# 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

内元 清貴† 馬 青† 村田 真樹†
小作 浩美† 内山 将夫† 井佐原 均†

本論文では，ME(最大エントロピー) モデルと書き換え規則を用いて固有表現を抽出する手法について述べる．固有表現の定義はIREX 固有表現抽出タスク (IREX-NE) の定義に基づくものとする．その定義によると，固有表現には一つあるいは複数の形態素からなるもの，形態素単位より短い部分文字列を含むものの 2 種類がある．複数の形態素からなる固有表現は，固有表現の始まり，中間，終りなどを表すラベルを 40 個用意し，各々の形態素に対し付与すべきラベルを推定することによって抽出する．ラベルの推定には ME モデルを用いる．この ME モデルでは学習コーパスで観測される素性と各々の形態素に付与すべきラベルとの関係を学習する．ここで素性とはラベル付与の手がかりとなる情報のことであり，我々の場合，着目している形態素を含む前後 2 形態素ずつ合計 5 形態素に関する見出し語，品詞の情報のことである．一方，形態素単位より短い部分文字列を含む固有表現は，ME モデルを用いてラベルを決めた後に書き換え規則を適用することによって抽出する．書き換え規則は学習コーパスに対するシステムの解析結果とコーパスの正解データとの差異を調べることによって自動獲得することができる．本論文では IREX-NE 本試験に用いられたデータに対し我々の手法を適用した結果を示し，さらにいくつかの比較実験から書き換え規則と精度，素性と精度，学習コーパスの量と精度の関係を明らかにする．

キーワード: 固有表現，最大エントロピー *(ME)* モデル，書き換え規則

# Named Entity Extraction Based on A Maximum Entropy Model and Transformation Rules

Kiyotaka Uchimoto† , Qing Ma† , Masaki Murata† , Hiromi Ozaku† ,
Masao Utiyama† and Hitoshi Isahara†

This paper describes a system for extracting named entities. The system is based on a ME (maximum entropy) model and transformation rules. Eight types of named entities are defined by IREX-NE, and each named entity consists of one or more morphemes, or it includes a substring of a morpheme. We define 40 named entity labels, which are at the beginning, the middle, or the end of a named entity, and extract a named entity which consists of one or more morphemes by estimating the labels according to the ME model. The trained ME model detects the relationship between features and named entity labels assigned to morphemes. The features are clues used for estimating labels. We use information about lexical items and parts-of-speech as features in the target morpheme. We also use information about lexical items and parts-of-speech in four morphemes, two on the left and two on the right of the target morpheme, as features. After estimating the named entity labels according to the

ME model, we extract a named entity, which includes a substring of a morpheme, by using transformation rules. These rules are automatically acquired by investigating the difference between named entity labels in a tagged corpus and those extracted by our system from the same corpus without tags. This paper also evaluates the relationships between transformation rules and accuracy, between features and accuracy, and between the amount of training data and accuracy by conducting several comparative experiments.

**KeyWords:** *named entity, maximum entropy (ME) model, transformation rule*

## 1 はじめに

固有表現 (NE = Named Entity) 抽出は情報抽出における基礎技術として認識されているだけでなく，形態素，構文解析の精度向上にもつながる重要な技術である．米国では 1980 年代から MUC(Message Understanding Conference)(SAIC 1998) のようなコンテストが行なわれ，その技術の向上が図られてきた．日本においても 1998 年からコンテスト形式のプロジェクト「IREX (Information Retrieval and Extraction Exercise)」が始められ，そのタスクの一つとして固有表現抽出が盛り込まれた．このタスクで固有表現として抽出するのは，「郵政省」のように組織の名称を表すもの，「小渕恵三」のように人名を表すもの，「神戸」のように地名を表すもの，「カローラ」のように固有物の名称を表すものおよび，「9 月 28 日」，「午後 3 時」，「100 万円」，「10%」のように日付，時間，金銭，割合を表す表現である．このように，固有名詞的表現だけでなく，時間表現，数値表現も抽出の対象としているため，本論文ではそれらをすべてまとめて固有表現と呼ぶ．このような固有表現は多種多様で，次々と新たに生み出されるためそのすべてを辞書に登録しておくことは不可能である．また，同じ表現でも，あるときは地名としてまたあるときは人名として使われるというようにタイプに曖昧性がある．そのため，テキストが与えられたときその中でどの部分がどのタイプの固有表現であるかを同定するのは容易ではない．

固有表現を抽出する方法には大きく分けると，人手で作成した規則に基づく方法と学習に基づく方法がある．固有表現の定義は抽出したものを何に応用するかによって異なってくるものであるため，前者の方法では定義が変わるたびに規則を人手で作成し直す必要がありコストがかかる．後者の方法は学習コーパスを作る必要があるが，データスパースネスに強い学習モデルを使えばそれほど大量のコーパスがなくても高い精度が得られる．そこで我々は後者の方法をとることにした．

この学習に基づく方法は英語での固有表現抽出の研究でも用いられている．例えば，HMM(Bikel, Miller, Schwartz, and Weischedel 1997; Miller, Crystal, Fox, Ramshaw, Schwartz, Stone, Weischedel, and the Annotation Group 1998)，決定木モデル (Cowie 1995)，ME(最大エントロピー) モデル (Borthwick, Sterling, Agichtein, and Grishman 1998a)，共起情報 (Lin 1998)，誤り駆動の書き換え規則 (Aberdeen, Burger, Day, Hirschman, Robinson, and Vilain 1995)

† 郵政省通信総合研究所, Communications Research Laboratory, Ministry of Posts and Telecommunications

などに基づくシステムがある．学習に基づく方法として MUC のコンテストで最も精度が高かったのは HMM に基づく Nymble という名のシステムである．このシステムは基本的に以下のような手法をとっている．まず学習では，MUC の NE タスクで定義された「PERSON」や「ORGANIZATION」などの固有表現およびそれ以外を表す「NOT-A-NAME」をそれぞれ状態として持つ状態遷移図を用意し，ある状態で，ある単語が入力されたときにどの状態に移るかを状態遷移確率として求める．そして，解析する際には，ビタビアルゴリズムを用いて，入力された単語列が辿り得る状態のパスうち，最適なパスを探索し，順次，辿った状態を出力することで固有表現を抽出する．他の学習手法を用いたシステムも確率の計算方法は違うが同様の手法をとっていることが多い．Borthwick らは，この学習に基づくシステムおよび人手で作成した規則に基づくシステムの中から，それぞれ MUC で比較的精度の高かったシステムを選びそれらを学習に基づく方法によって統合することによってより高い精度を得ている (Borthwick et al. 1998a)．あるデータに対しては人間のパフォーマンスを越えるような結果も得られている (Borthwick, Sterling, Agichtein, and Grishman 1998b)．

学習に基づく方法は固有表現抽出の研究以外に形態素解析や構文解析においてもよく用いられている (内元 馬 1999)．学習モデルとしては ME モデルを用いたものが優れた精度を得ていることが多く (Ratnaparkhi 1996, 1997; Uchimoto, Sekine, and Isahara 1999)，データスパースネスに強いため，我々は固有表現抽出においてもこの ME モデルを用いることにした．さらに後処理として，誤り駆動により獲得した書き換え規則を用いる．この書き換え規則を用いる手法は形態素解析でも用いられている (Brill 1995; 久光 丹羽 1998)．

固有表現の定義は IREX 固有表現抽出タスク (IREX-NE) の定義 (IREX 実行委員会 1999) に基づくものとする．その定義によると，固有表現には「日本」や「国立／公文書／館」(／は形態素の区切りを表す) のように一つあるいは複数の形態素からなるもの，あるいは「在米」の「米」，「兵庫／県内」の「兵庫県」のように形態素単位より短い部分文字列を含むものの 2 種類がある．

前者の固有表現は，固有表現の始まり，中間，終りなどを表すラベルを 40 個用意し，各々の形態素に対し付与すべきラベルを推定することによって抽出する．ラベルの推定には ME モデルを用いる．この ME モデルでは学習コーパスで観測される素性と各々の形態素に付与すべきラベルとの関係を学習する．ここで素性とはラベル付与の手がかりとなる情報のことであり，我々の場合，着目している形態素を含む前後 2 形態素ずつ合計 5 形態素に関する見出し語，品詞の情報のことである．ラベルを推定する際には，入力文を形態素解析し，ME モデルを用いてそれぞれの形態素ごとにそこで観測される素性から各ラベルの尤もらしさを確率として計算し，一文全体における確率の積の値が高くなり，かつラベルとラベルの間の連接規則を満たすように各々の形態素に付与するラベルを決める．一文における最適解の探索にはビタビアルゴリズムを用いる．

一方，後者の固有表現のように形態素単位より短い部分文字列を含む固有表現は上記の方法で

は抽出できないので，ME モデルを用いてラベルを決めた後に書き換え規則を適用することによって抽出する．書き換え規則は学習コーパスに対するシステムの解析結果とコーパスの正解データとの差異を調べることによって自動獲得することができる．一つあるいは複数の形態素からなる固有表現についても同様に書き換え規則を適用することは可能であるが，本論文では ME モデルについてはラベル付けの精度に重点を置き，書き換え規則についてはできるだけ簡便な獲得方法を用いて効果をあげることに重点を置く．

本論文では IREX-NE 本試験に用いられたデータに対し我々の手法を適用した結果を示し，さらにいくつかの比較実験から ME モデルにおける素性と精度の関係，学習コーパスの量と精度の関係，さらに簡便な方法を用いて自動獲得した書き換え規則がどの程度精度に貢献するかを明らかにする．

## 2　固有表現抽出アルゴリズム

### 2.1　アルゴリズムの概要

固有表現は IREX-NE の定義にしたがい，表 1 の 8 種類とする．この節ではこの表にあげた SGML タグを付与する方法について述べる．

表 1　固有表現のタグ

| | 開始位置タグ | 終了位置タグ | 例 |
|---|---|---|---|
| 固有名詞的表現 | | | |
| 組織名，政府組織名 | <ORGANIZATION> | </ORGANIZATION> | 郵政省，ニューヨーク大学，毎日新聞，ＩＲＥＸ実行委員会 |
| 人名 | <PERSON> | </PERSON> | 長尾眞，グリッシュマン，若ノ花 |
| 地名 | <LOCATION> | </LOCATION> | 日本，神戸，井の頭線，富士山 |
| 固有物名 | <ARTIFACT> | </ARTIFACT> | ノーベル賞，ＰＬ法案，特殊相対性理論，ペンティアム２００ＭＨｚ，カローラ |
| 時間表現 | | | |
| 日付表現 | <DATE> | </DATE> | ９月２８日，去年，ある秋 |
| 時間表現 | <TIME> | </TIME> | 午後５時２５分，未明，明け方 |
| 数値表現 | | | |
| 金額表現 | <MONEY> | </MONEY> | １ドル，数十兆円，五千から六千万円 |
| 割合表現 | <PERCENT> | </PERCENT> | ２０％，５割，５分の１，２倍 |

本手法では以下の手順で固有表現を抽出する．

(1)　テキストを形態素解析する．
実験では形態素解析に JUMAN(黒橋 長尾 1998) を用いた．例えば，“在米女性を中心に「人権を考える会」ができ，...” という部分は表 2 の第 1 行のように形態素ごとに区

切られ，それぞれの形態素ごとに第 2 行，第 3 行のような品詞の情報が得られる．

表 **2**　ME モデルを用いたラベル付与の例

| 見出し語 | | 在米 | 女性 | を | 中心 | に | 「 | 人権 | を | 考える | 会 | 」 | が | でき | ， | スコア |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 品詞 (大分類) | | 名詞 | 名詞 | 助詞 | 名詞 | 助詞 | 特殊 | 名詞 | 助詞 | 動詞 | 名詞 | 特殊 | 助詞 | 動詞 | 特殊 | |
| 品詞 (細分類) | | 普通名詞 | 普通名詞 | 格助詞 | 普通名詞 | 格助詞 | 括弧始 | 普通名詞 | 格助詞 | | 普通名詞 | 括弧終 | 格助詞 | | 読点 | |
| ラベルの候補 | 1 | OTHER | OTHER | OTHER | OTHER | OTHER | PRE | ORG:BEGIN | ORG:MIDDLE | ORG:MIDDLE | ORG:END | POST | OTHER | OTHER | OTHER | 0.8 |
| | 2 | OTHER | OTHER | OTHER | OTHER | OTHER | PRE | ART:SINGLE | POST | OTHER | OTHER | OTHER | OTHER | OTHER | OTHER | 0.7 |
| | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |

(表で「ORG」「ART」はそれぞれ「ORGANIZATION」「ARTIFACT」の略である．)

(2)　各形態素にラベルを付与する．
ラベルとしては，以下の合計 40 個を用意した．

(a)　IREX-NE で定義されている固有表現のタグに「OPTIONAL」を加えた 9 種類を，固有表現の始まり，中間，終り，単独に分けた 9×4=36 個．例えば人名のタグの場合，それぞれ「PERSON:BEGIN」「PERSON:MIDDLE」「PERSON:END」「PERSON:SINGLE」を用いる．このように分けたのは，複数の形態素が一つの固有表現を構成することがあることを考慮するためである．

「OPTIONAL」のタグはタグ付けが判定者にも困難な場合のために設けたものである．これも IREX-NE における定義にしたがっている．固有表現の判定は人間にも難しいことが多い．例えば，「東京高裁」は LOCATION か ORGANIZATION か，「日経平均株価」と言ったときの「日経」は ORGANIZATION とするべきかなどがそうである．このような場合，それぞれ「東京高裁」，「日経」にこのタグを付与し，固有表現としては抽出しない．この「OPTIONAL」をラベルとして考慮したのはその性質を学習することによって，例えば LOCATION か ORGANIZATION の判定が困難なものをいずれかのタグに分類してしまうのを避けることができると考えたためであ

る．

(b) 固有表現の前後の1形態素および固有表現に挟まれた1形態素を他の形態素と区別するための3個(「PRE」「POST」「MID」)．例えば，“昨日大阪と神戸で...”という部分では「大阪」と「神戸」がそれぞれ地名を表す固有表現であり，その前後の形態素は次のようにラベル付けされる．

“昨日(PRE)／大阪(LOCATION:SINGLE)／と(MID)／神戸(LOCATION:SINGLE)／で(POST)...”

(括弧内はそれぞれ前の形態素に付与されたラベルの候補)

この三つのラベル「PRE」「POST」「MID」を用いたのは，固有表現の前後の形態素(接辞など)は固有表現を抽出する際に手がかりとなることが多いため，次にあげる「OTHER」と区別する方が良いと考えたからである．

(c) 以上のどのラベルもつかない「OTHER」．

今，一文が $n$ 個の形態素からなるとする．手順(1)で得られた形態素解析結果を用いて，個々の形態素 $m_i(1 \leq i \leq n)$ にそれぞれ上記のラベルのいずれかを付与する．形態素 $m_i$ に付与するラベルはコーパスから学習したMEモデルから各ラベルを付与したときの尤もらしさを確率として計算しそれを基に決める．詳しくは，モデルについては2.2節で，最適解の探索アルゴリズムについては2.3節で述べる．

(3) 書き換え規則による後処理

JUMANの解析結果における形態素の境界とIREXで定義されている固有表現の境界は必ずしも一致しない．このような一致しない場合に対応するために書き換え規則を自動獲得し，獲得した規則を用いて後処理を行う．例えば，表 2 の「在米」に対しては以下のような書き換え規則が適用される．

| | | | | |
|---|---|---|---|---|
| 見出し語 | 在米 | ⇒ | 在 | 米 |
| 品詞(大分類) | 名詞 | | 名詞 | 名詞 |
| 品詞(細分類) | 普通名詞 | | 普通名詞 | 普通名詞 |
| ラベル | OTHER | | PRE | LOCATION:SINGLE |

書き換え規則の自動獲得手法については2.4節で述べる．

(4) ラベルを固有表現のタグに変換

すべてのラベルが決まったら，それぞれのラベルに対し手順(2)で定義したラベルの定義にしたがって，ラベルからIREX-NEで定義されたタグへと変換する．抽出したい固有表現は表 1 の8種類なので，最後に解析結果から「OPTIONAL」のタグを取り除く．

例えば，表 2 でスコアが最大であるラベル候補1 の場合，手順(3)の操作によって“在米(OTHER)”の部分が“在(PRE)米(LOCATION:SINGLE)”(括弧内はそれぞれ前の形態素に付

与されたラベルの候補) に書き換えられる．そして，ラベルをタグに変換することによって次のような出力を得る．

“在 <LOCATION> 米 < /LOCATION> 女性を中心に
「<ORGANIZATION> 人権を考える会 < /ORGANIZATION>」ができ，”

### 2.2　固有表現抽出に用いる確率モデル

この節では形態素に付与するラベルの尤もらしさを確率として計算するためのモデルについて述べる．モデルとしては，ME(最大エントロピー法) に基づく確率モデルを採用する．まず，ME の基本について説明し，その後，ME に基づく固有表現ラベル付与確率モデルおよびそのモデルをコーパスから統計的に学習する方法について述べる．

#### ME(最大エントロピー) モデル

一般に確率モデルでは，文脈 (観測される情報のこと) とそのときに得られる出力値との関係は既知のデータから推定される確率分布によって表される．いろいろな状況に対してできるだけ正確に出力値を予測するためには文脈を細かく定義する必要があるが，細かくしすぎると既知のデータにおいてそれぞれの文脈に対応する事例の数が少なくなりデータスパースネスの問題が生じる．

ME モデルでは，文脈は素性と呼ばれる個々の要素によって表され，確率分布は素性を引数とした関数として表される．そして，各々の素性はトレーニングデータにおける確率分布のエントロピーが最大になるように重み付けされる．このエントロピーを最大にするという操作によって，既知データに観測されなかったような素性あるいはまれにしか観測されなかった素性については，それぞれの出力値に対して確率値が等確率になるようにあるいは近付くように重み付けされる．このように未知のデータに対して考慮した重み付けがなされるため，ME モデルは比較的データスパースネスに強いとされている．このモデルは例えば言語現象などのように既知データにすべての現象が現れ得ないような現象を扱うのに適したモデルであると言える．

以上のような性質を持つ ME モデルでは，確率分布の式は以下のように求められる．文脈の集合を $B$，出力値の集合を $A$ とするとき，文脈 $b(\in B)$ で出力値 $a(\in A)$ となる事象 $(a,b)$ の確率分布 $p(a,b)$ を ME により推定することを考える．文脈 $b$ は $k$ 個の素性 $f_j (1 \leq j \leq k)$ の集合で表す．そして，文脈 $b$ において，素性 $f_j$ が観測されかつ出力値が $a$ となるときに 1 を返す以下のような関数を定義する．

$$g_j(a,b) = \begin{cases} 1, \text{ if } exist(b,f_j) = 1 \ \& \ \text{出力値} = a \\ 0, \text{ それ以外} \end{cases} \tag{1}$$

これを素性関数と呼ぶ．ここで，$exist(b,f_j)$ は，文脈 $b$ において素性 $f_j$ が観測されるか否かによって 1 あるいは 0 の値を返す関数とする．

次に，それぞれの素性が既知のデータ中に現れた割合は未知のデータも含む全データ中におい

ても変わらないとする制約を加える．つまり，推定するべき確率分布 $p(a,b)$ による素性 $f_j$ の期待値と，既知データにおける経験確率分布 $\tilde{p}(a,b)$ による素性 $f_j$ の期待値が等しいと仮定する．これは以下の制約式で表せる．

$$\sum_{a\in A,b\in B} p(a,b)g_j(a,b) = \sum_{a\in A,b\in B} \tilde{p}(a,b)g_j(a,b) \quad for\ \forall f_j\ (1 \leq j \leq k) \tag{2}$$

この式で，$p(a,b) = p(b)p(a|b) \approx \tilde{p}(b)p(a|b)$ という近似を行ない以下の式を得る．

$$\sum_{a\in A,b\in B} \tilde{p}(b)p(a|b)g_j(a,b) = \sum_{a\in A,b\in B} \tilde{p}(a,b)g_j(a,b) \quad for\ \forall f_j\ (1 \leq j \leq k) \tag{3}$$

ここで，$\tilde{p}(b)$，$\tilde{p}(a,b)$ は，$freq(b)$，$freq(a,b)$ をそれぞれ既知データにおける事象 $b$ の出現頻度，出力値 $a$ と事象 $b$ の共起頻度として以下のように推定する．

$$\tilde{p}(b) = \frac{freq(b)}{\sum_{b\in B} freq(b)} \tag{4}$$

$$\tilde{p}(a,b) = \frac{freq(a,b)}{\sum_{a\in A,b\in B} freq(a,b)} \tag{5}$$

次に，式 (3) の制約を満たす確率分布 $p(a,b)$ のうち，エントロピー

$$H(p) = -\sum_{a\in A,b\in B} \tilde{p}(b)p(a|b)\ log\,(p(a,b)) \tag{6}$$

を最大にする確率分布を推定するべき確率分布とする．これは，式 (3) の制約を満たす確率分布のうちで最も一様な分布となる．このような確率分布は唯一存在し，以下の確率分布 $p^*$ として記述される．

$$p^*(a|b) = \frac{\prod_{j=1}^{k} \alpha_{a,j}^{g_j(a,b)}}{\sum_{a\in A} \prod_{j=1}^{k} \alpha_{a,j}^{g_j(a,b)}} \quad (0 \leq \alpha_{a,j} \leq \infty) \tag{7}$$

ただし，

$$\alpha_{a,j} = e^{\lambda_{a,j}} \tag{8}$$

であり，$\lambda_{a,j}$ は素性関数 $g_j(a,b)$ の重みである．この重みは文脈 $b$ のもとで出力値 $a$ となることを予測するのに素性 $f_j$ がどれだけ重要な役割を果たすかを表している．訓練集合が与えられたとき，$\lambda_{a,j}$ の推定には Improved Iterative Scaling(IIS) アルゴリズム (Pietra, Pietra, and Lafferty 1995) などが用いられる．式 (7) の導出については文献 (Jaynes 1957, 1979) を参照されたい．

固有表現ラベル付与確率モデル

2.1 節に，個々の形態素に付与すべき固有表現のラベルを定義した．以降では，形態素にそれぞれのラベルを付与したときの尤もらしさを表す確率をラベルの付与確率と呼ぶ．

一文が $n$ 個の形態素からなるとき，形態素 $m_i(1 \leq i \leq n)$ にラベル $l_j(0 \leq j \leq 39)$ を付与するときの付与確率は，前節で述べた ME モデルの式（7）を用いて $p^*(l_j|F_i)$ で求められる．ここで $F$ は「見出し語：人権, 品詞 (大分類)：名詞, 品詞 (細分類)：普通名詞」などの素性の集合であり，個々の $m_i$ ごとに異なるため $F_i$ と表した．一文全体の付与確率は個々の確率の積で表す．

素性

基本的に学習コーパスから得られる形態素情報を素性として用いる．実験では，着目している形態素を含む前後 2 形態素ずつ合計 5 形態素に関する見出し語，品詞 (大分類，細分類) とした．品詞分類はそれぞれ大分類 15 個，細分類 48 個である．これは JUMAN のものにしたがった．学習コーパスに 1，2 回しか現れないような素性はノイズとなる可能性があるのでそれを避けるために頻度による素性選択を行なう．見出し語としては学習コーパス中に 5 回以上現れたものを用い，さらに式 (1) の素性関数としては学習コーパスに 3 回以上観測されたものを用いる．見出し語を 5 回以上現れたものとしたのはこれ以上少なくすると素性の数が増え現在のマシンパワーでは学習できなかったためである．素性としては他にもいろいろ考えられるが，今回の実験では学習コーパスから得られる情報でかつ着目している形態素の周辺の情報のみを用いた場合にどの程度の精度が得られるかを調べることに重きを置いた．

さらに学習コーパス以外から得られる情報の有効性も調べるために追加実験として，固有名詞に関する辞書情報を利用しその辞書に登録されているかどうかを素性として利用した場合の実験も行なった．これについては 3 節で実験結果をあげて考察する．

### 2.3 ビタビアルゴリズム

本節ではラベル付与に用いるビタビアルゴリズムについて説明する．このアルゴリズムは，スコアが一文全体で最適値となるようにラベルを付与するものである．

形態素 $m_i$ に対し，2.2 節で述べたラベル $l_j$ の付与確率 $p^*(l_j|F_i)$ の一文全体における掛け算 $\sum_{i=1}^{n} p^*(l_j|F_i)$ が最大になるように各ラベルを決める．ただし，表 3 の連接規則を満たすようにする．この表で，$ (文末)，# (文頭) は便宜上設けたもので実際に付与するラベルとは異なる．表 3 の連接規則は人手で作成した．

手順は以下の通りである．

(1) 文頭の形態素 $m_1$ に対し各ラベル $l_{j(1)}$ の付与確率 $p^*(l_{j(1)}|F_1)$ $(0 \leq j(1) \leq 39)$ を計算し，それぞれ各ラベルごとのスコア $S_1(l_{j(1)})$ とする．つまり，$S_1(l_1) = p^*(l_1|F_1)$, $S_1(l_2) = p^*(l_2|F_1), \ldots, S_1(l_{39}) = p^*(l_{39}|F_1)$ とする．

表 3　連接規則

| ラベル | 左方に接続可能なラベル | 右方に接続可能なラベル |
|---|---|---|
| $x$ | # (文頭), $x$, $y$, $x$:END, $y$:END, PRE, MID | $ (文末), $x$, $y$, $x$:BEGIN, $y$:BEGIN, POST, MID |
| $x$:BEGIN | # (文頭), $x$, $y$, $x$:END, $y$:END, PRE, MID | $x$:MIDDLE, $x$:END |
| $x$:MIDDLE | $x$:BEGIN, $x$:MIDDLE | $x$:MIDDLE, $x$:END |
| $x$:END | $x$:BEGIN, $x$:MIDDLE | $ (文末), $x$, $y$, $x$:BEGIN, $y$:BEGIN, POST, MID |
| MID | $x$, $x$:END | $x$, $x$:BEGIN |
| PRE | # (文頭), POST, OTHER | $x$, $x$:BEGIN |
| POST | $x$, $x$:END | $ (文末), PRE, OTHER |
| OTHER | # (文頭), POST, OTHER | $ (文末), PRE, OTHER |
| $ (文末) | $x$, $x$:END, POST, OTHER | |
| # (文頭) | | $x$, $x$:BEGIN, PRE, OTHER |

($x$, $y$ は IREX-NE で定義された 8 種類のタグおよび「OPTIONAL」に対応する．)

(2)　次の形態素 $m_2$ に対し各ラベルの付与確率 $p^*(l_{j(2)}|F_2)$ $(0 \leq j(2) \leq 39)$ を計算し，それぞれ各ラベルごとのスコアを

$$S_2(l_{j(2)}) \quad = \quad \max_{l_{j(1)}} \; p^*(l_{j(2)}|F_2) \times S_1(l_{j(1)})$$

とする．ただし，$l_{j(1)}$ と $l_{j(2)}$ が連接規則を満たすものに限る．

(3)　さらに次の形態素 $m_3$ に対しても同様に各ラベルの付与確率 $p^*(l_{j(3)}|F_3)$ $(0 \leq j(3) \leq 39)$ を計算し，それぞれ各ラベルごとのスコアを

$$S_3(l_{j(3)}) \quad = \quad \max_{l_{j(2)}} \; p^*(l_{j(3)}|F_3) \times S_2(l_{j(2)})$$

とする．ただし，$l_{j(2)}$ と $l_{j(3)}$ が連接規則を満たすものに限る．

(4)　同様のことを文末まで繰り返し，

$$S_n(l_{j(n)}) \quad = \quad \max_{l_{j(n-1)}} \; p^*(l_{j(n)}|F_n) \times S_{n-1}(l_{j(n-1)})$$

のうち最大のものを選ぶと，最適解であるラベルの並び $l_{j(1)}, l_{j(2)}, \ldots, l_{j(n)}$ が得られる．

### 2.4　自動獲得した書き換え規則による後処理

ME モデルを用いたラベル付けの処理が終った後で，形態素解析により得られる形態素が固有表現より長い場合に対処するため，予め用意しておいた書き換え規則を適用する．書き換え規則は Brill が品詞タグ付けに用いた (Brill 1995) のと同様の手法である誤り駆動で獲得する． Brill の規則獲得方法との違いは Brill がテンプレートを用いているのに対して我々は用いていない点である．我々の場合，書き換え規則は学習コーパスに対するシステムの解析結果とコーパスの正解データとの差異を調べることによって自動獲得することができる．差異の中から，コーパスでは同じ文字列に対応しているにもかかわらず形態素の数が異なる部分をすべて抽出し書き換え規則として利用する．ただし，前件部は同じであるが後件部が異なるような規則が複数獲得された場合は，最も頻度の高い規則のみを用いる．最も頻度の高い規則が複数種類ある場合，それらの規則はすべて捨てる．さらに，ここで獲得した規則を学習コーパスに対するシステムの解析結果に適用し，誤りとなる数が正解となる数以上であるものはすべて捨てる．例えば，以下のようなものが書き換え規則として獲得される．

| | 前件部 | | 後件部 | |
|---|---|---|---|---|
| 見出し語 | 在日 | | 在 | 日 |
| 品詞 (大分類) | 名詞 | ⇒ | 名詞 | 名詞 |
| 品詞 (細分類) | サ変名詞 | | 普通名詞 | 普通名詞 |
| ラベル | OTHER | | PRE | LOCATION:SINGLE |

## 3　実験と考察

### 3.1　実験データ

モデルの学習に用いたデータは， CRL(郵政省通信総合研究所) 固有表現データ， IREX-NE 予備試験トレーニングデータ， IREX-NE 予備試験データ， IREX-NE 本試験逮捕トレーニングデータの合計約 12,000 文である．試験に用いたデータは IREX-NE 本試験データである[1]．これらはすべて毎日新聞のデータに対して固有表現のタグが付与されたものである．以下で簡単にデータの説明をする．

#### 学習データ

学習コーパスの書式は SGML 形式で，各固有表現には表 1 のタグが付与されている．これらのタグ付コーパスからテキスト部分を取り出して形態素解析し，表 4 にあげる変換規則を用いて各形態素にラベルが付与されたものに変換した後，学習に用いた．

1 今回学習，試験に用いたデータはすべて IREX のホームページ (IREX 実行委員会 1999) から入手可能である．

表 4　タグからラベルへの変換規則

| 一つ前の形態素のラベル | 形態素と前後のタグ | | | | 形態素 (ラベル) |
|---|---|---|---|---|---|
| | $<x>$ | $m$ | $</x>$ | $\Rightarrow$ | $m$ ($x$:SINGLE) |
| | $<x>$ | $m$ | | $\Rightarrow$ | $m$ ($x$:BEGIN) |
| | | $m$ | $</x>$ | $\Rightarrow$ | $m$ ($x$:END) |
| | $</x>$ | $m$ | $<y>$ | $\Rightarrow$ | $m$ (MID) |
| | $</x>$ | $m$ | | $\Rightarrow$ | $m$ (POST) |
| | | $m$ | $<x>$ | $\Rightarrow$ | $m$ (PRE) |
| $x$:BEGIN | | $m$ | | $\Rightarrow$ | $m$ ($x$:MIDDLE) |
| $x$:MIDDLE | | $m$ | | $\Rightarrow$ | $m$ ($x$:MIDDLE) |
| | | $m$ | | $\Rightarrow$ | $m$ (OTHER) |

**CRL**(郵政省通信総合研究所) 固有表現データ　毎日新聞 1995 年 1 月 1 日から 10 日までの全記事，約 1 万文に対して，固有表現をタグ付けしたデータである．固有表現は IREX-NE の 1999 年 2 月 14 日に更新された定義に基づいている．

**IREX-NE** 予備試験トレーニングデータ　毎日新聞 1994 年 4 月 13 日の 46 記事，約 500 文に対して固有表現をタグ付けしたデータである．固有表現は IREX-NE の 1998 年 10 月 27 日に更新された定義に基づいている．

**IREX-NE** 予備試験データ　毎日新聞 1994 年 9 月 11 日の 36 記事，約 500 文に対して固有表現をタグ付けしたデータである．固有表現は IREX-NE の 1998 年 10 月 27 日に更新された定義に基づいている．

IREX-NE の定義は少しずつ更新されている．しかし，それらの定義の違いによるノイズの数は人間が学習データを作成するときに生じるノイズの数とさほど違いはないと考え，すべて学習に用いた．

本試験データ

1999 年 4 月 14 日から 5 月 13 日の毎日新聞のデータから選ばれた 91 記事で，ドメインを限らないもの 71 記事，約 400 文 (以下，「一般ドメイン」と呼ぶ) と「逮捕」にドメインを限ったもの 20 記事，約 100 文 (以下，「限定ドメイン」と呼ぶ) の 2 種類のデータからなる．それぞれのドメインのデータにおける固有表現の数は表 5 の通りである．

表 5　IREX-NE 本試験データに対する固有表現の内訳

| | ARREST | GENERAL |
|---|---|---|
| 固有表現 (NE) | 個数 (個) | 個数 (個) |
| ORGANIZATION | 74 | 361 |
| PERSON | 97 | 338 |
| LOCATION | 106 | 413 |
| ARTIFACT | 13 | 48 |
| DATE | 72 | 260 |
| TIME | 19 | 54 |
| MONEY | 8 | 15 |
| PERCENT | 0 | 21 |
| OPTIONAL | 8 | 86 |
| 合計 | 389 | 1510 |

### 3.2　実験結果

実験に用いた素性は形態素解析結果から得られる情報であり，着目している形態素を含む前後 2 形態素ずつ合計 5 形態素に関する見出し語，品詞 (大分類，細分類) である．見出し語としては学習コーパス中に 5 回以上現れた 12,368 個を用いた．品詞分類は JUMAN のものにしたがった．それぞれ大分類 15 個，細分類 48 個である．このうち，式 (1) の素性関数としては学習コーパスに 3 回以上観測されたもの 27,370 個を用いた．モデルの重み (式 (8) の $\lambda_{a,j}$) の学習には Ristad のツール (Ristad 1998) を利用した[2]．

次に我々の解析結果を表 6 に示す．この表で第 1 列と第 2 列は書き換え規則を利用したときの限定ドメインの試験 (ARREST) に対する結果とドメインを限定しない試験 (GENERAL) に対する結果である．第 3 列と第 4 列は書き換え規則を利用しなかったときのそれぞれのドメインに対する結果である．どちらのドメインに対しても特別なチューニングはしなかった．精度はどちらのドメインに対しても書き換え規則を用いたときの方が良く，F-measure [3]は限定ドメインに対して 83.91，一般ドメインに対して 79.42 であった．IREX-NE の試験では「OPTIONAL」のタグが振られたものについては，その範囲内にシステムがどのような結果を出そうとも，評価には反映されない．ただし，範囲外にずれて重なっている場合には不正解とされる．本論文においてもこの

2 現在このツールは公開されていない．参考文献 (Ristad 1997) を参照．

3 F-measure としては以下の定義のものを用いる．

$$\mathrm{F-measure} \quad = \quad \frac{2 \times \text{再現率} \times \text{適合率}}{\text{再現率} + \text{適合率}}$$

評価方法にしたがった[4]．

表 6 実験結果

| 固有表現 (NE) | 書き換え規則を利用 | | | | 書き換え規則を利用しない | | | |
|---|---|---|---|---|---|---|---|---|
| | ARREST | | GENERAL | | ARREST | | GENERAL | |
| | 再現率 (%) | 適合率 (%) | 再現率 (%) | 適合率 (%) | 再現率 (%) | 適合率 (%) | 再現率 (%) | 適合率 (%) |
| ORGANIZATION | 59.46 | 81.48 | 59.28 | 79.55 | 59.46 | 81.48 | 58.73 | 81.85 |
| PERSON | 84.54 | 84.54 | 76.92 | 83.87 | 84.54 | 84.54 | 76.92 | 83.87 |
| LOCATION | 83.02 | 81.48 | 76.27 | 84.45 | 73.58 | 77.23 | 69.73 | 82.52 |
| ARTIFACT | 61.54 | 66.67 | 35.42 | 50.00 | 61.54 | 66.67 | 35.42 | 50.00 |
| DATE | 97.22 | 97.22 | 91.15 | 94.80 | 97.22 | 97.22 | 90.38 | 94.76 |
| TIME | 94.74 | 100.00 | 87.04 | 94.00 | 94.74 | 100.00 | 87.04 | 94.00 |
| MONEY | 100.00 | 100.00 | 93.33 | 93.33 | 100.00 | 100.00 | 93.33 | 93.33 |
| PERCENT | - | - | 100.00 | 95.45 | - | - | 80.95 | 94.44 |
| 総合 | 81.75 | 86.18 | 74.50 | 85.03 | 79.18 | 85.08 | 72.19 | 84.96 |
| F-measure | 83.91 | | 79.42 | | 82.02 | | 78.05 | |

### 3.3 書き換え規則と精度

書き換え規則の適用対象となる固有表現は形態素単位より短い部分文字列を含むもので，本試験データでは限定ドメインに 18 個，一般ドメインに 79 個あった．いずれも本試験データ全体の約 5% に相当する．学習コーパスから得られた書き換え規則の数は 362 個であり，そのうち限定ドメインの試験には 9 個の規則が延べ 11 回適用され誤りが 1 個 (再現率 56%(10/18)，適合率 91%(10/11))，一般ドメインの試験には 12 個の規則が延べ 42 回適用され誤りが 10 個 (再現率 41%(32/79)，適合率 76%(32/42)) であった．誤りは以下のようなものであった．

- 本来抽出するべき固有表現の部分文字列を誤って抽出してしまう場合 (1 個)．
  「在日米軍横田基地」から「日」だけが LOCATION として抽出されていた．これは，IREX-NE の定義によると「在日米軍横田基地」が LOCATION として抽出されるべきであるが，ME モデルを用いたラベル付けによってうまく抽出できなかった結果，書き換え規則が適用され誤って抽出されてしまった例である．このような誤りをなくすためには，ME モデルを用いたラベル付けの精度を向上する必要がある．

4 OPTIONAL を正解とする評価もあり得るが，その場合，タスクは固有表現の抽出ではなく，人間が固有表現かどうか迷う部分の特定ということになる．したがって，その評価は他の固有表現の抽出結果の評価とは別にするべきである．本論文の目的は固有表現の抽出であるため，OPTIONAL を正解とする評価は対象外とした．

- 学習コーパスでは固有表現となっていたが，正解データでは固有表現となっていない場合 (10 個)．
  学習コーパスでは「邦人」の「邦」が LOCATION，「外相会談」の「外」が ORGANIZATION となっていたが，本試験の正解データでは固有表現とみなされていなかった．このような誤りをなくすためには学習コーパスの整備が必要である．

書き換え規則を用いることにより，F-measure で 2 ポイント (限定ドメイン) および 1.5 ポイント (一般ドメイン) 程度の精度向上がみられた．ここで用いた書き換え規則は形態素の境界と IREX で定義されている固有表現の境界が一致しない場合にのみ対応するために獲得したものである．一致する場合についても同様の書き換え規則を適用することは可能であるが，そうした場合の追加実験では F-measure で 72.23(限定ドメイン) および 73.12(一般ドメイン) と書き換え規則を用いない場合に比べてそれぞれ 10 ポイントおよび 5 ポイント程度精度が悪くなった．これは我々の用いた簡便な獲得手法では，ME モデルにより付与したラベルを精度良く書き換えられるほどの規則を獲得できないことを示している．しかし，得られた精度向上および規則獲得の簡便さを考慮すると，ME モデルで抽出できない部分を補う方法としては有効な方法であると言える．

本手法では形態素解析が終った後に固有表現のラベルを推定するが，形態素解析の段階で形態素の文法的属性 (品詞など) と固有表現のラベルを同時に推定するような方法をとることも考えられる．この場合でも書き換え規則は形態素解析の後処理として利用されている久光らの研究 (久光・丹羽 1998) と同様にして使えると考えられる．

### 3.4　素性と精度

実験に用いた素性の有効性を調べるために，それぞれの素性を削除したときの比較実験を行なった．比較実験ではすべて書き換え規則を用いた．結果を表 7 にあげる． 表中の F というのは F-measure のことで，精度の差というのは，着目している形態素とその前後 2 形態素ずつについて，見出し語，品詞大分類，品詞細分類のすべての情報を素性として用いたときの精度と比べたときの差を意味する．どの素性を削除した場合にも精度が悪くなっており，どの素性も精度の向上に貢献していることが分かる．特に見出し語は精度向上に著しく貢献している．

表 7 の下から三行は，前後の形態素情報の利用範囲を変更したときの結果であり，着目している形態素の情報のみ (表では「(0) のみ」と示す．)，着目している形態素とその前後の形態素の情報のみ (表では「(-1)(0)(1)」と示す．)，着目している形態素とその前後 2 形態素の情報 (表では「(-2) から (2)」と示す．表 6 に示した結果と同じ．)，着目している形態素とその前後 3 形態素の情報 (表では「(-3) から (3)」と示す．) をそれぞれ素性として用いたときの精度を表す．用いる情報が前後 2 形態素ずつより多くても少なくても精度が悪くなった．用いる情報を多くしたにもかかわらず精度が悪くなるのは，データスパースネスの問題が深刻になってくるためであると考えられる．

表 7　素性と精度の関係 (書き換え規則利用)

| 素性 | ARREST | | | | GENERAL | | | |
|---|---|---|---|---|---|---|---|---|
| | 再現率 (%) | 適合率 (%) | F | 精度の差 | 再現率 (%) | 適合率 (%) | F | 精度の差 |
| すべて | 81.75 | 86.18 | 83.91 | 0 | 74.50 | 85.03 | 79.42 | 0 |
| 見出し語のみ | 73.26 | 80.97 | 76.92 | -6.99 | 62.58 | 74.29 | 67.94 | -11.48 |
| 品詞大分類のみ | 5.40 | 70.00 | 10.02 | -73.89 | 2.85 | 42.16 | 5.33 | -74.09 |
| 品詞細分類のみ | 51.41 | 62.50 | 56.42 | -27.49 | 45.23 | 61.31 | 52.06 | -27.36 |
| 見出し語削除 | 51.41 | 63.49 | 56.82 | -27.09 | 46.16 | 65.45 | 54.14 | -25.28 |
| 品詞大分類削除 | 80.46 | 85.99 | 83.13 | -0.78 | 72.91 | 82.29 | 77.32 | -2.10 |
| 品詞細分類削除 | 76.09 | 87.57 | 81.43 | -2.48 | 66.89 | 82.72 | 73.97 | -5.45 |
| (0) のみ | 31.11 | 48.79 | 37.99 | -45.92 | 35.56 | 70.57 | 47.29 | -32.13 |
| (-1)(0)(1) | 76.86 | 84.46 | 80.48 | -3.43 | 72.32 | 85.11 | 78.20 | -1.22 |
| (-2) から (2) | 81.75 | 86.18 | 83.91 | 0 | 74.50 | 85.03 | 79.42 | 0 |
| (-3) から (3) | 80.72 | 85.09 | 82.85 | -1.06 | 73.38 | 84.19 | 78.41 | -1.01 |

### 3.5　学習コーパスと精度

この節では，学習コーパスと解析精度の関係について考察する．まず，図 1，図 2に学習コーパスとテストコーパスのそれぞれを解析した場合の学習コーパスの量と解析精度の関係をあげる．図の横軸は学習コーパスの文数，縦軸は F-measure を表す．学習コーパスの解析には限定ドメインとして IREX-NE 本試験逮捕トレーニングデータ，一般ドメインとして IREX-NE 予備試験データを用いた．図では限定ドメイン，一般ドメインに対するグラフにはそれぞれ「arrest」，「general」，書き換え規則を用いた場合と用いなかった場合にはそれぞれ「with_rules」，「without_rules」という表記を用いている．

テストコーパスに対する学習曲線 (図 2) を見ると，特に一般ドメインに対してはまだ精度は飽和していないようである．学習コーパスに対する学習曲線 (図 1) もわずかではあるが増加する傾向にある．したがって，少なくとも一般ドメインに対しては学習コーパスの量が増えればもう少し精度の向上が期待できそうである．

### 3.6　関連研究との比較

1999 年 5 月 13 日から 17 日にかけて，IREX-NE の本試験が行なわれた．試験は 13 日に実行委員長より問題が配布され，17 日までに各々のシステムのタグ付け結果を電子メイルで送り返すという形式で行なわれた．IREX-NE 本試験に参加したシステムは 15 システムであった．それらをパターン駆動型，学習型，それらの組み合わせの 3 種類に分類すると，我々のシステムは学習型

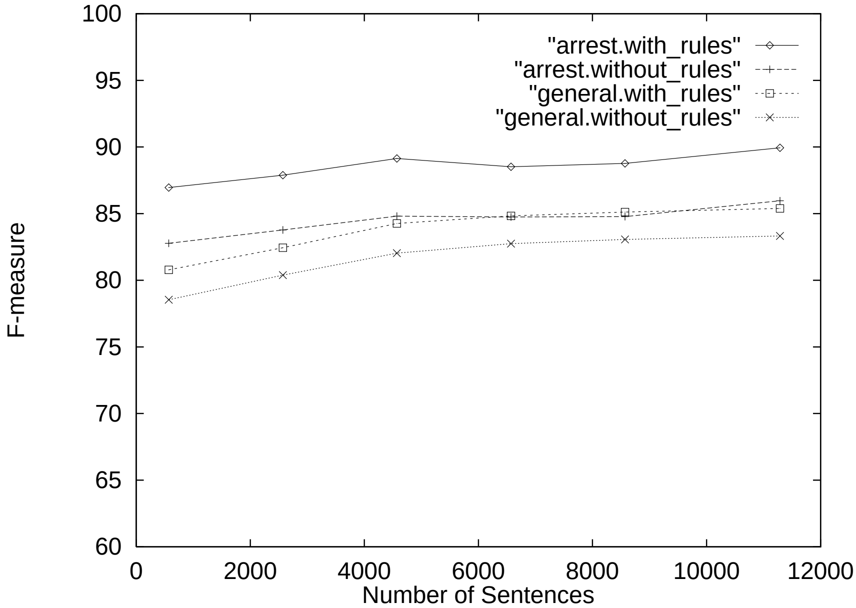


図 1　学習コーパスの量と精度の関係 (学習コーパスに対して)

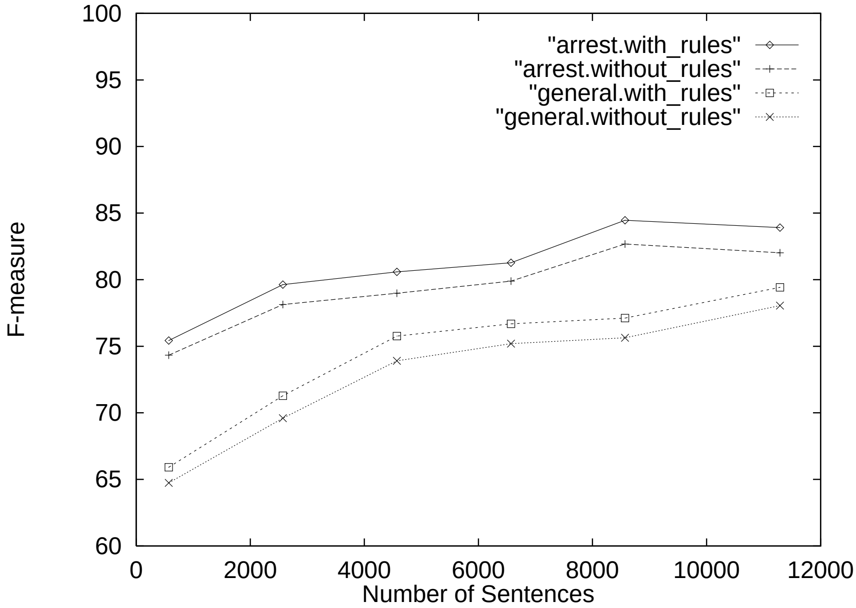


図 2　学習コーパスの量と精度の関係 (テストコーパスに対して)

に分類される．本節では主に他の学習型システムとの違いを説明し，そのうち重要であると思われる部分については追加実験を行なうことによりその違いがどの程度精度に影響を与えるかを調べる．

学習型のシステムのアプローチは我々のものも含めて四つあり，どれも基本的に関根らのとったアプローチ (Sekine, Grishman, and Shinnou 1998) に類似している．関根らのシステム (Sekine et al. 1998) を改良したものには Borthwick のシステム (Borthwick 1999)，野畑のシステム (野畑 1999)，新納 (新納 1999) のシステムがあり，それらと我々のシステムの違いを表にすると表 8 のようになる[5]．違いは主に学習モデル，形態素に付与するラベル (NE ラベル) の定義，素性，後処理にある．

以下では主に表 8 の違いに着目して考察する．

### 形態素に付与するラベル (NE ラベル) の定義について

我々の定義した NE ラベルは関根らのシステムに比べると 7 種類多い．これは，関根らのシステムよりさらに OPTIONAL(始まり，中間，終り，単独の 4 種類) および PRE， POST， MID のラベルを考慮したためである． OPTIONAL はタグ付けが判定者にも困難な場合のために設けられたものであり，その性質を学習することによって，例えば LOCATION か ORGANIZATION の判定が困難なものをいずれかのタグに分類してしまうのを避けることができると考えられる．PRE, POST, MID のラベルは固有表現の前後および固有表現の間の形態素に付与するように設けたものである．これは見方を変えると関根らが OTHER(あるいは NONE) としていたラベルを固有表現以外の部分の始まり (POST に対応)，中間 (OTHER に対応)，終り (PRE に対応)，単独 (MID に対応) に細分類したものであるとも言える．

OPTIONAL に関する 4 種類および PRE， POST， MID のラベルがどの程度精度に影響しているかを調べるために，それぞれのラベルを OTHER にマージして追加実験を行なった．その結果を表 9 ~表 11 にあげる．表の括弧内の数値は表 6 にあげた精度からの増減を表す．これらの実験では ME モデルによるラベル付けの精度の違いを調べることを目的としているためいずれも書き換え規則は用いていない．

表から分かるように OPTIONAL に関するラベルは期待していたほど精度に影響を与えていなかった． PRE， POST， MID のラベルは一般ドメイン (GENERAL) に対しては精度の向上に貢献しているのに対し，限定ドメイン (ARREST) に対してはその利用がかえって精度を低下させることになっている．固有表現ごとに精度の増減を調べてみると， PRE， POST， MID のラベルを OTHER にマージすることによって限定ドメインに対しては PERSON と LOCATION の精度が良くなっており，他の固有表現に対しては悪くなっていることが分かる．限定ドメインの内訳 (表 5) を見ると PERSON と LOCATION の個数が多く，これらの固有表現に対する抽出精

5 表で比較としてあげた精度 (F-measure) はすべて IREX ワークショップ予稿集において報告されたものである．

表 8　関連研究との比較

| | 学習モデル | NE ラベル | 素性 | 後処理 | F-measure |
|---|---|---|---|---|---|
| 我々のシステム | ME モデル | 40 種類．固有表現タグに OPTIONAL，OTHER を加えた 10 種類のそれぞれに対し，固有名詞の始まり，中間，終り，単独を表す 4 種類のラベルを用意する． | 着目している形態素を含む前後 2 形態素ずつ合計 5 形態素に関する見出し語，品詞 (大分類，細分類) | 自動獲得した書き換え規則 | 83.91(ARREST)，79.42(GENERAL) |
| Borthwick のシステム | ME モデル | (関根らのシステムと同じ) | (関根らのシステムと同じ) | 人手で作成した書き換え規則 | 85.02(ARREST)，77.37(GENERAL) |
| 野畑のシステム | 決定木モデル | 38 種類．8 種類の固有表現タグそれぞれに対し，固有名詞の始まり，中間，終り，単独を表す 4 種類のラベルを用意し，固有表現以外に対し OTHER というラベルを用意する．さらに，固有物名 (ARTIFACT), 地名 (LOCATION)，組織名 (ORGANIZATION) を細分類する． | 着目している形態素を含む前後 1 形態素ずつ合計 3 形態素に関する品詞 (大分類，細分類)，字種情報，単語リスト，統語的情報 | なし | 80.37(ARREST)，70.34(GENERAL) |
| 新納のシステム | n-gram モデル | 形態素ごとではなく文字ごとにラベルを付与する．36 種類．固有表現タグに固有表現以外を表す OTHER を加えた 9 種類のそれぞれに対し，固有名詞の始まり，中間，終り，単独を表す 4 種類のラベルを用意する． | 着目している文字を含む前後 1 文字ずつ合計 3 文字に関する文字，品詞情報 | なし | 58.46(ARREST)，60.36(GENERAL) |
| 関根らのシステム | 決定木モデル | 33 種類．8 種類の固有表現タグそれぞれに対し，固有名詞の始まり，中間，終り，単独を表す 4 種類のラベルを用意し，固有表現以外に対し OTHER というラベルを用意する． | 着目している形態素を含む前後 1 形態素ずつ合計 3 形態素に関する見出し語，品詞 (大分類，細分類)，字種情報，辞書情報 | なし | |

度が良くなったため全体の精度も良くなったと考えられる．PERSON と LOCATION に対する精度が良くなったのは PRE や POST などが数の多い PERSON や LOCATION の性質に引っ張られて PERSON や LOCATION の前後に位置しやすいラベルとして学習されたためである可能性が高い．PERSON や LOCATION については PRE や POST などをさらに細分類して PERSON:PRE や PERSON:POST のようなラベルを考えると良いかも知れない．このような細分類は，PERSON や LOCATION，特に PERSON は固有表現直後の形態素が「さん」や「氏」など特別な語であることが多く，他の固有表現についてそのような傾向は見られない (野畑 関根 1999) ことからも妥当な方法であると考えられる．しかし，抽出精度をもとに細分類を続けると本試験の

表 9　OPTIONAL のラベルを OTHER にマージした場合 (書き換え規則は用いない)

| 固有表現 (NE) | ARREST | | GENERAL | |
|---|---|---|---|---|
| | 再現率 (%) | 適合率 (%) | 再現率 (%) | 適合率 (%) |
| ORGANIZATION | 58.11 | 81.13 | 59.28 | 82.31 |
| PERSON | 82.47 | 84.21 | 76.33 | 83.50 |
| LOCATION | 73.58 | 78.00 | 69.98 | 82.81 |
| ARTIFACT | 61.54 | 66.67 | 33.33 | 50.00 |
| DATE | 97.22 | 97.22 | 90.00 | 95.12 |
| TIME | 94.74 | 100.00 | 87.04 | 94.00 |
| MONEY | 100.00 | 80.00 | 93.33 | 82.35 |
| PERCENT | - | - | 80.95 | 94.44 |
| 総合 | 78.41 (-0.77) | 84.72 (-0.36) | 72.12 (-0.07) | 85.01 (+0.05) |
| F-measure | 81.44 (-0.58) | | 78.04 (-0.01) | |

表 10　PRE, POST, MID のラベルを OTHER にマージした場合 (書き換え規則は用いない)

| 固有表現 (NE) | ARREST | | GENERAL | |
|---|---|---|---|---|
| | 再現率 (%) | 適合率 (%) | 再現率 (%) | 適合率 (%) |
| ORGANIZATION | 58.11 | 79.63 | 57.06 | 78.03 |
| PERSON | 86.60 | 88.42 | 74.26 | 83.39 |
| LOCATION | 74.53 | 81.44 | 68.04 | 84.13 |
| ARTIFACT | 53.85 | 63.64 | 31.25 | 48.39 |
| DATE | 97.22 | 95.89 | 88.46 | 93.50 |
| TIME | 94.74 | 100.00 | 94.44 | 98.08 |
| MONEY | 100.00 | 88.89 | 93.33 | 73.68 |
| PERCENT | - | - | 80.95 | 94.44 |
| 総合 | 79.43 (+0.25) | 86.55 (+1.47) | 70.53 (-1.66) | 84.19 (-0.77) |
| F-measure | 82.84 (+0.82) | | 76.76 (-1.29) | |

表 11　PRE, POST, MID, OPTIONAL のラベルを OTHER にマージした場合 (書き換え規則は用いない)

| 固有表現 (NE) | ARREST | | GENERAL | |
|---|---|---|---|---|
| | 再現率 (%) | 適合率 (%) | 再現率 (%) | 適合率 (%) |
| ORGANIZATION | 58.11 | 81.13 | 56.51 | 78.16 |
| PERSON | 86.60 | 88.42 | 74.56 | 83.44 |
| LOCATION | 75.47 | 80.81 | 68.04 | 84.13 |
| ARTIFACT | 53.85 | 63.64 | 31.25 | 50.00 |
| DATE | 97.22 | 95.89 | 88.08 | 93.47 |
| TIME | 94.74 | 100.00 | 94.44 | 98.08 |
| MONEY | 100.00 | 88.89 | 93.33 | 73.68 |
| PERCENT | - | - | 80.95 | 94.44 |
| 総合 | 79.69 (+0.51) | 86.59 (+1.51) | 70.40 (-1.79) | 84.30 (-0.66) |
| F-measure | 83.00 (+0.98) | | 76.72 (-1.33) | |

データに対しては精度が良くなるかもしれないが，他のデータに対しても良くなるとは言えなくなる．したがって，これ以上細分類して本試験のデータに対する精度を調べることはあまり意味がないと思われる．

Borthwick のシステムとの精度の差 (表 8) は主にこのラベルの定義の違いと素性の違いから生じていると考えられる．素性の違いについては後で述べる．一方，野畑は ARTIFACT，LOCATION，ORGANIZATION を細分類して F-measure で 2 ポイント程度精度が向上したと報告している．我々のシステムにおいてもどの程度精度に影響するかを調べたいところであるが，学習コーパスに付与されたラベルを人手で細分類する必要があるため同じ学習コーパスを作成するのは困難であると判断し，野畑の細分類に基づいて追加実験をするのは見合わせた．新納は形態素ごとではなく文字ごとに NE ラベルを付与する方法を提案した．この方法は形態素の区切りと固有表現の区切りが一致しない場合でも一つのモデルで固有表現を抽出できるという点で優れているが，精度は我々に比べて F-measure で 20 ポイント以上低い．この理由は，後に述べる素性に関連することであるが，新納の方法が文字 3-gram という少ない情報のみを用いて NE ラベルを推定しているためであると考えられる．3 文字ということは多くても 3 形態素の情報しか用いていないということである．我々の実験では，3.4 節でも述べたように着目する形態素およびその前後 2 形態素ずつの情報を用いた場合が最も精度が良いことから，新納の方法で我々のシステムと同程度の精度を得るためには少なくとも文字 5-gram 以上の情報を用いる必要があるだろう．しかし，文字 5-gram を得るには膨大な学習コーパスが必要であり，我々が実験に用いた学習コーパスだけでは我々のシステムと同程度の精度は得られないと予想される．

素性について

Borthwick や野畑は我々が用いた素性に加えて字種情報，統語的情報や辞書情報などを用いて精度を向上させている．このうち字種情報については我々の素性においても形態素の品詞情報としてある程度考慮されている．統語的情報については野畑のシステムで用いられている方法では人手の介入が必要であり，同じ条件での実験は困難である．辞書情報については我々の素性に加えて追加実験を行ない，精度に与える影響を調べた．

辞書情報としては Borthwick や野畑と同様に文献 (関根 1999) で公開されているものを用いた．これは組織名，地名に関する辞書で登録数は約 1,000 である．それに加えて，学習コーパスに 3 回以上出現した固有表現約 1,400 個 (ORGANIZATION: 272 個, PERSON: 336 個, LOCATION: 339 個, ARTIFACT: 45 個, DATE: 233 個, TIME: 31 個, MONEY: 21 個, PERCENT: 45 個, OPTIONAL: 56 個) を取り出しそれぞれの固有表現ごとに 9 種類の辞書を作成した．予めこれらの辞書に登録されている固有表現を JUMAN を用いて形態素解析し，各形態素に我々の定義した NE ラベルを付与しておく．素性としては，形態素の見出し語がこれらの辞書中でどのような NE ラベルが付与されているかという情報を用いた．つまり，我々が定義した 40 個のラベルの

各々について付与されているかいないかのそれぞれを素性として用いる．辞書中の形態素の見出し語の異なり数は合計約 10,000 個である．この素性を，着目している形態素のみについて利用した場合，着目している形態素を含む前後 1 形態素ずつ合計 3 形態素について利用した場合，着目している形態素を含む前後 2 形態素ずつ合計 5 形態素について利用した場合それぞれについて追加実験を行なった．それぞれの実験結果を表 12 ~表 14にあげる．結果は着目している形態素のみについて利用した場合が最も精度が良く，考慮する前後の形態素の数が増えるにつれて精度は悪くなった．これは辞書の登録数が問題になっている可能性が高い．辞書に登録されている固有名詞は高々 2,400 個程度であり，そのうち 1 形態素， 2 形態素， 3 形態素， 4 形態素以上からなるものはそれぞれ 745 個， 448 個， 125 個， 60 個である．例えば二つ前の形態素の見出し語が辞書にあるかないかという情報 (我々が辞書情報の素性として利用したもの) が有効なのは辞書に登録されている固有名詞の約 8%， 185(=125+60) 個についてのみであるということになる．今回，学習コーパス以外から得た辞書情報としては一般に公開されている 1,000 語程度の辞書を用いたが，一般に利用可能な大規模な固有名詞辞書があれば，辞書の登録数と精度の関係も調べてみたい．

次に，素性として一つ前の形態素に付与したラベルの情報を考慮したときの精度を調べた．一般に学習による形態素解析では一つあるいは二つ前の形態素に付与したラベルの情報を用いて次のラベルを決定することが多い．我々の手法においてこれと同様の情報がどの程度精度に影響を与えるかを調べることが目的である．

実験結果を表 15 にあげる．表から分かるようにどちらのドメインについても精度を下げる結果となった．特に再現率の低下が著しい．これは学習コーパスでは OTHER の隣は OTHER であることが多く，この連接関係が他のラベルとの連接に比べて学習されやすいため OTHER が連続する場合が最適解となることが多くなるためであると考えられる．

後処理について

Borthwick は我々と同様に形態素解析により得られる形態素が固有表現より長い場合に対処するために書き換え規則を用いて後処理をしている．この後処理により，どちらのシステムも F-measure で 2 ポイント程度精度が向上している．違いは Borthwick が日本語を母語とする人が人手で作成した規則を用いているのに対し，我々は学習コーパスから誤り駆動で自動獲得した規則を用いている点にある．異なるドメインのテキストが与えられたとき，できるだけコストを少なく学習し直すためには規則を自動獲得できる方が望ましい．

誤り駆動型学習を用いて固有表現を抽出するシステムには颯々野らのシステム (颯々野 塚本 1999) がある．このシステムは後処理として書き換え規則を適用することにより，形態素単位より短い文字列を含む固有表現だけでなく一つあるいは複数の形態素からなる固有表現も同様の手法を用いて抽出できる．彼らがベースラインとして用いているシステムの精度が F-measure で 40 程度であるため，我々のシステムをベースラインとして用いることによってより良い精度が得られる可

表 12　辞書情報を素性として考慮した場合 ((0) のみ考慮)

| 固有表現 (NE) | ARREST | | GENERAL | |
|---|---|---|---|---|
| | 再現率 (%) | 適合率 (%) | 再現率 (%) | 適合率 (%) |
| ORGANIZATION | 68.92 | 85.00 | 62.33 | 79.79 |
| PERSON | 83.51 | 85.26 | 77.22 | 83.92 |
| LOCATION | 83.96 | 84.76 | 76.76 | 86.38 |
| ARTIFACT | 61.54 | 80.00 | 35.42 | 48.57 |
| DATE | 97.22 | 97.22 | 90.77 | 94.78 |
| TIME | 94.74 | 100.00 | 90.74 | 94.23 |
| MONEY | 100.00 | 88.89 | 93.33 | 82.35 |
| PERCENT | - | - | 100.00 | 100.00 |
| 総合 | 83.55 (+1.80) | 88.08 (+1.90) | 75.89 (+1.39) | 85.20 (+0.17) |
| F-measure | 85.75 (+1.84) | | 80.17 (+0.75) | |

表 13　辞書情報を素性として考慮した場合((-1)(0)(1) を考慮)

| 固有表現 (NE) | ARREST | | GENERAL | |
|---|---|---|---|---|
| | 再現率 (%) | 適合率 (%) | 再現率 (%) | 適合率 (%) |
| ORGANIZATION | 71.62 | 81.54 | 63.16 | 79.17 |
| PERSON | 78.35 | 86.36 | 74.85 | 84.05 |
| LOCATION | 83.02 | 83.81 | 77.24 | 89.36 |
| ARTIFACT | 46.15 | 75.00 | 31.25 | 53.57 |
| DATE | 95.83 | 95.83 | 90.00 | 92.86 |
| TIME | 94.74 | 100.00 | 90.74 | 92.45 |
| MONEY | 100.00 | 100.00 | 93.33 | 63.64 |
| PERCENT | - | - | 100.00 | 100.00 |
| 総合 | 81.75 (±0) | 87.36 (+1.18) | 75.03 (+0.53) | 85.70 (+0.67) |
| F-measure | 84.46 (+0.55) | | 80.01 (+0.59) | |

表 14　辞書情報を素性として考慮した場合((-2)(-1)(0)(1)(2) を考慮)

| 固有表現 (NE) | ARREST | | GENERAL | |
|---|---|---|---|---|
| | 再現率 (%) | 適合率 (%) | 再現率 (%) | 適合率 (%) |
| ORGANIZATION | 68.92 | 79.69 | 62.05 | 78.32 |
| PERSON | 78.35 | 86.36 | 74.85 | 82.95 |
| LOCATION | 81.13 | 83.50 | 75.79 | 88.42 |
| ARTIFACT | 46.15 | 75.00 | 33.33 | 53.33 |
| DATE | 95.83 | 95.83 | 91.15 | 93.68 |
| TIME | 94.74 | 100.00 | 90.74 | 94.23 |
| MONEY | 100.00 | 88.89 | 93.33 | 73.68 |
| PERCENT | - | - | 100.00 | 100.00 |
| 総合 | 80.72 (-1.03) | 86.74 (+0.56) | 74.64 (+0.14) | 85.38 (+0.35) |
| F-measure | 83.62 (-0.29) | | 79.65 (+0.23) | |

表 15　一つ前の形態素に付与したラベルを素性として考慮した場合

| 固有表現 (NE) | ARREST | | GENERAL | |
|---|---|---|---|---|
| | 再現率 (%) | 適合率 (%) | 再現率 (%) | 適合率 (%) |
| ORGANIZATION | 60.81 | 80.36 | 55.12 | 79.28 |
| PERSON | 78.35 | 86.36 | 73.08 | 83.16 |
| LOCATION | 83.96 | 88.12 | 74.33 | 85.28 |
| ARTIFACT | 53.85 | 58.33 | 29.17 | 33.33 |
| DATE | 95.83 | 95.83 | 85.38 | 94.47 |
| TIME | 94.74 | 100.00 | 85.19 | 93.88 |
| MONEY | 100.00 | 100.00 | 93.33 | 93.33 |
| PERCENT | - | - | 95.24 | 95.24 |
| 総合 | 80.21 (-1.54) | 87.89 (+1.71) | 70.79 (-3.71) | 84.17 (-0.86) |
| F-measure | 83.87 (-0.06) | | 76.91 (-2.51) | |

能性が高いと考えられる．

**IREX-NE 本試験に参加したシステムの結果との比較**

IREX-NE 本試験での結果を表 16 にあげる．最も良い精度を出したシステムは人手により作成された規則に基づいている．我々のシステムは本試験ではシステム番号 1223 であり，精度は F-measure で限定ドメインに対して 74.90，一般ドメインに対して 72.18 であった．このように精度が悪かったのは「MIDDLE」に関するラベルの連接規則に洩れがあったためである．「MIDDLE」に関するラベル同士が連接可能であるという規則が欠如していたため，“比例 (ARTIFACT:BEGIN) ／代表 (ARTIFACT:MIDDLE) ／並立 (ARTIFACT:MIDDLE) ／制 (ARTIFACT:END)” のように 4 形態素以上からなる固有表現は抽出できなくなっていた．本論文ではその洩れを埋めたときの精度を示した．その精度は最高で F-measure で 85.75(限定ドメイン)，80.17(一般ドメイン) であった．これは IREX-NE 本試験で我々が用いた素性に加えて，3.6 節で述べた人名や組織名などの固有名詞辞書も利用したときの精度であり，悪くない精度であると考えている．我々の手法は学習コーパスがあれば人手のコストもかからないため，さまざまなドメインに対しても低コストでそれなりの精度が得られるものであると言える．

## 4　まとめ

本論文では ME モデルと書き換え規則を用いて固有表現を抽出する手法について述べた．IREX-NE の定義に基づくと固有表現には一つあるいは複数の形態素からなるものと形態素単位より短い部分文字列を含むものの 2 種類がある．前者の固有表現は，固有表現の始まり，中間，終りなどを表すラベルを 40 個用意し，それらのラベルを推定することによって抽出する．ラベルの推

表 16　IREX-NE 本試験結果

| システム番号 | F-measure ARR | F-measure GEN | システム番号 | F-measure ARR | F-measure GEN |
|---|---|---|---|---|---|
| 1201 | 54.17 | 57.69 | 1229 | 64.81 | 57.63 |
| 1205 | 78.08 | 80.05 | 1231 | 81.94 | 74.82 |
| 1213 | 59.87 | 66.60 | 1234 | 72.77 | 71.96 |
| 1214 | 80.37 | 70.34 | 1240 | 58.46 | 60.96 |
| 1215 | 74.56 | 66.74 | 1247 | 87.43 | 83.86 |
| 1223 | 74.90 | 72.18 | 1250a | 70.12 | 69.82 |
| 1224 | 77.61 | 75.30 | 1250b | 55.24 | 57.76 |
| 1227 | 85.02 | 77.37 | | | |

定にはコーパスから学習した ME モデルを用いる．後者の固有表現は書き換え規則を用いて抽出する．書き換え規則は学習コーパスに対するシステムの解析結果とコーパスの正解データとの差異を調べることによって自動獲得することができる．

書き換え規則，素性，学習コーパスの量についての条件を変えた比較実験により，形態素解析により得られる形態素が固有表現より長い場合に書き換え規則が有効であること，我々が考慮した素性，つまり，着目している形態素を含む前後 2 形態素ずつ合計 5 形態素に関する見出し語，品詞の情報が IREX-NE 本試験に用いられたテキストに対して有効であることを示すことができた．これらは学習コーパスのみから得られる情報であったが，それに加えて一般に公開されている人名や組織名などの固有名詞辞書も利用することにより，IREX-NE 本試験のデータに対する実験で，F-measure で 85.75(限定ドメイン)，80.17(一般ドメイン) の精度を得ることができた．

本手法の精度をさらに向上させるために必要であると考えているのは，以下の三点である．

- 素性の発見
  今回は利用しなかったような情報，例えば，係り受けの情報や照応関係などを素性として新たに考慮することによって，ラベル推定の精度が向上することが期待される．また，解析と同時に解析の過程で得られた情報を利用することも考えられる．これは今後の重要な課題である．
- コーパス，辞書の充実
  今回の実験では書き換え規則を利用したことによる誤り例を考察することで，学習コーパスの誤りが精度に影響することが分かった．また，辞書情報を考慮することで精度が向上することも分かった．コーパス修正や辞書作成にかかるコストを考えると，コーパスの誤りを自動あるいは半自動で修正する方法，辞書情報を自動あるいは半自動で獲得する方法を考案する必要がありそうである．

- 特定ドメインへのチューニング
  ドメインを限定すると，そのドメインに固有の固有表現パターンに，より特化した学習が可能であると考えられる．今回は限定ドメインに対して特にチューニングするようなことはしなかったが，限定ドメインに対しどこまでチューニングすることが可能かを調べたい．

謝辞

本研究を進めるにあたって有意義なコメントを下さったニューヨーク大学の関根聡助教授に心から感謝の意を表する．また，データの利用を許可して下さった毎日新聞社に感謝する．

## 参考文献

Aberdeen, J., Burger, J., Day, D., Hirschman, L., Robinson, P., and Vilain, M. (1995). "MITRE: Description of the ALEMBIC System used for MUC-6." In *Proceedings of the Sixth Message Understanding Conference (MUC-6)*, pp. 141–155.

Bikel, D. M., Miller, S., Schwartz, R., and Weischedel, R. (1997). "Nymble: a High-Performance Learning Name-finder." In *Proceedings of the Fifth Conference on Applied Natural Language Processing*, pp. 194–201.

Borthwick, A. (1999). "A Japanese Named Entity Recognizer Constructed by a Non-Speaker of Japanese." IREX ワークショップ予稿集, pp. 187–193.

Borthwick, A., Sterling, J., Agichtein, E., and Grishman, R. (1998a). "Exploiting Diverse Knowledge Sources via Maximum Entropy in Named Entity Recognition." In *Proceedings of the Sixth Workshop on Very Large Corpora*, pp. 152–160.

Borthwick, A., Sterling, J., Agichtein, E., and Grishman, R. (1998b). "NYU: Description of the MENE Named Entity System as Used in MUC-7." In *Proceedings of the Seventh Message Understanding Conference (MUC-7)*. http://www.muc.saic.com/.

Brill, E. (1995). "Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part-of-Speech Tagging." *Computational Linguistics*, **21** (4), 543–565.

Cowie, J. (1995). "CRL/NMSU Description of the CRL/NMSU System Used for MUC-6." In *Proceedings of the Sixth Message Understanding Conference (MUC-6)*, pp. 157–166.

久光徹 丹羽芳樹 (1998). "書き換え規則と文脈情報を用いた形態素解析後処理." 情報処理学会 自然言語処理研究会, **NL126** (8), 55–62.

IREX 実行委員会 (1999). "IREX homepage." http://cs.nyu.edu/cs/projects/proteus/irex/.

Jaynes, E. T. (1957). "Information Theory and Statistical Mechanics." *Physcal Review*, **106**, 620–630.

Jaynes, E. T. (1979). "Where do we Stand on Maximum Entropy?." In Levine, R. D. and Tribus, M. (Eds.), *The Maximum Entropy Formalism*, p. 15. M. I. T. Press.

黒橋禎夫 長尾眞 (1998). 日本語形態素解析システム JUMAN 使用説明書 version 3.6. 京都大学大学院情報学研究科.

Lin, D. (1998). "Using Collocation Statistics in Information Extraction." In *Proceedings of the Seventh Message Understanding Conference (MUC-7)*. http://www.muc.saic.com/.

Miller, S., Crystal, M., Fox, H., Ramshaw, L., Schwartz, R., Stone, R., Weischedel, R., and the Annotation Group (1998). "Algorithms that Learn to Extract Information BBN: Description of the Sift System as Used for MUC-7." In *Proceedings of the Seventh Message Understanding Conference (MUC-7)*. http://www.muc.saic.com/.

野畑周 (1999). "決定木を用いた学習に基づく固有表現抽出システム." IREX ワークショップ予稿集, pp. 201–206.

野畑周 関根聡 (1999). "固有表現抽出の難しさ." IREX ワークショップ予稿集, pp. 115–124.

Pietra, S. D., Pietra, V. D., and Lafferty, J. (1995). "Inducing Features of Random Fields." Tech. rep., Carnegie Mellon University. CMU-CS-95-144.

Ratnaparkhi, A. (1996). "A Maximum Entropy Model for Part-Of-Speech Tagging." In *Conference on Empirical Methods in Natural Language Processing*, pp. 133–142.

Ratnaparkhi, A. (1997). "A Linear Observed Time Statistical Parser Based on Maximum Entropy Models." In *Conference on Empirical Methods in Natural Language Processing*.

Ristad, E. S. (1997). "Maximum Entropy Modeling for Natural Language." ACL/EACL Tutorial Program, Madrid.

Ristad, E. S. (1998). "Maximum Entropy Modeling Toolkit, Release 1.6 beta." http://www.mnemonic.com/software/memt.

SAIC (1998). "MUC homepage." http://www.muc.saic.com/.

颯々野学 塚本浩司 (1999). "有限状態変換器の誤り駆動型学習を用いた固有表現抽出." 情報処理学会 自然言語処理研究会, NL132 巻, pp. 1–8.

関根聡 (1999). "Satoshi Sekine homepage." http://www.cs.nyu.edu/cs/projects/proteus/sekine/.

Sekine, S., Grishman, R., and Shinnou, H. (1998). "A Decision Tree Method for Finding and Classifying Names in Japanese Texts." In *Proceedings of the Sixth Workshop on Very Large Corpora*, pp. 171–178.

新納浩幸 (1999). "拡張文字ベースの HMM を利用した固有名詞抽出." IREX ワークショップ予稿集, pp. 151–157.

内元清貴 馬青 (1999). "形態素構文解析." 人文学と情報処理 (21), 13–29.

Uchimoto, K., Sekine, S., and Isahara, H. (1999). "Japanese Dependency Structure Analysis Based on Maximum Entropy Models." In *Proceedings of the Ninth Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pp. 196–203.

## 略歴

**内元 清貴:** 1994 年京都大学工学部卒業．1996 年同大学院修士課程修了．同年郵政省通信総合研究所入所，郵政技官．自然言語処理の研究に従事．言語処理学会，情報処理学会，ACL，各会員．

**馬 青:** 1983 年北京航空航天大学自動制御学部卒業．1987 年筑波大学大学院理工学研究科修士課程修了．1990 年同大学院工学研究科博士課程修了．工学博士．1990 ～ 93 年株式会社小野測器勤務．1993 年郵政省通信総合研究所入所，主任研究官．人工神経回路網モデル，知識表現，自然言語処理の研究に従事．日本神経回路学会，言語処理学会，電子情報通信学会，各会員．

**村田 真樹:** 1993 年京都大学工学部卒業．1995 年同大学院修士課程修了．1997 年同大学院博士課程修了，博士（工学）．同年，京都大学にて日本学術振興会リサーチ・アソシエイト．1998 年郵政省通信総合研究所入所．研究官．自然言語処理，機械翻訳，情報検索の研究に従事．言語処理学会，情報処理学会，人工知能学会，ACL，各会員．

**小作 浩美:** 1985 年郵政省電波研究所（現通信総合研究所）入所．研究官．自然言語処理の研究に従事．言語処理学会，情報処理学会，電子情報通信学会，各会員．

**内山 将夫:** 筑波大学第三学群情報学類卒業 (1992)．筑波大学大学院工学研究科博士課程修了 (1997)．信州大学工学部電気電子工学科助手 (1997)．郵政省通信総合研究所非常勤職員 (1999)．博士 (工学)．

**井佐原 均:** 1978 年京都大学工学部電気工学第二学科卒業．1980 年同大学院修士課程修了．博士（工学）．同年通商産業省電子技術総合研究所入所．1995 年郵政省通信総合研究所関西支所知的機能研究室室長．自然言語処理，機械翻訳の研究に従事．言語処理学会，情報処理学会，人工知能学会，日本認知科学会，ACL，各会員．

(1999 年 10 月 21 日 受付)
(1999 年 12 月 13 日 再受付)
(2000 年 1 月 7 日 採録)