タグなしコーパスによる形態素解析と仮名漢字変換の精度向上

森 信介　伊東 伸泰
日本 IBM 東京基礎研究所
〒 242-8502 大和市下鶴間 1623-14
{mori,iton}@trl.ibm.co.jp

あ ら ま し
確率的言語モデルを基礎とする自然言語処理において、タグが付与された学習コーパスは重要であり、これを増量することが精度向上につながることがわかっている。しかしながら有意な精度向上のためには、学習コーパスを指数関数的以上に増加させる必要があり、このために必要なコーパスにタグを付与するコストは無視できない程度になっている。このような背景のもと、本論文ではタグなしコーパスの利用による形態素解析と仮名漢字変換の精度向上について述べる。実験では、タグなしコーパスの利用により、確率的言語モデルの予測力やそれに基づく仮名漢字変換の精度は有意に向上し、タグなしコーパスは 0.87 倍の量のタグつきコーパスに匹敵したが、形態素解析の精度向上は微小であった。

キーワード　仮名漢字変換　確率的言語モデル　コーパス　形態素解析　タグなし

Improvement of POS tagger and Kana Kanji Converter by an Untagged Corpus

Shinsuke Mori, Nobuyasu Itoh
Tokyo Research Laboratory, IBM Japan
1623-14 Shimotsuruma Yamatoshi
Kanagawaken 242-8502 Japan
{mori,iton}@trl.ibm.co.jp

Abstract
A tagged corpus plays an important role in natural language processing based on a stochastic language model and increasing the corpus size improves the accuracy. It is, however, necessary for a meaningful improvement to increase a corpus size more than expornentially and an annotation cost needed for it is not negligiable. In this paper, we discuss the usage of an untagged corpus. In the experiments, using an untagged corpus improved the predictive power of a stochastic language model and the accuracy of a *kana-kanji* converter based on it. But for a tagger the improvement was slight.

Key Words　*Kana-kanji* converter, Stochastic Language Model, Corpus, Morphological analysis, Untagged

1 はじめに

コーパスに基づく言語処理は、その客観性と計算機能力の劇的な向上や機械可読の辞書や文章の増加により、言語処理の方法論として確固たる地位を築いている。このアプローチでは、入力と出力の対を大量に用意し、未知の入力に対する出力をその大量の例から推定する。これらの一つとして、文の生成確率を計算する確率的言語モデルを用いる方法がある。例えば、この方法による形態素解析 [1] では、まず、文を形態素列とみなす確率的言語モデルを、予め形態素に分割してあるコーパスから作成し、次に未知の入力に対して、その生成確率が最大となる形態素列を計算し形態素解析の結果とする。別の応用例である仮名漢字変換 [2] では、キーボードからの入力を形態素に対応させるモデルと文を形態素列とみなす確率的言語モデルを組み合わせて、キーボードからの入力に対応し、且つ日本語の文として尤もらしい文字列を出力する。

確率的言語モデルを応用した言語処理には他にもあるが、これらは日本語の文字列が出力となる認識系と、日本語の文字列が入力となる解析系に大別される。解析系は、言語の文字列を入力として、その内部構造などの情報を付与する。この例は、上述した形態素解析であり、他の解析系と異なるのは出力として付与される情報のみである。構文解析の出力は構文木であり、読み付与の出力は読みである。認識系は、キーボードの入力や音響特徴量などの言語の文字列に対応する信号を言語の文字列に変換する。この例は、上述した仮名漢字変換であり、他の認識系と異なるのは入力の記号列のみである。音声認識の入力は音響特徴量であり、文字認識の入力は画像特徴量であり、誤り訂正 [1] の入力は誤りを含む文字列である。

確率的言語モデルの評価基準として、個々の応用における精度を用いることもできるが、確率的言語モデル単体での評価基準としては、一般にクロスエントロピーが用いられる。これは、学習に用いなかったコーパスの文字あたりの情報量であり、この値が低い方が実際の文章をより良くモデル化していると言える。さらに、クロスエントロピーが低い確率的言語モデルを用いれば、認識系や解析系の精度がより高くなる傾向にある。確率的言語モデルのクロスエントロピーを低下させるために、さまざまなモデルやその改良が提案されているが、最も効果的な方法は、学習コーパスの量を増やすことである。しかし、各応用に用いられる確率的言語モデルの最小単位は形態素であり、学習コーパスの各文はこの単位に正しく分割されている必要がある。さらに形態素などの解析系では、各単位に品詞が付与されている必要もある。このような学習コーパスをある程度の量準備するには、相当のコストがかかるのみならず、学習コーパスが大きくなれば、それを増やすことによる効果は急速に減少する。したがって、形態素に分割され品詞が付与されたコーパス (タグつきコーパス) に比べて非常に低いコストで大量に利用可能なタグなしコーパスを有効に利用する方法が求められている。この方法として、EM アルゴリズムによる確率的言語モデルの改良 [3] や未知語の収集 [4][1] が提案されている。

本論文では、低いコストで大量に利用可能なタグなしコーパスを自動的に形態素解析した結果得られるコーパスを、確率的言語モデルの学習コーパスに加えることによる形態素解析と仮名漢字変換の精度向上について述べる。タグなしコーパスの利用については、自動解析の結果をすべて学習コーパスに加える方法と、解析結果に対する信頼度が一定の閾値以上の文のみを学習コーパスに追加する方法について検討する。実験の結果、自動形態素解析の出力を学習コーパスに追加することにより、クロスエントロピーは減少し、仮名漢字変換の精度は有意に向上した。より具体的には、15 万文のタグなしコーパスの自動形態素解析の出力を追加すれば、同量のタグつきコーパスを追加するほどの精度向上はないものの、10 万文のタグつきコーパスを追加するよりもさらに精度は向上した。一方、形態素解析に関しては、自動形態素解析の出力すべてを学習コーパスに追加することでは、精度は向上しなかった。そこで信頼度が一定の閾値以上の文のみを学習コーパスに追加する方法を試みたが、精度の向上はわずかであった。

自動解析の結果をすべて学習コーパスに加える方法において、クロスエントロピーの減少は言語モデルの改善を意味するが、形態素解析や仮名漢字変換という応用での改善には差がある。これは、確率的言語モデルを用いる応用が、認識系であるか解析系であるかの違いに由来すると考えられる。

2 確率的言語モデル

この節では、確率的言語モデルの一つであるクラス n-gram モデルとそれに基づく形態素解析と仮名漢字変換について述べる。

2.1 クラス n-gram モデル

クラス n-gram モデル [5] は、あらかじめ形態素 m をクラスと呼ばれるグループ c に分類しておき、先行するクラスの列を直前の事象とみなして分類する。そして、以下の式が示すように、まず次のクラス c_i を予測し、次にそのクラスから形態素 m_i を予測することを繰り返すことで、形態素の列 $\boldsymbol{m}$ として表現される文の出現確率を計算する。

ここで $\mathcal{M}_k$ は既知形態素の集合である。

$$M_{c,n}(\boldsymbol{m}) = \prod_{i=1}^{h+1} P_c(m_i|c_{i-k}c_{i-k+1}\cdots c_{i-1}) \quad (1)$$

$$P_c(m_i|c_{i-k}\cdots c_{i-2}c_{i-1}) = \begin{cases} \text{if} \quad m_i \in \mathcal{M}_k \\ \quad P(c_i|c_{i-k}\cdots c_{i-2}c_{i-1})P(m_i|c_i) \\ \text{if} \quad m_i \notin \mathcal{M}_k \\ \quad P(\mathtt{UM}_t|c_{i-k}\cdots c_{i-2}c_{i-1})M_{x,t}(m_i) \end{cases} \quad (2)$$

この式の中の $\mathtt{UM}_t$ は、品詞 t の未知語に対応するクラスである。未知語の表記は、文字 2-gram モデルなどに基づく未知語モデル $M_{x,t}$ を用いて予測される。また、c_j $(j \leq 0)$ は、文頭に対応する特別な記号である。これを導入することによって式が簡便になる。さらに、c_{h+1} は、語末に対応する特別な記号であり、これを導入することによって、すべての可能な文字列に対する確率の和が 1 となる [6]。

2.1.1 確率値の推定

確率 $P(c_i|c_{i-k}c_{i-k+1}\cdots c_{i-1})$ と $P(m_i|c_i)$ の値は、まず既知形態素集合を定義し、学習コーパスの未知形態素を未知形態素に対応する特別な記号に置き換えて頻度を計数し、最尤推定することで得られる。

$$P(c_i|c_{i-k}c_{i-k+1}\cdots c_{i-1}) = \frac{N(c_{i-k}c_{i-k+1}\cdots c_i)}{N(c_{i-k}c_{i-k+1}\cdots c_{i-1})} \quad (3)$$

$$P(m_i|c_i) = \frac{N(m_i, c_i)}{N(c_i)} \quad (4)$$

データスパースネスの問題に対処する方法として、補間を用いることができる。これは、次の式で表されるように、より信頼性が高いことが期待される、より低次のマルコフモデルの遷移確率を一定の割合で足し合わせるという操作を施すことをいう。

$$P'(c_i|c_{i-k}c_{i-k+1}\cdots c_{i-1}) = \sum_{j=0}^{k} \lambda_j P(c_i|c_{i-j}c_{i-j+1}\cdots c_{i-1})$$

$$\text{ただし } 0 \leq \lambda_j \leq 1,\ \sum_{j=0}^{k} \lambda_j = 1$$

2.2 形態素解析

日本語の形態素解析は、日本語のアルファベット $\mathcal{X}$ のクリーネ閉包に属する文 $\boldsymbol{x} \in \mathcal{X}^*$ を入力として、これを表記 $\mathcal{W} = \mathcal{X}^*$ と品詞 $\mathcal{T}$ の直積として定義される形態素 $\mathcal{M} = \{(w,t)|w \in \mathcal{W} \wedge t \in \mathcal{T}\}$ の列 $\boldsymbol{m}$ に分解して出力することと定義できる。このとき、出力される形態素列の表記の連接は、入力のアルファベット列に等しくなければならない。つまり、入力のアルファベット列 (長さ l) を $\boldsymbol{x} = x_1x_2\cdots x_l$ とし、出力の形態素列 (要素数 h) を $\boldsymbol{m} = m_1m_2\cdots m_h$ とすると以下の式が成り立つ必要がある。ただし、$\boldsymbol{w}(\boldsymbol{m})$ は形態素の連接 $\boldsymbol{m}$ の表記の連接を表わすものとする。

$$\boldsymbol{w}(\boldsymbol{m}) = w(m_1)w(m_2)\cdots w(m_h) = x_1x_2\cdots x_l = \boldsymbol{x}$$

一般に、これを満たす解は一意ではない。形態素解析の問題は、可能な解の中から人間の判断 (正解) に最も近いと推測される形態素列 (単語分割と品詞割り当て) を選択し出力することである。確率的言語モデルによる形態素解析では、この選択の基準として、大量の形態素解析済みコーパスからの推定値を用いる。

2.3 確率的言語モデルによる形態素解析

確率的形態素解析器は、品詞という概念を内包する確率的言語モデルを基にして、与えられた文字列 $\boldsymbol{x}$ に対する確率最大の形態素列 $\hat{\boldsymbol{m}}$ を計算し出力する。これは、以下の式で表される。

$$\begin{aligned} \hat{\boldsymbol{m}} &= \operatorname*{\mathbf{argmax}}_{\boldsymbol{w}(\boldsymbol{m})=\boldsymbol{x}} P(\boldsymbol{m}|\boldsymbol{x}) \\ &= \operatorname*{\mathbf{argmax}}_{\boldsymbol{w}(\boldsymbol{m})=\boldsymbol{x}} P(\boldsymbol{m}|\boldsymbol{x})P(\boldsymbol{x}) \quad (\because P(\boldsymbol{x}) \text{は} \boldsymbol{m} \text{によらず}) \\ &= \operatorname*{\mathbf{argmax}}_{\boldsymbol{w}(\boldsymbol{m})=\boldsymbol{x}} P(\boldsymbol{x}|\boldsymbol{m})P(\boldsymbol{m}) \quad (\because \text{ベイズの公式}) \\ &= \operatorname*{\mathbf{argmax}}_{\boldsymbol{w}(\boldsymbol{m})=\boldsymbol{x}} P(\boldsymbol{m}) \quad (\because P(\boldsymbol{x}|\boldsymbol{m}) = 1) \end{aligned}$$

この式の最後の $P(\boldsymbol{m})$ が品詞という概念を内包する確率的言語モデルである。このようなモデルとして、上述したクラス n-gram モデルを用いることができる。

2.4 仮名漢字変換

仮名漢字変換は、キーボードから直接入力することが可能な記号 $\mathcal{Y}$ の正閉包 $\boldsymbol{y} \in \mathcal{Y}^+$ からの、日本語のアルファベット $\mathcal{X}$ の正閉包 $\boldsymbol{x} \in \mathcal{X}^+$ への対応である。仮名漢字変換の入力の記号列は、一般的に、ユーザーが計算機に入力したい日本語文の読みである。このとき、複数の日本語文が同一の読みを共有する状況が頻繁に発生する。つまり、仮名漢字変換の読みに対応する日本語文 (変換候補) が複数あるという状況である。このような場合には、入力効率

を最大にするために、ユーザーが意図している日本語列に近いと推測される変換候補を順次出力する。したがって、仮名漢字変換は、キーボードから直接入力することが可能な記号列 (読み) から日本語文 (変換候補) の列への写像である。これは、以下の式のように示される。

$$\mathcal{Y}^+ \mapsto (\mathcal{X}^+,\ \mathcal{X}^+, \cdots,\ \mathcal{X}^+)$$

ここで、右辺の変換候補の数は入力の記号列に依存し、それらはユーザーが意図している日本語文に近いと推測される順に左から右へ並んでいるとする。

2.5 確率的言語モデルによる仮名漢字変換

上で定義したような仮名漢字変換を実現する方法の一つとして、確率的言語モデルを用いる方法 [2] がある。これは、基本的には音声認識と同じであるが、入力が音響特徴量の列ではなくキーボードから入力される記号列である点と最尤解だけでなくすべての候補をその尤度順に出力する点が異なる。この尤度は、キーボードからの入力の記号列が与えられたときの日本語文の条件付確率 $P(\boldsymbol{x}|\boldsymbol{y})$ である。したがって、確率的モデルによる仮名漢字変換器 wnm は、以下のような写像である。

$$wnm(\boldsymbol{y}) = (\boldsymbol{x}_1,\ \boldsymbol{x}_2,\ \ldots,\ \boldsymbol{x}_n)$$

$$\text{ただし}\quad i \leq j \Leftrightarrow P(\boldsymbol{x}_i|\boldsymbol{y}) \geq P(\boldsymbol{x}_j|\boldsymbol{y})$$

この式から、仮名漢字変換器の主要な役割は、各変換候補の確率値 $P(\boldsymbol{x}|\boldsymbol{y})$ の順序関係の算出であることがわかる。逆にこの順序関係を保持している限りにおいて、実際にはこの確率値以外の他の値を用いてもよいと結論できる。この点を考慮に入れて、以下の式のように確率的言語モデルの分離が行なわれる。

$$\begin{aligned}
&P(\boldsymbol{x}_i|\boldsymbol{y}) \geq P(\boldsymbol{x}_j|\boldsymbol{y}) \\
&\quad\Leftrightarrow\quad \frac{P(\boldsymbol{y}|\boldsymbol{x}_i)P(\boldsymbol{x}_i)}{P(\boldsymbol{y})} \geq \frac{P(\boldsymbol{y}|\boldsymbol{x}_j)P(\boldsymbol{x}_j)}{P(\boldsymbol{y})} \\
&\qquad\qquad (\because \text{ベイズの公式}) \\
&\quad\Leftrightarrow\quad P(\boldsymbol{y}|\boldsymbol{x}_i)P(\boldsymbol{x}_i) \geq P(\boldsymbol{y}|\boldsymbol{x}_j)P(\boldsymbol{x}_j) \qquad (5) \\
&\qquad\qquad (\because P(\boldsymbol{y}) \text{は} \boldsymbol{x}_i \text{や} \boldsymbol{x}_j \text{によらない})
\end{aligned}$$

この式において、日本語文 $\boldsymbol{x}$ の出現確率を表す $P(\boldsymbol{x})$ が確率的言語モデルであり、上述のクラス n-gram モデルを用いることができる。残りの $P(\boldsymbol{y}|\boldsymbol{x})$ は、日本語文 $\boldsymbol{x}$ が与えられたときのキーボードからの入力の記号列 (読み) の確率を表す。これは確率的仮名漢字モデルと呼ばれる。

2.6 確率的仮名漢字モデル

確率的仮名漢字モデル $P(\boldsymbol{y}|\boldsymbol{x})$ は、日本語文 $\boldsymbol{x}$ が与えられたときのキーボードからの入力の記号列 $\boldsymbol{y}$ の確率を表す。あらゆる可能な日本語文に対する入力記号列の確率を推定することは不可能であり、日本語文を形態素に分割し、それらの入力記号列との対応関係がそれぞれ独立であると仮定する。このとき、形態素列 $\boldsymbol{m}$ が与えられたときの入力記号列 $\boldsymbol{y}$ の確率的仮名漢字モデル M_{kk} による出現確率は以下の式で表される。

$$M_{kk}(\boldsymbol{y}|\boldsymbol{m}) = \prod_{i=1}^{h} P(\boldsymbol{y}_i|m_i) \qquad (6)$$

ここで、入力記号部分列 $\boldsymbol{y}_i$ は形態素 m_i に対応する入力記号列であり、以下の条件を満たす。

$$\boldsymbol{y} = \boldsymbol{y}_1\boldsymbol{y}_2\cdots\boldsymbol{y}_h$$

確率 $P(\boldsymbol{y}_i|m_i)$ の値は、形態素ごとに読み (入力記号列) が振られたコーパスから以下の式を用いて最尤推定することで得られる。

$$P(\boldsymbol{y}_i|m_i) = \frac{N(\boldsymbol{y}_i,\ m_i)}{N(m_i)} \qquad (7)$$

2.7 確率的言語モデルと確率的仮名漢字モデルの統合

すでに述べたように、確率的モデルによる仮名漢字変換において、変換候補に順序関係を与える尤度は、確率的言語モデルによる確率値と確率的仮名漢字モデルによる確率値の積で与えられる。したがって、式 (5) 中の $P(\boldsymbol{y}|\boldsymbol{x})P(\boldsymbol{x})$ は式 (1)(6) 及び $P(\boldsymbol{x}) \approx M_{c,n}(\boldsymbol{m})$ から以下のようになる。

$$\begin{aligned}
&P(\boldsymbol{y}|\boldsymbol{x})P(\boldsymbol{x}) \\
&\quad\approx\quad M_{kk}(\boldsymbol{y}|\boldsymbol{m})M_{c,n}(\boldsymbol{m}) \\
&\quad=\quad \prod_{i=1}^{h+1} M_{kk}(y_i|m_i)M_{c,n}(m_i|c_{i-k}\cdots c_{i-2}c_{i-1})
\end{aligned}$$

クラス n-gram モデルの既知語と未知語の場合分けの式 (2) と最尤推定の式 (3)(4)(7) を代入することで、この積の繰り返しの対象の式は、予測される形態素が既知か未知かに応じて以下のように計算される。

1. 既知形態素の場合 ($m_i \in \mathcal{M}_k$)

$$\begin{aligned}
&M_{kk}(\boldsymbol{y}_i|m_i)M_{c,n}(m_i|c_{i-k}\cdots c_{i-2}c_{i-1}) \\
&\quad=\quad \frac{N(c_{i-k}\cdots c_{i-1}c_i)}{N(c_{i-k}\cdots c_{i-2}c_{i-1})}\frac{N(m_i,c_i)}{N(c_i)}\frac{N(\boldsymbol{y}_i,m_i)}{N(m_i)} \\
&\quad=\quad \frac{N(c_{i-k}\cdots c_{i-1}c_i)}{N(c_{i-k}\cdots c_{i-2}c_{i-1})}\frac{N(\boldsymbol{y}_i,m_i)}{N(c_i)}
\end{aligned}$$

ここで、形態素とクラスの対応関係が多対一なので $N(m_i, c_i) = N(m_i)$ であることを用いている。

2. 未知形態素の場合 ($m_i \notin \mathcal{M}_k$)

$$M_{kk}(\boldsymbol{y}_i|m_i)M_{c,n}(m_i|c_{i-k}\cdots c_{i-2}c_{i-1}) = \frac{N(c_{i-k}\cdots c_{i-1}c_i)}{N(c_{i-k}\cdots c_{i-2}c_{i-1})}M_{x,t}(m_i)M_{kk}(\boldsymbol{y}_i|m_i)$$

この式の $M_{x,t}(m_i)M_{kk}(\boldsymbol{y}_i|m_i)$ の部分は各未知語の仮名漢字変換に対応する。この部分については、未知語の仮名漢字変換が困難であるという理由から入力記号列 $\mathcal{Y}$ 上の未知語モデル $M_{y,t}$ が代わりに用いられる。これは以下の式で与えられる近似である。

$$M_{x,t}(m_i)M_{kk}(\boldsymbol{y}_i|m_i) \approx M_{y,t}(\boldsymbol{y}_i)$$

このようなモデルは、学習コーパスの未知語を $\mathcal{Y}^+$ に変換しておき、通常のパラメータ推定を行なうことで容易に得られる。このようなモデルによる未知語の変換結果は入力の記号列と同じである。実際には、多くの未知語が片仮名列であることから、未知語を片仮名列として出力している。

3 タグなしコーパスの利用

前節で説明した確率的言語モデルによる形態素解析や仮名漢字変換では、パラメータ推定のために学習コーパスが必要である。このコーパスは、形態素への分割や読みなどの情報が付与されている必要がある。当然ながら、この学習コーパスは大きいほうが形態素解析や仮名漢字変換の精度が高いが、学習コーパスの大きさに対する精度の上昇は、学習コーパスの文字数の対数値に対しても比例よりも遅い。実際、ある程度の大きさの学習コーパスから推定された確率的言語モデルに基づく形態素解析や仮名漢字変換の精度を有意に向上するには、数倍以上の文に対して新たに形態素への分割や読みなどの情報を付与する必要がある。次節で述べる実験では、学習コーパスとして EDR コーパスの約 5 万文を用いた場合と約 20 万の文を用いた場合の形態素解析の再現率と適合率の平均は、それぞれ 92.22%と 93.19%である。つまり、約 15 万文に情報を付与することにより減少させることができた誤りは約 12%ということである。仮名漢字変換の同様の場合の誤り減少率は約 46%である。

一方で、形態素への分割や読みなどの情報が付与されていないタグなしコーパスは、多くの新聞などが機械可読の状態で容易に入手可能でり、量は膨大である。1 年分の日本経済新聞は 100 万以上の分を含んでいる。このような生コーパスを有効に利用できれば確率的言語モデルの様々な応用の精度を容易に向上させることが可能であると考えられる。

3.1 タグなしコーパスの利用

本論文で述べるタグなしコーパスの利用は、以下の 2 通りである。

1. 一定の学習コーパスから推定された確率的言語モデルに基づく形態素解析器による出力 (図 1 参照)

2. 上記の文のうちで、後述する信頼度の条件を満たす文

一定量の学習コーパスから推定された確率的言語モデルの予測精度を向上するためには、すでにある学習コーパスの切り分けや品詞付与の基準に沿ったコーパスを用意する必要がある。学習コーパスがある程度大きければ、それに基づく形態素解析の精度はかなり高いので、自動解析の結果が利用できると考えられる。しかしながら、形態素解析精度の向上を考えれば、自動解析の結果を無条件に学習コーパスに追加することは、有効であるとは考えられない。形態素解析という観点からは、タグなしコーパスから得られる情報は、文頭や句読点の前後の形態素境界程度であろう。したがって、解析誤りによる負の影響を排除するために、ある程度の精度が期待できる解析結果のみ利用すべきでる。このような観点から、信頼度を利用して文を選別する方法も試みることとした。また、タグなしコーパスの分野は、タグつきコーパスの場合と同様に、言語モデルを利用する予定の分野 (テストコーパスの分野) と大きく異ならない分野であることが望ましいと考えられる。

次節で述べる実験では、図 1 が示すように、タグつきの学習コーパスのみから推定した言語モデルに基づく形態素解析器 (Tagger A) と、同じタグつきの学習コーパスに加えて、タグなしコーパスの Tagger A による解析結果から推定した言語モデルに基づく形態素解析器 (Tagger B) の精度を比較する。仮名漢字変換器の場合もほぼ同様であるが、タグなしコーパスには読み (入力記号列) が振られていないので、確率的仮名漢字モデルはタグつきの学習コーパスのみから推定される。したがって、異なるのは確率的言語モデルの部分のみである。

3.2 信頼度

形態素解析の正解率との相関が見られた以下の 3 つの尺度を用いた。

- 未知語の数

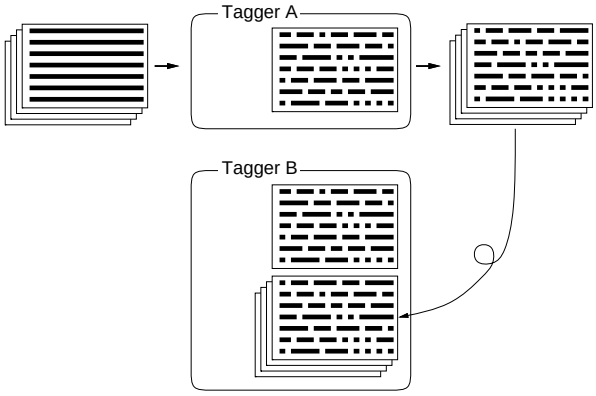

図 1: タグなしコーパスの利用

未知語とは、学習コーパスに出現していない形態素であり、未知語の前後では切り分けの誤りが多い。さらに、未知語は、各品詞に対応する特別な記号から生成されるので、その前後では表記レベルでの 2-gram が予測に利用されていないことを意味する。未知語の数が多いということは、形態素解析結果が誤りを含む可能性が高いということである。その一方、未知語を含む形態素解析結果を学習コーパスに追加することによってのみ言語モデルの語彙は増加する。したがって、少しの未知語を含む解析結果を利用することが、精度向上につながると考えられる。

- 生成確率

  ある確率的言語モデルによる生成確率が高い文は、その確率的言語モデルによれば適格文である可能性が高いことを意味する。確率的言語モデルがクラス 2-gram モデルである場合には、学習コーパスに出現しない 2-gram や 1-gram が相対的に少ないことを意味する。文長の影響を排除するために、生成確率の負対数値を文字数で割った値 (エントロピー) を用いることとした。したがって、この値が低い形態素解析結果を選択することとした。

- 第 2 候補の生成確率との比

  多くの場合、形態素解析の解の第 1 候補と第 2 候補は一部だけが異なり、他の部分が全く同じである。この場合に、第 1 候補の生成確率と第 2 候補の生成確率に大きな差がない場合には、その部分の曖昧性が非常に高いことを意味する。逆に、差が大きな場合には、解析結果の各部分の曖昧性が低いと考えられる。したがって、この差が大きい解析結果を選択することにした (計算の便宜上、差ではなく比を使っている)。

以上の値に閾値を設けて、タグなしコーパスの形態素解析結果のうちで、すべての条件を満たす文のみを用いることとする。この場合、図 1 中の Tagger B の学習コーパスには、Tagger A の出力のうちですべての条件を満たす文のみとなる。

4 評価

2 節で述べた確率的言語モデルに基づく形態素解析と仮名漢字変換器を実装し、3 節で説明した方法でタグなしコーパスを利用する場合と利用しない場合の精度を評価した。この節では、実験の条件とその結果を提示し、それに対する考察を述べる。

4.1 実験の条件

実験には EDR コーパス [7] を用いた。このコーパスの各文は、以下のように、入力記号列 (読み) が振られた形態素に分割されている。

１９８７/１９８７/数字　ネン/年/名詞　ノ/の/助詞
アタラシ/新し/形容詞　イ/い/語尾
ケイコウ/傾向/名詞　ハ/は/助詞　、/、/記号
ＩＢＭ/ＩＢＭ/名詞　ガ/が/助詞　ドレ/どれ/名詞
ダケ/だけ/助詞　セイヒン/製品/名詞　ノ/の/助詞
（後略）

まず、コーパスを 10 個に分割し、この内の 9 個を学習コーパスとし、残りの 1 個をテストコーパスとした。入力記号列と形態素の対応を記述する確率的仮名漢字モデルは、入力記号列と表記と品詞の組の列から学習する。クラス 2-gram モデルは、このうちの入力記号列を除いた表記と品詞の組の列から学習する。

4.2 確率的言語モデルの評価基準

確率的言語モデルの良否の尺度としては、クロスエントロピーが一般的である。これは、確率的言語モデルを M とし、テストコーパスを $\mathcal{S} = \{s_1,\ s_2,\ \ldots,\ s_k\}$ とすると以下の式で与えられる[1] 。ただし、$|s|$ は文 s の長さ (文字数) を表わす。

$$H(M,\mathcal{S}) = -\frac{\sum_{i=1}^{k} \log M(s_i)}{\sum_{i=1}^{k}(|s_i|+1)}$$

この値は、コーパス $\mathcal{S}$ をモデル M で符合化した時の文字あたりの平均符合長の下限であり、$\mathcal{S}$ として無作為に抽出された十分多数の文を選択すれば、複数のモデルの良否を比較するための尺度となる。定義から明らかなように、この値がより小さいほうがより良い言語モデルである。クロスエントロピーの意味で良い言語モデルを用いる方が、形態素解析や仮名漢字変換などの応用の精度が良いと考えられる。

4.3 形態素解析の評価基準

我々が用いた評価基準は、先行研究 [8] と同じ再現率と適合率である。これらは、次のように定義される。EDR コーパスに含まれる形態素数を N_{EDR}、解析結果に含まれる形態素数を N_{SYS}、分割と品詞の両方が一致した形態素数を N_{COR} とすると、再現率は N_{COR}/N_{EDR} と定義され、適合率は N_{COR}/N_{SYS} と定義される。例として、コーパスの内容と解析結果が以下のような場合を考える。

コーパス

外交/名詞 政策/名詞 で/助動詞/ は/助詞 な/形容詞 い/語尾

解析結果

外交政策/名詞 で/助詞 は/助詞 な/形容詞 い/語尾

この例において、分割と品詞の両方が一致した形態素は「は/助詞」と「な/形容詞」と「い/語尾」であるので、$N_{COR} = 3$ となる。また、コーパスには 6 つの形態素が含まれ、解析結果には 5 つの形態素が含まれているので、$N_{EDR} = 6,\ N_{SYS} = 5$ である。よって、再現率は $N_{COR}/N_{EDR} = 3/6$ となり、適合率は $N_{COR}/N_{SYS} = 3/5$ となる。

4.4 仮名漢字変換の評価基準

我々が用いた評価基準は、各文を一括変換することで得られる最尤解と正解の最長共通部分列 (longest common subsequence)[9] の文字数に基づく再現率と適合率である。EDR コーパスに含まれる文字数を N_{EDR} とし、仮名漢字変換結果に含まれる文字数を N_{SYS} とし、これらの最長共通部分列の文字数を N_{LCS} とすると、再現率は N_{LCS}/N_{EDR} と定義され、適合率は N_{LCS}/N_{SYS} と定義される。例として、コーパスの内容と変換結果が以下のような場合を考える。

コーパス

私 が 長 尾 真 で す 。

変換結果

渡 し が 長 尾 マ コ ト で す 。

この場合、最長共通部分列は「が長尾です。」の 6 文字であるので、$N_{LCS} = 6$ となる。コーパスに含まれる文字数は 8 であり、変換結果に含まれる文字数は 11 であるので、$N_{EDR} = 8,\ N_{SYS} = 11$ である。よって、再現率は $N_{LCS}/N_{EDR} = 6/8$ となり、適合率は $N_{LCS}/N_{SYS} = 6/11$ となる。

4.5 評価

表 1 は、学習コーパスの 1/4 をタグつきコーパスとみなし、残りの学習コーパスをタグなしコーパスみなして得られる確率的言語モデルのクロスエントロピーとこれに基づく形態素解析器や仮名漢字変換器のテストコーパスにおける精度を、タグなしコーパスを利用しない場合と比較した結果である。確率的言語モデルの評価基準であるクロスエントロピーは 2 倍のタグつきコーパスを利用する場合よりも低く、3 倍のタグつきコーパスを利用する場合よりも高い。この区間でタグつきコーパスの増加量とクロスエントロピーの減少量が比例するとして補間すると、140,267 文のタグなしコーパスを追加する効果は約 63,561 文 (約 0.45 倍) のタグつきコーパスを追加する効果に匹敵することになる。しかしながら、形態素解析の精度は、タグなしコーパスの自動解析結果を利用することにより、再現率が低下し適合率が上昇するという結果である。それぞれの変化は微小であり、再現率と適合率の平均を考えた場合には、僅かながら下がっている。学習コーパスの量を増加させれば、確率的言語モデルの記述量は大きくなるので、タグなしコーパスの自動解析結果をそのまま利用するのは形態素解析器にとって利点は全くない。仮名漢字変換の精度は、3 倍のタグつきコーパスを利用する場合よりも高く、4 倍のタグつきコーパスを利用する場合よりも低い。再現率と適合率の平均を精度と考え、この区間でタグつきコーパスの増加量と精度の向上が比例するとして補間すると、140,267 文のタグなしコーパスを追加する効果は約 122,228 文 (約

[1] 式の分母の +1 は文末記号に対応する。これは、s_x, s_y と $s_x s_y$ を区別するために必要である。

表 1: 学習コーパスサイズと精度

学習コーパスの文数	タグつき	46,755	46,755	93,512	140,267	187,022
	タグなし	140,267	0	0	0	0
クロスエントロピー		4.6947	4.9215	4.7310	4.6300	4.5655
形態素解析の精度	再現率	91.99%	92.14%	92.77%	93.10%	93.30%
	適合率	92.36%	92.30%	92.76%	92.94%	93.08%
仮名漢字変換の精度	再現率	93.41%	88.66%	91.62%	92.96%	93.70%
	適合率	94.61%	91.44%	93.44%	94.26%	94.79%

0.87 倍) のタグつきコーパスを追加する効果に匹敵するといえる。

タグなしコーパスの追加がクロスエントロピーの減少と仮名漢字変換の精度向上にのみつながっている。これは、双方ともに形態素の単位は便宜的に導入されているに過ぎず、問題にしているのは入力の文の出現確率であることに起因すると思われる。換言すれば、形態素列としての解釈に部分的な差異があっても、ある文字列 (文) が実際に出現したということが言語モデルや仮名漢字変換にはかなりの情報を持つということである。これは、他の認識系の応用にも当てはまるであろう。対して、形態素の単位が重要な役割を果たしている形態素解析では、タグなしコーパスの追加による効果は否定的である。形態素解析には、文の形態素列としての解釈が重要であり、ある文が実際に出現したということだけでは、文頭や句読点の前後の形態素の境界程度の情報しかなく、他の部分に含まれるであろう解析誤りが誤った情報となり、精度を下げている。

このように、形態素解析にとってはタグなしコーパスは、情報があるにせよ非常に少ない。そこで、解析誤りが少ないと思われる文を前節で述べた信頼度で選択し、これらの文のみを学習コーパスに追加する実験を行なった。自動解析の結果に課す条件は以下の通りである。各閾値は恣意的に決定したが、一応の理由を付記しておく。

- 数字を除く未知語が 1 つ

  多くの未知語が含まれる文の解析精度は低い。

- 出現する未知語の頻度は 5 以上

  頻度の低い未知語は解析誤りの可能性が高い。

- 文字あたりの平均エントロピーが 4 未満

  テストコーパスの文字あたりの平均エントロピーは 4.6〜4.8 程度である。

- 第 2 候補との文字あたりの平均エントロピー差が 0.5 より大きい

  文字あたりの平均エントロピーの 1 割程度。

EDR コーパスを用いた実験では、適合率が 92.16%で再現率が 92.31%となり、タグなしコーパスを用いない場合の精度 (適合率が 92.14%、再現率が 92.30%) よりもわずかに上昇した。同様の実験を日本経済新聞の 1996 年の記事約 1 万文からなるタグつきコーパスと同年の他の記事からなるタグなしコーパスに対して行なった結果でも、適合率は 95.31%から 95.33%と小幅ながら上昇し、再現率も 95.46%から 95.52%と小幅ながら上昇した。いずれのコーパスの場合も、精度は上昇しているがその幅は小さい。表 1 から分かるように、仮名漢字変換では約 15 万文のタグつきコーパスによる誤りの減少は約 42%であるのに対し、形態素解析ではこれが約 12%であるが、このことを考慮に入れても上昇幅は小さい。信頼度の尺度の改善やタグなしコーパスの増量などの余地はあるが、タグなしコーパスの利用によって形態素解析の精度を有意に向上させるのは容易ではないと考えられる。

5 おわりに

本論文では、低いコストで大量に利用可能なタグなしコーパスを自動的に形態素解析した結果得られるコーパスを、確率的言語モデルの学習コーパスに加えることによる形態素解析と仮名漢字変換の精度向上について述べた。タグなしコーパスの利用については、自動解析の結果をすべて学習コーパスに加える方法と、解析結果に対する信頼度が一定の閾値以上の文のみを学習コーパスに追加する方法について検討した。実験の結果、自動形態素解析の出力を学習コーパスに追加することにより、クロスエントロピーは減少し、仮名漢字変換の精度は有意に向上した。より具体的には、15 万文のタグなしコーパスの自動形態素解析の出力を追加すれば、同量のタグつきコーパスを追加する

ほどの精度向上はないものの、10 万文のタグつきコーパスを追加するよりもさらに精度は向上した。一方、形態素解析に関しては、自動形態素解析の出力すべてを学習コーパスに追加することでは、精度は向上しなかった。そこで信頼度が一定の閾値以上の文のみを学習コーパスに追加する方法を試みたが、精度の向上はわずかであった。

自動解析の結果をすべて学習コーパスに加える方法において、クロスエントロピーの減少は言語モデルの改善を意味するが、形態素解析や仮名漢字変換という応用での改善には差がある。これは、確率的言語モデルを用いる応用が、認識系であるか解析系であるかの違いに由来すると考えられる。

参考文献

[1] 永田昌明. 確率モデルによる日本語処理に関する研究. PhD thesis, 京都大学, 1999.

[2] 森信介, 土屋雅稔, 山地治, 長尾真. 確率的モデルによる仮名漢字変換. 情報処理学会論文誌, Vol. 40, No. 7, pp. 2946–2953, 1999.

[3] 竹内孔一, 松本裕治. Hmm による日本語形態素解析システムのパラメータ学習. 情報処理学会研究報告, 1995.

[4] 森信介, 長尾眞. n グラム統計によるコーパスからの未知語抽出. 情報処理学会論文誌, Vol. 39, No. 7, 1998.

[5] Peter F. Brown, Vincent J. Della Pietra, Peter V. deSouza, Jennifer C. Lai, and Robert L. Mercer. Class-based n-gram models of natural language. *Computational Linguistics*, Vol. 18, No. 4, pp. 467–479, 1992.

[6] King Sun Fu. *Syntactic Methods in Pattern Recognition*, Vol. 12 of *Mathematics in Science and Engineering*. Accademic Press, 1974.

[7] 日本電子化辞書研究所. EDR 電子化辞書仕様説明書, 1993.

[8] 永田昌明. Edr コーパスを用いた確率的日本語形態素解析. EDR 電子化辞書利用シンポジウム, pp. 49–56, 1995.

[9] Alfred V. Aho. 文字列中のパターン照合のためのアルゴリズム. コンピュータ基礎理論ハンドブック, I: 形式的モデルと意味論, pp. 263–304. Elseveir Science Publishers, 1990.