自然言語処理 141－13
（2001. 1. 25）

Support Vector Machineを用いた未知語の品詞推定

中川 哲治　　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

〒630-0101 奈良県生駒市高山町 8916-5

{tetsu-na,taku-ku,matsu}@is.aist-nara.ac.jp

品詞タグ付けにおいて，辞書に存在する語と比較して未知語に対する解析精度は低く，品詞タグ付けの全体の性能を左右する要因となる。本稿では，英語における未知語の品詞をSupport Vector Machineを用いて推定する手法を提案する。単語の部分文字列やコンテキストを素性として，高い精度で品詞を推定することができた。さらに同じ枠組みで，SVMを用いて品詞のタグ付けを行ったところ，100,000トークンの訓練データによる学習の結果，95.9%という従来手法以上の精度が得られた。

キーワード: 未知語, 品詞タグ付け, 機械学習, Support Vector Machine

Unknown Word Guessing with Support Vector Machines

Tetsuji Nakagawa　　Taku Kudoh　　Yuji Matsumoto

Graduate School of Infomation Science, Nara Institute Science and Technology

8916-5 Takayama, Ikoma, Nara 630-0101 Japan

{tetsu-na,taku-ku,matsu}@is.aist-nara.ac.jp

In part-of-speech tagging, the accuracy for unknown words is lower than for known words. In consideration of high accuracy rate of up-to-date statistical POS taggers, unknown words account for non-negligible cause of errors. This paper describes POS estimation for unknown words using Support Vector Machines. We achieved high accuracy of POS tag estimation using substrings and surrounding contexts as feature. Furthermore, we apply this method to a practical POS tagger, and achieved the accuracy of 95.9%, higher than conventional approaches.

Keywords : Unknown Word, POS tagging, Machine Learning, Support Vector Machines

1 はじめに

英語等における品詞タグ付けの際，辞書中には存在しない単語がしばしば出現する。そのような未知語に対しては，単語の出力確率などの統計情報がないために特別な処理を行わなければならない。品詞タグ付けの方法は，HMM モデル [3]，変換規則に基づく誤り駆動方式 [2]，決定木 [13]，最大エントロピー法 [11] など様々な方法が試みられており，非常に高い精度を達成できるようになっている。しかし，未知語に対する精度は辞書中に存在する単語に比べて低く，テストデータに対して十分な訓練データが無い場合には品詞タグ付けの精度低下の大きな原因になっている [1]。

未知語の品詞を推定する手法として，単語の語頭や語尾，前後の単語等のコンテキストを使う方法がよく知られている [9, 14]。Weischedel[16] らはある未知語 w が品詞 t である確率を，単語の語尾，ハイフンの有無，大文字の有無を素性として，次のような式で推定している

$$\begin{aligned} p(w|t) &= p(unknown\ word|t) \cdot p(capital|t) \\ &\quad \cdot p(endings/hypenation|t) \end{aligned}$$

この方法は，未知語の処理を確率的な枠組みの中で行うことができるが，単語の語尾に関しては，“-ed”，“-ion” などの特定の語尾を人手で選択しているため，英語以外の言語への適用を考える場合に問題となる。Brants は，品詞タグ付けプログラム TnT の未知語処理に，異なる長さの語尾の条件付き確率を線形補間したものを使用している。この方法は比較的精度が高く，英語以外の言語に対しても有効であることが報告されている [6]。Cucerzan ら [5] は語形変化の類似度の尺度を導入した方法を提案しており，この方法は語形変化する言語に対して大量のタグなしコーパスが利用できる場合に高い精度を得ている。他にも，規則に基づく方法 [8] などが試みられている。

本稿では，英語における未知語の品詞推定を，品詞タグ付けの後処理として行う場合を考え，Support Vector Machine(SVM) [4, 15] を用いて行う。SVM は二値分類を行う教師付き学習アルゴリズムで，高い汎化能力を持つことが知られている。多量の素性を扱うことができ過学習を起こしにくいことから自然言語処理にも応用され，その有効性が報告されている [18, 7]。SVM では素性の組合わせを自動的に展開して学習を行えるなどの特徴も持ち，未知語の品詞推定に用いることで高い精度が得られることが期待できる。未知語に対する品詞の推定を一般化して考えると，通常の品詞タグ付けも同じ枠組みで考えることができる。そこで，本稿では SVM を用いた品詞タグ付けの実験も行い，その結果についても報告する。

以下，2 節では SVM を用いた未知語の品詞推定と品詞タグ付けの手法を述べる。3 節で実験結果を示して考察を行い，4 節でまとめと今後の課題を述べる。

2 SVMによる未知語の品詞推定と品詞タグ付け

2.1 Support Vector Machine

Support Vector Machine(SVM) は二値分類を行う教師付き学習モデルである。L 次元の素性ベクトル $\mathbf{x} \in \mathbf{R}^L$ に対し，

$$\mathbf{w} \cdot \mathbf{x} + b = 0 \qquad \mathbf{w} \in \mathbf{R}^L, b \in \mathbf{R} \tag{1}$$

で表わされる分離平面を考え，+1(正例)，−1(負例) の値をとる y_i をラベルとして持つ訓練事例 $\mathbf{x}_i$ を次のように分類するとする。

$$y_i \cdot (\mathbf{w} \cdot \mathbf{x}_i + b) > 0 \tag{2}$$

このような分離平面は一般に無数存在するが (図 1 左)，SVM では分離平面とそれに最も近い事例間の距離 (マージン) を最大にするような $\mathbf{w}$ と b を求める (図 1 右)。このような平面はテストデータに対するエラーの期待値を最小にすることが示されており，素性の次元が大きい場合でも高い汎化能力を持つことができる。テスト事例 $\mathbf{x}$ が与えられた場合，次のような決定関数 $f(\mathbf{x})$ を使い，この値の正負によって事例のラベル y を決定する。

$$f(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x} + b \tag{3}$$

$$y = \mathrm{sgn}(f(\mathbf{x})) \tag{4}$$

線形分離不可能な場合を扱うために，素性ベクトルをより高次元の空間に写像し，そこで線形分離を行うことを考える。SVM におけるモデルの学習，事例の分類には，特徴ベクトル同士の内積しか使わないという性質があるため，2 つの特徴ベクトルに対し，これらの高次元空間上における写像の内積を与える関数を使用すれば，高次元空間を意識せずに計

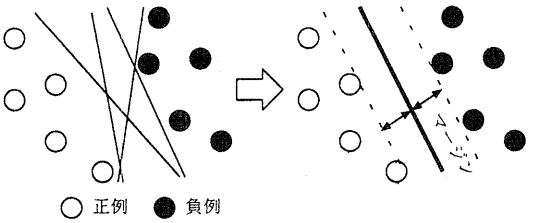

図 1: マージン最大化 (二次元空間における例)

表 1: 未知語の品詞推定における素性の例

品詞	t_{-1} =TO, t_{-2} =VBD, t_{-3} =PRP, t_{+1} =CD, t_{+2} =NNS, t_{+3} =IN
単語	w_{-1} =to, w_{-2} =returned, w_{-3} =she, w_{+1} =two, w_{+2} =days, w_{+3} =before
表層	^g, ^gr, ^gre, ^gree, e$, le$, lle$, ille$ 大文字

算を行うことができる。このような関数を kernel 関数と呼び，kernel 関数を使用することで素性ベクトルを実際に高次元へ写像することなく効率的に計算を行うことができる。kernel 関数はいくつかのものが知られているが，ここでは次のような式で定義される polynomial kernel を使用した。

$$K(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i \cdot \mathbf{x}_j + 1)^d \tag{5}$$

これは，各素性の d 個までの組み合わせを考えた ${}_{L+d}C_d$ 次元の空間で線形分離を行うことに相当する。

SVM は二値分類器であるため，そのままでは複数の品詞候補の中から一つを決定するようなマルチクラスの問題を扱うことができない。SVM でマルチクラス問題を扱う方法はいくつか考えられている [17] が，ここでは one-versus-rest と呼ばれる方法を用いる。これは k 個のクラスが存在した場合，学習時に一つのクラスとそれ以外の $(k-1)$ 個のクラスとを分類するような二値分類器を合計 k 個作り，テスト時は，テスト事例に対する決定関数の値 (分離平面からの距離) が最も大きくなるようなクラスを選択する方法である。

上で述べた多項式カーネルを使う場合，d というパラメータの自由度があるため，適切な値を選ぶ必要があるが，ここではテストデータに対するエラー率の期待値の上限を与える次の値を使う。

$$R_{emp} + \sqrt{\frac{h(\ln \frac{2l}{h} + 1) - \ln \frac{\eta}{4}}{l}} \tag{6}$$

R_{emp} は訓練データに対するエラー率，h はモデルの VC 次元，l は訓練データ数，η は危険率であり，この値は VC bound と呼ばれる。VC 次元はモデルのキャパシティに影響し，この値が大きくなれば複雑な訓練事例を分離できるようになるが，その事例に対して過学習する可能性も高くなる。ここではマルチクラスの SVM を使うため，k 個の分類器の VC bound の平均値をとるというヒューリスティクスを使った。

2.2 未知語の品詞推定

品詞の推定に使う素性は，前後 3 つの品詞，前後 3 つの単語，未知語の 4 文字以下の語頭と語尾，未知語が数字・大文字・ハイフンをそれぞれ含むかどうかを使用した。例として，

`she/PRP returned/VBD to/TO Greenville/未知語 two/CD days/NNS before/IN`

という文脈での未知語 `Greenville` に対する素性は表 1 のようになる。

このような素性を使い，訓練データ中の全ての単語を事例として各品詞に対する分類器を作り，未知語の品詞推定を行う。

2.3 品詞タグ付け

未知語の品詞推定では，前後の品詞が既知なものとして SVM により単語の品詞を求めるが，与えられた文の全ての品詞を求めることを考えれば品詞のタグ付けを行うことができる。ここでは以下のような 2 つの方法を試みた。

2.3.1 前方の品詞のみを使用する方法

一つめの方法では，注目している単語の前方の品詞のみを使用する。素性として，前後 2 つの単語，前 2 つの品詞，4 文字以下の語頭と語尾，大文字・数字・ハイフンをそれぞれ含むかどうかを用いた。未知語の品詞推定とは異なり前後 2 つまでの文脈しか参照

しないが，これはRatnaparkhiが最大エントロピー法による品詞タグ付けにおいて，未知語に対して使用した素性に合わせたものである[1]。

評価時の解析について，HMM等の確率を使ったモデルでは，一つの品詞系列に対してその生起確率を計算することができるため，Viterbiアルゴリズムにより最尤パスを求めることができる。一方SVMでは，そのままでは確率的な枠組みで使うことはできない。確率を扱うための手法が提案されてはいるが[10]，理論的な保証がないことと計算量を考え，先頭の単語から順に品詞を決定していくという方法をとることにした。この方法は，計算量が少ないという利点はあるが，後方の単語の品詞に関する情報を利用しないという問題がある。

解析時には，訓練データ中に現われた単語に対しては探索を絞り込むために辞書を用いて品詞の候補を与え，未知語に対しては全ての品詞を候補として与えた。この点を除けば，単語の出力確率等は使用せず語頭などの単語の表層情報しか使用していないため，未知語に対する特別な処理は行わずに解析を行うことができる。

2.3.2 前後の品詞を使用する方法

もう一つの方法として，前後の品詞を使用する。素性として，前後2つの単語，前後2つの品詞，4文字以下の語頭と語尾，大文字・数字・ハイフンをそれぞれ含むかどうかを用いた。

後方の品詞も素性とする場合，解析中の単語よりも先にある品詞を利用することになるが，それらは一般に未知である。Rothら[12]は，訓練データ中に現われた各単語に対して最も多かった品詞を辞書に記録しておき，その品詞を使って前後の品詞を使用した品詞タグ付けを行っている。その場合誤って付与した品詞がノイズとなり，正しい品詞が付与された場合に比べて精度が2%弱低下している。ここでは，次のような2つのパスからなる方法を用いた。まず1パス目では前後の品詞情報を全く使用せずに，前後の単語と表層情報だけから各単語の品詞を推定する。2パス目では，後方にある品詞については1パス目で推定した品詞を用いて，文の先頭から品詞を決定する。この方法はRothらの手法と比べて，前後の単語を考慮することや未知語に対しても特別な処理を必要としないなどの利点がある。

この方法でも，前方の品詞のみを使用する場合と同様に辞書を使用した。

[1]Ratnaparkhiは2つ前と1つ前の品詞の組を素性としているが，SVMではカーネル関数により素性の組合わせを考慮した学習が行えるためこれらは独立に扱った。

3 実験と考察

3.1 実験条件

実験には，Penn Tree-bankコーパスを使用した(品詞の数は50)。訓練用データは文単位でランダムに抽出し，約1000，10000，100000，1000000トークンのものを用意した。

未知語の品詞推定のための評価用データは，訓練データに現われなかった単語を使用した。単語の語尾を使って未知語処理を行うTnTを使った場合の結果と比較した。また，未知語の前後の品詞はTnTによってタグ付けしたものを使用した。

品詞タグ付けのための評価用データは，コーパスの全データから訓練データを除いた残りの約285000トークンを使用し，HMMによる品詞タグ付けを行うTnTの結果と比較した。

3.2 未知語の品詞推定

学習データ量を変えた場合の未知語に対する品詞推定の結果を表2に示す。ここで，SVMは本手法による精度，dはそのときのpolynomial kernelの次数，suffixはTnTによる精度を表わす。1,000,000トークンの場合，SVMによる方法は計算に時間がかかり過ぎるため，学習データに100,000トークンのみを使用して，1,000,000トークンの場合のテストデータの評価を行ったものである。また，未知語の前後の品詞にコーパス中の正しいものを与えた場合と，前後の品詞，前後の単語，未知語の表層のそれぞれの素性を使わなかった場合の精度を表3に示す。

これらを見ると，1,000トークンで学習した場合のSVMによる方法はあまり高い精度が得られていない。これは，10,000トークンの場合でも(VC Boundによって最適と判断した)polynomial kernelの次数が1であることから，$d = 1$のモデルは10,000トークンの訓練データに対して十分なキャパシティを持っており，学習データが1,000トークンしか無い場合には，モデルが複雑で過学習を起こしていることが考えられる。100,000トークンではsuffixに比べて高い精度

表 2: 未知語の品詞推定精度の比較

訓練トークン数	SVM	d	suffix
1,000	69.1%	1	69.4 %
10,000	81.9%	1	81.5 %
100,000	86.6%	2	83.3 %
1,000,000	(85.0%)*	(2)*	84.2 %

* 学習データは 100,000 トークンを使用

表 3: 素性を変えた場合の未知語の品詞推定

訓練トークン数	品詞正解	品詞なし	単語なし	表層なし
1,000	70.4%	64.2%	69.4%	34.7%
10,000	82.5%	75.2%	81.0%	34.0%
100,000	87.3%	79.8%	85.6%	35.0%

が得られている。さらに，本手法で 100,000 トークンを使って学習を行った方が，その 10 倍の 1,000,000 トークンを使って学習を行った suffix よりも精度が高く，SVM を用いた手法が有効であることを示している。

素性を変えた場合の結果に関して，表 3 の正解の品詞を与えた場合と表 2 の (TnT によって未知語の前後の品詞を付与された)SVM の精度を比べて，TnT の解析誤りによる精度の低下は 1%ほどしか起きていないことが分かる。また，表層情報が未知語の品詞推定に大きな影響を持っているが，前後の単語は品詞の推定にそれほど役に立っていないことが分かる。前後の品詞は未知語の品詞推定に対して大きな情報を持っているように思えるが，この結果からは表層情報ほど大きな影響は持っていないことがいえる。

3.3 品詞タグ付け

品詞タグ付けの実験結果を表 4 に示す。HMM は TnT による結果を表わす。SVM による結果では，前方の品詞のみを使った方法より，後方の品詞も使った場合が高い精度が得られている。前後の品詞を使った方法は HMM によるものと比べ，同等かそれ以上の精度が得られている。現在のところ前方の品詞のみを使った手法も前後の品詞を使った手法も，確率的方法による場合のようにパス全体の尤度を考慮することはせずにパスを決定するため，この解析方法を改良することでさらに精度の向上が得られる可能性がある。

表 4: 品詞タグ付けの精度

訓練	SVM				HMM
トークン数	前方の品詞	d	前後の品詞	d	
1,000	83.4	1	83.8	1	83.8
10,000	92.1	2	92.5	2	92.3
100,000	95.6	2	95.9	2	95.4

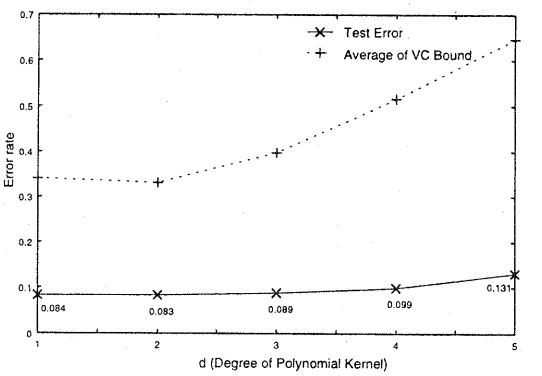

図 2: VC Bound によるテストエラーの予測

また訓練データ数が 10,000 トークンの場合の，polynomial kernel の次数 d を変えた場合のテストエラーと VC Bound の平均値との関係を図 2 に示す (前方の品詞のみを使用する方法で，テストエラーは 10,000 トークンの評価データに対する値)。これを見ると，VC Bound の平均値がテストエラーのピークをうまく予測できている。

4 まとめ

本稿では SVM を用いて未知語の品詞推定を行い，従来手法で大量のデータの学習によって実現された以上の精度を，より少数の学習データで達成できることを示した。同じ枠組みを品詞タグ付けに応用した結果，従来手法以上の精度が得られることを示した。今後の課題として，次のものが挙げられる。

(1) 英語以外の言語における未知語の品詞推定への適用

(2) 計算量の削減

(3) 素性の選択

(1) について，日本語の場合は分かち書きをしないために未知語の同定が難しく，本手法を簡単に適用することはできない。しかし，本手法は言語に依存する情報は用いていないため，ドイツ語等への適用は容易に行えると考えられる。

(2) について，SVM は最適な大域解を求めることができ，汎化能力が高く，素性が多くなっても学習時間はあまり影響されないという優れた特徴を持つ一方，訓練事例数の増加に対して計算時間は線形以上に増大するため，大量の訓練データに対しては現実的な時間で計算できないという問題がある。またモデルが複雑になると，分類にかかる計算量も大きくなる。

(3) について，SVM では polynomial kernel を使うことによって素性の組合わせを考慮した学習を行えるが不要な素性を除くことはできない。また，kernel の次数を上げればより多くの組合わせを見ることができるようになると同時にモデルが複雑になりテストエラーの期待値も増加するため，SVM を用いる場合でも素性を適切に選択する必要性がある。また，適切な素性を選択してモデルを小さくすることで，(2) の計算量の課題にもある程度対処できる。

謝辞

本研究では，品詞タグ付けプログラム TnT を使用しました。作者の Thorsten Brants に感謝します。

参考文献

[1] Thorsten Brants. TnT — A Statistical Part-of-Speech Tagger. In *Proceedings of ANLP-NAACL 2000*, pp. 224–231, 2000.

[2] Eric Brill. Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part-of-Speech Tagging. *Computational Linguistics*, Vol. 21, No. 4, 1995.

[3] Eugene Charniak, Curtis Hendrickson, Neil Jacobson, and Mike Perkowitz. Equations for Part-of-speech Tagging. In *Proceedings of the Eleventh National Conference on Artificial Intelligence*, pp. 784–789, 1993.

[4] C. Cortes and Vladimir N. Vapnik. Support Vector Networks. *Machine Learning*, Vol. 20, pp. 273–297, 1995.

[5] Silviu Cucerzan and David Yarowsky. Language Independent, Minimally Supervised Induction of Lexical Probabilities. In *Proceedings of ACL-2000*, pp. 270–277, 2000.

[6] Sašo Džeroski, Tomaž Erjavec, and Jakub Zavrel. Morphosyntactic Tagging of Slovene: Evaluating Taggers and Tagsets. In *Proceedings of LREC 2000*, 2000.

[7] Taku Kudoh and Yuji Matsumoto. Use of Support Vector Learning for Chunk Identification. In *Proceedings of the 4th Conference on CoNLL-2000 and LLL-2000*, pp. 142–144, 2000.

[8] Andrei Mikheev. Automatic Rule Induction for Unknown-Word Guessing. *Computational Linguistics*, Vol. 23, No. 3, 1997.

[9] Giorgos S. Orphanos and Dimitris N. Christodoulakis. POS Disambiguation and Unknown Word Guessing with Decision Trees. In *Proceedings of EACL '99*, pp. 134–141, 1999.

[10] John C. Platt. Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. In *Advances in Large Margin Classifiers*. MIT Press, 1999.

[11] Adwait Ratnaparkhi. A Maximum Entropy Model for Part-of-Speech Tagging. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 133–142, 1996.

[12] Dan Roth and Dmitry Zelenko. Part of Speech Tagging Using a Network of Linear Separators. In *Proceedings of the joint 17th International Conference on Computational Linguistics and 36th Annual Meeting of the Association for Computational Linguistics*, pp. 1136–1142, 1998.

[13] Helmut Schmid. Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of the International Conference on New Methods in Language Processing*, pp. 44–49, 1994.

[14] Scott M. Thede. Predicting Part-of-Speech Information about Unknown Words using Statistical Methods. In *Proceedings of COLING-ACL 1998*, pp. 1505–1507, 1998.

[15] Vladimir N. Vapnik. *The Nature of Statistical Learning Theory (Statistics for Engineering and Information Science)*. Springer, 1999.

[16] Ralph Weischedel, Marie Meeter, Richard Schwartz, Lance Ramshaw, and Jeff Palmucci. Coping with Ambiguity and Unknown Words through Probablistic Models. *Computational Linguistics*, Vol. 19, No. 2, 1993.

[17] J. Weston and C.Watkins. Support Vector Machines for Multi-Class Pattern Recognition. In *Proceedings of the Seventh European Symposium On Artificial Neural Networks*, 1999.

[18] 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, p. 1113, 2000.