確率モデルを用いた日本語形態素解析 (第2報)*

4AE−3

平沢 克宏 † 吉田 敬一 ‡

静岡大学大学院理工学研究科§

1 はじめに

統計を用いた形態素解析ではn-gramモデルやHMMモデルがよく使用され、英語においては高い精度で解析が行われている。しかし、単語の区切りを必要とする日本語に、単語の区切りを必要としない英語のモデルをそのまま適用するのには問題がある。本研究では、タグ付きコーパスから求められた確率をそのまま使用するのではなく、解析結果がコーパスの確率に近付くようにトレーニングを行うことにより、解析精度を向上させることを目的とする。

2 確率モデル

入力文字列 $S = s_1...s_m$ 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。形態素解析は単語列と品詞列の同時確率 $P(W,T)$ を最大化する単語分割と品詞分割の組を求める問題に帰着する [1]。一般に確率モデルにはbigramやtrigramなどのn-gramモデルが使われ、同時確率 $P(W,T)$ は次式で近似される [2]。

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

$P(t_i|t_{i-2},t_{i-1})$ は品詞 t_{i-2},t_{i-1} の後に品詞 t_i が出現する確率 (品詞三つ組確率) で $P(w_i|t_i)$ は品詞別単語出現確率である。品詞タグ付きコーパスにおいては単語の出現頻度をカウントする事により、以下の式により確率を求めることができる。

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

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

表 1: 形態素の数

コーパスの形態素数	解析結果の形態素数
24584	24556

表 2: 品詞の出現頻度

	コーパスの頻度	解析結果の頻度
名詞	0.249	0.252
助動詞	0.059	0.049
語尾	0.118	0.105

ここで C は出現回数を表し、$C(t_{i-2},t_{i-1})$ は品詞列 t_{i-2},t_{i-1} のが現れた回数、$C(w_i,t_i)$ は単語 w_i が品詞 t_i であった回数を表す。

3 パラメータの再推定

式(1)のモデルでは形態素の数が異なる解析候補の値を比較するため、分割数の少ない形態素列が優先されることになる。「最長一致」や「形態素数最小法」などの有効性が知られており、この性質がうまく取り入れられているため精度は高いものとなるが、解析結果に偏りが生じる。つまり形態素数が少ないものが優先されるため、単語分割において必要以上にまとめあげる傾向が生まれる [表1]。また文字数の少ない形態素が選ばれにくくなり助動詞などの出現頻度が小さくなる [表2]。そのためコーパスで求められた確率をそのまま使用するのではなく、値を調整する必要がある。本研究では解析結果の確率が、タグ付きコーパスで求められた確率に近付くようにトレーニングを行った。新しい確率を求めるのに次式を用いた。

$$EST\_P = EST\_P + (ORG\_P - OUT\_P) \quad (4)$$

*A Japanese Morphological Analysis Using Probability Models

†Katsuhiro Hirasawa

‡Keiichi Yoshida

§Graduate School of Science and Engineering, Shizuoka University

表 3: クローズドテストの結果

		初期確率 (コーパス)	トレーニング後
unigram	単語分割	0.950	0.971
	品詞付与	0.885	0.902
bigram	単語分割	0.968	0.980
	品詞付与	0.938	0.946
trigram	単語分割	0.971	0.980
	品詞付与	0.951	0.955

表 4: オープンテストの結果

		初期確率 (コーパス)	トレーニング後
unigram	単語分割	0.933	0.961
	品詞付与	0.865	0.887
bigram	単語分割	0.962	0.975
	品詞付与	0.928	0.937
trigram	単語分割	0.969	0.975
	品詞付与	0.945	0.949

ここでORG_Pはコーパスから求められた確率である。EST_Pには初期確率としてORG_Pが入る。式(1)でEST_Pを使い解析を行いその解析結果からOUT_Pを求め、式(4)を使い新たな確率EST_Pを求める。これを解析結果から求められた確率がコーパスで求められた確率に近付くまで繰り返す。unigram,bigram,trigramのそれぞれについて、式(4)で確率の再推定を行う。

4 実験

実験には日本電子化辞書研究所のEDRコーパスを用いた。このコーパスは新聞記事などの例文が約２１万文収録されており、形態素や構文の情報が付加されている。EDRコーパスの中からランダムに１万文(246536単語)を選んで初期確率を求め、そのうちの１千文(24854単語)をクローズドテストに用い、それ以外の１千文(24638単語)をオープンテストに用いた。クローズドテストでは式(4)を用いて２０回、値の再推定を繰り返した。オープンテストではトレーニングで推定された値とコーパスから求められた値を使って、解析を行い、結果を比較した[表4]。これらをunigram,bigram,trigramのそれぞれについて実験を行った。ここでは各文の形態素列の候補について式(1)の確率が最も高い解を採用する方式をとった[表3,表4]。それぞれの値はコーパスの正解データと一致した単語分割の数と品詞付与の数を、正解データに含まれる単語の数で割った値である。

5 評価と今後の課題

トレーニング後の確率を用いた場合、クローズドテスト、オープンテスト共に0.5%～2%解析精度を向上させる事ができた。unigram,bigram,trigramでは、trigramモデル精度が最も高かったが、向上の幅はunigramモデルが最も大きかった。これは１万文の学習では出現頻度の小さいtrigramについての情報が不十分であったためと考えられ、トレーニングデータを増やせばさらに精度の向上が可能と思われる。

また確率の再推定に式(4)を用いたが、この式はコーパスと結果との差を元の確率に加えるという単純なものなので、今後の検討が必要である。

本研究では解析結果の品詞の出現確率がコーパスから得られた確率に近付くようにトレーニングを行う事により、解析精度の向上を試みた。一般にタグ付きコーパスを使用する場合コーパスから確率が得られるためトレーニングは行われないが、式(1)のような異なった形態素数の積を比較するモデルでは確率の再推定が必要であり、その有効性を示した。

参考文献

[1] Charniak,E.Statistical language 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] 山本、増山：品詞・区切り情報を含む拡張文字の連鎖確率を用いた日本語形態素解析, 言語処理学会代３回年次大会発表論文集,421-424,1997.

[5] 平沢、吉田：確率モデルを用いた日本語形態素解析, 情報処理学会第53回全国大会 pp.2-5-6,1996