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

山本幹雄　増山正和

筑波大学

1．はじめに

英語の形態素解析においては、タグ付きコーパスを用いた確率的手法がほぼ確立している[Charniak 93]。英語の正書法は単語区切りを明示的に持つため、これらの手法は単語をベースとしている。このため、単語区切りを明示的に持たない日本語に応用する場合次のような問題がある。

1. 英語の場合、未知語があっても単語分割には影響を与えないが、日本語の場合は単語分割にまで影響を与えるので、精度への影響がより深刻である。
2. 日本語の場合は区切りの曖昧さがあるため、単語分割数が一定であることを前提とした英語のモデルをそのまま適用するには問題がある（この問題に関しては次章で詳しく述べる）。

本報告では、上記の1、2に関する解決法として文字をベースとした形態素解析手法を提案する。日本語の文字は一般に使われているもので約3,000種あり、また平均単語長も2文字程度であるため、1文字は単語に近い情報を持っていると思われる。文字をベースとして単語辞書を用いなければ、未知語の概念自体がなくなり、1の問題が解決される。また、文字の長さは1文において一定であるため、2の問題も解決される。

以下では、2章で英語の手法を日本語に適用した場合の問題点と提案手法、3章で評価実験、4章で関連研究について述べる。

2．拡張文字を用いた形態素解析

2.1 単語を基本とした確率モデルとその問題点

英語の確率モデルによる形態素解析では、品詞トライグラムと品詞で条件付けられた単語の出現確率を用いた以下のようなモデルが一般的である。

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

英語の形態素解析においては入力が単語列 $W=w_1w_2...w_{length(W)}$ であるため、この式の値を最大化するような $T=t_1t_2...t_{length(T)}$ を求める問題(tagging)となる。未知語が存在したとしても単語分割は所与であるため、未知語にはすべての可能なタグを可能性として考慮するだけでよい。しかし、日本語の場合は入力が文字列であり、単語列ではないので、単語分割Wも同時に変化させる必要がある。このため、未知語を考慮する場合は、タグの種類だけでなく、文中のあらゆる箇所で任意の長さの単語を仮定する必要がある。

また、分割を変化させるため、分割数が異なる解析候補の尤度を比較しなければならない。上記の単語をベースとした確率モデルでは品詞列のモデル化にマルコフ・モデルを使っているため、厳密には長さで条件付けられた確率を用いていることになる。英語の場合は長さは固定であるため、最大の値を求める場合この条件は問題とならない。しかし、日本語の場合は、異なる長さの品詞系列を比較しなければならないため、この条件が近似として入ってしまう。英語のtaggingモデルにはなかった近似が1段多く入ってしまうのである。

この近似によって、分割数が少ない形態素列（各形態素は長い）が優先されることが起こる。なぜならば、長い系列の方が可能な系列の数が多いので、1つの可能性あたりの平均的な確率が小さくなるためである。日本語の形態素解析においては、「最長一致」や「文節数最小」ヒューリスティックスの有効性が知られているが、これをうまく確率モデルに取り入れていると言えなくもない。しかし、問題は

このヒューリスティックスの強さを制御できない点である。ヒューリスティックスの強さは訓練データを反映しておらず、偶然によるものとなっている。

2.2 拡張文字連鎖モデル

本節では、単語をベースとした形態素解析における上記２つの問題点を解決するために拡張文字の連鎖確率を用いた手法を提案する。この手法は文字をベースに確率連鎖モデルを構成するため、入力文に対して系列の長さを一定に保ったまま可能な解析候補を比較できる。また、辞書を用いず文字ベースの連鎖確率だけを使うため、未知語の問題がある程度回避される。

我々のシステムでもマルコフ・モデルを用いるが、文字単位のモデルを使うという点で従来の方法と異なる。また、単なる文字ではなく、分割とタグの情報を含ませた拡張文字を使うことによって形態素解析を実現している。拡張文字の連鎖確率を用いた場合のp(W,T)は一般的に次のよう表現される。

$$p(W,T)=\prod_{i=1}^{length(W)+1} p\left(e_i|e_{i-1}e_{i-2}...e_{i-n+1}\right)$$

ここで、length(W)は入力文字列の長さ、nはn-gramのn、e_iは<W,T>の情報から決定される拡張文字である。どのような拡張文字を使うかによって、いくつかのバリエーションが考えられる。例えば、文字と直後の区切りと品詞(タグ)を組み合わせた拡張文字は次のようになる。

$$e_i=\left\langle c_i,d_i,t_{f(i)}\right\rangle$$

c_iは入力文字列の位置iにおける文字、d_iはc_iの直後における区切り情報、$t_{f(i)}$はc_iを含む単語のタグ情報である。タグ情報としては、品詞の他に読み情報や活用形などが考えられる。この拡張文字を使ったモデルを文字品詞モデルと呼ぶ。単語を区切るだけの場合はタグ情報を省略する。このモデルを文字境界モデルと呼ぶ。図１に訓練データが与えられた場合の拡張文字列例を示す。図の中で、区切り情報は、文字の直後に区切りがある場合は1、ない場合は0で表現されている。

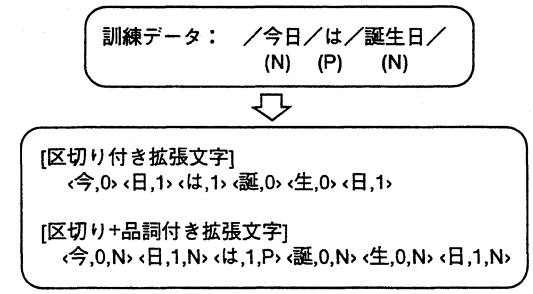

図１　訓練データに対する拡張文字列例

訓練用タグ付きコーパスから拡張文字列への変換は一意に可能であるため、これを用いて拡張文字列の連鎖確率(n-gram)を推定する。

形態素解析時には、与えられた入力文字列から可能な拡張文字の半順序構造を構成し、Viterbi アルゴリズムなどで最尤パスを決定すればよい。半順序構造の例を図２に示す。図から明らかなように、単語分割のみの場合は各文字２通りの可能性を考えるだけでよく、極めて高速な分割が可能である。

得られた最尤拡張文字列からタグ付き形態素列にもどすときは、区切りに関しては一意に定まるが、ある単語内での拡張文字の品詞が矛盾することもありえる。しかし、これは訓練データ中にそのような系列がありえないため、最尤となる可能性はきわめて低い。もし、未知文字などに対して矛盾する品詞が結果として出てしまった場合は、多数決や品詞の

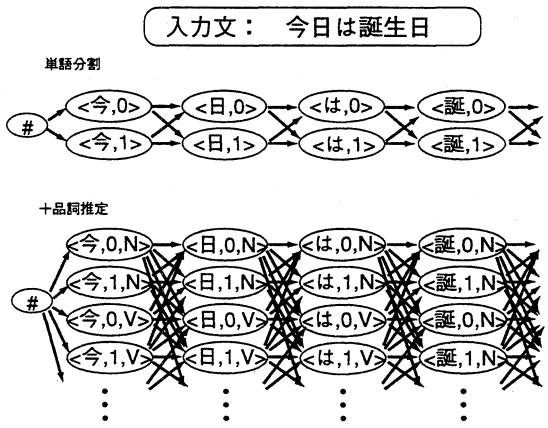

図２　拡張文字による形態素解析時の半順序構造

ユニグラムなどを用いて決定する。

3．実験

3.1 実験の概要

実験では提案手法として前節で述べた文字境界モデルと文字品詞モデルをトライグラムで実現したもの、従来法として2.1節で述べた(1)式を実現したものを用いた。コーパスは次節で述べる2種類を用いた。比較尺度として適合率（システムが出力した形態素のうち、正解の割合）を用い、単語分割性能と単語分割・品詞推定性能をクローズとオープンの実験で評価した。ただし、従来法に対して未知語処理の機構を実現しなかったため、オープン実験における未知語（訓練データから辞書を生成する）に対応できない。このため、従来法に対してはクローズ実験だけを行った。

3.2 コーパスとモデル推定

実験では、EDRコーパス[EDR95](以下、EDR)とATR対話データベース[江原他90]（旅行、電話対話）(以下、ADD)を用いた。EDRは、品詞15個、訓練データ197,744文、テストデータ1000文を使った。ADDは、品詞に活用形と活用型を加えた120種類のタグ、訓練データ8,821文、テストデータ500文を使った。それぞれ、テストデータを訓練にも使った実験をクローズ実験、使わなかった実験をオープン実験と呼ぶ。

また、モデルの確率モデルの推定にはCMU SLMツールキット[Rosenfeld95]を使ってback-offスムージング付きのトライグラムを作成した。

3.3 評価実験結果

実験結果を表1に示す。また、EDRコーパスを用いた場合の単語分割性能のグラフを図3、ADDコーパスを用いた場合の単語分割・品詞推定性能のグラフを図4に示す。

EDRにおける従来法が特に悪いが、これはEDRコーパスに15種類の品詞しかないため、品詞のトライグラムが十分な言語モデル能力を持たないためである。提案手法の方はEDRコーパスに関しては若干悪いが、比較的安定してよい結果を出している。これは、拡張文字による言語モデルの可能性を示している。なお、EDRコーパスの解析誤りに対する視察によれば、誤りの大部分はコーパス自体の揺れに起因しているようである。

品詞推定も同時に行った場合を分割だけの場合と比べると拡張文字を使う方法は1～2%程度、従来法では2～3%程度精度が低下した。オープン実験では、さらに3% 程度低下した。拡張文字のトライグラムはその中に辞書を持っており、同時に未知の単

表1　実験結果
(上段が分割適合率、下段が分割・品詞適合率)

手法	OPEN実験		CLOSE実験	
	EDR	ADD	EDR	ADD
文字境界モデル	95.63 -----	98.41 -----	97.80 -----	99.77 -----
文字品詞モデル	95.91 94.13	98.59 96.94	98.25 97.42	99.97 99.77
従来法(単語ベース)	-----	-----	95.65 92.55	99.52 97.82

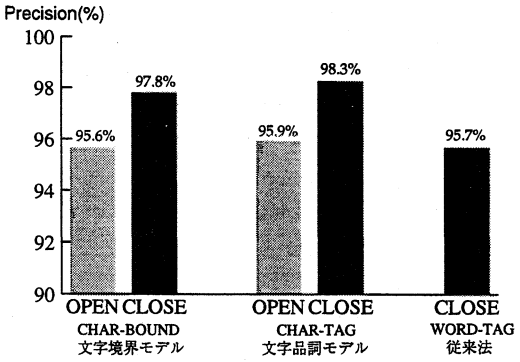

図3　EDRに対する分割性能

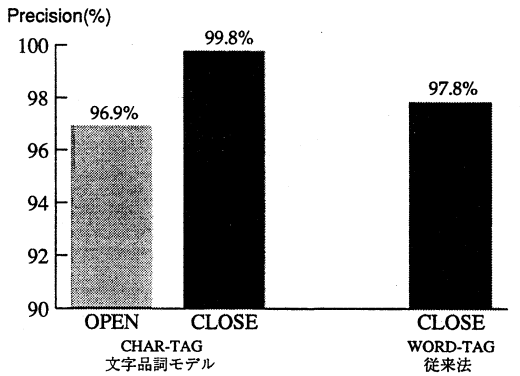

図4　ADDに対する分割・品詞推定性能

語のモデルとしても機能している。このため、訓練データにない未知語に対してもある程度の対応が可能であるが、さらにロバスト性を上げるためには文字のクラスタリングなどを行う必要がある。

拡張文字を使う方法は、単語分割では極めて高速であるが(SUN Ultra-Sparc140でEDRテスト用1000文約3秒)、品詞まで推定する場合は逆に極めて遅くなる(同条件で約670秒)。これは、トライグラムの場合で1文字当たりの可能な拡張文字数の3乗、バイグラムの場合で2乗のオーダで計算量に効いてくるためである。実験よりバイグラムでもそれほど性能が落ちないことが確認されているため(ADDオープン実験で分割・品詞適合率96.72%)、上記のクラスタリングの検討も含めて高速化を今後の課題としたい。

4．関連研究

文字の連鎖確率をベースとした言語モデルの研究は、文字認識(OCR)、音声認識における言語モデルにおいて見られる。しかし、これらの研究では文全体の文らしさをモデル化できればよく、単純な文字の連鎖確率を用いている研究がほとんどである。[山田他94]においては、音響モデルと効率良く組み合わせ、かつ音声認識結果をかな漢字混じり文に変換するために文字と読みをペアにした拡張文字のトライグラムが使われている。しかし、目的が異なるため、品詞情報や分割情報を文字に入れるというアイデアはない。

永田の単語モデル[Nagata94]は、日本語の未知単語を品詞ごとに区切り記号で始まる単語内の文字列によってモデル化している。単語境界を明示的に使用している点で似ているが、我々の手法は単語ではなく、単語列を直接モデル化している点で大きく異なっている。

Papageorgiouの手法は、単語境界あるいは非境界において前後の1文字づつ（計2文字）が出現する確率に、境界と非境界の時系列に対するバイグラムモデル（文字とは独立）を結合したものである。日本語の単語分割という目的に限って言えば、文字をベースとしたモデルによって明示的に単語境界をモデル化している点において、我々の方法と最も近い。我々の手法はPapageorgiouの手法をタグ付け可能なように拡張したと考えることもできる。しかし、単語分割という目的だけに限っても、Papageorgiouのモデルは拡張文字を用いる手法と比べて非常に荒い。文字と独立して単語の境界・非境界の連鎖をモデル化しているため、このモデルはすべての単語の長さに関する単一のモデルとなってしまっている。単語の長さは品詞や実際の単語（文字）に強く依存しているのであるが、これを無視しているため、システムの性能は単語の分割精度で約91%と、あまりよい結果は報告されていない。もちろん、実験に用いたコーパスが異なるため、正確な比較を行うことはできないが、3.3節で触れたバイグラムを用いた我々の実験結果よりもはるかに悪い。境界・非境界の連鎖モデルが文字と独立している限り、性能に限界があると思われる。

5．おわりに

提案手法は、単純ではあるが、かなりよい性能を持っていることが確認できた。品詞推定を行う場合の計算量が大きい点が問題であるので、これを軽減することが今後の課題である。

参考文献

[江原他90] 江原、井ノ上、幸山、長谷川、庄山、森元. 1990. ATR対話データベースの内容. ATR Technical Report, TR-I-0186.

[山田他94] 山田、松永、川端、鹿野. 1994. 音声認識における仮名・漢字文字連鎖確率に基づく統計的言語モデルの利用. 信学論(A)Vol.J77-A, No.2. pages 198-205.

[Charniak93] Eugene Charniak. 1993. Statistical Language Learning. MIT Press.

[EDR95] Japan Electronic Dictionary Research Institute. 1995. EDR Electronic Dictionary Version 2 Technical Guide. http://www.ijnet.or.jp/edr.

[Nagata94] Masaaki Nagata. 1994. A stochastic Japanese morphological analyzer using a forward-DP backward-A* N-best search algorithm. In Proceedings of COLING-94, pages 201-207.

[Papageorgiou94] Constantine P. Papageorgiou. 1994. Japanese word segmentation by hidden Markov model. In Proceedings of the Human Language Technology Workshop. pages 283-288.

[Rosenfeld95] R. Rosenfeld. 1995. The CMU statistical language modeling toolkit and its use in the 1994 ARPA CSR evaluation. In the Proceedings of ARPA Spoken Language Systems Technology Workshop. pages 47-50.