2K-01 後続の品詞を考慮した形態素解析*

長田 靖† 吉田 敬一‡

静岡大学大学院理工学研究科§

1 はじめに

確率を用いた形態素解析では、一般に、隠れマルコフモデル (HMM) や n-gram モデルが用いられている。従来の方法では、単語の品詞の決定にその単語の前方の単語列の品詞列のみを考慮している。しかし、後ろに出てくる品詞列も単語の品詞の決定に影響するはずである。そこで、本研究では、後続の品詞列を考慮し、一つ前の単語、一つ後の単語との連接の情報を加えることにより、形態素解析の精度を向上させることを目的とした。

2 確率的日本語形態素解析

入力文字列 S が単語列 $W = w_1, ..., w_n$ に分割され、品詞列 $T = t_1, ..., t_n$ が付与されるとすると、確率を用いた日本語形態素解析においては、正しい (最もありそうな) 単語列と品詞列に一番高い確率を割り当てることができれば良い。そのため、形態素解析は単語列と品詞列の同時確率 $P(W,T)$ を最大化するような単語と品詞の対の集合を求める問題に帰着する [1]。これには一般に trigram が使われ、同時確率 $P(W,T)$ は次式で近似される [2]。

$$P(W,T) = \prod_{i=1}^{n} P(t_i|t_{i-2}, t_{i-1})P(w_i|t_i) \qquad (1)$$

ここで、$P(t_i|t_{i-2}, t_{i-1})$ は品詞 t_{i-2},t_{i-1} の後に品詞 t_i が現われる確率で、$P(w_i|t_i)$ は品詞別単語出現確率である。これらは、品詞タグ付きコーパスを用いて出現頻度をカウントすることにより、以下の式で求めることができる。

$$P(t_i|t_{i-2}, t_{i-1}) = \frac{C(t_{i-2}, t_{i-1}, t_i)}{C(t_{i-2}, t_{i-1})} \qquad (2)$$

$$P(w_i|t_i) = \frac{C(w_i, t_i)}{C(t_i)} \qquad (3)$$

ここで C は出現回数を表す。

*A Morphological Analysis Based on Part of Speech Subseqent to the Object Word

†Osamu Nagata

‡Keiichi Yoshida

§Grasuate Scool of Science and Engineering,Shizuoka University

3 提案する手法

提案する手法では、確率の計算において、後続の品詞列を考慮したモデルを考える。本研究においても、従来の手法と同様に、形態素解析を単語列と品詞列の同時確率 P(W,T) を最大化する問題として定義する。式 (1) からわかるように、従来の手法では、ある品詞 t_i は前の 2 つの品詞からのみ影響を受けると仮定している。本研究では、後続の品詞を考慮するため、以下のように P(W,T) を定義する。

$$\begin{aligned} & P(W,T) \\ &= P(W = w_{1,n}, T = t_{1,n}) && (4) \\ &= P(w_1|w_{2,n}, T)P(w_2|w_1, w_{3,n}, T)\ldots \\ &\quad P(t_1|t_{2,n})P(t_2|t_1, t_{3,n})\ldots && (5) \\ &= \prod_{i=1}^{n} P(w_i|w_{1,i-1}, w_{i+1,n}, T) \\ &\quad P(t_i|t_{1,i-1}, t_{i+1,n}) && (6) \end{aligned}$$

ここで、$w_{1,n}$、$t_{1,n}$ は、各々$w_1, \ldots w_n$、$t_1, \ldots t_n$ をあらわすものとする。(4) 式から (5) 式への変形に際しては、単語の出現確率を、その単語を除いた全ての単語、および全ての単語の品詞から影響を受け、品詞の出現確率を、その品詞を除いた全ての品詞から影響を受けると仮定する。このままでは計算量が多すぎて実用的でないため、さらに単語の出現確率は、一つ前の単語、一つ後の単語、その単語の品詞のみから影響を受けると仮定し、品詞の出現確率は、二つ前までの品詞、二つ後までの品詞の影響を受けると仮定することにより、以下の近似式を得る。

$$\begin{aligned} & P(w_i|w_{1,i-1}, w_{i+1,n}, T) \\ &= P(w_i|w_{i-1}, w_{i+1}, t_i) && (7) \\ &= P(w_i|w_{i-1})P(w_i|w_{i+1})P(w_i|t_i) && (8) \end{aligned}$$

(8) 式は (7) 式の近似である。
同様に次の式を得る。

$$\begin{aligned} & P(t_i|t_{1,i-1}, t_{i+1,n}) \\ &= P(t_i|w_i, t_{i-2}, t_{i-1}, t_{i+1}, t_{i+2}) && (9) \\ &= P(t_i|w_i)P(t_i|t_{i-2}, t_{i-1})P(t_i|t_{i-1}, t_{i+1}) \\ &\quad P(t_i|t_{i+1}, t_{i+2}) && (10) \end{aligned}$$

よって、(8) 式と (10) 式を 6 式に代入することにより、以下の式を得る。

$$
\begin{aligned}
& P(W,T) \\
&= \prod_{i=1}^{n} P(w_i|w_{i-1})P(w_i|w_{i+1})P(w_i|t_i) \\
& \quad P(t_i|t_{i-2},t_{i-1})P(t_i|t_{i-1},t_{i+1}) \\
& \quad P(t_i|t_{i+1},t_{i+2}) \qquad (11)
\end{aligned}
$$

従来の手法では、(1) 式を用い、ヴィテルビアルゴリズムを用いて前向きに解析する。しかし、(11) 式からもわかるように、本研究では確率の計算に後ろの品詞を必要とするため、確率の計算の時点で後続の品詞が決定していないので、その手法を使うことはできず、全ての単語列と品詞列の組合せ (これをパスと呼ぶ) を求める必要がある。しかし、全てのパスを求めるとなると、その数は非常に膨大になってしまい、実装に耐えうるものではない。ところが、(1) 式と (11) 式からわかるように、従来の確率の式 (1) に含まれる項は本手法の確率の式 (11) の中に全て含まれている。そこで、以下の手順で解析を行う。

1. (1) 式で解析し、確率の高い上位 N 個を残し、順位付けする。
2. 残した N 個それぞれについて (11) 式で確率を再計算し、順位を並び変える。
3. 再計算した時の最も確率の高いものを解析結果とする。

4 実験

実験には日本電子化辞書研究所の EDR コーパスを用い、そのコーパスの中からランダムに 1 万文を選んでトレーニングを行った。トレーニングに用いた文の中からランダムに千文でクローズドテストを行い、それ以外の文の中からランダムに千文でオープンテストを行った。比較対象となる従来の手法としては、(1) 式を用いた。評価尺度として、以下のものを用いた。

$$
\text{単語の再現率} = \frac{\text{正解データと一致した単語分割の数}}{\text{正解データに含まれる単語の数}}
$$

$$
\text{単語の適合率} = \frac{\text{正解データと一致した単語分割の数}}{\text{解析結果に含まれる単語の数}}
$$

$$
\text{品詞の再現率} = \frac{\text{単語分割と品詞付与ともに一致した数}}{\text{正解データに含まれる単語の数}}
$$

$$
\text{品詞の適合率} = \frac{\text{単語分割と品詞付与ともに一致した数}}{\text{解析結果に含まれる単語の数}}
$$

単語の再現率、適合率は単語分割の精度を表し、品詞の再現率、適合率は品詞付与の精度を表している。結果を表 1、表 2 に示す。

表 1: クローズドテストの結果

	従来の手法	本手法
単語の再現率	0.933	0.977
単語の適合率	0.965	0.990
品詞の再現率	0.909	0.938
品詞の適合率	0.939	0.959

表 2: オープンテストの結果

	従来の手法	本手法
単語の再現率	0.908	0.920
単語の適合率	0.887	0.890
品詞の再現率	0.872	0.865
品詞の適合率	0.850	0.843

5 評価

クローズドテストにおいては、従来の手法よりも多くの情報を用いたため、全てにおいて 2%～4.4%程度の精度の向上が見られた。オープンテストでは、単語の再現率、適合率は精度が向上しているが、品詞付与については、どちらも 0.7%程度精度が落ちている。これは、多くの情報を用いたために、スパースデータの問題が生じたためと思われ、トレーニングデータを増やしたり、スムージングを行うことで解消できると考えられる。

参考文献

[1] Charniak,E.Stastical langage learning. MIT Press,Cambridge,1993.

[2] 永田 昌明：前向き DP 後向き A*アルゴリズムを用いた確率的日本語形態素解析システム，自然言語処理 101-10,pp.74-80,1994.

[3] 長尾 真 編：自然言語処理，岩波書店,1996.

[4] 平沢 克宏：確率モデルを用いた日本語形態素解析，静岡大学大学院理工学研究科 計算機工学専攻 修士論文，1998.