自然言語処理 108－3
(1995. 7. 20)

HMMによる日本語形態素解析システムのパラメータ学習

竹内孔一　松本裕治

奈良先端科学技術大学院大学 情報科学研究科
〒630-01 奈良県生駒市高山町 8916-5
E-mail : {kouit-t,matsu}@is.aist-nara.ac.jp

あらまし

日本語の形態素解析は自然言語処理を行なう上で最も基本的でかつ重要な処理である。我々の研究室で開発している形態素解析システムJUMANは、品詞の連接と単語に対してコストによる制約を与えることで曖昧性の絞り込みを行なっている。コスト値はJUMANの品質に大きな影響を与えるにも関わらず、人手で付与されるため最適化する機構は存在しなかった。そこで、本研究では、英語のタグづけなどで効果を発揮しているHMM（Hidden Markov Model）を用いて、コーパスによる学習を行なうことでJUMANのコスト値、すなわちパラメータの最適化を行なう。HMM学習の結果、現在のJUMANの解析精度を改善する結果が得られた。

キーワード　隠れマルコフモデル、学習、日本語形態素解析システムJUMAN

Learning Parameters of Japanease Morphological Analyzer based-on Hidden Markov Model

Kouichi Takeuchi　Yuji Matsumoto

Graduate School of Information Science,
Nara Institute of Science and Technology
8916-5, Takayama, Ikoma, Nara, 630-01, JAPAN
E-mail : {kouit-t,matsu}@is.aist-nara.ac.jp

Abstract

Morphological analysis is the first step toward the analysis of Japanese texts and one of the most important processes. So far, we have been developing the Japanease morphological analyzer JUMAN as a public-domain system. In JUMAN, ambiguities of morphological analysis are reduced by means of costs manually attached to the connectivity rules and words. The performance of JUMAN largely depends on those manually attached costs, while at present JUMAN has no facility to optimize the costs. This paper proposes a method for optimizing the costs (i.e. parameters) to be attached to the connectivity rules and words. The proposed method is based on hidden Markov model, which has proved effective in parameter estimation of English part-of-speech tagging. The result of experiments shows that the proposed optimization method improves the manually attached parameters.

key words　Hidden Markov Model, Learning, JUMAN

1 はじめに

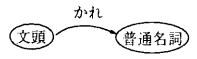
日本語の形態素解析は計算機処理を行なう上で最も基本的でかつ重要な処理である。我々の研究室で開発している形態素解析システムJUMAN[1]は品詞の連接と単語に対してコストによる制約を与えることで曖昧性の絞り込みを行なっている。このコスト値はJUMANの解析精度に大きな影響を与えるにも関わらず、連接規則とともにユーザーが値を決定して用いなくてはならないため目的とするコーパスに対してコスト値を人手で作成する必要があった。そこで、本研究では、英語のタグづけなどで効果を発揮しているHMM（hidden Markov model）を用いて、JUMANシステムに対応するHMMシステムを構築して、コーパスによる学習を行なうことでコスト値、すなわちパラメータの最適化を行なうことを目標とする。

このような機械に学習させる手法を用いたタグづけは英語で良く用いられており、特にHMMを用いた学習を行なわせた場合96%を越える高い結果が得られている[2]。しかし、日本語では語順の自由度が大きく、また、わかち書きがされていないため、問題は複雑である。Constantine[3]は解析済みコーパスとHMMを用いてJUMANの日本語のわかち書き修正を試みた。永田は日本語の形態素解析において確率的な言語モデルを用いてATRの解析済みコーパスを利用した学習を行なった[4][5]。学習は解析済みコーパスのみで行なうため、このコーパスが無い分野では学習することができない。また、解析済みコーパスが存在しても、日本語では、文法が統一されていないためそれを有効に利用できるとは限らない。中国語でも日本語と同様なわかち書きの問題がおこるがChang[6]らは中国語のタグづけにおいて前処理として単語のわかち書きを行なって9割を越す精度を得ている。我々が以下で提案するJUMAN-HMMシステムは初期値の獲得手段として少量のタグつきコーパスを用いるが学習にはタグなしのコーパスを学習することにより精度を向上させる。そのため、大規模なタグつきコーパスが存在しない分野に対しても学習を行なわせることができる。本報告ではHMMのモデルの確定を行なうために新聞記事に対して実験を行なった。

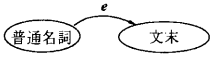
2 学習システム

2.1 JUMANのコスト計算について

JUMANは入力された文字列に対して可能な辞書引きを行なった後、すでに与えられている連接規則のコスト値と単語のコスト値の和によってその文章のコスト値が決まる。図1は例文「かれ」のコスト値を計算している。ここで、「文頭」及び「文末」という特別な品詞を用意している。また、文末に遷移する場合は空語「e」で遷移する。図より例文コスト値は140となる。



文頭 → 普通名詞		コスト10
普通名詞	かれ	コスト100
合計		コスト110



普通名詞 → 文末		コスト30
文末	e	コスト0
合計		コスト30

図1 JUMANのコスト計算

この図の例では解析結果は1つのパスを示しているが、JUMANにはコスト幅という概念があり、（最小のコスト値パス）＋（コスト幅）の範囲に入るパスを全て出力する。よってコスト幅を緩めると「動詞：かる」の命令形や、「か」「れ」のように別の形態素として解釈される結果が出力される。

2.2 HMMの定式化

HMMは英文におけるタグづけにおいてよく用いられ、高い精度の結果が得られているが、日本語文ではわかち書きされていないため、そのまま用いることができない。そこで、ある入力文字列Lから得られる1つの単語列を$w_{1,n} = w_1, w_2, ..., w_n$として、各単語に品詞系列$t = t_1, t_2, ..., t_n$を付与すると考えると、入力列$L$に対する確率は、$w_{1,n} \in L$のなかで可能な組合せを全て足し込めば良いから、

$$
\begin{aligned}
&P(L)\\
&= \sum_{w_{1,n+1}\in L} P(w_{1,n+1})\\
&= \sum_{w_{1,n+1}\in L} \sum_{t_{0,n+1}} P(w_{1,n+1}, t_{0,n+1})\\
&= \sum_{w_{1,n+1}\in L} \sum_{t_{0,n+1}} \prod_{i=1}^{n+1} P(w_i|t_i)P(t_i|t_{i-1}) \quad (1)
\end{aligned}
$$

となる。ここで、t_0は「文頭」、t_{n+1}は「文末」という品詞で、w_{n+1}は文末に遷移するために設けた空語である。これにより、全ての文章は必ず、品詞「文頭」から始まって「文末」で終了するlattice状の構造をとることになる。この構造は前節で説明したJUMANのコスト計算における連接規則のコスト値と単語のコスト値の和に対応している。また最適な単語系列と品詞系列を求めることは、与えられた入力文字列Lに対して、確率$P(w_{1,n}, t_{0,n+1})$を最大化する単語列と品詞列の組合せを求めることになり、JUMANのコスト最小の組合せに対応している。

2.3 学習について

通常のHMMは入力文字列を与えて、あらゆる可能な状態遷移のパスに対して（つまり、状態はhiddenとして）確率を計算する。我々はJUMANのコスト幅を緩めて出現する曖昧性の範囲内での状態遷移のパスについてのみ確率の足し込みを行なう。この制約により計算量が少なくて済むことと、明らかにおかしな解釈を学習の対象外にすることができる。再推定の場合、確率的回数のγをもとにして、品詞接続確率a_{ij}とある品詞jでの単語w_kの出力確率$b_j(w_k)$を再推定する。

$$
\begin{aligned}
&\gamma(i,j,w_k)\\
&= \frac{1}{P(L)} \sum_{w_{1,n}\in L} \sum_{t=0}^{n} \alpha_i(t)a_{ij}b_j(w_k)\beta_j(t+1) \quad (2)
\end{aligned}
$$

$$
a_{ij} = \frac{\sum_{w_k}\gamma(i,j,w_k)}{\sum_{w_k}\sum_j \gamma(i,j,w_k)} \quad (3)
$$

$$
b_j(w_k) = \frac{\sum_i \gamma(i,j,w_k)}{\sum_{w_k}\sum_i \gamma(i,j,w_k)} \quad (4)
$$

ここで、α, βは Cutting[2] らと同様である。$\alpha_i(t)$ の式は1文中のt番目の単語において「文頭」から品詞iまでの確率の総和で、$\beta_j(t+1)$は逆に「文末」から品詞jまでの確率の総和である。上式(2)(3)(4)はテキスト全文について計算し、再推定された確率値をコスト値に変換して図2の中間辞書に与える。これをJUMANの辞書に変換することで学習結果が反映される。

2.4 JUMAN-HMM システム

前節で示した、JUMANとHMMはコストと確率を対応させるとJUMANに対応するHMMが作成できる（表1）。JUMANで行なうコストの足し算は確率のかけ算になるので、確率の対数の絶対値をとった値がコストになる。

全体の構成を説明する。図2にJUMAN-HMMシステムの構成図を示す。この学習システムには大きく分けて2つのモジュールがある。1つは初期値を獲得する右側のモジュールで、他は左側のループでタグなしコーパスを用いて学習を行なうモジュールである。

表1 JUMANとHMMの対応

JUMAN システム	HMM モデル
品詞間の接続に関するコスト	状態間の遷移確率
ある品詞における単語のコスト	ある状態間の遷移においてある単語を出力する確率

まず、最初にHMMもJUMANも初期の連接確率（コスト）ならびに単語の生成確率（コスト）（以降、初期値と呼ぶ）を獲得するため図2右側のモジュールで初期値を獲得する。つまり、何らかの手法で得られたJUMANの文法体系に合致したタグつきコーパスからマルコフ学習によって単純に頻度を数えることで初期値を得る。すなわち（1）式で用いる最初の連接確率$P(t_i|t_{i-1})$と品詞別単語の生成確率$P(w_i|t_i)$を獲得している。得られた確率値をHMMに、コスト値に変換した値をJUMANにそれぞれ送り学習する準備が完了する。

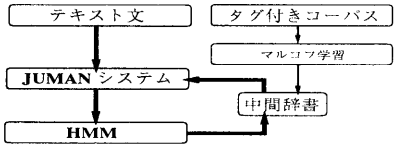


図2 JUMAN-HMM パラメータ学習システム

次に、大量の目的の分野のコーパスを用意してJUMANにある程度のコスト幅を許して曖昧性をふくめた解析を行なわせる。HMMはJUMANの解析結果を受けとり、持っている連接確率、単語の生成確率から出現頻度を計算し、コーパスの解析が終了した時点でそれぞれの確率値を再推定する。これが学習の1サイクルである。後は、再推定された確率値をコスト値に変換してJUMANに与え、同じコーパスに対して学習を繰り返す。値が安定してきたところで学習を打ち切る。このシステムではJUMANは辞書引きの役目だけを行ない、HMMは確率値を独自に保存している。よってJUMANの辞書にコスト値を書き込む場合、整数化による丸め誤差が生じるが、再推定の計算には保持している確率値をもとに計算する。

2.5 HMMノードの作成

JUMANの文法では形態素に対して、品詞、品詞細分類、活用型、活用形および表層の基本形まで指定することができる。原則として、HMMのノードはJUMANの品詞細分類を採用するが次のような例外を設けた。まず、前方の単語が活用する場合、その活用形が次の単語の連接確率に影響するので、同じノードでも前方からの連接確率を計算するときには品詞細分類まで観測し、後方から確率を計算する時には活用形まで観測する。また、助詞、助動詞はそれぞれの語の接続に特徴があるので表層語まで区別して個別のノードとした(図3)。

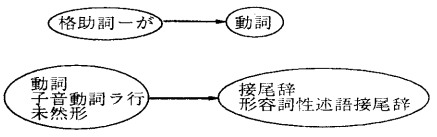


図3 HMMで用いる連接規則の例

2.6 未定義語の扱い

未定義語はJUMANも「サ変名詞」として扱っているので、文全体の確率を計算する時には「サ変名詞」で計算する。しかし、各連接確率を再評価する際に数えないことにする。未知語に対する特別な処理は今のところ特に行なっていない。

2.7 確率の最小値

図2に示した通り、HMMの学習は収束するまで何度か繰り返す。この際、ある連接規則ならびに単語の確率値が学習前に高くても、コーパス中に1回もその規則もしくは単語が出現されなければ原則として確率は0である。しかし、コスト値に変換する必要があることから実際には低い確率値を与えている。

3 実験

以下では2種類の実験を示す。1つは初期値の獲得に関する実験。他の1つはその中で精度の良い初期値をベースに朝日新聞社説9年分(1985～1993年)によるHMM学習を行なわせた。

3.1 初期値の獲得実験

前節でも述べたようにJUMAN-HMMシステムは初期値として連接確率と単語の確率を必要とする。この初期値は後のHMM学習に大きな影響を与えるため、非常に重要である。そこで、以下の4つの方法について獲得実験を行なった。

3.1.1 初期値の獲得方法

初期値の獲得には以下の4つの方法が考えられる。後の3.1.2節ではこの中の最初の3つの方法について初期値の獲得実験を行なった。

a. EDR解析済みコーパスを用いる方法

b. 朝日新聞の社説をJUMANに解析させた結果から初期確率を得る方法

c. 朝日新聞の社説を人手による解析を行なった結果を使用する方法

d. 現在のJUMANの連接コストと単語のコストを直接変換して初期値とする方法

aの場合、ＥＤＲの品詞体系はJUMANの文法体系と異なり、かつ直接変換するのは容易ではないため、わかち書き情報のみを用いて品詞を現在のJUMANに付与させた。そのため、このコーパスはJUMANの品詞付与の際のエラーを含んでいる。bの場合、朝日新聞の社説を現在のJUMANにそのまま解析させ、その結果を用いる。aに比べてわかち書き情報を用いてない分、正確さは劣るが、後の大規模学習に用いるコーパスと同じ社説なので整合性はよい。

cの場合は人手で解析を行なった結果から初期確率を求めた。そのため、量は少なく、300文、約6700形態素である。aからcの方法は、最終的にJUMANの文法体系に即したタグつきコーパスが獲得できているので、2.4節で述べたように、マルコフモデルによる学習を行ない、その結果をHMMに渡す。この場合獲得された連接規則は2.5節で述べた体系に整理されている。

dの場合は上記の３つと異なる。これは現在のJUMANが持っている人手で作成された連接規則とコスト値ならびに単語のコスト値を直接確率値に変換してHMMを作成する。よってこの場合3.1.2節の実験におけるこの初期値によるJUMANの精度は当然、表２の「現在のJUMAN」と等価になる。このモデルによるHMM学習の実験結果は4章で示す。

3.1.2 初期値の獲得結果

前節で示したaからcの方法で初期値を獲得したJUMANの解析精度を評価する。解析対象は朝日新聞の社説で行なう。その解析済み社説は２種類（300文(約6700形態素)と200文(約5000形態素)を用意した。このうち300文は方は前節で述べた方法cで初期値として使用した。後者の200文は初期確率の計算でも、後のHMM学習でも用いていない評価用のデータである。表２の評価ではcの場合の時のみinsideデータ[1]となる。

表２の数字は解析による誤った形態素数の割合を百分率で表している。形態素の正誤の判断は、一意に正解を記述している解析済みコーパスと比較して、品詞細分類又は活用する語は活用形まで観測して１つでも異なれば不正解とする。わかち書きが異なった場合の誤り数は、例えば１つの形態素で正解のところを２つに分割した場合は２として数えている。このため、名詞連続などの分割の違いによる誤りも数えられている。

[1]初期値に含まなかった解析済みコーパスによる評価の結果をoutsideデータと呼ぶことにする。

表２ 各初期値によるJUMANの誤り率

初期値獲得方法	300文(%)	200文(%)
a. ＥＤＲ(１万文)	14.8	16.9
b. 社説(約１万文)	8.1	12.5
c. 解析済み社説	4.9(inside)	8.7
社説(約6.5万文)	7.5	10.2
現在のJUMAN	7.9	9.2

3.1.3 初期値の獲得実験の考察

表２におけるaのＥＤＲ１万文を初期値にしたJUMANの誤り率は同じ文章量であるbの社説に比べて大きい。これは、ＥＤＲの文章は新聞記事と雑誌を含んでいることから、評価に用いた社説とＥＤＲとの間に単語や連接に関する統計的性質の違いが存在して上記のような結果が出現したと考えられる。

cの解析済み社説を用いた場合、単語数が極端に少ないにも関わらず、insideデータである300文の場合はもとより、学習に含んでいない200文の解析結果も良い。初期値として、目的とする分野の解析済みコーパスの威力が強いことがわかる。

以上の初期値実験を踏まえて、分野の異なるＥＤＲの初期値は用いないことにする。よって次の章では初期値の獲得方法b～dを元にHMMの学習実験を行なう。

3.2 HMM学習実験

3.2.1 初期値と学習のコーパス

初期値は前節で述べたb～dの方法を採用する。ただし、方法bにおいては、より初期値を良くするため社説3年分（約6.5万文)を用いた。この場合のJUMANの誤り率は表２の４行目に示している。次に、HMM学習のコーパスには朝日新聞の社説で９年分（約20万文）を採用する。学習の際、JUMANのコスト幅を用いて、あまりに無駄な解釈が出ない程度に制約した。

3.2.2 学習実験の結果および評価

学習結果を表3に示す。まず、解析済み社説を用いた場合の結果について考察する。学習に使用したinsideデータである300文の正解コーパスと比較すると表2の解析結果ではエラー率4.9%であったが学習すると5.4%にすこし誤り率が上がっている。しかしながら、学習に用いなかったoutsideデータを解析すると誤り率10.9%から8.8%に改善されている。HMM学習効果はoutsideデータで観測できたものの、insideでは確認できなかった。この点についてはさらなる考察が必要である。しかし、insideとoutsideの両方の誤り率は現在のJUMANとの誤り率と比べても低い。このことから今回提案したJUMAN-HMMシステムによってJUMANの解析精度の向上に効果があることが明らかになった。

表3 HMM学習後のJUMANの誤り率

初期値	300文(%)	200文(%)
b. 社説3年分	14.0	16.2
c. 解析済みの社説	5.4(inside)	8.2
d. 直接変換	13.4	14.8
JUMANの精度	7.9	9.2

方法dの場合の結果がかんばしくない。このdは現在のJUMANから連接規則とそのコスト値、並びに単語のコスト値を直接変換して初期確率としている。そのため、HMM学習前では現在のJUMANと同程度の解析精度を有していたはずであるが、表3はHMM学習により誤り率が増加したことを示している。これは、現在の人手で作成されたコスト値を変換して得た初期確率値では連接規則ならびに単語に対して第1候補には最大の確率が付与されるが、第2、第3候補についての統計的に合致した順序を保証していない。そのためHMM学習を行なうと第2、第3候補に誤る例が数多く出現してしまうため、正解だった第1候補まで確率値が下がって誤り率が増加した。

次に、bの初期値を社説3年分から獲得した場合のHMM学習結果について考察する。この結果もHMM学習により誤りが増加した。これは、表2の初期状態では誤り率が低いので第1候補の確率は正解にも合致してうまく獲得されていた。しかし、その第2番目の候補、第3番目の候補に現在のJUMANがよく犯す誤りが統計的に内在していたため、HMM学習によってそれらも強化されて誤りを導いてしまったと考えられる。実際、誤った事例を調べると現在のJUMANがよく犯す特徴的な誤りが含まれていた[1]。つまり、bの場合もdの場合もJUMANの持つコスト値の与え方の悪さが影響しており、HMM学習でこの種の誤りが自動的に修正されないことを示している。

以上示したように初期値の違いによってHMM学習の結果は大きく変化することが明らかになった。モデルの設定から考えればHMMはコーパス中で観測されなかった連接規則や単語は確率が最低値のままなので、未知の連接規則ならびに単語の頻度は低いままであるが、間違った初期値をシステムに提示するとそれをもとにHMM学習では増加が始まる。よって表2の初期値獲得の段階で、現在のJUMANに頼るbやdの場合の誤りの内容とcの解析済み社説の場合の誤りの内容には本質的な違いが存在する。つまり、前者では誤った初期値が獲得され、陽に誤っている。しかし、後者では少数の正解のみの初期値しか保持してないため、insideではほぼこのモデルの最適値まで誤りが減少し、outsideでは、初期値に含まれなかった未知の連接規則ならびに単語によって誤りが生じているのである。

まとめると、今回提案したJUMAN-HMMシステムでは少量の正解からスタートして、少量ではあるが正しい初期値が未知の連接規則ならびに単語の頻度に対して正しい方向に引き込む機構となっていることが言える。

4 今後の課題

4.1 モデルのチューニングについて

今回、示している実験結果はわかち書きを取り込んで学習している以外は普通のHMMとほぼ同様である。さらなるモデルの改善が必要である。例えば、HMM学習後の誤りとして「切り離せ(命令形)+ない(形容詞)」といった普通では考えられない連接規則が優先された結果が得られている。さらに、JUMANでは接続の役割を重視して考えられた品詞分類がいくつも存在し、明示してあるものがある。例えば「名詞接頭辞」は名詞

[1] 具体的には「いく」固有名詞 などが出現する

に接続する接頭辞であるが、HMMでは名詞以外の接続についても連接確率を与えてしまう。そこで、JUMANの連接規則には接続しないことを示す負の連接規則を記述することができるので、あらかじめ接続しない品詞の組合せには負の連接を明示的に与えて学習する方法が考えられる。このような連接規則はどんなコーパスの分野でも一般的に通用するので汎用的に適用可能である。

4.2 様々な分野に対する実験について

今回の実験では提案した学習モデルの確定を目的としたので新聞記事の分野に絞って実験を行なった。他分野での実験も行なう必要がある。例えば、対話文や判例文など、新聞の社説とはかなり異なる分野のテキストを対象に実験を行なってみたい。

5 まとめ

我々は少量のタグつきコーパスと大量のタグなしコーパスを利用して日本語形態素解析システムのパラメータ学習を行なうシステムを提案した。その結果、HMM学習によってわずかながら現在のJUMANより正解率が向上することが確認できた。また、以下に示す2点がHMM学習において必要であることが明らかになった。

- 確率的に求まる最適なパラメータ値は分野によって異なる。そのため、対象と同分野の解析済みコーパスの存在が少量であっても重要であること
- 解析済みコーパスについてはその量を増やすよりもコーパス中にある誤りの種類の絶対数が少ないことが重要であること

今回の評価の基準として、一意に解析した解析済みコーパスとの比較によって誤り率を得ているので絶対的な評価ではない。例えば、「名詞：考え」と「連用形：考え」（動詞の連用形の名詞的用法）のようにどちらが正解とも言い難い場合も一意に処理しているので、不利な評価を行なっている。今後はこの評価の方法についても考察していきたい。

参考文献

[1] 松本裕治、他、“日本語形態素解析システムJUMAN使用説明書2.0”、奈良先端大技術報告書、*NAIST-IS-TR94025*,(1994).

[2] Cutting, D., Kupiec, J., Pedersen, J., and Sibun, P., “A Practical Part-of-Speech Tagger”, *ANLP-92*, pp.133–143, 1992.

[3] Constantine P. Papageorgiou, “Japanese Word Segmentation by Hidden Markov Model”, *Proc. Human Language Technology*, pp.283–288(1994).

[4] Nagata, M, “A Stochastic Japanese Morphological Analyzer Using a Forward-DP Backward -A* N-Best Search Algorithm”, *Proc. Coling*, pp.201–207(1994).

[5] 永田昌明、“確率モデルによる自由発話の形態素解析”, 音声言語情報処理 *3-1*, pp.1–6(1994).

[6] Chang, C. and Chen, C., “HMM-based Part-of-Speech Tagging for Chinese Corpora”, *Proc. of the Workshop on Very Large Corpora*, pp.40-47, 1993.