情報学基礎 64－7
自然言語処理 145－7
（2001．9．10）

機械学習を用いたタイ語の品詞タグづけ

村田 真樹　　馬 青　　井佐原 均

独立行政法人 通信総合研究所
けいはんな情報通信融合研究センター
〒619-0289 京都府相楽郡精華町光台 2-2-2
TEL:0774-95-2424 FAX:0774-95-2429 {murata,qma,isahara}@crl.go.jp

あらまし

われわれはこれまで，できるだけ少量の訓練データで実用的な品詞タグづけシステムを構築する目的で伸縮性入力ニューロタガー，ニューラルネットワークと Brill の誤り駆動学習を融合したハイブリッド型ニューロタガーを提案してきた．これらのタガーでは多品詞語を含む小規模タイ語コーパスを学習データとして，それぞれ 94.4%，95.5% という高い精度 (多品詞語のみを測定対象) で品詞のタグづけをした．本研究では，新たに決定リスト，最大エントロピー法，サポートベクトルマシン法の三つの機械学習手法を用いるタイ語の品詞タグづけの手法を提案し，それらを用いて品詞タグづけを行なった．結果はサポートベクトルマシンがもっともよい精度 (96.1%) を出し，さらにタイ語の品詞タグづけの精度向上を実現することになった．本稿では，手法間の比較や精度向上の原因についての考察も行なう．

キーワード　機械学習，品詞タグづけ，タイ語，サポートベクトルマシン

POS Tagging in the Thai Language Using Machine-Learning Methods

Masaki Murata　Ma Qing　Hitoshi Isahara

Keihanna Human Info-communication Research Center,
Communications Research Laboratory
2-2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619-0289, Japan
TEL:+81-774-95-2424 FAX:+81-774-95-2429 {murata,qma,isahara}@crl.go.jp

Abstract

The elastic-input neuro tagger and the hybrid neuro tagger combined the neural network and Brill's error-driven learning have already been proposed for the purpose of constructing a practical tagger that uses as little training data as possible. When a small Thai corpus is used for training, these taggers have tagging accuracies of 94.4% and 95.5% (accounting only for the ambiguous words in part of speech), respectively. In this study, we developed new tagging methods using the three machine learning methods: decision-list, maximum entropy, and support vector machine methods. We then performed tagging experiments using these methods. Our results showed that the support vector machine method has the best precision (96.1%), and that it is capable of improving the accuracy of tagging in the Thai language. Finally, we theoretically examined the above-mentioned methods and discussed how improvements have been made.

key words　Machine Learning, Part of Speech Tagging, Thai Language, Support Vector Machine

1 はじめに

われわれはこれまで，できるだけ少量の訓練データで実用的な品詞タグづけシステムを構築する目的で伸縮性入力ニューロタガー[1]，ニューラルネットワークと Brill の誤り駆動学習を融合したハイブリッド型ニューロタガー[2] を提案してきた．これらのタガーでは多品詞語を含む小規模タイ語コーパスを学習データとして，それぞれ 94.4%，95.5% という高い精度 (多品詞語のみを測定対象) で品詞のタグづけをした．本研究では，新たに決定リスト，最大エントロピー法，サポートベクトルマシン法の三つの機械学習手法を用いるタイ語の品詞タグづけの手法を提案し，それらを用いて品詞タグづけを行なう．タイ語の品詞タグづけのための教師ありデータとしては文献[1, 2, 3, 4, 5, 6] のコーパスのものを用いる．

本稿で強調したいことをあらかじめまとめておくと，以下のようになる．

- 本論文ではサポートベクトルマシン法を用いてタイ語の品詞タグづけの問題を扱った．形態素解析を含む品詞タグづけの問題を機械学習手法で扱った研究は数多くあるが，サポートベクトルマシン法を用いた研究はあまり見ない．サポートベクトルマシン法は性能は高いが手法の利用コストが大きく大規模コーパスの場合は容易に動かせないという欠点がある．また，大規模コーパスの場合はわざわざサポートベクトルマシン法のような強力な学習手法を用いなくとも，HMM などの手法でも十分な精度が出ることが多い．これに対してタイ語のようにまだ大規模なコーパスが作成されていない言語では本論文のような手法が有効となる．
- サポートベクトルマシン法だけでなく，比較のため決定リスト法，最大エントロピー法でも実験を行ない，サポートベクトルマシン法が最もよい精度を出すことを確認した．本稿はこのような機械学習の性能比較のデータも提供するものである．
- サポートベクトルマシン法によって得られた精度が先行研究[2] のパーセプトロンと Brill の誤り駆動学習のハイブリッドシステムの精度をわずかではあるが上回った．先行研究の精度よりも高い精度を得たので，本研究によりタイ語の品詞タグづけの技術は向上したことになり，本研究はこの方面の学問および技術の向上に貢献したことになる．

2 品詞のタグづけの問題

本研究では，文からの単語への分割は扱わない．品詞のタグづけを行なうまでに文からの単語への分割は終了していることを前提とする[1]．この場合，文は以下の式で表される．

$$S = (w^1, w^2, \cdots, w^n) \tag{1}$$

ただし，w^i はその文の i 番目の単語である．品詞タグづけの問題は各単語に対して品詞のタグを付与することである．このため，品詞タグづけの結果は以下の式で表される．

$$T = (t^1, t^2, \cdots, t^n) \tag{2}$$

ただし，t^i は単語 w^i の品詞のタグである．本研究の目的は各単語にこのような品詞のタグを付与することである．品詞のタグの分類はあらかじめ定義しておく．このため，品詞のタグづけの問題は分類問題とみなすことができ，機械学習手法によって扱うことができるようになる．

3 機械学習手法

本稿では機械学習手法としては，以下の三つの方法を利用した[2]．

- 決定リスト法
- 最大エントロピー法
- サポートベクトルマシン法

本節ではこれらの機械学習手法の説明を行なう．

3.1 決定リスト法

これは，あらかじめ設定しておいた素性 $f_j (\in F, 1 \leq j \leq k)$ のうちいずれか一つのみを文脈として各分類の確率値を求めてその確率値が最も大きい分類を求める分類とする方法である[10]．ある文脈 b で分類 a を出力する確率は以下の式によって与えられる．

$$p(a|b) = \tilde{p}(a|f_{max}) \tag{3}$$

ただし，f_{max} は以下の式によって与えられる．

$$f_{max} = argmax_{f_j \in F} \ max_{a_i \in A} \ \tilde{p}(a_i|f_j) \tag{4}$$

また，$\tilde{p}(a_i|f_j)$ は学習データで素性 f_j を文脈に持つ場合の分類 a_i の出現の割合である．決定リストに基づく方法は簡便ではあるが，ある一つの素性のみを文脈として推定を行なうので，機械学習の手法としては少々貧弱なものとなっている．

[1] タイ語は日本語と同じく膠着語であり，形態素解析としては品詞タグづけの問題の他に単語分割の問題もある．本研究では単語分割の問題は扱わない．単語分割の問題も扱う場合は，単語の辞書引きなどを行なってあらゆる分割の可能性をあげて，一文全体での品詞タグづけや単語分割の確率もしくは尤度が最も高くなるようにビタビ探索を行なうとよいだろう．本研究はこのビタビ探索の際に用いることになるであろう，一構成要素の品詞タグづけの部分に着目したものとなっている．本稿でのアプローチは機械学習手法を用いるものなので，確率もしくは尤度のようなものは解析結果とともに与えられるために，ビタビ探索の一構成要素として容易に用いることができる．

[2] 機械学習手法としては，他に C4.5 などの決定木学習を利用する方法があるが，本稿では，種々の問題で決定木学習手法が他の手法に比べて劣っていること[7, 8, 9]，また，本稿で扱う問題は属性の種類の数が多く C4.5 が走るまで属性の数を減らすと精度が落ちるであろうことの二つの理由により，用いていない．

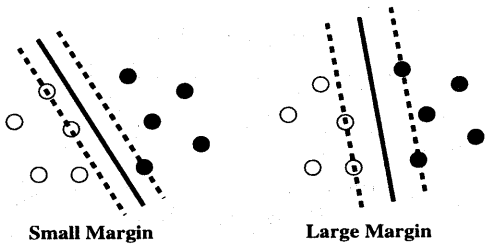

図 1: マージン最大化

3.2 最大エントロピー法

これは，あらかじめ設定しておいた素性 $f_j(1 \leq j \leq k)$ の集合を F とするとき，式 (5) を満足しながらエントロピーを意味する式 (6) を最大にするときの確率分布 $p(a,b)$ を求め，その確率分布にしたがって求まる各分類の確率のうち，もっとも大きい確率値を持つ分類を求める分類とする方法である[11].

$$\sum_{a\in A,b\in B} p(a,b)g_j(a,b) = \sum_{a\in A,b\in B} \tilde{p}(a,b)g_j(a,b) \quad (5)$$
$$for\ \forall f_j\ (1 \leq j \leq k)$$

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

ただし，A,B は分類と文脈の集合を意味し，$g_j(a,b)$ は 文脈 b に素性 f_j があってなおかつ分類が a の場合 1 となりそれ以外で 0 となる関数を意味する．また，$\tilde{p}(a,b)$ は，既知データでの (a,b) の出現の割合を意味する．

式 (5) は確率 p と出力と素性の組の出現を意味する関数 g をかけることで出力と素性の組の頻度の期待値を求めることになっており，右辺の既知データにおける期待値と，左辺の求める確率分布に基づいて計算される期待値が等しいことを制約として，エントロピー最大化 (確率分布の平滑化) を行なって，出力と文脈の確率分布を求めるものとなっている．

3.3 サポートベクトルマシン法

サポートベクトルマシン法は，空間を超平面で分割することにより 2 つの分類からなるデータを分類する手法である．このとき，2 つの分類が正例と負例からなるものとすると，学習データにおける正例と負例の間隔 (マージン) が大きいもの (図 1参照[3]) ほどオープンデータで誤った分類をする可能性が低いと考えられ，このマージンを最大にする超平面を求めそれを用いて分類を行なう．基本的には上記のとおりであるが，通常，学習データにおいてマージンの内部領域に少数の事例が含まれてもよいとする手法の拡張や，超平面の線形の部分を非線型にする拡張 (カーネル関数の導入) がなされたものが用いられる．この拡張された方法は，以下の識別関数を用いて分類することと等価であり，その識別関数の出力値が正か負かによって二つの分類を判別することができる[12, 13].

$$f(\mathbf{x}) = sgn\left(\sum_{i=1}^{l} \alpha_i y_i K(\mathbf{x}_i,\mathbf{x}) + b\right) \quad (7)$$

$$b = -\frac{max_{i,y_i=-1} b_i + min_{i,y_i=1} b_i}{2}$$

$$b_i = \sum_{j=1}^{l} \alpha_j y_j K(\mathbf{x}_j,\mathbf{x}_i)$$

ただし，$\mathbf{x}$ は識別したい事例の文脈 (素性の集合) を，$\mathbf{x}_i$ と $y_i(i = 1,...,l, y_i \in \{1,-1\})$ は学習データの文脈と分類先を意味し，関数 sgn は，

$$sgn(x) = \begin{array}{ll} 1 & (x \geq 0) \\ -1 & (otherwise) \end{array} \quad (8)$$

であり，また，各 α_i は式 (10) と式 (11) の制約のもと式 (9) の $L(\alpha)$ を最大にする場合のものである．

$$L(\alpha) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{l} \alpha_i\alpha_j y_i y_j K(\mathbf{x_i},\mathbf{x_j}) \quad (9)$$

$$0 \leq \alpha_i \leq C\ (i = 1,...,l) \quad (10)$$

$$\sum_{i=1}^{l} \alpha_i y_i = 0 \quad (11)$$

また，関数 K はカーネル関数と呼ばれ，様々なものが用いられるが本稿では以下の多項式のものを用いる．

$$K(\mathbf{x},\mathbf{y}) = (\mathbf{x}\cdot\mathbf{y}+1)^d \quad (12)$$

C,d は実験的に設定される定数である．本稿ではすべての実験を通して C,d はそれぞれ 1 と 2 に固定した．ここで，$\alpha_i > 0$ となる $\mathbf{x}_i$ は，サポートベクトルと呼ばれ，通常，式 (7) の和をとっている部分はこの事例のみを用いて計算される．つまり，実際の解析には学習データのうちサポートベクトルと呼ばれる事例のみしか用いられない．

サポートベクトルマシン法は分類の数が 2 個のデータを扱うもので，通常これにペアワイズ手法を組み合わせて用いることで，分類の数が 3 個以上のデータを扱うことになる[14].

ペアワイズ手法とは，N 個の分類を持つデータの場合，異なる二つの分類先のあらゆるペア (N(N-1)/2 個) を作り，各ペアごとにどちらがよいかを 2 値分類器 (ここではサポートベクトルマシン法[4]) で求め，最終的に

[3] 図の白丸，黒丸は，正例，負例を意味し，実線は空間を分割する超平面を意味し，破線はマージン領域の境界を表す面を意味する．

N(N-1)/2 個の 2 値分類器の分類先の多数決により，分類先を求める方法である．

本稿のサポートベクトルマシン法は，上記のようにサポートベクトルマシン法とペアワイズ手法を組み合わせることによって実現される．

4 素性 (解析に用いる情報)

前節で種々の機械学習の説明を述べたが，それぞれの手法ともに素性 (解析に用いる情報) を定義しなければ，その手法を用いることができない．本節ではその素性の説明を行なう．

2節の問題設定で述べたように，本稿の問題設定では，タイ語の文の単語分割の結果が入力として与えられたときに，その入力中の各単語に対して，その単語の品詞の分類を推定して出力することになっている．このため，解析に用いる情報，すなわち，素性は入力されるタイ語の文から取り出すことになる．

本稿では素性としては以下のものを定義する．

- **品詞情報**

  現在品詞を推定している単語と，その前の三つの単語とその後ろの三つの単語の，とりうる品詞[5](例：「名詞」「動詞」．5節で述べるタイ語コーパスでの素性の総数は 316 個．)

  ここで，各単語がとりうる品詞は単語辞書もしくはコーパスによりあらかじめ単語ごとに決められる．このとき，どの箇所の単語のとりうる品詞であるかの区別は行なう．

- **品詞順位情報**

  現在品詞を推定している単語と，その前の三つの単語とその後ろの三つの単語の，とりうる品詞およびその品詞の現在品詞を推定している単語における出現順位[6](例：「名詞 1 位」「動詞 2 位」．素性の総数は 782 個．)

  このとき，どの箇所の単語のとりうる品詞であるかの区別は行なう．

- **単語情報**

  現在品詞を推定している単語と，その前の三つの単語とその後ろの三つの単語 (例：「明日」「行く」．素性の総数は 15,763 個．)

  ここで，各単語ともどの箇所のものかの区別は行なう．

表 1: 実験結果

手法	精度
ベースライン	83.6%
HMM	89.1%
Brill の誤り駆動学習	93.5%
伸縮性 NN	94.4%
ハイブリッド型 NN	95.5%
決定リスト	83.6%
最大エントロピー法	95.3%
サポートベクトルマシン	96.1%

(品詞に曖昧性のある単語のみを測定対象とした．)

表 2: 単語情報を削った場合の実験結果

手法	精度
決定リスト	78.0%
最大エントロピー法	92.3%
サポートベクトルマシン	93.9%

(品詞に曖昧性のある単語のみを測定対象とした．)

5 実験

本節では，3節で述べた問題設定で，4節で述べた素性で3節で述べた機械学習手法を用いて行なった，タイ語の品詞タグづけの実験について述べる．

本研究の実験は文献[1, 2, 3, 4, 5, 6] で用いられているタイ語コーパスを用いて行なった．このコーパスは 10,452 文からなるコーパスで，学習用の 8,322 文とテスト用の 2,130 文にランダムに分割されている．また，学習セットとテストセットの文に含まれる品詞の曖昧な単語 (つまり，品詞タグづけの対象となる単語) の数は，それぞれ 22,311 個と 6,717 個である[7]．ここでいう品詞の曖昧な単語とは，複数の品詞の分類をとりうる単語のことを意味し，これ以外の単語については常に同じ品詞をとるためにわざわざ品詞タグづけを機械学習などで扱う必要も

[4] 本稿の 2 値分類器としてのサポートベクトルマシンは，工藤氏が作成した TinySVM[13] を利用している．

[5] 一般に現在処理している単語に対して前方に出現している単語については解析が終わっているはずなので，その単語については「とりうる品詞」のように可能な品詞でなく，実際にその文脈で使われている一つの品詞の情報を用いるだけでよい．実際，実験の節で述べる先行研究の手法はすべて前方での解析結果の品詞を使っている．しかし，本論文では処理の簡便さのためと，前方の解析自体誤る可能性があることを考慮し，前方の品詞も可能な品詞の情報を用いている．

[6] HMM などの統計的方法や先行研究の馬らの研究では，各単語がある品詞になりうる確率 (その単語での品詞の出現確率) も解析に用いている．しかし，本稿で扱うような素性を用いた機械学習 (決定リスト法や最大エントロピー法) の場合は，その素性の部分に確率のような連続値を用いるのは困難である．このため，出現確率のかわりに本稿では出現順位を用いている．順位の情報はたかだか品詞の曖昧性の数でしかなく大きな値はとらないために，本論文で用いるような機械学習の手法でも十分扱える．なお，本稿で用いる機械学習のうち，サポートベクトルマシン法は素性に実数を扱うことができる．本論文では決定リスト法と最大エントロピー法との比較のため，サポートベクトルマシン法でも出現確率でなく出現順位を用いるが，今後サポートベクトルマシン法で出現順位でなく出現確率を用いることも試す必要がある．

[7] 品詞に曖昧性のない単語も含めた総単語数はそれぞれ 124,331 個と 34,544 個である．

なく辞書引きをしてその品詞を割り当てるだけよい．このタイ語コーパスでは 47 個の品詞の分類が定義されている[15]．

実験結果を表 1に示す．表の「ベースライン」「HMM」「Brill の誤り駆動学習」「伸縮性 NN(伸縮性ニューラルネットワーク)」「ハイブリッド型 NN(ハイブリッド型ニューラルネットワーク)」の精度は先行文献[1, 2] のものを記述している．「ベースライン」は，文脈情報を用いず学習用のコーパスから得られた各単語でもっとも出現頻度の高い品詞を求める品詞として出力する方法である．「HMM」は隠れマルコフモデル (Hidden Markov Model) の略で，品詞を内部状態とする隠れマルコフモデルを用いて文レベルで品詞をタグづけする手法である．詳しくは文献[1] を参照のこと．「Brill の誤り駆動学習」は，文献の Brill の書き換え規則を学習コーパスから学習してそれを用いて解析する方法を意味する．「伸縮性 NN」は文献[1] で提案された方法であり，入力の文脈を可変にした 3 階層のパーセプトロンを用いたニューラルネットワークによる手法である．「ハイブリッド型 NN」は文献[2] で提案された方法であり，「Brill の誤り駆動学習」と「伸縮性 NN」を組み合わせた方法である．まず伸縮性 NN で解析を行ないその後 Brill の誤り駆動学習を行ないさらに精度をあげる手法となっている．「ハイブリッド型 NN」は本実験で用いたコーパスを用いた先行研究ではもっとも精度の高いものである．他の三つの手法 (決定リスト，最大エントロピー法，サポートベクトルマシン) の精度は本研究で行なった実験結果である．

本研究で用いた三つの手法の中ではサポートベクトルマシン法の精度 (96.1%[8]) がもっともよかった．この結果はわれわれの他の研究[16, 17] とも矛盾しない．また，サポートベクトルマシン法の精度は，先行研究のうちでもっとも精度のよかった「ハイブリッド型 NN」のもの (95.5%) よりも高い．先行研究の精度よりも高かったので，本研究によりタイ語の品詞タグづけの技術は向上したことになり，本研究はこの方面の学問および技術の向上に貢献したことになる．

次に手法間の考察を行なう．まず本稿で扱った三つの手法であるが，これらはそれぞれ全く同じ素性を用いたため，比較は厳密であり，本研究で扱った問題では手法の優劣は以下のようになっていると思われる．

サポートベクトル > 最大エントロピー
> 決定リスト

また，決定リストの精度は極めて悪くベースラインとほぼ同じ程度であった．これは今回の実験ではシステムの入力としては AND 素性 (素性の組) を用いなかったためで，ある一つの素性のみで推定する決定リスト法には不利な実験条件であったともいえる．AND 素性を利用すると決定リスト法でも少しは精度があがると思われる[9]．しかし，AND 素性としては無作為に作成すると素性の数が爆発的に増える問題がある．また少数の素性の組を AND 素性として追加するにしてもどのような素性の組を追加するか十分に検討する必要がある．これに対し，サポートベクトルマシン法と最大エントロピー法の場合は全素性を用いて推定するものになっている．また，サポートベクトルマシン法ではカーネル関数の係数 d を調節することで自動で AND 素性を考慮する枠組みになっている．AND 素性のことをわざわざ人手で考慮する必要がないという意味でもサポートベクトルマシン法は有効な機械学習手法であると考えられる．

次に先行研究の手法との比較を行なう．本研究と先行研究とでは用いた素性が一部一致しないところがあるために手法間の比較は慎重に行なう必要がある．まず，「Brill の誤り駆動学習」と「ハイブリッド型 NN」だが，これらは誤り駆動学習の際の規則のテンプレートにおいて品詞の情報だけでなく，単語の情報をすべて用いている．このため，本研究で用いた情報とほとんど同じものを使っていると考えてよい．このため，これらの方法は本研究の方法と比較可能であり，本研究の問題ではこの二つも含めた主要な学習アルゴリズム間の優劣は以下のようになっているということができる[10]．

サポートベクトル > ハイブリッド
> 最大エントロピー > Brill

次に「HMM」と「伸縮性 NN」だが，これらは単語情報は直接は用いていない．間接的に各単語でのある品詞が出現する出現確率を用いている程度である．ここで本稿でもこれらの手法とよくにた状況を作るために単語情報の素性を減らした実験を行なった．この結果を表 2にあげる．表のように単語情報を用いていたときよりも各手法ともに精度が下がっている．「伸縮性 NN」と単語情報なしの「サポートベクトルマシン法」を比較すると，「伸縮性 NN」が 94.4% で「サポートベクトルマシン

[8] 本稿であげる精度は多品詞語のみを測定対象とした場合のものである．一つの品詞しかとらない曖昧性のない単語も測定対象とした場合は 99.2% になる．

[9] 文節まとめ上げの問題で AND 素性を用いることで決定リストでも高い精度をだすことを示したものとして文献[8] がある．本稿では AND 素性を用いないために決定リスト法は精度は悪いが，用いる場合は精度がよくなる場合もあることに注意していただきたい．

[10] 厳密には「ハイブリッド型 NN」では AND 素性のようなものを用いており，最大エントロピー法でも AND 素性を用いると精度があがる可能性があるため，「ハイブリッド型 NN」と最大エントロピー法の順位は交替する可能性がある．

法」が 93.9% となり，「伸縮性 NN」の方が性能がよいということになる．しかし，厳密には「伸縮性 NN」では各単語でのある品詞が出現する出現確率を用いており，また「サポートベクトルマシン法」ではこれにかわる情報として「品詞順位情報」を用いているが，「品詞順位情報」の方が品詞の出現確率よりも情報が少ないので厳密な比較にはなっていない．とはいえ，この結果は「伸縮性 NN」の手法も「サポートベクトルマシン法」と同様に性能の高いことを推測させる[11]．「HMM」については単語情報を用いない「サポートベクトルマシン法」や「最大エントロピー法」よりも精度がかなり低いため，これらの手法よりは性能の低い手法であるということができる．

最後に本研究で精度向上を実現できた理由について考察する．本研究ではサポートベクトルマシン法を用いることで，タイ語の品詞タグづけの問題で先行研究の諸手法よりも高い精度を得た．HMM と 伸縮性 NN よりも高い精度を得たのは，単語情報を追加で用いたためであると思われる (HMM と 伸縮性 NN では上述のように単語情報を用いない．)．前後の文脈にこの単語があれば品詞はこれだという場合もあるであろうから単語情報が効く場合も多いことだろう．次に Brill の誤り駆動学習とハイブリッド型 NN との比較であるが，解析に用いる情報がほとんど同じであったことから手法自体がこれらの手法よりも性能がよいということが予想される．また，ハイブリッド型 NN は Brill の誤り駆動学習も一部含むために Brill の誤り駆動学習が性能が悪い場合ハイブリッド型 NN の性能も下がると思われる．本研究では単語情報を用いなおかつ性能のよいサポートベクトルマシン法を用いたことでよい精度をあげたものと思われる．今後の課題としては伸縮性 NN の手法で単語情報を用いてその精度がどれくらいになるかを試すなどのことが必要と思われる．

6 おわりに

本稿では，タイ語の品詞タグづけの問題を教師あり機械学習手法を用いて扱った．タイ語の品詞タグづけのための教師ありデータとしては文献[5, 6, 2] のコーパスのものを用いた．機械学習手法としては，決定リスト法，最大エントロピー法，サポートベクトルマシン法を試した．この結果，サポートベクトルマシン法が最もよい精度を得た．また，このサポートベクトルマシン法の精度は先行研究のパーセプトロン型のニューラルネットワークと Brill の誤り駆動学習のハイブリッドシステムの精度をわずかではあるが上まわった．また，先行研究を含めて各学習手法の比較と考察を行なった．そのなかでは，サポートベクトルマシン法の性能の良さを議論した．また，先行研究の一部である伸縮性 NN も性能が高い学習アルゴリズムと予想させる結果もあった．しかし，伸縮性 NN では単語情報を用いておらず本稿のサポートベクトルマシン法のものよりも精度が低かった．本研究は基本的に，単語情報を用いなおかつ性能のよいサポートベクトルマシン法を用いたことでよい精度をあげたものと思われる．今後の課題としては伸縮性 NN の手法で単語情報を用いてその精度がどれくらいになるかを試すなどのことが必要と思われる．

[11] ここでは素性が異なる手法を比較する議論を行なったが，素性を同じにして手法の比較をする実験を実際に行なった方がよいことはいうまでもない．

参考文献

(1) 馬青, 内元清貴, 村田真樹, 井佐原均, 品詞自動タグづけシステム — 伸縮性入力ニューロタガー —, 人工知能学会誌, Vol. 14, No. 6, (1999).

(2) 馬青, 内元清貴, 村田真樹, 井佐原均, ニューラルネットとルール手法を統合した品詞タグ付けシステム, 情報処理学会論文誌, Vol. 41, No. 10, (2000).

(3) Qing Ma, Kiyotaka Uchimoto, Masaki Murata, and Hitoshi Isahara, A multi-neuro tagger using variable lengths of contexts, *17th International Conference on Computational Linguistics (COLING-ACL'98)*, (1998), pp. 802–806.

(4) Qing Ma, Kiyotaka Uchimoto, Masaki Murata, and Hitoshi Isahara, Elastic neural networks for part of speech tagging, *IJCNN'99*, (1999).

(5) Qing Ma, Masaki Murata, Masao Utiyama, Kiyotaka Uchimoto, and Hitoshi Isahara, Part of speech tagging with mixed approaches of neural networks and transformation rules, *NLPRS'99 Workshop on Natural Language Processing and Neural Networks*, (1999).

(6) Qing Ma, Masaki Murata, Kiyotaka Uchimoto, and Hitoshi Isahara, Hybrid neuro and rule-based part of speech taggers, *Proceedings of the 18th International Conference on Computational Linguistics (COLING'2000)*, (2000).

(7) 村田真樹, 内元清貴, 馬青, 井佐原均, 学習による文節まとめあげ — 決定木学習，最大エントロピー法，用例ベースによる手法と排反な規則を用いる新手法の比較 —, 情報処理学会 自然言語処理研究会 NL128-4, (1998).

(8) 村田真樹, 内元清貴, 馬青, 井佐原均, 排反な規則を用いた文節まとめあげ, 情報処理学会論文誌, Vol. 41, No. 1, (2000), pp. 59–69.

(9) 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, Vol. 41, No. 4, (2000), pp. 1113–1123.

(10) David Yarowsky, Decision lists for lexical ambiguity resolution: Application to accent restoration in Spanish and French, *32th Annual Meeting of the Associtation of the Computational Linguistics*, (1994), pp. 88–95.

(11) Eric Sven Ristad, Maximum Entropy Modeling for Natural Language, (ACL/EACL Tutorial Program, Madrid, 1997).

(12) Nello Cristianini and John Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*, (Cambridge University Press, 2000).

(13) Taku Kudoh, Tinysvm: Support vector machines, (http://cl.aist-nara.ac.jp/ taku-ku//software/TinySVM/index.html, 2000).

(14) 工藤拓, 松本裕治, Support vector machine を用いた chunk 同定, 自然言語処理研究会 2000-NL-140, (2000).

(15) Thatsanee Charoenporn, Virach Sornlertlamvanich, and Hitoshi Isahara, Building a large thai text corpus - part-of-speech tagged corpus: Orchid -, *NLPRS'97*, (1997).

(16) 村田真樹, 馬青, 内元清貴, 井佐原均, サポートベクトルマシンを用いたテンス・アスペクト・モダリティの日英翻訳, 電子情報通信学会 言語理解とコミュニケーション研究会 NLC2000-78, (2001).

(17) 村田真樹, 内山将夫, 内元清貴, 馬青, 井佐原均, 種々の機械学習手法を用いた多義解消実験, 電子情報通信学会言語理解とコミュニケーション研究会 NLC2001-2, (2001).