接続コスト最小法による日本語形態素解析

Morphological Analysis of Japanese Sentences by Minimum Connective-Cost Method

1C−1

久光　徹　　新田　義彦
Toru HISAMITSU　Yoshihiko NITTA
日立製作所 基礎研究所
Advanced Research Laboratory, Hitachi Ltd.

形態素解析において生じる多数の解を，尤度により序列化して出力するための統一的な手法として「接続コスト最小法」を提案し，未登録語を含む教科書の文1000個を用いた実験結果を報告する．

［1］はじめに

日本語のように単語間に切れ目を置かない膠着言語の文の処理において，形態素解析は第一の関門である．形態素解析の解の個数は，一般に文字列の長さの指数関数となるため，解を効率よく尤度付けして出力する技術の確立が望まれる．個々の解の間の尤度を比較する手法は数多く提案されているが，尤度の高い順に解を導出するための計算量を評価した論文は少ない．その少数のうちの代表例として[4]があげられる．[4]は文節数最小法の基礎を与えるものであるが，全解を文節数により分類し，文節数の少ない解から出力するための解析表（付録参照）を，文字数 n に関して時間計算量$O(n^2)$で作製するアルゴリズムを与えている．しかし応用の観点からは，文節数だけでは尤度基準として弱いため，文節数最小解を求め，「自立語の後は付属語が来るものを優先する」などの基準を援用してさらに詳細な尤度付けをすることが多い．しかし，文節数最小解の個数は，一般に文字数の指数関数となるため，最尤解の出力に多大の時間を必要とする恐れがある．したがって，文節数最小法よりきめ細かい尤度付けができ，かつ，妥当な計算量が保証された手法を基礎付けることが望まれる．本報告では，そのような尤度付き形態素解析の手法を，実験結果と共に報告する．

［2］コスト付き形態素解析

尤度付きの形態素解析を形式的に表現するために記号を準備する．

Σ：文字の集合を表す有限集合．
W：Σの文字からなる語の有限集合．
C：カテゴリーコードを表す記号の有限集合．
D：見出し語とカテゴリーコードの対からなる有限集合．Dを辞書，Dの要素を形態素と呼ぶ．
Φ：カテゴリーコード間の接続可否を表現するための，C×Cから{0, 1} への関数．$c_1, c_2 \in C$に対して，$\Phi(c_1, c_2) = 0/1$ で，コードc_1を持つ形態素に，コードc_2を持つ形態素が接続不能／可能なことを示す．Φを接続テーブルと呼ぶ．
S：隣接する形態素間の接続が，接続テーブルにより許される形態素列の全体．すなわち，
$\{\langle s_1, c_1\rangle \cdots \langle s_n, c_n\rangle \mid$ すべてのjに対して $\langle s_j, c_j\rangle \in D,\ \Phi(c_j, c_{j+1}) = 1\}$．

以下では，Σ，W，C，Dは固定して考える．非負整数全体の集合をNで表し，記号 s は，常にΣの要素からなる文字列を表す．形態素解析とは，辞書Dと接続テーブルΦを参照して，文字列の可能な分割の仕方を与えることであるから，形式的には次のように書ける：

定義
形態素解析とは，s に対して，集合Seg(s)を与えることである．ここで，
$Seg(s) = \{\langle s_1, c_1\rangle \cdots \langle s_n, c_n\rangle \in S \mid$ 文字列としての連結 $s_1 \cdots s_n = s\}$．
Seg(s)の要素を形態素解析の解という．

SからNへの写像fが任意に与えられたとき，Seg(s)の要素は，fがとる値により分類できる．fをコスト関数，その値をコストと呼ぶ．解のコストが小さいほど尤度が高いと見なすことにより，尤度付きの形態素解析は，以上で準備した言葉を用いて，「コスト付き形態素解析」として定式化できる：

定義
fにもとづくコスト付き形態素解析とは，Nの任意の要素N，Kと文字列 s に対し，Seg(s)の要素をコストにより分類したとき，値が小さいものから高々N個までのグループに属する要素を，値が小さい順に高々K個出力することである．このとき，対{N, K}を，詳細度と呼ぶ．Nに制限を設けないとき，詳細度は{∞, K}と考える．詳細度{N, K}におけるNを深さと呼ぶ．

［3］接続コスト最小法

コスト付き形態素解析の効率はfに強く依存する．妥当な計算量が保証されるためには，適切なコスト関数を選ぶ必要がある．ここでは，そのようなコスト関数のひとつのクラスを与える．このクラスは，次の条件（C）を満たすコスト関数からなる：

【条件C】
関数$g_0: \Sigma^* \times C \to N$ と，第1変数に関して単調増加な関数$g_1: N \times (\Sigma^* \times C)^2 \to N$ が存在し，$\langle s_1, c_1\rangle \cdots \langle s_n, c_n\rangle \in S$に対し，fの値は次のように帰納的に定義できる：
$f(m_1) = g_0(m_1)$
$2 \leqq i \leqq n$ のとき
$f(m_1 \cdot m_2 \cdot \cdots \cdot m_i) = g_1(f(m_1 \cdot m_2 \cdot \cdots \cdot m_{i-1}), m_{i-1}, m_i)$．
ここで，$m_i = \langle s_i, c_i\rangle$．

定義
コスト付き形態素解析において，条件（C）を満たすコスト関数を用いる手法を，接続コスト法と呼ぶことにする．特に，深さN＝1の場合を，接続コスト最小法と呼ぶ．

接続コスト法の効率については付録に示した．接続コスト法は，従来の効率の保証された手法を包含する．以下，応用例の提示も兼ねて，それを示す．

以下では，コスト関数がカテゴリー間の連接のしやすさを示す関数Jを用いて次のように書ける場合を考える．この場合，条件（C）は満たされている：
$f(m_1) = g_0(c_1)$
$2 \leqq n$ のとき
$f(m_1 \cdot m_2 \cdot \cdots \cdot m_n\rangle) = \sum_{i=2}^{n} J(c_{i-1}, c_i)$
ここで，$m_i = \langle s_i, c_i\rangle$．$J(c_{i-1}, c_i)$を，カテゴリー$c_{i-1}$と c_iとの連接コストと呼ぶ．

以下ではΣ，W，C，D，Φを，それぞれ日本語の場合に特化して議論を進める．

1) コストによる分類をしない場合
最も特殊な場合として，コストによる分類をしない場合，$J \equiv 0$とすればよい．このときfをf_1と書くと，解を高々K個得るには，3項組$\langle D, \Phi, f_1\rangle$に関して，詳細度{1, K}のコスト付き形態素解析を実行すればよい．そ

の時間計算量は，付録の定理２より，入力文字列の長さ n に関して O(nK)で抑えられる．これは[1]の効率と等しい．（これをおおまかに説明すると，解析表の作製の時間計算量O(n)と，作製した解析表を用いてK個の解を出力するための時間計算量O(K)の積である．）

2) 文節数最小法

「文節数が最も少ない」ものが最も良い解であるとするよく知られた手法である．文節数は，各解を構成する形態素のコードのうち，「自立語」に対応するコードの数である．このとき，

$c_i \in$ 自立語のコードのとき $J(c_{i-1}, c_i) = 1$

$c_i \in$ 付属語のコードのとき $J(c_{i-1}, c_i) = 0$

とすることにより，文節数最小法を実現できる．文節数最小法は，[5]により未登録語に対処できるように拡張されている．紙面の都合上その詳細は述べないが，その場合も適当なJにより定まるfを用いて実現できる．これをf_2とする．f_2に関するコスト最小解を高々K個得るには，３項組＜D，Φ，f_2＞に関して，詳細度｛１，K｝のコスト付き形態素解析を実行すればよい．その時間計算量は，付録の定理２より，入力文字列の長さ n に関してO(nK)で抑えられる．全解から，コストの低い順に高々K個の解を得るには，３項組＜D，Φ，f_2＞に関する詳細度｛∞，K｝のコスト付き形態素解析を実行すればよい．付録の定理２より，時間計算量は$O(n^2K)$で抑えられる．これらは，[5]で述べられた効率と等しい．

3) さらに詳しい序列化を行う手法

接続コスト最小法の特徴をより有効に活かすために，カテゴリー間の接続しやすさを連接コストに反映することを考える．すなわちJの値に，

J(名詞，名詞) ＞ J(名詞，格助詞)

のような制約を設けることにより，解析表作製の終了と同時に，文節数最小法より詳細な尤度付けが得られるようにできる．

未登録語に対処するため，平仮名，片仮名，漢字等も形態素として扱い，これらを含む連接コストは，通常のカテゴリー間の連接コストより高く設定する．

このようにして定めたfをf_3とすると，３項組＜D，Φ，f_3＞に関する詳細度｛N，K｝のコスト付き形態素解析の時間計算量はO(nK)で抑えられる．詳細度｛∞，K｝の場合は，$O(n^2K)$で抑えられ，2)の効率と同等である．

［４］実験結果

前節のコスト関数f_1，f_2，f_3を用いて，実際の文章を解析した結果の概略を報告する．実験に用いた文は，[2]から無作為に選んだ1000文である．文単位での最小コスト解の個数，ブロック単位での最小コスト解の個数をまとめたのが表１である．ここでブロックとは，句読点や括弧など，形態素解析の曖昧さが生じない形態素で挟まれた部分を指す．文の平均文字数は44.5文字，ブロックの平均文字数は13.2文字である．

これらの文の特徴は，一般の技術文と比較して，平仮名書きの自立語（本実験で使用した辞書には登録されていない）が多いことである．その結果，未登録語が１文に１～２個存在するため，尤度付けを行わないf_1では，文全体の形態素解析の解の個数は非常に多くなる．

表１

コスト関数	ブロック単位の最小コスト解の平均個数	文単位の最小コスト解の平均個数
f_1	２５．３	3.49×10^3
f_2	１．７３	６．０１
f_3	１．２６	２．３５

注）f_1において，文単位での解の個数の平均をとる際，極端に解の個数が多い 20 文は除外した．そのうちの最大個数は，約2.5×10^9個であった．

表１から，f_2，f_3により，大幅な解の絞り込みが可能なことがわかる．

全ての場合について，最小コスト解の中に，未登録語以外の部分を正しく解析した結果が含まれていた．またf_3においては，平仮名書きされた未登録動詞の語幹が，ほとんどの場合正しく推定されていた．実験結果の詳しい解析については，別稿にて報告したい．

［５］おわりに

日本語形態素解析における尤度付き形態素解析の基礎として，「接続コスト法」を提案した．本手法は，妥当な計算量が保証された一般性の高い手法であり，適切なコスト関数を選ぶことにより，きめ細かく解を序列化することができ，未知語にも強いなどの特徴を持つ．

次の課題は，より容易なコスト関数の設定法の検討である．その手法と，今回の実験結果の詳細は別稿にて報告したい．

［付録］

fが条件（C）を満たす場合，詳細度｛N，K｝のコスト付き形態素解析を，作表型上向き横型探索で実現するアルゴリズムの時間計算量について述べる．この手法では，解析表というデータ構造を用いて，中間解析結果に関する情報をデータ共有により圧縮して保持し，それを利用して解を出力する．以下では，深さNを明示的なパラメータとした評価を示す．証明は[3]に示した．

定理１

i) fが条件（C）を満たすならば，深さNのコスト付き形態素解析のための解析表を生成する時間計算量は，入力文字列の長さ n に関して，O (nS(N)) で抑えられる．ここで，S(N)は，N個の数の sorting にかかる時間計算量の評価関数である．
（例えば，$S(N)=N\log_2 N$ or $N^{1.2}$）．

ii) 単調増大な関数h：N→Nが存在し，Seg(s)の各要素のコストが，s の長さ n に関して，h(n)で抑えられるとする．このとき，深さ∞のコスト付き形態素解析のための解析表を生成する時間計算量は，入力文字列の長さ n に関して O (nh(n)) で抑えられる．

定理２

i) ３項組＜D，Φ，f＞に関する詳細度｛N，K｝のコスト付き形態素解析の時間計算量は，入力文字列の長さ n についてO (nKS(N))，で抑えられる．ここで，S(N)は，定理１と同じ．

ii) 定理1-ii)の条件のもとで，３項組＜D，Φ，f＞に関する詳細度｛∞，K｝のコスト付き形態素解析の時間計算量は，入力文字列の長さ n について，O(nh(n)K)で抑えられる．

［謝辞］福岡大学日本語処理用辞書の使用を認めてくださった福岡大学・首藤 公昭教授に感謝します．

［参考文献］

[1] 杉村 領一 他：論理型形態素解析ＬＡＸ，Proc. of the Logic Programming Conf., 12.3(1988)

[2] 近角 聡信 他：精選理科Ｉ（生命・科学編，及び力学・生命科学編），東京書籍(1981)

[3] 久光 徹 他：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，電子情報通信学会研究会報告，NLC 90-8，pp17-24(1990)

[4] 吉村 賢治 他：文節数最小法を用いたべた書き日本語文の形態素解析，情報処理学会論文誌，Vol.23,No.6，pp40-46(1983)

[5] 吉村 賢治 他：未登録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,pp294-300(1989)