

条件付き確率最大法を利用した日本語形態素解析

1 B-1

久光 徹 新田 義彦

日立製作所 基礎研究所

1. はじめに

べた書き日本語文の形態素解析においては、一般にきわめて多数の解が生じるため、それらの中から適切な解を選択する必要がある。そのためにさまざまな解の尤度付け手法が提案されているが、単語（またはカテゴリー）接続に関するマルコフモデルを用いて解の尤度付けを行なう手法は、代表的なものの一つである[1][4]。本報告では、マルコフモデルを用いる手法において、与えられたマルコフモデルをより有効に利用し、モデルを変形することなく尤度付け精度を向上させる方法として、入力文字列に基づく各解の条件付き確率を利用する方法を提案する。以下では、マルコフモデル、条件付き確率の利用法、およびアルゴリズムについて簡単に述べた後、本手法により、最小コスト解の個数を（正解を含んだままで）低減できることを実験的に示す。更に、本手法を拡張して適用することにより、ある種の非確率的な尤度関数（例えば形態素数最小法、コスト最小法[5]）についても同様の効果が見られることを示す。

2. マルコフモデル

以下では次の形のマルコフモデルを考える：

$$\mathcal{M} = \{ \{c_j\}, \{w\}, \{p_{ij}\}, \{q_j(w)\} \};$$

$\{c_j\}$: カテゴリーの集合;

$\{w\}$: 語の集合;

p_{ij} : カテゴリー c_i から c_j への遷移確率;

$q_j(w)$: カテゴリー c_j からの w の生成確率;

カテゴリー:

品詞分類に対応するもので、文法コードの集合。ただし、単語を生成しない次の特別な要素を含む:

START: 文頭; COMMA: 句点 ", ";

PAREN-L: 左括弧 " ["; PAREN-R: 右括弧 "] ";

END: 読点 " . "。

3. 条件付き確率の利用

入力文字列を $s = \sigma_1 \dots \sigma_n$ とし、 $j-1$ 文字目までの部分文字列 $\sigma_1 \dots \sigma_{j-1}$ の解の一つを固定し、その末尾の語を w 、そのカテゴリーを $c(w)$ とする。 j 文字目からの単語候補を

$$\{v_{j1}^1, \dots, v_{jm1}^1, v_{j1}^2, \dots, v_{jm2}^2, \dots, v_{j1}^k, \dots, v_{jm_k}^k\}$$

とし、 $\{v_{ji}^1, \dots, v_{mi}^1\}$ の属するカテゴリーを c_i とすると（図1参照）、従来の手法では単語連鎖 $w \cdot v_{ji}^1$ に対して、マルコフモデルに基づく確率

$$p_{c(w)j} \times q_i(v_{ji}^1) \quad (1)$$

を与え、解に現れる各単語連鎖についてこれを掛け合わせたものをその解の生成確率とし、これを最大とする解を選択する。この手法を、ここでは確率最大法と呼ぶことにする。

ここで、同一の単語連鎖 $w \cdot v_{ji}^1$ であっても、 v_{ji}^1 以外の候補単語は入力文字列や解析位置により一般には異なるが、確率最大法の欠点は、上記の出現環境の違いを区別できないことである。そこで、 v_{ji}^1 の出現環境を尤度付けに反映するため、条件付き確率

$$\frac{p_{c(w)j} \times q_i(v_{ji}^1)}{\sum_{s=1}^k p_{c(w)s} \times \sum_{t=1}^{m_i} q_i(v_{jt}^1)} \quad (2)$$

を計算し、解に現れる各単語連鎖についてこれを掛け合わせたものを用いて尤度付けする手法が考えられる。この確率が最大となる解を選ぶ手法を、条件付き確率最大法と呼ぶことにする。

確率最大法の欠点はすでに述べたが、条件付き確率最大法も、単独で用いた場合、漢字熟語の不必要な細分化が発生する等の欠点がある。しかし、条件付き確率最大法を通常の最大確率法に融合させて利用する場合、両者の相補作用により尤度付け精度が向上すると期待される。

具体的には、各単語連鎖に対する接続コストとして、(1)、(2)の対数を取り符号を変えたコスト(1)', (2)'を付与する：

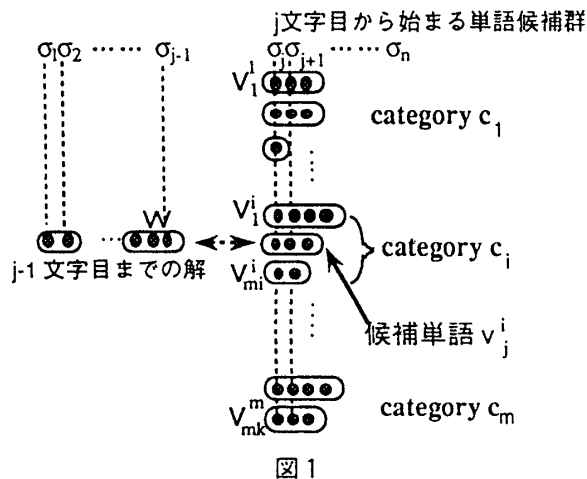
$$- \log_2 \{ p_{c(w)j} \times q_i(v_{ji}^1) \} \quad (1)'$$

$$-\log_2 \left\{ \frac{p_{c(w)i}}{\sum_{s=1}^k p_{c(w)s}} \times \frac{q_i(v_j^i)}{\sum_{t=1}^{m_i} q_i(v_t^i)} \right\} \quad (2)'$$

接続コストとして(1)'を用いるコスト関数を f_p , (2)'を用いるコスト関数を f_{CP} とし, これらを融合するため正数 α, β を用いて $f_{\alpha\beta}$ を次のように定義する:

$$f_{\alpha\beta} = \alpha f_p + \beta f_{CP}$$

このような $f_{\alpha\beta}$ を用いることにより, 条件付き確率最大法を取り入れることができる (α, β の値については後述する)。



4. アルゴリズム

アルゴリズムは, [3]に示したコスト N 位までの解をlattice状に保持するアルゴリズムの尤度計算部を多少変形して得られる. 解析表作成の計算オーダは, [3]とほぼ同様の議論により, 入力文字列の長さを n として, $O(n \log N)$ であることが示せる.

5. 実験結果

実験では, カテゴリー接続のみを利用する2種類の簡単なマルコフモデルを用いた:

- (A) 文法的なカテゴリーとして自立語, 付属語のみを考慮するモデル.
- (B) 21個からなるカテゴリーを持つ, (A)より詳細なモデル (詳細は省略する)。

また, 非確率的尤度関数でも, 接続コストの値が正値のみからなる場合, 解集合上にこの尤度関数と同一の順序付けを与えるマルコフモデルを構成できることが示せる. 一旦対応するマルコフモデルを構成すれば, 非確率的手法に対しても3で述べた手法が適用できる. 我々は, 2種類の非確率的手法に対応して, 次のモデル(C), (D)を構成した:

(C)形態素数最小法に対応するマルコフモデル.

(すなわち, 任意の単語連鎖を等確率とみなす)

(D) コスト最小法[5]における, 次の尤度関数

自立語・自立語 $\rightarrow 4$, 自立語・付属語 $\rightarrow 1$,

付属語・自立語 $\rightarrow 4$, 付属語・付属語 $\rightarrow 1$

と同一の順序を解集合に与えるマルコフモデル.

表1は, 上述の四つのモデルを用いて, [2]から抽出した100文に対して形態素解析を行った結果を, 確率最大法のみの場合 (すなわち $\beta=0$) と, 条件付き確率最大法を利用した場合 ($\beta>0$) で比較した結果である. 比較には次の二つの指標を用いた:

指標1: 100文の最小コスト解に含まれる正解数.

指標2: 100文の最小コスト解数の合計.

指標1は大きいほど, 指標2は小さいほど好ましい. 以下, 例えば表中のモデルAは, $\alpha=1, \beta=0$ とした場合であり, A'は, $\alpha=1, \beta=0.1$ とした場合である. 他の場合も同様.

モデル	A	A'	B	B'	C	C'	D	D'
指標1	69	69	90	93	64	64	66	66
指標2	210	170	113	113	1806	661	312	281

表1

表1から次のことがわかる:

- 1) 条件付き確率最大法を取り入れた場合, 最小コスト解の個数は減少する.
- 2) 条件付き確率最大法を取り入れても, 最小コスト解中の正解数は減らない.

実験により, 条件付き確率最大法を従来方式と融合した形で利用することにより, 最小コスト解の個数が低減するという意味で, 尤度付け精度が向上することが確認された.

6. おわりに

確率モデルを用いる形態素解析における条件付き確率最大法を提案し, その効果を示した. 他の指標を用いた比較, 平仮名列分割に適用した場合の効果の確認等が今後の課題である.

参考文献

- [1]下村 他: 最小コストパス探索モデルの形態素解析に基づく日本語文誤り検出の方式, 情報処理学会論文誌, Vol.33, No.4, pp.457-464 (1992)
- [2]近角 他: 精選理科I (生命・科学編, 及び力学・生命科学編), 東京書籍 (1981)
- [3]久光 他: 接続コスト最小法による日本語形態素解析の提案と計算量の評価について, 電子情報通信学会研究会資料NLC90-8 (1990)
- [4]松延 他: 確率文節文法による構文解析, 情報処理学会研究会資料, 56-3 (1986)
- [5]吉村 他: 未登録語を含む日本語文の形態素解析, 情報処理学会論文誌, Vol.30, No.3, pp.294-300 (1983)