Untagged-corpusを用いた形態素解析用HMMパラメータの一推定法

山本幹雄（筑波大学 電子・情報工学系）

1．はじめに

確率モデルに基づく形態素解析システムの各種パラメータをタグなしコーパスから推定できれば、次のような利点・利用法がある。(1)作成コストの大きなタグ付きコーパスを開発するためのブートストラップ的なツールとして使用できる。(2)タグ付きコーパスよりも桁違いに大きなコーパスを学習データとできる。(3)特定のドメインへのシステムの適応化に使える。

本報告では、タグなしコーパスからの新しい形態素解析用HMMパラメータ推定手法を提案し、予備実験結果について報告する。

2．従来方法と提案手法

タグなしコーパスからの形態素解析用確率モデルの推定法は、[Kupiec92]によって提案された方法が基本である。これは、HMMの観測シンボルを単語とし、状態が品詞を表現していると見なし、Baum-Welchのパラメータ推定アルゴリズム[中川88など]を用いて、結果的に品詞のbigramと品詞で条件付けられた単語の確率を得る方法である。まず、英語へ適用され、その有用性が実証されている[Cutting92]。また、[竹内95]では、Cuttingらの方法を単語境界が分からない日本語へ適用できるように拡張し、日本語の形態素解析で成果を上げている。

しかし、いずれの方法も学習によって得られるモデルがbigramであるため、タグ付きコーパスを用いた場合によく使われるtrigramよりもモデルの能力が低い。Kuipecの方法をtrigramに拡張することも可能であるが（批判的論文ではあるが[Merialdo94]で使われている）、HMMの状態として品詞の2つ組みを持たなければないため、状態遷移のパラメータが品詞の3乗となる。後で述べる実験で用いたような品詞数が260くらいのシステムになると推定すべき状態遷移のパラメータだけで$260^3 \fallingdotseq 17.6$M個となり、パラメータの個数が大きくなりすぎる。

一方、[周94]の報告では、HMMがbigramよりも能力が高く、trigramに匹敵する能力をより少ないパラメータ数で獲得できることを、タグ付きコーパスを学習データとする英語と日本語の品詞予測実験で示している。

以下では、学習されるモデルがN-gramではなくHMMであるような、タグなしコーパスからの確率モデルパラメータ推定手法を提案し、予備実験について報告する。

3．曖昧観測シンボルからのHMMパラメータ推定手法

3.1 HMMによる形態素解析

日本語における形態素解析では、単語分割と品詞の付与を同時に行わなければならない。ここで、単語の系列を$W=w_1,\ldots,w_n$ 品詞の系列を$T=t_1,\ldots,t_n$とすると、確率を使った形態素解析は単語列と品詞列の同時確率P(W,T)を最大化する問題に帰着される[永田94]。

$$(\hat{W},\hat{T}) = \arg\max_{W,T} p(W,T \mid S) = \arg\max_{W,T} p(W,T) \quad (1)$$

このときp(W,T)をどのように推定するかによっていくつかのモデルが考えられる。品詞N-gramを用いる場合を(2)式、HMMを用いる場合を(3)式に示す。(2)式で、N=1の場合は品詞bigram、N=2の場合は品詞trigramを使ったモデルと呼ばれる。

$$p(W,T) = \prod_{i=1}^{n} p(t_i \mid t_{i-N}\ldots t_{i-1}) p(w_i \mid t_i) \quad (2)$$

$$p(W,T) = \sum_{x} \prod_{i=0}^{n-1} a_{x(i),x(i+1)} b_{x(i+1)}(w_{i+1}, t_{i+1}) \quad (3)$$

$$= \sum_{x} \prod_{i=0}^{n-1} a_{x(i),x(i+1)} b'_{x(i+1)}(t_{i+1}) p(w_{i+1} \mid t_{i+1}) \quad (3)'$$

ここで、xはHMMの可能な状態間のパスを表わ

し、$x(i)$はあるパス上のi番めに遷移する状態である。$a_{x(i),x(i+1)}$は状態 $x(i)$から状態 $x(i+1)$への遷移確率であるが、特に$a_{x(0),x(1)}$は状態$x(1)$の初期状態確率$\pi_{x(1)}$を表わす。$b_{x(i)}(w,t)$、$b'_{x(i)}(t)$は状態$x(i)$で品詞がtの単語w、あるいは品詞tを出力する確率である。(3)式は出力シンボルとして単語と品詞のペアを考えた場合、(3)'はHMMの出力シンボルは品詞として、$p(w|t)$を乗ずることによってペアの出力確率を近似したものである。

HMMの場合、N-gramと異なり、状態は形態素系列をなんらかの意味で抽象化しているものと考えることができる。すなわち、可変長のN-gramともとらえることが可能である。

次節では、(3),(3)' 式のHMMのパラメータをタグなしコーパスから推定する方法を述べる。

3.2 パラメータ推定手法

3.2.1 形態素ネットワーク

タグなしデータが与えられると、まず辞書引きにより可能な形態素のネットワークを生成する。この際、連接制約によってある程度信頼性を上げることも可能ならば行った方がよい [竹内 95] 。得られた形態素ネットワークを以下のように定義する。

m_s：s番目の形態素（番号sを持った形態素）
w_sまたは$word(s)$：s番目の形態素の単語（見出し）
t_sまたは$tag(s)$：s番目の形態素のタグ（品詞）
$suc(s)$：形態素ネットワーク上で、m_sの後(文上で右)に接続している形態素の番号の集合
$pre(s)$：形態素ネットワーク上で、m_sの前(文上で左)に接続している形態素の番号の集合

形態素ネットワークの例は図1を参照のこと。

3.2.2 HMMパラメータの推定法

観測シンボル列が、形態素ネットワークとして曖昧に与えられた場合のHMMの状態遷移確率、出力確率、初期状態確率(それぞれ、$\boldsymbol{a},\boldsymbol{b},\boldsymbol{\pi}$)の再推定式を考える。基本的な考え方は、形態素ネットワークに対応したネットワーク状になったトレリスを考え、その上で従来の再推定式を拡張すればよい。すなわち、従来の前向き・後ろ向き確率が時間(位置)同期的に定義されていたものをネットワーク的に拡張し、形態素ネットワーク上の各形態素ごとに前向き・後ろ向き確率が定義される。

各形態素における前向き・後ろ向き確率は、初期値として文頭と文末の形態素 u と v に関して、$\alpha_u(j)=\pi_j b_j(w_u,t_u)$、$\beta_v(i)=1$を与えれば、次のように再帰的に定義される。

$$\alpha_r(j)=\sum_{s\in pre(r)}\sum_{i=1}^{N}\alpha_s(i)a_{ij}b_j(w_r,t_r) \tag{4}$$

$$\beta_s(i)=\sum_{r\in suc(s)}\sum_{j=1}^{N}a_{ij}b_j(w_r,t_r)\beta_r(j) \tag{5}$$

(4),(5) 式で求められた前向き・後ろ向き確率を用いて、HMM の各パラメータは以下のように再推定される。ここで、kは学習データとしての文の番号、p_kはその生起確率である。

$$\bar{a}_{ij}=\frac{\sum_{k=1}^{K}\frac{1}{p_k}\sum_{(r,s)\in\{(u,v)|u\in pre(v)\}}\alpha_r^k(i)a_{ij}b_j(t_s)\beta_s^k(j)}{\sum_{k=1}^{K}\frac{1}{p_k}\sum_{s}\alpha_s^k(i)\beta_s^k(i)} \tag{6}$$

$$\bar{b}_i(w,t)=\frac{\sum_{k=1}^{K}\frac{1}{p_k}\sum_{s\in\{r|word(r)=w,tag(r)=t\}}\alpha_s^k(i)\beta_s^k(i)}{\sum_{k=1}^{K}\frac{1}{p_k}\sum_{s}\alpha_s^k(i)\beta_s^k(i)} \tag{7}$$

$$\bar{\pi}_i=\frac{\sum_{k=1}^{K}\frac{1}{p_k}\sum_{s\in on(1)}\alpha_s^k(i)\beta_s^k(i)}{\sum_{k=1}^{K}\frac{1}{p_k}\sum_{s\in on(1)}\sum_{j=1}^{N}\alpha_s^k(j)\beta_s^k(j)} \tag{8}$$

3.3 スケーリング

前向き・後ろ向き確率を求める際に、語彙が大きく、入力文中の形態素連鎖が長いと、アンダーフローの問題が生じる。例えば、EDRコーパスのように形態素の単位が小さい場合、形態素数も多くなるため、アンダーフローが実際に生じる。ここでは、形態素ネットワークからのHMMパラメータ推定時のスケーリング手法を提案する。

以下のように、スケーリングのために用いる同期点という概念を定義する。

形態素の開始点：形態素の見出し語の先頭の文字が文上にある位置
形態素の範囲：形態素の見出し語の文字が文上にある位置の集合
同期点：形態素の開始点
同期点の集合：すべての形態素の開始点の集合
同期点の番号：同期点のうち、文上で左にあるものから順に1,2,3,...とする。

B：同期点の最大の番号

$on(q)$： 番号がqである同期点をその範囲に含む形態素の番号の集合

L_s：s番目の形態素の開始点に対応する同期点の番号

R_S：s番目の形態素の範囲内にある同期点のうち、最も大きな番号

図1に形態素ネットワークと同期点の例を示す。「はきもの」と「きもの」という形態素の可能性によって文字の位置4と5が同期点となっている。また、「はきもの」という単語は同期点3,4をその範囲に含み、Lsが3, Rsが4である。

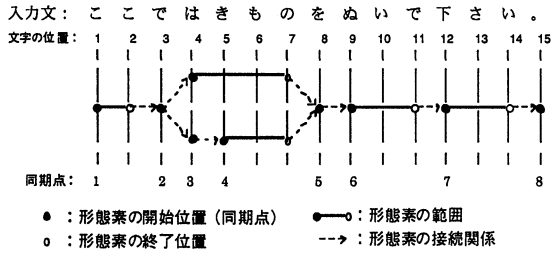

図1　形態素ネットワークと同期点

以上のような同期点に関する定義を用いると、スケーリングされた前向き確率は各形態素の各同期点ごとに次のように定義される。

$\hat{\hat{\alpha}}_{sl}(i)$：番号sの形態素の同期点$l$での状態$i$に関する（前向き確率の）中間段階確率

c_l：同期点lにおける scaling factor

$\hat{\alpha}_{sl}(i)$：番号sの形態素の同期点lでの状態iに関する scaling された前向き確率

初期値：（ここで1は数字であることに注意）

$$\hat{\hat{\alpha}}_{s1}(i)=\alpha_{s1}(i)=\pi_i b_i(w_s,t_s) \quad \text{where} \quad s\in on(1) \tag{9}$$

$$c_1=1/\sum_{s\in on(1)}\sum_{i=1}^{N}\hat{\hat{\alpha}}_{s1}(i) \tag{10}$$

$$\hat{\alpha}_{s1}(i)=c_1\hat{\hat{\alpha}}_{s1}(i) \quad \text{where} \quad s\in on(1) \tag{11}$$

再帰的な定義：（ここでlは英記号であることに注意）

$$\hat{\hat{\alpha}}_{sl}(j)=\begin{cases}\hat{\alpha}_{s,l-1}(i) & \text{if } L_s\neq l\\ \sum_{r\in pre(s)}\sum_{i=1}^{N}\hat{\alpha}_{r,l-1}(i)a_{ij}b_{tj}(w_s,t_s) & \text{if } L_s=l\end{cases} \tag{12}$$

$$c_l=1/\sum_{s\in on(l)}\sum_{i=1}^{N}\hat{\hat{\alpha}}_{sl}(i) \tag{13}$$

$$\hat{\alpha}_{sl}(i)=c_l\hat{\hat{\alpha}}_{sl}(i) \tag{14}$$

上記の scaling された前向き確率は同期点に関して左から右に位置同期的に計算することができる。

後ろ向き確率$\hat{\beta}_{sl}(i)$は、前向き確率を求めた際のスケーリングファクタcを用いて、同様に定義される。

このようにして求められた前向き・後ろ向き確率は以下のような性質を持っている。ここで、$\hat{\alpha}_s=\hat{\alpha}_{sR_s}$、$\hat{\beta}_s=\hat{\beta}_{sL_s}$とする。

$$\hat{\alpha}_s(i)a_{ij}b_j(w_r,t_r)\hat{\beta}_r(j)=\frac{1}{p_k}\alpha_s(i)a_{ij}b_j(w_r,t_r)\beta_r(j)$$

この性質をもとに再推定式を書き換えれば、スケーリングされると共にp_kを計算する必要がなくなる。ただし、$\hat{\alpha}_s^k\hat{\beta}_s^k\neq\alpha_s^k\beta_s^k/p_k$であるため、分母は若干複雑になることに注意。

4．実験

実験では、形態素ネットワークを 13 万単語の辞書を持つ Juman[松本 94]で生成し、3節で述べた再推定式を用いてパラメータの推定を行った。また、形態素解析を行うときも、同様に Juman によってまず形態素ネットワークを生成し、その中で最も確率の高いパスを Viterbi アルゴリズムによって決定し、形態素列を出力する。

データとしては、日経新聞 94 年版[日経 94]を用いた。学習用として、10 日分の記事 26108 文(長さ 150 文字以上の文は削除した結果)、テスト用として、学習用以外の 1 日分の記事から 100 文をランダムに取り出した。テスト用は人手によってタグ付けを行った(形態素数約 2500)。モデルの品詞としては、Juman の出力する品詞、品詞細分類、活用型、活用形の組み合わせ 264 種類を用いた。正解は(上記組み合わせとしての)品詞、見出し、基本形が一致した場合とした。ただし、正解の判断は、人手でどちらの品詞にするべきか悩んだものはどれでも正解ということにした。また、固有名詞で未定義語となるものが多かったが、これらは Juman が未定義語として出力するデフォルトの品詞（本実験ではサ変名詞）でも正解とした。適合率の計算は、「正解出力の数」／「テスト文の全(正解)形態素数」[永田 94]で求めた。

Juman が学習用の形態素ネットワークを生成するときの制約の大きさを表わすコスト幅は、テスト文に対して、適合率約 70%、再現率約 97%となるコスト幅 70 を固定で用いた。また、初期モデルとしてすべてのパラメータを等確率としたものを用いた。

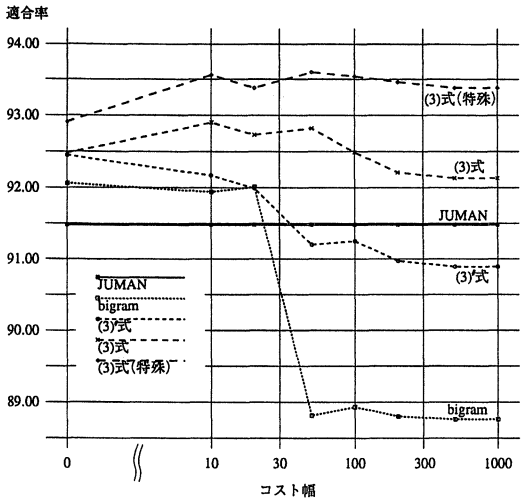

図2　実験結果

結果を図2に示す。縦軸は適合率、横軸はJumanが出力する候補としての形態素ネットワークの大きさを決定するコスト幅の値である。コスト幅が大きいほど、大きな形態素ネットワークが出力される。大きくなると正解が入っている可能性も大きくなるが、誤る可能性も大きくなる。この中から確率モデルを用いて正解を選択し、出力する。図中Jumanと書いてある直線は確率モデルを用いないオリジナルのJumanのテスト文に対する適合率である。bigramと書いてある線は、[竹内95]と同じように状態を品詞に対応させて、bigramを推定する手法による適合率である(再推定の繰り返しは8回)。ただし、[竹内95]と異なり、動詞に関する遷移の後ろと前で品詞を変えるなどの工夫はしていない。式(3)'と書いた線は式(3)'のモデルを学習したもの(10状態, 繰り返し5回)、式(3)と書いた線は式(3)のモデルを学習した(10状態,繰り返し5回)システムの結果である。式(3)のモデルは出力確率が単語と品詞のペアとなり、総単語数の状態数倍のパラメータが必要であるが、その分、モデルの能力が高く、以上3つの中では最もよい結果となった。

(3)式(特殊)とあるものは、学習の条件が違う。学習データとして2ヶ月分約15万文を用いて、学習の繰り返しごとにコスト幅を0から70まで徐々に増加させて(3)式のモデル(10状態)を学習させた。また、品詞数も約100個と少な目に設定してある（適合率の計算はまったく同条件）。これより、品詞数や学習の計画がモデルの精度に大きな影響を与えることが分かる(学習データの数はこれくらいの規模ではそれほど影響しないことが実験より分かっている)。これらの検討は今後の課題である。

5．おわりに

タグなしコーパスからのHMMパラメータ推定法を提案し、有効であることを示した。今後の課題としては、タグなしコーパスからの学習の限界を踏まえ[Merialdo94]、より精度の高いモデルとするために、タグなしコーパスとタグ付きコーパスから求めた2つのモデルの融合を検討する予定である。

謝辞

Juman を開発された京都大学 長尾研究室・奈良先端科学技術大学院大学 松本研究室の皆様、日頃議論していただく、筑波大学 知能情報・生体工学研究室の皆様、豊橋技術科学大学 中川研究室の皆様に感謝いたします。

参考文献

[Cutting92] D. Cutting, J. Kupiec, J. Pedersen and P. Sibun: A practical part-of-speech tagger, ANLP-92,pp.133-140, 1992.

[Kupiec92] J. Kupiec: Robust part-of-speech tagging using a hidden Markov model, Computer Speech and Language, Vol.6, pp.225-242, 1992.

[Merialdo94] B. Merialdo: Tagging English text with a probabilistic model, Computational Linguistics, Vol.20, No.2, pp.155-171, 1994.

[周 94] Min Zhou, 中川聖一: 日本語及び英語の確率言語モデルに関する検討, 「自然言語における学習」シンポジウム論文集, pp.57-64, 1994.11.

[竹内 95]竹内、松本:「HMMによる日本語形態素解析システムのパラメータ学習」、情報処理学会研究会報告, 自然言語処理研究会, NL-108-3, pp.13-19, 1995.7.

[中川 88]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.

[永田 94]永田昌明: 前向きDP後ろ向き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, 情報処理学会研究会報告, 自然言語処理研究会, NL-101-10, pp.73-80, 1994.5.

[松本 94] 松本祐治、他：「日本語形態素解析システムJUMAN使用説明書 version2.0」, 1994.

[日経 94] 日本経済新聞 CD-ROM 版(1994年版)、日本経済新聞社、1995.