情 報 学 基 礎 61−17
自 然 言 語 処 理 142−17
（2001. 3. 6）

Support Vector Machines を用いた日本語固有表現抽出

山田 寛康, 工藤 拓, 松本 裕治
奈良先端科学技術大学院大学 情報科学研究科
{hiroya-y, taku-ku, matsu}@is.aist-nara.ac.jp

本稿では, 機械学習アルゴリズム Support Vector Machines (SVMs) を用いて日本語固有表現抽出を学習する手法を提案し, 実験によりその有効性について検証する. 素性として語彙, 品詞, 文字種を使用し, Kernel 関数によって素性の組合せを考慮した学習を行なった. 本手法を IREX 日本語固有表現抽出タスクに適用し, CRL 固有表現データに対する交差検定を行った結果, F 値で 83.81 精度を得た.

キーワード:情報検索, 情報抽出, 固有表現抽出, サポートベクター学習

Japanese Named Entity Extraction using Support Vector Machines

YAMADA Hiroyasu, KUDO Taku, MATSUMOTO Yuji
Graduate School of Infomation Science, Nara Institute Science and Technology
{hiroya-y, taku-ku, matsu}@is.aist-nara.ac.jp

This paper investigates the effect of Support Vector Machines (SVMs) for extracting named entities in Japanese. We used a lexical entry, part-of-speech, type of strings as features and combination of all features by Kernel method. We applied the method to IREX NE task using CRL named entities data. The cross validation result of 83.81 shows the effect of the method.

Keywords : Information Retrieval, Information Extraction, Named Entity, Support Vector Learning

1 はじめに

人名・組織名といった固有表現を自動的に抽出する固有表現抽出 (Named Entity) の問題は, 情報検索, 情報抽出の基礎技術としてのみならず, 自然言語処理における形態素解析, 構文解析などの処理に大きな影響を及ぼすため, 重要な問題とされている.

従来, 様々な分野や新しい固有表現に柔軟に対応するために, 最大エントロピー法 [15], 決定木学習 [5], 決定リスト [4, 17, 11] などの機械学習を用いて, 固有表現規則を学習する手法が提案されている.

固有表現抽出規則を学習する場合, 学習に用いる素性としては, 語彙, 文字種, 品詞, などを用いるため, その素性空間は数万以上の高次元空間となり, 過学習に頑強なアルゴリズムが必要とされる. Support Vector Machines [9, 8] は, その汎化誤差が素性空間の次元に依存しないことが理論的に証明されており, 実験的にも Chunking [6] , 文書分類 [3, 16, 10, 7] などの高次元素性空間での学習を必要とするタスクで高い精度が報告されている.

本稿では, IREX 日本語固有表現抽出タスク [1] に対して, Support Vector Machines を用いて抽出規則を学習し, その有効性を検証する.

以下次節では, 日本語固有表現抽出について述べる. 3 節では学習アルゴリズムである Support Vector Machines について説明し, 4 節で固有表現抽出に Support Vector Machines を適用する方法について説明する. 5 節で実験と考察について報告し, 最後に 6 節でまとめと今後の課題について述べる.

2 日本語固有表現抽出

固有表現の表現方法

IREX 日本語固有表現タスクでは表 1 に示す 8 つの固有表現を定義している. 固有表現抽出は, ある単語が固有表現か否かを分割する Chunking 問題とみなすことができ, 一つ以上の形態素からなる固有表現を Chunking タスクで用いられる IOB1, IOB2, IOE1, IOE2 という 4 種類の記法で表現する [2] .

図 1 は"エリツィン大統領は五日午前零時を期して" という文中エリツィン:人名 (PERSON), 五日:日付 (DATE), 午前零時:時間 (TIME) という 3 つの固有表現に対して, IOB1, IOB2, IOE1, IOE2 それぞれの 表現法の違いを表している. IOB1 は 固有表現である単語にたいして I というタグで表し, 異なる固有表現が連続した場合を識別するために, 連続する固有表現の開始位置を B というタグで表現する. 固有表現以外の単語は O というタグを付ける. IOB2 は IOB1 とは違い, 固有表現の始めには必ず B というタグを付けることで, 固有表現を識別する. IOE1 , IOE2 はそれぞれ, IOB1,IOB2 が開始位置に注目していた替わりに, 終了位置に E というタグを付けて識別する.

表 1: IREX で使用する固有表現の種類その頻度

固有表現のタイプ		例
ARTIFACT	固有物名	ノーベル文学賞
DATE	日付表現	五月五日
LOCATION	地名	日本, 韓国
MONEY	金額表現	２０００万ドル
ORGANIZATION	組織名	社会党
PERCENT	割合表現	二〇％ , 三割
PERSON	人名	村山富市
TIME	時間表現	午前五時

以後本稿では混乱を避けるために, 固有表現の開始終了位置をあらわす B, I, O, E の表記を **Chunk タグ**と呼び, IREX で定義した 8 つの固有表現を**固有表現の種類**と呼ぶ. そして Chunk タグと 固有表現の種類が一つになった, B-DATE のような表記を**固有表現タグ**と呼ぶ.

形態素と固有表現の区切りの違い

日本語は英語などの言語とは違い, わかち書きを必要とする言語である. そのため日本語固有表現を抽出する場合に, わかち書きの位置と固有表現の開始, 終了位置が異なる場合があり, これに対処する必要がある.

本稿では訓練データで形態素解析器のわかち書きの区切りと固有表現の区切りの異なる単語が出現した場合, 固有表現の区切りに合わせる. またテスト事例でこの問題が発生した場合, 訓練データで区切りを合わせた単語に限り, 固有表現の区切りに合わせ対処した.

3 Support Vector Machines

固有表現抽出規則の学習に用いるアルゴリズムである Support Vector Machines(SVMs) の概要について説明する. 図 2 は SVMs の概要図を示す. いま, 訓練事

入力文	エリツィン	大統領	は	五	日	午前	零	時	を	期し	て
タグの種類	タグの種類による各単語の表現										
IOB1	I-PERSON	O	O	I-DATE	I-DATE	B-TIME	I-TIME	I-TIME	O	O	O
IOB2	B-PERSON	O	O	B-DATE	I-DATE	B-TIME	I-TIME	I-TIME	O	O	O
IOE1	I-PERSON	O	O	I-DATE	E-DATE	I-TIME	I-TIME	I-TIME	O	O	O
IOE2	E-PERSON	O	O	I-DATE	E-DATE	I-TIME	I-TIME	E-TIME	O	O	O

図 1: 固有表現の識別のためのタグ表現

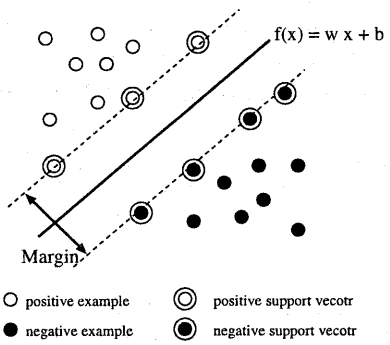

図 2: Support Vector Machines

例 として, n 次元の特徴ベクトル $\mathbf{x}_i$ と 正例 (+1), 負例 (-1) を表すラベル y_i のペアを考える. l 個の訓練事例は以下のように表すことができる.

$$(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \cdots, (\mathbf{x}_l, y_l), \quad \mathbf{x}_i \in \mathbf{R^n}, \mathbf{y_i} \in \{+\mathbf{1}, -\mathbf{1}\}$$

これらの訓練事例に対して, 式 1 に示す正例・負例を正しく分離する超平面を考える.

$$\mathbf{w} \cdot \mathbf{x} + b = 0, \mathbf{w} \in \mathbf{R}^n, b \in \mathbf{R} \qquad (1)$$

一般に訓練事例中の 正例, 負例 を正しく分離する超平面は無数に存在する. SVMs はマージンと呼ばれる図 2 で破線で描かれた超平面間の距離が最大になるような分離平面を見つけることで, 訓練事例に特化した過学習を回避する. 式 2 は , 正例側 (+), 負例側 (-) のそれぞれに対して, 分離超平面と並行で, 分離平面から等距離にある超平面を表し, これを用いてマージン ρ は式 3 で計算される.

$$\mathbf{w} \cdot \mathbf{x} + b = \pm 1, \quad \mathbf{w} \in \mathbf{R}^n, b \in R \qquad (2)$$

$$\rho = \frac{|\mathbf{w} \cdot \mathbf{x}_i + b - 1|}{\|\mathbf{w}\|} + \frac{|\mathbf{w} \cdot \mathbf{x}_i + b + 1|}{\|\mathbf{w}\|} = \frac{2}{\|\mathbf{w}\|} \qquad (3)$$

マージンを最大にすることは $\|\mathbf{w}\|$ を最小にすることであり, 式 4 の制約条件付き最適化問題を解くことと等価である.

$$\text{目的関数}: \quad \frac{1}{2}\|\mathbf{w}\|^2 \rightarrow \text{最小化}$$
$$\text{制約条件}: \quad y_i[(\mathbf{w} \cdot \mathbf{x}_i) + b] \geq 1, (i = 1, \cdots, l) \qquad (4)$$

これは 2 次計画問題として Lagrange の未定乗数法を用いて解くことが可能であり, Lagrange 乗数を α_i とすると, 式 5 に示す解を得る.

$$\mathbf{w} = \sum_{i=1}^{l} \alpha_i y_i \cdot \mathbf{x}_i \qquad (5)$$

ここで $\alpha_i > 0$ となる 訓練事例 $\mathbf{x_i}$ を Support Vector と呼ぶ. これらの事例は分類に影響を与える事例であり , α_i はその重みを表す. 未知の事例 $\mathbf{x}$ を正・負例に分類する決定関数 $f(\mathbf{x})$ を式 6 に示す.

$$f(\mathbf{x}) = sgn(\sum_{i=1}^{l} \alpha_i y_i \mathbf{x}_i \cdot \mathbf{x} + b) \qquad (6)$$

本稿で SVMs を固有表現抽出規則の学習に採用した主な理由は以下の 3 点である.

- Structural Risk Minimisation(SRM) よる高次元素性空間での過学習の回避
- Kernel 関数適用による素性の組合せを考慮した学習
- テスト事例に対する精度以外に, VC confidence, Leave-one-out bound などの学習モデルに対する指標

これらについて順に説明する.

SRMによる高次元素性空間での過学習の回避

訓練事例テスト事例が独立で同一の確率分布 $P(\mathbf{x}, y)$ から生成されたと仮定する. l 個の訓練事例から学習した仮説 f によって訓練事例を分類した場合のエラー率, 事例全体を分類した場合のエラー率をそれぞれ, $R_{emp}(f)$, $R(f)$ とする. これら二つは式 7, 8 で表され, また $1-\eta$ の確率で式 9 の関係が成り立つことが知られている [8]

$$R_{emp}(f) = \frac{1}{l}\sum_{i=1}^{l}\frac{1}{2}|f(\mathbf{x}_i) - y_i| \quad (7)$$

$$R(f) = \int \frac{1}{2}|f(\mathbf{x}) - y| dP(\mathbf{x}, y) \quad (8)$$

$$R \leq R_{emp} + \sqrt{\frac{h(\ln(\frac{2l}{h}) + 1) - \ln\frac{\eta}{4}}{l}} \quad (9)$$

式 9 で h は VC 次元と呼ばれる仮説 f の複雑さを表す指標でる, また式 9 の右辺第二項を VC confidence と呼び, 第一項の訓練事例に対するエラー率とは trade off の関係にある. 右辺全体を Structural Risk とよび, SRM は 右辺を最小にすることで未知の事例も含めた事例全体のエラー率を最小にする戦略である.

SVMs の VC 次元は, 全事例を囲む hypersphere の最小直径を D , マージンを ρ , 素性空間の次元数を d とすると, 式 10 によって上限される [8] .

$$h \leq min(\frac{D^2}{\rho^2}, d) + 1 \quad (10)$$

式 10 より, マージン ρ を最大にする SVMs の戦略は VC 次元を最小にすることと等価である. 仮説 f が訓練事例に対して正しく分離できると仮定すると, h の最小化は 式 9 の右辺第二項を最小にするため SVMs は SRM に基づく戦略といえる. また式 10 より, 素性空間の次元数 d が十分大きな値をとった場合には h の最小化は d に依存しないため, SVMs は素性空間の次元に依存しない汎化能力を持つといえる.

Kernel 関数の適用よる素性の組合せを考慮した学習

SVMs は学習における最適化問題を解く場合も, テスト事例を分類する決定関数にも事例間の内積が使用されるため, そこに Kernel 関数を適用する. Kernel 関数の適用により, 事例を高次元空間に写像しそこでの線形分離問題を解くことで, もとの空間での非線形問題に対応することができる.

Kernel 関数には様々なものが提案されているが, とくに d 次の Polynomial 関数 $K(x, y) = (x \cdot y + 1)^d$ は d 個の素性の組合せを考慮した次元への写像を意味し, これにより素性の組合せを考慮した学習が可能となる.

学習モデルを評価する指標

与えられた l 個の訓練事例のうち $l-1$ 個で学習し残りの 1 つをテスト事例とし分類するという処理を l 回繰り返す Leave-one-out (LOO) は, 学習モデルを評価する一つの方法である. SVMs では LOO のエラー率 E_{LOO} が式 11 で上限されることが知られている [8] .

$$E_{LOO} \leq \frac{number\ of\ Support\ Vectors}{l} \quad (11)$$

この他にも式 9 , 10 など示した VC confidence や VC 次元も, 学習モデルが未知の事例に対してどれだけ精度よく分類できるかを表す指標の一つであり, 実際に Chunking や未知語の品詞推定などのモデルの自動選択に利用され, よい精度が報告されている [12, 14].

本稿でも, 単にテスト事例に対する精度だけで学習モデルを評価するのではなく, これらの指標を基に学習に使用した素性などの違いによる比較, 考察を行なう.

4 SVMs の日本語固有表現抽出への適用

SVMs を日本語固有表現抽出タスクに適用する方法について説明する.

解析方向と素性

固有表現を抽出する場合, 文頭から文末の方向に, 順に単語の固有表現タグを推定する**右向き解析**と, 文末から文頭へ順に推定する**左向き解析**の 2 つを考える.

また各単語の素性としては, 語彙, 品詞の他に, 単語に含まれる文字の種類を考慮する. 今回文字種としては, カタカナ, 平仮名, 漢字, 記号, 数字, アルファベットの 6 つの種類を素性として考慮した.

位置	単語	品詞	文字種	固有表現タグ
1	エリツィン	名詞	カタカナ	B-PERSON
2	大統領	名詞	漢字	O
3	は	助詞	ひらがな	O
4	五	名詞	漢字	B-DATE
5	日	名詞	漢字	-
6	午前	名詞	漢字	-
7	零	名詞	漢字	-

図 3: 学習に使用する素性

訓練事例は, 解析方向順に, i 番目の単語の固有表現タグを分類クラスとし, 素性は $i-2$ から $i+2$ までの単語の語彙, 品詞, 文字種, $i-2$ から $i-1$ の 固有表現タグを使用し生成する.

テスト事例は, 語彙, 品詞, 文字種に関しては既知であるが, 固有表現タグについては未定であるため, 推定した固有表現を順次動的に追加し, 以降の解析の素性に利用する. 即ち, 解析方向順に i 番目の単語の素性は, $i-2$ から $i+2$ の語彙, 品詞, 文字種, $i-2$ から $i-1$ の固有表現は $i-2$,$i-1$ 番目の解析で推定した固有表現タグを使用する.

図 3 は, 入力文 "エリツィン大統領は五日午前零時" に対して, IOB2 Chunk タグで, 右向き解析した場合, 4 番目の単語"五"の素性を示す.

入力文が訓練データであれば, 固有表現タグ B-DATE が分類するクラスであり, 素性として, 2 から 6 までの語彙, 品詞, 文字種, 2 から 3 の固有表現タグを使用する. 同一の文がテストデータであれば, 4 番目の単語の固有表現タグを推定するために, 2 から 6 番目の語彙, 品詞, 文字種, 2 から 3 番目で推定した固有表現タグを素性とする.

二値分類から多値分類への拡張

SVMs は正例・負例を分類する二値分類器であるため, 固有表現規則を学習するためには複数のクラスに分類する多値分類に拡張する必要がある. これには 2 つの方法が提案されており, あるクラスかそれ以外かという二値分類器を分類するクラス数分構築する one class vs. all others 法と, 分類する k 個のクラスから 2 つのクラスに関する二値分類器を ${}_kC_2$ 個の構築する pairwise 法がある. 本稿では 工藤らと同様に pairwise 法を用いた [6]. pairwise 法では ${}_kC_2$ の分類器の多数決により最終的な分類クラスを決定する.

表 2: 固有表現の種類その頻度

固有表現のタイプ	出現頻度
ARTIFACT	747
DATE	3567
LOCATION	5463
MONEY	390
OPTIONAL	585
ORGANIZATION	3676
PERCENT	492
PERSON	3840
TIME	502

5 実験

5.1 データ

実験では CRL (郵政省通信総合研究所) 固有表現データを使用した[1]. CRL 固有表現データは, 毎日新聞 95 年度版 1174 記事, 10718 文に対して IREX で定義された固有表現タグが付与されている. 表 2 は CRL 固有表現データに出現した固有表現の種類とその頻度を示す. また形態素解析器は茶筌 [13] を使用した. テストデータに対する評価は $\beta=1$ の F 値を使用した. また学習したモデルの評価として 式 9 の右辺第二項である VC confidence と 式 11 の Leave-one-out bound を用いた. VC confidence の計算で $\eta=0.05$ とした. これら 2 つの指標は訓練データに対するエラー率が 0 である場合, その値が低いほどテスト事例に対してよい精度を期待できる.

5.2 単語を表す素性の違いによる精度の比較

予備実験として, 単語に関する素性の違いによる精度を調査した. Chunk タグを IOB2, 解析方向を右向きに固定し, 以下の 4 つのセットについて精度を比較した.

- (a) 語彙, 品詞大分類

[1]現在, IREX 日本語固有表現抽出タスク本試験データは参加者以外使用できない. そのため本稿では非参加者が使用可能でデータ量の多い CRL 固有表現データで実験を行なった.

- (b) 語彙, 品詞細分類
- (c) 語彙, 品詞大分類, 文字種
- (d) 語彙, 品詞細分類, 文字種

表 3: 素性の違いによる精度 (5 分割の交差検定)
(a): 語彙 + 品詞大分類, (b): 語彙 + 品詞細分類, (c): 語彙 + 品詞大分類 + 文字種, (d): 語彙 + 品詞細分類 + 文字種

	$F_{\beta=1}$ 値			
素性の種類	(a)	(b)	(c)	(d)
ARTIFACT	45.90	**46.78**	46.30	46.29
DATE	89.72	91.07	91.11	**91.75**
LOCATION	76.88	79.03	83.22	**83.49**
MONEY	91.08	91.34	**91.76**	91.13
OPTIONAL	39.85	45.75	54.36	**55.88**
ORGANIZATION	64.93	68.57	76.61	**76.89**
PERCENT	91.27	91.74	91.58	**92.24**
PERSON	69.57	72.80	83.90	**84.88**
TIME	88.91	89.38	**89.56**	**89.56**
総合	75.11	77.32	82.29	**82.73**
VC conf.	0.3950	0.3748	0.5621	0.4719
LOO bound	0.0471	0.0264	0.0317	**0.0220**

表 3 にデータを 5 分割した交差検定の結果を示す. 表 3 より, (d) 語彙, 単語の文字種 , 品詞細分類を素性とした場合に最も良い精度が得られた.

(b),(d) から, 品詞細分類を素性として追加することで, 精度が向上することが解る. これは名詞などの細分類には, 人名, 地名などの固有名詞に関する情報が付与されているため, 当然の結果とえいる.

文字種や, すべての品詞の細分類情報を素性として追加することで, 素性空間の次元はより高次元になるが, VC confidence , は, 品詞細分類を考慮しない場合よりも考慮した場合に下がり, LOO bound の値では最も高次元な素性空間である (d) で最小となった. これは SVMs による学習によって素性空間の次元に依存しない汎化能力の裏付けといえる.

以後の実験では各事例を表す素性として, (d) の 語彙, 品詞細分類, 文字種を用いることとする.

5.3 分割モデルと解析方向の違いによる精度の比較

各事例を表現する 素性を, 前後 2 単語の語彙, 品詞再分類, 文字種に固定し, Chunk タグと, 解析方向の違いによる精度を調査した. 表 4 に データを 5 分割した交差検定の結果を示す. 表 4 より, 全体の精度が最も高かったのは, IOB2 タグで左向き解析を行なった場合であった. また解析方向では, いずれのタグの種類でも左向き解析の場合がよい精度が得られた. これは本稿の固有表現抽出方法が, 解析の順に一意に決定されるため, 「A の B」 などの固有表現は, 右向き解析の場合, 単語 A の固有表現の種類によって決定されてしまうことが原因である. この場合主辞である B の固有表現タグを考慮できる左向き解析が適切である. また VC confidence , LOO bound の値も, IOE2 を除いたすべてのタグにおいて左向き解析のときに値が低く, このことからも固有表現抽出では左向き解析が重要であることがわかる.

5.4 Kernel 関数の違いによる精度の比較

Chunk タグを IOB2 , 事例の素性としては, 単語の語彙, 品詞細分類, 文字種とし, 解析方向を 左向きに固定し, Polynomial Kernel 関数の次元 d を変化させた場合の精度を調査した. これは素性を独立に扱う学習と, d 個の組合せを考慮した学習での比較を意味する. 評価は学習時間の都合上交差検定の 1 つのセットで行った. 結果を表 5 に示す.

表 5 より, 素性を独立に扱う学習より, 2 つ 素性の組合せを考慮したモデルで最も良い精度を得ることができた. しかし, VC confidence と LOO bound の値はこれら精度と矛盾し それぞれ $d = 4$,$d = 1$ で最小になった. 今回 式 9 の右辺第二項の VC confidence を使用したが, 右辺全体を使用した VC bound と呼ばれる値を求めることで, 原因の追求を行う予定である.

5.5 わかち書きの性能による精度の比較

本稿では, わかち書きと固有表現の区切りが異なる場合, 固有表現の区切りに合わせた. そしてテスト事例でこの問題が発生した場合, 訓練データで区切りを固有表現に合わせた単語のみに限り, 区切りを固有表現に合わせ対処した.

表 4: Chunk タグの違いと解析方向の違いによる精度比較

	$F_{\beta=1}$ 値							
タグの種類	IOB1		IOB2		IOE1		IOE2	
解析方向	右	左	右	左	右	左	右	左
ARTIFACT	47.19	47.66	46.79	48.33	47.23	**48.34**	44.82	46.37
DATE	91.98	**92.75**	91.89	92.73	91.92	92.59	91.28	92.39
LOCATION	82.74	83.17	82.48	**84.30**	82.60	82.98	82.52	82.33
MONEY	90.64	93.83	91.01	93.98	90.64	93.83	90.52	**94.10**
OPTIONAL	64.34	62.85	63.47	**64.93**	64.36	62.98	36.62	38.64
ORGANIZATION	76.20	76.80	75.83	**78.74**	76.47	76.88	75.48	75.98
PERCENT	89.90	**93.21**	91.73	92.08	89.79	**93.21**	89.26	92.50
PERSON	85.06	85.47	85.07	**85.68**	85.12	85.47	85.11	85.42
TIME	87.16	86.62	**88.61**	85.44	87.16	86.62	84.40	86.42
総合	82.51	83.02	82.41	**83.81**	82.52	82.97	81.47	81.94
VC Conf.	0.5045	0.4834	0.4705	0.4606	0.5040	0.4876	0.4903	**0.4548**
LOO bound	**0.0179**	**0.0179**	0.0217	0.0204	0.0182	0.0180	0.0216	0.0218

表 5: Polynomial Kernel の次元数による精度比較

	$F_{\beta=1}$ 値			
d	1	2	3	4
ARTIFACT	32.94	**46.49**	46.08	45.28
DATE	89.78	**91.83**	91.08	89.90
LOCATION	78.69	**82.48**	82.26	81.85
MONEY	87.76	87.76	87.76	87.76
OPTIONAL	48.89	**55.97**	53.94	52.10
ORGANIZATION	72.48	**79.10**	78.12	77.46
PERCENT	89.61	**91.49**	91.10	91.43
PERSON	82.81	85.38	**85.43**	84.86
TIME	82.40	**85.23**	82.20	83.26
総合	79.26	**82.81**	82.31	81.74
VC conf.	0.387	0.461	0.381	**0.352**
LOO bound	**0.0151**	0.0205	0.0261	0.0322

表 6: わかち書きの性能による精度比較

	$F_{\beta=1}$ 値		
	(A)	(B)	(C)
ARTIFACT	47.15	48.33	48.17
DATE	92.15	92.73	92.83
LOCATION	84.05	84.30	87.74
MONEY	93.98	93.98	94.50
OPTIONAL	54.95	64.93	71.47
ORGANIZATION	79.53	78.74	81.43
PERCENT	94.69	92.08	97.62
PERSON	85.94	85.68	86.38
TIME	89.33	85.44	89.98
総合	83.77	83.81	**85.92**

そこで, テストデータに対して (A) 形態素解析器のわかち書きをそのまま使用, (B) 訓練データで区切りを固有表現に合わせた単語のみを考慮 ,(C) 固有表現の区切りを正しく推定した場合, の 3 種類の精度を比較し SVMs による潜在的な精度を測った. 結果を表 6 に示す.

(A)(B) では大きな精度向上は見られなかったが, (C) の形態素解析器が固有表現の区切りを完全に推定した場合, F 値で 2 ポイントの上昇が見られた. 今後, わかち書きと固有表現の区切りが異なる場合, 形態素解析器の複合語処理などにより適切なわかち書きをすることで精度の向上が期待できる.

5.6 他手法との比較

IREX の本試験データを使用した他の学習法と比較した. 比較の対象としたのは, 内元らの最大エントロピーによる手法 [15] と, 颯々野らの可変長文脈を考慮した最大エントロピー法 [17] とした. 他手法の精度はいずれも本試験データの GENERAL と呼ばれるデータに対する F 値である. 我々は本試験データを使用できないため, 交差検定の各検定での精度を記載した.

同一データでないため完全な比較は不可能だが, SVMs で学習した場合の最低精度は 81.34 であり, 内元らの手法と同等以上の精度が期待できる. また颯々野らの手法は固有表現の長さにより可変長文脈を考慮しているため, 一部の精度で我々を上回っている. 今後 SVMs を用いて学習する場合でも, 固有表現の長さによる可変長文脈を考慮する必要がある.

表 7: 他手法との比較

	平均	1	2	3	4	5
SVMs	83.01	82.81	82.49	85.19	81.34	83.73
他手法	ME(内元)			ME(颯々野)		
	80.17			82.8		

6 まとめと今後の課題

本稿では, IREX 日本語固有表現タスクに対して, Support Vector Machines を用いて固有表現抽出規則を学習し, 抽出実験においてその有効性を示した. SVMs の素性空間に依存しない高い汎化能力により, 語彙, 品詞細分類, 文字種で表現した高次元素性空間で, 最良の精度を得ることができた. また固有表現抽出では, 「A の B」のような固有表現にて適切に対処する左向き解析がよりよい精度を得ることが解った. また SVMs に 多項式 Kernel 関数を適用することで素性間の組合せを考慮した学習を行ない, 固有表現抽出タスクでは 2 つの素性の組合せを考慮した場合, 最も高い精度を得ることができた. また VC confidence , LOO bound の 2 つのモデルに対する評価指標を用いて, これらの指標と選択した素性, 解析方向, Kernel 関数の違いによる精度との関係について考察した.

今後の課題としては, 颯々野らの提案した固有表現の長さによって可変文脈長を考慮した学習 [4, 17] があげられる. また実験により, 各固有表現の種類によって, 選択すべき最良の素性が異なることが確認できたため, VC confidence , LOO bound などを基に固有表現の種類によって素性を自動的に選択できるよう拡張する予定である.

参考文献

[1] IREX 実行委員会: IREX homepage, http://cs.nyn.edu/cs/projects/proteus/irex, 1999.

[2] Erik F. Tjong Kim Sang and Jorn Veenstra. Representing text chunks. In *In Proceedings of EACL'99*, pp. 173–179, 1999.

[3] Hirotoshi Taira , Masahiko Haruno. Feature Selection in SVM Text Categorization. In *AAAI-99/IAAI-99 Proceedings *Sixteenth National Conference on Artificial Intelligence / Eleventh Conference on Innovative Applications of Artificial Intelligence, Orlando, Florida*, pp. 480–486, 18-22 Jul 1999.

[4] Manabu Sassano, Takehito Utsuro. Named Entity Chunking Techniques in Supervised Learning for Japanese Named Entity Recognition. In *18th International Conference on Computational Linguistics*, pp. 705–711, 2000.

[5] Satoshi Sekine and Ralph Grishman and Hiroyuki Shinnou. A Decision Tree Method for Finding and Classifying Names in Japanese Texts. In *the Sixth Workshop on Very Large Corpora*, pp. 171–178, 1998.

[6] Taku Kudoh and Yuji Matsumoto. Use of Support Vector Learning for Chunk Identification. In *Computational Natural Language Learning (CoNLL-2000)*, pp. 142–144, 2000.

[7] Thorsten Joachims. Transductive Inference for Text Classification using Support Vector Machines. In *Machine Learning *Proceedings of the Sixteenth International Conference (ICML '99)*, pp. 200–209, 27-30 Jun 1998.

[8] Vladimir N. Vapnik. *Statistical Learning Theory*. A Wiley-Interscience Publication, 1998.

[9] Vladimir N. Vapnik. *The Nature of Statistical Learning Theory*. New York, 1995.

[10] Yiming Yang , Xin Liu. A Re-examination of Text Categorization Methods. In *SIGIR '99 *Proceedings of the 22nd International Conference on Research and Development in Information Retrieval, University of California, Berkeley*, pp. 42–49, 9 Aug 1999.

[11] 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. 情報処理学会研究会報告, No. 2000-NL-139, pp. 9–16, 2000.

[12] 工藤 拓, 松本 裕治. Support Vector Machine を用いた Chunk 同定. 情報処理学会研究会報告, No. NL-140-2, pp. 9–16, 2000.

[13] 松本 裕治, 北内 啓, 山下 達雄, 平野 善隆, 松田 寛, 浅原 正幸. 日本語形態素解析システム「茶筌」version 2.0 使用説明書第二版, 12 1999.

[14] 中川 哲治, 工藤 拓, 松本 裕治. Support Vector Machine を用いた未知語の品詞推定. 情報処理学会研究会報告, No. NL-141-13, pp. 77–82, 2001.

[15] 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, 第 7 巻, pp. 63–90, 2000.

[16] 平 博順 , 春野 雅彦. Support Vector Machine によるテキスト分類における属性選択. 情報処理学会論文誌, 第 41 巻, pp. 1113–1123, 4 2000.

[17] 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 j 評価. 情報処理学会研究会報告, No. 2000-NL-139, pp. 1–8, 2000.