

コスト最小法に基づく逐次確定型・形態素解析

6M-2

小松順子

(株) リコー 情報通信研究所

1 はじめに

コスト最小法に基づく形態素解析には、複数の解をもっともらしさの指標を付けて出力できる、形態素解析に有用なヒューリスティックや未登録語処理などを反映してコスト値をうまく設定してやれば、解析系を統一的な枠組みで扱えるという長所がある。しかし、多くの単語候補や単語候補間の接続情報を保持していなければならないので、他の方法に比べて計算量およびメモリ消費量が多い。

これに対して効率の良いアルゴリズム [1] が提案されているが、解析範囲を短く保つことができれば、さらに計算量とメモリを抑えられるし、リアルタイム性を要求されるアプリケーションには有利である。

そこで、単語候補間の文法的な接続チェックを行うと同時に、単語候補間の接続状態が文節の切れ目になるかどうかのチェックを行うことによって、逐次確定しながら解析する方法をとった。

2 基本アルゴリズム

ここでは、単語列 $W = (w_1, w_2, w_3, \dots, w_{n-1})$ のコスト $C_{path}(W)$ を次式のように、 W を構成する個々の単語のコスト $C_{word}(w_i)$ と隣接する単語間の接続のコスト $C_{cnct}(w_{i-1}, w_i)$ の総和で表す。

$$C_{path}(W) = \sum_{i=1}^{n-1} C_{word}(w_i) + \sum_{i=2}^{n-1} C_{cnct}(w_{i-1}, w_i)$$

すると、単語列 W の最後に単語 $w_n (n > 1)$ が接続してできた単語列 $W' = \langle W, w_n \rangle$ のコストは次のように帰納的に書ける。

$$C_{path}(W') = C_{path}(W) + C_{word}(w_n) + C_{cnct}(w_{n-1}, w_n)$$

A Morphological Analysis with Automatic Segmentation, Using Minimal Cost Method
Junko Komatsu
Information and Communication R&D Center
RICOH Co., Ltd.

従って、ある単語列のコストは過去の履歴を利用して求めることができる。そこで、文字列の先頭から解析を始めて、文字位置 i における部分解のコスト、末尾の単語候補、直前の部分解へのポインタを解析表に記録しながら解析を進める方法をとった。但し、実際には上位 N 位までに入る解が求まれば十分なので、文字位置 i までの部分解のコスト値が上位 N 位までに入るものだけを残す枝刈りを行った。

未登録語の場合を除いて、単語コストは頻度調査を基に求め、接続コストは経験的な値をトップダウンに与えた。[2]

3 未登録語処理

未登録語については、解析文字列の各文字 1 文字分 (句読点を除く) を表記とする未登録語候補を常に生成し、未登録語の単語コスト、および未登録語と通常単語 (辞書に登録されている単語) の接続コストを通常単語のものより大きめに設定することによって対処した。そして、最終的に得られた単語列の中に未登録語が連続する部分があった場合には、それらをまとめて 1 つの未登録語にする。

未登録語の単語コストには、通常単語より大きめの一定値を与えた。未登録語と他の単語候補との接続コストは、経験的に設定しておいた文字種の連鎖コストを基にして、次式のようにして求めた。 c_1, c_2 は単語 w_1, w_2 の表記のうち接続部分の文字の文字種を表す。 $C_{char}(c_1, c_2)$ は文字種の連鎖コスト、 C_{max} は文字種の連鎖コストの最大値。

$$C_{cnct}(w_1, w_2) = \begin{cases} \alpha \times C_{char}(c_1, c_2) & (w_1, w_2 \text{ 共に未登録語}) \\ \beta + \gamma \times (C_{max} - C_{char}(c_1, c_2)) & (w_1, w_2 \text{ の一方が未登録語}) \end{cases}$$

文字種の連鎖コストとは、単語表記における文字種の連鎖のしやすさを表す。例えば、カタカナ同士は連鎖しやすいがカタカナと平仮名は連鎖しにくいといった情報を反映したものである。

4 逐次確定する方法

4.1 逐次確定の条件

確定位置として妥当と思われるのは、全ての単語候補間の接続状態が文節の切れ目になる所である。そのためには、句読点で区切るのが簡単だが、句読点の有無は文章のスタイルに依存するので、解析範囲を短くできるとは限らない。そこで、次のような条件を満たす位置 $i-1 (i > 1)$ を確定位置とした。確定位置のチェックは、単語候補間の接続チェックと同時に行うことができる。

1. 文字位置 i をまたぐ単語候補が存在しない
2. 文字位置 $i-1$ で終わる単語候補と i で始まる単語候補との接続状態が全て文節の切れ目になる

文節の切れ目になるかどうかは、表1の条件によって決定した。未登録語は全て自立語であるという前提なので、未登録語同士の接続を除いては自立語と同じである。

表 1: 文節の切れ目の有無

前単語	後単語	文節の切れ目
自立語	自立語	有
自立語	未登録語	有
未登録語	自立語	有
自立語	付属語	無
未登録語	付属語	無
付属語	自立語	有
付属語	未登録語	有
付属語	付属語	無
未登録語	未登録語	無

未登録語同士の接続は単語列確定後、1単語にまとめられるので文字位置 i で未登録語同士の接続があるということは、 i をまたぐ単語候補が存在することに相当する。先に述べたように、未登録語候補は解析文字列の各文字について必ず1つ生成されるので、常に未登録語同士の接続が発生してしまい、永久に上記の条件が満たされなくなってしまうように思われるが、 $i-1$ で終わる単語候補は、上位 N 位までに入るパスに含まれるものに絞り込まれているので、実際には未登録語同士の接続はめったにおこらない。

4.2 確定後の処理

文字位置 $i-1 (i > 1)$ が確定位置になった場合は、 $i-1$ までの解を出力した後に、 $i-1$ までの履歴を破棄し、 i を解析文字列の先頭として解析を進める。その際、確定位置 $i-1$ までの部分解のうち、それに接続する単語候補が0個のものは、棄却されたことになるので解として出力しない。また、 i を解析文字列の先頭として解析を進める場合は、 $i-1$ までの部分解に接続しなかった単語候補は解析表に登録しない。

4.3 逐次確定の効果

解説文 61 文 (56.1 文字/文) を用いて、逐次確定の効果調べた結果を表2に示す。解析は上位5位までの候補を残しながら行った。正解率は、総単語数における正解単語 (1 位) の割合で、1 位の解が複数ある場合はその中に正解単語列が含まれていれば正解とした。

このテキストにおいては、解析精度をほとんど下げずに、平均解析文字数を約 1/3 にできた。

表 2: 逐次確定した場合としない場合との比較

	平均解析文字数	正解率 [%]
逐次確定した	9.0	96.8
句読点で区切った	25.2	96.9

5 おわりに

全ての単語候補間の接続状態が文節の切れ目になる所を確定位置とすることによって、句読点で区切る場合よりも解析範囲を短く保つことができた。これは処理量の削減に有効である。

今後は多くテキストを用いて、未登録語処理の評価と改良を行っていく予定である。

参考文献

- [1] 久光徹, 新田義彦. 接続コスト最小法による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 NIC 研資料, 90-8, 1990.
- [2] 佐藤奈穂子, 小松順子. コスト最小法を用いた形態素解析におけるコスト設定の一方. 情報処理学会第 47 回全国大会, 1993.