自然言語処理 122－6
（1997. 11. 20）

係り受けを用いた確率的言語モデル

森 信介 長尾 真
京都大学工学研究科
〒606-01 京都市左京区吉田本町
{mori,nagao}@kuee.kyoto-u.ac.jp

あらまし
本論文では，係り受けを用いた日本語の確率的言語モデルを提案する．このモデルにおける予測単位は文節の属性である。これは、各文節の内容語の主辞と機能語の主辞の直積として表される。文節の属性の間の関係は、確率文脈自由文法によって記述される。文節の属性からの文節の形態素列の予測には形態素 n-gram モデルを用いる。また、形態素列のに含まれる未知語の予測には文字 n-gram モデルを用いる。このモデルは、任意の文字列に対してその生成確率を計算できるという頑強性と、未知語から係り受けまでを一貫してモデル化しているという完全性を備えている。

キーワード 言語モデル 文節 係り受け 文脈自由文法 構文解析

A Stochastic Language Model using Dependency

Shinsuke Mori Makoto Nagao
Department of Electrical Engineering, Kyoto University
Yoshida-honmachi, Sakyo, Kyoto, 606-01 Japan
{mori,nagao}@kuee.kyoto-u.ac.jp

Abstract
In this paper, we present a stochastic language model for Japanese using dependency. The prediction unit in this model is an attribute of *bunsetsu*. This is represented by the product of the head of content words and that of function words. The relation between the attributes of *bunsetsu* is ruled by a context-free grammar. The word sequences are predicted from the attribute using word n-gram model. The spell of Unknow word is predicted using character n-gram model. This model is robust in that it can compute the probability of an arbitrary string and is complete in that it models from unknown word to dependency at the same time.

Key Words Language Model, Bunsetsu, Dependency, Contex-free Grammar, Parsing

1 はじめに

自然言語処理の有力な方法論としての確率的言語モデルは、音声認識に代表される認識系や形態素解析に代表される解析系などの様々な応用において、その有用性が確かめられている。この方法論では、いくつかのパラメータを持つ確率的言語モデルを構成し、未知の入力に対する予測力最大 (クロスエントロピー最小) を目的としてパラメータを推定する。それぞれの応用に個別に対処するのであれば、認識精度や解析精度を目的関数としてパラメータを推定する方が良いと考えられるが、このような方法は問題への依存が大きく、これを体系的に行なう方法は我々の知る範囲では存在しない。これに対して、確率的言語モデルという方法論は、多様な自然言語処理の応用のための枠組から、それらすべてに共通の言語に対する記述を分離し、確率的モデルの研究成果を活用して、体系的に個々の応用の精度を向上させることを可能にする。

このような枠組では、言語に対する仮説はアルファベット列から確率値への写像として表現されることだけが条件である。最初のモデルは、Shannon による連続する n 文字 (n-gram) の頻度を利用するモデルである [1]。具体的には、ある自然言語に属する文を大量に集めたコーパス中に、その言語を記述するために用いられる文字がどのように出現するかを観測し、その結果に基づいてモデルのパラメーターを決定する。これは、連続する要素の関係のみが記述できるという点で、現在実用となっている応用の言語モデルと同じである。しかし、離れた要素間の関係を仮定する方が記述が容易な言語現象もあり、このような現象を捉えたモデルを用いることによって、様々な応用の精度が向上すると考えられる。

本論文では，日本語を対象言語として、文節間の係り受けを用いる完全な確率的言語モデルを提案し、このモデルの予測力をクロスエントロピーで評価する。文節の種類は膨大なので、データスパースネスの問題を回避するために、単語 n-gram モデルに対して提案されているクラスの概念 [2] を応用した。つまり、各文節は内容語と付属語の最後の品詞から一意に計算されるクラスで代表し、このクラスの間での係り受けを文脈自由文法を [3, 4] を用いてモデル化した。クラスが与えられると、文節の具体的な内容語列と機能語列は、それぞれ独立に未知語モデルを含む形態素 n-gram モデル [5] によって生成される。

上述のモデルは、各文節の性質を品詞のみに依存するとしている。しかし、人手で与えた品詞は、確率的言語モデルという観点から最適であるとは限らない。このことは、形態素 n-gram モデルに対して、人手で与えた品詞と予測力を基準として推定したクラス分類を比較した報告 [6] において実験的に示されている。この報告を踏まえて、我々は、係り受けモデルに対しての予測力を基準とした形態素クラスタリングを提案する。また、このモデルの応用の一つとしての構文解析について述べる。

上述のモデルの有効性を確かめるため，EDR コーパス [7] を用いて実験を行なった．コーパスを 9 対 1 の比率で分割し、前者から言語モデルを推定し、後者に対してクロスエントロピーの計算を行った。この結果、品詞だけを用いた係り受けモデルのスクロスエントロピーは 5.3536 であり、形態素クラスタリングの結果を用いた係り受けモデルのスクロスエントロピーは 4.9944 であった。これは、我々の提案する形態素クラスタリングが、モデルを有意に改善することを示す。また、構文解析器を実装し、テストコーパスに対して構文解析を行った。この結果、予測力の改善から容易に推測されることではあるが、形態素クラスタリングは、構文解析の精度も有意に向上させることが分かった。

2 係り受けを用いた確率的言語モデル

この章では、言語学が提案する文節という単位を予測単位とし、係り受けという構造を内包するモデルを提案する。形式的には、このモデルは確率文脈自由文法に属する。文法の終端記号は文節の属性であり、これは内容語の主辞と機能語の主辞の直積として表される。属性からの文節の形態素列の予測には形態素 n-gram モデルを、未知語の文字の予測には文字 n-gram モデルを用いる。

2.1 文節モデル

日本語の文は、1 個以上の内容語と 0 個以上の機能語と句読点からなる文節と呼ばれる単位の連接とみなすことができる。これは、内容語の集合を $Cont$、機能語の集合を $Func$、句読点の集合を $Sign$ とすると以下の式で定義される。

$$Bnst = Cont^{+} Func^{*} \cup Cont^{+} Func^{*} Sign$$

ここで、$+$ と $*$ はそれぞれ正閉包とクリーネ閉包を表す。確率的言語モデルとして、この文節を予測単位とするモデルを構成することができる。このようなモデルとして、第一に考えられるのは、文節 n-gram モデルであろう。しかし、係り受けとして知られる複数の文節間の関係は、必ずしも連続した文節間のみではない。この関係をモ

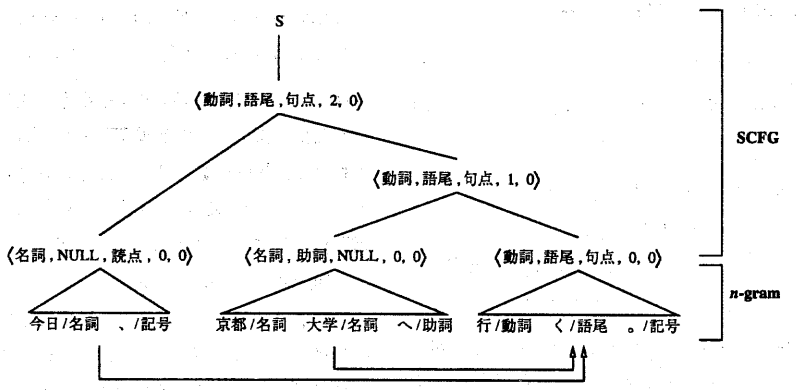

図 1: 文節単位の係り受けモデル

デル化するためには、確率正規言語に属する n-gram モデルでは不十分である。離れた要素間の関係を記述するために、さまざまな文法が提案されている。この章では、これらの文法の一つである確率文脈自由文法 (SCFG) [3, 4] を用いて文節間の係り受けをモデル化する。

確率文脈自由文法でまず問題となるのは、終端記号と非終端記号である。終端記号として、文節をそのまま用いることが考えられるが、この数は非常に大きく、データスパースネスの問題を引き起こす。そこで、クラス n-gram モデル [2] の考え方を応用して、文節を何らかのグループに分類し、これを終端記号とすることが考えられる。この分類には、以下で定義される属性を用いることにする。

$$attrib(b) \quad (1)$$
$$= \langle last(cont(b)),\ last(func(b)),\ last(sign(b)) \rangle$$

関数 $cont$, $func$, $sign$ は、それぞれ文節を引数として、その内容語列、付属語列、句読点を返す。また、関数 $last(m)$ は形態素列 m の最後の形態素の品詞を返す。空形態素列の場合は NULL を返す。文節の属性が与えられると、文節の具体的な内容語列と機能語列は、それぞれ独立に形態素 n-gram モデル [5] によって生成される。

2.2 係り受けのモデル

係り受けとして知られる文節間の関係を記述するために、一般的に認められている複数の係り受け関係の非交差を仮定し、文節の属性を終端記号とする確率文脈自由文法を導入する。日本語の係り受けの性質として、文において前に位置する文節が、後に位置する文節に係ることが分かっている。さらに、係り受け関係をすでに何が係っているかに依存しない二項関係であると仮定する。したがって、これを文脈自由文法の生成規則として表すと $B \Rightarrow AB$ という形式となる。ここで、A は係り文節を表す非終端記号であり、B は受け文節を表す非終端記号である。

非終端記号を終端記号と同じように文節の属性とすることもできるが、付加的な情報との直積とすることで、係り受けの性質を反映するように特殊化することもできる。文の位置という意味で近い文節間の係り受けは、遠い文節間の係り受けよりも高い頻度で生じることが分かっている [8] ので、この性質をモデルに組み込むために、いくつの文節を受けているかを付加的な情報として加えることとした。また、読点を含む文節は、それに先行する文節の大半を受けることが多いことが分かっている。この性質をモデルに組み込むために読点を含む文節を受けた数も付加的な情報として加えることとした。データスパースネスの問題に対処するために、これらの数には上限を設けた。受けた文節の数と受けた読点を含む文節の数をそれぞれ d, v とすると、終端記号の集合 T と非終端記号の集合 V は以下のように表される (図1参照)。

$$T = attrib(b) \times \{0\} \times \{0\}$$
$$V = attrib(b) \times \{1, 2, \cdots d_{max}\} \times \{0, 1, \cdots v_{max}\}$$

ここで、終端記号には係る文節がないという点に注意しなければならない。この結果、生成規則は以下のような形式になる。ただし、開始記号 S からの生成は例外であ

る。また、a は文節の属性を表すとする。

$$S \Rightarrow \langle a,\ d,\ v \rangle \tag{2}$$

$$\langle a_1,\ d_1,\ v_1 \rangle \Rightarrow \langle a_2,\ d_2,\ v_2 \rangle \langle a_3,\ d_3,\ v_3 \rangle \tag{3}$$

$$a_1 = a_3$$

$$d_1 = \min(d_3 + 1,\ d_{max})$$

$$v_1 = \begin{cases} \min(v_3 + 1,\ v_{max}) & \\ & \text{if } sign(a_2) = \text{読点} \\ v_3 & \text{otherwise} \end{cases}$$

ある文の属性列は、開始記号にこれらの生成規則を何回か適用して生成される。各生成規則には確率が付与されており、属性列の生成確率はこれらの積となる。この生成確率は、図1の例では、以下のように計算される。ただし、d_{max} および v_{max} は十分大きいとする。

$$\begin{aligned} & P(\langle \text{名詞, NULL, 読点, 0, 0} \rangle \\ & \quad \langle \text{名詞, 助詞, NULL, 0, 0} \rangle \langle \text{動詞, 語尾, 句点, 0, 0} \rangle) \\ & = P(S \Rightarrow \langle \text{動詞, 語尾, 句点, 2, 0} \rangle) \\ & \quad \times P(\langle \text{動詞, 語尾, 句点, 2, 0} \rangle \\ & \qquad \Rightarrow \langle \text{名詞, NULL, 読点, 0, 0} \rangle \langle \text{動詞, 語尾, 句点, 1, 0} \rangle) \\ & \quad \times P(\langle \text{動詞, 語尾, 句点, 1, 0} \rangle \\ & \qquad \Rightarrow \langle \text{名詞, 助詞, NULL, 0, 0} \rangle \langle \text{動詞, 語尾, 句点, 0, 0} \rangle) \end{aligned}$$

生成規則の確率値は、係り受けが付与されたコーパスからその頻度を計数し、以下の式を用いて最尤推定することで得られる。

$$\begin{aligned} & P(S \Rightarrow \langle a_1,\ d_1,\ v_1 \rangle) \\ & \overset{MLE}{=} \frac{N(S \Rightarrow \langle a_1,\ d_1,\ v_1 \rangle)}{N(S)} \\ & P(\langle a_1,\ d_1,\ v_1 \rangle \Rightarrow \langle a_2,\ d_2,\ v_2 \rangle \langle a_3,\ d_3,\ v_3 \rangle) \\ & \overset{MLE}{=} \frac{N(\langle a_1,\ d_1,\ v_1 \rangle \Rightarrow \langle a_2,\ d_2,\ v_2 \rangle \langle a_3,\ d_3,\ v_3 \rangle)}{N(\langle a_1,\ d_1,\ v_1 \rangle)} \end{aligned}$$

本研究では用いていないが、係り受けが付与されていないコーパスからのパラメータ推定の方法として、Inside-Outside アルゴリズム [9] と呼ばれる方法がある。

2.3 低頻度事象への対処

形態素や文字の n-gram モデルと同様に、確率文脈自由文法にも補間を導入することができる。文法 G_1 と文法 G_2 による生成規則の確率をそれぞれ P_1, P_2 とすると、これらを補間した確率 P は以下の式で与えられる。ただし、$A \in V$ かつ $\alpha \in (V \cup T)^*$ である。

$$\begin{aligned} & P(A \Rightarrow \alpha) = \lambda_1 P_1(A \Rightarrow \alpha) + \lambda_2 P_2(A \Rightarrow \alpha) \\ & \text{ただし } 0 \leq \lambda_j \leq 1\ (j = 1, 2) \text{ かつ } \lambda_1 + \lambda_2 = 1 \end{aligned} \tag{4}$$

文法 G_1 として文法 G_2 よりも凡化レベルの高い文法を選択すれば、文法 G_2 の低頻度事象の問題に対処していることになる。補間係数の値は、形態素 n-gram モデルや文字 n-gram モデルの場合と同じように、Held-out 法や削除補間法 [10] によって求めることができる。

3 形態素クラスタリング

これまでに説明したモデルは、文節の属性として内容語や附属語の品詞を用いていたが、これを形態素やそのクラスに変更することで、予測精度が向上すると考えられる。すでに説明したモデルは品詞というクラスの特別な例に基づいているので、文節の属性として内容語や附属語のクラスを用いるモデルへの変更の必要はなく、単に式 (1) の関数 $last$ を形態素列 $\boldsymbol{m}$ の最後の形態素のクラスを返すように変更すればよい。以下では、クラスタリングの目的関数と探索アルゴリズムについて述べる。

3.1 目的関数

形態素クラスタリングの目的は、クロスエントロピーという観点でより良い言語モデルを構成することである。同様の目的で、形態素 n-gram モデルを改善する報告はすでになされている [6]。この研究と異なるのは、確率的言語モデルだけである。したがって、クラス分類の目的関数は、以下の式のように削除補間を応用することで得られる平均クロスエントロピーである。

$$\overline{H} = \frac{1}{m} \sum_{i=1}^{m} H(L_i, M_i) \tag{5}$$

ここで、M_i は i 番目以外の $m-1$ の部分コーパスから推定された係り受けモデル (補間係数の推定も含む) であり、L_i は i 番目の部分コーパスを表す。

3.2 アルゴリズム

アルゴリズムは、文献 [6] の形態素 n-gram モデルの場合と異なり、トップダウンである。つまり、初期状態では同一の品詞に属する形態素は一つのクラスとなっており、繰り返し部分では図2のように各形態素の分離を試みる。形態素 n-gram の場合と同様に、計算量という観点から最適解を選択するということは不可能なので、貪欲アルゴリズムを用いることにした。このアルゴリズムは図3の通りである。なお、$\overline{H}$ は式 (5) で与えられる平均クロスエントロピーである。

計算量は、二番目の **foreach** での繰り返しの回数は形

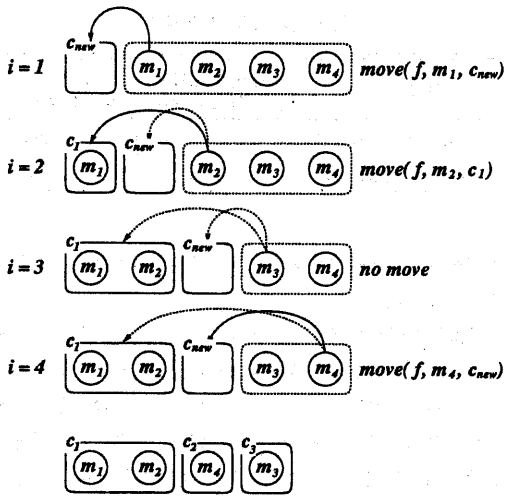

図 2: クラスタリングの概念図

態素数 $|\mathcal{M}|$ に比例し、argmin での繰り返しの回数はクラス数 $|\mathcal{C}|$ に比例するので、全体で $O(|\mathcal{M}| \cdot |\mathcal{C}|)$ である。クラス数 $|\mathcal{C}|$ は、全ての形態素が独立したクラスに分けられる場合に最大 ($|\mathcal{C}| = |\mathcal{M}|$) となり、全ての形態素が同一のクラスとなる場合に最小 ($|\mathcal{C}| = 1$) となる。従って、初期化における全体の計算量は、最良の場合が $O(|\mathcal{M}|)$ であり、最悪の場合が $O(|\mathcal{M}|^2)$ である。ただし、形態素の並べ替えや一番目の foreach の計算量は係数が非常に小さいと考えられるので、考慮に入れていない。

4 構文解析

日本語に対する構文解析とは、日本語の文 (文字列) を入力とし、これを文節に分割すると同時に文節間の係り受け関係を決定する処理である。この章では、これを実現する手法の一つとしての確率的構文解析とその基礎となる係り受けモデルにおける最尤解の探索方法について述べる。

4.1 確率的構文解析

日本語の構文解析は、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字列 $x \in \mathcal{X}^*$ を入力として、これを文節に分割し、それらの分割間の係り受け関係を出力することと定義できる。このとき、出力される文節列の表記の連接は、入力のアルファベット列に等しくなければならない。一般に、これを満たす解は一意ではない。構文解析の問題は、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構文を選択し出力することである。この選択

$\mathcal{M}$ を頻度の降順に並べ $m_1, m_2, \ldots, m_n$ とする
$c_1 := \{m_1, m_2, \ldots, m_n\}$
$C = \{c_1\}$
foreach i $(1, 2, \cdots, n)$
  $f(m_i) := c_1$
foreach i $(1, 2, \cdots, n)$
  $c := \mathbf{argmin}_{c \in C \cup c_{new}} \overline{H}(move(f, m_i, c))$
  if $(\overline{H}(move(f, m_i, c)) < \overline{H}(f))$ **then**
    $f := move(f, m_i, c)$
    if $(c = c_{new})$ **then**
      $C := C \cup \{c_{new}\}$

図 3: クラスタリングのアルゴリズム

の基準としては、文法家が自身の言語直観を頼りにした規則に基づく方法と大量の正解例 (構文解析済みコーパス) からの推定を基準にする方法がある。以下では、後者の一つである確率的構文解析について説明する。

確率的構文解析器は、係り受け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 x に対する確率最大の構文木 (図 1参照) を計算し出力する。これは、以下の式で表される。ただし、$w(T)$ は構文木 T の文節列の表記の連接を表す。

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

この式の最後の $P(T)$ が係り受け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、第2章で説明した品詞係り受けモデルやクラス係り受けモデルを用いることができる。

4.2 解探索のアルゴリズム

構文解析に用いる確率的言語モデルは、式 (2) の開始記号からの導出を除いて、式 (3) のような Chomsky 標準形 [11, 12] に制限されている。したがって、解探索のアルゴリズムには動的計画法の一種である CKY 法を、確率文脈自由文法に拡張したアルゴリズム [13] を用いることができる。ただし、例外である開始記号からの導出確率を最後

表 1: 実験に用いたコーパス

用途	文数	文節数	形態素数	文字数
学習	174,524	1,610,832	4,251,085	6,724,609
評価	19,397	178,415	471,189	744,332

に掛ける必要がある。CKY 法による文脈自由文法の構文解析の計算量は、入力の記号数を n として $O(n^3)$ である。確率を扱う拡張は、CKY 表に非終端記号とともにそこから部分文字列が生成される確率を記憶しておくことで実装されるので、計算量には影響しない。したがって、確率的構文解析の計算量は $O(n^3)$ である。文全体の生成確率を計算するためには、各終端記号の生成確率の初期値として、文節の属性から文節が生成される確率を与えておけば良い。

5 評価

以上で説明した品詞係り受けモデルとクラスタリングの結果を用いたクラス係り受けモデルを構成し、それぞれの予測力 (クロスエントロピー) を評価した。さらに、それぞれのモデルに対して、文節列からの最尤の構文木 (係り受け) を探索するアルゴリズムを実装し、構文解析の精度を評価した。この節では、この結果を提示し、それに対する考察を述べる。

5.1 実験の条件

実験には EDR コーパス [7] を用いた。これを 10 個に分割し、そのうちの 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。各コーパスの大きさを表 1に掲げた。EDR コーパスにあらかじめ付加された構造を係り受けに変換できない文が存在したので、コーパスの大きさは少し小さくなっている。アルファベット数は、文献 [5] と同じ 6,879 としている。モデルの説明では可変であった受けた文節の数の上限と受けた読点を含む文節の数の上限は共に 1 とした ($d_{max} = v_{max} = 1$)。これらを平均クロスエントロピーを基準として学習することも可能であるが、以下の実験では固定である。

5.2 予測力の評価

予測力の評価を目的として、推定されたモデルによるテストコーパスのクロスエントロピーを計算した。テストコーパスの係り受けは、コーパスにあらかじめ付加された構造を用いた。したがって、テストコーパスに含まれる文字列の出現確率は、その文字列のすべての生成方法による確率を合計した値ではなく、コーパスに示された生成方法のみによる値である。

表 2: 各モデルの予測力

言語モデル	非終端記号数 ＋終端記号数	クロス エントロピー
品詞係り受けモデル	576	5.3536
クラス係り受けモデル	10,752	4.9944

品詞係り受けモデルとクラス係り受けモデルを比較するために、これらを同じ学習コーパスから構成し、同じテストコーパスに対してクロスエントロピーを計算した。それぞれの言語モデルの構成の手順は以下の通りである。なお補間に用いた確率文脈自由文法は、生成規則の確率が等確率分布であること以外は係り受けモデルの文法と同じである。

- 品詞係り受けモデル
  1. 削除補間により式 (4) の補間係数を推定
  2. すべての学習コーパスを対象に生成規則の頻度を計数
- クラス係り受けモデル
  1. 削除補間により式 (4) の補間係数を推定
  2. 前章で述べた方法でクラス関数を推定
  3. 削除補間により式 (4) の補間係数を再推定
  4. すべての学習コーパスを対象に生成規則の頻度を計数

各モデルに含まれる文節モデルと未知語モデルは、文献 [5] と同じ形態素 2-gram モデルと文字 2-gram モデルである。各文節の主辞の形態素の予測までが係り受けモデルに含まれるとすれば、この部分のクロスエントロピーへの寄与は一定である。

表 2は各モデルのテストコーパスのクロスエントロピーである。クラス係り受けモデルは、品詞係り受けモデルよりも低いクロスエントロピーとなっている。このことから、提案手法による形態素クラスタリングは係り受けモデルにも有効であり、これにより推定されたクラスによるクラス係り受けモデルが、品詞係り受けモデルよりも、予測力という点で良い言語モデルであることが分かる。

付録 Aは得られたクラスタの例である。多くのクラスタが**クラスタ 1** のように我々の言語直観に照らし合わせて、納得できるクラスタであった。一方、**クラスタ 2** のように我々の言語直観に合致しないクラスタもあった。こ

表 3: クロスエントロピーの内訳

モデルの部分	クロスエントロピー
文節の主辞の予測	3.6652
形態素予測	0.8700
未知語の文字予測	0.4592
合計	4.9944

れは、我々が行なった形態素クラスタリングは、クラス係り受けモデルの改善という観点からのクラスタリングであることと、得られた形態素の分類が準最適解であることを考えると特に不自然ではないであろう。

クロスエントロピーの分枝性 [14] から、代入によって多段になっている確率的モデルによるクロスエントロピーは、各部分の寄与に分解されるので、独立に計算することができる。係り受けモデルは、文節の主辞以外の形態素列を予測する部分と、未知語の文字列を予測する部分に分解できる。表 3は、このようにして計算したクロスエントロピーの各部分の内訳である。この結果を見ると、テストコーパスに対するクロスエントロピーには、文節の主辞を予測する部分がかなり大きく寄与していることが分かる。よって、短期的により良い言語モデルを構成するためには、この部分を改良することが近道であると考えられる。形態素クラスタリングによるクラス係り受けモデルの構成は、この部分を有意に改善している。長期的には、文節の主辞以外の形態素列を予測する部分や未知語モデルを改善することが望ましい。

5.3 構文解析の精度の評価

係り受けを用いた確率的言語モデルの応用の例として構文解析を行った。我々の提案するモデルは、未知語から係り受けまでを一貫してモデル化しているので、未知語処理や形態素解析なども含めた、文字列からの構文解析を一括して行うことができる。しかし、この結果を評価することは容易ではない。よって、この節で述べる実験では、文節に分割された入力を仮定している。

我々が用いた評価基準は、文節単位の係り受けの正解率である。ただし、最後の文節は係り先を持たず、その直前の文節は必ず最後の文節に係るので、これらを評価の対象としていない。例として、コーパスの内容と解析結果が表4のような場合を考える。この例では、4つの文節がある。このうち、係り先があらかじめコーパスに与えられた

表 4: 評価基準の説明のための例

文節番号	係り先(正解)	係り先(結果)	文節
1	2	4	今日/名詞 と/助詞
2	4	4	明日/名詞 、/記号
3	4	4	京都/名詞 大学/名詞 へ/助詞
4	–	–	行/動詞 く/語尾 。/記号

表 5: 各係り受けモデルによる解析精度

言語モデル	クロスエントロピー	解析精度
品詞係り受けモデル	5.3536	68.77%
クラス係り受けモデル	4.9944	81.96%
無条件に次の文節を選択	–	53.10%

正解と一致した文節は、文節 2 と文節 3 である。文節 3 は、評価の対象ではないので、2 つのうち 1 つが正解であったことになる。よって、この例の正解率は 1/2 となる。

表 5は、品詞係り受けモデルとクラス係り受けモデルによるクロスエントロピーと構文解析の精度である。この結果から、形態素クラスタリングは、構文解析の精度を向上させることが分かる。これは、形態素解析の場合と同様に、クロスエントロピーの減少から予測される通りの結果である。

図 4は学習コーパスの大きさとクロスエントロピー及び解析精度の関係である。クロスエントロピーの下限は日本語のエントロピーである。これを考慮に入れると、クロスエントロピーはどちらのモデルでも十分な減少傾向にあるといえる。解析精度の上限はコーパスの誤りによって規定される。これを考慮にいれると、クラス係り受けモデルは学習コーパスの増加による精度向上が見込めるが、品詞係り受けモデルではそれがあまり見込めないことが分かる。しかしながら、クラス係り受けモデルの解析精度の増加量は、あまり大きくない。このことから、学習コーパスの増加による精度向上が見込めるような特殊化の可能性があることが分かる。これは、今後の課題である。その体系的方法として、様々な特殊化の方法を列挙して、学習コーパスの平均クロスエントロピーを用いて、真に有効と推測される特殊化を自動的に選択することが最短距離であろう。

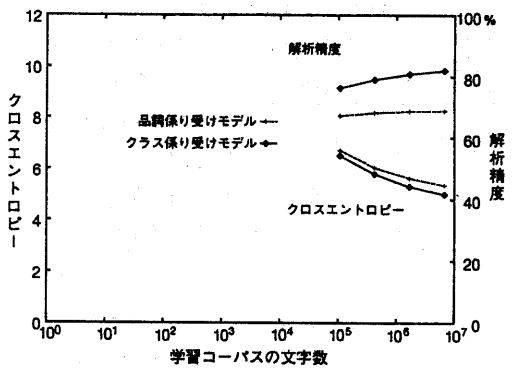

図 4: クロスエントロピーと解析精度の関係

6 おわりに

本論文では、文節の属性を利用した係り受けモデルを提案した。このモデルにより、未知語から係り受けまでを一貫してモデル化しているので、係り受けまでの言語現象を考慮にいれた音声認識や読み推定などを同時に行うことができる。次に、モデルの改善方法として、このモデルに対して準最適なクラス分類を求めるアルゴリズムについて述べた。このアルゴリズムは、クラス推定のためのコーパスを係り受けモデルの推定用のコーパスとは別に用意するというアイデアに基づいている。実験の結果、クラスタリングによる予測力の向上が観測された。これは、上述の応用を行った場合の精度向上につながる。さらに、モデルの応用の一例として、構文解析器について述べた。クロスエントロピーを基準とした形態素クラスタリングにより、構文解析の精度の向上が観測された。

謝辞

本研究を進めるに過程で、示唆に富んだコメントを頂いた日本アイ・ビー・エム東京基礎研究所の西村雅史氏と伊東伸泰氏に心から感謝する。また、本論文で報告している研究は文部省科学研究費補助金 (課題番号 00093069) の助成を受けている。ここに感謝の意を表する。

参考文献

[1] C. E. Shannon. Prediction and Entropy of Printed English. *Bell System Technical Journal*, Vol. 30, pp. 50–64, 1951.

[2] Peter F. Brown, Vincent J. Della Pietra, Peter V. deSouza, Jenifer C. Lai, and Robert L. Mercer. Class-Based *n*-gram Models of Natural Language. *Computational Linguistics*, Vol. 18, No. 4, pp. 467–479, 1992.

[3] King Sun Fu. *Syntactic Methods in Pattern Recognition*, Vol. 12 of *Matehmatics in Science and Engineering*. ACCADEMIC PRESS, 1974.

[4] C. S. Wetherell. Probabilistic Languages: A Review and Some Open Questions. *ACM Computing Surveys*, Vol. 12, No. 4, pp. 361–379, 1980.

[5] 森信介, 山地治. 日本語の情報量の上限の推定. 情報処理学会論文誌, Vol. 38, No. 11, 1997.

[6] 森信介, 西村雅史, 伊東伸泰. クラスに基づく言語モデルのための単語クラスタリング. 情報処理学会論文誌, Vol. 38, No. 11, 1997.

[7] 日本電子化辞書研究所. EDR 電子化辞書仕様説明書, 1993.

[8] MARUYAMA Hiroshi and OGINO Shiho. A Statistical Property of Japanese Phrase-to-Phrase Modifications. *Mathematical Linguistics*, Vol. 18, No. 7, pp. 348–352, 1992.

[9] John D. Lafferty. A Derivation of the Inside-Outside Algorithm from the EM Algorithm. Technical report, IBM T. J. Watson Research Center, 1993.

[10] Fredelick Jelinek, Robert L. Mercer, and Salim Roukos. Principles of Lexical Language Modeling for Speech Recognition. In *Advances in Speech Signal Processing*, chapter 21, pp. 651–699. Dekker, 1991.

[11] John E. Hopcroft and Jeffrey D. Ulman. オートマトン言語理論 計算論 I. サイエンス社, 1984.

[12] John E. Hopcroft and Jeffrey D. Ulman. オートマトン言語理論 計算論 II. サイエンス社, 1984.

[13] 北研二, 中村哲, 永田昌明. 音声言語処理. 森北出版, 1996.

[14] 堀部安一. 情報エントロピー論. 森北出版, 第 2 版, 1997.

A 得られたクラスタの例

クラスタ 1

[る / 語尾 た / 語尾 した / 語尾 える / 語尾 ます / 語尾]

クラスタ 2

[ば / 助詞 ても / 助詞 たり / 助詞 ながら / 助詞 だり / 助詞 ども / 助詞 たら / 助詞 ところで / 助詞 きゃ / 助詞 なら / 助詞 ては / 助詞 ったら / 助詞 やいなや / 助詞]