自然言語処理 112－6
（1996．3．14）

形態素 bi-gram と品詞 bi-gram の重ね合わせによる形態素解析

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
入力文を単語に分割し品詞を付加する形態素解析は、日本語処理における基本的な処理である。英語の品詞タグ付けと異なりコーパスに基づく形態素解析の試みはあまり行なわれていない。本論文では、コーパスに基づく形態素解析の新しい手法を提案する。我々の提案は次のように要約される。1) 各形態素の語彙化、2) 附属語列の登録、3) マルコフモデルの重ね合わせ。これらのアイデアは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。以上のアイデアに基づく形態素解析器を作成し、EDR コーパス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 形態素解析 コーパス 語彙化 附属語列 マルコフモデル

Japanese Morphological Analysis by Superposition of Morpheme Bi-gram and Part-of-Speech Bi-gram

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
Morphological analysis, which segments the input sentence into words and attaches part-of-speeches to them, is the most fundamental process of Japanese language processing. Contrary to English part-of-speech tagging, only few attempts have so far been made at corpus-based Japanese morphological analysis. In this paper we propose a novel method for corpus-based Japanese morphological analysis. Our proposals are summarized as follows: 1) lexicalization of all morphemes; 2) memorization of particle sequence; 3) superposition of Markov models. These ideas are so general that one can apply them to other corpus-based applications. We conducted experiments on EDR corpus and obtained the considerably high accuracy.
Key Words Morphological analysis, Corpus, Lexicalize, Particle sequence, Markov model

1 はじめに

日本語には単語間に明示的な区切りがない。そのため、入力文を単語に分割し、品詞を付加する形態素解析は日本語処理における基本的な処理である。このような視点から、今日までに多くの形態素解析器が人間の言語直観に基づき作成されている。一方、英語の品詞タグ付けではいくつかのコーパスに基づく方法が提案され、非常に高い精度を報告している [1]-[10]。今日、多くの研究者が、英語の品詞タグ付けに関してはコーパスに基づく手法が従来のヒューリスティックルールに基づく手法より優れていると考えるに至っている。

日本語の形態素解析に対しては、コーパスに基づく手法が従来のルールに基づく手法より優れていると考えるには至っていないようである。これは、形態素解析済みの大規模日本語コーパスが最近になってようやく現れたことを考えると極めて自然である。実際、コーパスに基づく形態素解析は、現在までのところ少数の研究がなされているのみである [11]-[14]。

本論文では、コーパスに基づく形態素解析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する。文献 [12] では、2 重マルコフモデルが用いられているが、我々の方法は単純マルコフモデルを用いる。実験の結果、本論文で提案する手法の精度は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ように要約される。

1. 各形態素の語彙化
2. 小辞列の登録
3. マルコフモデルの重ね合わせ

最初のアイデアにより、形態素間の連接確率がより精密に表現できる。二番目のアイデアは、いくつかの形態素列は一つの形態素とみなすことができるという直観に基づいている。これにより、単純マルコフモデルが入力文によっては、多重マルコフモデルと同等の能力をもつことが可能になる。これらのアイデアにより、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。最後のアイデアは、この問題に対する一つの解決策となっている。語彙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、語彙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、まず日本語形態素解析の概略と現在の技術水準について述べる。次に、我々のアイデアについて説明する。さらに、実際に形態素解析器を作成し EDR コーパスに対して行なった実験の結果を報告する。最後に、本研究の結論を述べる。

2 日本語の形態素解析

この節では、日本語形態素解析の処理過程の概略と、現在の日本語形態素解析の技術水準について述べる。

2.1 解析過程の概略

日本語には単語間に区切りがないので、文解析の最初に位置する形態素解析は、文字の連接である入力文を部分文字列に分割し、それぞれの部分文字列の品詞を決定する。この結果得られる文字列と品詞の組を形態素と呼ぶ。この分割には、一般に以下の条件が課せられる。

1. 各部分文字列には品詞が対応していること
2. 隣接する形態素は連接可能であること

最初の条件は、部分文字列を辞書で調べることでチェックする。このとき、各部分文字列の品詞が決定される。次の条件は、連接表を調べることでチェックする。この表は、全ての形態素の組がそれぞれ連接してよいか否かを記憶している。

上記の条件の下では、入力文に対して複数の解析が可能な場合がある。このため、実際の解析では、各形態素には形態素コストと呼ばれる数値が、各連接には連接コストと呼ばれる数値がそれぞれ与えられており、これらのコストの合計が最小になる組合せを最適解として出力する。単語列 $w_1 w_2 \cdots w_n$ に品詞列 $t_1 t_2 \cdots t_n$ が割り当てられる解析のコストは、以下のように表わされる。

$$\sum_{i=1}^{n} \{C(t_{i-1}, t_i) + M(w_i, t_i)\} + C(t_n, t_{n+1})$$

ここで、$C(t_{i-1}, t_i)$ は品詞 t_{i-1} と品詞 t_i がこの順で連接する場合の連接コストであり、$M(w_i, t_i)$ は単語 w_i が品詞 t_i として用いられる場合の形態素コストである。また、t_0 と t_{n+1} は、文頭と文末に対応する特別な品詞である。以上が解析の骨格であるが、付帯的な問題として、入力文が辞書に登録されていない形態素 (未知語) を含み、正しい解析結果が得られない場合がある。この問題に対処するために、ある条件を満たす文字列をある品詞とみなし、ある程度高いコストを与えるという処理を行なうのが

一般的である。この例として、片仮名の連接を品詞とみなすなどの規則が挙げられる。

コーパスに基づく方法では、形態素コストと連接コストを確率として表し、これらのパラメータをコーパスから推定する。一般に、このモデルとしてマルコフモデルが用いられる。これは、形態素解析の条件がマルコフモデルによく対応しているためである。つまり、各品詞を状態 t に対応させることで、連接コストが遷移確率 $P(t_i|t_{i-1})$ に対応する[1]。形態素コストは、各状態において形態素 w_i が出力される条件付確率 $P(w_i|t_i)$ に対応する。これらのパラメータは、一般に解析済みコーパスから品詞列や形態素の頻度を計算することで推定される。

$$P(t_i|t_{i-1}) = f(t_{i-1}t_i)/f(t_{i-1})$$
$$P(w_i|t_i) = f(w_i,t_i)/f(t_i)$$

ここで、$f(t_{i-1}t_i)$ は品詞 t_{i-1} と品詞 t_i がこの順に連続して現れる頻度であり、$f(t_i)$ は品詞 t_i の頻度である。また、$f(w_i,t_i)$ は単語 w_i が品詞 t_i として用いられている頻度である。人手で記述したルールによる方法と同様、複数の解析結果の中で最も確率が高い結果を最尤の結果として出力する。マルコフモデルを用いた場合、入力文 (文字列)$S = c_1c_2\cdots c_m$ が単語列 $W = w_1w_2\cdots w_n$ と品詞列 $T = t_1t_2\cdots t_n$ で構成される確率は以下の式で表わされる。

$$P(W,T|S) = \prod_{i=1}^{n} P(t_i|t_{i-1})P(w_i|t_i) \times P(t_{n+1}|t_n)$$

ここで、t_0 と t_{n+1} は、文頭と文末に対応する特別な品詞である。この値の最大値を与える単語列と品詞列の組み $(\hat{W},\hat{T})$ が形態素解析の結果として出力される。解の探索には動的計画法を用いることができ、入力の文字数 m に対して計算時間のオーダーが $O(m)$ となるアルゴリズムが提案されている [11]。

2.2 現在の技術水準

人手で記述したルールによる方法とコーパスに基づく方法の最大の違いは、コストの設定方法である。前者の方法では、コストを文法家や計算言語学の研究者が人手で与える。この方法では、ひとたびコストを決定した後も、解析結果に応じて調整を行なう。これは、この方法の利点であると同時に欠点でもある。つまり、適切に調整を行なうことで確実に精度を上げることができる反面、多くの簡単な誤りを訂正するための調整を行なった後に、さらに精度を上げるための微調整を新たな誤りを生み出すことなく行なうのは非常に困難な課題である。現在、広く使われている形態素解析システム JUMAN[15] の精度は、形態素単位の評価で 98% ～ 99% であるとされているが、かなりのコスト調整が行なわれており、人手によるコスト調整ではこれ以上の改善は見込めない状況である。

コーパスに基づく方法では、先行する 2 つの形態素を考慮に入れる 2 重マルコフモデルを用いた方法を EDR コーパス [16] に適用した結果、91% 前後の精度であったと報告されている [12]。ここで、この方法では、コーパスと正確に一致する形態素のみを正解とするため、人手で記述したルールによる方法と直接比較することはできないことに注意しなければならない。このほかに、隠れマルコフモデルを用いてモデル化し、解析結果が付与されていないコーパスから Forward-Backward アルゴリズムによって各パラメータを推定する方法が提案されている [14]。

いずれの方法においても、主な誤りは以下の 2 種類に分類される。

1. 漢字列の複合語の分割
2. 主に助詞や助動詞などからなる平仮名列の分割

最初の範疇の誤りは、複合語の構造に影響するだけで、構文解析などの後続する解析に対する影響は少ないと考えられる。また、形態素解析の結果として得られる単語間の境界はかなり正確なので、複合語の解析を独立に扱うことができる。二番目の範疇の誤りは、形態素解析の条件では平仮名列に対して非常に多い解釈が可能であることによる。これは、形態素解析の段階で解決すべき問題であり、主にこの誤りに対処する方法を次の節で提案する。

3 確率的日本語形態素解析の改良

我々は、確率的日本語形態素解析の精度を向上させるため、語彙化と附属語列の登録を行なうことを提案する。これらによりスパースデータの問題を生じるが、これはマルコフモデルの重ね合わせにより解決される。この節では、これらについて述べる。

3.1 語彙化と附属語列の登録

従来の方法では、各形態素の連接可能性はそれぞれが属する品詞で代表しているために、各形態素の個別の振舞いを記述できず、解析誤りの一つの原因となっている。我々は、これを解決することを目的として、連接可能性を各形態素に対して記述することを提案する。具体的には、マルコフモデルの状態を形態素に対応させることで実現され

[1] より一般的に k 重マルコフモデルを用いた場合、遷移確率は $P(t_i|t_{i-1}t_{i-2}\cdots t_{i-k})$ となる。

る。これは、従来の方法における品詞の分類を極限まで細分化したモデルと言える。理論的には、連接規則を人手で記述する方法にも応用できるが、現実的には、形態素のあらゆる組合せに対して連接コストを整合的に記述するのは不可能であり、コーパスに基づく統計的手法に対してのみ有効である。以下では、これを自立語に対して適用することを自立語の語彙化と呼ぶ。

もう一つの改良方法として、附属語列の登録を提案する。前節で、従来の形態素解析の主な誤りの原因として平仮名列の解析を挙げた。正しい解析結果を得ることが困難な平仮名列の多くは、活用語の語尾や助詞や助動詞の連接であるが、品詞レベルの連接規則では実際に出現する組合せよりもかなり多くの組合せを解析候補として生成してしまう。また、2連接に対する規則では不十分で、3連接やそれ以上の連接規則が必要な場合もある。我々は、ある種の平仮名列を登録することでこの問題に対処できると考えた。登録の対象としては、活用語の語尾と助詞と助動詞の最長の連接を選んだ。この理由は、これらを一つの形態素とみなすことにより、文節が複数の自立語と一つ以下の附属語の組合せとなり、後続の解析との整合性がよいと考えたことである。これを実現するために、解析済みコーパスに出現する最長の附属語列をマルコフモデルの異なる状態に対応させ、その平仮名列の従来の文法体系での解析結果を記憶しておくこととした。附属語列は、既存の品詞には属さないため、附属語列の登録は語彙化を必然的に伴う。以下ではこれを附属語の語彙化と呼ぶ。

次に、以上に述べたことの実現方法を、以下のような形態素解析済みの文を例として具体的に説明する。

例文 1

★ フランス (名詞) の (助詞) 外交政策 (名詞)
は (助詞) 借り物 (名詞) で (助動詞) は (助詞)
な (形容詞) い (形容詞語尾) ★

文頭と文末に便宜的に置かれた記号「★」は、文頭と文末に対応する品詞である。これを品詞が状態に対応するマルコフモデルでモデル化すると図1のようになる。この図では、遷移確率を省略している。このように、自立語も附属語も語彙化せずに得られるマルコフモデルを $M_{P,P}$ で表す。同じ例文に対して、自立語と附属語の両方を語彙化することで得られるマルコフモデル $M_{L,L}$ を図2に示す。このマルコフモデルは、語彙化していないマルコフモデル $M_{P,P}$ 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。

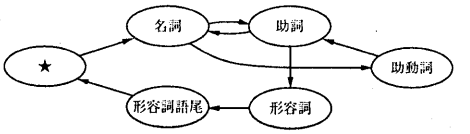

図 1: マルコフモデル $M_{P,P}$ の状態遷移図

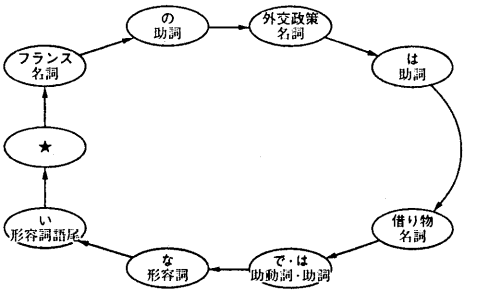

図 2: マルコフモデル $M_{L,L}$ の状態遷移図

一般にコーパスに基づく手法では、現象を何らかの基準でクラスに分類し、それぞれの現象をクラスで代表してモデル化する。これによって推定すべき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 [17]。しかし、クラスで代表することは何らかの情報を失うということを忘れてはならない。上述のマルコフモデルの場合には、テストコーパスに対する $M_{L,L}$ の受理率 (入力文の数に対する解析可能な文の数) は、$M_{P,P}$ の受理率よりもかなり低い一方、学習コーパスに対する $M_{L,L}$ の精度は、$M_{P,P}$ の精度よりもかなり高くなると予想される。

3.2 マルコフモデルの重ね合わせ

上で議論したことをふまえると、語彙化されていないマルコフモデルと語彙化されているマルコフモデルをうまく組み合わせることで、両方の利点を生かすことができると考えられる。この考えは、以下で定義されるマルコフモデルの重ね合わせによって実現される。

一般に、日本語の形態素解析などに用いられるマルコフモデルのパラメータは解析済みコーパスにおける品詞列や単語の頻度から最尤推定 [18] される。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、各形態素の頻度を記憶した形態素頻度行列 B の3項組み $M = (v, A, B)$ で表わすことができる。これらの各要素は、品詞列や単語のコーパスにおける

頻度から以下の式を用いて推定される。

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

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 が出力される条件付確率 $P(w_k|t_j)$ 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, ..., M_n と、それぞれの重み k_1, k_2, ..., 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。遷移頻度行列と形態素頻度行列は非常にスパースなので、ハッシュを用いて実装した。さらに、行列の添字となるハッシュのキーを、数字ではなく品詞名や単語とし、定義されていないキーに対してゼロを返すようにすることで、足し算を簡単に表すことができる。

上記の定義に従って、自立語と附属語の両方が語彙化されているマルコフモデル $M_{L,L}$ と両方とも語彙化されていないマルコフモデル $M_{P,P}$ を $M_{L,L}$ の重みを十分大きくして重ね合わせることで、$M_{L,L}$ が受理する文に対しては $M_{L,L}$ の解析結果を出力し、$M_{L,L}$ が受理しない文に対しては $M_{P,P}$ の解析結果を出力するマルコフモデルを得る。このマルコフモデルを $M_{SP'}$ とし、図3に示す。

マルコフモデル $M_{L,L}$ で受理されなかった文を詳しく調べると、受理されない原因となる未知語や未知の接続は文単位で見るときわめて少数であることが分かった。例えば、先に掲げた例文の「借り物」が「輸入品」に置き換わった以下の文では、「輸入品」が未知語であり、その前後の接続が未知であるために $M_{L,L}$ では受理できない。

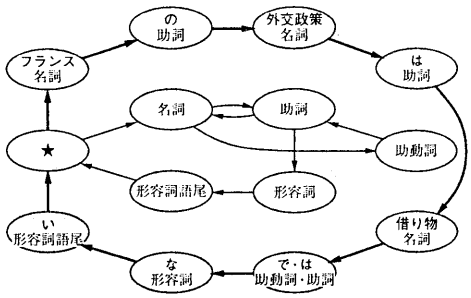

図 3: マルコフモデル $M_{SP'}$ の状態遷移図

例文 2

★ フランス (名詞) の (助詞) 外交政策 (名詞) は (助詞) 輸入品 (名詞) で (助動詞) は (助詞) な (形容詞) い (形容詞語尾) ★

マルコフモデル $M_{SP'}$ を用いた場合、図3から分かるように、$M_{L,L}$ の寄与による語彙化された外側のパスは「輸入品」の手前で探索が失敗するが、内側の $M_{P,P}$ の寄与による語彙化されていないパスを通ることで得られる解析結果を出力する。これは、$M_{L,L}$ と $M_{P,P}$ が「★」以外の共通の状態を持たないことによる。未知の文に対しても、語彙化されたマルコフモデルは語彙化されていないマルコフモデルよりも精度が高いと考えられるので、部分的にであっても語彙化されたマルコフモデルの寄与による解析結果を出力することが望ましい。このために、これらのマルコフモデルの間を橋渡しする以下のマルコフモデルを重ね合わせる。

- $M_{L,P}$: 自立語のみ語彙化して得られるマルコフモデル
- $M_{P,L}$: 附属語のみ語彙化して得られるマルコフモデル

先に掲げた例文から得られるこれらのマルコフモデルを図 4と図 5に示す。さらに、以上に述べたマルコフモデル $M_{L,L}$, $M_{L,P}$, $M_{P,L}$, $M_{P,P}$ を 100:10:10:1 で重ね合わせることで得られるマルコフモデルを $M_{SP} = 100M_{L,L} + 10M_{L,P} + 10M_{P,L} + M_{P,P}$ とし、これを図6に示す。次の節では、このマルコフモデルを用いて行なった実験結果を示すが、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[2]。

先に掲げた例文の「借り物」が「輸入品」に置き換わっ

[2] ある解析済みコーパスに対して最適な重みの組みは次のようにして求めることができる。それぞれの重みを変数として、正解の確率がその他の可能な解析の確率より大きくなるという条件により連立不等式を得る。得られた不等式により分割された重みの組みの部分空間の中で、精度が最高になる領域を求める。

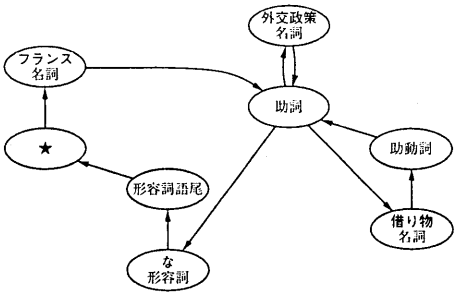

図 4: マルコフモデル $M_{L,P}$ の状態遷移図

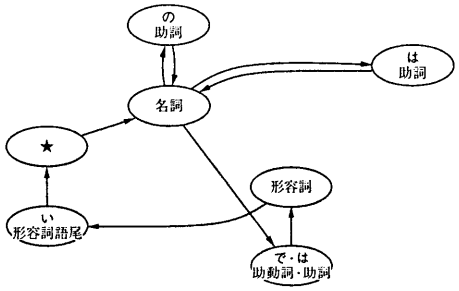

図 5: マルコフモデル $M_{P,L}$ の状態遷移図

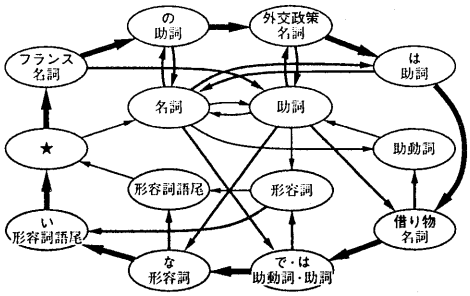

図 6: マルコフモデル M_{SP} の状態遷移図

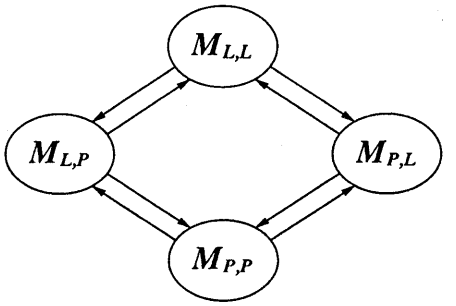

図 7: マルコフモデルの重ね合わせの概念図

た文に対して、マルコフモデル M_{SP} では、「は (助詞)」までは $M_{L,L}$ の寄与による語彙化された外側のパスを通り、$M_{P,L}$ の寄与によるパスを通って状態「名詞」を経て、状態「で・は (助動詞・助詞)」に遷移した後、再び $M_{L,L}$ の寄与による語彙化された外側のパスを通り最終状態「★」に到達する。内側のパスを多く通る解析も可能であるが、外側の状態への遷移確率が高くなるように重みづけされているので、外側のパスを多く通る解析結果が最尤であると推定される。同様に、附属語列が未知である場合には、$M_{L,P}$ の寄与によるパスを通ることになる。このように、重みづけによって 4 つのマルコフモデルを、図 7に示すように階層的に結合することが可能になる。この考えは、重ね合わせが定義されるあらゆるモデルに適用できるので、他のコーパスに基づく手法に応用できるという一般性がある。

4 実験結果とその評価

前節で述べた方法を用いて確率的形態素解析器を実装し、EDR コーパス [16] を用いて実験を行なった。コーパスをパラメータ推定用 (36,733 文、722,440 形態素) と精度評価用 (1,020 文、19,828 形態素) に分割し、前述の 5 つのマルコフモデルによる形態素解析の精度を計算した。品詞体系は基本的に EDR コーパスの品詞体系と同じであるが、動詞については概念辞書を用いて活用形を特定し分類した。以下では、その結果を提示し、これらの結果を評価する。

4.1 評価基準

我々が用いた評価基準は、文献 [12] で用いられた再現率と適合率であり、次のように定義される。EDR コーパスに含まれる形態素数を N_{EDR}、解析結果に含まれる形態素数を N_{SYS}、分割と品詞の両方が一致した形態素数を N_{COR} とすると、再現率は N_{COR}/N_{EDR} と定義され、適合率は N_{COR}/N_{SYS} と定義される。例として、コーパスの内容と解析結果が以下のような場合を考える。

コーパス

外交 (名詞) 政策 (名詞) で (助動詞) は (助詞)
な (形容詞) い (形容詞語尾)

解析結果

外交政策 (名詞) で (助詞) は (助詞)
な (形容詞) い (形容詞語尾)

この場合、分割と品詞の両方が一致した形態素は「は (助詞)」と「な (形容詞)」と「い (形容詞語尾)」である

表 1: マルコフモデルによる形態素解析の精度

重み				学習コーパス			テストコーパス		
$M_{L,L}$	$M_{P,L}$	$M_{L,P}$	$M_{P,P}$	再現率	適合率	受理率	再現率	適合率	受理率
1	0	0	0	98.88%	99.11%	100.00%	2.44%	91.15%	3.53%
0	1	0	0	95.14%	99.11%	100.00%	73.96%	91.65%	81.67%
0	0	1	0	96.77%	97.59%	100.00%	16.99%	88.87%	21.86%
0	0	0	1	93.24%	95.34%	100.00%	76.89%	90.67%	85.78%
100	10	10	1	98.79%	99.08%	100.00%	78.54%	91.24%	85.78%

表 2: 外部辞書と未知語モデルを持つマルコフモデルによる形態素解析の精度

重み				学習コーパス			テストコーパス		
$M_{L,L}$	$M_{P,L}$	$M_{L,P}$	$M_{P,P}$	再現率	適合率	受理率	再現率	適合率	受理率
100	0	0	0	98.88%	99.11%	100.00%	2.44%	91.15%	3.53%
0	10	0	0	91.11%	94.81%	100.00%	89.01%	92.48%	100.00%
0	0	10	0	96.77%	97.59%	100.00%	16.99%	88.87%	21.86%
0	0	0	1	87.25%	92.66%	100.00%	86.67%	91.53%	100.00%
100	10	10	1	98.75%	99.07%	100.00%	90.76%	92.74%	100.00%

外部辞書の重みは 0.1 であり、未知語モデルの重みは 0.01 である。

ので、 $N_{COR} = 3$ となる。また、コーパスには 6 つの形態素が含まれ、解析結果には 5 つの形態素が含まれているので、 $N_{EDR} = 6, N_{SYS} = 5$ である。よって、再現率は $N_{COR}/N_{EDR} = 3/6$ となり、適合率は $N_{COR}/N_{SYS} = 3/5$ となる。受理されない場合は、解析結果に含まれる形態素数を 0 とした。

4.2 実験結果とその評価

表 1に前節で述べたマルコフモデルの解析精度と受理された文の割合を示す。この表から、テストコーパスに対する受理率は $M_{L,L}$ が最も低く、 $M_{P,P}$ 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精度は $M_{L,L}$ が最も高く、 $M_{P,P}$ が最も低い。テストコーパスに対する M_{SP} の受理率は $M_{P,P}$ の受理率と同じであると同時に、学習コーパスに対する精度は $M_{L,L}$ の精度と同程度である。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すことなく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、EDR 辞書 [16] を 0.1 の重みで付け加えた。この辞書には約 25,000 の形態素が含まれていた。これにより、解析器はこれらの形態素の出現頻度を 0.1 とみなす。未知語モデルにより、解析器は英数字列を頻度 0.1 の名詞とし、数字列を頻度 0.1 の数字とし、漢字列と片仮名列を頻度 0.01 の名詞とし、平仮名列を頻度 1×10^{-10} の名詞とみなす。例えば、学習コーパスや辞書にない片仮名列「ディープブルー」が入力文に含まれるとすると、f(ディープブルー, 名詞) = 0.01 とみなす。

表 2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解析精度と受理率を示す。 $M_{L,L}$ と $M_{L,P}$ の解析精度と受理率が、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自立語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増加し、結果として再現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} の精度は、再現率が 90.66% で、適合率が 92.73% であり、文献 [12] の精度を双方で上回っている。全ての条件が同じというわけではないので単純な比較は適切ではないが、この結果は、本手法の優位性を実験的に示すと考えられる。

誤りの多くは、以下のように分類される。

1. 漢字列の複合語の分割
2. 文法家にも難解な微妙な品詞の違い
3. 本質的な誤り

最初の二つの誤りの原因は主に、コーパスの品詞付けが一貫していないことであると思われる。最後の種類の誤りだけを誤りとして、無作為抽出した 100 文を人手でチェックした結果、再現率は 99.41% で、適合率は 99.44% であった。同じ文の JUMAN [15] による出力を同じ基準で評価すると適合率 98.21% であった。このことから、本手法による形態素解析が十分実用的であると結論できる。

5 おわりに

本論文では、コーパスに基づく形態素解析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、本手法による形態素解析は高い精度を示し、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] Steven J. DeRose. Grammatical Category Disambiguation by Statistical Optimization. *Computational Linguistics*, Vol. 14, No. 1, pp. 31–39, 1988.

[2] 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.

[3] 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.

[4] 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.

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

[6] 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.

[7] 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.

[8] 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.

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

[10] 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.

[11] 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.

[12] 永田昌明. EDR コーパスを用いた確率的日本語形態素解析. EDR 電子化辞書利用シンポジウム, pp. 49–56, 1995.

[13] 丸山宏, 荻野紫穂, 渡辺日出雄. 確率的形態素解析. 日本ソフトウェア科学会第 8 回大会論文集, pp. 177–180, 1991.

[14] 竹内孔一, 松本裕治. HMM による日本語形態素解析システムのパラメータ学習. 情報処理学会研究報告, 1995.

[15] 松本裕治, 黒橋禎夫, 宇津呂武仁, 妙木裕, 長尾眞. 日本語形態素解析システム JUMAN 使用説明書 version 2.0. 京都大学工学部長尾研究室, 1994.

[16] 日本電子化辞書研究所. EDR 電子化辞書仕様説明書, 1993.

[17] 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.

[18] 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.