確率付決定木を用いた日本語形態素解析

柏岡 秀紀 , Stephen G. Eubank , Ezra W. Black
ATR 音声翻訳通信研究所

1 はじめに

現在，かなり精度の良い形態素解析システムが報告されている．形態素解析では，“形態素をどのような単位で扱うか”，“品詞タグとしてどのような体系を扱うか”は，重要な問題であり，処理する対象，目的などにより異なる．現在，各システムで使われている品詞タグの体系は，統一されたものではなく，形態素の単位についても，標準的な単位があるとはいえない．既存のシステムの体系と実際に利用したい体系の間には，差異が生じることが多い．その形態素や品詞体系の差異を吸収するために，既存システムの出力を修正する処理を組み込んだり，既存システムの辞書を利用する体系で整備し，品詞の接続関係を設定し直す必要が生じる．しかし，辞書の整備はコストがかかり，分野に依存した辞書の作成も，困難である．さらに，接続関係に付随するパラメータの人手による微妙な調節は困難である．そのため，これらの修正を行ない，現状で報告されている精度を保つことは難しい．そこで精度向上の1解決策に，形態素解析に必要な知識（形態素の特徴や接続関係）をコーパスから学習する，あるいは統計的に処理することが考えられる [6, 1, 2]．この種の学習は，現状のパラメータの調整を人手に変わって行なうという目的で提案されることが多く，形態素を扱う単位の変更や品詞の体系の修正に対する形態素解析に必要な知識の学習は，ほとんど行なわれていない．

本稿では，多様な形態素の単位や様々な品詞タグの体系で利用できる形態素解析手法を提案する．本手法は，形態素を構成する部分的な文字列，単語の階層的な分類，形態素・品詞の接続関係などの特徴に着目し，タグ付コーパスからその統計的特徴を学習し，形態素解析を行なう．

2 形態素解析の問題点

従来の形態素解析は，主に，1）入力文を形態素に分割し，2）各形態素に品詞タグの候補を与え，3）得られた形態素列，品詞タグ列の候補に対して優先度を調べ解とする，という3つの処理から構成されている．

1），2）の処理では，辞書[1]から得られる情報を利用している．そのため，辞書にない未知語や，見出しとして辞書に登録されていても品詞等の属性が登録されていない場合に，特別な処理を必要とする．未知語を減らすために，コーパスから未知語を抽出，収集する手法が提案されている [3, 4, 5]．しかし，正解となる形態素が全て辞書に登録されていても，3）の優先度によっては，誤った解が優先されることもある．

3）の処理では，各形態素の優先度，形態素の接続関係（主に，品詞等の属性による接続関係で制限），および各接続関係の優先度の設定が必要となる．既存のシステムでは，多くの場合，形態素の優先度や接続関係の優先度の判断は，独立に計算され，それらの優先度を組み合わせるための計算式を定義している．しかし，人手による調整が困難であり，コーパスを利用した手法が提案されている．各システムで利用されている接続関係は，主に連接する前後の品詞属性に対する情報であり，連接しない形態素間の接続関係を利用しているシステム [7] は，ごく少数であり，統一的な判断基準となる計算式の作成は困難である．

語の構成規則や語法の特徴から得られる情報を積極的に利用することで，辞書を利用せずに処理できれば，辞書の保守や未知語の問題を解消することができる．また，決定木 [9] を利用することで，多様な知識に関する判断基準に一貫性を持たすことができる．

3 決定木による形態素解析

前節で述べたように，語の構成規則などの特徴を決定木に用いた形態素解析処理の手法を提案する．これまでに英語に対して，語の構成規則などの特徴を利用した品詞付与システム作成し，この手法の有効性を実験により確認している [10]．英語の品詞付与システムでは，語の構成規則などの特徴を決定木の選択項目として用い，その処理を実現している．本稿で提案する形態素解析手法でも，同様に決定木を用い処理を実現する．

決定木は，対象を複数の属性とその属性値から，適切なクラスに分類する木構造のモデルであり，様々な特徴を対象の属性として利用することで，同じ枠組の中で利用することができ，一貫し

[1] ここでいう辞書とは，形態素の見出し，および見出しに対する品詞等の属性の組の情報が記述されているものとする．

たコストを与えることができる.

以下では，利用する特徴，決定木の学習，形態素解析における確率モデル，および処理のながれについて述べる.

3.1 利用する情報

決定木を利用する際には，決定木の選択項目として，どのような属性を対象の属性として利用するかが重要である．本手法では，決定木の選択項目として形態素の構成規則や語法の特徴から得られる情報を利用する．利用する特徴は，次の3つの観点により捉えることができる.

I 形態素の構成の特徴
II 単語の分類体系上の特徴
III 文脈による特徴

形態素の構成の特徴や単語の分類体系上の特徴は，従来の辞書情報を補う特徴として捉えることができる．また，従来よく利用されている隣接した品詞の接続関係を，隣接する2つの品詞の関係だけでなく，様々な接続関係を柔軟に捉えるようにするために，文脈による特徴として捉えた．以下に個々の観点による特徴について説明する.

形態素の構成の特徴

「形態素が“大きな”のように“な”で終る」というような部分的な文字列の特徴や，「形態素は4文字からなる」という長さの特徴，機能語を部分的にまとめた，ある意味では辞書的な特徴などを利用している．音声対話を対象とする場合には，文字種や句読点の情報は，利用できないが，現在の実験では，対話の書きおこしデータを利用しているので，文字種の特徴（カタカナだけ，漢字混じり等）や句読点の情報も利用している.

単語の分類体系上の特徴

ある種のシソーラスを利用しているもので，実験には，コーパスから相互情報量を利用して単語を階層分類したもの [8] を利用している.

文脈による特徴

「直前の形態素の品詞」などの直接つながっている前後の形態素だけでなく，「文頭の形態素の品詞」など文の構成に重要な役割を持つと思われる形態素や，「直前の形態素は“な”で終る形態素である」というように近くに現れた形態素に関する文字構成の特徴や素性などを利用している．また，熟語的に利用される形態素列も特徴として利用している.

3.2 決定木の学習

3.1で述べた特徴を選択項目として，形態素解析済みコーパスを用いて決定木学習をおこない，形態素への分割を目的とした決定木，および形態素への品詞付与を目的とした決定木を学習する．決定木学習では，予め用意された学習データから，利用する特徴を対象の属性とみなし，“対象，および，属性とその値”の組からなる情報を事例として用い，その事例の集合を，特定の属性の値で分類するために，有効な属性を用いて繰り返し分割することにより学習する (図1)．事例の集合を分割する属性の有効性を判断するために，評価尺度としてエントロピーを用いる.

分割の対象となる事例の集合に対して，分割後のエントロピーが最小となる属性を求め，ノード (ルートノード) の選択項目として，その属性を割り当てる．選ばれた属性により分割した各事例の集合に対して同様の処理を繰り返し行ない，決定木を学習する．事例集合の分割は，次の条件をすべて満たす限り続けられる.

1) 各ノードに含まれる事例数が一定数以上
2) 分割によるエントロピーの減少量がある基準値以上

分割が停止したノードを *leaf* とよび，その *leaf* ノードに存在する事例の集合から目的とする形態素あるいは品詞タグの確率を計算する[2].

3.3 統計的モデル

決定木を利用した形態素解析で用いている統計的モデルについて述べる.

日本語の入力文として $C = c_1\, c_2 \ldots c_n$ の N 個の文字列が，$W = w_1\, w_2 \ldots w_m$ の M 個の形態素に分割され，各形態素の品詞タグが，$\mathrm{T} = t_1\, t_2 \ldots t_m$ をとるとする[3].

形態素解析は，文字列 C に対して，単語列 W と品詞タグ列 T の同時確率 $P(W,T|C)$ を最大化する形態素列と品詞タグ列を求める問題として捉えられる．本手法では，$P(W,T|C)$ を，以下の式により近似する.

$$P(W,T|C) = \prod_{i=1}^{M} P(w_i, t_i | w_1, ..., w_{i-1}, t_1, ..., t_{i-1}, C)$$

$P(w_i, t_i | w_1, ..., w_{i-1}, t_1, ..., t_{i-1}, C)$ を，$i-1$ 番目までの形態素列および品詞タグ列を考慮した形態素 w_i の出現確率 $P(w_i | w_1, ..., w_{i-1}, t_1, ..., t_{i-1}, C)$

[2]実際のシステムでは，スムージングを行なっている．また，エントロピー基準以外の学習の実験も試みている.

[3]t_i は，i 番目の形態素の品詞タグ

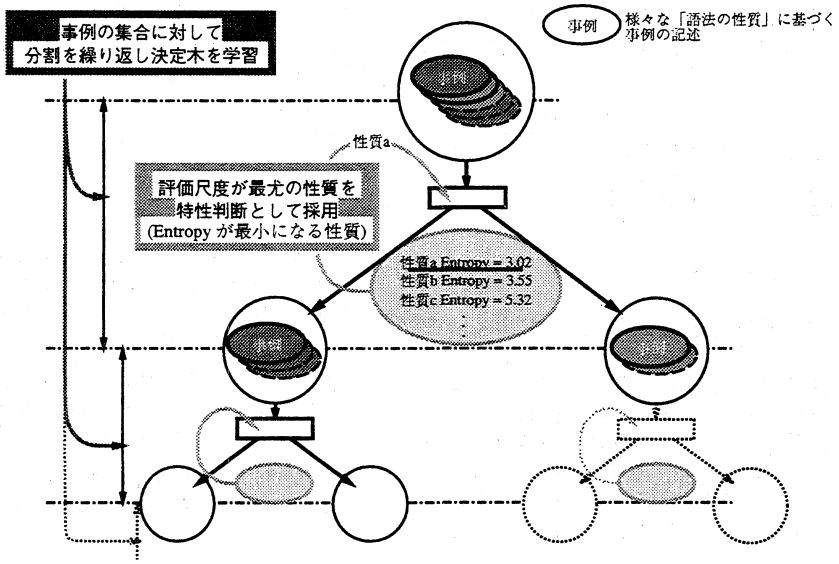

図 1: 決定木学習の過程

と単語 w_i まで考慮した品詞タグ t_i の出現確率 $P(t_i|w_1,...,w_i,t_1,...,t_{i-1},C)$ の積とする．この形態素 w_i の出現確率と品詞タグ t_i の出現確率は，それぞれ形態素への分割を目的とした決定木，および形態素への品詞付与を目的とした決定木において計算する．

これは，形態素への分割を目的とした決定木において，$(i-1)$ 番めの形態素まで品詞を付与した状態で，i 番めの形態素候補が分類される $leaf$ ノードを $leaf(Lw_i)$ とする．Lw_i の確率分布を，p_{Lw_i} により表現し，i 番めの形態素が w_i である確率を，決定木の条件付確率分布を用いて以下のように近似する．

$$p(w_i|w_1,...,w_{i-1},t_1,...,t_{i-1},C) \approx p_{Lw_i}(w_i) \quad (1)$$

同様に，形態素への品詞付与を目的とした決定木において，i 番めの形態素が分類される $leaf$ ノードを $leaf(Lt_i)$ とする．

$$p(t_i|w_1,...,w_i,t_1,...,t_{i-1},C) \approx p_{Lt_i}(t_i) \quad (2)$$

本手法では，形態素，および品詞タグの出現をマルコフ情報源として取り扱っていない．従って，十分に長い文において，文の最初の形態素やその品詞タグに依存して文末の形態素およびその品詞タグを導くことが，原理的には可能である．

3.4 処理のながれ

提案する形態素解析システムでは，一文字目から順に，1）形態素の候補を調べ，2）各形態素候補に品詞タグを割り当て，次の文字を読み込み，1）2）を繰り返す．

品詞タグは，いわゆる品詞，活用種類，活用型などの属性の組合せで表現し，形態素に品詞タグを割り当てる際には，2段階の処理により，各形態素の品詞タグを決定している．第1段目の処理として，形態素に k 個の品詞の候補を与えるもので，第2段目の処理として，形態素に与えられた k 個の品詞の各々に対して，品詞タグの候補を l 個与える．

このような処理では，少なくとも1形態素に対する品詞候補だけでも，$k*l$ の候補があり，それまでに得られている可能性のある形態素列および品詞タグ列全てに関して，考慮する必要があり探索範囲が非常に膨大になる．そこで，stack decoder アルゴリズム [11, 12] を用いて，確率が最大となる品詞列を探索している．

4 実験

現在，形態素に分割するための決定木を学習するモジュールを作成しているところである．ここでは，正しく形態素分割行なわれた場合の品詞タグを付与するための実験結果について示す．

品詞付与実験

本実験では，対象としたテキストは，ATR で収集している旅行会話に関する対話データの一部

で，用例主導翻訳のための形態素解析済みデータを用いた．学習データ，評価データの語数，文数を表1に示す．

	語数	文数
学習データ	146453	11282
スムージング	28933	2231
評価データ	55544	4147

表 1: 実験データ 1

品詞タグは，変換主導による翻訳システムのために作成された品詞体系に基づく18品詞に活用種類等の属性を付加した209品詞タグの体系を用いた．その結果，91.6%の正答率が得られた．この正答率は，学習データから作成した辞書(出現した形態素と付与されている品詞タグとの組合せ)を利用して，各形態素にもっとも高頻度の品詞タグを付与するという手法での正答率とほとんど差がなかった．

5 考察

実験では，品詞タグを付与するための決定木の選択項目に基本的な特徴のみを用いており，特徴を増やすことにより，より精度を高めることができると考えている．また，本実験で用いた形態素，品詞タグの体系は，変換主導翻訳システムのために開発された体系であり，一つの形態素の持つ品詞タグは，ほぼ一つ[4]であるため，辞書として，品詞を持つことは非常に有効な手段となる．しかしながら，正しい形態素に対する品詞付与の実験において，辞書を利用せずに決定木を利用した結果と，辞書を利用した結果がほぼ同等の結果を得ることができていることから，決定木による手法は有効であると考えられる．

6 おわりに

本稿では，確率付決定木を用いた形態素解析手法を提案した．決定木において，語法の特徴や形態素の構成の特徴の統計的特徴を利用することで，多様な形態素の単位，および様々な品詞タグの体系において，利用することができると思われる．現在，形態素への分割のための決定木を構築し，利用するモジュールを作成中であり，形態素分割での有効性を確かめる必要がある．しかし，品詞付与に対しては，英語において有効性が確認されており，日本語に対しても，本手法の有効性を示唆する実験結果を得た．

[4]本実験で用いた学習データでは，一単語，平均1.01の異なる品詞タグを持つ

今後は，まず，形態素分割での有効性を実験により確認し，学習データ量，および利用する特徴量との関係について調べるとともに，学習された決定木の修正等についても考察を加えていくつもりである．

参考文献

[1] M. Yamamoto: "A Re-estimation Method for Stochastic Language Modeling from Ambiguous Observations", Fourth Workshop on Very Large Corpora, pp.155–167, 1996.

[2] 竹内，松本: "HMMによる日本語形態素解析システムのパラメータ学習"，情報処理学会，95–NL–108，pp.13–19，1995.

[3] 森，長尾: "n グラム統計によるコーパスからの未知語抽出"，情報処理学会，95–NL–108，pp.7–12，1995.

[4] 中渡瀬: "正規化頻度による形態素境界の推定"，情報処理学会，96–NL–113，pp.13–18，1996.

[5] 永田: "単語頻度の期待値に基づく未知語の自動収集"，情報処理学会，96–NL–116，pp.13–20，1996.

[6] M. Nagata: "A Stochastic Japanese Morphological Analyzer Using a Forward–DP Backward–A* N–Best Search Algorithm", COLING–94, pp.201–207, 1969.

[7] 渕 武志，松岡 浩司，高木 伸一郎: "保守性を考慮した日本語形態素解析システム"，情報処理学会，97–NL–117，pp.59–66，1997.

[8] A. Ushioda: "Heirarchical Clustering of Words and Application to NLP Tasks", Fourth Workshop on Very Large Corpora, pp.28–41, 1996.

[9] L. Breiman, J. Friedman, R. Olshen, and C. Stone: "*Classification and Regression Trees*", Wadsworth & Brooks/Cole, Monterey, CA. 1984.

[10] E. Black, S. Eubank, 柏岡: "The Non–Dictionary: Description and Evaluation of a Dictionaryless Semantic and Syntactic Tagger for Unrestricted English Text"，言語処理学会第3回年次大会発表論文集，1997.

[11] F. Jelinek: "A fast sequential decoding algorithm using a stack", *IBM Journal of Research and Development*, 13:675–685. 1969.

[12] D. Paul: "Algorithms for an optimal a^* search and linearizing the search in the stack decoder", *Proceedings of the June 1990 DARPA Speech and Natural Language Work shop*. 1990.