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法

6M－3

佐藤奈穂子　小松順子

（株）リコー　情報通信研究所

1　はじめに

コスト最小法を用いた形態素解析において、コストは候補の絞り込みに大きく影響する。

コスト設定の主な手法としては、単語あるいは単語間の接続関係を分類し、それに対してトップダウンに初期値を与えて修正を繰り返す方法 [1][2] と、確率モデルに基づく方法 [3][4] が提案されている。前者は、必要に応じて意図的にコスト値の操作ができ、メインテナンスしやすいという利点があるが、単語の分類が大まかで、単語固有の尤もらしさが反映されにくいという問題点がある。後者は、値が実際的ではあるが、厳密な生起確率を求めることが困難なため近似的であり、統計をとるサンプル文の量や質に影響されやすい。更に、部分的な値の操作がしにくいというメインテナンス上の問題もある。

そこで、本方法では単語列のコストを、構成単語のコストと隣接する単語間の接続のコストの総和で表す [5]。単語コストは個々の単語の出現頻度を基に統計的に求め、接続コストは接続のパタンそれぞれにトップダウンに経験値を与えた。

2　単語コストの設定

単語コストとは、その語の出現しやすさを表した値である。単語の出現しやすさは、その語の性格や入力テキストの分野に依存する。今回は、単語コストを新聞の語彙調査 [6] の結果得られた頻度に基づいて設定した。

単語コスト $C_{word}(w_i)$ は、対象語を w_i、w_iの品詞を $hinA$、w_iの補正頻度値を $n(w_i)$ として次式により求めた。補正頻度値とは頻度がつかなかったり、頻度が0の語に対処するために全ての頻度に1加算した値のことである。

$$C_{word}(w_i) = -\left(\alpha \times \log\left(\frac{n(w_i)}{\sum_{w_i \in hinA} n(w_i)}\right)\right)$$

この式によりコストを求めると、助詞・助動詞のように1品詞1単語の機能語のコストは0になる。

3　接続コストの設定

接続コストは、隣接する単語の接続のしやすさを表した値である。一般に品詞の接続関係は文法的に決まっており、その接続の強さは入力テキストの分野に左右されることは少ないと考えられる。そこで品詞の接続関係をいくつかのパタンに分類し、それぞれに対して接続の強さに応じた多段階の値を主観で与えた。

接続パタンは、まず接続関係を「接続する」「稀に接続する」「接続しない」に大別した。「稀に接続する」は接続しにくいが、しないわけではない関係である。更に、「接続する」ものを文法的な知見に基づき表1のように細分した。

語構成上の接続は、接続してまた一つの単語を構成するもので、接続の強さはかなり強いと考えられる。但し、一字名詞同士の複合語は熟語として辞書登録されている場合が多いので生成しにくくする必要があるため別分類とした。

文節内接続は、形態上自立語と付属語の接続と付属語連鎖に分けられる。前者は付属語が必ず自立語を伴って文節を生成することから接続は強いと考えられる。後者には助詞連鎖、助動詞連鎖、助詞と助動詞の接続などがある。

文節間接続には、連体修飾関係と連用修飾関係がある。前者は大抵直後を修飾するが、後者は常に隣接して現われるとは限らないため別分類とした。また、特殊パタンとは助動詞「だ」の連用形「で」＋動詞「ある」など慣用的な言い回しである。

接続コストとして概ね文節内接続には文節間接続より小さめの値を与えた。語構成上の接続・特殊パタンには、文節内接続とほぼ同等の値を与えた。

A Method of Cost Assignment for Morphological Analysis Based on Minimal Cost Method

Nahoko Sato　Junko Komatsu

Information and Communication R&D Center

RICOH Co., Ltd.

表 1: 接続パタン

語構成上の接続	接頭辞と自立語
	自立語と接尾辞
	複合名詞
	1字名詞同士の複合
	複合用言
文節内接続	自立語と助詞
	体言／副詞と助動詞
	用言と助動詞
	付属語連鎖
	単語と句読点類
文節間接続	連体修飾
	連用修飾
	特殊パタン
	句読点類と単語
	その他

但し一字名詞同士の複合は大きめの値を与えた。また、「稀に接続する」ものには「接続する」ものより大きめの値を与え、「接続しない」ものには無限大相当の値を与えた。

4 解析精度の評価

本方法で設定したコストを用いて評価実験を行なった。評価用テキストにはA新聞（社会分野）、B講演集（国際政治分野）、C教科書（数学分野）の異分野3種（合計173文、5,396語、8,909文字）を用いた。未登録語の割合は平均4.1%であった。解析用辞書は約58,000語である。

正解率と平均最小コスト解数を表2に示す。正解率は、テキスト中の総単語数のうち表記、品詞、読みが合っている単語の数の割合を言う。1位の解が複数出力された場合は、その中に正解単語列が含まれていれば正解とした。未登録語に関しては、正しく範囲が区切られていれば正解とした。平均最小コスト解数は一つの解析範囲あたり最小コストの単語列がいくつ出現するかの平均であり、解の絞り込みの度合いを示す。

表2によると、テキストAの正解率が最も高く、テキストCの正解率が低い。これは新聞の語彙調査で得られた単語コストとの分野の適合性がAでは高く、Bでは低いことが原因である。例えば、「角」の読みは、数学分野では「かく」を優先したいが、新聞の語彙調査による頻度は、「かど」が最も高い。

表 2: 実験結果

テキスト	A	B	C	平均
正解率（1位）[%]	98.5	97.2	95.4	97.0
平均最小コスト解数	1.29	1.37	1.25	1.30

また、誤解析の内訳の64%は未登録語によるものだった。「中（名詞）-東（名詞）」のように本来一単語である未登録語が既存の単語の連鎖として解析されていた例があった。今後は未登録語との兼ね合いを含めた接続コストの最適化が必要である。

最小コスト解数の平均は1.30であった。格助詞「で」と断定の助動詞「だ」の連用形「で」など頻繁に出現するが文脈によらねば判別できない例や、「日本」など二種類の読みがある例がほとんどだった。

5 おわりに

今回は単語の出現頻度に基づき単語コストを設定し、品詞間の接続の強さという観点から接続のパタンを分類して、それに対して主観的に値を与えて接続コストを設定したが、まずまずの結果が得られた。

今後は用例を使い、主観による接続パタンの分類やコスト値の妥当性を検証・修正する予定である。更にコスト値の最適化の方法も検討していきたい。また、他のコスト設定法との比較も行なっていきたい。

参考文献

[1] 吉村他. コスト最小法を用いた日本語文の形態素解析. 情報処理学会 *NL* 研資料, 60-1,1987.

[2] 小松他. コスト最小法形態素解析のコストルールの作成方法. 情報処理学会 *NL* 研資料, 85-1,1991.

[3] 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文誤り検出の一方式. 情報処理論文誌, 33(4),1992.

[4] 江原. 漢字仮名混じり文の形態素解析におけるペナルティー値付与の一方法. 情報処理学会第 *35* 回全国大会, 1987.

[5] 小松. コスト最小法に基づく逐次確定型・形態素解析. 情報処理学会第 *47* 回全国大会, 1993.

[6] 電子協. 日本語処理技術に関する調査研究. 1976.