自 然 言 語 処 理 139－1
（2000．9．21）

統計的日本語固有表現抽出における固有表現まとめ上げ手法とその評価

颯々野 学
富士通研究所
sassano@flab.fujitsu.co.jp

宇津呂 武仁
豊橋技術科学大学 情報工学系
utsuro@ics.tut.ac.jp

本論文では，統計的手法に基づく日本語固有表現のまとめ上げの問題に対して，性能を大きく左右する四つの要因，i) 固有表現まとめ上げ状態の表現法，ii) 現在位置の周囲の形態素を何個まで考慮するか，iii) 個々の形態素の素性，iv) 統計的学習法，について，これまで日本語固有表現のまとめ上げにおいてはその有効性が確認されていない幾つかの方式を実験的に評価し，その得失について報告する．特に，ii) について，現在位置の形態素が，いくつの形態素から構成される固有表現の一部であるかを考慮して学習を行なう可変長モデルを新たに提案する．実験の結果，先行研究で用いられた固定長モデルの性能を大きく上回る結果が得られ，可変長モデルの有効性が確認できた．

Named Entity Chunking Techniques and their Evaluation in Japanese Statistical Named Entity Recognition

Manabu Sassano
Fujitsu Laboratories, Ltd.
sassano@flab.fujitsu.co.jp

Takehito Utsuro
Department of Information and Computer Sciences,
Toyohashi University of Technology
utsuro@ics.tut.ac.jp

This paper focuses on the issue of named entity chunking in Japanese named entity recognition. We discuss the issues of i) encoding schemes of named entity chunking states, ii) models of preceding/subsequent morphemes as contextual clues, iii) features of each morpheme, and iv) statistical learning techniques. Especially, as for ii), we propose a novel model, namely the *variable length model*, which incorporates richer contextual information as well as patterns of constituent morphemes within a named entity. We also show that that the proposed model outperforms previous approaches.

1 はじめに

固有表現抽出は，情報検索・抽出，機械翻訳，自然言語理解など自然言語処理の応用的局面における基礎技術として重要な技術の一つである．英語においては，特に米国において，MUC(Message Understanding Conference，例えば，MUC-7 [MUC98]) コンテストにおける課題の一つとして固有表現抽出がとりあげられ，集中的に研究が行なわれてきた．また，最近では，日本語においても，MET (Multilingual Entity Task, 例えば，MET-1 [Maiorano96]，MET-2 [MUC98]) や IREX(Information Retrieval and Extraction Exercise) ワークショップ [IREX 実行委員会 99] などのコンテストにおいて，固有表現抽出が課題の一つに取り上げられている．

日本語解析においては，近年，高性能な形態素解析システムが利用可能となってきており，日本語固有表現抽出においても，形態素解析の結果得られる情報が非常に有用であるため，日本語固有表現抽出の研究においては，前処理として形態素解析を行なうことが標準的となっている．そして，形態素解析結果の形態素列に対して，人手で構築されたパターンマッチング規則や統計的学習によって得られた固有表現抽出規則を適用することにより，固有表現が抽出される [IREX 実行委員会 99]．人手で規則を記述するアプローチと，人手で構築された正解データから統計的手法により規則を学習するアプローチを比べると，後者のアプローチでは，共通の訓練用コーパスが公開されている場合に人手のコストをかけずに規則を自動学習できるという利点がある．また，対象とする分野や固有表現の定義が変わった場合でも，一度訓練用のコーパスを整備してしまえば，その訓練コーパスの傾向に沿った規則が自動学習できるという利点もある．そこで，本研究では，統計的学習により固有表現抽出規則を自動学習するアプローチを採用する．

ここで，日本語の固有表現抽出を難しくしている問題の一つに，形態素の区切り単位と固有表現の区切り単位の違いの問題がある．例えば，我々が，IREX ワークショップの訓練データを日本語形態素解析システム BREAKFAST [颯々野 97] によって形態素解析し，その結果の形態素の区切り単位と固有表現の区切り単位を比較・分析したところ，固有表現の約半数近くは，形態素とは異なる区切り単位によって定義されており，一つの形態素とは一対一には対応しなかった．また，それらの，一つの形態素に一対一に対応しない固有表現について，固有表現の区切り単位と形態素の区切り単位の対応パターンを分析したところ，その 90%近くについては，一つの固有表現の開始および終了位置が，いずれかの形態素の開始位置または終了位置と一致し，一つの固有表現が複数の形態素から構成されていることがわかった (2 節参照)．

これらの調査結果から，日本語の固有表現抽出の問題の大部分は，高性能な形態素解析の処理を前処理として，その結果の形態素列に対して，一つもしくは複数の形態素をまとめ上げる処理を行ない，同時にまとめ上げられた形態素列がどの種類の固有表現を構成しているかを同定するというプロセスにより解決可能であることがわかる．ここで，一つもしくは複数の形態素のまとめ上げの処理においては，英語の単名詞句まとめ上げ (base NP chunking) を統計的学習により行なう手法 (例えば，[Ramshaw95] など) が利用できる．それらの単名詞句まとめ上げ手法の基本的な考え方は，現在注目している位置にある単語およびその周囲の単語を考慮しながら，現在位置の単語が単名詞句の一部となりうるかどうかを判定するというものである．実際に，IREX ワークショップの固有表現抽出タスクにおいて，統計的学習を用いて好成績をおさめたシステム [Borthwick99, 内元 00] は，基本的にはこのような考え方に沿って，一つもしくは複数の形態素をまとめ上げると同時に固有表現の種類を同定するという方式で設計されたものである．

ここで，統計的学習に基づいて日本語固有表現のまとめ上げを行なう処理の性能を決定する要因について考えてみると，大きく分けて以下の四つの要因が考えられる．i) 固有表現まとめ上げ状態の表現法，ii) 現在位置の周囲の形態素を何個まで考慮するか，iii) 個々の形態素の素性，iv) 統計的学習法．このうち，IREX ワークショップで好成績をおさめたシステム [Borthwick99, 内元 00] において採用された方式を挙げると，i) については，英語単名詞句まとめ上げ [Ramshaw95] で用いられた，現在位置の単語が単名詞句の一部となるか否かを状態とする (Inside/Outside 法，3.2.1 節参照) ような粗い表現法ではなく，まとめ上げ状態をより細分化した表現法 (Start/End 法，3.2.2 節参照) が用いられた．また，ii) については，現在位置の形態素の前後一形態素ずつを考慮するもの [Borthwick99] と，前後二形態素ずつを考慮するもの [内元 00] とがある．iii) については，個々の形態素の語彙・品詞・文字種などが用いられた．iv) については，いずれも最大エントロピー法を用いている．

先行研究 [Borthwick99, 内元 00] におけるこれらの成果をふまえて，本論文では，上記の要因のうち，i)，ii)，iv) について，これまで日本語固有表現のまとめ上げにおいてはその有効性が確認されていない幾つかの方式を実験的に評価し，その得失について報告する．まず，i) としては，英語単名詞句まとめ上げ [Ramshaw95] で用いられた，現在位置の単語が単名詞句の一部となるか否かを状態とする粗い表現法 (Inside/Outside 法) を評価する．次に，ii) としては，先行研究 [Borthwick99, 内元 00] でやられたように，現在位置の形態素がどれだけの長さの固有表現を構成するのかを全く考慮せずに，常に現在位置

表 1: 日本語固有表現の種類およびその頻度

種類	頻度 (%)	
	訓練データ	評価データ
ORGANIZATION	3676 (19.7)	361 (23.9)
PERSON	3840 (20.6)	338 (22.4)
LOCATION	5463 (29.2)	413 (27.4)
ARTIFACT	747 (4.0)	48 (3.2)
DATE	3567 (19.1)	260 (17.2)
TIME	502 (2.7)	54 (3.5)
MONEY	390 (2.1)	15 (1.0)
PERCENT	492 (2.6)	21 (1.4)
合計	18677	1510

の形態素の前後二形態素ずつまでを考慮して学習を行なう (固定長モデル, 3.3.1 節参照) のではなく, 現在位置の形態素が, いくつの形態素から構成される固有表現の一部であるかを考慮して学習を行なう方法 (可変長モデル, 3.3.2 節参照) を新たに提案し, その効果を検証する. 最後に, iv) としては, 先行研究 [Borthwick99, 内元 00] において最大エントロピー法を用いて試されたように, 各素性を独立とみなして学習を行ない適用時には重み付きであらゆる素性を考慮するのではなく, 複数の素性の組合わせの結合事象を素性選択し, 適用時には, 選択されなかった素性の情報を考慮しないという素性選択型の学習法の一つである決定リスト学習の性能を評価する[1]. なお, 実験においては, これらの新たに評価する設定, および, 先行研究 [Borthwick99, 内元 00] で用いられた設定のほぼ全ての組合わせを評価する.

これらの実験の結果は, 5 節で詳細に述べるが, 最も注目すべき結果として, ii) の「現在位置の周囲の形態素を何個まで考慮するか」という問題に対して, 我々が新たに提案した可変長モデル (現在位置の形態素が, いくつの形態素から構成される固有表現の一部であるかを考慮して学習を行なう方法) が, 先行研究 [Borthwick99, 内元 00] の固定長モデル (現在位置の形態素がどれだけの長さの固有表現を構成するのかを全く考慮せずに, 常に現在位置の形態素の前後二形態素ずつまでを考慮して学習を行なう方法) の性能を大きく上回る結果が得られた. これは, IREX ワークショップの公式スコアとしても, 統計的学習を用いるシステムのなかでは最高の成績に相当し, 日本語固有表現まとめ上げの問題に対して, 可変長モデルの有効性を実験的に確認することができた.

[1] 本研究の次の段階の研究として, ブートストラップを用いることにより, 人手解析済コーパスを用いず固有表現抽出規則を学習する方式の研究を行なっている [Utsuro00a, 宇津呂 00b] が, ブートストラップ方式においては, 従来, 決定リスト学習もしくはそれと類似の手法が用いられてきた [Yarowsky95, Collins99, Cucerzan99] ため, 本研究においても, 素性選択型の学習法としては決定リスト学習を採用した. また, 同様の素性選択型の学習法として決定木学習があるが, 決定リスト学習においては, あらゆる素性の組合わせを選択の候補として素性探索を行なうことが容易であり, 素性の組合わせの探索空間は決定木学習よりも広いと考えられるため, 決定リスト学習の方を用いている.

表 2: 形態素と固有表現の対応パターン

対応パターン		固有表現タグ頻度 (%)	
1 対 1		10480 (56.1)	
$n(\geq 2)$ 形態素 対 1 固有表現	$n = 2$	4557 (24.4)	7175 (38.4)
	$n = 3$	1658 (8.9)	
	$n \geq 4$	960 (5.1)	
その他		1022 (5.5)	
合計		18677	

2 日本語固有表現抽出

2.1 IREX ワークショップの固有表現抽出タスク

IREX ワークショップの固有表現抽出タスクでは, 表 1 に示す八種類の固有表現の抽出が課題とされた [IREX 実行委員会 99]. 表 1 には, 主催者側から提供された訓練データの主要部分を占める CRL(郵政省通信総合研究所) 固有表現データ (毎日新聞 1,174 記事の固有表現をタグ付け), および本試験データのうちの一般ドメインのもの (毎日新聞 71 記事の固有表現をタグ付け) について, 八種類の固有表現数を調査した結果を示す.

2.2 形態素と固有表現の対応パターン

次に, 上記の IREX ワークショップの固有表現抽出タスクの訓練データを形態素解析システム BREAKFAST[颯々野 97][2] で形態素解析し, その結果の形態素と固有表現の対応パターンを調査した結果を表 2 に示す. これからわかるように, 半分近くの固有表現については, 形態素と固有表現が一対一に対応しないことがわかる. また, そのうちの 90%近くについては, 一つの固有表現の開始および終了位置が, いずれかの形態素の開始位置または終了位置と一致し, 一つの固有表現が複数の形態素から構成されていることがわかる. 図 1 にこのような場合の例を示す. また, 表 2 の「その他」の場合の多くは, 一つ以上の固有表現が一つの形態素の一部となる場合である. 例えば, 「訪米」という形態素に対して, その一部である「米」のみが LOCATION(地名) であるという例がこれに相当する. この「その他」の場合の固有表現については, その割合が少なく, また, 先行研究 [内元 00] において, ある程度の割合で抽出できることがわかっているので, 本論文における考慮の対象には含めない.

3 固有表現まとめ上げおよび分類

本節では, 日本語固有表現まとめ上げおよび分類手法を定式化する. 特に, 固有表現の構成素となる各形態素の周囲の形態素のモデル化に関して, 既存の固有表現抽出手法よりもより多くの情報を参照するモデルを提案する.

[2] BREAKFAST の品詞タグの種類数は約 300 であり, 新聞記事に対しては 99.6%の品詞正解率である.

表 3: 固有表現まとめ上げ状態の表現法

固有表現タグ 形態素列	…	M	<ORG> M	M	<LOC> M	M	M	<LOC> M	M	…
Inside/Outside 法		O	ORG_I	O	LOC_I	LOC_I	LOC_I	LOC_B	O	
Start/End 法		O	ORG_U	O	LOC_S	LOC_C	LOC_E	LOC_U	O	

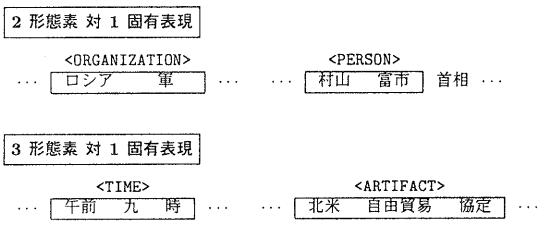

図 1: 複数形態素が一つの固有表現に対応する例

3.1 問題設定

まず，本論文における日本語固有表現まとめ上げおよび分類の問題を以下のように定義する．いま，以下に示すような形態素列が与えられているとする．

$$\begin{array}{c}\text{(左側文脈)} \qquad\qquad \text{(右側文脈)} \\ \cdots M_{-k}^{L} \cdots M_{-1}^{L} \;\; M_0 \;\; M_1^{R} \cdots M_{l}^{R} \cdots \\ \uparrow \\ \text{(現在位置)}\end{array}$$

ここで，現在の位置が形態素 M_0 のところであるとすると，日本語固有表現まとめ上げおよび分類の問題とは，この現在位置の形態素 M_0 に，まとめ上げ状態 (詳細は 3.2 節で述べる) および固有表現タイプを付与することである．

本論文の統計的固有表現抽出においては，訓練データからの教師あり学習により固有表現抽出モデルを学習する．その際には，各固有表現がどの形態素から構成されているかという情報が利用可能で，そのような情報を用いて固有表現抽出モデルを学習する．例えば，以下の例では，現在の位置に相当する形態素 M_i^{NE} が m 個の形態素からなる固有表現の一部であるという情報が利用可能である．

$$\begin{array}{ccc}\text{(左側文脈)} & \text{(固有表現)} & \text{(右側文脈)} \\ \cdots M_{-k}^{L} \cdots M_{-1}^{L} & M_1^{NE} \cdots M_i^{NE} \cdots M_m^{NE} & M_1^{R} \cdots M_{l}^{R} \cdots \\ & \uparrow & \\ & \text{(現在位置)} & \end{array} \qquad (1)$$

3.2 固有表現まとめ上げ状態の表現法

本論文では，固有表現まとめ上げの際のまとめ上げ状態の表現法として，以下の二つの方法を採用しその性能を評価した．この二つの方法は，いずれも，日本語固有表現抽出あるいは英語単名詞句まとめ上げにおいてよく研究されている方法である．これらの二種類の方法により日本語固有表現のまとめ上げを行なう様子を表 3 に示しておく．

3.2.1 Inside/Outside 法

この方法は英語の単名詞句まとめ上げでよく用いられる方法の一つである [Ramshaw95]．単名詞句まとめ上げの場合には，まとめ上げ状態として以下の三種類の状態を設定する．

O – 現在位置の単語はどの単名詞句にも含まれない．
I – 現在位置の単語は一つの単名詞句の一部である．
B – 現在位置の単語は，ある単名詞句の直後の位置する別の単名詞句の先頭の単語である．

本論文では，この方法を固有表現まとめ上げおよび分類に適用し，状態 I および B をそれぞれ八種類の固有表現タイプに細分類する[3]．結果として，この表現法では，固有表現まとめ上げ状態として，$2 \times 8 + 1 = 17$ の状態を設定する．

3.2.2 Start/End 法

この方法は，日本語固有表現抽出の既存の手法 [Sekine98, Borthwick99, 内元 00] において用いられた方法で，各固有表現タイプについて，以下の四種類のまとめ上げ状態を設定する．

S – 現在位置の形態素は，一つ以上の形態素から構成される固有表現の先頭の形態素である．
C – 現在位置の形態素は，一つ以上の形態素から構成される固有表現の先頭・末尾以外の中間の形態素である．
E – 現在位置の形態素は，一つ以上の形態素から構成される固有表現の末尾の形態素である．
U – 現在位置の形態素は単独で一つの固有表現を構成する．

また，固有表現を構成しない形態素のための状態として以下の状態を設定する．

O – 現在位置の形態素はどの固有表現にも含まれない．

結果として，この表現法では，固有表現まとめ上げ状態として，$4 \times 8 + 1 = 33$ の状態を設定する．

[3] 現在位置の形態素が，固有表現タイプ x に対するまとめ上げ状態 x_B を付与されるのは，タイプ x の固有表現 NE_1 の直後に同じタイプ x の別の固有表現 NE_2 が位置し，現在位置の形態素がその直後の固有表現 NE_2 の先頭の形態素である場合のみである．

3.3 周囲の形態素のモデル化

次に，本論文では，現在位置の形態素に対して固有表現のまとめ上げ状態を付与する際に，周囲のどれだけの形態素を考慮するか，つまり周囲の形態素をどのようにモデル化するかについて，大きく以下の二種類のモデルに分けてその性能を考察する．

3.3.1 固定長モデル

一つ目のモデルは，現在位置の形態素がどれだけの長さの固有表現を構成するのかを全く考慮せずに，固有表現まとめ上げ状態を付与するモデルである．これは，学習時においても，現在の形態素が，いくつの形態素からなる固有表現の一部であるか (式 (1) 参照) といった情報を全く考慮せず学習を行なうモデルである．このモデルにおいては，以下に示すように，現在位置の形態素 M_0 の左側および右側の文脈中の形態素については，学習時においても適用時においても，常に固定された数の形態素だけを考慮する．

(左側文脈) (右側文脈)
$\cdots M^L_{-k} \cdots M^L_{-1}\ M_0\ M^R_1 \cdots M^R_l \cdots$
↑
(現在位置)

本論文ではこのモデルのことを，**固定長モデル**と呼ぶ．

例えば，[Sekine98, Borthwick99] の手法のモデルは固定長モデルの一種で，左側および右側について常に一つずつの形態素のみを考慮している．本論文では [Sekine98, Borthwick99] の手法のモデルを 3 グラムモデルと呼ぶ．

(左側文脈) (現在位置) (右側文脈)
$$\cdots\ M_{-1}\qquad M_0\qquad M_1\ \cdots \tag{2}$$

また，[内元 00] の手法のモデルも固定長モデルの一種で，左側および右側について常に二つずつの形態素を考慮している．本論文では [内元 00] の手法のモデルを 5 グラムモデルと呼ぶ．

(左側文脈) (現在位置) (右側文脈)
$$\cdots\ M_{-2}M_{-1}\qquad M_0\qquad M_1M_2\ \cdots \tag{3}$$

3.3.2 可変長モデル

一方，もう一つの，本論文で新たに提案する方のモデルは，学習時において，現在位置の形態素が，いくつの形態素から構成される固有表現の一部であるか (式 (1) 参照) を考慮して学習を行なうモデルで，これを**可変長モデル**と呼ぶことにする．より具体的には，以下に示すように，現在位置の形態素 M^{NE}_i が m(ただし本論文では 3 以下) 個の形態素からなる固有表現の一部であるときに，固有表現を構成する形態素およびその左右の二個ずつの形態素を考慮して学習を行なうというモデルである．つまり，現在注目している固有表現の長さ m に応じて，考慮する周囲の形態素の総数が可変となる．

(左側文脈) (固有表現) (右側文脈)
$$\cdots\ M^L_{-2}M^L_{-1}\quad M^{NE}_1 \cdots M^{NE}_i \cdots M^{NE}_{m(\leq 3)}\quad M^R_1M^R_2\ \cdots \tag{4}$$
↑
(現在位置)

また，現在位置の形態素 M^{NE}_i が 4 個以上の形態素から構成される固有表現の一部であるときには，本論文では，固有表現を構成するとみなす形態素数を 3 に限定するという近似を行なう．例えば，以下のように，現在位置の形態素 M^{NE}_i が 4 個の形態素から構成される固有表現の一部である場合を考える．

(左側文脈) (固有表現) (右側文脈)
$\cdots\ M^L_{-2}M^L_{-1}\quad M^{NE}_1M^{NE}_2M^{NE}_3M^{NE}_4\quad M^R_1M^R_2\ \cdots$
↑
(現在位置)

この場合，固有表現を構成する末尾の形態素 M^{NE}_4 が，あたかも固有表現の直後の右側文脈に存在する形態素であるかのようにみなされ，以下のように近似されてモデル化される[4]．

(左側文脈) (固有表現) (右側文脈)
$\cdots\ M^L_{-2}M^L_{-1}\quad M^{NE}_1M^{NE}_2M^{NE}_3\quad M^{NE}_4M^R_1\ \cdots$
↑
(現在位置)

4 固有表現抽出規則の統計的学習

前節で述べた固有表現まとめ上げおよび分類のモデルに基づいて，本節では，固有表現抽出規則の統計的学習法を簡単に説明する．

4.1 素性およびクラス

各形態素の素性として用いるのは，語彙と品詞の組，文字種と品詞の組，品詞の三種類である．品詞は，形態素解析システム BREAKFAST の約 300 種類を用いる．また，文字種は，平仮名・片仮名・漢字・数字・英語アルファベット・記号，およびそれらの組合わせを用いる．実際

[4] この際の近似のされ方を，固有表現部分とみなされる三形態素の列で示すと，i) 現在位置の形態素が固有表現の先頭である場合は $M^{NE}_1M^{NE}_2M^{NE}_3$ (現在位置)↑，ii) 現在位置の形態素が固有表現の末尾である場合は $M^{NE}_{m-2}M^{NE}_{m-1}M^{NE}_m$ (現在位置)↑，iii) その他の場合は $M^{NE}_{i-1}M^{NE}_iM^{NE}_{i+1}$ (現在位置)↑，となる．

表 4: F 値 ($\beta = 1$) で評価した実験結果

		固定長モデル		
		3 グラム	5 グラム	可変長モデル
決定リスト	Inside/Outside	72.9	73.7	74.3
	Start/End	72.7	72.0	72.1
最大エントロピー	Inside/Outside	70.4	70.6	72.5
	Start/End	80.2	80.6	**85.9**

にこれらの各形態素の素性をどのように組合わせて統計的学習を行なうかは，次節で述べる各学習法により異なる．クラスは，Inside/Outside 法あるいは Start/End 法に応じて，3.2 節で述べた各固有表現まとめ上げ状態がその値となる．

4.2 統計的学習法

決定リスト学習法および最大エントロピー法を用いる．

4.2.1 決定リスト学習

決定リスト学習法としては，[Yarowsky94] の方法を用いる (詳細は [Sassano00] 参照)．素性の基本的な考え方としては，3.3 節の固定長モデル・可変長モデルのいずれにおいても，参照する全ての形態素の素性の組合わせを素性候補とする．その際，現在位置の形態素については何らかの素性を考慮する必要があるが，その他の周囲の文脈の形態素の素性については，情報を省略することが可能である．

4.2.2 最大エントロピー法

最大エントロピー法としては，[内元 00] の方法を用いる．個々の形態素の素性としては，4.1 節で述べたものを用いる．ただし，[内元 00] の基準に従い，語彙としては，訓練データとして用いる CRL 固有表現データ中で，固有表現の位置およびその前後二形態素ずつの範囲における出現頻度が 5 以上の語彙のみを用い，また，素性とクラスの組に相当する素性関数としては，頻度 3 以上のもののみを用いる．

5 実験および評価

5.1 概要

CRL 固有表現データを訓練データとし，本試験データのうちの一般ドメインのものを評価データとして，{Inside/Outside 法,Start/End 法} × { 固定長モデル (3 グラム，5 グラム)，可変長モデル } × { 決定リスト学習，最大エントロピー法 } の組み合わせについて，固有表現抽出規則の学習を行ないその性能を F 値 ($\beta = 1$) で評価した結果を表 4 に示す．ただし，以下で示す数値は全て，表 2 の「形態素と固有表現の対応パターン」のうち「その他」(5.5%) に該当する固有表現を除外して測定した値である．

これらの組合わせのうち，最も高い性能を示したのは，Start/End 法+可変長モデル+最大エントロピー法という組合わせ (太字) であった．統計的学習法としては，最高の性能において最大エントロピー法が決定リスト学習を上回る結果となった．これは，日本語固有表現抽出のタスクにおいては，決定リスト学習のように，複数の素性の組合わせの結合事象を素性選択し，適用時には，選択されなかった素性の情報を考慮しないモデルよりも，最大エントロピー法のように，各素性を独立とみなして学習を行なうが，適用時には重み付きであらゆる素性を考慮するモデルの方が高い性能を示すことを意味する．また，決定リスト学習においては，Inside/Outside 法と可変長モデルの組合わせが，また，最大エントロピー法においては，Start/End 法と可変長モデルの組合わせが，それぞれ最高の性能を示したことから，基本的には，可変長モデルが固定長モデルよりも高い性能を示すことが確認できた[5]．

さらに，学習法と固有表現まとめ上げ状態の表現法との組合わせとしては，決定リスト学習においては Inside/Outside 法が，また，最大エントロピー法においては Start/End 法が，それぞれより高い性能を示した．ここで，この理由を説明するために，それぞれのモデルの汎化の度合いについて考えてみる．まず，Inside/Outside 法と Start/End 法では，固有表現まとめ上げ状態の数が異なっており，状態数の少ない Inside/Outside 法の方がより粗いモデル化を，逆に状態数の多い Start/End 法の法がより細かいモデル化をしていると言える．一方，学習法については，上述したように，決定リスト学習が複数素性の結合事象を考慮するのに対して，最大エントロピー法では素性は独立とみなされていることから，決定リスト学習がより細かいモデル化を，最大エントロピー法がより粗いモデル化をしていると言える．これらのことから，決定リスト学習においては，決定リスト学習+Start/End 法という組合わせがモデルとしては細か過ぎるために，Inside/Outside 法がより高い性能を示したと推測できる．一方，最大エントロピー法におい

[5] なお，先行研究 [内元 99] においては，最大エントロピー法と Start/End 法の組合わせにおいて，現在位置の形態素の前後三形態素までを考慮するモデル (7 グラムモデル) を評価した結果，5 グラムモデルの性能を下回るという結果が得られている．つまり，5 グラムモデルの性能を上回るためには，固定長モデルにおいて単に考慮する形態素数を増やすのではなく，可変長モデルによって周囲の形態素のモデル化を柔軟に調整することが不可欠であると結論づけることができる．

表 5: 固定長モデル (5 グラム) と可変長モデルの比較 (Start/End 法+最大エントロピー)

		n 形態素 対 1 固有表現				
		$n \geq 1$	$n = 1$	$n = 2$	$n = 3$	$n \geq 4$
固定長モデル (5 グラム)	F 値 ($\beta = 1$)	80.6	83.1	86.9	68.2	46.2
	(適合率)	(84.2)	(83.1)	(89.2)	(74.4)	(79.6)
	(再現率)	(77.2)	(83.2)	(84.6)	(62.9)	(32.5)
可変長モデル	F 値 ($\beta = 1$)	85.9	86.6	89.7	80.9	72.4
	(適合率)	(88.6)	(88.0)	(92.4)	(82.0)	(84.4)
	(再現率)	(83.5)	(85.3)	(87.2)	(79.7)	(63.3)

表 6: 文字種および OTHER の細分化の効果 (Start/End 法+最大エントロピー, F 値)

		細分化あり	細分化なし
固定長モデル (5 グラム)	文字種あり	81.2 (77.9)	80.6 (77.3)
	文字種なし	80.0 (76.8)*	78.9 (75.6)
可変長モデル	文字種あり	**86.2 (82.8)**	85.9 (82.5)
	文字種なし	85.3 (81.9)	85.3 (81.8)

(括弧内: 表 2 の「その他」(5.5%) を除外しない F 値)

ては，逆に，最大エントロピー法+Inside/Outside 法という組合わせがモデルとしては粗過ぎるために，Start/End 法がより高い性能を示したと推測できる．

5.2 固定長モデル (5 グラム) と可変長モデルの比較

次に，Start/End 法+最大エントロピーの組合わせにおいて，固定長モデル (5 グラム) と可変長モデルの F 値 ($\beta = 1$)・適合率・再現率を，固有表現を構成する形態素の数ごとに測定した結果を表 5 に示す．これから分かるように，$n = 3, n \geq 4$ といった長い固有表現において可変長モデルの効果が顕著に現れている．また，適合率・再現率ともにすべての長さにおいて，可変長モデルが固定長モデルを上回っていることが分かる．

5.3 文字種および OTHER の細分化の効果

最後に，[内元 00] の実験結果との比較を行なうための追加実験を行なった．[内元 00] の実験と我々の実験との違いとして，[内元 00] では文字種の情報は使っていないが，固有表現を構成しない形態素のクラスを細分化し，固有表現の直前，直後，直前かつ直後，およびその他の四クラスを設定している．このうち，固有表現を構成しない形態素のクラスの細分化については，F 値で 1.33 ポイントの効果があるとしている．そこで，ここでは，最大エントロピー法において，文字種の情報を利用するか否か，および固有表現を構成しない形態素のクラスを細分化するか否かの四通りの実験を行ない，それぞれについて，固定長モデル (5 グラム) および可変長モデルの F 値 ($\beta = 1$) を測定した．この結果を，表 6 に示す．ここで，表中で括弧内に示した値は，表 2 の「その他」(5.5%) に該当する固有表現を除外しないで測定した F 値，すなわち IREX ワークショップの公式スコアに相当する値である．

この結果から分かるように，可変長モデルにおいて，文字種を利用し，かつ，固有表現を構成しない形態素のクラスを細分化した場合 (太字) が最も性能が高い．しかも，IREX ワークショップの公式スコアに相当する F 値 (82.8) の方は，統計的学習を用いるシステム中の最高の成績 ([内元 00] の 79.42) を上回っている．

また，[内元 00] の実験の設定は，文字種を利用せず，固有表現を構成しない形態素のクラス (OTHER) を細分化した場合 ('*' マーク) に相当する．[内元 00] の実験でこの F 値 (76.8) に相当するのは，「その他」の 5.5%のデータのための書き換え規則を利用する前の段階の性能で，その F 値は 78.05 である．また，「その他」の 5.5%のデータのための書き換え規則を経た後の F 値は 79.42 である．我々の実験の F 値 (76.8) と [内元 00] の実験の F 値 (78.05) の差の原因としては，[内元 00] の実験では，訓練データとして，CRL 固有表現データに加えて約 1,000 文ほどのデータを追加していること，および形態素解析システムの品詞の違い (我々が約 300 種類の品詞を用いているのに対して，[内元 00] の実験では大分類 15 個，細分類 48 個を用いている) が考えられる．仮に，我々が，この二つの違いを取り込み，かつ可変長モデルで実験を行ない，かつ，「その他」の 5.5%のデータのための書き換え規則まで用いたとすると，IREX ワークショップの公式スコアとして，$82.8 + (79.42 - 76.8) = 85.42$ 程度の性能が期待できる[6]．

[6] IREX ワークショップの固有表現抽出タスクの最高性能は，人手によって構築されたシステムによって達成された 83.86 である [IREX 実行委員会 99].

6 おわりに

本論文では，統計的手法に基づく日本語固有表現のまとめ上げの問題に対して，性能を大きく左右する四つの要因，i) 固有表現まとめ上げ状態の表現法，ii) 現在位置の周囲の形態素を何個まで考慮するか，iii) 個々の形態素の素性，iv) 統計的学習法，について，これまで日本語固有表現のまとめ上げにおいてはその有効性が確認されていない幾つかの方式を実験的に評価し，その得失について報告した．特に，ii) について，現在位置の形態素が，いくつの形態素から構成される固有表現の一部であるかを考慮して学習を行なう可変長モデルを新たに提案した．また，実験の結果，先行研究 [Borthwick99, 内元 00] で用いられた固定長モデルの性能を大きく上回る結果が得られ，可変長モデルの有効性が確認できた．現在の性能をさらに向上させる可能性のある有望な方式として，i) 委員会方式など複数のモデルの結果を統合する方式，ii) ブースティング [Freund99] や変形に基づく学習 [Brill95, 颯々野 99] などの誤り駆動学習の方式，などが存在するので，今後はそれらの方式の効果を検証する．

謝辞

本研究の成果の多くは，筆者らが米国ジョンズホプキンス大学計算機科学科に客員研究員として滞在中に得られたものである．本研究に対し多くの貴重なコメントを頂いた同大学 David Yarowsky 教授に感謝する．また，最大エントロピー法を用いた実験に協力して頂いた，郵政省通信総合研究所 内元清貴氏に感謝する．

参考文献

[Borthwick99] Borthwick, A.: 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 (1999).

[Brill95] Brill, E.: Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part-of-Speech Tagging, *Computational Linguistics*, Vol. 21, No. 4, pp. 543–565 (1995).

[Collins99] Collins, M. and Singer, Y.: Unsupervised Models of Named Entity Classification, *Proceedings of the 1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pp. 100–110 (1999).

[Cucerzan99] Cucerzan, S. and Yarowsky, D.: Language Independent Named Entity Recognition Combining Morphological and Contextual Evidence, *Proceedings of the 1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pp. 90–99 (1999).

[Freund99] Freund, Y. and Schapire, R.: (訳: 安倍 直樹): ブースティング入門, 人工知能学会誌, Vol. 14, No. 5, pp. 771–789 (1999).

[IREX 実行委員会 99] IREX 実行委員会（編）: IREX ワークショップ予稿集 (1999).

[Maiorano96] Maiorano, S.: The Multilingual Entity Task (MET): Japanese Results, *Proceedings of TIPSTER PROGRAM PHASE II*, pp. 449–451 (1996).

[MUC98] MUC: *Proceedings of the 7th Message Understanding Conference (MUC-7)* (1998).

[Ramshaw95] Ramshaw, L. and Marcus, M.: Text Chunking using Transformation-Based Learning, *Proceedings of the 3rd Workshop on Very Large Corpora*, pp. 83–94 (1995).

[颯々野 97] 颯々野学, 斎藤由香梨, 松井くにお: アプリケーションのための日本五形態素解析システム, 言語処理学会第 3 回年次大会論文集, pp. 441–444, 言語処理学会 (1997).

[颯々野 99] 颯々野学, 塚本浩司: 有限状態変換器の誤り駆動学習を用いた固有表現抽出, *IPSJ SIG Notes*, Vol. 99, No. (99–NL–132), pp. 1–8 (1999).

[Sassano00] Sassano, M. and Utsuro, T.: Named Entity Chunking Techniques in Supervised Learning for Japanese Named Entity Recognition, *Proceedings of the 18th COLING*, pp. 705–711 (2000).

[Sekine98] Sekine, S., Grishman, R. and Shinnou, H.: A Decision Tree Method for Finding and Classifying Names in Japanese Texts, *Proceedings of the 6th Workshop on Very Large Corpora*, pp. 148–152 (1998).

[内元 99] 内元清貴, 村田真樹, 小作浩美, 馬青: ME モデルと書き換え規則に基づく固有表現抽出 — IREX–NE 本試験における評価 —, IREX ワークショップ予稿集, pp. 133–140 (1999).

[内元 00] 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).

[Utsuro00a] Utsuro, T. and Sassano, M.: Minimally Supervised Japanese Named Entity Recognition: Resources and Evaluation, *Proceedings of the 2nd International Conference on Language Resources and Evaluation*, pp. 1229–1236 (2000).

[宇津呂 00b] 宇津呂武仁, 颯々野学: ブートストラップによる低人手コスト日本語固有表現抽出, 情報処理学会研究報告, Vol. 2000, No. (2000–NL–139) (2000).

[Yarowsky94] Yarowsky, D.: Decision Lists for Lexical Ambiguity Resolution: Application to Accent Restoration in Spanish and French, *Proceedings of the 32nd Annual Meeting of ACL*, pp. 88–95 (1994).

[Yarowsky95] Yarowsky, D.: Unsupervised Word Sense Disambiguation Rivaling Supervised Methods, *Proceedings of the 33rd Annual Meeting of ACL*, pp. 189–196 (1995).