Probabilistic Decision-Tree Tagging Without A Dictionary *

1L－1

柏岡 秀紀 , Ezra W. Black , Stephen G. Eubank †
ATR 音声翻訳通信研究所‡
e-mail: {kashioka,black,eubank}@itl.atr.co.jp

1 はじめに

現在，かなり精度のよいTaggingシステム [1, 2] が報告されている．多様な分野で Taggingを用いる場合，分野の変更による未知語の処理や，使われる品詞体系の変更により辞書のメンテナンスが問題となる．本稿では，英語を対象として，品詞付与のための知識を辞書を用いずにコーパスから得るTagging 手法を提案する．

本手法は，決定木学習アルゴリズムにより，コーパスから得られる知識を用いて確率付決定木 [3] を構成し，Taggingを行なう．確率付決定木で用いられる属性は，言語学的な特徴やコーパスから得られる統計的な特徴を用いる．実験システムにおける複数の品詞体系での実験結果について報告するとともに，日本語を対象とした場合の課題についても考察する．

2 確率決定木による Tagging

従来の Taggingでは，辞書を引くことで品詞候補を制限し，その中から，前後に現れる語との関係などを考慮して，もっとも適切な品詞を選択するという手法が一般的である．しかし，辞書の作成や保守にかかるコストの問題となる．また，辞書項目に無い語 (未知語) や辞書の品詞候補にない品詞として使われた語に対しては，特別な処理が必要とされる．

本稿で提案する確率付決定木を用いた手法では，単語の品詞を決定するために，辞書を用いないため，辞書の作成や保守にかかるコストは問題にならない．確率付決定木を，品詞付与済みテキストを用いた学習により構築する．そのために，品詞付与済みテキストがあれば，品詞体系に柔軟に対応できる．また，確率を用いて，品詞列の優先順位を自動的に決定することができる．

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る．Taggingにおいては，対象が各単語に，クラスが品詞に相当する．属性としては，各単語の綴の特徴や文内の使われ方による特徴，単語の相互情報量を用いた階層的分類などを用いる．

以下に，決定木を構築する決定木学習アルゴリズム，および，品詞付与アルゴリズムについて述べる．

決定木学習アルゴリズム

決定木学習では，各属性の有効性を他の属性と独立に計算し，クラスの決定のための効率的な属性による分類順序を木構造として構築する．属性の有効性は，その属性による分類後のエントロピーにより評価する．有効性の計算のために，学習データから各語について“属性とその属性値，品詞”の組からなる情報 (event) をとりだしておく．

具体的には，全ての event の集合に対して，分類後のエントロピーが最小となる属性を求め，最初のノードに割り当てる．この属性の属性値により，event の集合を分割し，対応する子ノードを作る．各々の子ノードにおいて，同様の処理を繰り返し行なうことにより，木構造を構築する．分割の停止条件は，各ノードに含まれる event 数が一定数以下，あるいは分割による有効性が一定基準以下[1]とする．ここで，分割されないノードを *leaf* とよぶ．学習された決定木の *leaf* では，与えられた event の集合から各品詞の確率を計算する[2]．

実際のシステムでは，上に述べたアルゴリズムにしたがって，2段階の決定木を作成している．1段目は，粗く分類した品詞 (以後 GPOS とする) [3]のための決定木であり，2段目として，GPOS の品詞毎に実際の品詞を決定するための決定木を作成する．

品詞付与アルゴリズム

Taggingは，入力文を左から右に処理し，結合確率を最大にする品詞列を出力する．入力文が，$w_1\ w_2\ ...\ w_N$ のような N 個の単語からなり，品詞列 $\{\hat{t}_1, \hat{t}_2, ..., \hat{t}_N\}$ [4]が得られたとすると，結合確率は以下のようになる[5]．

$$P \equiv p(\hat{t}_1, \hat{t}_2, ...\hat{t}_N | w_1, w_2, ...w_N) \quad (1)$$

$$= \prod_{i=1}^{N} p(\hat{t}_i | w_1, ..., w_N, \hat{t}_1, ..., \hat{t}_{i-1}) \quad (2)$$

そこで，文脈に依存する属性をもちいて，決定木の $leaf(L)$ を導き，L に関連した確率分布を，$\hat{p}_L$ により表現し，決定木の条件付分布を用いて以下のように近似する．

$$L_i \equiv \text{文脈} w_1, ..., w_N, \hat{t}_1, ..., \hat{t}_{i-1} \text{において導かれた} leaf \quad (3)$$

$$p(\hat{t}_i | w_1, ..., w_N, \hat{t}_1, ..., \hat{t}_{i-1}) \approx \hat{p}_{L_i}(\hat{t}_i) \quad (4)$$

つまり，最大化すべき結合確率は以下のようになる．

$$P \approx \prod_{i=1}^{N} \hat{p}_{L_i}(\hat{t}_i) \quad (5)$$

各語の品詞付与では，2段階の処理を行なっている．

1. GPOS の各品詞の確率を計算する．
2. GPOS の各品詞に対応する決定木を用いて，品詞の確率を計算する．

各語の確率の計算では，それまでに得られている可能性のある品詞列を全て考慮する必要がある．細かな品詞体系を扱う場合，探索範囲が膨大になるため，本システムでは，stack decoder アルゴリズム [4, 5] を用いて，確率が最大となる品詞列を探索している．

---

*Probabilistic Decision-Tree Tagging Without A Dictionary
†Hideki Kashioka, Ezra W. Black, Stephen G. Eubank
‡ATR Interpreting Telecommunications Research Labs.

[1]分類語のエントロピーと現状のエントロピーとの差がある一定量を越えない場合
[2]実際のシステムでは，スムージングを行なっている．
[3]実際の品詞の属性の一つに対応している
[4]$\hat{t}_i$ は，i 番目の単語の品詞
[5]本手法では，品詞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，文の最初の語とその品詞に依存して最後の単語の品詞を導くことが，原理的には可能である．

3 実験

本実験では，異なる品詞体系による差異，学習データのとり方による差異を比較する．異なる品詞体系として，表1の3種類の体系[6]で実験を行なった．学習データのとり方として，文単位でランダムに集めた場合と，文章単位でランダムに集めた場合の実験を行なった．

品詞体系	品詞総数
UPENN	48
ATR Syntax	441
ATR Full	2600

表 1: 取り扱った品詞体系

UPENN のデータは，Wall-Street-Journal を対象として，100万語強の学習データ，5万語のテストデータを用いた．ATR {Syntax, Full} のデータは，文単位では，40万語の学習データ，1万語のテストデータを，文章単位では，約30万語の学習データ，6万語のテストデータを用いた．表2に，実験結果を示す．

品詞体系	選択	ALL	KW			UW
			ALL	KT	UT	
UPenn	文	96.0	96.7	99.6	61.0	91.9
ATR Syntax	文	92.6	94.7	95.2	52.2	82.9
	文章	90.8	93.8	94.6	41.2	79.6
ATR Full	文	76.5	79.4	83.6	8.5	63.7
	文章	71.8	76.8	81.7	8.2	53.9

KW：既知語，UW：未知語，KT：既知の品詞，UT：未知の品詞

表 2: 本手法による正答率

表2では，学習データに現れた単語を KW，現れなかった単語を UW として別々に精度を調べた．また，KW の中では，その単語と正解の品詞の組合せが，学習データ上に現れている場合を KT，現れていない場合を UT としている．

また，精度を図る基準として，各語について最も頻度の高い品詞を付与した場合の精度を表3に示す．

品詞体系	学習データの選択	正答率 %	perplexity	
			学習データ	テストデータ
UPenn	文	89.6	1.18	1.16
ATR Syntax	文	82.4	1.30	1.19
	文章	83.6	1.30	1.26
ATR Full	文	69.3	1.73	1.36
	文章	69.3	1.72	1.57

表 3: 基準となる精度

いずれの結果も，全体では，基準となる精度を越えている．また，語と品詞の未知の組合せのものに関しては，良い精度が出ていない．語と品詞の未知の組合せに対する精度の向上が望まれる．そのために，誤ったテストデータを調べ，新たな属性を加えるという手法がある．また，学習データの増加も考えられる．

ATR {Syntax, Full} の場合，文を単位とした場合より，文章を単位とした場合のほうが正答率が下がっている．これは，学習量の差とも考えられるが，文章としてまとまった特徴 (使用する単語の分布や，言いまわしなど) があり，文単位で学習データを選択した場合には影響のない，学習データとテストデータ間の特徴の差が影響しているとも考えられる．

これらの実験に関しては，学習データの量が問題となる．どれだけの学習データが必要となるかは重要な問題であり，品詞の細かさなどとの関連を考慮して，明確にすべき点である．

4 考察

現在のシステムでは，単語の分割が済まされたものとして，処理をしている．そのために，日本語に適用するには，単語分割モジュールをシステムに組み込む必要がある．それにともない，複数の分割候補に対して優先度を考慮する必要がある．品詞体系は，柔軟に対応できる．ただし，学習データとして品詞付与されたデータが必要となる．現在，どの程度の量の学習データが必要かということに関しては，明確な答がない．今後，日本語への適用を試みるとともに，品詞の細かさや学習データ量と精度の関係を，実験を通じて議論していきたい．

5 おわりに

本稿では，確率付決定木を用いた Tagging 手法について述べた．本手法では，必要な知識を学習データから取り出すために，辞書の保守や品詞体系に捕らわれることなく利用することができる．その例として，3種類の品詞体系での実験結果を示し，語と品詞の未知の組合せの精度に問題が残るが，他の部分では，基準とした精度を上回っていることを示した．今後，品詞体系，学習量，精度の関係を明確にすることが必要であり，日本語への適用を試みる．

参考文献

[1] E. Brill. "Some Advances in Transformation-Based Part of Speech Tagging." *Proc. of the Twelfth National Conference on Artificial Intelligence*, pages 722-727, AAAI, 1994.

[2] B. Merialdo. Tagging English Text with a Probabilistic Model. *Computational Linguistics*, 20.2:155-171. 1994.

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

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

[5] 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.

[6] UPenn は，ペンシルベニア大学の提供している TreeBank のデータで用いられている体系，ATR Full は，ATR で作成した TreeBank で，意味カテゴリを含む品詞体系を用いており，ATR Syntax は，その意味カテゴリを削除したサブセットの品詞体系である．