HMMを用いた形態素解析のパラメータ学習

1R－4

竹内孔一　松本裕治
奈良先端科学技術大学院大学

1 はじめに

日本語の形態素解析は自然言語処理を行なう上で最も基本的でかつ重要な処理である。我々の研究室が公開している形態素解析システム JUMAN[1] は品詞の連接と単語に対してコストによる制約を与えることで曖昧性の絞り込みを行なっている。しかし、このコスト値は対象とするテキストの分野によって左右されるが、それを最適化する機構が存在しなかった。

そこで本報告では、最近、機械学習などで良く用いられている HMM を利用して、JUMAN システムに対応する HMM システムを構築して、タグ付きタグ無しコーパスによる学習を行なうことで、コスト値、すなわちパラメータの最適化を行なう。

2 HMM による学習システム

まず、今まで人手で決定してきた、コスト値は全て廃止し、ＥＤＲのタグつきコーパスを利用して初期の連接と単語の確率を獲得する。その後、タグのついていないコーパスについて HMM による学習を数回繰り返して行なう。

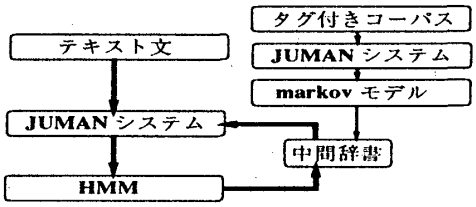

図1 HMM によるパラメータ学習モデル

2.1 初期値の獲得

EDR タグつきコーパスを利用してＨＭＭの初期確率、すなわち連接確率と単語の確率を獲得する。そのためには、（１）JUMAN の文法体系に即したタグつきコーパスを獲得する。（２）タグつきコーパスの結果をもとに、markov モデルによって、単純に頻度をカウントすることで、連接確率と単語の確率を獲得する。という２つのステップを必要する。

Learning Costs of Japanease Morphological Analyzer Using Hidden Markov Model
Kouichi Takeuchi, Yuji Matsumoto
Nara Institute of Science and Technology

2.1.1 タグつきコーパスの獲得

EDR のタグつきコーパスは文法体系は JUMAN のものと異なるため、なんらかの方法で JUMAN の文法体系に変換する必要がある。そこで、EDR コーパスからわかち書き情報を利用して現在ある JUMAN に解析させて自動的にタグづけを行なう。ただし、品詞情報を JUMAN に与えないので、正確なタグづけが行なわれる保証はない。

2.1.2 markov モデルによる初期値の獲得

前節で得られたタグつきコーパスから markov モデルによって初期値を得る。すなわち（１）式で用いる最初の連接確率 $P(t_i|t_{i-1})$ と品詞別単語の確率 $P(w_i|t_i)$ は出現頻度を数えるだけで求めることができる [2]。この結果を HMM の初期確率とする。

2.2 HMM の構築

2.2.1 JUMAN に対応した HMM

HMM は英文におけるタグづけにおいてよく用いられ、高い精度の結果が得られているが、日本語文ではわかち書きされていないため、そのまま用いることができない。そこで、ある入力文の文字列 L から得られる１つの単語列を $w_{1,n} = w_1, w_2, ..., w_n$ として、各単語に品詞系列 $t = t_1, t_2, ..., t_n$ を付与すると考えると、入力文 L に対する確率は、$w_{1,n} \in L$ のなかで可能な組合せを全て足し込めば良いから、

$$
\begin{aligned}
P(L) &= \sum_{w_{1,n+1} \in L} P(w_{1,n+1}) \\
&= \sum_{w_{1,n+1} \in L} \sum_{t_{0,n+1}} P(w_{1,n+1}, t_{0,n+1}) \\
&= \sum_{w_{1,n+1} \in L} \sum_{t_{0,n+1}} \prod_{i=1}^{n+1} P(w_i|t_i) P(t_i|t_{i-1}) \quad (1)
\end{aligned}
$$

となる。ここで、t_0は'文頭'、t_{n+1}は'文末'という品詞で、w_{n+1}は文末に遷移するために設けた空語である。これにより、全ての文章は必ず、品詞'文頭'から始まって'文末'で終了する lattice 構造をとることになる。よって最適な単語系列と品詞系列を求めることは、与えられた入力文字列 L に対して、確率 $P(w_{1,n}, t_{0,n+1})$ を最大化する単語列と品詞列の組合せを求めることになる。

2.2.2 学習について

通常のHMMは入力文字列を与えて、あらゆる可能な状態遷移のパスに対して（つまり、状態はhiddenとして）確率を計算する。我々はJUMANのコスト幅を緩めて出現する曖昧性の範囲内での状態遷移のパスについてのみ確率の足し込みを行なう。この制約により計算量が少なくて済むことと、明らかにおかしな解釈を学習の対象外にすることができる。再推定の場合、確率的回数のγをもとにして、品詞接続確率a_{ij}とある品詞jでの単語w_kの出力確率$b_j(w_k)$を再推定する。

$$\gamma(i,j,w_k) = \frac{1}{P(L)} \sum_{w_{1,n} \in L} \sum_{t=0}^{n} \alpha_i(t) a_{ij} b_j(w_k) \beta_j(t+1) \quad (2)$$

$$a_{ij} = \frac{\sum_{w_k} \gamma(i,j,w_k)}{\sum_{w_k} \sum_j \gamma(i,j,w_k)} \quad (3)$$

$$b_j(w_k) = \frac{\sum_i \gamma(i,j,w_k)}{\sum_{w_k} \sum_i \gamma(i,j,w_k)} \quad (4)$$

ここで、α, βはCutting[3]らと同様である。$\alpha_i(t)$の式は1文中のt番目の単語において'文頭'から品詞iまでの確率の総和で、$\beta_j(t+1)$は逆に'文末'から品詞jまでの確率の総和である。上式(2)(3)(4)はテキスト全文について計算し、再推定された確率値をコストに変換して図1の中間辞書に与える。これをJUMANの辞書に変換することで学習結果が反映される。

3 学習実験

3.1 方法

markovモデルとHMMの効果を測るために、EDRの7万5千文のタグつきコーパスからmarkovモデルにより初期値を獲得してHMM学習を行なう場合と、1万文だけで行なう場合の2つの実験を行なった。HMMの学習回数は朝日新聞の社説3年分（約8万文）について5回学習させた。

3.2 学習実験の結果

上記の学習の結果、もとのJUMAN以外に、markovモデルで学習したJUMAN、HMMで学習したJUMANの3つのモデルがある。これら各モデルに対する評価として、学習に使わなかった社説1カ月分を解析させて、解析結果の違う部分に対してその正解率で比較する。表1に正解率のパーセンテージを示す。正解率が足して100にならないのは両方とも不正解の場合が存在するためである。また、表1ではもとのJUMANを(or)、markovモデルを(mr)、HMMの場合を(hm)のように表記している。

表1 各モデルの社説解析結果の差（数値は%）

	or 対 mr	mr 対 hm	or 対 hm
EDR7.5万文	58 − 41	43 − 49	46 − 48
EDR1.0万文	60 − 33	36 − 51	47 − 48

まず、markovモデルについて比較してみると7.5万文で初期値を獲得した方が1万文の場合より良いという当然の結果が表れている。しかし、どちらの場合も、もとのJUMANの解析結果の方が優れていた。次に、HMMの学習結果をみると、7.5万文で初期値を得た方はmarkovモデルに対してあまり解析が向上していないが、1万文の方は大きく向上している。結局、もとのJUMANと解析結果を比較すると、どちらの場合も少し良くなる程度であった。実際、どちらのHMMの解析結果もほぼ同じであったことから、初期のコーパスの大きさによらず学習がほぼ収束したことがわかった。この結果から、タグつきコーパスが少ない分野においても、HMM学習により解析精度を向上させることができることがわかった。

4 まとめと今後の課題

以上、HMMを用いた形態素解析のパラメータ学習システムを提案した。しかしながらHMMの解析精度はまだ低い。これは学習時のコーパスの量を増やすことで精度の向上が期待できるが、コーパスの量を大きくすると何度も学習させるためかなりの時間がかかることになる。

今回は新聞記事に関して学習を行なったが、他の分野に対して実際この手法を用いて学習実験を行なう必要がある。

参考文献

[1] 松本裕治、他、"日本語形態素解析システムJUMAN使用説明書2.0"、奈良先端大技術報告書、*NAIST-IS-TR94025*,(1994).

[2] Nagata, M, "A Stochastic Japanese Morphological Analyzer Using a Forward-DP Backward -A* N-Best Search Algorithm", *Proc. Coling*, pp.201–207(1994).

[3] Cutting, D., Kupiec, J., Pedersen, J., and Sibun, P., "A Practical Part-of-Speech Tagger", *ANLP-92*, pp.133–143, 1992.