp.13–18 (2007).
17) 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).
18) 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).
19) 福島健一，鍛治伸裕，喜連川優：日本語固有表現抽出における超大規模ウェブテキストの利用，電子情報通信学会第 19 回データ工学ワークショップ（DEWS2008），pp.A3–3 (2008).
20) 福岡健太：Semi-Markov conditional random fields を用いた固有表現抽出に関する研究，修士論文，奈良先端科学技術大学院大学情報学研究科 (2006).
21) 国立国語研究所：分類語彙表，大日本図書株式会社 (2004).
22) 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，Vol.44, No.3, pp.970–979 (2003).
23) 佐竹正臣，白井清昭，奥村　学：照応関係を考慮した新聞記事の固有表現抽出，言語処理学会第 8 回年次大会発表論文集 (2002).
24) 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).
25) 工藤　拓：MeCab. http://mecab.sourceforge.jp/
26) 黒橋禎夫，河原大輔：日本語形態素解析システム JUMAN version 6.0 使用説明書，京都大学大学院情報学研究科 (2007).
27) 黒橋禎夫，河原大輔：日本語構文解析システム KNP version 3.0 使用説明書，京都大学大学院情報学研究科 (2007).
28) 松本裕治，高岡一馬，浅原正幸：形態素解析システム『茶筌』version 2.4.3 使用説明書 (2008).

(平成 20 年 6 月 12 日受付)
(平成 20 年 9 月 10 日採録)

笹野　遼平

1981 年生．2004 年東京大学工学部電子情報工学科卒業．2006 年東京大学大学院情報理工学系研究科修士課程修了．現在，同大学院博士課程在学中．省略解析，照応解析の研究に従事．

黒橋　禎夫（正会員）

1966 年生．1994 年京都大学大学院工学研究科電気工学第二専攻博士課程修了．博士（工学）．2006 年 4 月より京都大学大学院情報学研究科教授．自然言語処理，知識情報処理の研究に従事．