語彙化マルコフモデルによる英語品詞タグ付け

森 信介　　長尾 眞
京都大学工学研究科
〒606-01 京都市左京区吉田本町
{mori,nagao}@kuee.kyoto-u.ac.jp

あらまし
本論文では、我々が提案し、日本語形態素解析に応用したマルコフモデルの語彙化を、英語の品詞タグ付けに応用した結果を報告する。マルコフモデルの語彙化における変更点は次のように要約される。 1) 各品詞の語彙化、 2) 単語列の登録。これらの改良により、品詞単位のマルコフモデルより学習コーパスに忠実なマルコフモデルが構成される。一方で、未知の入力に対する頑強性が損なわれる。この問題は、語彙化したマルコフモデルと品詞単位のマルコフモデルを重ね合わせることにより解決される。品詞タグ付コーパス Wall Street Journal に対して実験を行なった結果、非常に高い精度を得た。
キーワード 品詞タグ付け コーパス 語彙化 連語 マルコフモデル

English Part-of-Speech Tagging Using Lexicalized Markov Model

Shinsuke Mori　　Makoto Nagao
Department of Electrical Engineering, Kyoto University
Yoshida-honmachi, Sakyo, Kyoto, 606-01 Japan
{mori,nagao}@kuee.kyoto-u.ac.jp

Abstract
In this paper, we report the results of lexicalized Markov model applied to English part-of-speech tagging, which we have proposed and applied to Japanese Morphological analysis. The instructions to lexicalize Markov model are summarized as follows: 1) lexicalization of part-of-speech; 2) memorization of word sequence. This provides us with more faithful Markov models to the learning corpus. On the other hand, the lexicalized Markov models may be less robust to unknown corpora. This problem is solved by superposition of lexicalized Markov models and part-of-speech-based Markov model. We conducted experiments on Wall Street Journal corpus and obtained the considerably high accuracy.
Key Words Part-of-speech tagging, Corpus, Lexicalize, Ideom, Markov model

1 はじめに

今日までに、英語の品詞タグ付けの問題に対して、多くの研究がなされている。最近の研究のほとんどはコーパスに基づく方法であり、以下のように分類される。

1. 確率を用いる方法 [1]–[15]

この方法では、マルコフ情報源を用いて言語をモデル化する。このモデルでは、品詞が状態に対応し、各状態から単語と品詞の対が出力される。状態遷移確率や出力文字 (単語と品詞の対) の出現確率は、あらかじめコーパスから推定しておく。入力文 (単語列) が与えられると、この単語列を出力する状態遷移列 (品詞列) の中から、状態遷移確率と出現確率の積を最大にする状態遷移列を計算する。

2. 規則を用いる方法 [16]–[21]

この方法では、ある位置の単語の品詞を決める規則をあらかじめ用意しておき、入力文にこれらの規則を適用して各単語の品詞を決定する。これらの規則は、あらかじめ用意した規則のテンプレートとタグ付きコーパスから具体的に決定される。

3. ニューラルネットを用いる方法 [22]–[25]

この方法では、ある単語の品詞を決定するためにニューラルネットを用いる。具体的には、入力層のニューロンを注目している単語との相対的な位置関係と品詞に対応させ、周辺の単語がある品詞に属する確率に比例するように入力層のニューロンを活性化する。活性伝播の結果得られる出力層の活性度をもとに、注目している単語の品詞を決定する。

これらの他に、コーパスに基づいて学習した決定木を用いる方法 [26] や最初に規則を用いて解析し、曖昧性が残る場合にのみ確率を用いるという方法 [27] がある。また、文献 [28] では、規則を用いる方法と確率を用いる方法の比較を行なっている。

本論文では、我々が提案し日本語の形態素解析に応用した手法 [29] を英語の品詞タグ付けに応用し、 Wall Street Journal [30] に対して行なった実験の結果を報告する。一般的に、確率的方法では二重マルコフモデルが用いられているが、我々は単純マルコフモデルを用いた。実験の結果、本論文で提案する手法の精度は二重マルコフモデルを用いた方法と同等かそれ以上であることがわかった。本手法の中心となるアイデアは次の 3 つに要約される。

1. **状態の語彙化**

これまでのマルコフモデルによる品詞タグ付けの研究では、パラメータの数を削減し、スパースデータの問題をできる限り避けるため、品詞を状態に対応させている [31] 。しかし、品詞体系が完全でないため、異なった振舞いをする単語が同じ品詞に属することがある。この場合、コーパスの性質が正確にモデルに反映されない。この問題は、文献 [17] の中ですでに指摘されており、規則に基づく方法 [16] に対して、規則を語彙化するという解決方法を提案している。確率を用いる方法では、助動詞の細分類 [3] や前置詞の細分類 [28] を行った研究があるが、統一的な取り扱いができていない。この問題を解決するため、我々は品詞と単語の対を状態に対応させることを提案した。これにより、確率を用いる方法でも、語彙化を統一的に扱うことが可能になる。

2. **特定の単語列の登録**

単語列の中には、常に同じ品詞列として出現し、一つの単語のように振舞うものがある。単語列の登録とは、このような単語列を一つの状態に対応させることを意味する。これにより、単純マルコフモデルが、入力文によっては、より高次のマルコフモデルと同等の能力をもつことが可能になる。同じような変更として文献 [3] [8] では、冠詞の直後の形容詞と名詞に特別の状態を割り当てることを提案している。また、文献 [13] では、ダイバージェンス (Kullback-Leibler 情報量) を基準として獲得した文脈木を用いることで、可変長の文脈を用いるマルコフモデルによる品詞タグ付けを提案している。

3. **マルコフモデルの重ね合わせ**

状態の語彙化と特定の単語列の登録より、学習コーパスの性質はより正確にマルコフモデルに反映されるが、スパースデータの問題の原因になりかねないことも事実である。この問題は、マルコフモデルの重ね合わせにより解決される。特殊化されたマルコフモデルと、通常の品詞に基づくマルコフモデルを重ね合わせることにより、通常の品詞に基づくマルコフモデルより精密であると同時に、特殊化されたマルコフモデルよりも頑強なマルコフモデルが得られる。

これらのアイデアは英語の品詞タグ付けや日本語の形態素解析に特有ではなく、他のコーパスに基づく手法に応用できるという点にも注意しておかなければならない。

以下の節では、まずマルコフモデルを用いた品詞タグづけの定式化について述べる。次に、我々の提案を詳しく説明する。さらに、これを実装して実験を行なった結果を報告する。最後に、本研究の結論を述べる。

2 確率モデルによる品詞タグ付け

単語間に区切りのある言語の確率モデルによる品詞タグ付けは、ある単語列 W にある品詞列 T が割り当てられる確率 $P(T|W)$ を最大にする品詞列 $\hat{T}$ を求めることと定義される。条件付き確率 $P(T|W)$ はベイズの法則を用いることで以下のように書き換えられる。

$$P(T|W) = \frac{P(W|T)P(T)}{P(W)}$$

この式の分母は T によらないので、求める品詞列は以下の式で与えられる。

$$\begin{aligned}\hat{T} &= \operatorname*{argmax}_T \frac{P(W|T)P(T)}{P(W)} \\ &= \operatorname*{argmax}_T P(W|T)P(T)\end{aligned}$$

この式に現れる $P(T)$ と $P(W|T)$ は単語列を $W = w_1 w_2 \cdots w_n$ とし、品詞列を $T = t_1 t_2 \cdots t_n$ とすると、次のように表される。

$$\begin{aligned}P(T) &= p(t_1)\prod_{i=2}^{n} p(t_i|t_1 t_2 \cdots t_{i-1}) \\ P(W|T) &= \prod_{i=1}^{n} p(w_i|t_1 t_2 \cdots t_i)\end{aligned}$$

一般に、これらの確率はコーパスから推定することで与えられる。十分な大きさの品詞タグ付きコーパスがある場合は、以下の式で与えられる最尤推定 [32] が用いられる[1]。式に現れる f は、品詞列や単語と品詞列の組のコーパスにおける頻度である。

$$\begin{aligned}p(t_i|t_1 t_2 \cdots t_{i-1}) &= \frac{f(t_1 t_2 \cdots t_i)}{f(t_1 t_2 \cdots t_{i-1})} \\ p(w_i|t_1 t_2 \cdots t_i) &= \frac{f(w_i, t_1 t_2 \cdots t_i)}{f(t_1 t_2 \cdots t_i)}\end{aligned}$$

この式から分かるように、品詞列の長さ i が大きくなると頻度 f の値は小さくなるので、推定される確率値の信頼性は小さくなる。このように、与えられるコーパスの大きさが有限であるという制限から、以下の近似を用いる。

$$p(t_i|t_1 t_2 \cdots t_{i-1}) \approx p(t_i|t_{i-k} t_{i-k+1} \cdots t_{i-1}) \quad (1)$$

$$p(w_i|t_1 t_2 \cdots t_i) \approx p(w_i|t_i) \quad (2)$$

式 (1) から、品詞の間の関係は、ある時点 i での品詞の確率分布が時点 $i-k$ から時点 $i-1$ までの長さ k の品詞の履歴にのみ依存する k 重マルコフ連鎖となっていることが分かる。また、式 (2) から、ある時点 i での単語の出現確率は、その時点での品詞 t_i にのみ依存する。以上のことから、品詞を状態に対応させることで、品詞タグ付けの問題は、以下のように定式化される。

1. k 重マルコフ情報源のパラメータを推定する。
2. 入力単語列 W に対して確率 $P(W|T)P(T)$ を最大にする状態遷移の列 T を求める。

確率 $P(W|T), P(T)$ は、上述の近似を用いて

$$\begin{aligned}P(T) &= p(t_1)\prod_{i=2}^{k} p(t_i|t_1 t_2 \cdots t_{i-1}) \\ &\quad \times \prod_{i=k+1}^{n} p(t_i|t_{i-k} t_{i-k+1} \cdots t_{i-1}) \\ P(W|T) &= \prod_{i=1}^{n} p(w_i|t_i)\end{aligned}$$

となる。さらに、便宜的に状態 $t_{-k+1}, t_{-k+2}, \ldots, t_0$ を導入し、最初の単語を読み込む時点での品詞の履歴をこれらの連接とすることで

$$P(T) = \prod_{i=1}^{n} p(t_i|t_{i-k} t_{i-k+1} \cdots t_{i-1})$$

となる。よって最大化すべき確率は次の式で与えられる。

$$\begin{aligned}&P(W|T)P(T) \\ &= \prod_{i=1}^{n} p(w_i|t_i) \prod_{i=1}^{n} p(t_i|t_{i-k} t_{i-k+1} \cdots t_{i-1}) \\ &= \prod_{i=1}^{n} p(w_i|t_i) p(t_i|t_{i-k} t_{i-k+1} \cdots t_{i-1}) \quad (3)\end{aligned}$$

また、遷移確率と出現確率は以下の式で与えられる。

$$\begin{aligned}P(t_i|t_{i-k} t_{i-k+1} \cdots t_{i-1}) &= \frac{f(t_{i-k} t_{i-k+1} \cdots t_i)}{f(t_{i-k} t_{i-k+1} \cdots t_{i-1})} \\ P(w_i|t_i) &= \frac{f(w_i, t_i)}{f(t_i)}\end{aligned}$$

コーパスにおける品詞列の頻度 $f(t_{i-k} t_{i-k+1} \cdots t_i)$ は、品詞単位の n-gram 統計によって与えられる。このとき、各文の前に便宜的に導入した状態列 $t_{-k+1} t_{-k+2} \cdots t_0$ を表わす品詞を付加しておかなければならない。また、単語と品詞の組の頻度 $f(w_i, t_i)$ は、単語と品詞の両方を区別して出現頻度を計数することで得られる。

式 (3) の値を最大にする状態遷移列 (最尤な経路) の探索には、動的計画法の一種である Viterbi アルゴリズム [34] が用いられる。これは、ある時点である状態に到達する最尤な経路と累積確率値 (部分解) を記憶しておくと、次の時点のある状態の部分解は、これを参照することで求められるという性質を利用している。

[1] 十分な量の品詞タグ付きコーパスがない場合は forward-backward アルゴリズム [33] が用いる。

3 品詞タグ付けのためのマルコフモデルの改良

我々は、マルコフモデルによる日本語の形態素解析の精度を向上させるため、状態の語彙化と特定の形態素列の登録を提案した [29]。この節では、これらの提案を説明し、これが英語の品詞タグ付けにも応用できることを示す。説明には単純マルコフモデルを用いるが、多重マルコフモデルにも適用できる。

3.1 語彙化と単語列の登録

従来のマルコフモデルによる品詞タグ付けの研究では、品詞を状態に対応させている。しかし、品詞体系が完全でないため、異った振舞いをする単語が同じ品詞に属することがある。これらの単語の遷移確率を品詞で代表して記述すると、コーパスから得られる情報を失う可能性がある。本研究では、品詞と単語の対を状態に対応させ、より精密に状態遷移確率を表現する。もう一つの改良として、単語列の登録を行なう。登録の対象としては、内部に句を含まない名詞句と形容詞句と副詞句および動詞列を選んだ[2]。これらは、構文解析済みコーパスから取り出される。単語列は既存の品詞には属さないため、単語列の登録は語彙化を必然的に伴う。この結果、複数の単語が一つの品詞に対応することがあるので、最適な品詞列の探索には Viterbi アルゴリズムを拡張したアルゴリズム [35] を用いた。

次に、以上に述べたことを、構文解析済みの文から変換された次の文を例として具体的に説明する[3]。各単語の後ろの記号は品詞である (表 1参照)。

```
( NP Areas/NNS ) ( Other of/IN ) ( NP
the/DT factory/NN ) ( VP were/VBD ) ( ADJP
particularly/RB dusty/JJ ) ( Other ./. )
```

例として、名詞句 NP を語彙化する場合を考える。この場合、通常の品詞に加えて、名詞句 “the/DT factory/NN” と “Areas/NNS” が状態に対応するので、次のような品詞単位の 2-gram の頻度が得られる。

$$f(\bigstar \cdot \texttt{Areas/NNS}) = 1$$
$$f(\texttt{Areas/NNS} \cdot \texttt{IN}) = 1$$
$$f(\texttt{IN} \cdot \texttt{the/DT factory/NN}) = 1$$
$$f(\texttt{the/DT factory/NN} \cdot \texttt{VBD}) = 1$$
$$f(\texttt{VBD} \cdot \texttt{RB}) = 1$$
$$f(\texttt{RB} \cdot \texttt{JJ}) = 1$$
$$f(\texttt{JJ} \cdot .) = 1$$

ここで、「★」は便宜的に導入された状態である。また、単語 (列) と品詞の組の頻度は次のようになる。

$$f(\texttt{Areas}, \texttt{Areas/NNS}) = 1$$
$$f(\texttt{of}, \texttt{IN}) = 1$$
$$f(\texttt{the factory}, \texttt{the/DT factory/NN}) = 1$$
$$f(\texttt{were}, \texttt{VBD}) = 1$$
$$f(\texttt{particularly}, \texttt{RB}) = 1$$
$$f(\texttt{dusty}, \texttt{JJ}) = 1$$
$$f(.,.) = 1$$

これらの頻度を大規模な学習コーパスに対して計数することにより、名詞句を語彙化したマルコフモデルが得られる。実験では、語彙化する単語列を変えることで得られる以下のマルコフモデルを用いた。

マルコフモデル	名詞句	動詞列	その他
$M_{P,P,P}$	品詞	品詞	品詞
$M_{L,P,P}$	語彙化	品詞	品詞
$M_{P,L,P}$	品詞	語彙化	品詞
$M_{P,P,L}$	品詞	品詞	語彙化
$M_{L,L,L}$	語彙化	語彙化	語彙化

日本語の形態素解析の場合と同様に、全ての状態を語彙化したマルコフモデル $M_{L,L,L}$ は、語彙化されていないマルコフモデル $M_{P,P,P}$ よりも学習コーパスの性質を忠実に反映していると考えられる。一方で、未知のコーパスに対する $M_{L,L,L}$ の受理率 (入力文の数に対する解析可能な文の数) は、$M_{P,P,P}$ の受理率よりもかなり低いくなると予想される。それぞれの長所を生かすために、次の項で定義する重ね合わせを用いる。このとき、全ての状態が品詞に対応するマルコフモデル $M_{P,P,P}$ と全ての状態が語彙化されているマルコフモデル $M_{L,L,L}$ を重ね合わせても、便宜的に導入された状態以外の状態を共有しないため、文の途中で $M_{L,L,L}$ の状態を部分的に通るということができない。語彙化された状態と語彙化されていない状態を持つマルコフモデル $M_{L,P,P}, M_{P,L,P}, M_{P,P,L}$ は、$M_{P,P,P}$ と $M_{L,L,L}$ の両方と状態を共有しているので、文の途中での $M_{P,P,P}$ と $M_{L,L,L}$ の間の遷移を可能にする。これらのマルコフモデルを重ね合わせることで得られるマルコフモデルは、語彙化されていないマルコフモデル $M_{P,P,P}$ と同じ頑強性を持つと同時に、語彙化されたマルコフモデル $M_{L,L,L}$ と同程度の精度を持つと考えられる。

[2] 動詞列とは、内包する名詞句や前置詞句を除いた動詞句とする。

[3] NP: 名詞句, VP: 動詞列, ADJP, 形容詞句, Other: その他

表 1: Penn Treebank の品詞タグ

1.	CC	Coordinating conjunction
2.	CD	Cardinal number
3.	DT	Determiner
4.	EX	Existential *there*
5.	FW	Foreign word
6.	IN	Preposition or subordinating conj.
7.	JJ	Adjective
8.	JJR	Adjective, comparative
9.	JJS	Adjective, superlative
10.	LS	List item marker
11.	MD	Modal
12.	NN	Noun, singular or mass
13.	NNS	Noun, plural
14.	NNP	Proper noun, singular
15.	NNPS	Proper noun, plural
16.	PDT	Predeterminer
17.	POS	Possessive ending
18.	PRP	Personal pronoun
19.	PRP$	Possessive pronoun
20.	RB	Adverb
21.	RBR	Adverb, comparative
22.	RBS	Adverb, superlative
23.	RP	Particle
24.	SYM	Symbol
25.	TO	*to*
26.	UH	Interjection
27.	VB	Verb, base form
28.	VBD	Verb, past tense
29.	VBG	Verb, gerund or present participle
30.	VBN	Verb, past participle
31.	VBP	Verb, non-3rd person singular present
32.	VBZ	Verb, 3rd person singular present
33.	WDT	Wh-determiner
34.	WP	Wh-pronoun
35.	WP$	Possessive wh-pronoun
36.	WRB	Wh-adverb
37.	#	Pound sign
38.	$	Dollar sign
39.	.	Sentence-final punctuation
40.	,	Comma
41.	:	Colon, semi-colon
42.	(	Left bracket character
43.	)	Right bracket character
44.	"	Straight double quote
45.	‘	Left open single quote
46.	“	Left open double quote
47.	’	Right close single quote
48.	”	Right close double quote

3.2 マルコフモデルの重ね合わせ

最尤推定により得られるマルコフモデルは、各状態の頻度を記憶した状態頻度ベクトル v と、それぞれの状態間の遷移頻度を記憶した遷移頻度行列 A と、単語品詞対の頻度を記憶した出現頻度行列 B の 3 項組み $M = (v, A, B)$ で表わすことができる。

$$\begin{aligned} v_i &= f(t_i) \\ A_{i,j} &= f(t_i t_j) \\ B_{i,j} &= f(w_i, t_j) \end{aligned}$$

これらを用いると、状態 t_i から t_j への遷移確率 $P(t_j|t_i)$ および、状態 t_j において単語 w_k が出力される出現確率 $P(w_k|t_j)$ は以下の式で表される。

$$\begin{aligned} P(t_j|t_i) &= \frac{f(t_i t_j)}{f(t_i)} = \frac{A_{i,j}}{v_i} \\ P(w_k|t_j) &= \frac{f(w_k, t_j)}{f(t_j)} = \frac{B_{k,j}}{v_j} \end{aligned}$$

マルコフモデルの重ね合わせは、マルコフモデル $M_1,\ M_2,\ \ldots,\ M_n$ と、それぞれの重み $k_1,\ k_2,\ \ldots,\ k_n$ が与えられたとき、重ね合わせの結果得られるマルコフモデルを $M_{SP} = (v_{SP}, A_{SP}, B_{SP})$ として、以下のように定義される。

$$\begin{aligned} v_{SP} &= k_1 v_1 + k_2 v_2 + \cdots + k_n v_n \\ A_{SP} &= k_1 A_1 + k_2 A_2 + \cdots + k_n A_n \\ B_{SP} &= k_1 B_1 + k_2 B_2 + \cdots + k_n B_n \end{aligned}$$

足し算を行なう際に、各添字に対応する状態や単語が、全てのマルコフモデルに対して同じである必要があることに注意しなければならない。次の節で述べる実験では、$M_{SP} = 256M_{L,L,L} + 16M_{L,P,P} + 16M_{P,L,P} + 16M_{P,P,L} + M_{P,P,P}$ とした。重みは恣意的に決められており、最適化を行なった結果得られた値ではない[4]。

4 実験結果とその評価

前節で述べた方法を用いて品詞タグ付けシステムを実装し、Wall Street Journal [30] を用いて実験を行なった。表 1は、このコーパスに用いられている品詞タグの一覧である。コーパスをパラメータ推定用 (50,773 文、842,052 単語) と精度評価用 (1,493 文、24,759 単語) に分割し、実験を行なった。

表 2に前節で述べた 6 つのマルコフモデルによる英語の品詞タグ付けの、単語単位の正解率と入力文の受理率

[4] ある解析済みコーパスに対して最適な重みの組みは EM アルゴリズムを用いて求めることができる

表 2: マルコフモデルによる品詞タグ付けの精度

重み					学習コーパス		テストコーパス	
$M_{L,L,L}$	$M_{L,P,P}$	$M_{P,L,P}$	$M_{P,P,L}$	$M_{P,P,P}$	正解率	受理率	正解率	受理率
1	0	0	0	0	99.77%	100.00%	96.76%	2.55%
0	1	0	0	0	98.78%	100.00%	94.65%	15.07%
0	0	1	0	0	98.38%	100.00%	94.82%	45.48%
0	0	0	1	0	97.21%	100.00%	96.70%	70.19%
0	0	0	0	1	96.99%	100.00%	97.01%	72.20%
256	16	16	16	1	99.73%	100.00%	97.02%	72.20%

表 3: 外部辞書と未知語モデルを持つマルコフモデルによる品詞タグ付けの精度

重み					学習コーパス		テストコーパス	
$M_{L,L,L}$	$M_{L,P,P}$	$M_{P,L,P}$	$M_{P,P,L}$	$M_{P,P,P}$	正解率	受理率	正解率	受理率
256	0	0	0	0	99.77%	100.00%	96.76%	2.55%
0	16	0	0	0	98.76%	100.00%	92.33%	84.13%
0	0	16	0	0	96.98%	100.00%	93.46%	100.00%
0	0	0	16	0	96.96%	100.00%	96.07%	100.00%
0	0	0	0	1	96.99%	100.00%	96.61%	100.00%
256	16	16	16	1	99.73%	100.00%	96.64%	100.00%

外部辞書の重みは 1/16 であり、未知語モデルの重みは 1/256 である。

を示す。この表から、テストコーパスに対する受理率は $M_{L,L,L}$ が最も低く、$M_{P,P,P}$ が最も高いことが分かる。これとは反対に、学習コーパスに対する正解率は $M_{L,L,L}$ が最も高く、$M_{P,P,P}$ が最も低い。テストコーパスに対する M_{SP} の受理率は $M_{P,P,P}$ の受理率と同じであると同時に、学習コーパスに対する正解率は $M_{L,L,L}$ の正解率と同程度である。これは、理論的に予測される結果と符合し、マルコフモデルの重ね合わせによって、精度を落さずスパースデータの問題を解決できることを示している。

しかしながら、M_{SP} の受理率は十分とはいえない。我々は、これがテストコーパスに含まれる未知語に起因すると考え、各マルコフモデルに外部辞書と未知語モデルを付け加え、同じ実験を行なった。外部辞書として、英語の形態素解析器 [36] に附属している辞書を用いた。この辞書の単語の中で、未知語として出現する可能性がある品詞に属する 240,147 個の単語を 1/16 の重みで付け加えた[5]。これにより、品詞タグ付けシステムは、これらの単語の出現頻度を 1/16 とみなす。未知語モデルは、単純な正規表現で実装されており、与えられた未知語の文字種や語尾などの情報と外部辞書を用いて、可能な品詞を出現頻度 1/256 として返す。派生や屈折を考慮した上で辞書で調べても見つからない場合は、単数名詞とする。例えば、学習コーパスや辞書にない単語 “zamindar” が入力文に含まれ、名詞であると推定されたとすると、$f(\texttt{zamindar}, \texttt{NN}) = 1/256$ とみなす。表 3に外部辞書と未知語モデルを付け加えた各マルコフモデルの正解率と受理率を示す。$M_{L,L,L}$ の正解率と受理率が、外部辞書と未知語モデルを付け加えても変わらないのは、これらのマルコフモデルは全ての状態が語彙化されており、名詞や動詞などの辞書に記述された単語の品詞に対応する状態を持たないためである。その他のマルコフモデルでは、外部辞書と未知語モデルによって受理率が高くなっている。

外部辞書と未知語モデルを付け加えた M_{SP} の正解率は 96.64% である。文献 [17] で提案された語彙化した規則を用いる方法を、同じ学習コーパスで学習し、同じテストコーパスに対して実験した結果得られた正解率は 96.07% であった。この文献では、語彙化した規則を用いる方法が確率を用いる方法よりも高い精度であったと報告している。このことを考えると、語彙化したマルコフモデルによ

[5] 対象とした品詞タグは JJ, JJR, JJS, NN, NNS, RB, RBR, RBS, VB, VBD, VBG, VBN, VBP, VBZ である

る方法は最も精度の高い英語の品詞タグ付けの方法の一つであると結論できる。

5 おわりに

本論文では、マルコフモデルを用いた英語品詞タグ付けの新しい手法を提案し、これを実装して実験を行なった結果を報告した。実験の結果、我々が提案する英語の品詞タグ付けの方法は高い精度を示し、我々が提案する手法の有効性が確かめられた。

参考文献

[1] Kenneth Ward Church. A Stochastic Parts Program and Noun Phrase Parser for Unrestricted Text. In *Proceedings of the Second Conference on Applied Natural Language Processing*, pp. 136–143, 1988.

[2] Steven J. DeRose. Grammatical Category Disambiguation by Statistical Optimization. *Computational Linguistics*, Vol. 14, No. 1, pp. 31–39, 1988.

[3] Julian Kupiec. Augmenting a Hidden Markov Model for Phrase-Dependent Word Tagging. In *Proceedings of the DARPA Speech and Natural Language Workshop*, pp. 92–98, 1989.

[4] Carl G. de Marcken. Parsing the LOB corpus. In *Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics*, pp. 243–251, 1990.

[5] Marie Meteer, Richard Schwartz, and Ralph Weischedel. POST: Using Probabilities in Language Processing. In *International Joint Conference on Artificial Intelligence*, pp. 960–965, 1991.

[6] Marie Meteer, Richard Schwartz, and Ralph Weischedel. Studies in Part of Speech Labelling. In *Proceedings of the DARPA Speech and Natural Language Workshop*, pp. 331–336, 1991.

[7] Doug Cutting, Julian Kupiec, Jan Pedersen, and Penelope Sibun. A Practical Part-of-Speech Tagger. In *Proceedings of the Third Conference on Applied Natural Language Processing*, pp. 133–140, 1992.

[8] Julian Kupiec. Robust Part-of-Speech Tagging using a hidden Markov Model. *Computer Speech and Language*, Vol. 6, , 1992.

[9] 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.

[10] Ralph Weischedel, Marie Meteer, Richard Schwartz, Lance Ramshaw, and Jeff Palmucci. Coping with Ambiguity and Unknown Words through Probabilistic Models. *Computational Linguistics*, Vol. 19, No. 2, pp. 359–382, 1993.

[11] David Elworthy. Does Baum-Welch Re-estimation Help Taggers? In *Proceedings of the Fourth Conference on Applied Natural Language Processing*, pp. 53–58, 1994.

[12] Yi-Chung Lin, Tung-Hui Chiang, and Keh-Yih Su. Automatic Model Refinement – with an Aplication to Tagging. In *Proceedings of the 15th International Conference on Computational Linguistics*, pp. 148–153, 1994.

[13] Hinrich Schütze and Yoram Singer. Part of Speech Tagging Using a Variable Memory Markov Model. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, pp. 181–187, 1994.

[14] Bernard Merialdo. Tagging English Text with a Probabilistic Model. *Computational Linguistics*, Vol. 20, No. 2, pp. 155–171, 1994.

[15] Evangelos Dermatas and George Kokkinakis. Automatic Stochastic Tagging of Natural Language Texts. *Computational Linguistics*, Vol. 21, No. 2, pp. 137–163, 1995.

[16] Eric Brill. A Simple Rule-Based Part of Speech Tagger. In *Proceedings of the Third Conference on Applied Natural Language Processing*, pp. 152–154, 1992.

[17] Eric Brill. Some Advances in Transformation-Based Part of Speech Tagging. In *Proceedings of the Twelveth National Conference on Artificial Intelligence*, pp. 722–727, 1994.

[18] Eric Brill. A Report of Recent Progress in Transformation-Based Error-Driven Approach. In *Proceedings of the ARPA Workshop on Human*

Language Technology, pp. 256–261, 1994.

[19] Eric Brill. Unsupervised Learning of Disambiguation Rules for Part of Speech Tagging. In *Proceedings of the Third Workshop on Very Large Corpora*, pp. 1–13, 1995.

[20] Emmanuel Roche and Yves Schabes. Deterministic Part-of-Speech Tagging with Finite-State Transducers. *Computational Linguistics*, Vol. 21, No. 2, pp. 228–253, 1995.

[21] 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, pp. 543–565, 1995.

[22] Masami Nakamura and Kiyoshiro Shikano. A Study of English Word Category Prediction Based on Neural Networks. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pp. 731–734, 1989.

[23] Julian Benello, Andrew W. Mackie, and James A. Anderson. Syntactic Category Disambiguation with Neural Networks. *Computer Speech and Language*, Vol. 3, pp. 203–217, 1989.

[24] Masami Nakamura, Katsuteru Maruyama, Takeshi Kawabata, and Kiyoshiro Shikano. Neural Network Approach to Word Category Prediction for English Texts. In *Proceedings of the 13th International Conference on Computational Linguistics*, pp. 213–218, 1990.

[25] Helmut Schmid. Part-of-Speech Tagging with Neural Networks. In *Proceedings of the 15th International Conference on Computational Linguistics*, pp. 172–176, 1994.

[26] Ezra Black, Fred Jelinek, John Lafferty, Robert Mercer, and Salim Roukos. Decision Tree Models Applied to the Labeling of Text with Parts-of-Speech. In *Proceedings of the DARPA Speech and Natural Language Workshop*, pp. 117–121, 1992.

[27] Pasi Tapanainen and Atro Voutilainen. Tagging accurately – Don't guess if you know. In *Proceedings of the Fourth Conference on Applied Natural Language Processing*, pp. 47–51, 1994.

[28] Jean-Pierre Chanod and Pasi Tapanainen. Tagging French – comparing a statistical and a constraint-based method. In *Proceedings of the Seventh European Chapter of the Association for Computational Linguistics*, pp. 149–156, 1995.

[29] 森信介, 長尾眞. 形態素 bi-gram と品詞 bi-gram の重ね合わせによる形態素解析. 情報処理学会研究報告, 1996.

[30] Mitchell P. Marcus and Beatrice Santorini. Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*, Vol. 19, No. 2, pp. 313–330, 1993.

[31] Peter F. Brown, Vincent J. Della Pietra, Peter V. deSouza, Jenifer C. Lai, and Robert L.Mercer. Class-Based n-gram Models of Natural Language. *Computational Linguistics*, Vol. 18, No. 4, pp. 467–479, 1992.

[32] 中川聖一. 確率モデルによる音声認識. 電子情報通信学会, 1988.

[33] L. E. Baum. An Inequality and Associated Maximization Technique in Statistical Estimation for Probabilistic Functions of Markov Process. *Inequalities*, Vol. 3, pp. 1–8, 1972.

[34] Andrew J. Viterbi. Error Bounds for Convolutional Codes and an Asymptotically Optimal Decoding Algorithm. *IEEE Transactions on Information Theory*, pp. 260–269, 1967.

[35] Masaaki Nagata. A Stochastic Japanese Morphological Analyzer Using a Forward-DP Backward-A* N-Best Search Algorithm. In *Proceedings of the 15th International Conference on Computational Linguistics*, pp. 201–207, 1994.

[36] Daniel Karp, Yves Schabes, Martin Zaidel, and Dania Egedi. A Freely Available Wide Coverage Morphological Analyzer for English. In *Proceedings of the 14th International Conference on Computational Linguistics*, pp. 950–955, 1992.