正誤判別規則学習を用いた 複数の日本語固有表現抽出システムの出力の混合

宇津呂 武仁[†]　颯々野　学[††]　内元　清貴[†††]

本論文では，日本語固有表現抽出の問題において，複数のモデルの出力を混合する手法を提案する．一般に，複数のモデル・システムの出力の混合を行なう際には，まず，できるだけ振る舞いの異なる複数のモデル・システムを用意する必要がある．本論文では，最大エントロピー法に基づく統計的学習による固有表現抽出モデルにおいて，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を考慮して学習を行なう可変 (文脈) 長モデルと，常に現在位置の形態素の前後数形態素ずつまでを考慮して学習を行なう固定 (文脈) 長モデルとの間のモデルの挙動の違いに注目する．そして，複数のモデルの挙動の違いを調査し，なるべく挙動が異なり，かつ，適度な性能を保った複数のモデルの出力の混合を行なう．次に，混合の方式としては，複数のシステム・モデルの出力 (および訓練データそのもの) を入力とする第二段目の学習器を用いて，複数のシステム・モデルの出力の混合を行なう規則を学習するという混合法 (stacking 法) を採用する．第二段目の学習器として決定リスト学習を用いて，固定長モデルおよび可変長モデルの出力を混合する実験を行なった結果，最大エントロピー法に基づく固有表現抽出モデルにおいてこれまで得られていた最高の性能を上回る性能が達成された．

キーワード:　日本語固有表現抽出，複数システム混合，*stacking*，可変文脈長，最大エントロピー法，決定リスト学習

Learning to Combine Outputs of Multiple Japanese Named Entity Extractors

Takehito Utsuro[†] , Manabu Sassano[††] and Kiyotaka Uchimoto[†††]

In this paper, we propose a method for learning a classifier which combines outputs of more than one Japanese named entity extractors. The proposed combination method belongs to the family of *stacked generalizers*, which is in principle a technique of combining outputs of several classifiers at the first stage by learning a second stage classifier to combine those outputs at the first stage. Individual models to be combined are based on maximum entropy models, one of which always considers surrounding contexts of a fixed length, while the other considers those of variable lengths according to the number of constituent morphemes of named entities. As an algorithm for learning the second stage classifier, we employ a decision list learning method. Experimental evaluation shows that the proposed method achieves improvement over the best known results with Japanese named entity extractors based on maximum entropy models.

KeyWords:　*Japanese named entity extraction, system combination, stacking, variable context length, maximum entropy model, decision list learning*

1　はじめに

これまで，機械学習などの分野を中心として，複数のモデル・システムの出力を混合する手法がいくつか提案され，その効果が報告されている．それらの成果を背景として，近年，統計的手法に基づく自然言語処理においても，複数のモデル・システムの出力を混合する手法を様々な問題に適用することが試みられ，品詞付け (van Halteren, Zavrel and Daelemans 1998; Brill and Wu 1998; Abney, Schapire and Singer 1999)，名詞句等の句のまとめ上げ (Sang 2000; 工藤, 松本 2000)，構文解析 (前置詞句付加含む) (Henderson and Brill 1999; Abney et al. 1999; 乾, 乾 2000; Henderson and Brill 2000) などへの適用事例が報告されている．一般に，複数のモデル・システムの出力を混合することの利点は，単一のモデル・システムでは，全ての現象に対して網羅的かつ高精度に対処できない場合でも，個々のモデル・システムがそれぞれ得意とする部分を選択的に組み合わせることで，全体として網羅的かつ高精度なモデル・システムを実現できるという点にある．本論文では，日本語固有表現抽出の問題に対して，複数のモデルの出力を混合する手法を適用し，個々の固有表現抽出モデルがそれぞれ得意とする部分を選択的に組み合わせることで，全体として網羅的かつ高精度なモデルを実現し，その効果を実験的に検証する．

一般に，日本語固有表現抽出においては，前処理として形態素解析を行ない，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 1999)．特に，統計的学習によって得られた固有表現抽出規則を用いる場合には，形態素解析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手順が一般的である (Sekine, Grishman and Shinnou 1998; Borthwick 1999; 内元, 馬, 村田, 小作, 内山, 井佐原 2000; Sassano and Utsuro 2000; 颯々野, 宇津呂 2000; 山田, 工藤, 松本 2001)．このとき，実際のまとめ上げの処理は，現在注目している位置にある形態素およびその周囲の形態素の語彙・品詞・文字種などの属性を考慮しながら，現在位置の形態素が固有表現の一部となりうるかどうかを判定することの組合わせによって行なわれる．

一方，一般に，複数のモデル・システムの出力を混合する過程は，大きく以下の二つの部分に分けて考えることができる．

(1)　できるだけ振る舞いの異なる複数のモデル・システムを用意する．(通常，振る舞いの酷似した複数のモデル・システムを用意しても，複数のモデル・システムの出力を

† 豊橋技術科学大学 工学部 情報工学系
, Department of Information and Computer Sciences, Toyohashi University of Technology
†† 富士通研究所, Fujitsu Laboratories, Ltd.,
††† 独立行政法人 通信総合研究所 けいはんな情報通信融合センター, Keihanna Human Info-Communications Research Center, Communications Research Laboratory, Independent Administrative Institution

混合することによる精度向上は望めないことが予測される．)

(2) 用意された複数のモデル・システムの出力を混合する方式を選択・設計し，必要であれば学習等を行ない，与えられた現象に対して，用意された複数のモデル・システムの出力を混合することを実現する．

複数の日本語固有表現抽出モデルの出力を混合するにあたっても，これらの (1) および (2) の過程をどう実現するかを決める必要がある．

本論文では，まず，(1) については，統計的学習を用いる固有表現抽出モデルをとりあげ，まとめ上げの処理を行なう際に，現在位置の周囲の形態素を何個まで考慮するかを区別することにより，振る舞いの異なる複数のモデルを学習する．そして，複数のモデルの振る舞いの違いを調査し，なるべく振る舞いが異なり，かつ，適度な性能を保った複数のモデルの混合を行なう．特に，これまでの研究事例 (Sekine et al. 1998; Borthwick 1999; 内元他 2000; 山田他 2001) 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，常に現在位置の形態素の前後二形態素 (または一形態素) ずつまでを考慮して学習を行なうモデル (固定長モデル，3.5.1 節参照) だけではなく，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モデル (可変長モデル (Sassano and Utsuro 2000; 颯々野, 宇津呂 2000)，3.5.2 節参照) も用いて複数モデルの出力の混合を行なう．

次に，(2) については，重み付多数決やモデルの切り替えなど，これまで自然言語処理の問題によく適用されてきた混合手法を原理的に包含し得る方法として，stacking 法 (Wolpert 1992) と呼ばれる方法を用いる．stacking 法とは，何らかの学習を用いた複数のシステム・モデルの出力 (および訓練データそのもの) を入力とする第二段の学習器を用いて，複数のシステム・モデルの出力の混合を行なう規則を学習するという混合法である．本論文では，具体的には，複数のモデルによる固有表現抽出結果，およびそれぞれの固有表現がどのモデルにより抽出されたか，固有表現のタイプ，固有表現を構成する形態素の数と品詞などを素性として，各固有表現が正しいか誤っているかを判定する第二段の判定規則を学習し，この正誤判定規則を用いることにより複数モデルの出力の混合を行なう．

以下では，まず，2 節で，本論文の実験で使用した IREX(Information Retrieval and Extraction Exercise) ワークショップ (IREX 実行委員会 1999) の日本語固有表現抽出タスクの固有表現データについて簡単に説明する．次に，3 節では，個々の固有表現抽出モデルのベースとなる統計的固有表現抽出モデルについて述べる．本論文では，統計的固有表現抽出モデルとして，最大エントロピー法を用いた日本語固有表現抽出モデル (Borthwick 1999; 内元他 2000) を採用する．最大エントロピー法は，自然言語処理の様々な問題に適用されその性能が実証されているが，日本語固有表現抽出においても高い性能を示しており，IREX ワークショップの日本語固有表現抽出タスクにおいても，統計的手法に基づくシステムの中で最も高い成績を達成し

ている (内元他 2000)．4 節では，複数のモデルの出力の正誤判別を行なう規則を学習することにより，複数モデル出力の混合を行なう手法を説明する．本論文では，正誤判別規則の学習モデルとしては，決定リスト学習を用い，その性能を実験的に評価する．

以上の手法を用いて，5 節で，複数の固有表現抽出結果の混合法の実験的評価を行ない，提案手法の有効性を示す．(内元他 2000) にも示されているように，固定長モデルに基づく単一の日本語固有表現抽出モデルの場合は，現在位置の形態素の前後二形態素ずつを考慮して学習を行なう場合が最も性能がよい．また，5 節の結果からわかるように，この，常に前後二形態素ずつを考慮する固定長モデルの性能は，可変長モデルに基づく単一のモデルの性能をも上回っている (なお，(颯々野, 宇津呂 2000) では，最大エントロピー法を学習モデルとして可変長モデルを用いた場合には，常に前後二形態素ずつを考慮する固定長モデルよりも高い性能が得られると報告しているが，この実験結果には誤りがあり，本論文で示す実験結果の方が正しい．)．ところが，可変長モデルと，現在位置の形態素の前後二形態素ずつを考慮する固定長モデルとを比較すると，モデルが出力する固有表現の分布がある程度異なっており，実際，これらの二つのモデルの出力を用いて複数モデル出力の混合を行なうと，個々のモデルを上回る性能が達成された．5 節では，これらの実験について詳細に述べ，本論文で提案する混合法が有効であることを示す．

2　日本語固有表現抽出

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 (MUC 1998)) 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 (Maiorano 1996), MET-2 (MUC 1998)) や IREX ワークショップ (IREX 実行委員会 1999) 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．

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 1999)．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 — 現，独立行政法人 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．

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

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	

2.2　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, 斎藤, 松井 1997)[1] 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，一つの固有表現が複数の形態素から構成されている場合は 90%近く (7175/ (7175+1022) = 87.5%) を占めており，これらの固有表現については，各固有表現の区切り位置はいずれかの形態素の区切り位置と一致している，すなわち，固有表現の開始位置は，先頭の構成要素となる形態素の開始位置と，また，固有表現の終了位置は，末尾の構成要素となる形態素の終了位置と，それぞれ一致する．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 (内元他 2000) 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．

1 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．

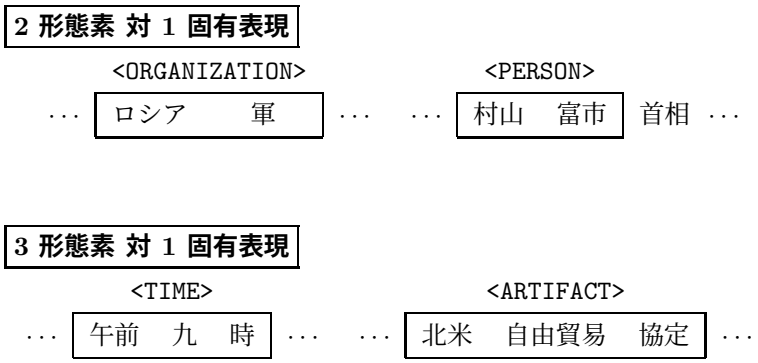

図 1　複数形態素が一つの固有表現に対応する例

3　最大エントロピー法を用いた固有表現抽出

本節では，まず，ベースモデルとなる，最大エントロピー法を用いた日本語固有表現抽出の手法 (Borthwick 1999; 内元他 2000) を定式化する．

3.1　問題設定

ここでの固有表現抽出の問題は，固有表現まとめ上げおよび固有表現タイプ分類の問題ととらえることができる．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．

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

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固有表現タイプ分類の問題とは，この現在位置の形態素 M_0に，まとめ上げ状態および固有表現タイプ (詳細は 3.3 節で述べる) 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．

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

また，次節で述べる最大エントロピー法を用いて固有表現抽出モデルを学習する際には，現在位置および周囲の形態素の素性 (3.4 節) を条件として，現在位置の形態素に固有表現まとめ上げ状態およびタイプ (3.3 節) をクラスとして付与するための条件付確率モデルを最大エントロピー法により学習する．

なお，通常，学習された確率モデルを適用して，形態素に固有表現まとめ上げ状態および固有表現タイプを付与することにより，固有表現の抽出を行なう場合は，一文全体で，固有表現まとめ上げ状態および固有表現タイプの確率を最大とする固有表現の組合わせを求める必要がある．本論文では，この最適解探索の方法としては，(内元他 2000) のものをそのまま用いている．

3.2　最大エントロピー法

最大エントロピー法は，文脈を規定する制約を素性として与え，与えられた素性のもとでエントロピーを最大化するという条件によって求められる確率モデルである．確率モデルの学習においてエントロピーを最大化することにより，与えられた制約を満たす最も一様なモデルが学習されるため，データの過疎性に強いという特徴を持つ．

ここでは，与えられた訓練集合から，文脈 $x(\in \mathcal{X})$ においてクラス $y(\in \mathcal{Y})$ を出力するプロセスの確率的振舞い，すなわち条件付確率分布 $p(y \mid x)$ を最大エントロピー法に基づいて推定する方法の概略を説明する．

まず，訓練集合中の事象 (x, y) の観測値を大量に集め，$freq(x, y)$ を事象 (x, y) の訓練集合中での生起頻度として，訓練集合中の経験的確率分布$\tilde{p}(x, y)$ を以下のように推定する．

$$\tilde{p}(x, y) \equiv \frac{freq(x, y)}{\sum_{x,y} freq(x, y)}$$

次に，訓練集合中のどのような現象に注目して確率分布を推定するのかを表す二値の関数 $f(x, y)$ を導入し，これを素性関数と呼ぶ．具体的には，各素性関数 f_iについて，この関数が真となる事象 x および yの集合 V_{xi}および V_{yi}が規定されていると考え，この集合にしたがって素性関数 f_iが以下のように定義される．

$$f_i(x, y) = \begin{cases} 1 & (x \in V_{xi}\text{かつ } y \in V_{yi}\text{の場合}) \\ 0 & (\text{それ以外の場合}) \end{cases}$$

表 3　固有表現まとめ上げ状態の表現法

固有表現タグ			<ORG>			<LOC>			<LOC>		
形態素列	$\cdots$	M	M	M	M	M	M	M	M	M	$\cdots$
固有表現まとめ上げ状態		O	ORG_U	O	LOC_S	LOC_C	LOC_E		LOC_U	O	

また，一般に確率モデル学習の際には，大量の素性からなる素性の候補集合 $\mathcal{F}$から，活性化された素性の部分集合 $\mathcal{S}(\subseteq \mathcal{F})$ が選択され，これらによって事象 (x, y) および確率分布 $p(y \mid x)$ が記述される．

次に，実際に確率モデル学習を行う際には，活性化された素性集合 $\mathcal{S}$中の各素性 f_iについて，学習すべき確率分布 $p(y \mid x)$ による素性 f_iの期待値 (左辺) と経験的確率分布$\tilde{p}(x, y)$ による素性 f_iの期待値 (右辺) が等しいとする以下の制約等式を課す．

$$\sum_{x,y} \tilde{p}(x)p(y \mid x)f_i(x, y) \quad = \quad \sum_{x,y} \tilde{p}(x, y)f_i(x, y) \quad \text{for } \forall f_i \in \mathcal{S}$$

そして，これらの制約等式を満たす確率分布 $p(y \mid x)$ のうちで，以下の条件付エントロピー $H(p)$ を最大にする最も「一様な」モデルが，求めるべきモデル p_*であるとする．

$$\begin{aligned} H(p) \quad &\equiv \quad -\sum_{x,y} \tilde{p}(x)p(y \mid x) \log p(y \mid x) \\ p_* \quad &= \quad \operatorname*{argmax}_{p \in \mathcal{C}(\mathcal{S})} H(p) \end{aligned} \tag{2}$$

(2) 式を満たす確率分布は必ず存在し，それは以下の確率分布 $p_\lambda(y \mid x)$ で記述される．

$$p_\lambda(y \mid x) \quad = \quad \frac{\exp\Big(\sum_i \lambda_i f_i(x, y)\Big)}{\sum_y \exp\Big(\sum_i \lambda_i f_i(x, y)\Big)}$$

ただし，λ_iは各素性 f_iのパラメータである．また，実際にエントロピーを最大にする最適なパラメータ$\lambda^*{}_i$を推定するには，Improved Iterative Scaling(IIS) アルゴリズム (Della Pietra, Della Pietra and Lafferty 1997; Berger, Della Pietra and Della Pietra 1996) と呼ばれるアルゴリズムが用いられる．

3.3　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，日本語固有表現抽出の既存の手法 (Sekine et al. 1998; Borthwick 1999; 内元他 2000) において用いられた

Start/End 法を採用する[2]．この方法では，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，二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，三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，二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．こ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に示す．

3.4　各形態素の素性

各形態素の素性としては，以下の三種類のものを用いる[3]．

(1)　語彙 — 訓練コーパス中で，固有表現の位置および周囲二形態素以内に 5 回以上出現した 2,052 語彙[4]．

(2)　品詞 — 形態素解析システム BREAKFAST の約 300 種類の品詞．

(3)　文字種 — 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．

3.5　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，以下の二種類のモデルを用いる．

3.5.1　固定 (文脈) 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

2 この他に，まとめ上げ問題でよく用いられる Inside/Outside 法が知られているが，最大エントロピー法との組合わせで日本語固有表現抽出を行なう場合は Start/End 法よりも性能が劣る (颯々野, 宇津呂 2000)．

3 これらの素性のうち，語彙素性を抽出する条件は (内元他 2000) に従っている．また，品詞素性については，(内元他 2000) とは，利用している形態素解析システムの品詞体系が異なっているため，異なった素性になっている．さらに，(内元他 2000) では，素性として文字種は用いていないが，文字種を用いた方が高い性能が得られることが分かっている (颯々野, 宇津呂 2000)．

4 例えば，頻度上位 10 位以内のものは，助詞 6 種類，括弧等の記号 3 種類，読点，11〜20 位は，助詞 3 種類，助動詞 1 種類，句点，助数詞 (“年”，“日”)，接尾辞 (“さん”)，地名 (“日本”)，時相名詞 (“昨年”)，21〜30 位は，助詞 3 種類，助動詞 2 種類，助数詞 (“%”，“円”)，接尾辞 (“氏”)，地名 (“ロシア”，“米国”) であった．

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 (3.1 節，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．

$$\begin{array}{cc} \text{(左側文脈)} & \text{(右側文脈)} \\ \cdots M_{-k} \cdots M_{-1} & M_0 \ M_1 \cdots M_l \cdots \\ & \uparrow \\ & \text{(現在位置)} \end{array}$$

本論文ではこのモデルのことを，**固定長モデル**と呼ぶ．本論文では特に，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をいくつ考慮するかに応じて，左右二形態素ずつを考慮する 5 グラムモデル

$$\begin{array}{ccccc} & \text{(左側文脈)} & \text{(現在位置)} & \text{(右側文脈)} & \\ \cdots & M_{-2}M_{-1} & M_0 & M_1M_2 & \cdots \end{array}$$

左右三形態素ずつを考慮する 7 グラムモデル

$$\begin{array}{ccccc} & \text{(左側文脈)} & \text{(現在位置)} & \text{(右側文脈)} & \\ \cdots & M_{-3}M_{-2}M_{-1} & M_0 & M_1M_2M_3 & \cdots \end{array}$$

左右四形態素ずつを考慮する 9 グラムモデル

$$\begin{array}{ccccc} & \text{(左側文脈)} & \text{(現在位置)} & \text{(右側文脈)} & \\ \cdots & M_{-4}M_{-3}M_{-2}M_{-1} & M_0 & M_1M_2M_3M_4 & \cdots \end{array}$$

を用いる．

3.5.2　可変 (文脈) 長モデル

一方，もう一つのモデルは，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 (颯々野, 宇津呂 2000; Sassano and Utsuro 2000)．

モデルの学習

学習時には，現在位置の形態素が固有表現を構成しない場合には，5 グラムモデルと同じく，現在位置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．一方，現在位置の形態素 M_i^{NE}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は，

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．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)

$$\cdots \quad M_{-2}^{L} M_{-1}^{L} \quad M_{1}^{NE} \cdots M_{i}^{NE} \cdots M_{m(\leq 3)}^{NE} \quad M_{1}^{R} M_{2}^{R} \quad \cdots$$

↑

(現在位置)

また，現在位置の形態素 M_i^{NE}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，以下の手順で，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．

(1)　現在位置の形態素が固有表現の先頭である場合は，先頭から三形態素のみが固有表現を構成するとみなし，四番目以降の形態素については右側文脈であるとみなす．

(左側文脈)　　(固有表現)　　(右側文脈)

$$\cdots \quad M_{-2}^{L} M_{-1}^{L} \qquad M_{1}^{NE} M_{2}^{NE} M_{3}^{NE} \qquad M_{4}^{NE} M_{?}^{?} \quad \cdots$$

↑

(現在位置)

(2)　現在位置の形態素が固有表現の末尾である場合は，末尾の三形態素のみが固有表現を構成するとみなし，末尾の三形態素以外については左側文脈であるとみなす．

(左側文脈)　　(固有表現)　　(右側文脈)

$$\cdots \quad M_{?}^{?} M_{m-3}^{NE} \qquad M_{m-2}^{NE} M_{m-1}^{NE} M_{m}^{NE} \qquad M_{1}^{R} M_{2}^{R} \quad \cdots$$

↑

(現在位置)

(3)　その他の場合は，現在位置の形態素およびその前後一形態素ずつのみが固有表現を構成するとみなし，それ以外の形態素については左側もしくは右側文脈であるとみなす．

(左側文脈)　　(固有表現)　　(右側文脈)

$$\cdots \quad M_{?}^{?} M_{?}^{?} \quad M_{i-1}^{NE} M_{i}^{NE} M_{i+1}^{NE} \quad M_{?}^{?} M_{?}^{?} \quad \cdots$$

↑

(現在位置)

例えば，以下のように，現在位置の形態素 M_i^{NE}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)

$$\cdots \quad M_{-2}^L M_{-1}^L \quad M_1^{NE} M_2^{NE} M_3^{NE} M_4^{NE} \quad M_1^R M_2^R \quad \cdots$$

↑

(現在位置)

この場合，固有表現を構成する末尾の形態素 M_4^{NE}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．

(左側文脈)　(固有表現)　(右側文脈)

$$\cdots \quad M_{-2}^L M_{-1}^L \quad M_1^{NE} M_2^{NE} M_3^{NE} \quad M_4^{NE} M_1^R \quad \cdots$$

↑

(現在位置)

モデルの適用

モデルの適用時には，現在位置の形態素がどのような固有表現を構成するかという情報が利用できないので，固定長の 9 グラムモデルの場合と同様に，現在位置の形態素，および，左右四形態素ずつの素性を考慮してモデルの適用を行なう[5]．

3.5.3　周囲の形態素の素性

前節までで述べた固定長モデルおよび可変長モデルにおいて，特に現在位置の周囲の形態素の素性について，3.4 節で述べた素性のうちの全部または一部のみを用いるモデルとして，以下の三種類のモデルを設定し，これらについて実験的評価を行なう[6]．

- 全素性を用いるモデル．
- 周囲の形態素 $M_{l(\leq -3)}$および $M_{r(\geq 3)}$については，語彙素性および品詞素性のみを考慮するモデル．
- 周囲の形態素 $M_{l(\leq -3)}$および $M_{r(\geq 3)}$については，語彙素性のみを考慮するモデル．

なお，(内元他 2000) と同様に，周囲の複数の形態素の素性を結合した結合素性は用いていない．

5 可変長モデルでは，モデルの学習時と適用時で考慮する素性の集合が異なっているので，単独での性能は高くないが，抽出される固有表現の分布が固定長モデルとは異なっている (5.1 節参照)．

6 実際に，実験で用いた訓練コーパスから学習したモデルのうち，全素性を用いた 5 グラムモデルの素性数は 13,200，素性関数の数は 31,344(頻度 3 以上)，全素性を用いた 9 グラムモデルの素性数は 15,071，素性関数の数は 35,311(頻度 3 以上) であった．

4　正誤判別規則学習を用いた複数システム出力の混合

4.1　訓練・評価データセット

本論文の複数システム出力の混合法では，以下の三種類の訓練・評価データセットを用いる．

(1)　TrI: 個々の固有表現抽出モデルを学習するための訓練データセット．

(2)　TrC: 複数システムの出力の正誤判別規則を学習するための訓練データセット．

(3)　Ts: 複数システムの出力の正誤判別規則を評価するための評価データセット．

4.2　訓練および評価手続きの概要

まず，以下に，訓練データセット TrIおよび TrCを用いて，複数システムの出力の正誤判別規則を学習するため手続きの概要を示す．

(1)　訓練データセット TrIを用いて，個々の固有表現抽出モデル $NEext_i$ $(i = 1, \ldots, n)$ を学習する．

(2)　個々の固有表現抽出モデル $NEext_i$ $(i = 1, \ldots, n)$ を，それぞれ，訓練データセット TrCに適用し，各固有表現抽出モデル $NEext_i$につき，抽出結果の固有表現リスト $NEList_i(TrC)$ をそれぞれ一つずつ得る．

(3)　訓練データセット (テキスト)TrC中での各固有表現の出現位置の情報を用いて，抽出結果の固有表現リスト $NEList_i(TrC)$ $(i{=}1, \ldots, n)$ を，複数システム間 $(i{=}1, \ldots, n)$ で整列し，訓練データセット TrCの事象表現 $TrCev$を作成する．

(4)　訓練データセット TrCの事象表現 $TrCev$を教師あり訓練データとして，複数システムの出力の正誤判別規則 $NEext_{cmb}$を学習する．

次に，評価データセット Ts に，学習された正誤判別規則 $NEext_{cmb}$を適用する手順の概要を示す．

(1)　個々の固有表現抽出モデル $NEext_i$ $(i = 1, \ldots, n)$ を，それぞれ，評価データセット Ts に適用し，各固有表現抽出モデル $NEext_i$につき，抽出結果の固有表現リスト $NEList_i(Ts)$ をそれぞれ一つずつ得る．

(2)　評価データセット (テキスト)Ts 中での各固有表現の出現位置の情報を用いて，抽出結果の固有表現リスト $NEList_i(Ts)$ $(i{=}1, \ldots, n)$ を，複数システム間 $(i{=}1, \ldots, n)$ で整列し，評価データセット Ts の事象表現 $Tsev$を作成する．

(3)　複数システムの出力の正誤判別規則 $NEext_{cmb}$を評価データセット Ts の事象表現 $Tsev$に適用し，性能を測定する．

4.3　データ構造

本節では，訓練データセット TrCの事象表現 $TrCev$，あるいは，評価データセット Ts の事象表現 $Tsev$のデータ構造を説明し，複数システムの出力の正誤判別規則を学習する際の素性・クラスについて述べる．以下では，訓練データセット TrCの事象表現 $TrCev$を例にして説明する．

4.3.1　事象

訓練データセット TrCの事象表現 $TrCev$は，訓練データセット (テキスト)TrC中での各固有表現の出現位置の情報を用いて，抽出結果の固有表現リスト $NEList_i(TrC)$ $(i{=}1,\ldots,n)$ を複数システム間 $(i{=}1,\ldots,n)$ で整列することにより作成される．ここで，整列結果の事象表現 $TrCev$は，セグメントの列 $Seg_1,\ldots,Seg_N$で表現され，各セグメント Seg_jは，整列された固有表現の集合 $\{NE_1,\ldots,NE_{m_j}\}$ によって表現される．

$$\begin{aligned} TrCev &= Seg_1,\ldots,Seg_N \\ Seg_j &= \{NE_1,\ldots,NE_{m_j}\} \end{aligned}$$

ただし，この整列の際には，少なくとも一つの形態素を共有する複数の固有表現は，同じセグメントに含まれなければならない，という制約が課せられる．

次に，各セグメント Seg_j中の固有表現の集合 $\{NE_1,\ldots,NE_{m_j}\}$ は，固有表現の事象表現の集合 $\{NEev_1,\ldots,NEev_{l_j}\}$ に変換され，これにより，各セグメント Seg_jは事象表現 $SegEv_j$に変換される．

$$SegEv_j = \{NEev_1,\ldots,NEev_{l_j}\} \tag{3}$$

ここで，各事象表現 $NEev_{k_j}$は，以下の二種類のうちのどちらかに対応し，それぞれ異なったデータ構造を持つ．

i)　そのセグメント中で少なくとも一つのシステムにより出力された固有表現の事象表現．

ii)　そのセグメント中で一つも固有表現を出力しなかった一つのシステムに関する情報を表す事象表現．

i) のタイプの事象表現 $NEev_{k_j}$は以下のようなデータ構造を持つ．

$$\begin{aligned} NEev_{k_j} = \Big\{ & systems = \langle p,\ldots,q\rangle,\ mlength = x \text{ morphemes}, \\ & NEtag = \cdots,\ POS = \cdots,\ class_{NE} = +/- \Big\} \end{aligned}$$

ここで，“$systems$” はこの固有表現を出力したシステムの指標のリストを，“$mlength$” はこの固有表現を構成する形態素の数を，“$NEtag$” はこの固有表現のタイプを，“POS” はこの固有表現を構成する形態素の数の品詞のリストを，それぞれ表す．また，“$class_{NE}$” は，正解デー

タと比較して，この固有表現が正解であるか (“+”)，それとも，システムによる誤出力であるか (“–”) を示す.

一方，ii) のタイプの事象表現 $NEev_{k_j}$ は，このセグメント中で，指標 r を持つシステムが固有表現を出力しなかったことを示す，以下のようなデータ構造を持つ.

$$NEev_{k_j} \quad = \quad \Big\{ systems = \langle r \rangle,\ class_{sys} = \text{“no output”} \Big\} \tag{4}$$

4.3.2 クラス

複数システムの出力の正誤判別を行なう規則は，式 (3) で定義されるセグメントの事象表現 $SegEv_j$ を一つの事象単位として，学習および適用が行なわれる．ここで，正誤判別規則の学習および適用の際には，セグメント $SegEv_j$ 中の固有表現を各システムごとにまとめて，システム単位で正誤のクラスを参照する．そこで，式 (3) で定義される一つのセグメントの事象表現 $SegEv_j$ に対して，各システム i ごとにまとめた以下のクラス表現を設定し，正誤判別規則の学習および適用を行なう.

$$\begin{aligned} class^1_{sys} \quad &= \quad \begin{cases} +/-,\ \ldots,\ +/- \\ \text{“no output”} \end{cases} \\ &\cdots \\ class^n_{sys} \quad &= \quad \begin{cases} +/-,\ \ldots,\ +/- \\ \text{“no output”} \end{cases} \end{aligned} \tag{5}$$

ここで，一般に，一つのセグメント中で，各システムは一つも固有表現を出力しない場合もあれば，複数の固有表現を出力する場合もありえるので，各システム i のクラス $class^i_{sys}$ は上記のような表現になる[7].

4.3.3 複数システムの出力の正誤判別規則

次に，前節の事象のデータ構造を用いて，複数システムの出力の正誤判別を行なう規則について説明する．複数システムの出力の正誤判別を行なう規則は，式 (3) で定義されるセグメントの事象表現 $SegEv_j$ を一つの事象単位として，各システム i ごとに，式 (5) で示すクラス $class^i_{sys}$ を判別するという形式をとる．この正誤判別規則の学習の際には，式 (3) で定義されるセグメントの事象表現 $SegEv_j$ から，次節で説明する素性を抽出し，この素性を用いて各システム i ごとのクラス $class^i_{sys}$ を判別する規則を学習する (4.4 節)．この正誤判別規則の適用の際にも，事象表現 $SegEv_j$ から抽出される素性を用いて各システム i ごとにクラス $class^i_{sys}$ を判別する (4.5 節).

7 実際に，実験で用いた訓練コーパスから学習した正誤判別規則において，クラスの種類が最も多かったのは，システム数 $n=2$ の場合で，“+”，“++”，“+ + +”，“+ + ++”，“+ + −”，“+−”，“+ − −”，“+ − −−”，“−”，“−−”，“− − −”，“no output”，の 12 通りであった.

4.3.4 素性

式 (3) で定義されるセグメントの事象表現 $SegEv_j$から抽出される一つの素性 fは，システムの指標のリスト $\langle p, \ldots, q \rangle$，および，固有表現の素性表現 Fの組 $\langle systems = \langle p, \ldots, q \rangle,\ F \rangle$ の集合によって表現される．

$$f \quad = \quad \left\{ \langle systems = \langle p, \ldots, q \rangle,\ F \rangle,\ \cdots,\ \langle systems = \langle p', \ldots, q' \rangle,\ F' \rangle \right\} \tag{6}$$

このうち，一つの組 $\langle systems = \langle p, \ldots, q \rangle,\ F \rangle$ は，指標 $p, \ldots, q$に相当する (複数の) システムによって出力された一つの固有表現が，素性表現 Fを持つことを表している．固有表現の素性表現 Fは，集合 $\{mlength = \cdots, NEtag = \cdots, POS = \cdots\}$ の巾集合の任意の要素，あるいは，そのセグメント中で指標 $p, \ldots, q$に相当する (複数の) システムが固有表現を出力しなかったことを表す集合の形式，のいずれかで表現される．

$$F \quad = \quad \left\{ \begin{array}{l} \left\{ mlength = \cdots,\ NEtag = \cdots,\ POS = \cdots \right\} \\ \left\{ mlength = \cdots,\ NEtag = \cdots \right\} \\ \left\{ mlength = \cdots,\ POS = \cdots \right\} \\ \left\{ NEtag = \cdots,\ POS = \cdots \right\} \\ \left\{ mlength = \cdots \right\} \\ \left\{ NEtag = \cdots \right\} \\ \left\{ POS = \cdots \right\} \\ \emptyset \\ \left\{ class_{sys} = \text{“no outputs”} \right\} \end{array} \right.$$

正誤判別規則の学習時には，式 (3) で定義されるセグメントの事象表現 $SegEv_j$から，式 (6) の形式のあらゆる可能な素性 fのうち，以下の制約を含むいくつかの制約を満たすものだけが抽出される[8]．詳細については，次節の例を参照．

i) システムの指標のリスト $\langle p, \ldots, q \rangle$ については，その固有表現を出力した全てのシステムの指標を記すこととし，部分リストの形式は許さない．

ii) 一つのシステムが，一つのセグメント中で複数個の固有表現を出力した場合は，一つの素性 f中で，それらの複数の固有表現のうちの一部のものだけの情報を記述することは許さない．それらの全ての固有表現について何らかの情報を記述するか，どの固有表現についての情報も記述しないかのどちらかである．

8 実際に，実験で用いた訓練コーパスから学習した正誤判別規則においては，固有表現を構成する形態素数 “$mlength$” の値は 18 通り，固有表現のタイプ “$NEtag$” の値は 8 通り，固有表現を構成する形態素の品詞のリスト “POS” の値は 4926 通りであった．また，システム数 $n = 2$ の場合で，可能な素性 fの数の最大数は，112,114 であった．

4.3.5　例

4.3.1 節の手続きにしたがって，二つのシステムの固有表現抽出結果を整列し，その整列結果を事象表現に変換する例を表 4に示す．また，4.3.2 節および 4.3.4 節の手続きにしたがって，それらの事象表現からクラスおよび素性を抽出する例を表 5に示す．

表 4では，形態素解析の結果の形態素列に対して，システム 0 およびシステム 1 の二つのシステムがそれぞれ単独で出力した固有表現を,「単独システムの固有表現出力」の欄に示す．それらの単独システムの固有表現出力を整列した結果は，$SegEv_i \sim SegEv_{i+3}$の四つのセグメントに分割されており，これらのセグメントを事象表現に変換した結果が「事象表現」の欄に示されている．各セグメントの特徴を簡単にまとめると以下のようになる．

- $SegEv_i$：システム 0 が連続する二つの固有表現を出力したのに対して，システム 1 はそれらをまとめて一つの固有表現として出力している．正解データとの比較では，システム 1 の出力結果の方が正解である．このセグメントの事象表現は，いずれかの単独システムから出力された三つの固有表現の事象表現から構成されている．
- $SegEv_{i+1}$：システム 1 のみが固有表現を出力したが，この固有表現は誤出力である．このセグメントの事象表現は，システム 0 からの出力がなかったことを表す事象表現と，システム 1 が出力した一つの固有表現の事象表現から構成されている．
- $SegEv_{i+2}$：システム 0 が一形態素から構成される一つの固有表現を出力したのに対して，システム 1 はその形態素を含む三形態素から構成される一つの固有表現を出力した．正解データとの比較では，システム 1 の出力結果の方が正解である．このセグメントの事象表現は，各々の単独システムから出力された二つの固有表現の事象表現から構成されている．
- $SegEv_{i+3}$：システム 0，システム 1 ともに二形態素から構成される同一の固有表現を出力した．正解データとの比較では，この固有表現は正解である．このセグメントの事象表現は，この一つ固有表現の事象表現から構成されている．

次に，表 5においては，まずクラスについては，これらの各セグメントの事象表現において，各システムが出力した固有表現のクラス (もしくは出力がなかったことを表す事象表現のクラス) をシステムごとにまとめたものになっている．一方，素性の方は，各セグメントについて，以下の制約を満たす可能な素性の一覧を表現したものになっている．

- $SegEv_i$：システム 0 は，このセグメント中で二つの固有表現を出力しているが，この二つの固有表現のうちの一つだけの情報を記述した素性は許容しない．
- $SegEv_{i+1}$：ある単独システムからの出力がなかったことだけを記述した素性は許容しない．例えば，$\left\{ \langle systems = \langle 0 \rangle, class_{sys} = \text{“no outputs”} \rangle \right\}$ という素性は許容しない．
- $SegEv_{i+3}$：システムの指標のリストにおいては，このセグメントの固有表現を出力した二つのシステムの指標 0 および 1 の両方を必ず記述する．

表 4　複数システムの出力の混合のための事象表現の例

セグメント	形態素列 (品詞)	単独システムの固有表現出力		事象表現
		システム 0	システム 1	
	⋮			
$SegEv_i$	来年 (時相名詞) 10 月 (時相名詞)	来年 (DATE) 10 月 (DATE)	来年 10 月 (DATE)	$\{systems=\langle 0\rangle, mlength=1,$ $NEtag=$DATE, $POS=$時相名詞, $class_{NE}=-\}$ $\{systems=\langle 0\rangle, mlength=1,$ $NEtag=$DATE, $POS=$時相名詞, $class_{NE}=-\}$ $\{systems=\langle 1\rangle, mlength=2,$ $NEtag=$DATE, $POS=$時相名詞-時相名詞, $class_{NE}=+\}$
	⋮			
$SegEv_{i+1}$	生殖 (名詞) 医療 (名詞) 技術 (名詞)		生殖医療技術 (ARTIFACT)	$\{systems=\langle 0\rangle,$ $class_{sys}=$ “no outputs”$\}$ $\{systems=\langle 1\rangle, mlength=3,$ $NEtag=$ARTIFACT, $POS=$名詞-名詞-名詞, $class_{NE}=-\}$
	について (助詞相当) ⋮ 調査 (サ変名詞) は (提題助詞)			
$SegEv_{i+2}$	厚生省 (固有名詞) 研究 (サ変名詞) 班 (名詞)	厚生省 (ORGANI- ZATION)	厚生省研究班 (ORGANI- ZATION)	$\{systems=\langle 0\rangle, mlength=1,$ $NEtag=$ORGANIZATION, $POS=$固有名詞, $class_{NE}=-\}$ $\{systems=\langle 1\rangle, mlength=3,$ $NEtag=$ORGANIZATION, $POS=$固有名詞-サ変名詞-名詞, $class_{NE}=+\}$
	((記号) 主任 (人称名詞) 研究者 (人称名詞) 、(読点)			
$SegEv_{i+3}$	山田 (人名) 太郎 (人名)	山田太郎 (PERSON)	山田太郎 (PERSON)	$\{systems=\langle 0, 1\rangle, mlength=2,$ $NEtag=$PERSON, $POS=$人名-人名, $class_{NE}=+\}$
	⋮			

表 5　表 4の事象表現の例から抽出される素性およびクラス

事象表現	素性	クラス
$\{systems=\langle 0\rangle, mlength=1,$ $NEtag=$DATE, $POS=$時相名詞, $class_{NE}=-\}$ $\{systems=\langle 0\rangle, mlength=1,$ $NEtag=$DATE, $POS=$時相名詞, $class_{NE}=-\}$ $\{systems=\langle 1\rangle, mlength=2,$ $NEtag=$DATE, $POS=$時相名詞-時相名詞, $class_{NE}=+\}$	$\{\langle systems=\langle 0\rangle,\ F\rangle, \langle systems=\langle 0\rangle,\ F'\rangle,$ $\langle systems=\langle 1\rangle,\ F''\rangle\}$　または $\{\langle systems=\langle 0\rangle,\ F\rangle, \langle systems=\langle 0\rangle,\ F'\rangle\}$ または　$\{\langle systems=\langle 1\rangle,\ F''\rangle\}$ ただし F, F'は　$\{mlength=1, NEtag=$DATE, $POS=$時相名詞$\}$ の巾集合の任意の要素 F''は　$\{mlength=2, NEtag=$DATE, $POS=$時相名詞-時相名詞$\}$ の巾集合の任意の要素	$class^0_{sys}=--$ $class^1_{sys}=+$
$\{systems=\langle 0\rangle,$ $class_{sys}=$“no outputs”$\}$ $\{systems=\langle 1\rangle, mlength=3,$ $NEtag=$ARTIFACT, $POS=$名詞-名詞-名詞, $class_{NE}=-\}$	$\{\langle systems=\langle 0\rangle, class_{sys}=$“no outputs”$\rangle,$ $\langle systems=\langle 1\rangle,\ F\rangle\}$ または　$\{\langle systems=\langle 1\rangle,\ F\rangle\}$ ただし Fは　$\{mlength=3, NEtag=$ARTIFACT, $POS=$名詞-名詞-名詞$\}$ の巾集合の任意の要素	$class^0_{sys}=$ “no outputs” $class^1_{sys}=-$
$\{systems=\langle 0\rangle, mlength=1,$ $NEtag=$ORGANIZATION, $POS=$固有名詞, $class_{NE}=-\}$ $\{systems=\langle 1\rangle, mlength=3,$ $NEtag=$ORGANIZATION, $POS=$固有名詞-サ変名詞-名詞, $class_{NE}=+\}$	$\{\langle systems=\langle 0\rangle,\ F\rangle, \langle systems=\langle 1\rangle,\ F'\rangle\}$ または　$\{\langle systems=\langle 0\rangle,\ F\rangle\}$ または　$\{\langle systems=\langle 1\rangle,\ F'\rangle\}$ ただし Fは　$\{mlength=1,$ $NEtag=$ORGANIZATION, $POS=$固有名詞$\}$ の巾集合の任意の要素 F'は　$\{mlength=3,$ $NEtag=$ORGANIZATION, $POS=$固有名詞-サ変名詞-名詞$\}$ の巾集合の任意の要素	$class^0_{sys}=-$ $class^1_{sys}=+$
$\{systems=\langle 0, 1\rangle, mlength=2,$ $NEtag=$PERSON, $POS=$人名-人名, $class_{NE}=+\}$	$\{\langle systems=\langle 0, 1\rangle,\ F\rangle\}$ ただし Fは　$\{mlength=2, NEtag=$PERSON, $POS=$人名-人名$\}$ の巾集合の任意の要素	$class^0_{sys}=+$ $class^1_{sys}=+$

4.4　学習アルゴリズム

教師あり学習法としては，決定リスト学習を用いる[9]．決定リスト (Rivest 1987; Yarowsky 1994) は，ある素性のもとでクラスを決定するという規則を優先度の高い順にリスト形式で並べたもので，適用時には優先度の高い規則から順に適用を試みていく．本論文では，各規則の優先度として，素性 fの条件のもとでの，システム i のクラス $class^i_{sys}$の条件付確率 $P(class^i_{sys}=c_i \mid f)$ を用い，この条件付確率順に決定リストを構成する．ただし，決定リストを構成する際には，素性 fの条件のもとでの，システム i のクラス $class^i_{sys}$の頻度 $freq(f, class^i_{sys})$ に下限 L_fを設け，

$$freq(f, class^i_{sys}) \quad \geq \quad L_f \tag{7}$$

の条件を満たす規則だけを用いて決定リストを構築する．頻度の下限 L_fは，各規則の条件付確率 $P(class^i_{sys}=c_i \mid f)$ を推定する際に使用したデータセット以外のデータセットに対して，正誤判別規則の性能を最大にする値を用いる．

4.5　正誤判別規則の適用による複数システム出力の混合

学習された正誤判別規則を適用することにより複数システムの出力の混合を行なう場合は，式 (3) と同じ形式のセグメントの事象表現

$$SegEv_j \quad = \quad \{NEev_1, \ldots, NEev_{l_j}\}$$

に対して，決定リストの形式の正誤判別規則が参照され，素性 fの条件のもとでの，システム i のクラス $class^i_{sys}$の条件付確率 $P(class^i_{sys}=c_i \mid f)$ の推定値を得る．そして，

(1)　複数のシステムによって出力された単一の固有表現は，同一の正誤クラスを持つ．

(2)　少なくとも一つの形態素を共有する複数の固有表現が，正のクラス (“+”) を持ってはならない．

という二つの制約のもとで，全システムについての条件付確率 $P(class^i_{sys}=c_i \mid f)$ の積を最大化するクラス割当ての組合わせが求められ，これが，セグメント中で各システム i $(i=1,\ldots,n)$ が出力した固有表現への正誤クラスの判別結果$\hat{class}^1_{sys}, \ldots, \hat{class}^n_{sys}$となる[10]．

$$\hat{class}^1_{sys}, \ldots, \hat{class}^n_{sys} \quad = \quad \underset{c_i, f_i}{\operatorname{argmax}} \prod_{i=1}^{n} P(class^i_{sys}=c_i \mid f_i)$$

9 5.3 節では，最大エントロピー法を用いて正誤判別規則学習を行なった結果との比較を行なっている．本論文では，実装が容易，学習が高速で，かつ，一定の性能を達成できるという理由で決定リスト学習を適用したが，より高性能な他の様々な教師あり学習法を適用することも十分可能である．

10 システム i について，決定リスト中に照合する判別規則が存在しない場合には，そのシステムが出力した固有表現を誤出力 (“–”) とみなしている．

5　実験および評価

本節では，IREX ワークショップの固有表現抽出タスクの訓練データおよび試験データを用いて，複数の固有表現抽出結果の混合法の実験的評価を行なった結果について述べる．以下では，訓練データとして用いている CRL 固有表現データの一般ドメインのものを D_{CRL}，評価データとして用いている本試験データのうちの一般ドメインのものを D_{formal}と記す．ただし，いずれも，表 2の「その他」のものは除いている．

5.1　各モデル単独の出力の比較

本節では，3.5 節で述べた各モデル単独の性能について述べ，各モデルの出力を比較する．実験に用いたモデルは，3.5.1 節の固定長モデルとしては，5 グラムモデル，7 グラムモデル，9 グラムモデル，および，3.5.2 節の可変長モデルである．また，7 グラムモデル，9 グラムモデル，および，可変長モデルについては，3.5.3 節の三種類の素性の設定も区別して実験を行なった．

まず，表 6に，個々の固有表現抽出モデルを学習するための訓練データセット TrIを D_{CRL} とした場合の，本試験データ D_{formal}に対する各モデルの F 値 ($\beta = 1$) を示す．この結果からわかるように，単独のモデルでは，5 グラムモデルが最も高い性能を示す．また，7 グラムモデルおよび 9 グラムモデルは，素性の設定に関わらず，ほぼ同等の性能を示している．

次に，最も性能のよい 5 グラムモデルの出力と，他のモデルの出力との違いを調べるために，5 グラムモデル以外の各モデルの出力について，5 グラムモデルの出力との和集合を求め，本試験データ D_{formal}の正解データに対する再現率を算出した．また，5 グラムモデル以外の各モデルの誤出力と 5 グラムモデルの誤出力の間の重複率

$$\text{誤出力の重複率} \quad = \quad \frac{\text{二つのモデルの誤出力間で重複する固有表現数}}{\text{5 グラムモデルの誤出力の固有表現数}}$$

を求めた．これらの結果を表 7に示す．特に，和の再現率が最も高く，誤出力の重複率が最も低い結果 (この場合は，可変長モデル (形態素 $M_{l(\leq -3)}$，$M_{r(\geq 3)}$の素性=全て) との差分) を**太字**で示す．

表 6および表 7の結果から分かるように，7 グラムモデルおよび 9 グラムモデルは，5 グラムモデルと比べて出力の和集合の再現率が低く，かつ誤出力の重複率も高いことから，相対的に 5 グラムモデルと似通ったモデルであると言える．一方，可変長モデルは，7 グラムモデルおよび 9 グラムモデルと比べて，相対的に 5 グラムモデルとの類似性が小さいことがわかる．特に，誤出力の重複率が比較的小さい点が目立つ．

表 6　本試験データ D_{formal}に対する各モデル単独の性能 (F 値 ($\beta = 1$) (再現率/適合率) (%))

	形態素 $M_{l(\leq -3)}$，$M_{r(\geq 3)}$の素性		
	全て	語彙+品詞	語彙
7 グラムモデル	80.78 (78.44/83.27)	80.81 (78.44/83.33)	80.71 (78.51/83.03)
9 グラムモデル	80.13 (77.87/82.54)	80.53 (78.22/82.98)	80.53 (78.37/82.82)
可変長モデル	45.12 (51.50/40.15)	77.02 (75.86/78.21)	75.16 (73.78/76.58)
5 グラムモデル	**81.16 (78.87/83.60)**		

表 7　5 グラムモデルの出力と各モデルの出力との差分 (和の再現率/誤出力の重複率) (%)

	形態素 $M_{l(\leq -3)}$，$M_{r(\geq 3)}$の素性		
	全て	語彙+品詞	語彙
7 グラムモデル	79.8/85.2	79.8/85.2	79.7/91.2
9 グラムモデル	79.7/84.7	79.7/86.1	79.5/90.7
可変長モデル	**82.6/27.3**	81.4/63.4	80.4/72.7

5.2　複数システムの出力の混合の性能評価

5.2.1　評価方法

次に，7 グラムモデル，9 グラムモデル，可変長モデルについて，それぞれ，3.5.3 節の三種類の素性の設定を区別して，合計 9 種類のモデルを考え，その各々について，5 グラムモデルの出力との間で混合を行ない，その性能を評価した．ただし，個々の固有表現抽出モデルを学習するための訓練データセット TrI，複数システムの出力の正誤判別規則を学習するための訓練データセット TrC，4.4 節の (7) 式の頻度閾値 L_fの設定の組合わせとしては，以下の二通りについて評価を行なった．なお，複数システムの出力の正誤判別規則を評価するための評価データセット Ts については，いずれも，本試験データ D_{formal}を用いた．

(a)　TrI:　D_{CRL}から 200 記事 D^{200}_{CRL}を除いた残り $D_{CRL} - D^{200}_{CRL}$
　　TrC:　D_{CRL}中の 200 記事 D^{200}_{CRL}
　　L_f:　$D_{CRL} - D^{200}_{CRL}$中の 200 記事に対して，正誤判別規則の性能を最大にする値

(b)　$TrI = TrC = D_{CRL}$
　　L_f:　(a) と同じ値

このうち，設定 (a) は，二つの訓練データセット TrIと TrCについて，重複のないデータセットを用いたものに相当する．ただし，利用可能なデータ量に限界があることから，混合のための正誤判別規則学習の訓練データセット TrCのサイズが小さくなっている．一方，設定 (b) の方は，個々の固有表現抽出モデルを訓練データ TrI自身に適用したインサイド適用の結果を利用した混合となるが，混合のための正誤判別規則学習の訓練データセット TrCのサイズは設定

表 8　5 グラムモデルの出力と各モデルの出力の混合結果の性能 (F 値 ($\beta = 1$) (再現率/適合率) (%))

(a)　$TrI = D_{CRL} - D^{200}_{CRL}$，$TrC = D^{200}_{CRL}$ (D_{CRL}中の 200 記事)

	形態素 $M_{l(\leq -3)}$，$M_{r(\geq 3)}$の素性		
	全て	語彙+品詞	語彙
7 グラムモデル	81.54 (78.15/85.23)	81.53 (77.79/85.65)	80.60 (77.08/84.46)
9 グラムモデル	81.31 (77.58/85.41)	81.26 (77.51/85.40)	80.60 (77.08/84.46)
可変長モデル	**83.43 (80.23/86.89)**	81.55 (76.29/87.58)	81.85 (78.51/85.49)

(b)　$TrI = TrC = D_{CRL}$

	形態素 $M_{l(\leq -3)}$，$M_{r(\geq 3)}$の素性		
	全て	語彙+品詞	語彙
7 グラムモデル	81.97 (78.51/85.76)	81.83 (78.22/85.78)	81.58 (78.51/84.90)
9 グラムモデル	81.53 (77.79/85.65)	81.66 (78.15/85.50)	81.52 (78.51/84.76)
可変長モデル	**84.07 (81.45/86.86)**	83.07 (79.94/86.44)	82.50 (79.87/85.31)

(a) よりもずっと大きい[11].

5.2.2　評価結果

評価結果を表 8に示す．この結果から分かるように，設定 (a) と (b) を比べると，一律に，設定 (b) の方が高い性能が得られている．このことから，正誤判別規則の学習において，たとえ，インサイド適用の結果しか利用できなかったとしても，混合のための正誤判別規則学習の訓練データセット TrCのサイズはできるだけ大きい方がよいことがわかる．特に，設定 (b) においては，どの混合結果においても 5 グラムモデル単独の性能を上回っていることから，混合規則学習のための十分な訓練データがあれば，混合により多少なりとも個々のモデルの出力の性能を向上できることが予想される．

また，設定 (b) の場合，7 グラムモデル，9 グラムモデルといった固定長モデルの出力と 5 グラムモデルの出力を混合した場合よりも，可変長モデルの出力と 5 グラムモデルの出力を混合した場合の方が圧倒的に高い性能向上を達成している．この結果は，表 7の差分の傾向と合致しており，5 グラムモデルとの類似性が相対的に小さい可変長モデルの出力との混合において，より高い性能向上が得られている．また，可変長モデル同士の間で，形態素 $M_{l(\leq -3)}$，$M_{r(\geq 3)}$ の素性の設定が異なる場合を比較しても，この傾向が成り立っており，5 グラムモデルとの類似性が小さいほど混合結果における性能向上は大きい．これらの結果から，出力の和の再現率が高く，誤出力の重複率が小さくなるような，なるべく類似性の小さい複数の日本語固有表現抽

11 ここで，厳密に 4.2 節の評価手続きに従うと，評価手順 (1) において，評価データセット Ts に対する固有表現抽出結果のリスト $NEList_i(Ts)$ $(i = 1, 2)$ を得る場合には，訓練の段階で用いた個々の固有表現抽出モデル $NEext_i$ $(i = 1, 2)$ と同じものを用いる必要がある．しかし，本論文では，設定 (a) と (b) の間で，混合を行なう前の固有表現抽出結果のリスト $NEList_i(Ts)$ $(i = 1, 2)$ を統一して，同一の条件で評価を行なうことを優先した．そのため，設定 (a) において用いる固有表現抽出結果のリスト $NEList_i(Ts)$ $(i = 1, 2)$ としては，設定 (b) と同じく，D_{CRL}の全体を用いて学習された各固有表現抽出モデルを適用して得られたものを用いた．訓練データが D_{CRL}であるか $D_{CRL} - D^{200}_{CRL}$であるかの違いによる固有表現抽出モデルの性能の差はそれほど大きくないので，このことによる影響は小さいと考えられる．

出モデルの出力を用意して，本論文の手法により出力の混合を行なえば，単独のモデルの出力の性能向上が期待できることがわかる．

5.2.3　固有表現の形態素長/種類ごとの分析

次に，5 グラムモデルの出力と可変長モデルの出力の混合の場合について，固有表現を構成する形態素数ごと，および，固有表現の種類ごとに，単独モデルの出力および混合結果の性能 (F 値，再現率，適合率) を列挙したものを，それぞれ，表 9，および，表 10に示す．なお，表中で，固有表現を構成する形態素数ごと，あるいは，固有表現の種類ごとに，最も高い F 値を達成した結果をそれぞれ**太字**で示す．

表 9から分かるように，どの可変長モデルの出力との混合においても，ほぼ全ての形態素長の固有表現において，5 グラムモデル単独の出力の再現率・適合率をともに上回っている．特に，最高の性能を示している「5 グラムモデル+可変長モデル (全て)」の結果においては，5 グラムモデルからの性能向上の度合は，形態素長が長くなるほど大きいことから，可変長モデルでしか出力されなかった長い固有表現を，混合によってうまく抽出できていることがわかる．

また，表 10からは，どの可変長モデルの出力との混合においても，ほぼ全ての種類の固有表現において，5 グラムモデルの出力の再現率・適合率とほぼ同等かそれ以上の性能が得られている．そのうち，TIME，MONEY，PERCENT の三種類については，他の種類と比較して，訓練データ・評価データともその頻度が小さく，また，5 グラムモデルにおける性能もかなり高いことから，改善の余地があまりなかったと考えられる．ただし，その場合でも，混合結果においては，可変長モデルの低い性能の悪影響を受けることなく，5 グラムモデルの高い性能が反映されている．

5.2.4　単独モデル・混合結果の出力のパターンの分析

5 グラムモデルの出力と可変長モデルの出力の混合の場合について，各単独モデルの出力における固有表現の有無，および，混合結果における固有表現の有無と，正解データにおける固有表現の有無のパターンの割合を調査した結果を表 11に示す．表中で,「有」「無」は，それぞれ，単独モデルの出力，混合結果，正解データに固有表現が存在する場合，および，存在しない場合を表す．例えば,「有」「有」「有」「有」のパターンは，両方の単独モデルの出力にその固有表現が存在し，混合結果においてもその固有表現が出力され，かつ，それが正解データにも存在する正解の固有表現である場合に相当する．また，割合 (%) の計算においては，両方の単独モデルの出力の和における固有表現数を分母，それぞれのパターンに該当する固有表現数を分子として，割合 (%) を計算している．さらに，混合における正誤判別結果が正解であるか否かについては，混合結果および正解データにおける出力の有無が一致する場合は正誤判別が正解，一致しない場合は正誤判別が誤りであるので,「正誤判別率」の欄にそれぞれの率を示した．

表 9　混合結果の性能: 固有表現の形態素長ごと，$TrI = TrC = D_{CRL}$
(F 値 ($\beta = 1$) (再現率) (適合率) (%))

	n 形態素 対 一固有表現				
	$n \geq 1$	$n = 1$	$n = 2$	$n = 3$	$n \geq 4$
5 グラムモデル	81.16 (78.87) (83.60)	83.60 (84.97) (82.28)	86.94 (85.90) (88.00)	68.42 (63.64) (73.98)	50.59 (35.83) (86.00)
可変長モデル (全て)	45.12 (51.50) (40.15)	53.77 (38.69) (88.14)	56.63 (71.37) (47.93)	33.74 (57.34) (23.91)	16.78 (40.00) (10.62)
可変長モデル (語彙+品詞)	77.02 (75.86) (78.21)	81.86 (78.57) (85.44)	79.96 (84.82) (75.63)	63.19 (63.64) (62.76)	50.52 (40.83) (66.22)
可変長モデル (語彙)	75.16 (73.78) (76.58)	79.11 (87.05) (72.49)	83.02 (81.13) (85.00)	50.46 (38.46) (73.33)	22.38 (13.33) (69.57)
5 グラムモデル + 可変長モデル (全て)	**84.07** **(81.45)** **(86.86)**	85.06 (85.12) (84.99)	**88.96** **(87.42)** **(90.56)**	**75.19** **(69.93)** **(81.30)**	**65.96** **(51.67)** **(91.18)**
5 グラムモデル + 可変長モデル (語彙+品詞)	83.07 (79.94) (86.44)	84.97 (84.52) (85.41)	87.29 (85.68) (88.96)	72.80 (66.43) (80.51)	63.04 (48.33) (90.63)
5 グラムモデル + 可変長モデル (語彙)	82.50 (79.87) (85.31)	**85.11** **(86.76)** **(83.52)**	87.73 (86.12) (89.41)	71.04 (64.34) (79.31)	50.89 (35.83) (87.76)

形態素 $M_{l(\leq -3)}$，$M_{r(\geq 3)}$ の素性の設定が異なる場合についてこの結果を比較すると,「5 グラムモデル+可変長モデル (全て)」において判別正解率が高くなっているが，これは,「可変長モデル (全て)」の性能が極端に悪く,「可変長モデル (全て)」のみが出力した固有表現の多くが誤りであり，その判別が比較的容易であったからである．全体では，どの可変長モデルの出力との混合においても，5 グラムモデルの出力を覆すことで正解となった場合 (「無」「有」「有」「有」および「有」「無」「無」「無」) が数%あり，これが，5 グラムモデルからの性能向上に寄与している．その一方で，判別誤りの内訳をみると，その多くは，誤出力の検出が十分できなかっ

表 10 混合結果の性能: 固有表現の種類ごと，$TrI = TrC = D_{CRL}$ (F 値 ($\beta = 1$) (再現率) (適合率) (%))

	ORGANI-ZATION	PER-SON	LOCA-TION	ARTI-FACT	DATE	TIME	MONEY	PER-CENT
5 グラムモデル	67.74 (58.45) (80.53)	81.82 (79.88) (83.85)	77.04 (71.91) (82.96)	30.43 (29.17) (31.82)	91.49 (88.85) (94.29)	**93.20** **(88.89)** **(97.96)**	**92.86** **(86.67)** **(100.00)**	87.18 (80.95) (94.44)
可変長モデル (全て)	35.48 (37.40) (33.75)	48.45 (48.52) (48.38)	38.47 (32.93) (46.26)	5.80 (22.92) (3.32)	78.60 (81.92) (75.53)	56.90 (61.11) (53.23)	60.61 (66.67) (55.56)	87.18 (80.95) (94.44)
可変長モデル (語彙+品詞)	65.30 (57.34) (75.82)	78.56 (77.51) (79.64)	72.46 (66.59) (79.48)	26.92 (29.17) (25.00)	88.51 (88.85) (88.17)	77.36 (75.93) (78.85)	80.00 (80.00) (80.00)	**89.47** **(80.95)** **(100.00)**
可変長モデル (語彙)	63.96 (54.57) (77.25)	76.81 (78.40) (75.28)	72.29 (68.52) (76.49)	25.00 (20.83) (31.25)	86.96 (84.62) (89.43)	54.21 (53.70) (54.72)	73.33 (73.33) (73.33)	81.08 (71.43) (93.75)
5 グラムモデル + 可変長モデル (全て)	**72.18** **(62.88)** **(84.70)**	84.15 (81.66) (86.79)	**79.58** **(73.61)** **(86.61)**	**38.71** **(37.50)** **(40.00)**	**92.86** **(90.00)** **(95.90)**	**93.20** **(88.89)** **(97.96)**	**92.86** **(86.67)** **(100.00)**	87.18 (80.95) (94.44)
5 グラムモデル + 可変長モデル (語彙+品詞)	70.19 (60.66) (83.27)	83.41 (81.07) (85.89)	78.22 (72.15) (85.39)	35.29 (31.25) (40.54)	92.64 (89.62) (95.88)	92.16 (87.04) (97.92)	**92.86** **(86.67)** **(100.00)**	87.18 (80.95) (94.44)
5 グラムモデル + 可変長モデル (語彙)	68.82 (59.28) (81.99)	**84.46** **(82.84)** **(86.15)**	77.50 (72.15) (83.71)	31.46 (29.17) (34.15)	91.85 (88.85) (95.06)	**93.20** **(88.89)** **(97.96)**	**92.86** **(86.67)** **(100.00)**	**89.47** **(80.95)** **(100.00)**

表 11 単独モデル・混合結果の出力のパターンの分析結果

5 グラムモデルと可変長モデル (形態素 $M_{l(\leq -3)}$，$M_{r(\geq 3)}$ の素性=全て) の出力の混合														
単独モデルの出力の有無	5 グラムモデル	有	有	無	有	有	無	有	有	無	有	有	無	
	可変長モデル	有	無	有	有	無	有	有	無	有	有	無	有	
混合結果の出力の有無		有			無			有			無			
正解データにおける有無		有			無			無			有			
割合 (%)		28.0	18.2	1.5	0.04	1.8	42.5	2.4	4.8	0	0	0	0.7	
正誤判別率 (%) (判別数/出力数)		(判別正解率) 92.1 (2194/2382)						(判別誤り率) 7.9 (188/2382)						

5 グラムモデルと可変長モデル (形態素 $M_{l(\leq -3)}$，$M_{r(\geq 3)}$ の素性=語彙+品詞) の出力の混合														
単独モデルの出力の有無	5 グラムモデル	有	有	無	有	有	無	有	有	無	有	有	無	
	可変長モデル	有	無	有	有	無	有	有	無	有	有	無	有	
混合結果の出力の有無		有			無			有			無			
正解データにおける有無		有			無			無			有			
割合 (%)		67.8	4.4	1.7	0.2	2.6	10.3	8.9	2.6	0.1	0	0.7	0.7	
正誤判別率 (%) (判別数/出力数)		(判別正解率) 87.1 (1315/1510)						(判別誤り率) 12.9 (195/1510)						

5 グラムモデルと可変長モデル (形態素 $M_{l(\leq -3)}$，$M_{r(\geq 3)}$ の素性=語彙) の出力の混合														
単独モデルの出力の有無	5 グラムモデル	有	有	無	有	有	無	有	有	無	有	有	無	
	可変長モデル	有	無	有	有	無	有	有	無	有	有	無	有	
混合結果の出力の有無		有			無			有			無			
正解データにおける有無		有			無			無			有			
割合 (%)		67.3	6.1	1.1	0.1	1.5	10.6	10.4	2.5	0	0	0.1	0.4	
正誤判別率 (%) (判別数/出力数)		(判別正解率) 86.6 (1297/1497)						(判別誤り率) 13.4 (200/1497)						

た場合で，ほとんどの場合，少なくとも 5 グラムモデルはその誤りの固有表現を出力している．このことから，より効果的な素性を用いる，あるいは，より高性能な学習器を用いるなどして，誤出力検出の精度を向上させることにより，適合率を向上できる余地があることがわかる．

5.3 最大エントロピー法による正誤判別規則学習

最後に，正誤判別規則学習の学習法の比較のために，最大エントロピー法を用いて正誤判別規則学習を行なった．

まず，最大エントロピー法を適用するために，4.3.1 節の (3) 式の事象表現 $SegEv_j$ を，以下のように変換する．

$$SegEv_j \quad = \quad \{NEListev_{p,\ldots,q}, \ldots, NEListev_{p',\ldots,q'}\} \tag{8}$$

ここで，各事象表現 $NEListev_{p,\ldots,q}$ は，システムの指標のリストごとに固有表現をまとめたもので，固有表現のリストの事象表現に相当する[12]．4.3.1 節の場合と同様に，以下の二種類のどちらかに対応し，それぞれ異なったデータ構造を持つ．

i) そのセグメント中で少なくとも一つのシステムにより出力された固有表現のリストの事象表現．

ii) そのセグメント中で一つも固有表現を出力しなかった一つのシステムに関する情報を表す事象表現．

i) のタイプの事象表現 $NEListev_{p,\ldots,q}$ は以下のようなデータ構造を持つ．

$$\begin{aligned} NEListev_{p,\ldots,q} \quad = \quad \Big\{ & systems = \langle p, \ldots, q \rangle,\ mlengthList = y, \ldots, z \text{ morphemes}, \\ & NEtagList = \cdots,\ POSList = \cdots, \\ & classList_{NE} = +/-, \ldots, +/- \Big\} \end{aligned} \tag{9}$$

このデータ構造は，4.3.1 節の (4) 式のデータ構造とほぼ同じであるが，固有表現のリストを表現するために，各素性に相当する情報が全てリスト表現になっている点が異なる．一方，ii) のタイプの事象表現 $NEList_r$ は，4.3.1 節の (4) 式と同じく，以下のデータ構造で表現される．

$$NEListev_r \quad = \quad \Big\{ systems = \langle r \rangle,\ class_{sys} = \text{“no output”} \Big\} \tag{10}$$

このような事象表現を用いて正誤判別規則の学習および適用を行なう際には，上述の (8) 式の事象表現を事象の単位とし，4.3.2 節の場合と同様に，各システム i ごとにまとめた以下のクラス表現を設定し，各システム i ごとにクラスの判別を行なうための正誤判別規則の学習およ

12 最大エントロピー法の適用における事象表現の形式は，4.3.1 節の決定リスト学習の場合の事象表現の形式とは異なっているが，最大エントロピー法における素性の表現能力を必要以上に制限しているわけではない．決定リスト学習において可能な素性を表現する際にも，4.3.4 節の i) および ii) の二つの制約を課しているため，素性の表現能力について両者の間に意図的な差はない．

表 12　5 グラムモデル/その他の各モデルの出力の最大エントロピー法による混合結果の性能 (F 値 ($\beta = 1$) (再現率/適合率) (%))

(a)　$TrI = TrC = D_{CRL}$，結合素性なし			
	形態素 $M_{l(\leq -3)}$，$M_{r(\geq 3)}$ の素性		
	全て	語彙+品詞	語彙
7 グラムモデル	81.81 (78.80/85.07)	81.70 (78.51/85.16)	81.47 (78.58/84.58)
9 グラムモデル	81.21 (78.01/84.68)	81.38 (78.30/84.73)	81.46 (78.51/84.63)
可変長モデル	81.12 (76.65/86.15)	81.48 (77.36/86.06)	81.37 (78.37/84.61)

(b)　$TrI = TrC = D_{CRL}$，結合素性あり			
	形態素 $M_{l(\leq -3)}$，$M_{r(\geq 3)}$ の素性		
	全て	語彙+品詞	語彙
7 グラムモデル	81.71 (78.72/84.93)	81.58 (78.37/85.07)	81.35 (78.44/84.49)
9 グラムモデル	81.16 (78.08/84.50)	81.22 (78.37/84.28)	81.29 (78.58/84.19)
可変長モデル	80.94 (76.65/85.74)	81.40 (77.29/85.98)	81.24 (78.01/84.75)

び適用を行なう．

$$
\begin{aligned}
class^1_{sys} &= \begin{cases} +/-,\ \ldots,\ +/- \\ \text{“no output”} \end{cases} \\
&\cdots \\
class^n_{sys} &= \begin{cases} +/-,\ \ldots,\ +/- \\ \text{“no output”} \end{cases}
\end{aligned}
$$

その際には，(9) 式の固有表現のリストの事象表現 $NEListev_{p,\ldots,q}$ の $mlengthList$，$NEtagList$，$POSList$，および，(10) 式の固有表現の事象表現 $NEList_r$ の $class_{sys}$ を，それぞれ文脈 x とし，上式の，各システムごとにまとめた正誤のクラスのリストを付与するための条件付確率モデルを，最大エントロピーモデルとして学習する．この最大エントロピーモデルは，各システム i ごとに個別にモデルの学習・適用を行なう．

このような方法で，7 グラムモデル，9 グラムモデル，可変長モデルについて，それぞれ，3.5.3 節の三種類の素性の設定を区別して，合計 9 種類のモデルを考え，その各々について，5 グラムモデルの出力との間で混合を行ない，その性能を評価した．ただし，$TrI = TrC = D_{CRL}$ とし，評価データセット Ts は本試験データ D_{formal} とした．最大エントロピーモデルの素性関数の頻度に下限を設け，評価データセット Ts に対して最も高い性能が得られた場合の結果を表 12(a) に示す．また，決定リスト学習との間で条件を揃えるために，4.3.4 節の (6) 式の形式の決定リスト学習の素性のうち，上述の実験結果 (a) では用いていなかった結合素性を追加して最大エントロピーモデルの学習および適用を行なった結果を表 12(b) に示す．この場合は，決定リスト学習における各規則の条件付確率 $P(class^i_{sys} = c_i \mid f)$ に下限を設け，評価データセット Ts に対して最も高い性能が得られた場合の結果を示している．

表 12の (a) と (b) の結果を比較すると，結合素性を用いた場合の方が性能が悪くなっている．また，いくかの結果を除いて，5 グラムモデルの性能からの向上はみられるものの，決定リ

スト学習による可変長モデルの出力との混合の場合のような高い性能向上は達成できていない．

この理由の一つとしては，最大エントロピーモデルと決定リスト学習の間のモデルの形式の違いの影響が挙げられる．最大エントロピーモデルは，あらゆる素性とクラスとの相関をそれぞれ別個のパラメータとし，モデル内では全パラメータを考慮する形式のモデルになっている．一方，決定リスト学習は，各々のクラス決定において最も寄与する素性の組合わせのみを考慮し，他の素性は全く考慮しない．したがって，素性間で寄与する度合の差がわずかしかない場合でも，決定リスト学習では，最も寄与する素性の組合わせのみが考慮されるのに対して，最大エントロピーモデルでは，全素性の寄与を総合的に考慮する．本論文の正誤判別規則学習による混合の問題では，素性の種類が比較的少なく，特に高頻度な素性[13] は，実際にクラス判別に寄与する度合に関係なく，どの事象においても常に一定の値以上の重みを持つと考えられる．そのような問題の場合には，最大エントロピーモデルのように全素性の寄与を総合的に考慮する学習法でなく，決定リスト学習のように各々のクラス決定に最も寄与する素性の組合わせのみを考慮する学習法が適していると考えられる．

逆に，正誤判別規則学習による混合の前段階である，形態素への固有表現まとめ上げ状態付与の問題の場合には，(颯々野, 宇津呂 2000; Sassano and Utsuro 2000) に示されるように，決定リスト学習よりも最大エントロピーモデルの方が高い性能を示している．この問題の場合には，素性の種類が比較的多く，極端に高頻度な素性も少ないことから，最大エントロピーモデルのように全素性の寄与を総合的に考慮する学習法が適していると考えられる．

6　関連研究

6.1　複数モデルの出力の混合法

1 節で述べたように，一般に，複数のモデル・システムの出力を混合する過程は，大きく以下の二つの部分に分けて考えることができる．

(1)　できるだけ振る舞いの異なる複数のモデル・システムを用意する．

(2)　用意された複数のモデル・システムの出力を混合する方式を選択・設計し，必要であれば学習等を行ない，与えられた現象に対して，用意された複数のモデル・システムの出力を混合することを実現する．

ここで，これまで自然言語処理の問題に適用された混合手法においては，これらの (1) および (2) の過程について，大体以下のような手法が用いられていた．

まず，(1) については，大きく分けて以下のような手法がある．

i)　学習モデルが異なる複数のシステム等 (原理的には，人手による規則に基づくシステムとデータからの学習に基づくシステム，などの組合わせも可能)，ある程度振る舞

13 例えば，複数の情報の結合でなく単独の情報のみから構成される素性など．

いの異なる既存のシステムを用意する (van Halteren et al. 1998; Brill and Wu 1998; Henderson and Brill 1999; 乾, 乾 2000; Sang 2000).

ii) i) と似ているが，学習モデルは単一のものを用い，データの表現法 (具体的には，まとめ上げ問題におけるまとめ上げ状態の表現法) として複数のものを設定することにより，複数の出力を得る (Sang 2000; 工藤, 松本 2000).

iii) 単一の学習モデルを用いるが，訓練データのサンプリングを複数回行なうことにより複数のモデルを学習する bagging 法 (Breiman 1996a) を用いる (Henderson and Brill 2000)，あるいは，単一の学習モデルを用い，誤り駆動型で訓練データ中の訓練事例の重みを操作しながら学習と適用を繰り返すことにより，各サイクルの誤りに特化した複数のモデル (およびそれらの重み) を学習する boosting 法 (Freund and Schapire 1999) を用いる (Haruno and Matsumoto 1997; Haruno, Shirai and Oyama 1999; Abney et al. 1999; Henderson and Brill 2000).

これに対して，本論文においては，振る舞いの異なる複数のモデルを得る方法として，学習モデルは単一のものを用い，固有表現まとめ上げの際に考慮する周囲の形態素の個数を区別することで複数のモデルを得るという方法をとった．この方法は，上記のうちでは，ii) でとられた方法と比較的似ている．

次に，(2) については，大きく分けて以下のような手法がある[14].

i) 重み付多数決など，何らかの多数決を行なうもの (Breiman 1996a; van Halteren et al. 1998; Brill and Wu 1998; Henderson and Brill 1999; 乾, 乾 2000; Sang 2000; Henderson and Brill 2000; 工藤, 松本 2000).

ii) 複数のシステム・モデルの重みに応じて採用するシステムの切り替えを行なうもの (Henderson and Brill 1999; 乾, 乾 2000).

iii) 原理的に，上記の i) および ii) を包含し得る方法として，複数のシステム・モデルの出力 (および訓練データそのもの) を入力とする第二段の学習器を用いて，複数のシステム・モデルの出力の混合を行なう stacking 法 (Wolpert 1992)，あるいは，それと同等の方法に基づくもの (van Halteren et al. 1998; Brill and Wu 1998; Sang 2000).

これらの方法のうち，本論文では，原理的に，i) および ii) を包含し得る iii) の stacking 法を用いている．特に，本論文では，個々のシステムの出力する重みの情報は利用せず stacking を行なっているので，規則に基づくシステムなどで重みを出力しない場合でも，そのまま本論文の手法を適用することができる．これに対して，重み付多数決や重みを用いたシステム切り替えの場合は，システム数が少なく (例えば，二種類のシステムの混合の場合)，かつ，個々のシステムが重みを出力しない場合などでは，適用が困難になると考えられる．また，通常の bagging 法や boosting 法を適用する場合でも，第一段としては何らかの学習モデルを採用する必要があ

14 boosting は，複数のモデルを組合わせる際の重みまで含めて，全体として誤りが減少するように複数モデルの生成法が設計されているので，以下の分類には含めない.

るが，本論文の混合法にはそのような制約はないので，原理的には，第一段として任意のシステムを採用することが可能である．

6.2 Stacking 法

次に，本節では，stacking 法についての関連研究，および，stacking 法と同等の手法を自然言語処理におけるシステム混合の問題に適用している研究事例について述べる．

stacking 法は，(Wolpert 1992) によってその枠組みが提案され，その後，機械学習の分野においていくつかの応用手法が提案されている (Breiman 1996b; Ting and Witten 1997; Gama 2000)．例えば，(Breiman 1996b) は，回帰法を用いた stacking を提案している．(Ting and Witten 1997) は，第一段の学習器として，決定木学習，ナイーブベイズ，最近隣法を用い，第二段の学習器として，決定木学習，ナイーブベイズ，最近隣法，線形回帰法の一種を用いた実験を行ない，性能の比較をしている．一方，(Gama 2000) は，それまで提案された stacking 法を，n 段の学習器の連鎖に拡張し，第 k $(1 < k \leq n)$ 段の学習器は，第一段から第 $k-1$ 段までの全ての学習器の入出力データを素性として学習を行なうというカスケード法を提案している．特に，それまでの stacking 法は，第一段の学習器の出力のみを入力素性として第二段の学習器の学習を行なうものがほとんどであったのに対して，カスケード法では，前段までの学習器の出力だけでなく，入力素性もあわせて利用する点が特徴的である．

一方，自然言語処理におけるシステム混合の問題に stacking 法と同等の手法を適用している研究事例[15] としては，英語品詞付けにおいて，最大エントロピー法，変形に基づく学習，トライグラムモデル，メモリベース学習を第一段の学習器とし，決定木学習，メモリベース学習法などを第二段の学習器として stacking を行なうもの (Brill and Wu 1998; van Halteren et al. 1998)，英語名詞句まとめ上げにおいて，七種類の学習器を第一段に用い，決定木学習，メモリベース学習法を第二段の学習器として stacking を行なうもの (Sang 2000) などがある．これらの事例においては，いずれも，第一段の入力素性および出力を用いて第二段の学習器の学習を行なった結果も報告している．また，(Borthwick, Sterling, Agichtein and Grishman 1998) は，英語の固有表現抽出において，単一の最大エントロピーモデルの素性として，通常の固有表現まとめ上げ・タイプ分類に用いる素性とあわせて，他の既存のシステムの出力を素性として用いて，個々の単語に固有表現まとめ上げ状態・タイプ分類を付与するための分類器の学習を行なっている．一方，(Freitag 2000) は，情報抽出におけるテンプレート・スロット埋め問題において，ナイーブベイズ法，帰納的論理プログラミング法などを第一段の学習器とし，回帰法を第二段の学習器として stacking を行なっている．ここでは，第二段の学習器の入力は，第一段の学習器の出力のみとなっている．

これらの事例と比較すると，本論文の日本語固有表現抽出の問題においては，第一段の学習

15 “stacking” という用語を用いていない事例も多い．

器は，個々の形態素に固有表現まとめ上げ状態・タイプ分類を付与するための分類器の学習を行なっているのに対して，第二段の学習器は，個々のシステムの固有表現抽出結果，および，第一段の学習器の入力となった素性 (の一部) を入力として，個々のシステムの固有表現抽出結果の正誤を判定するための分類器の学習を行なっている．このように，本論文の stacking 法では，第一段と第二段の学習器の学習の単位が異なっている点が変則的である．ただし，このような構成をとることにより，第一段としては，任意の固有表現抽出システムを用いることが可能となっている．また，(Borthwick et al. 1998) と比較すると，(Borthwick et al. 1998) では，本論文の第二段に相当する学習器が，個々の単語に固有表現まとめ上げ状態・タイプ分類を付与するための分類器の学習を行なっている点が異なっている．

6.3　統計的手法に基づく日本語固有表現抽出

統計的手法に基づく日本語固有表現抽出の研究事例としては，我々がベースとした，最大エントロピー法を用いるもの (内元他 2000) の他に，決定木学習を用いるもの (Sekine et al. 1998; 野畑 1999)，最大エントロピー法を用いるもの (Borthwick 1999)，決定リスト学習を用いるもの (Sassano and Utsuro 2000)， SVM(support vector machines) を用いるもの (山田他 2001) などがある．これらは，いずれも，単一の学習モデルを用いている．決定リスト学習を用いる事例 (Sassano and Utsuro 2000) では，可変長文脈素性を用いることにより，固定長モデルの性能の上回る結果が得られているが，ベースとなる決定リスト学習の性能は最大エントロピー法の性能よりも劣っている．その他の事例では，いずれも，固定長文脈素性を用いている．

また，stacking 法の研究事例においては，異なる数種類の学習器を第一段に用いるという構成が多く見られ，一定の効果が報告されているので，上記の複数の学習器を第一段として stacking 法を行なうことにより，精度の向上が期待できる可能性がある．その他には，(山田他 2001) で報告されているように，解析の方向を文頭から文末と文末から文頭の二通り設定し，解析済の固有表現のタグを素性として利用する方法により，振る舞いの異なった出力が得られる可能性があり，stacking 法でその出力を利用することで，精度の向上が期待できる可能性がある．また，(磯崎 2000) では，決定木学習により学習された可読性の高い規則や人手による付加制約等を適用して複数の固有表現候補を生成し，最長一致法により複数の候補の選別を行なっている．ここで，複数の候補の選別を行なう際に，本論文の混合法を適用することにより，誤出力の棄却まで含めたより一般的な選別が自然な形で実現できる可能性があると考えられる．

7　おわりに

本論文では，日本語固有表現抽出の問題において，複数のモデルの出力を混合する手法を提案した．まず，最大エントロピー法に基づく統計的学習による固有表現抽出モデルにおいて，

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を考慮して学習を行なう可変長モデルと，常に現在位置の形態素の前後数形態素ずつまでを考慮して学習を行なう固定長モデルとの間のモデルの挙動の違いに注目し，なるべく挙動が異なり，かつ，適度な性能を保った複数のモデルの出力の混合を行なった．混合の方式としては，複数のシステム・モデルの出力 (および訓練データそのもの) を入力とする第二段の学習器を用いて，複数のシステム・モデルの出力の混合を行なう規則を学習するという混合法 (stacking 法) を採用した．第二段の学習器として決定リスト学習を用いて，固定長モデルおよび可変長モデルの出力を混合する実験を行なった結果，最大エントロピー法に基づく固有表現抽出モデルにおいてこれまで得られていた最高の性能を上回る性能が達成された．

今回の実験では，固定長モデル同士は出力される固有表現の分布がお互いに似通っており，可変長モデル同士も使用する素性の集合に包含関係があることから，出力する固有表現の傾向が大きく異なるモデルは，固定長モデルと可変長モデルの二種類だけであると仮定した．そのため，評価実験においても，二つのモデルの出力の混合の結果のみを報告したが，今後は、傾向の大きく異なる三種類以上のモデルの出力に対して，本論文の混合手法の有効性を評価したいと考えている．

また，本論文の手法は，個々の単独システムに何らかの固有表現候補を出力させて，それらの固有表現候補を取捨選択するという方法であるので，再現率の観点からは，個々の単独システムの出力の和の再現率が上限となってしまう．したがって，本論文の方法によってより高い性能の固有表現抽出を実現するためには，個々の単独システムが少しでも多くの固有表現候補を出力することが不可欠である．今後は，既存のどの固有表現抽出モデルを用いても抽出が失敗する固有表現の特性を分析し，できるだけ網羅的に固有表現候補を出力し，その結果を本論文の混合法で利用する方式について検討を行なう予定である．その際，網羅的に固有表現候補を出力するためには，まず，何らかの方法によって，広範なテキストから固有表現候補を収集して蓄積する必要があるが，ここでは，新聞記事や WWW 上のテキスト等の大規模テキストから未知語を獲得する，あるいは専門用語を抽出するなどの手法の適用が有効であると考えている．

参考文献

Abney, S., Schapire, R. E., and Singer, Y. (1999). “Boosting Applied to Tagging and PP Attachment.” In *Proceedings of the 1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pp. 38–45.

Berger, A. L., Della Pietra, S. A., and Della Pietra, V. J. (1996). “A Maximum Entropy Approach to Natural Language Processing.” *Computational Linguistics*, **22** (1), 39–71.

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. (1998). "Exploiting Diverse Knowledge Sources via Maximum Entropy in Named Entity Recognition." In *Proceedings of the 6th Workshop on Very Large Corpora*, pp. 152–160.

Breiman, L. (1996a). "Bagging Predictors." *Machine Learning*, **24** (1), 123–140.

Breiman, L. (1996b). "Stacked Regressions." *Machine Learning*, **24** (1), 49–64.

Brill, E. and Wu, J. (1998). "Classifier Combination for Improved Lexical Disambiguation." In *Proceedings of the 17th COLING and the 36th Annual Meeting of ACL*, pp. 191–195.

Della Pietra, S., Della Pietra, V., and Lafferty, J. (1997). "Inducing Features of Random Fields." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **19** (4), 380–393.

Freitag, D. (2000). "Machine Learning for Information Extraction in Informal Domains." *Machine Learning*, **39** (2/3), 169–202.

Freund, Y. and Schapire, R. (1999). "(訳: 安倍 直樹): ブースティング入門." 人工知能学会誌, **14** (5), 771–789.

Gama, J. (2000). "Cascade Generalization." *Machine Learning*, **41** (3), 315–343.

Haruno, M. and Matsumoto, Y. (1997). "Mistake-Driven Mixture of Hierarchical Tag Context Trees." In *Proceedings of the 35th Annual Meeting of ACL and the 8th Conference of EACL*, pp. 230–237.

Haruno, M., Shirai, S., and Oyama, Y. (1999). "Using Decision Trees to Construct a Practical Parser." *Machine Learning*, **34** (1/2/3), 131–149.

Henderson, J. C. and Brill, E. (1999). "Exploiting Diversity in Natural Language Processing: Combining Parsers." In *Proceedings of the 1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pp. 187–194.

Henderson, J. C. and Brill, E. (2000). "Bagging and Boosting a Treebank Parser." In *Proceedings of the 1st Conference of NAACL*, pp. 34–41.

乾孝司, 乾健太郎 (2000). "確信度つき委員会方式による部分係り受け解析." 言語処理学会第 6 回年次大会論文集, pp. 471–474. 言語処理学会.

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

磯崎秀樹 (2000). "固有表現抽出のための可読性の高い規則の自動生成." 情報処理学会研究報告, **2000** (2000–NL–140), 69–76.

工藤拓, 松本裕治 (2000). "Support Vector Machines を用いた Chunk 同定." 情報処理学会研究報告, **2000** (2000–NL–140), 9–16.

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

MUC (1998). *Proceedings of the 7th Message Understanding Conference (MUC-7)*.

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.

Rivest, R. L. (1987). “Learning Decision Lists.” *Machine Learning*, **2**, 229–246.

Sang, E. F. T. K. (2000). “Noun Phrase Recognition by System Combination.” In *Proceedings of the 1st Conference of NAACL*, pp. 50–55.

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

颯々野学, 斎藤由香梨, 松井くにお (1997). “アプリケーションのための日本語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.

颯々野学, 宇津呂武仁 (2000). “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.” 情報処理学会研究報告, **2000** (2000–NL–139), 1–8.

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

Ting, K. M. and Witten, I. H. (1997). “Stacked Generalization: when does it work?.” In *Proceedings of the 15th IJCAI*, pp. 867–871.

内元清貴, 馬青, 村田真樹, 小作浩美, 内山将夫, 井佐原均 (2000). “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, **7** (2), 63–90.

van Halteren, H., Zavrel, J., and Daelemans, W. (1998). “Improving Data Driven Wordclass Tagging by System Combination.” In *Proceedings of the 17th COLING and the 36th Annual Meeting of ACL*.

Wolpert, D. H. (1992). “Stacked Generalization.” *Neural Networks*, **5**, 241–259.

山田寛康, 工藤拓, 松本裕治 (2001). “Support Vector Machines を用いた日本語固有表現抽出.” 情報処理学会研究報告, **2001** (2001–NL–142), 121–128.

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

略歴

宇津呂 武仁: 1989 年京都大学工学部 電気工学第二学科 卒業．1994 年同大学大学院工学研究科 博士課程電気工学第二専攻 修了．京都大学博士（工学）．同年，奈良先端科学技術大学院大学 情報科学研究科 助手．1999 年∼2000 年，米国ジョンズ・ホプキンス大学計算機科学科客員研究員．2000 年，豊橋技術科学大学 工学部情報工学系 講師，現在に至る．自然言語処理の研究に従事．言語処理学会，情報処理学会，人工知能学会，日本ソフトウェア科学会，日本音響学会，ACL，各会員．

颯々野 学: 1991 年京都大学工学部 電気工学第二学科卒業．同年より富士通研究所研究員，現在に至る．1999 年∼2000 年，米国ジョンズ・ホプキンス大学計算機科学科客員研究員．自然言語処理の研究に従事．言語処理学会，情報処理学会，各会員．

内元 清貴: 1994 年京都大学工学部 電気工学第二学科卒業．1996 年同大学院工学研究科 修士課程電気工学第二専攻修了．同年郵政省通信総合研究所入所．現在，独立行政法人通信総合研究所研究員．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．

(2001 年 7 月 4 日 受付)
(2001 年 8 月 24 日 再受付)
(2001 年 10 月 5 日 採録)