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

竹元義美[†] 福島俊一[††] 山田洋志[†††]

日本語テキストからの情報抽出の基盤技術として，組織名・人名・地名・固有物名・日付・時刻・金額・割合表現を高精度で分類抽出する，固有表現抽出システムを開発した．本システムは，形態素解析を利用して入力文を単語分割し，固有表現辞書とパターンマッチルールとを適用することでテキスト中の固有表現を判定するというベーシックなアプローチをとっている．辞書の充実とルールの整備を基本方針として抽出精度の改善を進め，辞書の増強と辞書情報の詳細化，人手によるルール作成を行った．また，辞書を充実させても生じる課題として，複合語の一部となる固有名詞判定と未知語・多義語の固有名詞判定とに工夫を加えた．前者は，複合語を分割して複合語中の固有名詞を判定することにより，固有名詞の抽出洩れを救済する．後者は，ルールで判定した固有名詞で信頼度の高いものをもとに，未知語・多義語となった固有名詞の省略表現を判定する．IREX-NE コーパス（トピックを限定しない一般的な内容の記事）を用いた精度評価を実施し，F 値で 83.86 という精度を得た．また，導入したルール・処理の効果も分析し，有効性を確認した．

A Japanese Named Entity Extraction System Based on Building a Large-scale and High-quality Dictionary and Pattern-matching Rules

Yoshikazu Takemoto,[†] Toshikazu Fukushima[††] and Hiroshi Yamada[†††]

We have developed a Named Entity extraction system from Japanese text. “Named Entities”, i.e., proper names and temporal/numerical expressions are considered as the essential elements for extracting information. The system employs a conventional method that it divides input Japanese text into words and parts of speech by morphological analysis and extracts each Named Entity by referencing dictionaries and applying pattern-matching rules. In order to improve the system’s accuracy, we aim to build a large-scale and high-quality dictionary and rules. Both the dictionary and rules have been produced manually, because we believe that a hand-made dictionary or rules have better quality than those that are made automatically. We also focused our attention on two points for cases that cannot be covered by the dictionary. One is to extract proper names from compound words, and the other is to designate unknown or vague words as proper names. For the first point, our system divides compound words and determines proper names within them. Thus, omissions of proper names in compound words can be eliminated. For the second point, our system recognizes abbreviations of proper names, which tend to be unknown or vague, using reliable proper names. For the IREX-NE corpus, our system has accomplished 83.86 as F-measure score.

1. はじめに

電子化テキストが大量に作成されるようになり，膨大な量のテキストからユーザの要求に合ったものを見つけ出す情報検索技術に加えて，テキスト中からユーザに必要な情報のみを取り出す情報抽出技術の重要性が高まっている[1)~5)]．情報抽出では，抽出したい情報の枠（テンプレート）を定義しておき，そのテンプレートを埋めるという処理で必要な情報を取り出すのが一般的な形態である．情報抽出のためには，テキス

† 日本電気特許技術情報センター情報サービス事業部インフォメーションサービス部
Information Services Department, Information Services Division, NEC Patent Service, Ltd.
†† NEC 情報通信メディア研究本部インターネットシステム研究所
Internet Systems Research Laboratories, Computer & Communication Media Research, NEC Corporation
††† NEC 第二システム事業本部オープン共通システム開発部
Open Systems Development Department, 2nd Systems Operations Unit, NEC Corporation

ト中のキー要素を判別することと，それらの間の関係を認定することとが必要になる．キー要素の判別に関して，できるだけテンプレートに依存しない形で実現しようというのが固有表現抽出である．ここで，固有表現とは，キー要素として一般性のある，人名・地名・組織名などの固有名詞や時間表現・数値表現を総称したものを指す．固有表現を正確にもれなく抽出できることが情報抽出にとって第1の要件である．

しかし，日本語テキストを対象とした情報抽出では，固有表現抽出の精度が十分でないことが課題となっている．英語テキストを対象とした米国の情報抽出評価会 MUC-6[6)☆]では，評価に参加したチームの多くが90%以上の固有表現抽出精度を達成しており，英語テキストからの固有表現抽出は実用レベルの精度に達しているという見解を得た[2)]．一方，MUC-6の課題を英語以外の言語に適用する形で開催された情報抽出評価会 MET-1[7),8)]では，日本語の固有名詞抽出の評価に参加した7チームのうち，再現率・適合率とも90%を超えたチームはわずか1つであり，日本語テキストからの固有表現抽出精度は，まだ十分でないといえる．こうした背景から，MUCの流れを汲み日本で独自に情報抽出評価会 IREX-NE[9)]が開催されるに至った．

日本語テキストから固有表現を抽出するオーソドックスな手法は，日本語テキストを形態素解析し，その結果に対して固有表現抽出用の辞書やパターンマッチルールを適用して固有表現を判定することである[10),11)]．パターンマッチルールは，固有名詞と共起しやすい語や時間・数値の単位を表す語に着目したパターンに基づくマッチングを行うためのルールである．構文解析や係り受け解析などの手法も試されているが，情報抽出には，構文解析のような深い自然言語処理よりも，辞書・パターンマッチルールなどを利用した浅い処理の方が有効であるというのが最近の共通の認識となってきている[1)]．筆者らも，MUC（MET）への参加を通じて[12)]，質・量ともに充実した辞書およびルールを作成することが高い抽出精度達成への最も確実なアプローチであると強く認識した．

本稿では，日本語テキストを対象に，形態素解析・固有表現辞書・パターンマッチルールを用いるアプローチに従った，筆者らの固有表現抽出システムについて述べる．本システムでは，辞書・ルールを充実させることで精度を高め，IREX-NEでは参加チームで最良の抽出精度を達成した[13)]．本稿では，これまで積み上げてきた辞書・ルールの改良内容を報告し，それら各々の効果を分析する．

以下では，まず2章で固有表現の定義について述べる．そして，3章で本システムの設計方針を述べ，4章でシステム構成と各モジュールについて説明する．5章では，IREX-NEのコーパスをベンチマークとして評価した結果の詳細および考察について述べ，6章で今後の課題について述べる．

2. 固有表現の定義

固有表現とは，固有名詞（組織名・人名・地名・固有物名），時間表現（日付・時刻），数値表現（金額・割合）といった情報抽出のキー要素を指す．しかし，固有名詞・時間表現・数値表現の各々がどこまでの表現を含むかについては曖昧な面がある．MUC（MET），IREX-NEのような評価会において正解判定をするためには厳密な定義が必要である．IREX-NEの評価会に先立っては，固有表現の定義に関する真剣な議論が交わされ[14)]，参加者の多くの合意が得られる定義が公開された．本稿でも固有表現は，このIREX-NEでの定義に基づくものとする．

以下，8種類の固有表現について，IREX-NEで公開された定義[15)]に基づき，簡単に説明する．IREX-NEのタスクは，与えられた日本語テキストに対して，上記8種類の固有表現を見つけ，SGML形式のタグで囲むことである．固有表現の一例を**表1**に示す．より具体的な定義や例については，文献15)を参照されたい．

- 組織名：複数の人間で構成され，共通の目的を持った組織などの名称を指す．株式会社などの会社，固有の政府組織，学校，軍，スポーツチーム，国際組織，労働組合，工場，ホテル，空港，病院，教会やなんらかの目的を持ったグループなどもその対象が組織としての意味で使われている文脈においては組織名とする．
- 人名：固有の人を指す名前．役職名，敬称などは

表1 固有表現の種類
Table 1 Examples of named entity.

	種類	タグ	例
固有名詞	組織名	ORGANIZATION	NEC，西武ライオンズ
	人名	PERSON	クリントン，山田太郎
	地名	LOCATION	東京都，新大阪駅
	固有物名	ARTIFACT	カローラ，芥川賞
時間	日付表現	DATE	5月14日，6月下旬
	時刻表現	TIME	午後5時15分
数値	金額表現	MONEY	500億円，1ドル
	割合表現	PERCENT	120%，5分の1

☆ 米国では国防省の一機関である ARPA が主催となり，情報抽出に関して 1987 年から MUC，情報検索に関して 1992 年から TREC というワークショップが開催されている．

人名に含めない．

- 地名：固有の場所を指す名前．大陸，国名，地域名，都市名，地方名，県名，町名，村名，道路名，住所，駅名，線路名，モニュメント，海洋名，湾，運河，川名，池名，湖名，島，公園，山，砂漠の名前などを含む．
- 固有物名：人間の活動によって作られた具体物，抽象物を含む物の固有の名前．商標，賞名，著作権，知的所有権が主張可能であるような作品名，出版物，成果物，法律名，法案名，条約名，理論名など．
- 日付表現：絶対的な表現や，基点が明確であり絶対的な時間が分かるような相対的な表現で，その単位が 24 時間以上であるものを指す．
- 時刻表現：絶対的な表現や，基点が明確であり絶対的な時間が分かるような相対的な表現で，その単位が 24 時間より短いもの．
- 金額表現：金額を表す表現．
- 割合表現：割合を表す表現．

3. 高精度固有表現抽出へのアプローチ

本章では，まず，3.1 節で高精度固有表現抽出のために取り組むべき 4 つの課題について述べる．3.2 節では，その課題へのアプローチについて述べる．

3.1 固有表現抽出精度向上のための課題

固有表現抽出の第 1 の課題は，固有表現辞書の充実，とくに固有名詞辞書の充実である．固有名詞には新語が多く，形態素解析の誤り（未知語）の大きな原因となっているため，定期的に固有名詞を辞書に拡充していく必要がある．

第 2 の課題は，固有表現辞書を増強していく際の辞書データ間の競合解消である．固有名詞辞書作成に関して，人物や企業のデータベースから自動的に人名や企業名の辞書データを作成することも可能である．しかし，辞書データ間の競合の問題が大きくなり，人手をかけずに自動的に作成した辞書では，規模が大きくなるほど抽出精度が悪くなるという報告がある[16]．固有名詞は，一般名詞と競合したり（例：“林”，“政治”，“日光”，“巨人”），他の固有名詞と競合したり（例：“福島”，“ワシントン”）する場合が多数発生する．そこで，辞書データを増強するだけでなく，ルールで競合解消するための辞書情報を充実させることを考える必要がある．

しかし，辞書を充実させるアプローチにも限界がある．第 3 の課題は，辞書を充実させても生じる課題の 1 つとして，複合語の一部となる固有名詞が抽出の洩れとなる問題である．複合語（長単位語）の辞書登録は，形態素解析の精度や効率をチューニングする技法としてしばしば用いられる．ところが，複合語中に固有名詞が含まれる場合，たとえば “上院議員” “タイ国民” のような複合語が形態素解析の辞書に登録されている場合，形態素解析結果は 1 語にまとまってしまい，複合語中に含まれる組織名の “上院” や地名の “タイ” が抽出の洩れとなる．

第 4 の課題は，辞書をいかに充実させても生じるもう 1 つの課題で，未知語・多義語の問題である．とくに，固有名詞は，一度出現した後に同じ話題で省略表現となって出現することがしばしばあり，固有名詞の省略表現は，未知語あるいは多義語となりやすい．

3.2 各課題へのアプローチ

3.2.1 辞書・ルールの充実

第 1 の課題としてあげた固有名詞辞書の充実および第 2 の課題としてあげた競合解消の問題のために，固有名詞辞書の増強および辞書情報の詳細化を行った．辞書の増強は，過去の新聞記事などから簡単なパターンマッチプログラムで粗く抽出したものを人手で見直すことにより行い，9 万語強の固有名詞辞書を構築した．辞書情報の詳細化は，ルールを詳細化して抽出精度を高めるためである．たとえば他の固有表現や一般名詞などとの競合情報を付与することで競合解消を行うルールを記述できるようにした．

また，ルールは，3 段階で適用することで，固有表現の判定精度を高めるように工夫した．つまり，低信頼語検出ルール，固有表現判定ルール，判定結果調整ルールという順序で適用する．低信頼語検出ルールは，未知語が既登録語の連続として判定される箇所を特定するためのルールである．未知語の多くが固有名詞なので，固有名詞の判定を行う前段階で未知語らしい箇所を特定できていることが望ましい．固有表現判定ルールは，固有表現判定を行う核となるルールであり，単語の品詞情報や未知語情報と共起語のパターンなどから固有表現を判定するルールである．判定結果調整ルールは，固有表現判定ルールによる結果をさらにリファインするルールである．抽出目的や規準に応じて判定の最終調整を行うために用いる．これら 3 つのルールについて，ルールを作成するもとになった例とルールの数を**表 2** に示す．

低信頼語検出ルールは，たとえば，“三（数詞）/星（名詞）/電子（名詞）”☆という形態素解析結果を解析の信頼性が低い箇所だと判定するようなルールで，短い

☆ ’/’ は，形態素解析結果の区切りを示すことにする．

表 2 ルールの例
Table 2 Examples of rule.

ルールの種類	ルールの数	ルールの例と適用例
低信頼語検出	2	■ 1 文字の漢字語の連続箇所を低信頼語と判定. (例) “三 (数詞)/星 (名詞)/電子 (名詞)” に対して “三星” を低信頼語と判定.
固有表現判定	合計 117	■ “未知語/人名共起語” のパターンの未知語は人名と判定.
人名	22	(例) “ムバラク (未知語)/大統領 (人名共起語)”
地名	31	の “ムバラク” を人名と判定.
組織名	58	■ “固有名詞/一般名詞またはサ変名詞の連続/組織名共起語” のパターンは組織名と判定.
固有物名	2	(例) “東京都 (固有名詞)/行政 (名詞)/改革 (サ変)/委員会 (組織名共起語)”
時間表現	2	のパターンで, “東京都行政改革委員会” を組織名と判定.
数値表現	2	
判定結果調整	10	■ “地名の連続/数詞/の/数詞/の/数詞” を地名と判定. (例) “宮前区 (地名)/宮崎 (地名)/4 (数詞)/の/1 (数詞)/の/1 (数詞)” のパターンで, “宮前区宮崎 4 の 1 の 1” を地名と判定.

既登録語が連続する箇所を低信頼語として検出するようにした. 1 文字の漢字語が連続する箇所, あるいは, “リス (名詞)/トラ (名詞)” のように短い (3 文字以下とした) 片仮名語が連続する箇所を検出する 2 つのルールを作成した.

固有表現判定ルールでは, 主に, 固有名詞を判定するキーワードに着目した. たとえば, “大統領”, “社長” のような単語は, 人名の後に付きやすく, 人名を判定するためのキーワードとなる. 以下では, 固有名詞の後に付き, 固有名詞を判定するためのキーワードになるものを固有名詞共起語と呼ぶ. 固有名詞共起語は, 共起する固有名詞の種類に応じて, 人名共起語, 地名共起語, 組織名共起語, 固有物名共起語などに分けることができる. たとえば, “ムバラク (未知語)/大統領 (人名共起語)” という形態素解析結果に対して, 未知語と人名共起語のパターンに該当したとき未知語の箇所を人名と判定するルールを作成した. ただし, 実際の例では, “ムバラク (未知語)/エジプト (地名・国名)/大統領 (人名共起語)” のように, 人名と人名共起語の間にいくつかの単語が入るケースもあり, このようなバリエーションを考慮して一般化したルールを作成していった. 作成したルールの数は, 判定対象ごとに分類すると, 人名 22, 地名 31, 組織名 58, 固有物名 2, 時間表現 2, 数値表現 2, 合計 117 である.

判定結果調整ルールは, たとえば, “宮前区 (地名)/宮崎 (地名)/4 (数詞)/の/1 (数詞)/の/1 (数詞)” のような住所表現を地名としてまとめるという IREX-NE の定義に基づいており, “地名の連続/数詞/の/数詞/の/数詞” というパターンを地名と判定するようなルールがある. ルールの数は, 合計 10 である.

ルールは, 固有表現抽出の再現率・適合率をバランス良く向上させるために, F 値を高める方針で作成した☆. 過去の新聞記事における固有表現の具体的な出現パターンを人手で一般化してルールを作成し, 開発用コーパス☆☆を対象とした F 値測定により, 作成したルールの妥当性を検証しながら精度を高めていった. ルールの作成方法として, 正解コーパスから機械学習でルールを獲得する試みもある[17]. しかし, MET-2 の結果では辞書と人手で作成したルールを用いたシステムの方が高精度を得ており, 筆者らも人手でルールを作成し, 高精度な固有表現抽出システムの開発を目指した.

3.2.2 複合語内の固有名詞判定

第 3 の課題としてあげた複合語の一部となる固有名詞判定のために, 複合語を分割して複合語に含まれる固有名詞を判定する処理を導入した. 3.1 節で述べた複合語の例に対して,「“上院議員” → “上院” + “議員”」「“タイ国民” → “タイ” + “国民”」のように複合語を分割することで, 固有名詞の抽出洩れを救済するようにした.

江里口ら[8]のシステムでは, 形態素解析の結果をそれ以上分割する必要がある場合にその分割パターンを定義して分割する機能を持つ. この機能は, 本処理と同等の仕組みとなっている. 江里口らが用いた形態素解析システムは, 形態素の単位で入力文を分割する. したがって, 出力の形態素をさらに細かく分割する必要がある場合 (“来日” から地名 “日” を判定する場合など) にこの機能を用いている. 形態素解析の辞書には, 複合語の単位で基本語が含まれていることも多く, 筆者らのシステムでは, 複合語を分割するためのデー

☆ 固有表現抽出における再現率・適合率・F 値の定義については 5 章を参照.

☆☆ 開発用コーパスとは, ルールや辞書の作成の参考にしたテキストのことである. これらは, IREX 実行委員会から提供された.

タ（分割パターン）も含めた．たとえば，“タイ国民”という複合語を “タイ” と “国民” に分割し，“タイ” が地名であると判定できるように複合語分割用辞書に含めている．“タイ” のように，多義語に関しては，“タイ” を単独で地名と判定するよりは，分割して得られた “タイ” を地名と判定する方が精度は良くなる．

3.2.3 固有名詞省略表現判定

第 4 の課題については，3.2.1 項の辞書情報により競合解消することに加えて，固有名詞の省略表現判定に焦点を当てた．そのために，ルールで判定した固有名詞と未知語・低信頼語・多義語とを文字列照合し，前方一致または完全一致するものを同定する処理を導入した．固有名詞の省略表現には，(a) “明治生命” が “明治” のように前方部分のみ残して後方が省略表記される場合，(b) 逆に後方部分のみ残して前方が省略表記される場合，(c) “東京電力” が “東電” のように途中部分が飛び飛びで省略表記される場合などが考えられる．このうち，本判定処理では (a) のみを対象とした．事前分析により，(a) は頻度も多く，曖昧性のために正しく判定できずに問題になるケースが多いことが分かったためである．それに対して，(b) のような省略のされ方はきわめてまれであった．また，(c) のケースは (b) ほどは少なくないが，(c) の省略表現が他の一般名詞や固有表現との曖昧性を持つケースは少なく，辞書を充実させる対処法が有効に働く．(a) のケースは，省略表現が辞書に登録されていても，他の一般名詞や固有表現との曖昧性が生じやすく，本項で述べたような省略判定処理が必要になる．

また，誤って判定した固有名詞を省略表現の判定に用いると，誤りを増幅して悪影響を与える．そこで，事前分析を行い，誤抽出がとくに少ないルールを選び出し，そのルールで抽出された固有名詞に絞って，上で述べたような省略表現判定処理を適用するようにした．たとえば表 2 で示した “東京都行政改革委員会” を組織名と判定するルールでは，“米（固有名詞）/捜査（サ変）/機関（組織名共起語）” のようなパターンも組織名と誤って判定する．“東京都行政改革委員会” の例では，“東京都” は地名以外に曖昧性を持たないので，同じ記事中の “東京都” を組織名と判定してしまうことはないが，“米捜査機関” の例では，“米” が多義語のため，同じ記事中の “米” を組織名と誤判定してしまうことになる．このように，誤って判定した固有名詞を省略表現の判定に用いると，誤りを増幅して悪影響を与えるため，それを避けるように配慮した．

江里口ら[8]のシステムでは，すでに組織名として特定された語と未知語とを文字列比較することにより，組織名の省略語を特定する機能を実現している．さらに，すでに組織名と特定された語の部分文字列からなる正規表現をダイナミックに作成して省略表記された組織名を特定する処理についても実現している[18]．後者は，無条件に文字列照合をしているので，かえって精度が悪くなる恐れがある．一方で，筆者らのシステムでは，3.2.1 項で述べたように再現率と適合率のバランスを重視するルール作成方針に基づいて事前分析を行い，正式名称の候補を確度が高いルールで判定された固有名詞に絞り，略称の候補は未知語だけでなく低信頼語・多義語としたため，より洩れが少なく正確である．また，組織名だけでなく，人名・地名も対象としている．

4. 固有表現抽出システム

本システムの構成を**図 1** に示す．入力文解析部，固有表現情報付与部，ルール適用部，固有名詞省略表現判定部，判定結果出力部，の 5 つの基本的なモジュールからなる．以下，4.1 節から 4.5 節で各モジュールについて説明する．4.6 節では，入力文の例をもとに，固有表現判定の全体の流れを説明する．

4.1 入力文解析部

入力文解析部では，形態素解析部により入力文を形態素解析して品詞情報付きの単語列に分割し，複合語分割部により形態素解析部の結果のうち固有名詞を含む複合語をさらに分割する．形態素解析は，社内で開発されたものを用いた．形態素解析用辞書には，精度や効率を向上させるために，短単位の単語だけでなく長単位の単語（複合語）も含まれている．それら複合語のうちでとくに固有名詞を構成要素に含むものを人手で選別し，形態素解析用辞書とは別に，各複合語がどのような構成要素に分割されるかを記述した辞書を作

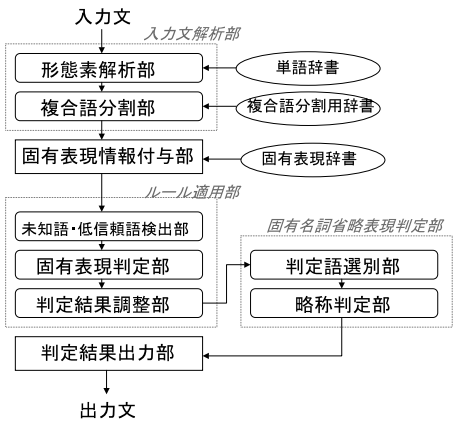

図 1 固有表現抽出システムの構成
Fig. 1 Our named entity extraction system configuration.

表 3 固有表現辞書の内訳
Table 3 Details of named entity dictionary.

分類 1	分類 2	分類 3	例	語数
固有名詞	組織名	会社，団体，施設	NEC，自民党	24990
	地名	国，県，都市，建物	日本，ニューヨーク	36714
	人名	姓，名，フルネーム	松田，聖子	27564
	固有物名		Lavie，放送法	1343
固有名詞共起語	人名，地名，組織名，固有物名		社，委員会，市，高原氏，大統領，賞，条約	1110
時間表現	日付表現，時刻表現		今年，月曜，正午	155
時間単位	日付単位，時刻単位		年，月，時，分	22
数値表現	金額表現，割合表現		半分	4
数値単位	金額単位，割合単位		円，ドル，パーセント	80
不要語			元，系，各，初代	151
合計				92133

成した（以下，複合語分割用辞書と呼ぶ）．したがって，形態素解析結果のうちで複合語分割用辞書とマッチした箇所は，固有名詞を含む複合語だということになる．たとえば “上院議員” のような複合語に関して，複合語分割用辞書に「“上院議員” → “上院” + “議員”」のような記述があれば，複合語分割部では “→” の左側の語を右側の語に分割する．分割した結果の語が固有表現辞書にあれば固有表現だと判定ができる．また，“来日” “渡米” のような基本語から地名の “日” “米” を抽出する必要がある場合には，「“来日” → “来” + “日”」のように複合語分割用辞書に定義しておけばよい．後で “日” “米” が複合語分割により得られた単語であるという情報を用いて，これらを地名と判定する．“来日” などの短単位の単語に含まれるケースは，想定されるパターンを人手で複合語分割用辞書に追加した．作成した複合語分割用辞書の登録語数は，557である．

4.2 固有表現情報付与部

固有表現情報付与部では，入力文解析部の結果の単語列に対して，固有表現辞書を参照して固有表現辞書情報を付与する．ここで付与された情報をもとにルール適用部以降で固有表現を判定する．

固有表現辞書は，固有名詞，固有名詞共起語，数値・時間表現，不要語の辞書データからなる．**表 3** に固有表現辞書の内訳と例を示す．固有名詞共起語とは，3.2.1 項で述べたように，固有名詞の前後に付いて固有名詞を判定するのに手がかりとなる語である．人名，地名，組織名，固有物名の分類がある．不要語とは，固有名詞共起語が付いても固有名詞になりにくい語である．“元” のような接辞語や，“初代” のように固有名詞（人名）との曖昧性を持つが共起語が付いても固有名詞となるケースが少ないと考えられる単語を不要語として集めた．たとえば，“村山（人名，地名；多義）/元（不要語）/首相（人名共起語）” というパターンに対して，“首相（人名共起語）” から “元（不要語）” をスキップして “村山（人名，地名；多義）” を人名と判定する．また，“トヨタ（固有名詞）/系（不要語）/販売（サ変）/会社（組織名共起語）” というパターンに対して，組織名共起語が付いても “系（不要語）” が途中にあれば “トヨタ系販売会社” を組織名と判定しないというような制御（ルールの記述）を可能にする．

さらに，辞書情報として，強制判定情報を付与した．強制判定情報は，(1) 強制的に固有名詞と判定する，(2) 強制的に固有名詞と判定しないようにする，(3) 強制判定の対象外とする，の 3 値をとる．デフォルト値は (3) である．(1) は，共起語なしで出現した場合でも固有名詞と判断してよさそうな語（人名の “クリントン”，組織名の “NEC” など）に付与する．テキスト中に共起語などが出現せずルールの適用外となるケースについては，辞書情報をもとに固有名詞判定を行う必要があり，(1) の情報が付与されている語のみ固有名詞と判定する．(2) は，曖昧性のある固有名詞に付与する（例：“林”，“政治”，“福島”）．曖昧性のある語が共起語なしで単独で出現した場合には，ルール適用部において固有名詞と判定しないようにルールで抑制する．辞書を増強していくと競合が発生しやすくなるため，(2) の情報はルールを記述するうえで有効である．

また，固有名詞は，表 3 に示すように 3 分類に詳細化している．たとえば，組織名データを会社名・団体名・施設名，地名データを国名・国名省略語・県名・主要都市名・建造物名，人名データを苗字・名前・フルネームのように細分化している．辞書情報を詳細化することによりルール適用部におけるルールも詳細に記述できる．たとえば，“千葉（人名・姓または地名）/太郎（人名・名）” の “千葉” に曖昧性があっても姓と名の連続なら人名と判定するルールを記述できる．一

方，“北方（人名・姓；一般名詞との多義）/高原（人名・姓；一般名詞との多義）” では，姓と名の連続でないので人名とは判定しない．

4.3 ルール適用部

ルール適用部では，固有表現情報付与部で付与した辞書情報とその並びのパターンに基づくルールを適用して固有表現を判定する．未知語・低信頼語検出部，固有表現判定部，判定結果調整部という 3 フェーズからなる．

未知語・低信頼語検出部では，固有表現情報付与部の結果に未知語・低信頼語検出ルールを適用して，未知語および解析の信頼性が低い箇所（以下，低信頼語と呼ぶ）を検出する．未知語検出は，未知語を含む同一字種区間を未知語にまとめる処理で，一般の形態素解析でも用いられている．低信頼語検出は，未知語が既登録語の連続として解析された箇所を低信頼語として検出する処理である[19]．つまり，見かけ上は解析が成功しているが，実際は無意味な単語の羅列となっている箇所である．たとえば，漢字 1 文字の連続や 3 文字以下の短い片仮名語の連続を信頼性が低い箇所として低信頼語にまとめる．会社名の “三星電子” や人名の “アイエロ” が辞書に未登録のために “三（数詞）/星（名詞）/電子（名詞）” “アイ（名詞）/エロ（名詞）” のように分割された場合，“三星電子” “アイエロ” を低信頼語として検出する．

固有表現判定部では，固有表現判定ルールを適用して固有名詞・時間表現・数値表現を判定する．ルールは，(A) 共起語や単位に基づくもの，(B) 辞書情報の並びに基づくもの，(C) 表層のパターンに基づくものなどがある．(A) の例としては，“福島（人名または地名）/氏（人名共起語）” という解析結果のパターンに対して，“氏” という人名共起語が付いていれば，その前の “福島” が人名または地名の曖昧性があっても人名と判定するようなルールがある．(B) の例として，“米（国名省略語）/、（読点）/英（国名省略語）” のように，“国名省略語，読点，国名省略語” という辞書情報パターンの国名省略語 “米” “英” を地名と判定するルールがある．(C) の例として，“ゲーム/機/メーカー/、/ソニー・コンピュータエンタテインメント（未知語）” のように「“メーカー” “、” “未知語”」という表層パターンの未知語を地名と判定するルールがある．このような基本的なルールは，文献 8)，11)，12)，18)，20)，21) などでも用いられている．

判定結果調整部では，固有表現判定部の結果に調整ルールを適用してリファインする．たとえば，“宮前区宮崎 4 の 1 の 1” のような住所は，固有表現判定部で “宮前区” “宮崎” が地名であると判定されたうえで，判定結果調整部で地名の連続と数字，助詞 “の” を地名としてまとめる．

4.4 固有名詞省略表現判定部

固有名詞省略表現判定部では，ルール適用部で判定した固有表現のうち判定の信頼性が高いものをもとに，固有名詞の省略表現（略称）を同定する．判定語選別部と略称判定部からなる．判定語選別部では，辞書・ルールで判定した信頼度の高い固有名詞のリスト（正式名称リスト）と，略称の候補とする未知語・低信頼語・多義語のリスト（略称リスト）とを文章番号付きで作成する．略称判定部では，略称リストと正式名称リストとを照合して，文章番号が同じで，略称リストの語と完全一致または先頭一致する語が正式名称リストにある場合に，正式名称リストの語の判定結果を略称リストの語の判定結果に与える．たとえば，“横浜” には地名と組織名の曖昧性がある．“横浜市”（地名）に関する記事で，“横浜市” を正式名称リストに入れ，多義語の “横浜” を略称リストに入れて先頭照合することで，“横浜” を地名と判定する．

4.5 判定結果出力部

4.4 節までの判定結果をもとに，入力文に対し，8 種類（人名，組織名，地名，固有物名，日付，時刻，金額，割合）のタグを付与して出力文を生成する．

4.6 処 理 例

本節では，入力文の例「当時の橋本龍太郎通産相は‥‥．橋本発言を‥‥．」をもとに，固有表現判定の全体の流れを説明する．4.1〜4.5 節の各部における処理例を**図 2** に示す．

まず，**形態素解析部**で入力文を単語に分割し，各単語に品詞情報を付与する．**複合語分割部**では，形態

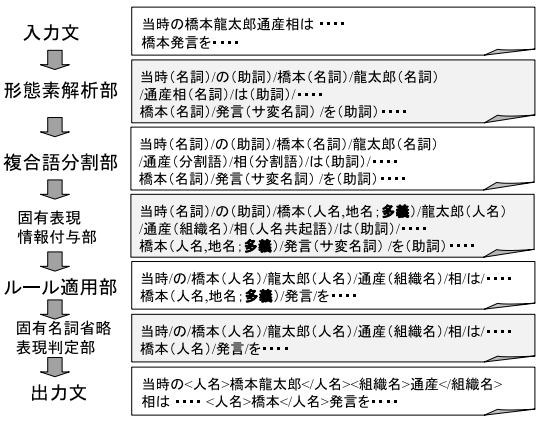

図 2 固有表現判定の処理例

Fig. 2 An example of our named entity extraction process.

素解析部の結果の “通産相” という単語をさらに “通産” “相” に分割し，複合語分割により得られた語（以下，分割語と呼ぶ）という情報を付与する．**固有表現情報付与部**では，固有表現辞書中の情報を各単語に与える．“橋本” には人名と地名の曖昧性があるので両方の情報を与える．また，分割語である “通産” には組織名情報を，“相” には人名共起語情報を与える☆．**ルール適用部**では，固有表現を判定するためのルールを適用して固有表現を判定する．この例では，“相”（人名共起語）をキーとするルールを適用している．具体的には，「未知語または固有名詞多義語＋特定の単語の連続＋人名共起語」というパターンで「未知語または固有名詞多義語」（上例の最初の “橋本”）を人名と判定する．「特定の単語の連続」とは，人名と人名共起語との間には特定の単語が入りやすいことを考慮している[11]．たとえば，組織名（上例の “通産”）や，接辞（例：橋本元首相），国名（例：クリントン米大統領）などの語が入りやすい．**固有名詞省略表現判定部**では，ルール適用部で判定した信頼性の高い固有名詞（上例で人名 “橋本龍太郎”）と判定の曖昧性が残る固有名詞（上例で後の “橋本”）とを文字列照合し，前方一致するので，後者の “橋本” を “橋本龍太郎” の省略表現と同定して人名と判定する．以上の流れからタグ付きの出力文を得る．

5. 評　　価

本章では，IREX-NE の本試験に使われたテキストを評価用コーパスとした評価結果を示す．本章の評価結果は，IREX-NE に参加した時点でのシステムのものである．評価用コーパスは，1999 年 4，5 月の毎日新聞から選定された 71 記事で，トピックを限定しない一般的な内容のものである．評価用コーパスに対する正解（固有表現が正しくタグ付けされたもの）は，あらかじめ人手で作成されている．

評価は，正解付きテキストと本システムの出力した結果とを比較し，抽出精度として再現率と適合率，および F 値を算出することで行う．再現率は，全正解中でシステムが抽出できた正解の割合を示す．適合率は，システムが出力した結果中の正解の割合を示す．通常，再現率と適合率との関係は，一方が良くなると他方が悪くなるという関係にある．F 値は，情報検索の評価でよく用いられる指標で，再現率 R と適合率 P とを重み b により統合した評価値であり，式 (1) で求められる．

$$F = \frac{(1+b^2)PR}{b^2P+R} \tag{1}$$

本章では，5.1 節で評価用コーパスにおける各固有表現の抽出精度を示す．5.2 節では，辞書・ルールによる効果を調べる．5.3 節では，とくにシステムの各構成部の効果を調べる．そして，5.4 節で全体的な誤りの傾向について述べる．

表 4 各固有表現の抽出精度
Table 4 Accuracy of our named entity extraction system.

	GLD	SYS	COR	R	P
組織	361	373	288	79.8	77.2
人名	338	324	290	85.8	89.5
地名	413	387	339	82.1	87.6
固有物	48	20	15	31.3	75.0
日付	260	275	242	93.1	88.0
時刻	54	60	47	87.0	78.3
金額	15	15	13	86.7	86.7
割合	21	17	16	76.2	94.1
合計	1510	1471	1250	82.8	85.0
F 値					83.86

5.1 固有表現抽出精度

評価用コーパスに対する本システムの固有表現抽出精度を**表 4** に示す．表 4 で，GLD は人手判定による正解数，SYS はシステムの抽出数，COR はシステムの正解数である．R は再現率であり，GLD に対する COR の割合である．P は適合率であり，SYS に対する COR の割合である．F 値は，再現率 R と適合率 P とを同じ重みとした場合，つまり式 (1) で b = 1 とした場合の式 (2) で求められる．

$$F = \frac{2PR}{P+R} \tag{2}$$

表 4 に示すように，総合の精度（F 値）は 83.86 であり，IREX-NE では最良の精度であった．表 4 の結果を次のように考察した．

- 固有物名抽出が最も難しかった．固有物名に該当する作品や製品の名前は，一般名詞や他の固有表現との解釈の曖昧性が発生しやすく，辞書に網羅することも現実的でない（例：油絵の作品名「ヨーロッパ調街並み」）．本システムでは，簡単な辞書登録とルール以外，特別な対策を行わなかったが，固有物名の件数は他の固有名詞に比べて少なく全体の精度にはあまり大きな影響を与えていない．
- 時間表現，組織名は，適合率より再現率が高い．文脈を解釈するケースがあるためだと考えられる．時間表現では，文脈を考慮せず，ほとんどパターンマッチで抽出しているため，“…から二十六年” のような時間の長さの表現を誤抽出してしまう．

☆ IREX-NE の定義では，“通産” を “通産省” の略と見なし，組織名として抽出するという決まりとなっている．

表 5 評価結果
Table 5 Evaluation of each component of our system.

評価条件 A	A1	A2			A3		A4	
評価条件 B			B1	B2	B3	B4	B5	B6
再現率	38.7	56.6	76.2	76.3	77.3	81.2	82.4	82.9
適合率	31.7	51.3	84.0	84.3	85.2	85.3	84.9	84.9
F 値	34.9	53.8	79.9	80.1	81.1	83.2	83.7	83.9

また，“関西国際空港” のように，文脈により組織名や地名となるケースも判断が難しい．

5.2 辞書・ルールによる効果

精度向上に対する固有表現辞書およびルールの寄与を調べるため，下記条件 A1〜A4 のもとで精度を比較評価した．

- （A1）辞書のみを用いたシステム
- （A2）A1 に時間・数値表現判定ルールを追加
- （A3）A2 に固有名詞判定ルールを追加
- （A4）完全なシステム

A1 は，まったくルールを用いないシステムに相当する．ルールを用いないといっても，固有名詞では辞書で競合がある場合はいずれかに決める必要があるので，組織名・地名・固有物名・人名という優先順で決定した．A2 は，時間・数値表現抽出ルールのみを追加した．時間・数値表現抽出ルールは，固有名詞抽出ルールに比べてかなり単純なので，これらのルールを区別してとらえた．A3 は，固有表現抽出ルールをすべて利用するシステムである．A4 は，A3 に複合語分割と固有名詞省略表現判定の機能を追加した完全なシステムである．評価結果を**表 5**（評価条件 A）に示す．

表 5（評価条件 A）の結果を以下のように考察した．

- 辞書だけを用いることで，38.7%の再現率が得られ，再現率に最も寄与するのは辞書であることが分かる．一方，適合率に関しては，辞書だけを用いることで 31.7%の適合率を得た．
- 適合率が最も向上するのは，固有名詞判定ルールを追加する場合で，33.9%向上している（A2→A3）．適合率への寄与が大きいのはルールであることが分かる．
- F 値に着目すると，辞書による効果が最も大きく，辞書のみを用いたシステム A1 で 34.9 の F 値を得た．
- 時間・数値表現抽出ルールは，単純なパターンマッチルールであるにもかかわらず効果が大きく，追加することにより，再現率が 17.9%，適合率が 19.6%，F 値が 18.9 向上した（A1→A2）．

5.3 システム各部の評価

システム各構成部の効果を調べるため，下記条件 B1〜B6 のもとで精度を比較評価した．

- （B1）ルールで，未知語検出部と固有表現判定部のみを用いたシステム
- （B2）B1 に低信頼語検出部を追加
- （B3）B2 に判定結果調整部を追加（=A3）
- （B4）B3 に複合語分割部を追加
- （B5）B4 に固有名詞省略表現判定部を追加（=A4）
- （B6）B5 の辞書に正解語データを追加

B1 は，ルールとして，形態素解析で一般に用いられている未知語処理と固有表現判定のみを用いた場合である．B2 は，低信頼語検出ルールの効果，B3 は，判定結果調整ルールの効果を調査するための条件である．B3 は，5.2 節の評価の A3 に相当する．B4 は，複合語分割部の効果を，B5 は，固有名詞省略表現判定部の効果を調査するための条件である．B5 は，5.2 節の評価の A4 に相当する．B6 では，開発用コーパスから必要な辞書データ（正解語）703 語を加えた．評価結果を表 5（評価条件 B）に示す．

表 5（評価条件 B）の結果を以下のように考察した．

- F 値に着目すると，システムの各部とも精度向上に寄与している．効果が最も大きかったのは，複合語分割部で，再現率が 3.9%向上し，F 値が 2.1 向上した（B3→B4）．適合率が最も向上したのは，判定結果調整部の導入によるもので，0.9%の向上だった（B2→B3）．
- 固有名詞省略表現判定部の効果として，F 値で 0.5 の向上があった（B4→B5）．副作用として，誤った判定結果を覚えてしまう場合と，部分一致するものをそのまま略称と同定すること自体に問題がある場合とが考えられる．しかし，条件をかなり絞ったためか，副作用は全体的に少なかった．たとえば，“米国” “社会党” が出現する記事で “米” は地名，“社会” は組織名と判定することで，適合率は 0.4%低下したものの，再現率が 1.2%向上した．
- 低信頼語検出ルールの寄与は期待より少なく，F 値で 0.2 向上だった（B1→B2）．評価用コーパスでは，“リス（名詞）/トラ（名詞）” を低信頼語として検出したのはよいが，「“企業” “の” 低信頼語」

表 6 開発用コーパスにおける精度（F 値のみ）
Table 6 Accuracy improvement on training corpus.

評価条件 A	A1	A2			A3		A4	
評価条件 B			B1	B2	B3	B4	B5	B6
テキスト A	51.1	69.5	85.9	86.2	86.2	89.6	89.8	93.9
テキスト B	41.3	56.5	77.2	77.5	79.1	81.9	82.9	87.2
テキスト C	38.9	56.0	75.6	76.0	77.2	80.4	81.0	82.5

というパターンで低信頼語を組織名と判定するルールにより，“リストラ” を組織名と誤抽出してしまう副作用も見られた．

また，参考として，3 種類の開発用コーパス（テキスト A，B，C）における精度（F 値）を**表 6** に示す．テキスト A は IREX-NE 訓練用テキスト 46 記事，テキスト B は 1998 年 11 月に実施された IREX-NE 予備試験用テキスト 36 記事，テキスト C は郵政省通信総合研究所が IREX-NE の定義に基づき作成した CRL コーパス 1460 記事である．いずれも出典は毎日新聞の 1994 年ないし 1995 年版である．

開発用コーパスのうち，テキスト A，B は，ルールを作成するために用い，比較的大規模なコーパスであるテキスト C は，テキスト A，B で作成したルールを検証し，副作用があった場合の調整に用いた．テキスト A は，開発の早い段階で利用し，十分に分析してルールを作成したので最も精度が高い．テキスト B は，テキスト A に比べると精度はあまり良くない．特殊なルールで対応すべきケースが多く，そのような特殊なルールはテキストが変われば効果が少なくなると考え，作成しなかったためである．

5.4 誤りの全体的な傾向

誤りの全体的な傾向について述べる．誤りには，誤抽出と抽出洩れがある．

誤抽出の主な原因は，(1) 文脈を考慮していないための過剰抽出，(2) ルールの悪影響，である．(1) で文脈とは，単語や複合語のレベルではなく，文や文章全体から得られる意味レベルの解釈である．たとえば，“横田基地で開催” という文脈においての “横田基地” は地名であり，“横田基地の裁判闘争” という文脈においての “横田基地” は組織名となる．(1) は，5.1 節で述べたとおり，とくに時間表現・数値表現の抽出については，単純なパターンマッチを行っているため，“1 年が経過” のような時間の長さを表す “1 年” を日付表現と判断してしまったり，“人一倍” の “一倍” を割合表現と判断してしまったりといった過剰抽出が発生する．また，“横田基地”，“羽田空港” のように，文脈により組織名や地名となるものの判定も難しく，本システムでは無条件に一方に判定してしまうため，誤抽出となることがある．(2) は，共起語に基づくルールによる誤抽出が多い．「固有名詞＋一般（サ変）名詞＋組織名共起語」のパターンを組織名と判定するルールで，組織名共起語 “機関” から “英国捜査機関” を過剰抽出するようなケースである．

抽出洩れの主な原因は，パターンマッチルールの記述不足である．“聟（むこ）島列島” のように固有名詞の途中に読み仮名が入ったり，“自民，自由，公明 3 党” “7，8 の両日” のように係り受けレベルで判定する必要があるような複雑なパターンまでは記述できていなかった．

6. 今後の課題

筆者らのシステムは，IREX-NE の本試験で最高の精度（F 値 83.86）を得た．しかし，1 章で述べたように，英語テキストを対象とした固有表現抽出精度が 90%を超えていることと比較すると日本語での精度はまだ十分とはいえない．

今後さらに精度を高めていくための課題として以下のようなものがあげられる．

- 5.1，5.4 節で述べたように，文や文章の意味レベルを解釈した判定が難しい．この精度を高めるには，後に付く助詞や動詞などにまで着目したパターンマッチルールないし構文解析が必要となる．
- 5.2 節の評価から，辞書の効果が最も大きいことが分かる．とくに固有名詞はこれからも増強していく必要がある．今後は，本システムで抽出した固有名詞を辞書構築にうまくフィードバックする仕組みを考える必要がある．
- また，5.2 節の評価から固有名詞抽出ルールの効果も大きいことが分かる．表 5，表 6 を比べると，評価用コーパスに対する精度と開発用コーパスに対する精度は大差ないため，汎用的なルールを作成できたと考えられる．ただし，低信頼語検出ルールで悪影響が見られるように，ルールの細かいレベルでの調整が必要である．効果の大きい基本的なルールは，ほぼそろったと考えられるので，細かいルールを，たとえば，正解コーパスから自動獲得する仕組みを検討する必要がある．

- 英語に比べて日本語での固有表現抽出が難しい理由は，日本語には分かち書きの習慣がないために単語認定（形態素解析）が難しいこともあげられる．形態素解析自体の未知語の問題や分割の曖昧性の解消に取り組むことも重要である．

7. おわりに

日本語テキストからの情報抽出の基盤技術として，組織名・人名・地名・固有物名・日付・時刻・金額・割合表現を高精度で分類抽出する，固有表現抽出システムを開発し，情報抽出評価会 IREX-NE のコーパスをベンチマークとして評価を行った．本システムは，形態素解析を利用して入力文を単語分割し，固有表現辞書とパターンマッチルールとを適用することでテキスト中の固有表現を判定するというベーシックなアプローチをとっている．筆者らは，高い抽出精度を達成するために，辞書の充実とルールの整備を基本方針として，辞書の増強と辞書情報の詳細化，人手によるルール作成を行った．また，辞書を充実させても生じる課題として，複合語の一部となる固有名詞判定と未知語・多義語の固有名詞判定とに工夫を加えた．前者は，複合語を分割して複合語中の固有名詞を判定することにより，固有名詞の抽出洩れを救済する．後者は，ルールで判定した固有名詞で信頼度の高いものをもとに，未知語・多義語となった固有名詞の省略表現を判定する．その結果，IREX-NE コーパス（トピックを限定しない一般的な内容の記事）に対して 83.86 の高い精度（F 値）を得た．

謝辞 本研究の最初の機会を与えてくださった NEC ソリューションズ第一パーソナル事業本部パーソナルメディア開発本部の村木一至本部長代理をはじめ，MET-1 におけるシステムの基本設計にあたって貴重な意見をくださった追手門学院大学文学部英語文化学科講師の福島孝博氏に感謝します．ならびに，ツールや辞書データの利用にご協力くださった，NEC 情報通信メディア研究本部マルチメディア研究所の奥村明俊マネージャー，池田崇博氏に感謝します．

参考文献

1) 関根：テキストからの情報抽出，情報処理，Vol.40, No.4 (1999).
2) 若尾：英語テキストからの情報抽出，情報処理学会自然言語処理研究会 114-12 (1996).
3) 松尾，木本：抽出パターンの階層的照合に基づく日本語テキストからの内容抽出法，情報処理学会論文誌，Vol.36, No.8 (1995).
4) 井出，藤吉，永井，中村，野村：テンプレートを用いた新聞記事からの製品情報抽出システム，情報処理学会自然言語処理研究会 115-12 (1996).
5) Cowie, J. and Lehnert, W.: Information Extraction, *Comm. ACM*, Vol.39, No.1 (1996).
6) *Proc. 6th Message Understanding Conference* (*MUC-6*), Morgan Kaufman Publishers Inc. (1996).
7) *Proc. Tipster Text Program* (*Phase II*), DARPA (1996).
8) 江里口，木谷：パターンマッチング手法による名称特定処理の有効性の検討，情報処理学会自然言語処理研究会 115-10 (1996).
9) 関根，井佐原：IREX 情報検索，情報抽出コンテスト，情報処理学会自然言語処理研究会 127-15 (1998).
10) 宮崎：係り受け解析を用いた複合語の自動分割法，情報処理学会論文誌，Vol.25, No.6 (1984).
11) 高木，安田，島崎，池原：日本語処理における固有名詞実在性検定方式の検討，情報処理学会第 35 回全国大会予稿集，6S-3 (1987).
12) Takemoto, Y., Wakao, T., Yamada, H., Gaizauskas, R. and Wilks, Y.: Description of NEC/Sheffield System Used For MET Japanese, *Proc. Tipster Text Program* (*Phase II*) (1996).
13) 竹元，福島，山田，奥村，池田：固有表現抽出システムの開発と IREX-NE における評価，IREX ワークショップ予稿集 (1999).
14) 関根，江里口：固有表現の定義の困難さ—IREX における NE 定義の事例から，言語処理学会第 5 回年次大会，B2-1 (1999).
15) http://cs.nyu.edu/cs/projects/proteus/irex/
16) 落谷：組織名抽出のための知識収集，言語処理学会第 5 回年次大会，A2-2 (1999).
17) Sekine, S.: NYU: Description of the Japanese NE System Used For MET-2 (1998). http://www.muc.saic.com/
18) 木谷：固有名詞の特定機能を有する形態素解析，情報処理学会自然言語処理研究会 92-90 (1992).
19) 山田：統計情報を用いた日本語形態素解析，言語処理学会第 3 回年次大会 (1997).
20) 竹元，山田，若尾：日本語新聞記事からの固有名詞情報抽出，情報処理学会第 53 回全国大会予稿集，7L-3 (1996).
21) 福本，下畑，桝井：固有名詞抽出における日本語と英語の比較，情報処理学会自然言語処理研究会 126-15 (1998).

(平成 12 年 3 月 1 日受付)
(平成 13 年 3 月 9 日採録)

竹元 義美（正会員）

1990 年九州大学工学部情報工学科卒業．同年 NEC 入社．現在,（株）日本電気特許技術情報センターに出向中，同社情報サービス事業部インフォメーションサービス部主任．自然言語処理，情報検索，情報抽出等の分野に興味を持ち，これらを応用したサービスシステムの開発に従事．

福島 俊一（正会員）

1982 年東京大学理学部物理学科卒業，NEC 入社．現在，同社情報通信メディア研究本部インターネットシステム研究所研究マネージャー．日本語テキスト処理，自然言語解析，情報検索等の分野に関心を持ち，現在はとくに WWW サーチエンジンの研究開発に注力している．情報処理学会平成 4 年度論文賞，第 6 回坂井記念特別賞，第 45 回全国大会奨励賞，第 53 回全国大会優秀賞を受賞．人工知能学会，言語処理学会，情報知識学会，情報科学技術協会，ACM 各会員．工学博士．

山田 洋志（正会員）

1962 年生．1987 年北海道大学工学研究科電子工学専攻修士課程修了．同年 NEC 入社．現在，第二システム事業本部オープン共通システム開発部マネージャー．オープン系システムの開発に従事．言語処理学会会員．