自然言語処理 123－2
（1998．1．19）

解析誤りデータを用いた
コスト最小法形態素解析のコスト関数の構成方法

小松 英二
沖電気工業株式会社 マルチメディア研究所
komatsu@hlabs.oki.co.jp

本論文では，コスト最小法形態素解析においてコスト設定規則が付与するコストの値を，コーパスを用いて，自動的に決定する方法について述べる。コスト決定に用いるデータとしては，形態素プログラムによる単語分割と，コーパスにある単語分割を比較し，一致しない文について、解析結果の単語分割と，コーパスの単語分割を組にしたものを用いる。収集したデータについて，コスト規則適用回数のベクトルの差を計算し，これに文識別子を付与したものを標本とする。すべての標本の平均を，コストの値とした。３００文のテキストを用いて評価を行い，コーパスとの自動比較による評価で，約９０％の精度を得た。

Automatic Construction of Cost Function for Minimum Cost Method Morphological Analyzer using Failure Data

Eiji KOMATSU
Media Laboratories , Oki Electric Industry Co. Ltd,

This paper describes a automatic method to decide cost values used by cost rules of minimum cost method morphological analyzer, using corpora. Data used are pairs of a failure result and a result in corpora. Then samples are made from data as difference vectors of frequencies of cost rule applications to above two results. The average of samples is used as cost values. The evaluation is done by 300 sentences, and success ratios by automatic evaluation using corpora reach about 90%.

1. はじめに

Viterbi アルゴリズム[1]（[1]）を用いた形態素解析における最適解の選択方法には，コスト最小法，確率モデル（[2]，[3]）を用いた手法，及び，これらを融合した方法（[4]）等がある。確率モデルを用いた最適解の選択方法では，コーパスを用いて自動的に確率パラメータを決定することができるが，コスト最小法については，自動的にコストを決定するための確立された方法はない。
本稿では，コスト最小解として出力された解析誤りと，コーパスの単語分割とから作られたデータを用いて，コスト最小法において用いられるコストを自動的に決定する方法について述べる。なお，説明の

---

1 Viterbi アルゴリズム：すべての可能な単語分割をグラフスタック，又は，フォーレストと呼ばれるグラフ構造の形式で生成し，グラフのノード，及び，アークに付与された評価値の総和が最大（確率モデルの場合），又は，最小（コスト最小法の場合）となるパスを選択するアルゴリズム

便宜上，形態素解析結果として，グラフ形式に圧縮された表現でなく，単語列を扱うが，方法は，一般のViterbiアルゴリズムで用いることができる。

2. コスト最小法のモデル

コスト最小法のモデルを定義する。コスト最小法では，「今日は天気が良い。」のような入力文に対して，以下の例のように，複数の単語分割が候補として生成される。

単語分割の例：

単語分割１：(文頭)/今(名詞)/日(名詞)/は(助詞)/天(名詞)/気(名詞)/が(助詞)/良(形容詞)/い(語尾)/。(句点)/(文末)

単語分割２：(文頭)/今日(名詞)/は(助詞)/天気(名詞)/が(助詞)/良(形容詞)/い(語尾)/。(句点)/(文末)

ただし，「(文頭)」，「(文末)」については，コスト計算の処理の都合上追加する単語である。

単語分割に対して，以下の例のような，コスト設定規則によりコストが与えられる。Viterbi アルゴリズムを用いるため，規則は，単語，及び，単語接続に関するものに制限する。

コスト設定規則の例：

コスト設定規則１：形容詞ならば，単語にコストＣ１を与える。
コスト設定規則２：名詞ならば，単語にコストＣ２を与える。
コスト設定規則３：助詞ならば，単語にコストＣ３を与える。
コスト設定規則４：文節区切でない単語区切りならば，単語接続にコストＣ４を与える。
コスト設定規則５：文節区切りである単語区切りならば，単語接続にコストＣ５を与える。
コスト設定規則６：１文字漢字名詞連続ならば，単語接続にコストＣ６を与える。

このとき，単語分割全体のコストの総和を計算する関数を，*コスト関数*と呼ぶことにすると，コスト関数は以下のように表わせる。ただし，コスト設定規則が与えるコストのベクトルを*コストベクトル*，コスト設定規則の１文での適用回数のベクトルを，*コスト規則適用回数ベクトル*と呼ぶことにする。

コスト関数：

$$f(X)=CX^{T}=\sum_{i=1}^{n} C_i X_i \qquad (2.1)$$

C：コストベクトル（C1，C2，．．．，Cn），Ciは，規則iが与えるコスト
X：コスト規則適用回数ベクトル（X1，X2，．．．，Xn），Xiは，規則iの適用回数
X^{T}は，Xの転置行列を表わす。

3．コストベクトルの決定方法

3．1 コーパス・データ・標本

コストベクトル決定に用いる，コーパス，データ，データから作られる標本について説明する。

まず，コーパスの例を示す。

コーパスの例：
入力文：今日は天気が良い。
単語分割：(文頭)/今日(名詞)/は(助詞)/天気(名詞)/が(助詞)/良(形容詞)/い(語尾)/。(句点)/(文末)
文識別子：０００００００１

次に，データ，および，標本を示す。形態素解析がコスト最小解として出力した単語分割が，コーパスにある単語分割と一致しなかったとする。この誤りを修正するように，コストベクトルを修正することを考える。このとき，誤ったコスト最小解の単語分割と，コーパスの単語分割，及び，文識別子を組にしたデータを作る。以下に，解析誤りデータの例を示す。

解析誤りデータの例：
誤ったコスト最小解の単語分割：
(文頭)/今(名詞)/日(名詞)/は(助詞)/天(名詞)/気(名詞)/が(助詞)/良(形容詞)/い(語尾)/。(句点)/(文末)
コーパスの単語分割：
(文頭)/今日(名詞)/は(助詞)/天気(名詞)/が(助詞)/良(形容詞)/い(語尾)/。(句点)/(文末)
文識別子：０００００００１

次に，コストベクトル変更のために，解析誤りデータの２つの単語分割についての，コスト規則適用回数ベクトルを計算し，解析結果の単語分割のベクトルから，コーパスの単語分割のベクトルを引いたベクトルを計算する。この，差ベクトルと文識別子の組を，コスト決定に用いる標本とする。以下に，標本の例を示す。例では，単語分割１，単語分割２に，コスト設定規則１～６を適用した場合のコスト規則適用回数ベクトルを用いた。

標本の例：
コスト規則適用回数ベクトルの差：(1, 4, 2, 10, 2, 2) − (1, 2, 2, 8, 2, 0)
= (0, 2, 0, 2, 0, 2)
文識別子：０００００００１

コーパスのすべての文について，このような標本を収集する。ここで，解析誤りデータ，及び，標本は，形態素解析で用いているコストベクトルの値に依存しており，コストベクトルの値が変われば，異なったデータ，標本が得られる。

3. 2 コストベクトルの決定式

いま，ある標本が得られたとする。誤りを修正するためには，f（誤りの単語分割のコスト規則適用回数ベクトル）＞f（コーパスの単語分割のコスト規則適用回数ベクトル）となるように，コストベクトルを変更しなければならない。f(X)は線形であるから，f（誤りの単語分割のコスト規則適用回数

ベクトル）－f（コーパスの単語分割のコスト規則適用回数ベクトル）＝f（誤りの単語分割のコスト規則適用回数ベクトルー コーパスの単語分割のコスト規則適用回数ベクトル）＝f（標本）となり，得られた標本に対して，f(X)＞0となることが必要である。

図3．1に，2 次元の場合について，コストベクトルを何回か変更して得られた標本と，あるf(X)の符号を示す。丸印は，標本を表す。f(X)が正になる領域は，図3．1に示すような半空間となる。このとき，できるだけ多くの標本に対して，f(X)＞0となるように，Cを決定すれば，多くの誤りが修正できることになる。

本稿では，2 次元における考察から，やや恣意的ではあるが，f(X)として，原点を通り，標本の平均のベクトルに垂直な平面が境界となるようなコスト関数を用いることにする。このような関数のコストベクトルは，以下の，（3．1）で与えられる。

$$C = (1/N) \sum_{i \in S} s_i \qquad (3.1)$$

S：収集された標本の全体

s_i：Sに含まれる標本。（$s_{i1}, s_{i2}, \ldots, s_{in}$）

N：収集された標本の総数

図3．2に，（3．1）により決定されるf(X)の符号を示す。

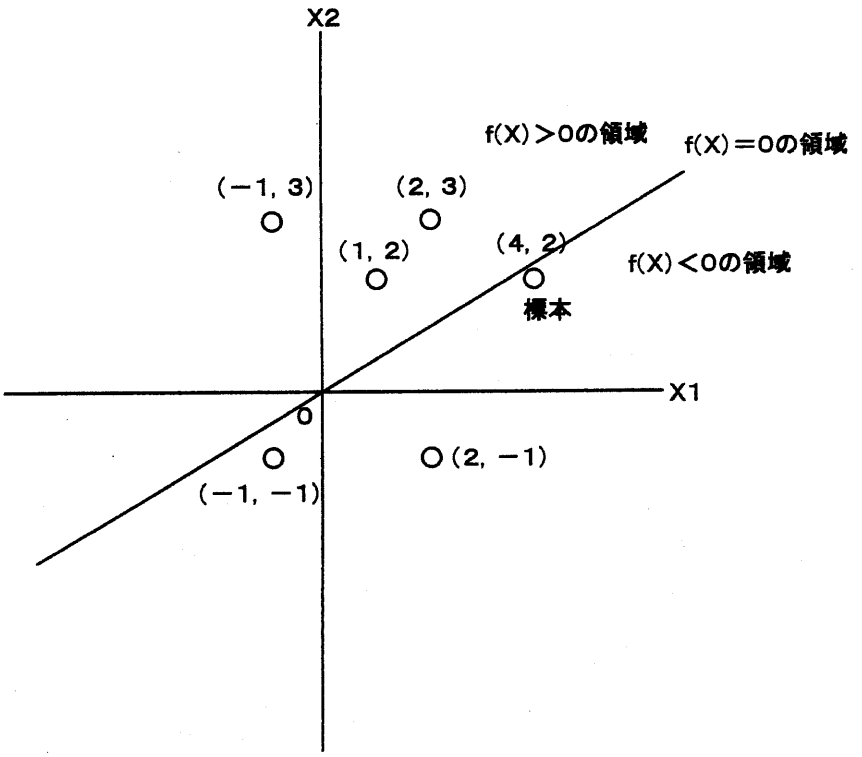

図3．1 f(X)の符号

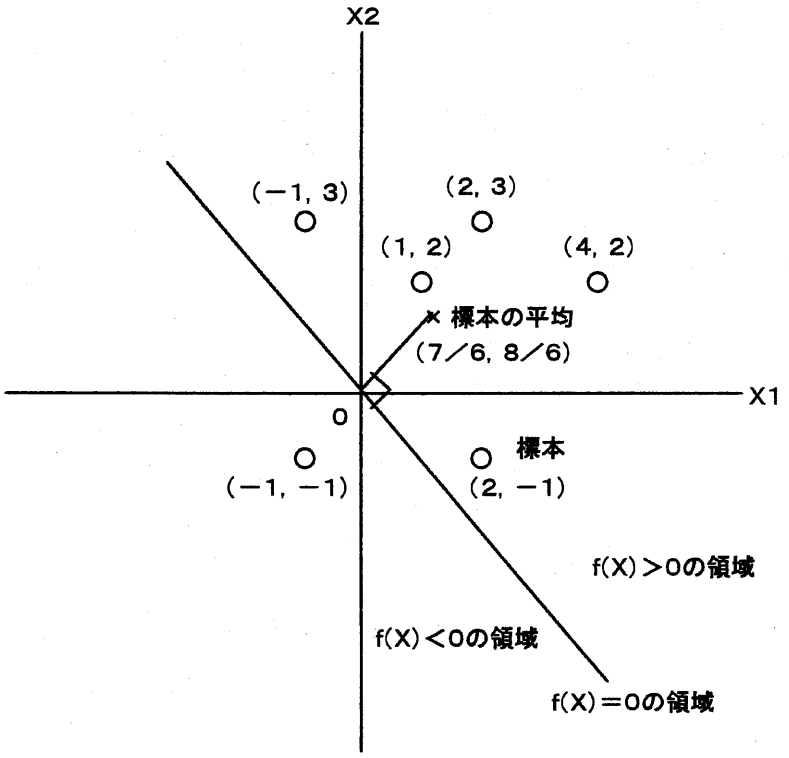

図3．2 （3．1）によって決定されるf(X)の符号

4. コストベクトル決定アルゴリズム

コストベクトル決定のアルゴリズムを示す。 Sは，収集した標本の集合，C（n）は，n回目のループで決定したコストベクトルを表わす。

［1］n＝0，S＝φ，C(0)＝（1，1，．．．，1）とする。
［2］nを1インクリメントする。
［3］C(n−1)を用いたコスト関数により，コーパスの文をすべてコスト最小法により形態素解析する。
［4］形態素解析結果である単語分割とコーパスの単語分割を比較して，解析誤りデータを収集し，標本を作成して，Sに追加する。ただし，同じ文識別子のついた標本は重複して追加しない。
［5］（3．1）式により，コストベクトルC(n)を決定する。
［6］C(n−1)とC(n)が等しくなければ，［2］へ戻る。
［7］C(n)を，最終的なコストベクトルとする

1つのコーパスで得られたコストベクトルをC(0)として，別のコーパスについて，コストベクトル決定処理を継続することも考えられるが，コストベクトルを変更した場合，最初のコーパスからは，変更に伴って得られるはずの標本が得られないため，本稿の方法では，コーパスを追加した場合は，全コーパスについて，コストベクトル決定をやり直すことにする。

5. コストベクトル決定プログラム

図5．1に，コストベクトルを決定するプログラムの構成，及び，データの流れを示す。形態素解析プログラムは，コーパス格納部からすべての入力文を取り出し，コスト設定規則格納部のコスト設定規則，及び，コストベクトル格納部のコストベクトルを用いたコスト関数により形態素解析を行い，単語分割を判別関数決定プログラムに渡す。単語分割には，文識別子を付与しておく。判別関数決定プログラムは，受け取った単語分割をコーパスの単語分割と比較し，一致しなければ，解析誤りデータを作り，さらに，コスト設定規則を用いて標本を作成し，標本格納部に追加する。同じ標本が既に標本格納部にあれば，標本は追加しない。コストベクトル決定プログラムは，すべての単語分割を処理し終わったら，標本格納部の全標本から，コストベクトルを計算し，新しい識別子をつけて，コストベクトル格納部に登録する。コストベクトルの識別子は，4節のアルゴリズムのループカウンタｎを用いる。このような処理を，コストベクトルの値が変化しなくなるまで繰り返す。

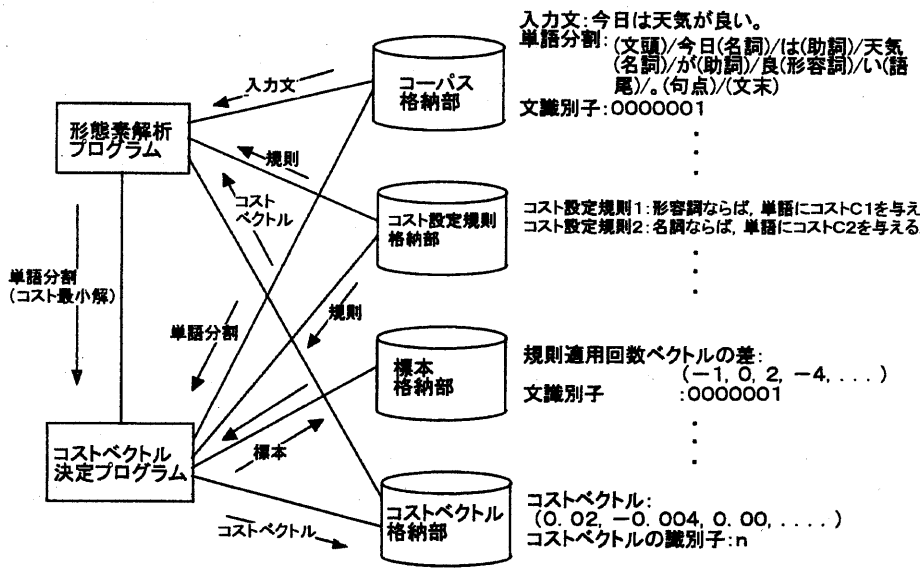

図5．1 プログラムの構成，及び，データの流れ

6. 実験

トレーニング用テキストとして，各１００文からなるテキストＴ１，Ｔ２，Ｔ３，計３００文を用いて評価を行った。テキストは，トレーニングの効率を高めるために，事前に，読点で分割した。読点は，両方の文に含ませた。図6．1は，文分割の例である。コスト設定規則は，約２２００個を用意した。このうち，約２０００が単語接続の品詞の組み合わせについての規則であり，残りが，文字数，字種，品詞の細分類に関する規則である。コーパスの単語分割には，品詞以外に，接続情報を記述した。

評価の方法としては，3つのテキストを，組み合わせを変えながら，トレーニング用と，評価用に分けて評価した。解析精度は，コーパスとの比較により，自動的に計算したため，普通名詞とサ変動詞の違いや，接続情報の違い等により，間違いとして認定されている場合が多く，人手によるチェックより，数パーセント低めになっている。

表6．1は，トレーニングにおける解析精度の変化である。図6．2は，Ｔ１をトレーニングテキストとしたときの，クローズドデータの解析精度の変化のグラフを示す。他のテキストについても，ほぼ，変化を示す。表6．2は，オープンデータによる解析精度である。

文１：今日は、天気良いので、公園に行きました。

↓

文１：今日は、。

文２：、天気良いので、。

文３：、公園に行きました。

図６．１ 文の分割の例

表６．１ トレーニングにおけるクローズドデータの解析精度の変化

トレーニングテキスト	解析精度	コストベクトルの更新回数					
		0	1	2	3	4	5
Ｔ１	再現率[1]	46.94	69.86	89.15	89.02	89.02	
	適合率[2]	46.25	73.42	90.63	90.63	90.63	
Ｔ２	再現率	46.64	69.03	86.40	87.78	87.97	
	適合率	46.67	73.95	89.06	89.97	90.06	
Ｔ３	再現率	48.01	69.85	87.64	87.60	87.82	87.82
	適合率	47.73	74.01	88.91	88.88	89.09	89.09
Ｔ１＋Ｔ２	再現率	46.79	69.52	87.61	87.83	87.83	
	適合率	46.46	73.72	89.72	90.01	90.01	
Ｔ２＋Ｔ３	再現率	47.25	69.36	87.80	87.85	87.85	
	適合率	47.14	73.96	89.53	89.58	89.58	
Ｔ３＋Ｔ１	再現率	47.42	69.42	87.49	89.00	87.91	88.85
	適合率	46.91	73.47	89.20	90.23	89.71	90.18

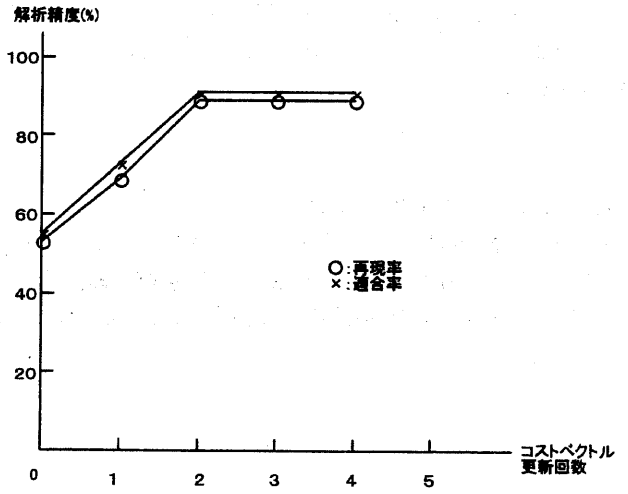

図６．２ Ｔ１のトレーニング時の解析精度の変化のグラフ

---

1 再現率(recall)：形態素解析が生成した正しい単語数÷コーパスの単語分割の単語数

2 適合率(precision)：形態素解析が生成した正しい単語数÷形態素解析が生成した単語分割の単語数

表６．２オープンデータによる評価結果

トレーニングテキスト	評価テキスト	解析精度	
		再現率	適合率
Ｔ１	Ｔ２	86.15	88.82
Ｔ１	Ｔ３	85.36	87.31
Ｔ２	Ｔ１	90.30	91.43
Ｔ２	Ｔ３	87.67	88.95
Ｔ３	Ｔ１	90.36	91.49
Ｔ３	Ｔ２	87.78	89.97
Ｔ１＋Ｔ２	Ｔ３	86.20	88.11
Ｔ２＋Ｔ３	Ｔ１	90.50	91.55
Ｔ３＋Ｔ１	Ｔ２	87.65	89.70

7. まとめ

コストベクトル決定の方法については，やや恣意的な決め方であるが，実験では，約９０％の精度が得られており，自動評価による解析精度の低下，コスト規則の改良等による精度向上の可能性を考慮すると，本方式はコスト決定として有効性があると考えられる。ただし，個々の誤りを見ると，方式に依存すると思われる誤りもあり，改善の余地がある。トレーニングテキストが１００文の場合と，２００文の場合とで，ほぼ精度に変わりがないことから，比較的少ないコーパスで，コストが決定できるといえる。

8. おわりに

コスト最小法においてコストを決定する関数を自動的に構成する方法について述べた。今後は，コーパスの大規模化，コストベクトルの計算式・標本の収集方法の論理的妥当性の検証，及び，種類の異なるテキストでの評価等を考えている。

[参考文献]

[1] 北，中村，永田，音声言語処理，森北出版，1996

[2] 竹内，松本，HMMを用いた形態素解析のパラメータ学習，情報処理学会第５０回全国大会，1995

[3] 永田，ＥＤＲコーパスを用いた確率的日本語形態素解析，EDR電子化辞書利用シンポジウム，1995

[4] 山下，松本：コスト最小法と確率モデルの統合による形態素解析，情報処理学会自然言語処理研究会研究報告，1997