HMMを用いた形態素解析

村上仁一 嵯峨山茂樹
(ATR 自動翻訳電話研究所)

4F−1

1 はじめに

形態素解析は、従来から対話、翻訳、校正などの目的のために、自然言語処理研究の一つの分野として研究が続けられている。これらの方式の多くは、予め単語を構文的意味的なカテゴリに分類してカテゴリー間の接続ルールや係受けルールを記述しておく必要がある[1]。しかし、実際の日本語では単語の境界が明確でないことや単語の多品詞性や曖昧な係受けなどの問題があるため、精密なルールの作成は容易でない。

そこで、本論文では隠れマルコフモデル(HMM)を用いた日本語の形態素解析方法を提案する。HMMにはBaum-Welchの学習アルゴリズムが知られているためテキストデータからモデルのパラメータが学習できる。そのため、文法としてのルールも品詞ラベルが振られたテキストデータが与えられなくても形態素解析ができる可能性がある。

最後にこのモデルに基づいて実験を行なった。ここで用いたモデルは、かなり単純なモデルであるが、実験の結果は、単純なモデルとしては良好な解析結果を得た。

2 Ergodic HMMを用いた形態素解析

2.1 Ergodic HMMを用いた形態素解析の概念

形態素解析は、漢字かな文を単語に分けて品詞ラベルを付与することであるが、日本語における単語の境界の曖昧性や未知語の問題を避けるため、本稿では漢字かな文字単位に品詞ラベルを付与することを目的にした。そして、文法としてルールの代わりに統計的な情報を利用する形態素解析方法を考えた。

日本語では各々の品詞に依存して漢字仮名文字の出現頻度に偏りがある。例えば助詞は、「は」「が」などの仮名の出現頻度が高く漢字は出現しない。また名詞は、漢字の出現頻度が高く仮名の出現頻度は低い。また、品詞間の遷移確率にも偏りがある。例えば名詞の後に助詞が遷移しやすい。

このような性質に着目して、日本語を、品詞の初期確率 π_i と、品詞間の遷移確率 a_{ij} と、各品詞の漢字かな文字の出力確率 $b_j(o_t)$ のパラメータを持つ確率付きの有限状態オートマトンでモデル化する。このモデルを用いて、任意の漢字かな列に対して最も高い尤度で出力する品詞系列を計算することによって、漢字かな文字に対する品詞が特定できる。品詞ラベルが付与された大量のテキストデータが与えられれば、以上のパラメータ値は求めることができる。

品詞ラベルが付与されていないテキストデータのみが与えられた場合は、HMMを用いる。HMM[2]は、確率的性質を持つ信号源がMarkov的に切替えられて非定常信号源を表現しているモデルで、与えられた学習データの尤度を最大化するようにパラメータを再推定するBaum-Welchの学習アルゴリズムがある。このモデルにはいくつかの種類があり、音声認識の分野では、Left-right HMMが良く利用されているが、図1のような全状態が全状態に接続されているモデルを特にergodic HMMと呼んでいる。

このergodic HMMは構造的には確率付き有限オートマトンと同じ構造を持つため、日本語のテキストデータをBaum-Welchの学習アルゴリズムを用いて学習したならば、学習後のモデルは、状態は品詞に、状態遷移確率は品詞間の遷移確率に、シンボル出力確率は各品詞の漢字かな文字の出力確率に対応づけて考えることができる。

つまり、言語モデルとしてergodic HMMを用いることによって、大量のテキストデータがあれば、品詞ラベルも従来の形態素解析において必要とされていたルールも必要とせずに形態素解析ができる可能性がある。

2.2 Ergodic HMMを用いた形態素解析の手順

ここでは、Ergodic HMMを用いた形態素解析の手順を説明する。

1. 初期モデルのパラメータの計算。
   Baum-Welchの学習アルゴリズムは、学習データの尤度を最大にするようにパラメータを再学習するアルゴリズムであるため、最初に初期モデルとしてパラメータを設定しておく必要がある。そこで、予め各品詞の漢字かな文字の頻度を実験的に求めておき、HMMの状態と品詞の対応を決めて、初期モデルの各状態のシンボル出力確率を、対応する漢字かな文字の頻度に設定する。
2. Baum-Welchの学習
   次に大量のテキストを学習データとしてBaum-Welchの学習アルゴリズムを用いてパラメータを計算する。このとき、初期モデルにおける状態と品詞の対応は保存されることが期待される。この学習の結果、例えば図1のパラメータが得られたとする。
3. 形態素解析
   最後に、学習されたergodic HMMを用いて形態素解析をおこなう。図1のモデルで“春がきた”と考えよう。このモデルでは、“春”が状態1、“が”が状態2、“き”および“た”は状態3から出力したときに文の最大の生成尤度を得る。初期モデルにおいて状態1は名詞、状態2は助詞、状態3は動詞に対応していたとすると、“春”は名詞、“が”は助詞、“き”と“た”は動詞と形態素解析ができる。

任意の漢字かな文を入力して状態遷移系列を計算するアルゴリズムとしては、文の最大の生成尤度を出力する状態遷移系列を求めるViterbiアルゴリズムと各文字の最大確率の状態を選択して状態遷移系列を求めるForwardアルゴリズムがある[2]。

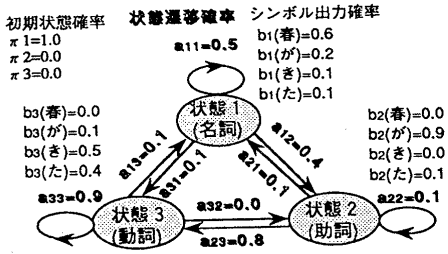

図1: Ergodic HMMを用いた形態素解析の例

3 Ergodic HMMを用いた形態素解析の実験

3.1 形態素解析の実験条件

ergodic HMMを用いた形態素解析の解析精度を知るために、パラメータや形態素解析の方法を変えて、以下の8個(2×2×2)の組み合わせについて実験を行なった。

"Hidden Markov Model applied to Morphological Analysis,"
Jin'ichi MURAKAMI and Shigeki SAGAYAMA
ATR Interpreting Telephony Research Laboratories, Kyoto, Japan.

(a) パラメータの計算方法
  i. 品詞ラベルが付与されているテキストデータから直接算出
  ii. Baum-Welch の学習アルゴリズムによる学習
(b) 形態素解析の計算方法
  i. Viterbi アルゴリズム
  ii. Forward アルゴリズム
(c) テストデータ
  i. パラメータの計算に使用したテキストデータ (closed データ)
  ii. パラメータの計算に使用しなかったテキストデータ (open データ)

また、Baum-Welch の学習アルゴリズムを用いたときの初期モデルの状態遷移確率および初期状態確率は共に均一の値 (1/114) とし、シンボル出力確率は 6000 単語の辞書データを利用して単語を品詞ごとに集め、各品詞ごとに漢字かな 1 文字の出力確率を計算して、この値を利用した。

その他の実験条件を表 1に、実験に用いたテキストデータの一部を表 2に示す。

表 1: 実験条件

HMM の状態数	114
HMM のシンボル数	約 3000 (漢字 JIS 1 級)
HMM の種類	全遷移型 状態出力タイプ
HMM の学習終了条件	16 回学習
テキストデータの種類	国際会議の申し込みの対話文
品詞数	114 種類 (活用形、活用型を含む)
学習データ	124175 文字 (品詞既知)
テストデータ closed	130 文 約 3500 文字
テストデータ open	130 文 約 1500 文字

表 2: テキストデータ (例)

文字	品詞	文字	品詞	文字	品詞	文字	品詞
は	感動詞	え	間投詞	ら	代名詞	訳	固有名詞
い	感動詞	ー	間投詞	第	接頭語	電	固有名詞
も	感動詞	っ	間投詞	1	数詞	話	固有名詞
し	感動詞	と	間投詞	回	接尾語	国	固有名詞
も	感動詞	そ	代名詞	の	格助詞	際	固有名詞
し	感動詞	ち	代名詞	通	固有名詞	会	固有名詞

3.2 形態素解析の実験結果

実験結果を表 3に示す。この結果から以下のことがわかった。

(a) 形態素解析の解析精度は、品詞既知のテキストデータからパラメータを計算して形態素解析を Forward アルゴリズムで計算した場合が最も高く、open データにおいて 70.2% を得た。
(b) Baum-Welch 学習によってパラメータを計算したときの形態素解析の解析精度は、品詞既知のテキストデータから計算した場合より低く 44.8% であった。
(c) いずれの実験でも Forward アルゴリズムを用いて得られた解析精度は、Viterbi アルゴリズムより高かった。

表 3: Ergodic HMM による形態素解析の実験結果 解析精度 (%)

学習方法	形態素解析	open データ	closed データ
直接算出	Viterbi	62.5%	64.2%
直接算出	Forward	70.2%	70.2%
Baum-Welch	Viterbi	36.0%	39.8%
Baum-Welch	Forward	44.8%	43.6%

表 4に、品詞ラベルが付与されているテキストデータからパラメータを直接計算して形態素解析を Forward アルゴリズムで計算したときの open データにおける形態素解析の誤出力の種類を、誤りの多い方から示した。この表から普通名詞をサ変名詞とする誤りが最も多いことがわかる。

表 4: Ergodic HMM を用いた形態素解析の誤出力 (品詞既知,Forward,open データ)

正解	出力	誤出力に対する出現率
普通名詞	サ変名詞	8.9%
助動詞・終止	格助詞	3.8%
助動詞・終止	助動詞・連用	3.0%
普通名詞	接尾語	2.8%
普通名詞	数詞	2.8%

3.3 2nd-order HMM を用いた形態素解析

ここでは、より高い解析精度をめざして 2 つ前の状態まで考慮する 2nd-order HMM[3] を用いて形態素解析の実験を試みた。ただし、Baum-Welch の学習に時間がかかるため、パラメータは品詞が付与されているテキストデータから直接計算した。この結果を表 5にしめす。表 3と比較すると、いずれの場合も 2nd-order HMM は単純な HMM より高い解析精度を得ている。また表 6に、形態素解析を Forward アルゴリズムで計算したときの open データにおける誤出力の種類を示した。表 4と比較すると、誤りの種類が変化していて、普通名詞を固有名詞と誤出力している場合が多い。

表 5: 2nd-order HMM による形態素解析の実験結果 解析精度 (%)

学習方法	形態素解析	open データ	closed データ
直接算出	Viterbi	84.9%	83.1%
直接算出	Forward	82.2%	83.3%

表 6: 2nd-order HMM を用いた形態素解析の誤出力 (品詞既知,Forward,open データ)

正解	出力	誤出力に対する出現率
普通名詞	固有名詞	12.8%
普通名詞	数詞	3.8%
普通名詞	接尾語	3.1%
本動詞連用 5 段	普通名詞	3.1%
接尾語	普通名詞	3.1%

4 まとめ

本論文では、漢字かな 1 文字に対応した確率付き有限状態オートマトンおよび ergodic HMM を用いた形態素解析の手法を提案し、その解析精度を調べた。この結果、状態遷移確率とシンボル出力確率を品詞が付与されているテキストデータから直接計算したときは 70.2% が得られたが、品詞未知のテキストデータから Baum-Welch の学習アルゴリズムを用いたときは 44.8% であった。また 2nd-order HMM を使用したときは 82.2% の高い解析精度が得られた。これらの実験結果から、ergodic HMM のような簡単な言語モデルを用いても、ある程度の解析精度が得られることがわかった。しかし Baum-Welch の学習アルゴリズムを用いたときの解析精度は、まだ、かなり低く、学習方法に改善の余地があると思われる。今後、このモデルと従来の形態素解析の方法を組み合わせることにより、未知語にも対処できるので、より高い解析精度が得られる可能性がある。

参考文献

[1] 長尾 真, " 日本語情報処理," 社団法人電子通信学会, pp.63-64 (1984).
[2] L.R.Rabiner, B.H.Juang, "An Introduction to Hidden Markov Models," IEEE ASSP MAGAZINE pp.4-16 (Jan. 1986).
[3] Yang He, "Extended Viterbi Algorithm for Second Order Hidden Markov Process," Proc. IEEE 9th Int. Conf. on Pattern Recognition, pp. 718-720, Rome, Italy (1988).