1C－1

HMMを用いた日本語形態素解析システムの確率情報分析*

木村 健† 乾 伸雄† 野瀬 隆† 小谷 善行† 西村 恕彦†
東京農工大学 工学部 電子情報工学科

1 はじめに

HMMを用いた形態素解析システムは，英語文での解析結果が極めて良好であるにもかかわらず，日本語文に対してそのまま適用した場合，あまりよい結果が得られないことが多い．日本語の単語の区切りが明確でないことなどが，解析成功率を下げている大きな原因と考えられる．

本論文はHMMの確率情報を求める時点で，後向きに(横書きでいうと右から左に)品詞間の遷移確率を求める3種の方法を示す．この方法は，日本語の句において中心となる単語に他の単語が係りやすいという性質を利用し，品詞間の依存関係を確率情報に取り入れやすくしている．

2 システムの概要

本方法によるわれわれの形態素解析システムは，確率情報を品詞付きコーパスから獲得する部分と，得られた確率情報を用いて形態素解析を行なう部分に分かれている．確率情報は，頻度の形で二つのデータベースファイルに格納される(図1)．

データベース1 品詞の遷移頻度

データベース2 単語の出現頻度

品詞の遷移頻度については，となり合う三つの品詞について求めたものを使用する(tri-gram)．単語の出現頻度は，一つの品詞に対しての頻度を使用する(後述)．品詞体系は，RWCコーパスの品詞系[2]をもとにした，九つの素性構造からなるものを使用する．

*A Parameter Learning Technique for Japanese Morphological Analyzer based-on Hidden Markov Model.

†Takeshi KIMURA, Nobuo INUI, Takashi NOSE, Yoshiyuki KOTANI, Hirohiko NISIMURA, Department of Computer Science, Tokyo University of Agriculture and Technology

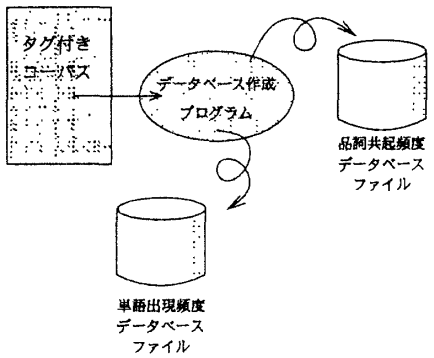

図1: 確率情報データベースの獲得

形態素解析は，HMMを扱えるようにした最小コスト法で行なう．

3 確率情報の獲得方法

連続するn個の品詞からなる列を$(T_1, T_2, \ldots, T_n)$とするとき，品詞の遷移確率は次式より推定できる．

$$P(T_3|(T_1,T_2)) = \frac{\mathbf{num}(T_1,T_2,T_3)}{\mathbf{num}(T_1,T_2)} \quad (1)$$

$\mathbf{num}(T_x)$は，品詞T_xの頻度を示している．また，任意の単語をW_i $(W_i \in W_{\mathrm{corpus}})$とすれば，

$$P(W_i|T_1) = \frac{\mathbf{num}(W_i,T_1)}{\mathbf{num}(T_1)}$$

である．W_{corpus}は，確率情報を得るために使うコーパス中に含まれる，すべての単語の集合である．

ここで，式(1)について，入力単語が，先頭から順にW_{i1}, W_{i2}, W_{i3}であるとき，遷移確率$P(T_{i3}|(T_{i1},T_{i2}))$ではなく，

$$P(T_{i1}|(T_{i2},T_{i3})) = \frac{\mathbf{num}(T_{i1},T_{i2},T_{i3})}{\mathbf{num}(T_{i2},T_{i3})} \quad (2)$$

を得ることを考える．これは文末から文頭への方向を持つ遷移確率である．ここから式(1)の確率を前

方向確率，式 (2) を**後方向確率**と呼ぶことにする．単語の出現頻度は，データベース 1 の品詞の出現頻度 **num**(T_k) と，データベース 2 の単語と品詞の組の頻度 **num**(W_j, T_k) から求められる。

4 形態素解析

Viterbi のアルゴリズムを使って，最適解を求める (図 2)．具体的には，次の目的関数を最大にする品詞列 $(T_1, T_2, \ldots, T_n)$ を求める．

$$P(T_0)P(T_1|T_0)P(W_1|T_1) \times \prod_{i=2}^{n-1} P(W_i|T_i)P(T_i|(T_{i-2}, T_{i-1})) \times P(W_n|T_n)P(T_n|T_{n-1})P(T_{n+1}) \quad (3)$$

(T_0 と T_{n+1} は，文頭，文末というカテゴリ)

実際は，コストとして確率値の対数を与え，最小コスト問題として考えられる．コスト値は負になるが，遷移確率と単語出現確率の積は，対数の和として扱う．

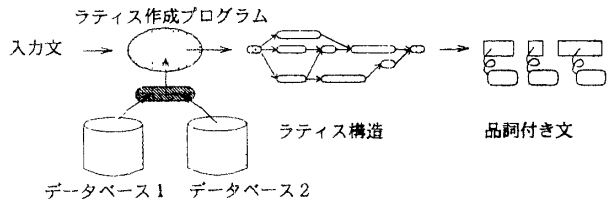

図 2: 形態素解析

5 後方向確率の適用

4 章の形態素解析に，後方向確率を形態素解析に適用する，3 種の方法を述べる．

5.1 逆方向への解析

普通の解析と同じように，ただし，文末から解析を行なう．このとき後方向確率を，前方向確率と同様に適用する．本質的に，前から解析を行なうのとほとんどかわらないシステムとなる．

5.2 後方向確率による品詞候補の修正

まず前方向から解析を行なう．ある入力文字列中の文字間において，前の二つの品詞 T_{i-2}, T_{i-1} より品詞 T_i の候補が M 個得られたとする．これを $T_i^{1\ldots M}$ と表す．推定された品詞について，その次の二つの単語候補すべてを調べ，さらにその品詞の候補を調べる．こうして，$T_{i+1}^{1,2,\ldots,L}$ と $T_{i+2}^{1,2,\ldots,K}$ が求まったと仮定する．

得られた品詞 bi-gram のすべての組み合わせについて，推定された品詞と品詞 bi-gram の後方向確率を調べる．

$$R(j = 1, 2, \ldots, O) = P(T_i^{Fa(j)}|(T_{i+1}^{Fb(j)}, T_{i+2}^{Fc(j)}))$$

O は $M \times L \times K$ であり，順序列 $Fa(j), Fb(j), Fc(j)$ はほかのすべての $j' \neq j$ による順序列と，重ならない．$R(j)$ が最も大きくなるように，品詞 T_i を求める．

5.3 後方向確率，前方向確率による A^*アルゴリズム

始めに，5.1 章のように，後方向確率を使ってラティス構造を構築する．次に，文頭からも同様にして部分最小解を求め解析を行なう．ここで，ヒューリスティック関数を文末からの部分最小解とし，文頭からの部分最小解と足し合わせたものを，目的関数とする．なお，文頭からの文字数に依存した関数で重み付けを行なう．

6 まとめ

5.1 章の方法は，従来の HMM 形態素解析システムに対してほとんど変更を加える必要がなく，実現が容易である．5.2 章の方法は，文頭から文末へ，一回の探索で最適解を見つけることができる．5.3 章の手法は，文中の位置に敏感な HMM 形態素解析システムをつくることができる．以上の三つの方法について，実装評価を予定している．

参考文献

[1] 永田 昌明．前向き DP 後向き A^*アルゴリズムを用いた確率的日本語形態素解析システム，情報処理学会研究報告，自然言語処理研究会，NL-101-10，pp.73-80，1994.5.

[2] 井佐原 均 ほか．RWC データベースワークショップ・テキストグループ 平成 7 年度報告書 (RWC コーパス付属ドキュメント)．

[3] A.Kempe. Probablistic Tagging with Feature Structures. COLING94,pp.172–176.