自 然 言 語 処 理 85－1
（1 9 9 1. 9. 1 9）

コスト最小法形態素解析のコストルールの作成方法

小松 英二　　安原 宏
日本電子化辞書研究所　沖電気工業株式会社

コスト最小法形態素解析のコストルールの作成方法について述べる。コストルールの条件部は、既存の経験的規則や解析の失敗例を基に作成する。コストルールが与えるコストの値は、形態素解析の実行結果の誤り部分と正解とのペアを制約条件として決定する。本稿の方法を実際の形態素解析プログラムに適用した例も示す。

A method to construct cost evaluation rules for morphological analyzers

Eiji Komatsu
Japan Electronic Dictionary Research Institute,LTD.

Hiroshi Yasuhara
Oki Electric Industry,Co.,LTD.

C/O system laboratory,Oki electric industry co.,LTD.
11-22,Shibaura 4-Chome,Minato-ku,Tokyo,Japan
komatsu@edr6r.edr.co.jp

11-22,Shibaura 4-Chome,Minato-ku,Tokyo,Japan
Yasuhara@okilab.oki.co.jp

In this paper ,we propose a method to construct cost evaluation rules for Japanese morphological analyzers . As for condition parts of rules,our method makes them from existing inductive facts or examples of failure of a morphological analyzer. As for values of cost that each rule gives, our method decides values by using pairs of an error part of an result by a morphological analyzer and a right solution, as constraints.We also show examples that we applyed this method to our own morphlogical analyzer.

1．はじめに

日本語形態素解析の曖昧性の解消方法には、最長一致法、二文節最長一致法、ヒューリスティック的方法、文節数最小法、コスト最小法等がある[1]。これらの方法は、処理上の違いはあるが、論理的にはすべて最後のコスト最小法の特別な場合である。

コスト最小法とは、全ての解を候補として生成し、単語及び単語の接続等に対して一定のコストを与え、コストの和が最も小さくなる単語列を解として選ぶ方法である。コスト最小法では、全解が必要なため、全解を効率よく生成するグラフスタック等の圧縮表現が研究されている[2,3,4,5]。また、コストの与え方[6,7]、形態素解析の曖昧性調査結果[8,9]についても多くの研究がある。

筆者らは、コスト最小法形態素解析プログラムを作成したが、コストを与える規則（以下、コストルールと呼ぶ）の作成にあたって、次のような2つの問題が生じた。

(1) 単語のどのような属性に着目してコストを与えるか
(2) 各コストルールの与えるコストの値をいくつにするか

(1)については、以下で述べるようにコストの値の決定方法を工夫することにより、逐次必要に応じて新しい属性を追加することで解決した。

(2)については、プログラム作成上必須であり、なんらかの定式的な方法を考えなければならなかった。各ルールの与えるコストの値の決定は、大量のデータによって統計的に決定することもできるが、筆者らのプログラムでは、単語同士の接続にもコストを与えるため、属性の組み合わせのデータも必要となり、すべての組み合わせの頻度を調べることは非常に労力が大きく、得られた頻度の意味付けもむずかしいことが予想される。さらに形態素解析プログラムのインプリメント上も記憶容量の点から効率的でない。

このような理由から、筆者らの作成した形態素解析プログラムでは、個々の実例を制約条件としてコストの値を決定する方法を検討した。

本論文は、沖電気工業(株)在職中の成果に基づいて書かれたものである。

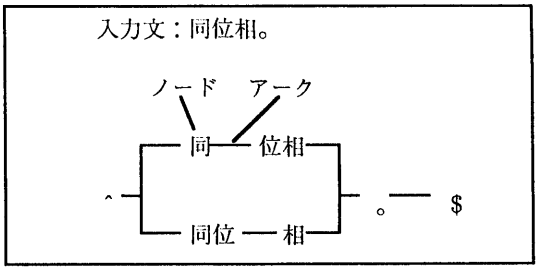

図1　グラフスタックを用いた形態素解析結果の例

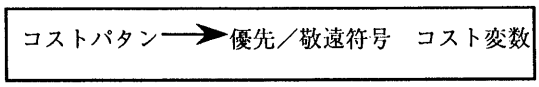

図2　コストルールの論理的なフォーマット

2．コスト最小法形態素解析プログラム

コストルール作成の対象として想定しているコスト最小法形態素解析プログラムについて述べる。

全解の表現形式としては、グラフスタック[1]を用いた。グラフスタックは、複数解の共通部分を共有することにより記憶の節約を行うデータ構造である。図1にグラフスタックを用いた形態素解析結果の例を示す。以下、グラフスタックの単語をノード、単語同士の接続をアークと呼ぶことにする。「^」と「$」は、それぞれ文頭と文末を表すために便宜上つけ加えられるノードである。

コストは、ノード及びアークに対して与え、最もコストの和が小さくなる経路を選び、その経路の単語列を形態素解析結果として出力する。

辞書は、沖電気の機械翻訳用辞書に3文字から6文字のひらがな表記を加えたものを用いている。品詞数は約40である。

コストルールは、手続きとして実現したが、論理的には図2のようなプロダクションルールと等価である。以下の説明では、この論理的な形式のルールを扱う。ルールの左辺は、見出し、品詞、字種、文字数等必要に応じてあらゆる情報を用いて記述した条件部であり、ルールの右辺は、各ルールの与えるコストに対応する変数である。左辺をコストパタン、右辺をコスト変数と呼ぶことにする。コスト変数は、3節で述べる方法により、形態素解析実行時には確定した値に置き換えて用

いる。コスト変数は、ルール数を減らす目的のため、0を基準としており、優先したいパタンにはコスト変数にマイナスの符号を与え、敬遠したいパタンにはプラスの符号を与えている。この符号を、*優先／敬遠符号*と呼ぶことにする。

コストルールは、ノード用ルールとアーク用ルールを作成した。図3にコストルールの例を示す。例えば、同じ見出しの語について固有名詞より普通名詞を積極的に用いたい場合には、ノード用ルールとして、品詞を用いて、図3のルール1のようなルールを作る。また、一文字漢字列へ分割する防ぎたい場合には、アーク用ルールとして、文字数と字種を用いて、図3のルール2のようなルールを作る。

ルール1：普通名詞を優先するルール
普通名詞 ⟶ －C1

ルール2：1文字漢字への分割を敬遠するルール
1文字漢字－1文字漢字 ⟶ ＋C2

図3　コストルールの例

誤り：同位(普通名詞)/相(普通名詞)
正解：同(接頭語)/位相(普通名詞)

誤り：私(普通名詞)/生(普通名詞)/活(普通名詞)
正解：私生活(普通名詞)

誤り：日本(固有名詞)/人(普通名詞)
正解：日本人(普通名詞)

図4　制約データの例

誤り：同位(普通名詞)/相(普通名詞)
正解：同(接頭語)/位相(普通名詞)

↓ 誤りと正解にコストルールを適用し、コスト変数の和で置き換える。

C1＋2C3－C6＋C7　＞C4＋C3－C8

↓ 変数を整理する。

C1＋C3＋C7＋C8　＞　C4＋C6

図5　制約データから制約不等式への変換の例

3．コストルールの作成方法

2節のような形態素解析プログラムのコストルールの作成方法について述べる。上記のコストルールの作成には、次の2段階の作業が必要である。

(1) コストパタン、優先／敬遠符号の決定
(2) コスト変数の値の決定

以下、それぞれの作業について述べる。

（1）コストパタン・優先／敬遠符号の選択

まず、従来の調査により得られている規則・内観に基づく規則を集め、コストパタンを作成する。また、解析結果の誤りと正解を弁別するのに必要なルールを逐次加えていく。コストパタンの選択は、作成者の主観・知識に依存する面が大きいため、コストパタンの妥当性の裏付けは別途行っておく必要が有る。ただし、不適切なパタンを登録しても、(2)で述べるコスト変数決定過程において、0に近い値が与えられるため、不適切なパタンは自然に排除できる。

（2）コスト変数の値の決定

(1)で作成したコストパタンに対応するコスト変数の値を決定する方法について述べる。

形態素解析の結果に誤りがある場合に、誤った解析結果を取り出し、正しい解析結果を加えたペアを作成する。これを、*制約データ*と呼ぶことにする。図4に制約データの例を示す。

制約データは、さらに、コストルール作成用のプログラムによりコスト評価を行い、誤りと正解の両方をコスト変数の和に置き換える。正解のコストは誤りのコストより小さくなければならないから、前者が後者より小さくなるように不等号で結び、コスト変数の不等式に変換する。この変換された不等式を*制約不等式*と呼ぶことにする。図5に制約データから制約不等式への変換の例を示す。

また、制約データを蓄積したものを*制約データベース*と呼ぶことにする。

コスト変数の決定は、まず、コスト変数に適当な初期値を設定し、形態素解析を行い、制約データを収集し、制約データベースに追加する。制約データベースの制約データをすべて制約不等式に変換し、これらを満たすコスト変数の値を決定し、得られたコスト変数を用いて、形態素解析を

行う。この作業を繰り返し、制約データを収集しながら、コスト変数値を修正していく。もし、制約データベースに対してコスト変数の値が存在しなくなった場合には、制約データを削除するか、(1)に戻って、コストルールを変更（追加・削除・修正）する。コストルールの変更があった場合には、制約不等式を新たに生成する必要がある。図６に上記の処理の概略フローを示す。また、図７にコスト変数の決定の各処理の内容を示す。

４．コストルールの作成例

３節の方法で、コスト変数の値を決定するようすを実例で示す。図８に初期状態を示す。コストルールは、文節数最小、単語数最小の規則に、字種による未知語の規則を加えた簡単な規則である。この時点でできるだけ多くのルールを用意しておいた方が望ましい。例文１として、図９のような文を選ぶ。コスト変数の初期値として全て１を設定する。

プログラムのコスト変数を修正し、１回目の解析を行う。図１０に例文１－１回目の実行結果を示す。図１１のような過程を経て、コスト変数の値を修正する。

プログラムのコスト変数の値を新しい値に変更し、形態素解析を行い２回目の解析を行う。図１２に例文１－２回目の実行結果を示す。図１３のように、新たに得られた制約データを制約データベースに加え、コスト変数の値を決定する。コスト変数の値を変更して形態素解析を実行すると、図１４のように正しい結果が得られる。

次に例文２として図１５のような文を選ぶ。図１６は例文２－１回目の実行結果である。今度は、図１７に示すように、新たに得られた制約不等式が解けないため、図１８のようなルールを加えると、図１９のようにコスト変数の値が決定でき、図２０のように形態素解析の結果も正しくなる。この場合、例文１の結果が正しいことは保障するためには再度チェックもする必要がある。図２１にコストルールの最終状態を示す。また、図２２に制約データベースの最終状態を示す。

上記の例では、優先／敬遠符号は＋しか現われていないが「－」の場合もある。また、コストルールの修正や制約データの削除などが必要になる場合もある。制約データ間の矛盾は制約不等式においてチェックする。

実際のコスト作成に当たっては、いくつかの例文をまとめて選んで制約データを作成した。また、上記の例では、誤りを直すために主観的にもっともらしいと思われるルールを逐次作成したが、実際のルール作成時には、最初に用意するルールと同様に十分に調査をしたうえで、確実度の高いルールを追加していくことがコスト変数を早く収束させるために必要である。

５．おわりに

コスト最小法形態素解析のコストルールの使われ方はルール同士の兼合により結果的に非決定的にな利、誤りに対応してプログラムを修正することとは個となる。また、逆にルールの妥当性をコストの値として得ることができるため、統計的データ収集の前段階として、必要な属性の洗いだしや基礎的なルールの抽出に有効であると考えている。

本稿の方法により、約６０ルールが得られ、新聞記事で９５％以上の解析精度を得ている。

本稿の方法の応用として、構文解析のコストルールの作成、あるいは、文書要約における重要文の抽出のルール作成などを考えている。

［参考文献］

[1]：長尾真監修：日本語情報処理、(社)電子通信学会、１９８４

[2]杉村：グラフスタックを用いた日本語形態素解析、情報処理学会第３７回全国大会、１９８８

[3]杉村他：論理型形態素解析ＬＡＸ、Proceedings of the logic programming conference '88、１９８８

[4]久保他：ＬＴＢ形態素解析ＬＡＸの開発環境、情報処理学会第３７回全国大会、１９８８

[5]吉村他：未知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、１９８９

[6]久光他：接続コスト最小法による日本語形態素解析、情報処理学会第４２回全国大会、１９９１

[7]N.Maruyama,etc：A Japanese sentence analyzer、IBM J. RES. DEVELOP. Vol.32 No.2、１９８９

[8]芦沢他：日英機械翻訳用前編集支援システム(2)形態素の曖昧性の検出方式、情報処理学会第３６回全国大会、１９８９

[9]坂本：文節の認定、日本語情報処理シンポジウム報告書、１９７８

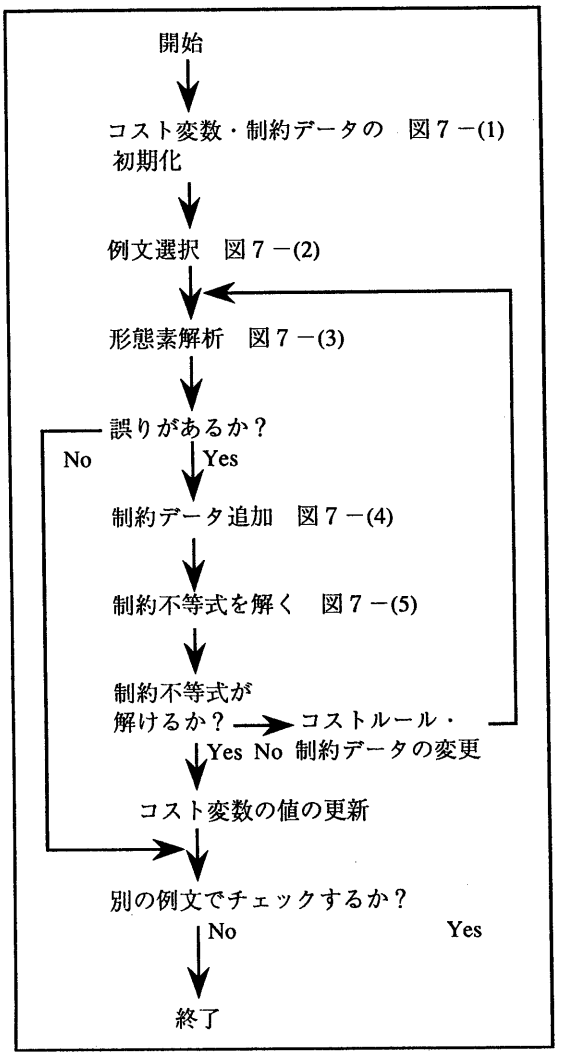

図６　コスト変数の決定の処理フロー

1)コストルールのコスト変数に初期値を設定する。初期値は、既にコスト変数の値が決定されている場合はその値を用い、新たに決定する場合には適当な正値を初期値として設定する。

制約データベースとしては、それまでにコスト変数の値を決定するために用いた全ての制約データを用いる。

(2)例文は１文ずつ選んでも、まとめて数文選んでもよい。

(3)既に得られているコスト変数の値を用いて、形態素解析を行う。

(4)形態素解析の結果に誤りがある場合は、誤りの前後を含む部分を取り出し、制約データベースに追加する。

(5)制約データベースの制約データをすべて制約不等式に変換し、連立不等式を満たす値の組を１つ決定する。制約不等式を満たす解がない場合は、互いに矛盾する制約データを発見し、制約データの変更を行う。変更の方法は次の３つが考えられる。

コストルールの追加：
誤りと正解を弁別する要因に関するコストルールがないと判断される場合に、新たな要因に関するコストルールを追加する。

コストルールの削除：
確信度の低いルールは、コスト変数の収束を遅らすため、削除する。

コストルールの修正：
コストルールのパタンが例外を含んだり、優先／敬遠の指定が間違っていると判断される場合に、ルールのパタンが指定する条件を厳しくしたり、優先／敬遠の指定を変更したりする。

制約データの削除：
他の制約データと矛盾するがどちらの制約データも同じ位の頻度で出現し、かつ、コストルールの変更でも対処できないと思われる場合には、制約データを削除する。

図７　コスト変数の決定の各処理の内容

コストルール：
ルール１：文節数の少ない解を選ぶルール
単語－単語，｛文節の切れ目｝　→＋Ｃ１
ルール２：単語数の少ない解を選ぶルール
単語－単語　→＋Ｃ２
ルール３：未知語の範囲を決定するルール
未知語－未知語，｛同一字種｝　→＋Ｃ３
ルール４：未知語の範囲を決定するルール
未知語－未知語，｛異なる字種｝→＋Ｃ４
ルール５：未知語の範囲を決定するルール
登録語－未知語，｛同一字種｝　→＋Ｃ５
ルール６：未知語の範囲を決定するルール
登録語－未知語，｛異なる字種　→＋Ｃ６
ルール７：未知語の範囲を決定するルール
未知語－登録語，｛同一字種　→＋Ｃ７
ルール８：未知語の範囲を決定するルール
未知語－登録語，｛異なる字種｝→＋Ｃ８

コスト変数：
Ｃ１＝１，Ｃ２＝１，Ｃ３＝１，Ｃ４＝１，
Ｃ５＝１，Ｃ６＝１，Ｃ７＝１，Ｃ８＝１

制約データベース：空

図８　初期状態

例文１：為替相場をいかにして安定させるか。

図９　例文の選択（例文１）

形態素解析結果：
為替相場(普通名詞)/をいかにして(未知語)/安定(サ変名詞)//させ(下一段動詞)/る(一段動詞語尾)/か(終助詞)/。(記号)

図１０　実行結果（例文１－１回目）

追加した制約データ：
誤り：為替相場(普通名詞)/をいかにして(未知語)/安定(サ変名詞)
正解：為替相場(普通名詞)/を(格助詞)/いかに(副詞)//し(サ変動詞)/て(接続助詞)/安定(サ変名詞)
誤り：安定(サ変名詞)//させ(下一段動詞)/る(一段動詞語尾)
正解：安定(サ変名詞)/さ(サ変名詞語尾)/せ(助動詞)/る(一段動詞語尾)

制約不等式：
2C2+C6+C8 > C1+5C2 ==> C6+C8>C1+3C2
C1+2C2 > 3C2 ==> C1>C2

コスト変数の値：
Ｃ１＝２，Ｃ２＝１，Ｃ３＝１，Ｃ４＝１，
Ｃ５＝１，Ｃ６＝３，Ｃ７＝１，Ｃ８＝３

図１１　コスト変数決定（例文１－１回目）

形態素解析結果：
為替相場(普通名詞)/を(格助詞)//いか(普通名詞)/にして(未知語)/安定(未知語)/させる(未知語)/か(終助詞)/。(記号)

図１２　実行結果（例文１－２回目）

追加した制約データ：
誤り：いか(普通名詞)/にして(未知語)/安定(未知語)/させる(未知語)/か(終助詞)
正解：いかに(副詞)//し(サ変名詞)/て(接続助詞)//安定(サ変名詞)/　さ(サ変名詞語尾)/せ(助動詞)/る(一段動詞語尾)/か(終助詞)

制約不等式：
2C2+C6+C8 > C1+5C2 ==> C6+C8>C1+3C2
C1+2C2 > 3C2 ==> C1>C2
・・・以上前の制約データからの制約不等式
4C2+2C3+2C4+C5+C7 > 2C1+7C2
==> 2C3+2C4+C5+C7 > 2C1+3C2
・・・追加した制約データからの制約不等式

コスト変数の値：
Ｃ１＝１，Ｃ２＝１，Ｃ３＝１，Ｃ４＝１，
Ｃ５＝２，Ｃ６＝３，Ｃ７＝２，Ｃ８＝３

図１３　コスト変数決定（例文１－２回目）

形態素解析結果：
為替相場(普通名詞)/を(格助詞)//いかに(副詞)//し(サ変動詞)/て(接続助詞)//安定(サ変名詞)//させ(下一段動詞)/る(一段動詞語尾)/か(終助詞)/。(記号)

誤りなし。

図１４　実行結果（例文１－３回目）

例文２：そして累積債務を抱えた開発途上国対策をどうするか。

図１５　例文の選択（例文２）

形態素解析解析結果：
そして(順接接続詞)//累積(サ変名詞)/債務(普通名詞)/を(格助詞)//抱え(普通名詞)/た(接頭語)/開発途上国(普通名詞)/対策(普通名詞)/を(格助詞)//どう(副詞)//する(サ変動詞)/か(終助詞)/。(記号)

図１６　実行結果（例文２－１回目）

追加した制約データ：

誤り：抱え(普通名詞)/た(接頭語)/開発途上国(普通名詞)
正解：抱え(下一段動詞)/た(助動詞)/開発途上国(普通名詞)

制約不等式：
2C2+C6+C8 > C1+5C2 ==> C6+C8>C1+3C2
C1+2C2 > 3C2 ==> C1>C2
4C2+2C3+2C4+C5+C7 > 2C1+7C2
==>2C3+2C4+C5+C7 > 2C1+3C2
・・・以上は前の制約データからの制約不等式
2C2 > 2C2 ==> 0>0
・・・追加した制約データからの制約不等式

コスト変数は決定できない。

図１７　コスト変数決定（例文２－１回目）

ルール９：/*普通名詞と接頭語の接続を敬遠するルール*/
@名詞－@接頭語 → ＋Ｃ９

注意）@は、複数の品詞のグループを表わす。

図１８　追加したコストルール

制約不等式：
2C2+C6+C8 > C1+5C2 ==> C6+C8>C1+3C2
C1+2C2 > 3C2 ==> C1>C2
4C2+2C3+2C4+C5+C7 > 2C1+7C2
==> 2C3+2C4+C5+C7 > 2C1+3C2
・・・以上は前の制約データからの制約不等式
2C2 +C9> 2C2 ==> C9>0
・・・追加した制約データからの制約不等式

Ｃ１＝１，Ｃ２＝１，Ｃ３＝１，Ｃ４＝１，
Ｃ５＝２，Ｃ６＝３，Ｃ７＝２，Ｃ８＝３，
Ｃ９＝１

図１９　コスト変数決定
（例文２－１回目、ルール９追加後）

形態素解析結果：
そして(接続詞)//累積(サ変動詞)/債務(普通名詞)/を(格助詞)//抱え(下一段動詞)/た(助動詞)//開発途上国(普通名詞)/対策(普通名詞)/を(格助詞)//どう(副詞)//する(サ変動詞)/か(終助詞)/。(記号)

誤りなし。

図２０　実行結果（例文２－２回目）

コストルール：
ルール１：文節数の少ない解を選ぶルール
単語－単語，｛文節の切れ目｝　→＋Ｃ１
ルール２：単語数の少ない解を選ぶルール
単語－単語　→＋Ｃ２
ルール３：未知語の範囲を決定するルール
未知語－未知語，｛同一字種｝　→＋Ｃ３
ルール４：未知語の範囲を決定するルール
未知語－未知語，｛異なる字種｝→＋Ｃ４
ルール５：未知語の範囲を決定するルール
登録語－未知語，｛同一字種｝　→＋Ｃ５
ルール６：未知語の範囲を決定するルール
登録語－未知語，｛異なる字種　→＋Ｃ６
ルール７：未知語の範囲を決定するルール
未知語－登録語，｛同一字種　→＋Ｃ７
ルール８：未知語の範囲を決定するルール
未知語－登録語，｛異なる字種｝→＋Ｃ８
ルール９：名詞と接頭語の接続のルール
@名詞－@接頭語　→＋Ｃ９

コスト変数：
Ｃ１＝１，Ｃ２＝１，Ｃ３＝１，Ｃ４＝１，
Ｃ５＝２，Ｃ６＝３，Ｃ７＝２，Ｃ８＝３，
Ｃ９＝１

図２１　コストルール・コスト変数の最終状態

誤り：為替相場(普通名詞)/をいかにして(未知語)/安定(サ変名詞)
正解：為替相場(普通名詞)/を(格助詞)/いかに(副詞)//し(サ変動詞)/て(接続助詞)/安定(サ変名詞)

誤り：安定(サ変名詞)//させ(下一段動詞)/る(一段動詞語尾)
正解：安定(サ変名詞)/さ(サ変名詞語尾)/せ(助動詞)/る(一段動詞語尾)

誤り：いか(普通名詞)/にして(未知語)/安定(未知語)/させる(未知語)/か(終助詞)
正解：いかに(副詞)//し(サ変名詞)/て(接続助詞)//安定(サ変名詞)/　さ(サ変名詞語尾)/せ(助動詞)/る(一段動詞語尾)/か(終助詞)

誤り：抱え(普通名詞)/た(接頭語)/開発途上国(普通名詞)
正解：抱え(下一段動詞)/た(助動詞)/開発途上国(普通名詞)

図２２　制約データベースの最終状態