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
豊橋技術科学大学 工学部 情報工学系
〒 441-8580 愛知県豊橋市天伯町雲雀ヶ丘 1-1
utsuro@ics.tut.ac.jp

颯々野 学
富士通研究所
sassano@jp.fujitsu.com

内元 清貴
独立行政法人 通信総合研究所
uchimoto@crl.go.jp

あらまし: 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．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) を採用する．第二段の学習器として決定リスト学習を用いて，最大エントロピー法に基づく固有表現抽出モデルのうち，お互いに挙動の異なる固定文脈長モデルと可変文脈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
Department of Information and Computer Sciences,
Toyohashi University of Technology
Tenpaku-cho, Toyohashi, Aichi, 441-8580, Japan
utsuro@ics.tut.ac.jp

Manabu Sassano
Fujitsu Laboratories, Ltd.,
sassano@jp.fujitsu.com

Kiyotaka Uchimoto
Communications Research Laboratory
uchimoto@crl.go.jp

Abstract: In this paper, we propose a method for learning a classifier which combines outputs of more than one Japanese named entity extractors. 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. Experimental evaluation shows that the proposed method achieves improvement over the best known results with named entity extractors based on maximum entropy models.

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

1 はじめに

近年，統計的手法に基づく自然言語処理において，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 Halteren98, Brill98]，名詞句等の句のまとめ上げ [Sang00]，構文解析 [Henderson99, 乾 00] 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提案する．

まず，ベースとなる単独の日本語固有表現抽出モデルとしては，最大エントロピー法に基づく日本語固有表現抽出モデル [内元 00] を用いる．特に，これまでの研究事例 [内元 00, 山田 01] 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，3.4.1 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 [颯々野 00]，3.4.2 節参照) も用いて複数モデルの出力の混合を行なう．次に，複数のモデルの出力を混合する方式としては，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 [Wolpert92] 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，決定リスト学習により，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する．

[内元 00] 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 (5 グラムモデル) が最も性能がよい．また，5 節の実験では，そ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る[1] ．ところが，5 グラムモデルと可変長モデルとでは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．

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 日本語固有表現抽出

2.1 IREX ワークショップ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 — 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．

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が，これらについては，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

[1] [颯々野 00] 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．

[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	···
固有表現まとめ上げ状態		O	ORG_U	O	LOC_S	LOC_C	LOC_E	LOC_U	O	

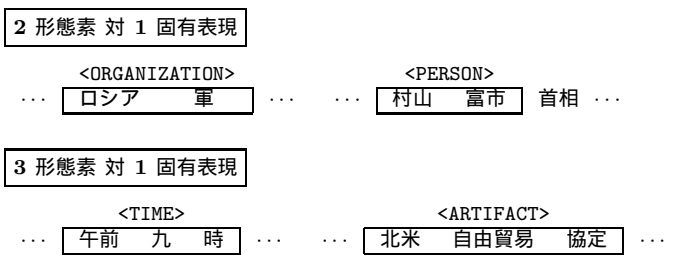

図 1: 複数形態素が一つの固有表現に対応する例

いるので，本論文における考慮の対象には含めない．

3 最大エントロピー法を用いた固有表現抽出

本節では，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 [内元 00] 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．

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

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 3.2 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では，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^{NE}_i 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．

$$\underset{\text{(左側文脈)}}{\cdots M^L_{-k} \cdots M^L_{-1}} \quad \underset{\text{(固有表現)}}{M^{NE}_1 \cdots M^{NE}_i \cdots M^{NE}_m} \quad \underset{\text{(右側文脈)}}{M^R_1 \cdots M^R_l \cdots} \tag{1}$$
$$\uparrow$$
$$\text{(現在位置)}$$

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 (3.3 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 (3.2 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．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 [内元 00]．

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，日本語固有表現抽出の既存の手法 [内元 00] において用いられた Start/End 法を採用する．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，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．この方法により固有表現のまとめ上げを行なう様子を表 3 に示す．

3.3 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]

i) 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，ii) 品詞 — 形態素解析システム BREAKFAST の約 300 種類の品詞，iii) 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．

3.4 周囲の形態素のモデル化

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については，以下の二種類のモデルを用いる．

[3] これらの素性のうち，語彙素性を抽出する条件は [内元 00] に従っている．また，品詞素性については，[内元 00] 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 00] 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 00]．

3.4.1 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．

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

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，左右三形態素ずつを考慮する 7 グラムモデル，左右四形態素ずつを考慮する 9 グラムモデルを用いる．

3.4.2 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 00]．

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[4] ．

$$\begin{array}{ccc}\text{(左側文脈)} & \text{(固有表現)} & \text{(右側文脈)} \\ \cdots \quad M_{-2}^L M_{-1}^L & M_1^{NE} \cdots M_i^{NE} \cdots M_{m(\leq 3)}^{NE} & M_1^R M_2^R \quad \cdots \\ & \uparrow & \\ & \text{(現在位置)} &\end{array}$$

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] ．

3.4.3 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，3.3 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] i) 全素性を用いるモデル，ii) 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，iii) 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．

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，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．

[4] 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の場合は，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．i) 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．ii) 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．iii) 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．

[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 節)．

[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．

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

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．

$$SegEv_j = \{NEev_1, \ldots, NEev_{l_j}\} \quad (2)$$

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 $NEev_{k_j}$ は以下のデータ構造を持つ．

$$NEev_{k_j} = \Big\{ systems = \langle p, \ldots, q \rangle,\ mlength = x \text{ morphemes},\ NEtag = \cdots,\ POS = \cdots,\ class_{NE} = +/- \Big\} \quad (3)$$

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，以下のデータ構造を持つ．

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

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 (2) 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 (2) 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[7] ．

$$class^i_{sys} = \begin{cases} +/-, \ldots, +/- \\ \text{“no output”} \end{cases} \quad (i = 1, \ldots, n) \quad (5)$$

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 (2) 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 (2) 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)．

4.3.4 素性

式 (2) 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．

$$f = \Big\{ \langle systems = \langle p, \ldots, q \rangle,\ F \rangle, \ \cdots, \ \langle systems = \langle p', \ldots, q' \rangle,\ F' \rangle \Big\} \quad (6)$$

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．素性表現 F は，集合 $\{mlength = \cdots, NEtag = \cdots, POS = \cdots\}$ の

[7] 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．

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)	$\{systems=\langle 0\rangle, mlength=1, NEtag=\text{DATE}, POS=\text{時相名詞}, class_{NE}=-\}$ $\{systems=\langle 0\rangle, mlength=1, NEtag=\text{DATE}, POS=\text{時相名詞}, class_{NE}=-\}$ $\{systems=\langle 1\rangle, mlength=2, NEtag=\text{DATE}, POS=\text{時相名詞-時相名詞}, class_{NE}=+\}$
	⋮			
$SegEv_{i+1}$	生殖 (名詞) 医療 (名詞) 技術 (名詞)		生殖医療技術 (ARTIFACT)	$\{systems=\langle 0\rangle, class_{sys}=\text{“no outputs”}\}$ $\{systems=\langle 1\rangle, mlength=3, NEtag=\text{ARTIFACT}, POS=\text{名詞-名詞-名詞}, class_{NE}=-\}$
	について (助詞相当) ⋮			
$SegEv_{i+3}$	山田 (人名) 太郎 (人名)	山田太郎 (PERSON)	山田太郎 (PERSON)	$\{systems=\langle 0,1\rangle, mlength=2, NEtag=\text{PERSON}, POS=\text{人名-人名}, class_{NE}=+\}$
	⋮			

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 $\{class_{sys}=\text{“no outputs”}\}$ のいずれかで表現される．正誤判別規則の学習時には，セグメントの事象表現 $SegEv_j$ (式 (2)) から，式 (6) の形式のあらゆる可能な素性 f のうち，いくつかの制約を満たすものだけが抽出される[8] ．

4.3.5 例

二つの単独システムの固有表現出力を整列した結果 ($SegEv_i \sim SegEv_{i+3}$ の四つのセグメント) を事象表現に変換した結果を表 4 の「事象表現」の欄に示す．

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 [Yarowsky94] を用いる[9] ．決定リストは，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 規則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 (2) と同じ形式のセグメントの事象表現 $SegEv_j = \{NEev_1,\ldots,NEev_{l_j}\}$ 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，i) 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，ii) 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せが求められ，これが，セグメント中で各システム i 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] ．

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

[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．

[9] 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，$class^i_{sys}=-$ とみなしている．

表 5: 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 6: 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 実験および評価

IREX ワークショップの固有表現抽出タスクの訓練・試験データ (表 2 の「その他」除く，CRL 固有表現データ (一般ドメイン): D_{CRL}，本試験データ (一般ドメイン): D_{formal}) を用いて，本論文の混合法の評価を行った．

5.1 各モデル単独の出力の比較

実験に用いたモデルは，3.4.1 節の固定長モデル (5/7/9 グラムモデル)，および，3.4.2 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，3.4.3 節の三種類の素性の設定も区別して実験を行なった．表 5 に，$TrI = D_{CRL}$ $Ts = D_{formal}$ の場合の各モデルの F 値 ($\beta = 1$) を示す．単独のモデルでは 5 グラムモデルが最も高い性能を示し，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．表 6 には，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} の正解データに対して算出した再現率を示す．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 (二つのモデルの出力間で重複する固有表現数/ 5 グラムモデルの出力の固有表現数) も示す (表中，太字: 和の再現率が最も高く，誤出力の重複率が最も低い結果)．7/9 グラムモデルと比較して，可変長モデルは 5 グラムモデルとの類似性が小さく，特に，誤出力の重複率が比較的小さい点が目立つ．

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

7 グラムモデル，9 グラムモデル，可変長モデルにおいて 3.4.3 節の三種類の素性の設定を区別した合計 9 種類のモデルの各々について，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．

表 7: 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) $TrI = D_{CRL} - D^{200}_{CRL}$, $TrC = D^{200}_{CRL}$

	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) 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 記事に対する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

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．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 (a) よりもずっと大きい．

5.2.2 評価結果

表 7 の評価結果の設定 (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，たとえ，インサイド適用の結果を利用した混合になったとしても，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．また，設定 (b) の場合，固定長モデルとの混合よりも，可変長モデルとの混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 6 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

表 8: 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			
	$n = 1$	$n = 2$	$n = 3$	$n \geq 4$
5 グラムモデル	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)	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)	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)	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)	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)	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)	**85.11** **(86.76)** **(83.52)**	87.73 (86.12) (89.41)	71.04 (64.34) (79.31)	50.89 (35.83) (87.76)

が得られている．また，表 8 では，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，固有表現の形態素長が長くなるほど，5 グラムモデルからの性能向上の度合が大きくなっており，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．

6 関連研究

stacking 法は，[Wolpert92] 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．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 [Brill98, van Halteren98]，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 [Sang00] などがある．[Borthwick98] 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，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．したがって，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．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[11] ．

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の正誤を判別する規則を学習することにより，複数モデルの出力を混合する手法を提案した．

参考文献

[Borthwick98] Borthwick, A., Sterling, J., Agichtein, E. and Grishman, R.: Exploiting Diverse Knowledge Sources via Maximum Entropy in Named Entity Recognition, *Proc. 6th Workshop on VLC*, pp. 152–160 (1998).

[Brill98] Brill, E. and Wu, J.: Classifier Combination for Improved Lexical Disambiguation, *Proc. 17th COLING and 36th ACL*, pp. 191–195 (1998).

[Henderson99] Henderson, J. C. and Brill, E.: Exploiting Diversity in Natural Language Processing: Combining Parsers, *Proc. 1999 EMNLP and VLC*, pp. 187–194 (1999).

[乾 00] 乾孝司, 乾健太郎：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 (2000).

[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).

[Sang00] Sang, E. F. T. K.: Noun Phrase Recognition by System Combination, *Proc. 1st NAACL*, pp. 50–55 (2000).

[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 (1997).

[颯々野 00] 颯々野学, 宇津呂武仁：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, Vol. 2000, No. 2000–NL–139, pp. 1–8 (2000).

[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).

[van Halteren98] van Halteren, H., Zavrel, J. and Daelemans, W.: Improving Data Driven Wordclass Tagging by System Combination, *Proc. 17th COLING and 36th ACL* (1998).

[Wolpert92] Wolpert, D. H.: Stacked Generalization, *Neural Networks*, Vol. 5, pp. 241–259 (1992).

[山田 01] 山田寛康, 工藤拓, 松本裕治：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, Vol. 2001, No. 2001–NL–142, pp. 121–128 (2001).

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

[11] 関連手法の一つである boosting 法では，単一の学習モデルを用いて，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) を学習し，それらの重み付き和により混合を行う．boosting 法では，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．