後方文脈を考慮した係り受けモデル

内元 清貴[†] 村田 真樹[†]

関根 聡^{††} 井佐原 均[†]

係り受け解析は日本語解析の重要な基本技術の一つとして認識されている。依存文法に基づく日本語係り受け解析では、文を文節に分割した後、それぞれの文節がどの文節に係りやすいかを表す係り受け行列を作成し、一文全体が最適な係り受け関係になるようにそれぞれの係り受けを決定する。本論文ではそのうち、係り受け行列の各要素の値を計算するためのモデルについて述べる。アプローチとしては、主にルールベースによる方法と統計的手法の二つのものがあるが、我々は利用可能なコーパスが増加してきたこと、規則の変更に伴うコストなどを考慮して、統計的手法をとっている。統計的手法では行列の各要素の値は確率値として計算される。これまでよく用いられていたモデル(旧モデル)では、その確率値を計算する際に、着目している二つの文節に係るか係らないかということのみを考慮していた。本論文では、着目している二つの文節(前文節と後文節)だけを考慮するのではなく、前文節と前文節より文末側のすべての文節との関係(後方文脈)を考慮するモデルを提案する。このモデルを ME(最大エントロピー)に基づくモデルとして実装した場合、旧モデルを同じく ME に基づくモデルとして実装した場合に比べて、京大コーパスに対する実験で、全く同じ素性を用いているにもかかわらず係り受け単位で 1% 程度高い精度 (88%) が得られた。

キーワード: 係り受け解析, 構文解析, 最大エントロピー (ME) モデル, 後方文脈

Dependency Model Using Posterior Context

KIYOTAKA UCHIMOTO[†], MASAKI MURATA[†], SATOSHI SEKINE^{††}
and HITOSHI ISAHARA[†]

Dependency structure analysis is one of the basic techniques in Japanese sentence analysis, and the Japanese dependency structure is usually represented by relationships between phrasal units called ‘bunsetsu.’ This analysis is a two-step procedure, and the first step is to prepare a dependency matrix in which each element represents how likely it is that one bunsetsu depends on another. The second step of the analysis finds an optimal set of dependencies for the entire sentence. In this paper we discuss a model used in the first step, a model for estimating dependency likelihood. There are two approaches to estimating the dependency likelihood: rule-based, and statistical. We take the statistical approach because electrically available corpora are getting large, and changing hand-crafted rules is costly. In our approach the value of each element in a dependency matrix is an estimated probability. A statistical model (here called the “old model”) that considers only the relationship between two bunsetsu when estimating those probabilities was earlier proposed. In this paper we propose a new model that estimates dependency likelihood by considering not only the relationship between two bunsetsu but also the relationship between the left bunsetsu and all of the bunsetsus to its right. Our implementation of this model

is based on the ME (maximum entropy) model. When tested with the Kyoto University corpus the dependency accuracy obtained with this model was 88%, which is about 1% higher than that obtained with the old model even using exactly the same features.

KeyWords: *dependency analysis, parser, maximum entropy (ME) model, posterior context*

1 はじめに

係り受け解析は日本語解析の重要な基本技術の一つとして認識されている。係り受け解析には、日本語が語順の自由度が高く省略の多い言語であることを考慮して依存文法 (dependency grammar) を仮定するのが有効である。依存文法に基づく日本語係り受け解析では、文を文節に分割した後、それぞれの文節がどの文節に係りやすいかを表す係り受け行列を作成し、一文全体が最適な係り受け関係になるようにそれぞれの係り受けを決定する。

依存文法による解析には、主にルールベースによる方法と統計的手法の二つのアプローチがある。ルールベースによる方法では、二文節間の係りやすさを決める規則を人間が作成する (黒橋, 長尾 1992; 白井, 池原, 横尾, 木村 1995)。一方, 統計的手法では, コーパスから統計的に学習したモデルをもとに二文節間の係りやすさを数値化して表す (Collins 1996; 藤尾, 松本 1997; 春野, 白井, 大山 1998; 江原 1998; 白井, 乾, 徳永, 田中 1998)。

我々は、ルールベースによる方法ではメンテナンスのコストが大きいこと、また統計的手法で利用可能なコーパスが増加してきたことなどを考慮し、係り受け解析に統計的手法を採用することにした。統計的手法では二文節間の係りやすさを確率値として計算する。その確率のことを係り受け確率と呼ぶ。これまでよく用いられていたモデル (旧モデル) では、係り受け確率を計算する際に、着目している二つの文節に係るか係らないかということのみを考慮していた。本論文では、着目している二つの文節 (前文節と後文節) だけを考慮するのではなく、前文節と前文節より文末側のすべての文節との関係 (後方文脈) を考慮するモデルを提案する。このモデルは以下の二つの特徴を持つ。

- (1) 二つの文節 (前文節と後文節) 間の関係を、「間」 (前文節が二文節の間の文節に係る) か「係る」 (前文節が後文節に係る) か「越える」 (前文節が後文節を越えてより文末側の文節に係る) かの三カテゴリとして学習する。 (旧モデルでは二文節が「係る」か「係らないか」の二カテゴリとして学習していた。)
- (2) 着目している二つの文節の係り受け確率を求める際に、その二文節に対しては「係る」確率、二文節の間の文節に対しては前文節がその文節を越えて後文節に係る確率 (「越える」の確率)、後文節より文末側の文節に対しては前文節がその文節との間にある後文節

† 郵政省通信総合研究所, Communications Research Laboratory, Ministry of Posts and Telecommunications
†† ニューヨーク大学 コンピュータサイエンス学科, Computer Science Department, New York University

に係る確率 (「間」の確率) をそれぞれ計算し, それらをすべて掛け合わせた確率値を用いて係り受け確率を求める. (旧モデルでは, 着目している二文節に係る確率を計算し, 係り受け確率としていた.)

このモデルを ME(最大エントロピー) に基づくモデルとして実装した場合, 旧モデルを同じく ME に基づくモデルとして実装した場合に比べて, 京大コーパスに対する実験で, 全く同じ素性を用いているにもかかわらず係り受け単位で 1% 程度高い精度 (88%) が得られた.

2 係り受け確率モデル

統計的日本語係り受け解析では, 二文節間の係りやすさは確率値で表される. この確率値は係り受け確率モデルから計算される.

2.1 係り受け確率モデル (旧モデル)

この節ではこれまでに依存文法に基づく係り受け解析によく用いられているモデル (Collins 1996; 藤尾, 松本 1997; 春野他 1998; 内元, 関根, 井佐原 1999) について説明する.

入力文 S が与えられると, S は n 個の文節 $b_1, \dots, b_n$ に一意に分割されると仮定し, S をそれらの順序付き集合 $B = \{b_1, \dots, b_n\}$ で表す. そして, 文全体の係り受け関係 D はそれぞれの文節 $b_i (i = 1, \dots, n-1)$ を係り元の文節とする係り受け関係 D_i の順序付き集合 $D = \{D_1, \dots, D_{n-1}\}$ で表されると仮定する. さらに文節の集合 B が決まると, それぞれ文節 $b_i (1 \leq i \leq n-1)$ と文節 $b_m (m > i, 2 \leq m \leq n)$ に関して観測される素性 $F_{i,m}$ が一意に決まると仮定し, 文節の集合 B を素性の集合 F

$$F = \{F_{1,2}, F_{1,3}, \dots, F_{i,m}, \dots, F_{n-1,n}\} \quad (1)$$

で表す. 統計的係り受け解析とは, S が与えられたときに文全体の係り受けが D となる確率 $P(D|S)$ が最も高くなるものを全体の係り受け関係とする処理のことである. つまり,

$$\begin{aligned} D_{best} &= \operatorname{argmax}_D P(D|S) \\ &= \operatorname{argmax}_D P(D|B) \\ &= \operatorname{argmax}_D P(D|F) \end{aligned} \quad (2)$$

となるような D_{best} を求めることに相当する.

日本語の係り受けには, 主に以下の特徴があるとされている.

- (i) 係り受けは前方から後方に向いている. (後方修飾)
- (ii) 係り受け関係は交差しない. (非交差条件)
- (iii) 係り要素は受け要素を一つだけ持つ.

以降では、これらの特徴を満たすような D_{best} を求めることを考える。まず、式 (2) の $P(D|F)$ は以下のように変形できる。

$$\begin{aligned} P(D|F) &= P(D_1, \dots, D_{n-1}|F) \\ &= P(D_{n-1}|F) \times P(D_{n-2}|D_{n-1}, F) \times P(D_{n-3}|D_{n-2}, D_{n-1}, F) \\ &\quad \times \dots \times P(D_1|D_2, \dots, D_{n-1}, F) \end{aligned} \quad (3)$$

この式で各々の係り受けつまり $D_1, \dots, D_{n-1}$ が独立であると仮定すると、 $P(D|F)$ は以下のよう
にそれぞれの文節に対する係り受けの確率の積で表せる。

$$P(D|F) = \prod_{i=1}^{n-1} P(D_i|F) \quad (4)$$

ここで、 $D_{i,i+j}$ を後で定義するように文節 b_i と文節 b_{i+j} の間の関係を表すフラグとし、 D_i を
文節 b_i と文節 b_{i+j} ($1 \leq j \leq n-i$) との間の関係の集合として、以下のように表す。ここでは、
上述の (i) の特徴を仮定している。

$$D_i = \{D_{i,i+1}, D_{i,i+2}, \dots, D_{i,n}\} \quad (5)$$

すると、式 (4) から以下の式が導ける。

$$P(D|F) = \prod_{i=1}^{n-1} P(D_{i,i+1}, \dots, D_{i,n}|F) \quad (6)$$

旧モデルでは、 $D_{i,i+j}$ として文節 b_i が文節 b_{i+j} に係るか否かの 1, 0 の二値をとると仮定して
いた。文節 b_i と係り受けの関係にある文節が文節 b_i の次から数えて $dep(i)$ ($1 \leq dep(i) \leq n-i$)
番目の係り先の候補であるとき、上述の (iii) の特徴、つまり係り要素は受け要素を一つだけ持
つということを仮定すると、

$$D_{i,i+l} = \begin{cases} 0 & (l \neq dep(i), 1 \leq l \leq n-i) \\ 1 & (l = dep(i)) \end{cases}$$

となる。よって、式 (6) は以下のように変形できる。

$$\begin{aligned} P(D|F) &= \prod_{i=1}^{n-1} P(D_{i,i+dep(i)} = 1|F) \times P(D_{i,i+1} = 0|D_{i,i+dep(i)} = 1, F) \\ &\quad \times P(D_{i,i+2} = 0|D_{i,i+1} = 0, D_{i,i+dep(i)} = 1, F) \\ &\quad \times \dots \times P(D_{i,n} = 0|D_{i,n-1} = 0, \dots, D_{i,i+1} = 0, D_{i,i+dep(i)} = 1, F) \end{aligned} \quad (7)$$

$D_{i,i+dep(i)} = 1$ のとき必ず $D_{i,i+j} = 0$ ($j \neq dep(i)$) となるので、

$$\begin{aligned}
P(D_{i,i+1} = 0 | D_{i,i+dep(i)} = 1, F) &= 1 \\
P(D_{i,i+2} = 0 | D_{i,i+1} = 0, D_{i,i+dep(i)} = 1, F) &= 1 \\
&\vdots \\
P(D_{i,n} = 0 | D_{i,n-1} = 0, \dots, D_{i,i+1} = 0, D_{i,i+dep(i)} = 1, F) &= 1
\end{aligned}$$

となり, 式 (7) は次のように表せる.

$$P(D|F) = \prod_{i=1}^{n-1} P(D_{i,i+dep(i)} = 1|F) \quad (8)$$

さらに, $F_{i,m}$ はそれぞれ独立で, 文節 b_i と文節 $b_{i+dep(i)}$ との関係 $D_{i,i+dep(i)}$ は $F_{i,i+dep(i)}$ のみによって決まり, 他の $F_{i,i+j} (j \neq dep(i))$ とは独立であると仮定する. すると, 式 (8) は, 以下のように変形できる.

$$P(D|F) = \prod_{i=1}^{n-1} P(D_{i,i+dep(i)} = 1 | F_{i,i+dep(i)}) \quad (9)$$

式 (9) と式 (2) とから D_{best} が導かれる.

2.2 後方文脈を考慮した係り受け確率モデル

この節では, 我々が提案するモデルについて説明する. 旧モデルでは, 二つの文節の関係を「係る」か「係らない」かの二カテゴリとして学習し, それらの二文節に係る確率を計算して係り受け確率としていた. 我々のモデルでは, (A) 二つの文節 (前文節と後文節) 間の関係を, 「間」か「係る」か「越える」かの三カテゴリとして学習し, (B) 着目している二つの文節の係り受け確率を求める際に, その二文節に対しては「係る」確率, 二文節の間の文節に対しては前文節がその文節を越えて後文節に係る確率 (「越える」の確率), 後文節より文末側の文節に対しては前文節がその文節との間にある後文節に係る確率 (「間」の確率) をそれぞれ計算し, それらをすべて掛け合わせた確率値を用いて係り受け確率を求める.

このモデルでは, $D_{i,i+j}$ の仮定が旧モデルにおけるものと異なる. $D_{i,i+j}$ としては, 「越える」, 「係る」, 「間」を表す 0, 1, 2 の三値をとると仮定する. この点がこのモデルの特徴の一つである. 文節 b_i と係り受けの関係にある文節が $dep(i)$ ($1 \leq dep(i) \leq n-i$) 番目の係り先の候補であるとき, 2.1 節の (iii) の特徴, つまり係り要素は受け要素を一つだけ持つということを仮定すると,

$$D_{i,i+l} = \begin{cases} 0 & (1 \leq l < dep(i)) \\ 1 & (l = dep(i)) \\ 2 & (dep(i) < l \leq n-i) \end{cases}$$

となる．よって，式 (6) は以下のように変形できる．

$$\begin{aligned}
 P(D|F) = & \prod_{i=1}^{n-1} \left(\prod_{j=1}^{dep(i)-1} P(D_{i,i+j} = 0 | \{D_{i,i+k} = 0 | 1 \leq k < j\}, F) \right. \\
 & \times \prod_{j=dep(i)+1}^{n-i} P(D_{i,i+j} = 2 | \{D_{i,i+k} = 0 | 1 \leq k < dep(i)\}, \\
 & \quad \{D_{i,i+k} = 2 | j < k \leq n-i\}, F) \\
 & \times P(D_{i,i+dep(i)} = 1 | \{D_{i,i+k} | 1 \leq k < dep(i), dep(i) < k \leq n-i\}, F) \Big) \quad (10)
 \end{aligned}$$

$D_{i,i+l}$ の値が決まるのは， $l < m$ を満たすような m に対し $D_{i,i+m} = 0$ になる場合か， $l > m$ を満たすような m に対し， $D_{i,i+m} = 2$ になる場合か， $D_{i,i+dep(i)} = 1$ となる $dep(i)$ が決まる場合かのいずれかである．これらの条件を満たさないように両端から順に係り受け関係 $D_{i,i+l}$ を確率式の前件部に移していけば，式 (10) の最後の項を除くそれぞれの項の確率値は一意には決まらない．式 (10) の最後の項は $D_{i,i+l}$ の値がすべて決まると $dep(i)$ が決まるので確率値は 1 になる．式 (10) のその他の項については， $\{D_{i,i+k} = 0 | 1 \leq k < j\}$ の間の独立性，および $\{D_{i,i+k} = 0 | 1 \leq k < dep(i)\}, \{D_{i,i+k} = 2 | j < k \leq n-i\}$ の間の独立性を仮定すると，式 (10) は次のように表せる．

$$P(D|F) = \prod_{i=1}^{n-1} \left(\prod_{j=1}^{dep(i)-1} P(D_{i,i+j} = 0 | F) \times \prod_{j=dep(i)+1}^{n-i} P(D_{i,i+j} = 2 | F) \right) \quad (11)$$

2.1 節の旧モデルにおける仮定と同様に $F_{i,m} (m = 1, \dots, n)$ はそれぞれ独立で， $D_{i,i+j}$ つまり文節 b_i と文節 b_{i+j} との関係は， $F_{i,i+j}$ のみによって決まると仮定する．すると，式 (11) は，以下のように変形できる．

$$P(D|F) = \prod_{i=1}^{n-1} \left(\prod_{j=1}^{dep(i)-1} P(D_{i,i+j} = 0 | F_{i,i+j}) \times \prod_{j=dep(i)+1}^{n-i} P(D_{i,i+j} = 2 | F_{i,i+j}) \right) \quad (12)$$

この式の $P(D|F)$ を $P_{new_model}(D|F)$ とし，旧モデルの式 (9) における $P(D|F)$ を $P_{old_model}(D|F)$ とする． $P_{old_model}(D|F)$ では素性 $F_{i,i+dep(i)}$ が用いられているが， $P_{new_model}(D|F)$ では用いられておらず， $P_{new_model}(D|F)$ では素性 $F_{i,i+dep(i)}$ 以外の素性が用いられているが， $P_{old_model}(D|F)$ では用いられていない．したがって， $P_{old_model}(D|F)$ と $P_{new_model}(D|F)$ は相補的な関係にあるため，この二つを組み合わせる．すると以下の式が得られる．

$$\begin{aligned}
 P(D|F)^2 &= P_{old_model}(D|F) \times P_{new_model}(D|F) \\
 &= \prod_{i=1}^{n-1} \left(\prod_{j=1}^{dep(i)-1} P(D_{i,i+j} = 0 | F_{i,i+j}) \right. \\
 & \quad \times P(D_{i,i+dep(i)} = 1 | F_{i,i+dep(i)}) \Big)
 \end{aligned}$$

$$\times \prod_{j=dep(i)+1}^{n-i} P(D_{i,i+j} = 2|F_{i,i+j}) \quad (13)$$

本節のモデルはこの式の平方根をとることによって係り受け確率 $P(D|F)$ を求めるものである．この $P(D|F)$ と式 (2) とから D_{best} が導かれる．

図 1 係り受け確率の求め方の例

実際にこのモデルから係り受け確率がどのように求まるかを図 1 を用いて説明する．図 1 はある文節 b_i の係り先の候補が 5 個あったときにそれぞれの候補に係るとしたときの係り受け確率を計算している様子を表している．このとき、文節 b_i とそれぞれの候補との関係がそれぞれ「越える」、「係る」、「間」となる確率として表 1 のような値が得られたと仮定している．

表 1 「越える」, 「係る」, 「間」となる確率の例

候補	越える	係る	間
候補 1	$P(D_{i,i+1}=0 F_{i,i+1})=0.6$	$P(D_{i,i+1}=1 F_{i,i+1})=0.4$	$P(D_{i,i+1}=2 F_{i,i+1})=0$
候補 2	$P(D_{i,i+2}=0 F_{i,i+2})=0.6$	$P(D_{i,i+2}=1 F_{i,i+2})=0.3$	$P(D_{i,i+2}=2 F_{i,i+2})=0.1$
候補 3	$P(D_{i,i+3}=0 F_{i,i+3})=0.3$	$P(D_{i,i+3}=1 F_{i,i+3})=0.5$	$P(D_{i,i+3}=2 F_{i,i+3})=0.2$
候補 4	$P(D_{i,i+4}=0 F_{i,i+4})=0.1$	$P(D_{i,i+4}=1 F_{i,i+4})=0.5$	$P(D_{i,i+4}=2 F_{i,i+4})=0.4$
候補 5	$P(D_{i,i+5}=0 F_{i,i+5})=0$	$P(D_{i,i+5}=1 F_{i,i+5})=0.4$	$P(D_{i,i+5}=2 F_{i,i+5})=0.6$

例えば, 候補 3 を係り先だと仮定したとき, 候補 1, 候補 2 は越えて, 文節 b_i と候補 4, 候補 5 の間に係る確率は, 式 (9) の文節 b_i に関する項を用いて

$$\begin{aligned}
P(D_i|F)^2 &= \prod_{j=1}^2 P(D_{i,i+j}=0|F_{i,i+j}) \times P(D_{i,i+3}=1|F_{i,i+3}) \times \prod_{j=4}^5 P(D_{i,i+j}=2|F_{i,i+j}) \\
&= P(D_{i,i+1}=0|F_{i,i+1}) \times P(D_{i,i+2}=0|F_{i,i+2}) \times P(D_{i,i+3}=1|F_{i,i+3}) \\
&\quad \times P(D_{i,i+4}=2|F_{i,i+4}) \times P(D_{i,i+5}=2|F_{i,i+5}) \\
&= 0.6 \times 0.6 \times 0.5 \times 0.4 \times 0.6 = 0.0432
\end{aligned}$$

つまり,

$$P(D_i|F) = \sqrt{0.0432} = 0.208$$

のように計算され, これが最も高い. 各々の係る確率だけを考えた場合にはそれぞれの $P(D_{i,i+j}=1|F_{i,i+j})$ を比較することになり, 候補 3 と候補 4 の確率がどちらも 0.5 のため決まらないが, この方法によると, 候補 3 を優先的に係り先とすることになる. 一文全体の確率はそれぞれの文節について求めた係り受け確率の積で表され, その積の値が最も高くなるようにそれぞれの係り受けを決めることになる.

我々は, 式 (2) において $P(D|F)$ を最大にする係り受け関係の集合 D_{best} を求めるために, 文末から文頭に向けて解析することにより, 効率良く組み合わせの数を減らしながら一文全体の係り受けを決定する方法を提案している (関根, 内元, 井佐原 1999). この方法では解の探索をビームサーチにより行う. この方法によると決定的に解析を行ってもビーム幅を広くしたときとほとんど同じ精度が得られることが実験により分かっている. そこで, D_{best} を求めるためにこの方法を採用する. このとき, 上述のモデルに 2.1 節で述べた (ii) の特徴, つまり非交差条件を仮定すると, 文節 b_i の係り先の候補はすでに解析の終わった文節 b_{i+1} から文節 b_n までの係り受け関係 $D_{i+1}, \dots, D_n$ に依存して制限されることになる. つまり, 非交差条件のために文節 b_i の係り先の候補とはなり得ない文節 b_j に対しては, 係る確率は 0 になり, 「越える」か「間」かについては b_j の前後の文節のうち非交差条件を満たす文節と文節 b_i との関係が決まれば一意に決まり確率は 1 になる. 文節 b_i の係り受け確率 $P(D_i|F)$ は文節 b_i のすべての係り先の候補について確率値を足すと 1 になるように正規化する.

実際にこのモデルから係り受け確率がどのように求まるかを図2を用いて説明する。図2は図1において非交差条件を考慮した場合の係り受け確率の計算の仕方を表している。ここで、ある文節 b_i より後方の文節について、破線の矢印で表されるような係り受け関係が決まったものと仮定している。このとき、候補3と候補4は非交差条件を満たさないために文節 b_i の係り先の候補とはなり得ない。また、文節 b_i とそれぞれの候補との関係としては、図1と同様に表1の値が得られたと仮定している。例えば、候補5を係り先だと仮定したとき、図2の一番下の例のように、候補1、候補2に対しては越える確率、候補5に対しては係る確率を用いてそれぞれ掛け合わせ、その平方根をとることにより係り受け確率が得られる。一文全体の確率はそれぞれの文節について求めた係り受け確率の積で表され、その積の値が最も高くなるように各々の係り受けを決めることになる。

3 実験結果

この節では、新モデル(後方文脈を考慮したモデル)と旧モデルとの比較実験を行う。実験に用いたコーパスは、京大コーパス (Version 2) (黒橋, 長尾 1997) の一般文の部分で、基本的に学習には1月1日と1月3日から8日までの7日分 (7,958 文)、試験には1月9日の1日分 (1,246 文) を用いた。

2.1節に述べた旧モデルと2.2節に述べた新モデル(後方文脈を考慮したモデル)のそれぞれを文献(内元他 1999)と同様に ME モデルとして実装し、テストコーパスに対する係り受け解析の精度を調べた。係り受け解析の実験に用いた素性は、文献(内元他 1999)のものと同一ものとした。これは表3の基本素性と呼ばれるものとそれらの組み合わせである。このうち、学習コーパス中に4回以上現れた素性約 38,000 個を用いている。

表2 解析結果

モデル	係り受け正解率	文正解率
新モデル	87.93% (9904/11263)	43.58%(540/1239)
旧モデル	87.02% (9801/11263)	40.68%(504/1239)
ベースライン	64.09% (7219/11263)	6.38% (79/1239)

解析結果を表2に示す。ここで、係り受けの正解率というのは文末の一文節を除く残りすべての文節に対して、係り先を正しく推定していた文節の割合を求めたものである。また、文正解率というのは文全体の解析が正しいものの割合を意味する。表2の第1行および第2行はそれぞれ新モデル、旧モデルを用いて京大コーパス1月9日の1,246文を解析した結果である。いずれも、コーパスの形態素情報、文節区切情報を入力として、文節間係り受けの解析を決定的に(ビーム幅 $k=1$) 行なった。ビーム幅を大きくしても精度にほとんど違いはなかったため、決

図2 係り受け確率の求め方の例

定的に解析した結果のみを示した。ベースラインとしては各文節がすべて隣に係るとしたときの精度をあげた。新モデルとしては2.2節に述べた後方文脈を考慮したモデルの精度をあげた。

3.1 旧モデルとの比較

本節では、2.1節に述べた旧モデルと2.2節に述べた新モデル(後方文脈を考慮したモデル)をそれぞれ理論と実験の観点、学習の観点から比較する。

[理論と実験の観点から]

式(13)は式(9)を包含するものであり、式(9)に比べるとより多くの文節との関係(素性 $F_{i,i+j}$ で表される)が考慮されている。ただし、式(10)から式(11)を導くときに用いている独

表 3 学習に利用した素性 (基本素性)

基本素性 (43 種類)	
素性名	素性値
前 (後) 文節主辞見出し	(2204 個)
前 (後) 文節主辞品詞 (Major)	動詞 名詞 ... (11 個)
前 (後) 文節主辞品詞 (Minor)	普通名詞 数詞 ... (24 個)
前 (後) 文節主辞活用 (Major)	母音動詞 ... (30 個)
前 (後) 文節主辞活用 (Minor)	語幹 基本形 命令形 ... (60 個)
前 (後) 文節語形 (String)	とにも ... (73 個)
前 (後) 文節語形 (Major)	助詞 子音動詞カ行 ... (43 個)
前 (後) 文節語形 (Minor)	格助詞 基本連用形 ... (102 個)
前 (後) 文節助詞 1(String)	から まで へ ... (63 個)
前 (後) 文節助詞 1(Minor)	(無) 格助詞 副助詞 (5 個)
前 (後) 文節助詞 2(String)	けど まま や よ か ... (63 個)
前 (後) 文節助詞 2(Minor)	格助詞 副助詞 (4 個)
前 (後) 文節句読点の有無	(無) 読点 句点 (3 個)
前 (後) 文節括弧開の有無	(無) 「 ’ (“ [... (14 個)
前 (後) 文節括弧閉の有無	(無) 」 ’) ”] ... (14 個)
文節間距離	A(1) B(2~5) C(6 以上) (3 個)
文節間読点の有無	無 有 (2 個)
文節間”は”の有無	無 有 (2 個)
文節間括弧開閉の有無	無 開 閉 開閉 (4 個)
文節間前文節同一語形の 有無	無 有 (2 個)
主辞品詞 (Major)	動詞 名詞 ... (11 個)
主辞品詞 (Minor)	普通名詞 数詞 ... (24 個)
主辞活用 (Major)	母音動詞 ... (30 個)
主辞活用 (Minor)	語幹 基本形 命令形 ... (60 個)
文節間後文節同一主辞の 有無	無 有 (2 個)
語形 (String)	とにも ... (73 個)
語形 (Major)	助詞 子音動詞カ行 ... (43 個)
語形 (Minor)	格助詞 基本連用形 ... (102 個)

立性の仮定は、実際の現象そのままではなく近似になっているので、旧モデルに比べると近似の部分が多い。しかしながら、同じ素性を用いた実験 (表 2) で、新モデルは旧モデルに比べて 1%程度良い結果を得ている。これは多少近似があっても実際に係り受け確率の計算に多くの情報を考慮している新モデルの方が良いということを示している。

次に、図 3 に文節長と解析精度の関係をあげる。この図から、どの文節数に対しても新モデルの精度は旧モデルの精度とほぼ同等以上であることが分かる。

[学習の観点から]

学習には学習コーパス中で非交差条件を満たす任意の二文節を用いる。旧モデルでは各二文節に対し「係る」と「係らない」の二つのカテゴリを学習しているのに対し、新モデルでは「越える」と「係る」と「間」の三つのカテゴリを学習している。一般に学習するカテゴリを多くするとデータスパースネスになりやすいが、新モデルでは三つのカテゴリに分けてもデータス

図3 文節長と解析精度の関係

パースネスの問題は生じない。これは新モデルで「越える」と「間」の二つのカテゴリに分けた、旧モデルの「係らない」というカテゴリにはもともと十分な学習データがあったためである。例えば、ある文節の係り先の候補が10個あるときには、そのうち1個だけが「係る」に対するデータであり、残りの9個は「係らない」に対するデータである。ここで「係らない」を「越える」と「間」の二つに分けても、「係る」に比べるとそれぞれ十分な量の学習データがある。

次に、新モデルが旧モデルに比べて優れていることを定量的に示すデータを図4にあげる。これはそれぞれのモデルに対し、学習コーパスの量と解析精度の関係をプロットしたものである。学習コーパスの量にかかわらず、新モデルの方が旧モデルに比べて常に1%程度精度がよいことが分かる。

3.2 その他のモデルとの比較

統計的な手法では、ルールベースに比べて並列構造や従属節間の係り受け関係に対する解析誤りが多い。西岡山らは、この後者の問題を取り上げ、二つの文節の関係が係るか係らずに越えるかを学習するモデルを提案した(西岡山, 宇津呂, 松本 1998)。このモデルを用いることにより、二つの文節だけでなくその二文節とそれらにある文節との関係も扱えるようになる。本論文で我々が提案したモデル(後方文脈を考慮したモデル)はさらにその二文節とそれらよりも文末に近い側の文節との関係も扱うため、彼らのモデルよりも多くの情報を考慮していることになる。

図 4 学習コーパスの量と解析精度の関係

表 4 解析結果

モデル	係り受け正解率	文正解率
「係る」と「越える」	85.40% (9618/11263)	41.40%(513/1239)
「係る」と「係らない」	86.95% (9793/11263)	40.27%(499/1239)
「間」「係る」「越える」	87.93% (9904/11263)	43.58%(540/1239)

表 4 の一行目に西岡山らのモデルを用いたときの実験結果を示す。実験に用いた素性，コーパスは 3 章の最初に説明したものと同じである。後方文脈を考慮したモデルを用いた実験のときと異なるのは、「係る」と「越える」の二つのカテゴリを学習するモデルを用いている部分のみである。表 4 より定量的にも，後方文脈を考慮したモデルのように「間」というカテゴリも考慮した方がよいことが分かる。

次に，後方文脈を考慮したモデルにおいて三カテゴリを学習する必要があることを示す。後方文脈を考慮したモデルでは特徴 (1) としてあげたように二文節間の関係を「間」か「係る」か「越える」かの三カテゴリとして学習する。この三カテゴリのうち二つのカテゴリ「間」と「越える」を，旧モデルの二カテゴリのうち「係らない」によって代用させたモデルを考える。このモデルは，係り受け確率を求める際に，着目している二つの文節（前文節と後文節）だけを考慮するのではなく，前文節と前文節より文末側のすべての文節との関係（後方文脈）を考慮して

いる点が旧モデルとは異なる。表 4 の二行目にこのモデルを用いたときの実験結果を示す。この表より、「間」と「越える」の違いは区別して学習するべきであることが分かる。

4 おわりに

係り受け解析は日本語解析の重要な基本技術の一つとして認識されている。依存文法に基づく解析には、主にルールベースによる方法と統計的手法の二つのアプローチがあるが、我々は利用可能なコーパスが増加してきたこと、規則の変更に伴うコストなどを考慮して、統計的手法をとっている。

統計モデルとしてこれまでよく用いられていたもの(旧モデル)では、係り受け確率を計算する際に、着目している二つの文節に係るか係らないかということのみを考慮していた。本論文では、着目している二つの文節(前文節と後文節)だけを考慮するのではなく、前文節と前文節より文末側のすべての文節との関係(後方文脈)を考慮するモデルを提案した。このモデルを ME(最大エントロピー)に基づくモデルとして実装した場合、旧モデルを同じく ME に基づくモデルとして実装した場合に比べて、京大コーパスに対する実験で、全く同じ素性を用いているにもかかわらず係り受け単位で 1% 程度高い精度 (88%) が得られた。また、後方文脈を考慮したモデルの精度は旧モデルに比べて、どの文長に対してもほぼ常に良く、学習コーパスの量によらず常に 1% 程度良かった。

参考文献

- Collins, M. (1996). "A New Statistical Parser Based on Bigram Lexical Dependencies." In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 184-191.
- 江原暉将 (1998). "最大エントロピー法を用いた日本語文節間係り受け整合度の計算." 言語処理学会 第 4 回 年次大会 発表論文集, pp. 382-385.
- 藤尾正和 松本裕治 (1997). "統計的手法を用いた係り受け解析." 情報処理学会 自然言語処理研究会 NL117-12, 83-90.
- 春野雅彦, 白井諭, 大山芳史 (1998). "決定木を用いた日本語係り受け解析." 情報処理学会論文誌, **39** (12), 3177-3186.
- 黒橋禎夫 長尾眞 (1992). "長い日本語文における並列構造の推定." 情報処理学会論文誌, **33** (8), 1022-1031.
- 黒橋禎夫 長尾眞 (1997). "京都大学テキストコーパス・プロジェクト." 言語処理学会 第 3 回 年次大会 発表論文集, pp. 115-118.

西岡山滋之, 宇津呂武仁, 松本裕治 (1998). “コーパスからの日本語従属節係り受け選好情報の抽出.” 情報処理学会 自然言語処理研究会 NL126-5, pp. 31-38.

関根聡, 内元清貴, 井佐原均 (1999). “文末から解析する統計的係り受け解析アルゴリズム.” 自然言語処理, **6** (3), 59-73.

白井清昭, 乾健太郎, 徳永健伸, 田中穂積 (1998). “統計的構文解析における構文的統計情報と語彙的統計情報の統合について.” 自然言語処理, **5** (3), 85-106.

白井諭, 池原悟, 横尾昭男, 木村淳子 (1995). “階層的認識構造に着目した日本語従属節間の係り受け解析の方法とその精度.” 情報処理学会論文誌, **36** (10), 2353-2361.

内元清貴, 関根聡, 井佐原均 (1999). “最大エントロピー法に基づくモデルを用いた日本語係り受け解析.” 情報処理学会論文誌, **40** (9), 3397-3407.

略歴

内元 清貴: 1994 年京都大学工学部卒業. 1996 年同大学院修士課程修了. 同年郵政省通信総合研究所入所. 研究官. 自然言語処理の研究に従事. 言語処理学会, 情報処理学会, ACL, 各会員.

村田 真樹: 1993 年京都大学工学部卒業. 1995 年同大学院修士課程修了. 1997 年同大学院博士課程修了, 博士 (工学). 同年, 京都大学にて日本学術振興会リサーチ・アソシエイト. 1998 年郵政省通信総合研究所入所. 研究官. 自然言語処理, 機械翻訳, 情報検索の研究に従事. 言語処理学会, 情報処理学会, 人工知能学会, ACL, 各会員.

関根 聡: 1987 年東京工業大学応用物理学科卒. 同年松下電器東京研究所入社. 1990-1992 年 UMIST, CCL, Visiting Researcher. 1992 年 MSc. 1994 年から New York University, Computer Science Department, Assistant Research Scientist. 1998 年 PhD. 同年から Assistant Research Professor. 自然言語処理の研究に従事. 情報処理学会, 人工知能学会, 言語処理学会, ACL 会員.

井佐原 均: 1978 年京都大学工学部電気工学第二学科卒業. 1980 年同大学院修士課程修了. 博士 (工学). 同年通商産業省電子技術総合研究所入所. 1995 年郵政省通信総合研究所関西支所知的機能研究室室長. 自然言語処理, 機械翻訳の研究に従事. 言語処理学会, 情報処理学会, 人工知能学会, 日本認知科学会, ACL, 各会員.

(2000 年 1 月 6 日 受付)

(2000 年 4 月 13 日 再受付)

(2000 年 4 月 24 日 採録)