# コスト最小法形態素解析のコストの学習方法

小松　英二
沖電気工業株式会社　マルチメディア研究所

## 1 はじめに

コスト最小法形態素解析で用いるコストを決定する方法として、解の候補の選好によりコストの値に対応するパラメータの制約不等式を集め、不等式全体の特殊解をコスト値とする方法を提案した [1], [2]。本稿では、上記の方法をを概整合ラベリング問題として定式化し、上記のコストの決定方法の妥当性について検討する。

## 2 コスト最小法形態素解析の概整合ラベリング問題としての定式化

形態素解析、構文解析等の自然言語処理アルゴリズムは、整合ラベリング問題として定式化できる [3]。整合ラベリング問題は、対象の局所的な制約に基づいて局所解の候補を求め、さらに対象全体として矛盾のない局所解の組合わせを求める問題である。さらに、局所的な候補に「誤差」と呼ばれる評価値を付与し、誤差の総和が最も小さい組合わせを解とする問題を概整合ラベリング問題と呼ぶ [4]。

概整合ラベリング問題は、(U,L,T,W,E)の５つ組で定義される。Uはユニット、Lはラベル、Tはユニット拘束関係、Wはラベリング誤差関数、Eは誤差の許容値と呼ばれる。誤差は、ラベリング誤差関数により、局所的な候補に与えられる評価値である。局所的な候補の組合わせには候補の誤差の和が誤差として与えられる。このような誤差の和を集積誤差と呼ぶ。このような定義のもとで、各ユニットに拘束関係を満たし、集積誤差が許容値に収まり、かつ、集積誤差が最も小さくなるようなラベルの組をユニットに付与するという問題である。定式化は、入力文毎に与えられる。以下に、「今日」という例文についての定式化を示す。ただし、誤差については、入力文と独立にパラメータとして用意されているとする。#^、#$は、それぞれ、文頭と文末を表わすダミーノードで、文頭と文末であることの誤差を定義するために追加した。ユニットは、文字列及び入力文での先頭の文字位置の組とした。

ユニット：

U={[#^,1],[今日,2],[今,2],[日,3],[#$,4]}

ラベル：

L={[#^,文頭],[今日,普通名詞],[今日,時詞],[今,普通名詞],[今,時詞],[日,単位],[日,普通名詞],[日,後置助数詞],[#$,文末]}

ユニット拘束関係：

T={t1,t2,t3,t4,t5}

t1=([#^,1],[今日,2]),t2=([今日,2],[#$,4]),t3=([#^,1],[今,2]),t4=([今,2],[日,3]),t5=([日,3],[#$,4])

ラベリング誤差関数：

W={W1,W2,W3,W4,W5}

W1:{([#^,文頭],[今日,普通名詞])→e1,([#^,文頭],[今日,時詞]→e2)},

W2:{([今日,普通名詞],[#$,文末])→e3,([今日,時詞],[#$,文末])→e4)},

W3:{([#^,文頭],[今,普通名詞])→e5,([#^,文頭],[今,時詞])→e6},

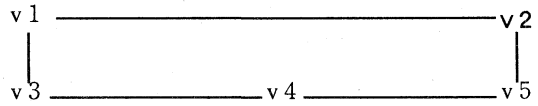
W4:{([今,普通名詞],[日,普通名詞])→e7,([今,時詞],[日,普通名詞])→e8,([今,普通名詞],[日,後置助数詞])→e9,([今,時詞],[日,後置助数詞])→e10}

W5:{([日,普通名詞],[#$,文末])→e11,([日,後置助数詞],[#$,文末])→e12}

拘束ネットワーク：

t1=([#^,1],[今日,2]),
W1:([#^,文頭],[今日,普通名詞])→e1
([#^,文頭],[今日,時詞])→e2

t2=([今日,2],[#$,4])
W2:([今日,普通名詞],[#$,文末])→e3
([今日,時詞],[#$,文末])→e4)




t3=([#^,1],[今,2]),
W3:([#^,文頭],[今,普通名詞])→e5
([#^,文頭],[今,時詞])→e6

t5=([日,3],[#$,4])
W5:([日,普通名詞],[#$,文末])→e7
([日,後置助数詞],[#$,文末])→e8

t4=([今,2],[日,3]),
W4:([今,普通名詞],[日,普通名詞])→e9
([今,時詞],[日,普通名詞])→e10
([今,普通名詞],[日,後置助数詞])→e11
([今,時詞],[日,後置助数詞])→e12}

誤差の許容値：
設定しない。

概整合ラベリング問題を解くには、拘束ネットワークにおいて、２つのノードを併合して、新しいユニット組とラベリング関数を作成することを繰り返す。一般的には、集積誤差が許容値を超えた場合には枝刈りをするが、形態素解析では組合わせの数が現実的な数で収まり枝刈りの必要がないため、誤差の許容値は考えなくてよい。従って、全ての局所解の組合わせが生成される。最終的な解を共通部分を圧縮したグラフ形式で書くと図２－１のようになる。これはコスト最小法形態素解析のグラフスタックと同じ表現であり、誤差がアークのコストに対応することが分かる。誤差は、最も優先する組合わせの誤差を０とするが、ここでは、誤差の許容度を用いないため、そのようにはなっていない。

コストの定義は、マルコフ過程の遷移確率、制約の優先度[5]等、様々な方法が可能であるが、本稿の方法では、コストを図２－２のc1～c11のようなパラメータの和で表わすことにしている。各コストは、品詞の組合わせ、字種、文字数の組み合わせのパタンに対して与えられると定義する。ラベリング関数の与える誤差e1,e2, ..., e12は、図２－３に示すように、これらのパラメータの和として定義される。パタンは、必要に応じて連接情報、単語見出しなども加えた定義も許している。パタンは、排他的でない場合も許しているため、上記のe7, e8, ..., e11のように、１つの誤差が複数のパラメータの和となる場合もある。パラメータは、３節で説明する制約不等式の特殊解の値で置き換えられ、形態素解析の実行時には、誤差は具体的な数値となっている。

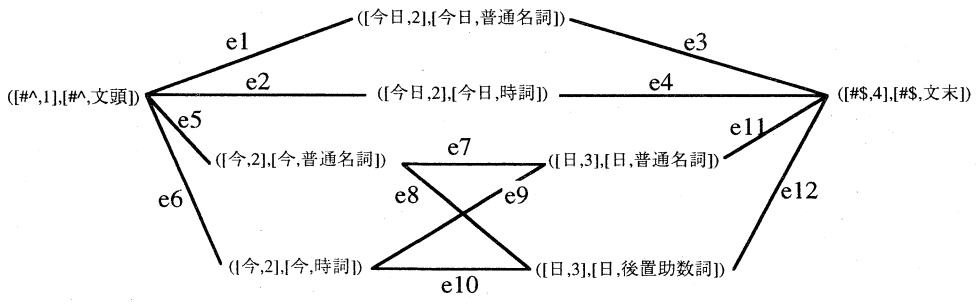



図２－１ 中間解

| パラメータ名 | パタン | パラメータ名 | パタン |
|---|---|---|---|
| p1 | 文頭－普通名詞 | p7 | 普通名詞－後置助数詞 |
| p2 | 文頭－時詞 | p8 | 時詞－後置助数詞 |
| p3 | 普通名詞－文末 | p9 | 後置助数詞－文末 |
| p4 | 時詞－文末 | p10 | 漢字－漢字（単語の接合部の字種） |
| p5 | 普通名詞－普通名詞 | p11 | 漢字１文字－漢字１文字 |
| p6 | 時詞－普通名詞 | | |

図２－２ コストを定義するパラメータ

e1=p1, e2=p2, e3=p3, e4=p4, e5=p1, e6=p2, e7=p5+p10+p11, e8=p6+p10+p11,
e9=p7+p10+p11, e10=p8+p10+p11, e11=p3, e12=p9

図２－３ 誤差とコストのパラメータの関係

### 3 誤差の決定方法

２節に用いた定式化を例として、誤差を定義するパラメータの値を決定する方法について述べる。今、「今日」という文を入力したときに、「#^(文頭)/今(普通名詞)/日(普通名詞)/#$(文末)」という結果が得られたとする。一方、別解に「#^(文頭)/今日(普通名詞)/#$(文末)」という解があり、この方が正解であるとする。組合わされた局所解の候補の誤差の和を集積誤差と定義すると、誤りと正解の集積誤差は、次のように表わされる。

誤り：p1＋p3＋p5＋p10＋p11

正解：p1＋p3

誤りより正解の集積誤差が小さいことから、次のような不等式ができる。

p1＋p3＋p5＋p10＋p11 ＞ p1＋p3

この式を整理すると次のようになる。

p5＋p10＋p11 ＞ 0

この式を選好を表わす制約不等式とする。

このようにして集めた制約不等式の特殊解によりパラメータの値を決定し、さらに、パラメータの和としてラベリング誤差関数の誤差の値を決定する。

ただし、作成された全ての制約不等式を満たす解は必ず存在するとは限らない。この理由として、第１に、パラメータ全体が構成する意味空間の次元が、すべての入力文に対して誤りと正解を弁別するのには不十分であること、第２に、拘束ネットワークが十分に大局的な組合わせを表現していないことがある。本稿での実験では、パラメータの数を大きくし、さらに、制約不等式が解けなくなったときに、詳細なパタンに対応するパラメータを追加することにより、上記の第１の原因を極力除去することで近似的に解決することにした。制約不等式を追加すると解がなくなる場合は、制約不等式を破棄することにした。

制約不等式は、１文毎に、すべての誤りと正解とのペアに対して作成しているわけではないため、２つの文の両方から、互いに矛盾する制約不等式が得られてしまい、本来どちらかが正しく解析されるにもかかわらず、２文とも解析が失敗するような制約不等式が得られてしまう危険性をはらんでいるが、この点については個別に対応することとし、別途検討中である。破棄した制約不等式に対応する誤りと正解の組は今後の課題として収集した。

### 4 誤差決定プログラム

３節で述べた方法の具体的なイメージとして、図４－１に、パラメータの値を決定する誤差決定プログラムの構成を示す。正解、誤りの判定は、コーパスとの比較により行う。パラメータを変更した場合には、制約不等式を全て作り直す必要があるため、原文、誤りの解析結果、及び、正解の解析結果を１組にして、「実例」として保存している。「制約不等式解決器」は、連立一次不等式を解くプログラムで、制約不等

式の特殊解を求める。「コストのパラメータ」には、制約不等式の特殊解をパラメータの値として保存する。例文は、学習用のトレーニングデータと評価用の評価用データがあり、５０文程度の文（句読点までを１文として分割した）からなるトレーニングデータに対して、パラメータの値が変化しなくなった時点で、誤差決定を中止し、評価用データを用いて解析性能を評価する。以上の処理は、全く人手を介さずに行うことができる。

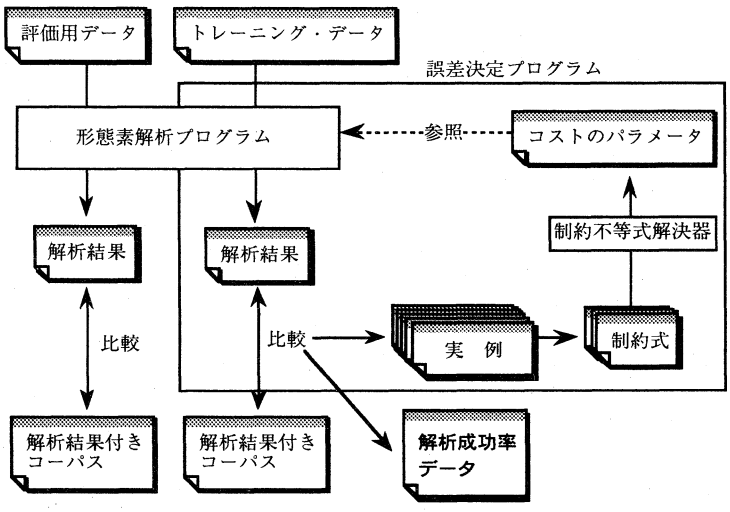



図４－１ 誤差決定プログラムの構成

## 5 おわりに

コスト最小法形態素解析を、概整合ラベリング問題として定式化し、本稿で用いたコストの決定方法が、集積誤差の制約不等式を解いていることを示した。整合ラベリング問題として定式化すると、ラベリング誤差関数の変わりに、ラベル拘束拘束関係と呼ばれる制約が用いられるが、この場合は、組合わせを減らすための制約を１か０の真偽値をとるラベルの組合わせとして記述する必要がある。コスト最小法形態素解析は、概整合ラベリング問題に対応するため、制約を選好として記述することができている。本稿で用いた方式の今後の具体的な解析成功率の向上の手段としては、(1) パラメータの数を増やす、(2) 形態素解析で大域的な誤差評価を行えるような処理を組み込むことにより、大域的な組合わせを考慮した誤差を定義できるようにすることが考えられる。(2) については、構文解析、文脈処理等を利用することが考えられる。また、構文解析についても同様の定式化に基づく解析方法が実現可能であると考えている。

## 参考文献


［１］小松，安原：コスト最小法形態素解析のコストルールの作成方法，自然言語処理研究会資料，８５－１　（１９９１）

［２］小松，安原：コスト最小法形態素解析のコストルールの作成実験，自然言語処理研究会資料，１０５－１　（１９９４）1］渡部，丸山：制約依存文法に基づいた日本語解析支援システム，自然言語処理研究会資料，６９６　（１９８８）

［３］西原，松尾，池田：概整合ラベリング問題における併合法の最適化と効率評価，人工知能学会誌，Ｖｏｌ．６　Ｎｏ．１（１９９１）

［４］西原，松尾：整合ラベリング問題における併合解法の並列化について，人工知能学会誌，Ｖｏｌ．３　Ｎｏ．２（１９８８）

［５］長尾　確：制約と選好としての知識を動的に統合して行う構造的多義性の解消，情報処理学会論文誌，Ｖｏｌ．３２　Ｎｏ．１０　（１９９１）