確率モデルを用いた日本語形態素解析*

1L－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

意味を担う最小の言語要素を形態素と呼び、言語処理において、語幹・接辞・語形変化などを同定することを形態素解析と呼ぶ。従来は規則に基づく方法が主流であったが、近年、統計モデルを用いた方法が提案されている。統計を用いた形態素解析では一般にHMM(隠れマルコフモデル)が使用され、高い解析精度が得られることが報告されている。本研究ではHMMにd-bigram[3]の概念を導入した確率モデルを構築し、その有用性を調べ、より高い精度で形態素解析を行いうることを示す。

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]。ここで同時確率 $P(W,T)$ を近似するのにどのような確率モデルを使用するかが重要な問題となる。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]。

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

$P(t_i|t_{i-2},t_{i-1})$ 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。n-gramモデルは $O(x^n)$(x は品詞の数)の記憶容量を必要とし、また n の値の増加に従いスパースデータが発生しやすくなる傾向から考えてnの値を大きくすることはあまり現実的でない。

そこで本研究では距離の離れた形態素間の情報を利用するためにd-bigramの概念を採り入れた確率モデルにより同時確率 $P(W,T)$ を次式で近似する。

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

$P(t_i|t_{i-d})$ は品詞 t_{i-d} の距離d後に品詞 t_i が現れる確率、$P(w_i|t_i)$ は品詞別単語出現確率を表している。$P(t_i|t_{i-d})$ と $P(w_i|t_i)$ はコーパスからそれぞれ事象の出現回数をカウントすることにより以下の式を用いて推定できる。

$$P(t_i|t_{i-d}) = \frac{C(t_i,t_{i-d},d)}{C(t_{i-d})} \qquad (3)$$

$$P(w_i|t_i) = \frac{C(w_i,t_i)}{C(t_i)} \qquad (4)$$

ここで C は出現回数を表し、$C(t_i,t_{i-d},d)$ は品詞 t_{i-d} の距離d後に品詞 t_i 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 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。このコーパスに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(245983単語)を選んでトレーニングを行い、そのうちの1千文(24524単語)をクローズドテスト、それ以外の1千文(24638単語)をオープンテストに用いた。

アルゴリズムは文の先頭から走査していき、文末において式(2)の確率が最も高い解を採用する方式をとった。

最初に形態素の決定に際し距離の離れた品詞がどの程度の影響を持つかを調べるために次式で近似される確率モデルでdの値を変化させながら実験を行った。

$$P(W,T) = \prod_{i=1}^{n} P(t_i|t_{i-d})P(w_i|t_i) \qquad (5)$$

式(5)はある品詞が出現する確率を距離d離れた品詞との関係のみによって求めることを意味し、その結果

---

*A Japanese Morphological Analysis Using Statistical Models

†Katsuhiro Hirasawa

‡Keiichi Yoshida

§Graduate School of Science and Engineering, Shizuoka University

	単語分割	品詞付与
d=0	96.12%	88.89%
d=1(bigram)	96.74%	94.30%
d=2	96.20%	92.13%
d=3	96.16%	90.90%
d=4	96.17%	90.25%
d=5	96.18%	90.07%

表 1: 距離 d の品詞のみによる解析結果

	単語分割	品詞付与
d=1(bigram)	96.74%	94.30%
d=2	97.05%	94.56%
d=3	97.11%	93.50%
d=4	97.05%	92.57%
d=5	97.00%	91.91%
trigram	97.08%	94.82%

表 2: クローズドテストの解析結果

を表 1 に示す。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。

表 1 の傾向として d=1 の時の精度が最も良く d の値を増加させるにつれ次第に悪くなっている。しかしどの値も d=0 の品詞の相関を使用しない場合よりも高い値であった。このことから品詞の決定には距離の離れた形態素の情報も手掛かりとなるが、距離が近いほど与える影響が大きいといえる。

次に式 (2) を用いて同様に 1000 文のテストデータに対して実験を行った。また比較のために式 (1) の trigram を使った確率モデルとも比較した。クローズドテストの結果を表 2 に示す。

単語分割に関しては d=2 以降はどれも bigram よりも高く、 d=3 の時が最も高い値であった。品詞付与は trigram モデルの精度が最も良く、次に d=2 の時であった。 d=3 以降は次第に精度が悪くなって行った。これは距離が離れる程、品詞間の影響が小さくなりノイズが大きくなるためと思われる。
同様の傾向がオープンテストの結果にも見られた。

	単語分割	品詞付与
d=1(bigram)	95.13%	92.75%
d=2	95.71%	93.00%
trigram	95.03%	91.76%

表 3: オープンテストの解析結果

4 評価

d=2 の時に trigram よりも少ない情報量にもかかわらず、ほぼ同じような精度であったことから式 (2) の有効性が確認された。またオープンテストではスパースデータの問題のため d=2 の時に trigram よりも高い値であった。

本研究では n-gram モデルでは距離の離れた形態素の情報を利用するために式 (2) のような確率モデルを構築した。表 1 から距離の離れた形態素も有効な情報を持っていることが予測されたが、残念ながら d の値を大きくすることによる精度の向上は達成できなかった。
式 (2) ではただ単に $P(t_i|t_{i-d})$ の積をとり、距離による影響を無視している。距離が近い程関係が強く、距離が遠い程ノイズが大きくなるならば、距離による重み付けが必要である。また d の値を大きくしても品詞別単語出現確率 $P(w_i|t_i)$ の影響が小さくならないような式の工夫も必要である。

今回は確率モデルの有効性を調べるために前向き探索のみの単純なアルゴリズムを用いた。距離の離れた形態素間の情報を有効に使用し、できるだけノイズを減らせるような情報の形態とアルゴリズムが構築できればより一層の解析精度の向上が可能と思われる。
また日本語よりも英語の方が離れた形態素間の規則性が強いと思われるので EDR 英語コーパスについても同様の実験を行いその傾向を調べてみたい。

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] 佐野 智久 ほか : d-bigram を用いた自然言語文評価に関する実験, 情報 処理学会第 51 回全国大会,Vol.3,pp.3-4,1995.

[4] 延澤 志保 ほか : d-bigram を用いた形態素解析, 情報処理学会第 51 回全 国大会, Vol.3,pp.25-26,1995.