自然言語処理 113－2
音声言語情報処理 11－11
（1996. 5. 28）

曖昧観測シンボル列からのHMMパラメータ推定法とその形態素解析への応用

山本幹雄
筑波大学 電子・情報工学系

曖昧な観測シンボル系列から確率モデルを推定する手法を検討した。我々の提案している推定手法は、N-gramだけでなく、HMM(Hidden Markov Model)も推定可能である。手法を評価するために、タグなしコーパスと辞書だけから日本語の品詞列あるいは単語列の確率モデルを推定し、日本語の形態素解析システムに応用した。タグなしコーパスから確率モデルを推定する場合、ノイズが大きな問題となるが、本報告ではさらに信頼性係数を導入し、ノイズによる影響の軽減を試みた。形態素解析による実験結果から、HMMと信頼性係数の有効性を確認した。モデルの違いによる性能の差は、ノイズが大きい場合はbigramなどの単純なモデルが良く、信頼性係数によってノイズの影響がある程度押えられるとHMMが良い性能を示した。しかし、信頼性係数を最大限に利用した場合は、確率モデルの違いによる性能の差を、今回の実験では示すことはできなかった。これは、ノイズによる手法の限界が、確率モデルの性能を下回ってしまったためと考えられる。

A Re-stimation Method of HMM Parameters from Ambiguous Observation and Its Application to Morphological Analyzer

Mikio Yamamoto
University of Tsukuba

This paper describes a reestimation method for stochastic language models such as the N-gram model and the HMM (Hidden Markov Model) from ambiguous observations. It is applied to the model estimation for a tagger from untagged corpus. We also introduce credit factors of training data to improve the reliability of the estimated models. In experiments, we show that the extended algorithm can estimate the HMM as well as the N-gram model from an untagged unsegmented Japanese corpus and the credit factor is effective for improving the model accuracy. However, the HMM estimated using the credit factor isn't better than the N-gram model using the credit factor.

1．はじめに

確率モデルに基づく手法は、一般に、精度が高く、かつタグの付け方が一貫したタグ付きコーパスを大量に必要とする。しかし、そのようなコーパスを大量に作成することは困難である。また、対象とするシステムのタグ体系がコーパスのタグ体系と一致していない場合も問題となる。これらの問題はタグなしコーパスからの推定が可能であれば、ある程度解消される可能性がある。

本報告では、タグなしコーパスと辞書のみから日本語の確率モデルに基づく形態素解析システムのパラメータを推定する手法を検討する。特に以下のことを検討した。

(1) タグなしコーパスから推定した各種確率モデルの比較(bigram, trigram, HMM)
(2) タグなしコーパスを利用する場合のノイズの問題を軽減する信頼性係数の検討

以下、2節で従来のタグなしコーパスからの確

率モデルの推定手法を検討し、信頼性係数の導入を行う。3節では、信頼性係数を考慮したHMMの推定手法を説明する。これは、N-gram の推定を特殊な場合として含んでいる。4節では、実際の信頼性係数を付与する方法を説明し、5節で実験の結果を述べる。

2．従来の推定手法と問題点

2.1 従来手法

タグなしコーパスからの形態素解析用確率モデルの推定法は、[Kupiec92]によって提案された方法が基本である。これは、HMMの観測シンボルを単語とし、状態が品詞を表現していると見なし、Baum-Welch のパラメータ推定アルゴリズム[中川 88 など]を用いて、結果的に品詞の bigram と品詞で条件付けられた単語の確率を得る方法である。まず、英語へ適用され、その有用性が実証されている[Cutting92]。また、[竹内 95]では、Cutting らの方法を単語境界が分からない日本語へ適用できるように拡張し、日本語の形態素解析で成果を上げている。また、[Nagata 96]も、日本語のような単語境界が分からない言語における、タグなしコーパスから N-gram を推定する手法を提案している。[Merialdo94]は Kupiec の方法を trigram に拡張した実験を報告している。

以上述べた方法は、すべて N-gram をタグなしコーパスから推定する手法であるが、我々は N-gram ではなく HMM を推定する手法をすでに提案し、その有効性を示している[山本 96]。HMMの状態はシンボル系列のクラスターを表現しているため、可変長の N-gram と考えることができる。また、N-gram は HMM の状態が特定の品詞、あるいは品詞系列に決定されていると考えることができるので、N-gram の推定は HMM を推定する手法の特殊な場合と考えることができる。

2.2 信頼性係数

[Merialdo94]と[Elworth94]は、タグなしコーパスからの形態素解析用の確率モデル推定において、訓練データに対する尤度の改善が、必ずしも形態素解析精度の改善を意味しないことを実験的に示した。

最大の問題は、タグなしコーパスから抽出した訓練用データ中のノイズがかなり大きいことである。正解の候補を含むためにはそれよりもはるかに多い誤りを含んだ候補も訓練データとして用いなければならない。正解が与えられていない以上、ノイズをなくすることは本質的に不可能である。

さらに、もう一つの問題は、前節で述べた推定手法が、ノイズを含んだデータの確率を等しい重みで上げるという意味になっていることである。このため、正解よりもはるかにノイズの方が多ければ、正解だけに対する尤度を上げることが困難になっていると想像される。しかし、タグなしコーパスからのパラメータ推定は、ある程度の性能の形態素解析システムがあり、その性能を向上させる場合に多く使われる。この場合、対象となっている形態素解析システムを使って候補に信頼性を付与することができる。

本報告では、前節の推定手法に加え、訓練データの信頼性という観点からの改良を試みた。提案手法は、ラティスで表現された訓練データ（形態素の候補を表わす）の各枝にその形態素の正しさの可能性を表わす信頼性係数を付与し、確率的カウントを計算するときにパスの重みとして乗算する。このようにして求められた確率に基づく最尤推定を行えば、曖昧なデータ中の信頼性に応じて重み付けをした尤度を最大化していることになる。

次節では、単語境界が分からない言語において、信頼性付きの曖昧観測シンボル系列からのHMMパラメータ再推定手法を述べる。信頼性がない場合の推定手法とほぼ同じであるため、詳しくは[山本 96]を参照のこと。

3．信頼性係数付きの曖昧観測シンボルからのHMMパラメータ推定手法

3.1 HMMによる形態素解析

日本語における形態素解析では、単語分割と品詞の付与を同時に行わなければならない。ここで、単語の系列を $W=w_1,\ldots,w_n$ 品詞の系列を $T=t_1,\ldots,t_n$ とすると、確率を使った形態素解析は単語列と品

詞列の同時確率 P(W,T)を最大化する問題に帰着される[永田 94]。

$$(\hat{W},\hat{T}) = \arg\max_{W,T} p(W,T \mid S) = \arg\max_{W,T} p(W,T) \quad (1)$$

p(W,T)をHMMでモデル化すると、(2)式、および(2')式のようになる。

$$p(W,T) = \sum_{x} \prod_{i=0}^{n-1} a_{x(i),x(i+1)} b_{x(i+1)}(w_{i+1}, t_{i+1}) \quad (2)$$

$$= \sum_{x} \prod_{i=0}^{n-1} a_{x(i),x(i+1)} b'_{x(i+1)}(t_{i+1}) p(w_{i+1} \mid t_{i+1}) \quad (2')$$

ここで、xはHMMの可能な状態間のパスを表わし、$x(i)$はあるパス上のi番めに遷移する状態である。$a_{x(i),x(i+1)}$は状態$x(i)$から状態$x(i+1)$への遷移確率であるが、特に$a_{x(0),x(1)}$は状態$x(1)$の初期状態確率$\pi_{x(1)}$を表わす。$b_{x(i)}(w,t)$、$b'_{x(i)}(t)$は状態$x(i)$で品詞がtの単語w、あるいは品詞tを出力する確率である。(2)式は出力シンボルとして単語と品詞のペアを考えた場合、(2')はHMMの出力シンボルは品詞として、$p(w|t)$を乗ずることによってペアの出力確率を近似したものである。

次節では、(2)式のHMMのパラメータをタグなしコーパスから推定する方法を述べる。

3.2 形態素ネットワーク

タグなしデータが与えられると、まず辞書引きにより可能な形態素のネットワークを生成する。得られた形態素ネットワークを以下のように定義する。

m_s：s番目の形態素（番号sを持った形態素）
w_sまたは$word(s)$：s番目の形態素の単語（見出し）
t_sまたは$tag(s)$：s番目の形態素のタグ（品詞）
$suc(s)$：形態素ネットワーク上で、m_sの後(文上で右)に接続している形態素の番号の集合
$pre(s)$：形態素ネットワーク上で、m_sの前(文上で左)に接続している形態素の番号の集合
credit(r,s)：r番目の形態素の後にs番目の形態素が接続する信頼性係数
#：文頭を表わす記号

3.3 HMMパラメータの推定法

観測シンボル列が、形態素ネットワークとして曖昧に与えられた場合のHMMの状態遷移確率、出力確率、初期状態確率(それぞれ、$\boldsymbol{a},\boldsymbol{b},\boldsymbol{\pi}$)の再推定式を考える。基本的な考え方は、形態素ネットワークに対応したネットワーク状になったトレリスを考え、その上で従来の再推定式を拡張すればよい。すなわち、従来の前向き・後ろ向き確率が時間(位置)同期的に定義されていたものをネットワーク的に拡張し、形態素ネットワーク上の各形態素ごとに前向き・後ろ向き確率が定義される。

各形態素における前向き・後ろ向き確率は、初期値として文頭と文末の形態素uとvに関して、$\alpha_u(j) = \pi_j b_j(w_u, t_u) credit(\#, u)$、$\beta_v(i) = 1$を与えれば、次のように再帰的に定義される。

$$\alpha_r(j) = \sum_{s \in pre(r)} \sum_{i=1}^{N} \alpha_s(i) a_{ij} b_j(w_r, t_r) credit(s,r) \quad (3)$$

$$\beta_s(i) = \sum_{r \in suc(s)} \sum_{j=1}^{N} a_{ij} b_j(w_r, t_r) \beta_r(j) credit(s,r) \quad (4)$$

(3),(4) 式で求められた前向き・後ろ向き確率を用いて、HMMの各パラメータは以下のように再推定される。ここで、kは訓練データとしての文の番号、p_kはその正規化係数としての生起確率である。各パスに信頼性係数が係っているので、正規化係数としての生起確率も信頼性係数が係ったものでなければならない。

$$\bar{a}_{ij} = \frac{\sum_{k=1}^{K} \frac{1}{p_k} \sum_{r \in pre(s)} \alpha_r^k(i) a_{ij} b_j(t_s) \beta_s^k(j) credit(r,s)}{\sum_{k=1}^{K} \frac{1}{p_k} \sum_{s} \alpha_s^k(i) \beta_s^k(i)} \quad (5)$$

$$\bar{b}_i(w,t) = \frac{\sum_{k=1}^{K} \frac{1}{p_k} \sum_{word(s)=w, tag(s)=t} \alpha_s^k(i) \beta_s^k(i)}{\sum_{k=1}^{K} \frac{1}{p_k} \sum_{s} \alpha_s^k(i) \beta_s^k(i)} \quad (6)$$

$$\bar{\pi}_i = \frac{\sum_{k=1}^{K} \frac{1}{p_k} \sum_{s \in on(1)} \alpha_s^k(i) \beta_s^k(i)}{\sum_{k=1}^{K} \frac{1}{p_k} \sum_{s \in on(1)} \sum_{j=1}^{N} \alpha_s^k(j) \beta_s^k(j)} \quad (7)$$

前向き・後ろ向き確率を計算する場合に、たまにアンダーフローを生じる場合がある。その場合は、形態素ネットワーク上に同期点という概念を導入することにより、スケーリングが可能となり、問題を避けることができる。また、一般にスケーリングすることによって、正規化も同時に行える

ため都合がよい。詳しくは[山本96]を参照されたい。

4．信頼性係数の付与手法

本実験では形態素ネットワークの生成のためにJumanを用いているため、Jumanのコストという概念に基づいた信頼性係数の付与の方法を試みた。Jumanは形態素解析中に、形態素候補の悪さの可能性を表わすコストによって、候補の枝刈りを行いつつ形態素解析を行う。形態素ネットワークのある点に接続できる形態素候補は、その中で最良の形態素のコストからの差がある与えられたコスト幅以上であれば、枝刈りされる。Jumanに大きな枝刈り用のコスト幅を与えると、大きな形態素ネットワークを出力する。また、小さい値を与えると、Jumanのコストに基づいた正しそうな形態素による小さな形態素ネットワークを出力する。

信頼性係数の決定は、あるテスト文に対する、あるコストを持つ候補の適合率とした。ただし、すべてのコストについて精度を調べるのは困難であるため、あるコストの範囲における適合率を調査した。例えば、コストが70-100の場合の信頼性係数は、コスト幅100で生成される形態素候補の中から、コスト幅70で生成される候補を削除し、残った候補の適合率である。適合率は、適合率の逆数だけ候補が集まると1個正しいことを意味している。表1に各コストにおける適合率を示す。実際の実験では、1/0.5c+1.19という式で近似した。ここで、cはコストである。

表1 Jumanのコストと適合率

cost	0	1-10	11-20	21-50	51-100	101-200	201-500
適合率	0.84	0.16	0.13	0.069	0.074	0.008	0.002

5．実験

5.1 実験システム・データ・評価方法

実験では、形態素ネットワークを13万単語の辞書を持つJuman[松本 94]で生成し、3節で述べた再推定式を用いてパラメータの推定を行った。また、形態素解析を行うときも、同様にJumanによってまず形態素ネットワークを生成し、その中で最も確率の高いパスをViterbiアルゴリズムによって決定し、形態素列を出力する。

データとしては、日経新聞94年版[日経 94]を用いた。学習用として、10日分の記事26108文(長さ150文字以上の文は削除した結果)、テスト用として、学習用以外の1日分の記事から100文をランダムに取り出した。テスト用は人手によってタグ付けを行った(形態素数約2500)。モデルの品詞としては、Jumanの出力する品詞、品詞細分類、活用型、活用形の組み合わせを104種類に分類した(264種類でも実験したが、bigramとHMMの実験で104種類の方がよい結果が得られたので省略する)。正解は(上記組み合わせとしての)品詞、見出し、基本形が一致した場合とした。ただし、人手で正解を作成するとき、どちらの品詞にする

表2　各確率モデルによる形態素解析の適合率

モデル	信頼性	繰り返し回数 (ピーク/飽和)	コスト幅 0	10	20	50	100	200	500	1000
品詞bigram	なし	10(ピーク)	92.4	92.5	92.4	90.6	90.8	90.7	90.5	90.5
	step	12(飽和)	92.6	92.7	92.6	91.0	91.1	91.0	91.0	91.0
	あり	19(飽和)	92.6	93.0	93.1	93.4	93.5	93.4	93.3	93.3
品詞trigram	なし	13(ピーク)	92.6	91.2	91.2	88.6	88.3	87.6	87.3	87.3
	step	17(飽和)	92.5	91.1	91.1	88.7	88.5	87.8	87.7	87.7
	あり	24(ピーク)	92.7	92.9	92.9	93.3	93.0	92.5	92.3	92.3
品詞HMM	なし	4(ピーク)	92.7	92.4	92.2	91.3	90.7	89.9	86.5	86.5
	step	10(ピーク)	92.8	92.6	92.4	92.4	92.5	92.3	92.1	92.1
	あり	3(ピーク)	92.4	93.1	92.9	93.0	93.0	92.8	92.6	92.6
単語HMM	なし	5(ピーク)	92.8	93.1	93.0	92.9	91.1	88.0	81.4	81.4
	step	3(ピーク)	92.8	93.2	93.0	93.2	93.0	92.9	92.7	92.7
	あり	3(ピーク)	92.8	93.4	93.3	93.5	93.4	93.3	93.1	93.1

べきか悩んだものはどれでも正解ということにした。また、固有名詞で未定義語となるものが多かったが、これらはJumanが未定義語として出力するデフォルトの品詞（本実験ではサ変名詞）でも正解とした。適合率の計算は、「正解出力の数」／「テスト文の全(正解)形態素数」[永田 94]で求めた。

Jumanが出力する訓練用の形態素ネットワークは3種類を用いた。出力する候補の大きさを決定するコスト幅を500と70にしたものと、コスト幅70のものに4節で述べた信頼性係数を付与したものである。コスト幅500では、可能な候補のほとんどすべてが出力されるので、全く信頼性係数を使用していないものに対応する。コスト幅70のものは、コスト70以上の候補に0、70以下のものに1の信頼性係数を付けたものと同じである。これをstep関数型の信頼性係数と呼ぶことにする。

初期モデルとしてすべてのパラメータを等確率としたものを用いた。

5.2 実験結果

表2に各推定モデルによる形態素解析の精度を示す。信頼性の列は前節で述べた、訓練用のデータの3種類に対応している。「なし」がコスト幅500、「setp」がコスト幅70で出力した場合の形態素ネットワーク、「あり」はコスト幅70で出力した形態素ネットワークに4節で述べた信頼性係数を付与したものである。繰り返し回数は、再推定の繰り返し回数である。適合率が最大となった繰り返し回数と、そのときの適合率を右に示した。また、繰り返し回数欄の括弧の中の「ピーク」と「飽和」は、その回数でピークに達して、その後適合率が下がったことを、またはそれ以降、適合率が変化しなくなったことを、それぞれ表わす。N-gramで信頼性係数を使用した場合、飽和している場合が多いことが分かる。Merialdoらの実験では一般に適合率は下がることが示されているので、飽和したということは信頼性係数の有効性を意味している。

訓練用データの違いによって、3つに分けたグラフを図1〜3に示す。表2の信頼性の列の種類

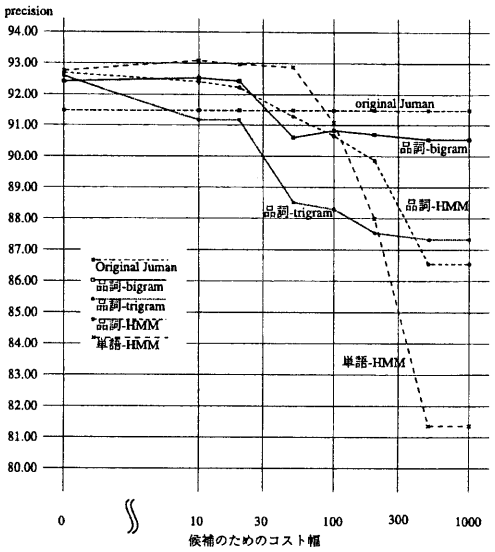

図1 信頼性係数なしの場合の結果

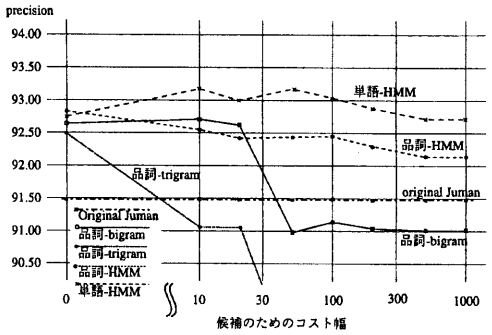

図2 step型信頼性係数の場合の結果

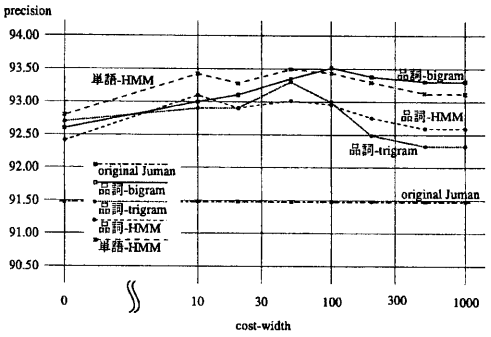

図3 信頼性係数付きの場合の結果

に従って、図1は「なし」、図2は「step」、図3が「あり」に対応する。縦軸は適合率、横軸はJumanが出力する候補としての形態素ネットワークの大きさを決定するコスト幅の値である。コスト幅が大きいほど、大きな形態素ネットワークが出力される。大きくなると正解が入っている可能性も大きくなるが、誤る可能性も大きくなる。この中から確率モデルを用いて正解を選択し、出力する。図中Jumanと書いてある直線は確率モデルを用いないオリジナルのJumanのテスト文に対する適合率である。

図1より、信頼性係数がない場合は、あまりよい結果となっていないことが分かる。この場合は、モデル化能力が一番弱いbigramが一番よい結果となっている。これは、能力の高いモデルはノイズも一緒にモデル化してしまうため、ノイズに弱いと考えられる。

図2より、ある程度信頼性の高い結果だけを使った場合、モデル化能力の高いHMMなどがよくなることが分かる。

図3では、各モデルに関して最もよい結果となり、信頼性係数の有効性を示していると言えるが、最もモデル化能力の弱いbigramがかなりよくなり、他のモデルと同等となっている。これは、HMMやtrigramに対しては訓練データが少なすぎるということも考えられる。しかし、trigarmとHMMに関して、約45万文を用いた実験を行ったが改善は見られなかった。残る可能性は、今回の信頼性係数を用いたタグなしコーパスからの推定の限界である。今回の信頼性係数を用いたデータからは93%前後の精度を得るのが限界で、すべてのモデルがそのレベルまで達したと解釈できる。また、HMMよりも一般に能力は高いと思われるtrigramが実験全体を通してよい結果を得ることができなかった。検討は今後の課題である。

6．おわりに

タグなしコーパスと辞書のみから各種確率モデルを推定し、形態素解析で評価した。また、信頼性係数を導入し、信頼性係数がタグなしコーパスを使用する場合に有効であることを示した。

今後の課題としては、形態素解析の実験が比較的低いレベルでの実験であったため、高いレベル(適合率95%前後)でも同じように形態素解析システムの性能を改善できるかどうかを調べなければならない。また、タグなしコーパスから推定されたモデルとタグ付きコーパスから推定されたモデルの融合を検討したい。

謝辞

Jumanを開発された京都大学 長尾研究室、奈良先端科学技術大学院大学 松本研究室の皆様、日頃議論していただく、筑波大学 知能情報・生体工学研究室の皆様、豊橋技術科学大学 中川研究室の皆様に感謝いたします。また、データ使用を許可していただいた、日本経済新聞社に感謝いたします。

参考文献

[Cutting92] D. Cutting, J. Kupiec, J. Pedersen and P. Sibun: A practical part-of-speech tagger, ANLP-92,pp.133-140, 1992.

[Kupiec92] J. Kupiec: Robust part-of-speech tagging using a hidden Markov model, Computer Speech and Language, Vol.6, pp.225-242, 1992.

[Merialdo94] B. Merialdo: Tagging English text with a probabilistic model, Computational Linguistics, Vol.20, No.2, pp.155-171, 1994.

[Nagata 96] Masaaki Nagata: Automatic extraction of new words from Japanese texts using generalized forward-backward search, Conference on Empirical Methods in NLP, 1996 (to appear).

[竹内 95]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」、情報処理学会研究会報告, 自然言語処理研究会, NL-108-3, pp.13-19, 1995.7.

[中川 88]中川聖一: 確率モデルによる音声認識、電子情報通信学会, 1988.

[永田 94]永田昌明: 前向きDP後ろ向き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, 情報処理学会研究会報告, 自然言語処理研究会, NL-101-10, pp.73-80, 1994.5.

[松本 94] 松本祐治、他：「日本語形態素解析システムJUMAN使用説明書 version2.0」, 1994.

[日経 94] 日本経済新聞CD-ROM版(1994年版)、日本経済新聞社、1995.

[山本 96] 山本幹雄: Untagged-corpusを用いた形態素解析用HMMパラメータの一推定法、言語処理学会第2回年次大会発表論文集, pp.61-64, 1996.3.