形態素単位の係り受けによる構文解析

森 信介 西村 雅史 伊東 伸泰 荻野 紫穂 渡辺 日出雄
日本IBM 東京基礎研究所
〒242-8502 神奈川県大和市下鶴間 1623-14

1 はじめに

本論文では、日本語を対象言語とする生成的な確率的言語モデルとそれに基づく構文解析器について述べる。このモデルでは、文は形態素の列とみなされ、それらの間での係り受け関係を確率的文脈自由文法で記述される。このモデルを用いた構文解析器を作成し、日本経済新聞からならるコーパスからパラメータを推定し、形態素列を入力とする構文解析実験を行なった結果、係り受け単位で88.6%の解析精度を得た。また、文字列を入力とする構文解析結果を形態素解析の結果としてして評価すると、形態素2-gramモデルによる結果よりも良く、構文情報が形態素解析の精度を向上させることが実験的に示された。

2 係り受けに基づく確率的言語モデル

この節では、我々が提案する係り受けに基づく確率的言語モデルについて述べる。このモデルでは、文は形態素の列 ($m = m_1 m_2 \cdots m_n$) とみなされ、それらの間での係り受け関係を確率文脈自由文法 (SCFG) で記述する。終端記号は形態素であり、非終端記号は形態素と受ける形態素の数との組である。

2.1 文のモデル

形態素間の係り受けとして知られる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、形態素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、文中で前に位置する形態素が、後に位置する形態素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と、係り受け関係を表す導出規則は $B \Rightarrow AB$ という形式となる。ここで、A は係り形態素を表す終端記号または非終端記号であり、B は受け形態素を表す終端記号または非終端記号である。

非終端記号を終端記号と同じように形態素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、モデルが係り受けの性質をより反映するように特殊化することもできる。我々が提案するモデルでは、いくつの形態素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。これは、文中の出現位置が近い形態素間の係り受けは、遠い形態素間の係り受けよりも高い頻度で生じる [1] ので、この性質をモデルに組み込むこためである。データスパースネスの問題に対処するために、受けている形態素の数に上限を設けた。形式的定義を簡便にするために、終端記号も形態素と受けている形態素の数との組とする。図1は、ある文の係り受け構造とそれ対応する文脈自由文法の導出木である。終端記号には係る形態素がないという点に注意しなければならない。

以上で述べた確率文脈自由文法の終端記号の集合 T と非終端記号の集合 V は、受けている形態素の数を d、その上限を d_{max} として、以下のように表される。

$$T = \mathcal{M} \times \{0\}$$
$$V = \mathcal{M} \times \{1, 2, \cdots, d_{max}\}$$

ここで、$\mathcal{M}$ は形態素の集合を表す。さらに、導出規則は以下のような形式になる。

$$(m_1, d_1) \Rightarrow (m_2, d_2)(m_3, d_3)$$

ここで、右辺の右側の形態素が主辞であるから $m_1 = m_3$ であり、1つの形態素を受けるので $d_1 = \min(d_3 + 1, d_{max})$ である。また、開始記号は $(BT, 1)$ である。

ある文は、開始記号にこれらの導出規則を何回か適用した結果得られる形態素列として生成される。各導出規則には確率が付与されており、形態素列の生成確率はこれらの積となる。

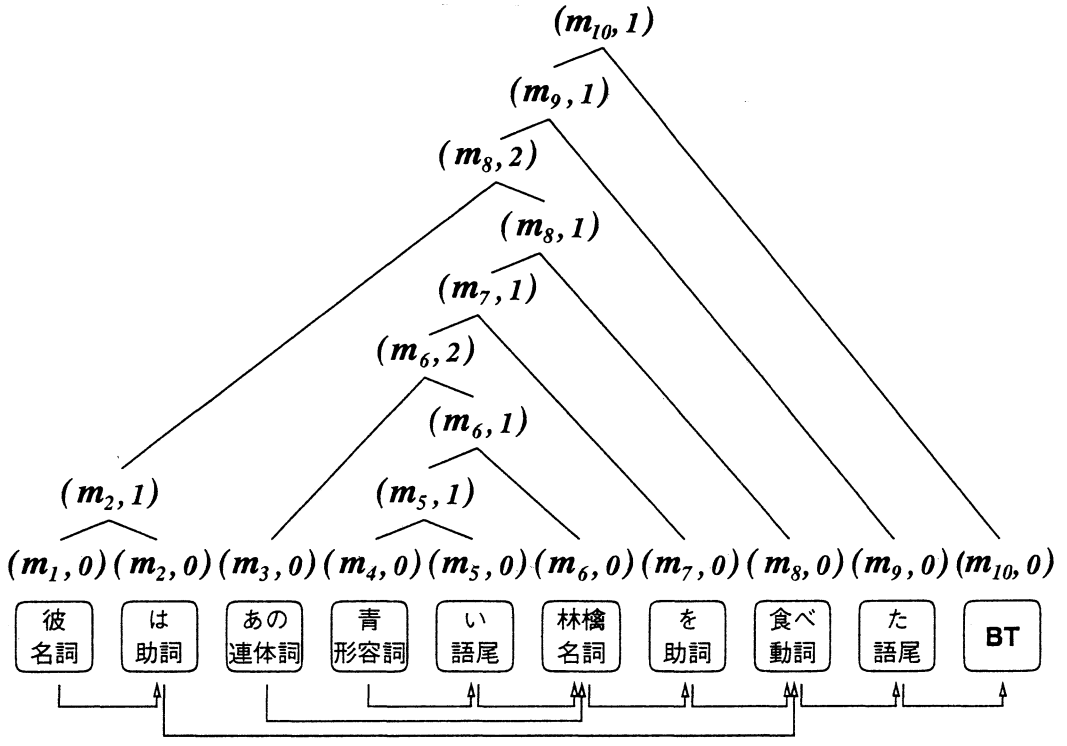

図 1: 文脈自由文法による導出

2.2 低頻度事象への対処

低頻度事象に対処するため補間を導入する。導出規則から分かるように、受けている形態素の数を無視すれば、修飾語が主辞から予測されるという意味で形態素 2-gram に類似している。したがって、形態素 n-gram モデルにおいてなされるように、より低次のモデルと補間することで低頻度事象に対処することができる。具体的には、以下の式が示すように形態素 1-gram モデルと形態素 0-gram モデル (一様分布) に類似するモデルとの補間を行なう。

$$P(B \Rightarrow AB) \quad (1)$$
$$= \lambda_2 P_2(AB|B) + \lambda_1 P_1(A) + \lambda_0 \frac{1}{|T| + |V|}$$

ただし、$0 \leq \lambda_j \leq 1$ $(j = 1, 2, 3)$ かつ $\lambda_1 + \lambda_2 + \lambda_3 = 1$ である。これら補間係数の値は、n-gram モデルの場合と同じように、削除補間法 [2] によって求めることができる。

2.3 パラメータ推定

各導出規則の確率値は、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。

$$P((m_1, d_1) \Rightarrow (m_2, d_2)(m_3, d_3)) \overset{MLE}{=} \frac{f((m_1, d_1) \Rightarrow (m_2, d_2)(m_3, d_3))}{f((m_1, d_1))}$$

ここで、$f(x)$ は事象 x の学習コーパスでの頻度を表す。

2.4 未知語モデル

我々の言語モデルが未知形態素を扱うことができるように、文字 2-gram からなる未知語モデルを付加した。文中の形態素が語彙にない場合、その品詞を終端記号として確率文脈自由文法により生成し、次いで未知語モデルが品詞から文字列 $(x = x_1 x_2 \cdots x_m)$ を以下

の式を用いて生成する。

$$P(\boldsymbol{x}|POS) = \prod_{i=1}^{m+1} P_{POS}(x_i|x_{i-1})$$

ここで、$x_0 = x_{m+1} = \mathrm{BT}$ は形態素の境界を示す特別の記号である。これにより、すべての可能な文字列に対する確率値の合計が1になる。

削除補間による補間係数の推定では、学習コーパスを複数の部分コーパスに分割する。実験では、語彙は部分コーパスの2つ以上に出現する形態素とした。未知語モデルのパラメータは、1つの部分コーパスにしか出現しない形態素から、品詞毎に以下の式を用いて推定される。

$$P_{\mathrm{POS}}(x_i|x_{i-1}) \stackrel{\mathrm{MLE}}{=} \frac{f_{\mathrm{POS}}(x_i, x_{i-1})}{f_{\mathrm{POS}}(x_{i-1})}$$

文字2-gramモデルも文字1-gramモデルと文字0-gramモデル(一様分布)と補間する。補間係数は削除補間法[2]により推定する。

3 構文解析

一般的に、構文解析器は形態素列に分割された文を入力とし、その構造を出力する。これ対して、我々が提案する構文解析器は、文字列を入力として、単語への分割と品詞の付与を構文解析と同時に行なうことができる。この節では、前節で述べた言語モデルに基づく構文解析器について説明する。

3.1 確率的構文解析器

確率的言語モデルに基づく構文解析器は、文字列($\boldsymbol{x}$)を与えられると、確率が最大となる構造を以下の式に従って出力する。

$$\begin{aligned}
\hat{T} &= \operatorname*{\mathbf{argmax}}_{\boldsymbol{w}(T)=\boldsymbol{x}} P(T|\boldsymbol{x}) \\
&= \operatorname*{\mathbf{argmax}}_{\boldsymbol{w}(T)=\boldsymbol{x}} P(T|\boldsymbol{x})P(\boldsymbol{x}) \\
&= \operatorname*{\mathbf{argmax}}_{\boldsymbol{w}(T)=\boldsymbol{x}} P(\boldsymbol{x}|T)P(T) \quad (\because \text{Bayes' formula}) \\
&= \operatorname*{\mathbf{argmax}}_{\boldsymbol{w}(T)=\boldsymbol{x}} P(T) \quad (\because P(\boldsymbol{x}|T) = 1)
\end{aligned}$$

ここで、$\boldsymbol{w}(T)$ は導出木 T の形態素列の文字列の連接である。最後の行の $P(T)$ は確率的言語モデルである。これは、我々の構文解析器では、2節で述べた確率文脈自由文法によって計算される導出木 T の確率である。

表1: コーパス

	文数	形態素数	文字数
学習	1,072	30,292	46,212
テスト	119	3,268	4,909

3.2 解探索のアルゴリズム

構文解析に用いる確率文脈自由文法の導出規則は、Chomsky標準形に制限されている。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[3]を用いることができる。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。確率を扱うための拡張は、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現されるので、計算量には影響しない。

4 評価

2節で述べた、確率文脈自由文法に基づくモデルを構成し、3節で説明した解探索アルゴリズムを用いる構文解析器を作成し、テストコーパスに対する構文解析の実験を行なった。この節では、この結果を提示し、その評価を行なう。

4.1 実験の条件

実験には、日本経済新聞の記事からなるコーパス(表1参照)を用いた。各文は、形態素に分割され、構文構造が付与されている。また、モデルの説明においては可変であった、受けている形態素の数の上限を9とした($d_{max} = 9$)。

4.2 評価

形態素単位の係り受けモデルによる構文解析器を作成し、精度の評価を行なった。文字列に対する構文解析も可能であるが、形態素への分割の誤りが生じ、結果の評価が困難であるため、構文解析の精度の評価には形態素列を入力とした場合の出力を用いた。精度は、以下の式で示されるように、推定した係り受け関係の数に対する、正しい係り受け関係の割合である。ここで、正しい係り受け関係とは、コーパスに付与された係り受け関係と同じでることを意味する。

$$\text{解析精度} = \frac{\text{係り先が正しい形態素の数}}{\text{形態素の数}}$$

係り先がない最後の形態素と係り先が明らかな最後から2番目の形態素は、評価の対象としていない。

表2は、文脈自由文法に基づくモデルによると、すべての形態素が次の形態素に係るとするベースラインとの構文解析の精度である。学習コーパスのサイズが不十分でるが、ベースラインの誤りの約半数が訂正されている。今後、学習コーパスを増やすことによって精度向上が期待できる。なお、従来の文節を単位とする係り受け解析と比較に関しては、名詞が直後の助詞に係る場合などの容易な係り受けを含む一方、複合語の解析などの困難な係り受けを含むという相違があり、精度の比較は容易ではない。

また、文字列に対する構文解析の結果を形態素解析の結果とみなして、形態素単位での再現率と適合率 [4] を形態素 2-gram モデルと比較した。

表3は、提案モデルと形態素 2-gram モデルに基づく形態素解析器の結果である。再現率と適合率の双方において提案モデルの精度は、形態素 2-gram モデルによる精度よりもよく、構文情報が形態素解析の精度を向上させることが実験的に示された[1]。これは、形態素解析と構文解析を同時に行なうことで実現されており、生成的な構文モデルの長所の1つである。

5 結論

本論文では、形態素を単位とする係り受け構造に基づく確率的言語モデルについて述べた。形態素間の関係は確率文脈自由文法で記述される。このモデルを用いた構文解析器を作成し、日本経済新聞からならる1,072文のコーパスからパラメータを推定し、119文に対して構文解析実験を行なった結果、係り受け単位での精度は88.6%であった。また、文字列を入力とする構文解析結果を形態素解析の結果としてして評価すると、形態素 2-gram モデルによる結果よりもよく、構文情報が形態素解析の精度を向上させることが実験的に示された。

[1] 東京工業大学の白井清昭氏は ATR 対話コーパスの例文500文(平均単語数10程度)に対して同様の結果を得ている。

表 2: 構文解析の精度

言語モデル	解析精度
確率文脈自由文法	88.6%
ベースライン *	80.0%

* それぞれの形態素は次の形態素に係るとする

表 3: 形態素解析の精度

言語モデル	再現率	適合率
確率文脈自由文法	89.9% $\left(\frac{2937}{3268}\right)$	90.5% $\left(\frac{2937}{3247}\right)$
形態素 2-gram	89.4% $\left(\frac{2920}{3268}\right)$	90.2% $\left(\frac{2920}{3237}\right)$

参考文献

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

[2] 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.

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

[4] 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.