文節間係り受け距離の統計的性質を用いた 日本語文の係り受け解析

張 玉 潔[†] 尾 関 和 彦[†]

日本語における2文節間の係り受け頻度は、その距離に依存することが知られている。すなわち、文中の文節はその直後の文節に係ることが最も多く、文末の文節に係る場合を除いては、距離が離れるにしたがってその頻度が減少する。この統計的性質は、日本語文の係り受け解析においてしばしば用いられるヒューリスティクス：「文中の文節は係り得る文節の中で最も近いものに係る」の根拠となっている。しかし、このヒューリスティクスは、日本語に見られるこのような統計的性質の一部しか利用していない。したがって、係り受け距離の頻度分布をもっと有効に利用することにより、解析性能が向上する可能性がある。本研究では、ATR 503 文コーパスから抽出した係り受け距離の頻度分布に基づいて2文節間の係り受けペナルティ関数を定義し、「総ペナルティ最小化法」を用いて係り受け解析実験を行なった。その結果を、上のヒューリスティクスに基づく決定論的解析法による解析結果と比較したところ、かなりの解析性能向上が認められた。また、係り文節を分類し、その種類別に抽出した係り受け頻度の情報を用いることにより、さらに解析性能を改善できることが明らかになった。

キーワード： 係り受け解析, 係り受けの整合性, 係り受け距離, 係り受け規則, 統計的言語知識, 総ペナルティ最小化

Dependency Analysis of Japanese Sentences Using the Statistical Property of Dependency Distance between Phrases

YUJIE ZHANG[†] and KAZUHIKO OZEKI[†]

It is well known that the frequency of phrase-pairs in modifier-modified relation depends on the distance between the phrases constituting the pair in the Japanese language. That is, a phrase in a sentence modifies its immediate successor most frequently, and the frequency of modification decreases as the distance between the modifier phrase and the modified phrase increases unless the modified phrase is the last one in the sentence. This paper discusses a method of exploiting this statistical knowledge for dependency analysis of Japanese sentences. The *minimum total penalty method* was used for dependency analysis in this work. The method requires a penalty function, which specifies the association strength between phrases. Several penalty functions were defined based on the frequency distribution of dependency distance extracted from ATR 503-sentence corpus, and the analysis performances were compared. Another experiment was conducted by using a deterministic dependency analysis method for comparison. It is concluded that the knowledge of the dependency distance distribution is effective, and that detailed knowledge of the dependency distance distribution extracted for each modifier phrase group is still more effective for improving the analysis performance.

KeyWords: *dependency analysis, degree of dependency, dependency distance, dependency rule, statistical linguistic knowledge, minimization of total penalty*

1 まえがき

日本語文の表層的な解析には、係り受け解析がしばしば用いられる。係り受け解析とは、一つの文の中で、どの文節がどの文節に係る(広義に修飾する)かを定めることであるが、実際に我々が用いる文について調べて見ると、2文節間の距離とそれらが係り受け関係にあるか否かというこの間に統計的な関係のあることが知られている。すなわち、文中の文節はその直後の文節に係ることがもっとも多く、文末の文節に係る場合を除いては距離が離れるにしたがって係る頻度が減少する(丸山・荻野 1992)。

係り受け距離に関するこのような統計的性質は「どの文節も係り得る最も近い文節に係る」というヒューリスティクス(黒橋・長尾 1994)の根拠になっていると思われる。しかし実際には「最も近い文節に係ることが多い」とは言え、「最も近い文節にしか係らない」というわけではない。したがって、係り受け距離の統計的性質をもっと有効に利用することにより、係り受け解析の性能を改善できる可能性がある(丸山・荻野 1992)。

本論文では、総ペナルティ最小化法(松永・好田 1986; 尾関 1987)を用いて、係り受け距離に関する統計的知識の、係り受け解析における有効性を調べた結果について報告する。総ペナルティ最小化法においては、2文節間の係り受けペナルティの総和を最小化する係り受け構造が解析結果として得られる。ここでは、係り受け距離に関する統計的知識を用いない場合と、そのような知識を用いて係り受けペナルティ関数を設定するいくつかの方法について、解析結果を比較した。また、「係り得る最も近い文節に係る」というヒューリスティクスを用いた決定論的解析法(黒橋・長尾 1994)についても解析結果を求め、上の結果との比較を行った。学習データとテストデータを分離したオープン実験の結果や統計的知識を抽出するための学習データの量が解析結果に与える効果についても検討した(張・尾関 1995)。

2 係り受け距離の統計的知識

丸山らは新聞記事データベースを用いて係り受け距離とその頻度の関係を調査し、それを表す近似式を見い出した(丸山・荻野 1992)。

我々はATR 音声データベース(セット B)(阿部, 勾坂, 梅田, 桑原 1990)に含まれる 503 文コーパスについて丸山らと同様の調査を行った。コーパスの概要を表 1 に示す。全体の 503 文は A~J までの 10 グループに分割されており、A~I のグループには各 50 文、グループ J には 53 文が含まれている。各文節には、それを受ける文節との間の距離を表すラベルが付けられてい

† 電気通信大学 情報工学専攻,
Course in Computer Science and Information Mathematics, The University of Electro-Communications

る。文の係り受け構造はこれらの値によって知ることができる。

表1 ATR 音声データベース (セット B) 中のコーパス

文体	文数	文節数/文	係り受け数
新聞記事, 教科書	503	6.8	2923

2.1 係り受け距離の頻度分布

文 $x_1x_2\dots x_N$ (x_k は文節) において x_i が x_j に係るとき, x_i と x_j の係り受け距離を $j-i$ と定義する。また, $N-i$ を x_i のレンジと呼ぶ。

表2は係り受け距離の頻度分布を距離4以下について示したものである。レンジは考慮していない。距離が11以上の係り受けは存在しなかった。このコーパスでは距離が1の係り受けが全体の64.6%を占めている。文の構成要素(日本語における文節や英語における単語など)が、隣接する構成要素を修飾する傾向は他の言語においても見られる。例えば、英語においても、文中の単語が直後の単語を修飾する (*Right Association*) 頻度は67%に及ぶことが報告されている (Hindle and Rooth 1993)。

表2 係り受け距離の頻度分布

係り受け距離	係り受け数 (相対頻度)
1	1889 (64.6%)
2	487 (16.7%)
3	243 (8.3%)
4	136 (4.7%)

2.2 係り受け距離の頻度を表す近似式

レンジごとに求めた係り受け距離の頻度分布を表3に示す。

表3 レンジごとの係り受け距離の頻度分布

レンジ	係り受け距離					
	1	2	3	4	5	6
2	278(55.7%)	221(44.3%)				
3	268(55.6%)	51(10.6%)	163(33.8%)			
4	260(58.6%)	66(14.9%)	12(2.7%)	106(23.8%)		
5	209(57.4%)	54(14.8%)	21(5.8%)	10(2.8%)	79(19.2%)	
6	141(55.5%)	37(14.6%)	23(9.1%)	6(2.4%)	1(0.4%)	46(18%)

ただし、レンジが6以下のものだけを掲げた。このデータを用いて文献(丸山・荻野 1992)と同じ方法により距離頻度の近似式

$$P(k) = \begin{cases} ak^b, & 1 \leq k \leq r-1 \text{ のとき;} \\ 1 - \sum_{j=1}^{r-1} P(j), & k = r \text{ のとき} \end{cases} \quad (1)$$

(k は係り受け距離, r は係り文節のレンジ)

をあてはめたところ $a = 0.58$, $b = -1.886$ が得られた。これは、丸山らの結果 $a = 0.54$, $b = -1.896$ と近いものである。

2.3 係り文節の種類による係り受け距離の頻度分布のちがい

前節に述べた係り受け距離の頻度分布は、文節の種類を考慮に入れずに求めた、全文節に対する平均的なものであるが、係り受け距離の頻度分布は係り文節の種類に依存することが予想される。そこで、ここでは係り文節をその末尾の形態素によって表4に示す基準により約100種類に分類し、その種類別に頻度分布を求めた。品詞属性はコーパスの説明書(阿部他 1990)によった。

表4 係り文節の分類基準

末尾の 形態素	自立語	活用語：品詞属性と活用形により分類
		非活用語：品詞属性により分類
	附属語	活用語：品詞属性と活用形により分類
		非活用語：形態素と品詞属性により分類

また、受け文節はそれが文末であるか非文末であるかによって区別した。そして係り文節の種類別に、また受け文節が文末か非文末かを区別して係り受け距離の頻度を求めた。具体的な計算は4.2の3に述べる定義式によって行なう。

距離の頻度分布が係り文節の種類に大きく依存する例を表5に示す。格助詞「が」の係り受け距離の頻度分布は式(1) ($a = 0.58$, $b = -1.886$)に近いが、接続助詞「が」の場合には距離2で頻度最大になり、ほかの距離の頻度は一様になっている。このような頻度分布を式(1)のような単調減少関数で近似することには無理がある。したがって、本研究では係り文節の種類別に求めた頻度分布を、そのまま係り受け解析のための情報として用いた。

表5 係り文節の種類により係り受け距離の頻度分布が異なる例

係り文節の種類	距離 1	距離 2	距離 3	距離 4
格助詞「が」	67.9%	21.9%	5.4 %	3.1%
接続助詞「が」	11.8%	35.3%	11.8 %	11.8%

3 係り受け解析法

係り受け距離の統計的性質を利用することにより、どの程度、解析性能が向上するかを調べるため、総ペナルティ最小化法(松永・好田 1986; 尾関 1987)を用いて係り受け解析の実験を行った。また、これと比較するため、決定論的解析法による解析も行った。ここではこれらの解析法について簡単に説明する。

3.1 総ペナルティ最小化法

文節列が“正しい”文を構成するためには文節間に以下の条件を満たす係り受けが存在する必要があると考えられている(吉田 1972)。

唯一性 文末の文節以外は、必ずそれより後ろにある文節のいずれかに唯一に係る。

非交差性 係り受けは交差しない。

整合性 2文節間に係り受けが成立するためには、それらを構成する形態素の品詞や活用形、意味などが整合しなくてはならない。

総ペナルティ最小化法においては2文節間の整合性を程度の問題と考え、それをペナルティ関数で表す。そして、唯一性と非交差性を満たす係り受け構造の中でペナルティの総和が最小になるものすべてを解析結果として出力する。この計算は動的計画法の原理を用いることにより効率よく実行できる(尾関 1987)。ペナルティ関数を適切に設定することにより、2文節間の種々の関係を係り受け解析に利用することができると考えられる。本研究では、後で述べるように係り受け距離の頻度分布に基づいてペナルティ関数を設定した。

3.2 決定論的解析法

この解析法(黒橋・長尾 1994)においては、整合性を程度の問題とは捉えず、整合するかしないかのいずれかであるとする。解析は文末から順に、その文節を受ける文節を決定することにより行われる。非交差性、整合性を満たす受け文節の候補が複数個ある時は、最も距離が近い文節を採用する。解析の途中で受け文節が見出せない時には、その時点で解析不能となる。本研究で用いたアルゴリズムを図1に示す。

4 実験と結果

係り受け距離の頻度情報に基づいて、いくつかのペナルティ関数を定義し、総ペナルティ最小化法による係り受け解析を行った。結果を正解検出率、一意正解率、曖昧度減少率および平均候補数によって評価し、ペナルティ関数による結果の違いを比較検討した。また、決定論的解析法による解析も行い、総ペナルティ最小化法の結果と比較した。

N : 文の文節数 ;
 $kakari_n$: 係り文節番号 ($1 \leq kakari_n \leq N - 1$) ;
 uke_n : 受け文節番号 ($2 \leq uke_n \leq N$) ;
 $uke_number[i]$: 文節番号 i の文節を受ける文節の番号 ;
 $kakari_uke_if(i, j)$: 文節番号 i の文節が文節番号 j の文節に係り得るか否かを
 チェックする関数 .

```

program deterministic_analysis ;
  var kakari_n, uke_n : integer ;
      uke_number [ i ] : array [ 1 .. N ] of integer ;
begin
  uke_number [ N ] := N + 1 ;
  uke_number [ N - 1 ] := N ;
  for kakari_n := N - 2 downto 1 do
    begin
      uke_n := kakari_n + 1 ;
      while (uke_n <= N) and
        (kakari_uke_if ( kakari_n , uke_n ) = false)
      do uke_n := uke_number [ uke_n ] ;
      if uke_n = N + 1
      then begin
        write ( ' failed ' ) ;
        uke_number [ kakari_n ] := kakari_n + 1
      end
      else uke_number [ kakari_n ] := uke_n
      end
    end
  end .

```

図 1 本研究で用いた決定論的解析アルゴリズム

4.1 係り受け規則

文献 (黒橋・長尾 1994) を参考にして, 2 文節間の形態素による整合条件を表 6 に示すように定めた.

表 6 係り受け規則

係り文節の後部の形態素	受け文節の頭部の形態素						
	名詞	名詞+ 判定詞	動詞	述語の イ・ナ 形容詞	連体形の イ 形容詞	連体形の ナ 形容詞	連用形の イ・ナ 形容詞
連体詞	+						
活用語の連用形		+	+	+			
活用語の基本形・タ形	+						
副詞		+	+	+	+	+	+
助詞							
の	+	+	+	+	+		
が、に、より		+	+	+	+	+	
へ		+	+				
を			+				
他の助詞		+	+	+			

実際には、この条件を、「受け文節の頭部の形態素」としての「名詞」、「名詞+判定詞」、「動詞」のさまざまな変形に対処できるように補強して用いる。例えば、「二時半ごろだった」(「数詞+接尾語+名詞+名詞+助動詞+助動詞」)は「名詞+判定詞」として認識されるようにしている。ここではこの条件を係り受け規則と呼ぶ。この規則は総ペナルティ最小化法と決定論的解析法の両方において共通に使用した。

4.2 ペナルティ関数の定義

係り文節 x , 受け文節 y に対してペナルティ関数 $F(x, y)$ を次のように定義する。

$$F(x, y) = \begin{cases} -\log P(x, y), & P(x, y) > 0 \text{ のとき;} \\ c, & P(x, y) = 0 \text{ のとき} \end{cases} \quad (2)$$

ここで、 $P(x, y)$ は x が y に係る頻度から定まる値である。その具体的な定め方は後で述べる。文節 x, y が係り受け規則を満たさない場合には $P(x, y) = 0$ と定義する。このときのペナルティ値 c は非零の $P(x, y)$ で生成される最大ペナルティ値より十分大きな値に設定する。

これにより、文節 x, y が係り受け規則を満たさない場合には大きなペナルティが課せられる。また、係り受け規則を満たす場合には、 $P(x, y)$ が大きい程、ペナルティは小さくなる。 $P(x, y)$ として次の4種類の関数を考え、それぞれに対して実験を行った。

1 距離情報なし

$$P_1(x, y) = \alpha$$

α は正の定数である。すなわち、 (x, y) が一様分布することを仮定する。

2 距離の頻度近似式

$$P_2(x, y) = P(d(x, y))$$

ここで、 $d(x, y)$ は x と y の係り受け距離であり、 $P(\cdot)$ は式 (1) である。

3 係り文節の種類別に求めた頻度分布

2.3 で述べたように、係り受け距離の頻度分布は係り文節の種類に依存する。また、係り受け距離が同じでも受け文節が文末か非文末かによって係り受け頻度は大きく異なる (丸山・荻野 1992)。そこで係り文節の種類別に、また受け文節が文末か非文末かを区別して、以下のように係り受け距離の頻度分布を求めた。

係り文節の種類 (表 4) を番号 m で表す。そして、 $T(m)$ を係り文節が第 m 種の文節であるような係り受け文節対の全体とする：

$$T(m) = \{ (u, v) \mid \text{文節 } u \text{ は文節 } v \text{ に係る, } u \text{ は第 } m \text{ 種の係り文節} \}$$

$T(m)$ の中で、係り文節と受け文節の距離が k であり、受け文節が文末であるようなものの全体を $S_l^m(k)$ 、同じく受け文節が非文末であるようなものの全体を $S_n^m(k)$ とする：

$$S_l^m(k) = \{ (u, v) \mid (u, v) \in T(m), d(u, v) = k, v \text{ は文末文節} \},$$

$$S_n^m(k) = \{ (u, v) \mid (u, v) \in T(m), d(u, v) = k, v \text{ は非文末文節} \}$$

当然、次の関係がある：

$$T(m) = \cup_k (S_l^m(k) \cup S_n^m(k))$$

そして、以下の式により文節 x が文節 y に係る相対頻度の推定値 $P_3(x, y)$ を計算する：

$$P_3(x, y) = \begin{cases} |S_l^{m(x)}(d(x, y))| / |T(m(x))|, & y \text{ が文末文節のとき;} \\ |S_n^{m(x)}(d(x, y))| / |T(m(x))|, & y \text{ が非文末文節のとき} \end{cases}$$

ここで $m(x)$ は係り文節 x の種類を表し、 $|\cdot|$ は集合の要素数を表す。すなわち、 $P_3(x, y)$ は、 $d(x, y) = k$ とするとき、 x と同じ種類の文節を係り文節とする係り受け文節対の中で、距離が k であるようなものが、どれだけの割合で存在するかを示す量であり、 y が文末か非文末かに分けて計算される。ここでは、係り受け文節対の出現頻度が、係り文節の種類、係り受け距離、および受け文節が文末か非文末かの三つの変数に依存して定まる分布モデルが仮定されていることになる。

4 補間した頻度分布

係り受け規則がコーパスを完全にカバーしていないため、ある文節を受ける文節が文末までに存在しないことがある。また係り受け規則で許されても、コーパスの希薄性によって、係り受け頻度が 0 となる場合がある。そこで、 P_3 に次のような一種の補間を施し、その結果を $P_4(x, y)$ とする。

(1) 上のような問題がない場合：

$$P_4(x, y) = P_3(x, y)$$

(2) 上のような問題がある場合:

(a) 文節 x を受ける文節が文末までに存在しない場合:

文節 x がどの文節に係るかを係り受け距離の頻度分布に基づくヒューリスティクスで定める。

すなわち、文節 x に対して、上の 3 で推定された係り受け頻度が最大となる後続文節を求め、 x がその文節に係ることを許す。また、他の文節に係ることは許さない。これは、文節 x が後続文節 y に係ることが係り受け規則の上では許されなくても、もし文節間距離が $d(x, y)$ に等しい係り受け文節対の出現頻度が大きければ、それを許そうという考え方である。その際、出現頻度が最も大きい後続文節だけに対して文節 x が係ることを許し、他の文節に対しては許さないことにする。

実際の計算は次のように行なう。

文節 x に対して、最大係り受け頻度を与える後続文節を z とする:

$$z = \arg \max \{ P_3(x, y) \mid y \text{ は } x \text{ の後続文節} \}$$

このような z が複数個あるときは、それらをすべて求める。そして、文節 x が後続文節 y に係る頻度 $P_4(x, y)$ を次のように設定する:

$$P_4(x, y) = \begin{cases} P_3(x, y), & y = z \text{ のとき;} \\ 0, & y \neq z \text{ のとき} \end{cases}$$

(b) x が y に係ることが、係り受け規則で許されるにもかかわらず、 $P_3(x, y)$ が 0 になる場合:

$$P_4(x, y) = \beta$$

ただし、 β は

$$\max_{u, v} \{ -\log P_3(u, v) \} < -\log \beta < c \quad (3)$$

を満たすような値に設定する。ここで、最左辺の最大値は文節 u が文節 v に係ることが係り受け規則で許されるような u, v のすべての組についてとる。これはいわゆる底上げ (flooring) の技法である。このとき、 x が y に係るペナルティは $-\log \beta$ となるが、式 (3) は、このペナルティが係り受け規則により係り受けが許されない文節対のペナルティよりはるかに小さく、また、係り受け頻度から定まる最大ペナルティよりは大きくなるように β の値を設定することを意味する。

4.3 解析実験

解析実験は学習データから係り受け距離の頻度情報を抽出する学習ステップと、その頻度情報に基づいて設定したペナルティ関数を用いてテストデータを係り受け解析し、結果を評価する解析ステップから成る。

まず学習法と各種のパラメータ設定法について説明する。

P_1 を用いる場合には、頻度情報は全く使用しない、したがって学習は不要である。 α はどのような値に設定しても解析結果は同じになる。

P_2 はパラメトリックな分布モデルである。学習ステップにおいては、学習データを用いてパラメータ a, b を推定する。

P_3 はノンパラメトリックな分布モデルである。学習ステップにおいては、学習データを用いて、係り文節の種類 m と係り受け距離 k に対して $S_l^m(k)$ と $S_n^m(k)$ を求め、記憶する。解析ステップにおいては、各文節 x, y に対し、 $S_l^{m(x)}(k)$ と $S_n^{m(x)}(k)$ から定義式に基づいて $P_3(x, y)$ の値を計算し、ペナルティ関数の値を設定する。

P_4 は基本的には P_3 から定まるので学習する必要はない。 β の値は式(3)を満たす限りどのような値に設定しても結果に大きな差はないと考えられるので、式(3)を満たす値を任意に選んで使用した。

ペナルティ関数を式(2)によって定義するとき、定数 c を定める必要がある。これは文節 x が文節 y に係ることが係り受け規則により許されないとき、あるいは文節 x が文節 y に係る頻度が0になるとき (P_4 によって補間されない限り) 文節 x が文節 y に係ることを禁止するような大きなペナルティを与えるためのものである。したがって、これは ∞ とも言うべきものであり、十分大きな値に設定すれば、どのような値に設定しても解析結果に差はない。

このような学習法およびパラメータ設定法を用いて次のような3種類の実験を行った。

実験1 $P_1 \sim P_4$ を用いてペナルティ関数を定義し、総ペナルティ最小化法による解析実験を行った。この実験では係り受け距離の頻度情報を抽出するための学習データとテストデータを分離せず、どちらも503文すべてを用いた。比較のため同じデータを用いた決定論的解析法による解析実験も併せて行った。

実験2 実験1で総合的に最も良い結果が得られたのは P_4 で定めたペナルティ関数である。これを用いて、学習データとテストデータを分離した解析実験を行った。10グループ $A \sim J$ の一つをテストデータとし、残りを学習データとした。テストデータを A から J まで変えて10回の実験を行った。

実験3 学習データの量と解析結果との関係調べるため、テストデータをグループ J に固定し、学習データを $A, A \cup B, A \cup B \cup C, \dots$ と漸次増加させる実験を行った。純粋に学習データの量が解析結果に与える効果を見出すため、補間していない P_3 を用いてペナルティ関数を定めた。

4.4 解析結果

解析結果について述べるため、まず記法と用語を定義する。

- M : 評価に用いるテスト文の総数 ;
- S_i : 番号 i のテスト文 ;
- L_i : 文 S_i の文節数 ;

文 S_i に対する係り受け構造の中の係り受けの総数は $L_i - 1$ に等しい。

- D_i : 長さ L_i の文節列上に存在し得る係り受け構造の総数 ;

D_i はカタラン数と呼ばれる数列になり、次の式によって計算することができる :

$$D_i = \frac{1}{L_i} 2^{(L_i-1)} C_{(L_i-1)}$$

ここで $2^{(L_i-1)} C_{(L_i-1)}$ は $2^{(L_i-1)}$ 個のものから (L_i-1) 個のものをとる組合せの数を表す。この式は再帰式

$$D_i = \begin{cases} 1, & i = 1 \text{ のとき;} \\ \sum_{1 \leq j \leq i-1} D_{i-j} D_j, & i \geq 2 \text{ のとき} \end{cases}$$

が成り立つ (尾関 1987) ことと、母関数の手法 (伊理他 1987) を用いることにより示すことができる。

- R_i : 文 S_i の解析結果、すなわち係り受け構造候補の集合 ;
- K_{ij} : 文 S_i に対する j ($1 \leq j \leq |R_i|$) 番目の解析候補の中でコーパスのラベルに示されるものと一致する係り受けの数。

評価方法

結果の評価は、(1) 2 文節間の係り受けがどの程度正しく検出されたか、(2) 係り受け構造がどの程度正しく検出されたか、という二つの観点から行った。また、文の検出率が高くても一つの文に対する解析結果の候補数が多ければ良い解析法とは言えない。このため、候補数がどれだけ絞られるかを評価することとした。さらに、解析を行うことにより、情報理論的な曖昧さがどれだけ減少するかを調べ、全体的な解析効率を評価した。これらの評価を行うため、以下のようないくつかの評価尺度を定義した。

(1) 係り受けの正しさに着目した評価尺度

$$\text{文 } S_i \text{ の係り受け検出率} = \frac{\sum_{j=1}^{|R_i|} K_{ij}}{(L_i - 1) \times |R_i|}$$

$$\text{係り受け検出率} = \frac{\sum_{i=1}^M \sum_{j=1}^{|R_i|} K_{ij}}{\sum_{i=1}^M (L_i - 1) \times |R_i|}$$

係り受け検出率は、解析結果がラベルと部分的に一致する度合を示す数字である。このような部分的な一致も、結果を意味理解に利用する場合などには有用と思われる。

決定論的解析法を用いた解析実験においては、解析の途中で、ある文節を受ける文節が存在しない時には、直後の文節を受け文節として解析を続けた。その結果をもとにして係り受け検出率を計算した。

(2) 係り受け構造の正しさに着目した評価尺度

(a) 文検出数と文検出率

$$\text{文検出数}_k = \sum_{i: |R_i| \leq k} I_i$$

ここで

$$I_i = \begin{cases} 1, & R_i \text{の中にラベルで指定される係り受け構造と一致する候補が存在する場合;} \\ 0, & \text{他の場合} \end{cases}$$

文検出数_kは、出力された解析結果の候補数がk以下であり、かつ、その中にコーパス中のラベルで指定される係り受け構造と一致するものが含まれるようなテスト文の数である。また、これをテスト文の総数に対する比で表わしたものを、文検出率_kと呼ぶ：

$$\text{文検出率}_k = \frac{\text{文検出数}_k}{M}$$

文検出数_∞、文検出率_∞をそれぞれ単に文検出数、文検出率という。また、文検出数₁、文検出率₁をそれぞれ一意正解数、一意正解率と呼ぶことにする。一意正解数は解析結果が一意的に決定し、それがコーパスのラベルと一致したテスト文の数である。

(b) 平均候補数

$$\text{平均候補数} = \frac{\sum_{i=1}^M |R_i| \times I_i}{\sum_{i=1}^M I_i}$$

(c) 曖昧度減少率

文 S_i の係り受け構造の候補数の対数をその文の曖昧度と定義すると、

$$\text{解析前の曖昧度} = \log D_i$$

$$\text{解析後の曖昧度} = \begin{cases} \log |R_i|, & R_i \text{の中にラベルで指定される係り受け構造と} \\ & \text{一致する候補が存在する場合;} \\ \log D_i, & \text{他の場合} \end{cases}$$

$$= \log (D_i + (|R_i| - D_i) I_i)$$

となる。これを用いて、次の量を定義する。

$$\text{曖昧度減少率} = \frac{\sum_{i=1}^M \log D_i - \sum_{i=1}^M \log (D_i + (|R_i| - D_i) I_i)}{\sum_{i=1}^M \log D_i}$$

曖昧度減少率は文の統語的な曖昧さが解析により減少する割合を表す。

実験結果と分析

(1) 実験 1 の結果を表 7 に示す。また、同実験において候補数を制限したときの文検出数 (率) を表 8 に示す。

P_1 を用いた場合は、他の場合と比べて、文検出率が高い反面、平均候補数が非常に大きい。また、表 8 から文検出率 $_k$ ($k = 10, 1$) は $P_2 \sim P_4$ を用いた場合の方が、 P_1 を用いた場合より高い。したがって、係り受け距離の情報は候補数を絞るのに有効であることがわかった。

P_3 を用いた場合は、 P_2 を用いた場合と比較して、表 7 における文検出率、および表 8 における文検出率 $_k$ ($k = 50, 10, 1$) が高い。したがって、係り文節の種類別に求めた係り受け距離の頻度情報は、係り文節を分類せず全体で求めた情報よりも、各文検出率 $_k$ を高めるのに有効である。また、その結果、曖昧度が減少している。

P_4 を用いた場合は、 P_3 を用いた場合と比べて、表 7 における係り受け検出率と表 8 における一意正解率が高くなっている。したがって、補間は係り受け検出率と一意正解率を上げる効果があることが検証された。文検出率は低下したが、 P_2 を用いた場合よりは高く、平均候補数は P_3 を用いた場合の半分以下になっている。したがって、距離の頻度分布を補間することにより、あ

表 7 実験 1 の結果

ペナルティ関数	文検出数 文検出率 (%)	平均 候補数	曖昧度 減少率	係り受け 検出率 (%)
P_1 (距離情報なし)	432(85.9)	41.3	0.40	61.5
P_2 (近似式)	247 (49.1)	2.5	0.36	75.2
P_3 (種類別)	321(63.8)	2.9	0.52	74.4
P_4 (補間)	287(57.1)	1.1	0.47	87.1
決定論的解析法	193(38.4)	1.0	0.28	81.9

表8 候補数を制限したときの文検出数(率)

ペナルティ関数	文検出数 _k (率%)		
	k = 50	k = 10	一意正解数(率) (k = 1)
P ₁ (距離情報なし)	364(72.4)	235(46.7)	31(6.2)
P ₂ (近似式)	246(48.9)	244(48.5)	195(38.8)
P ₃ (種別)	319(63.4)	315(62.6)	237(47.1)
P ₄ (補間)	287(57.1)	287(57.1)	264(52.5)
決定論的解析法	193(38.4)	193(38.4)	193(38.4)

る程度係り受け規則の不完全さとコーパスの希薄性の問題を軽減できることがわかった。また、平均候補数は1.1まで絞られており、87.1%というかなり高い係り受け検出率が得られることから、この解析法を意味理解のための部分解析法として使用できる可能性がある。

決定論的解析法を用いた解析実験と比較すると、 $P_1 \sim P_4$ を用いた場合は表7の平均候補数が大きくなったかわりに文検出率がかなり向上している。さらに、表8において、 $P_2 \sim P_4$ を用いた場合は決定論的解析法を用いた場合の一意正解率を越えるとともに各文検出率_kが高くなっている。とくに P_4 を用いた場合は決定論的解析法を用いた場合と平均候補数がほとんど等しく、各文検出率_k、曖昧度減少率、および係り受け検出率が高くなっている。したがって、係り受け距離の統計的知識を利用した総ペナルティ最小化法により、決定論的解析法を用いた場合に比べて解析性能を向上させることができる。

(2) 実験2の結果を表9に示す。この実験は学習データとテストデータを分離した、いわゆるオープン実験である。文検出率と一意正解率はクローズ実験である実験1の結果に比べると若干低下し、平均候補数もやや増加している。しかし、1.2という平均候補数は、この解析結果に対してさらに何らかの後続処理を行う場合、処理量の上で問題となる数ではなく、表7に示される決定論的解析法の結果と比べると、文検出数、一意正解数、曖昧度減少率など全てが高い。した

表9 実験2の結果

テストグループ	A	B	C	D	E	F	G	H	I	J	合計
文数	50	50	50	50	50	50	50	50	50	53	503
文検出数(率)	25	27	20	26	25	16	26	26	29	41	261(52%)
一意正解数(率)	21	22	19	19	22	13	18	24	23	37	218(43%)
平均候補数	1.3	1.3	1.1	1.5	1.1	1.2	1.4	1.1	1.2	1.1	1.2
曖昧度減少率	0.39	0.41	0.30	0.49	0.33	0.28	0.41	0.42	0.44	0.68	0.42

がって、未知文の係り受け解析に対しても (1) と同様の結論が導かれる。

(3) 実験 3 の結果を表 10 に示す。これもオープン実験である。学習データの量が増加するに従って、一意正解数、曖昧度減少率が向上するとともに、平均候補数は減少する。文検出数はほぼ一定に保たれる。したがって、学習データ量を増加させることはこのような解析法の性能向上に有効である。より大きなコーパスを使用することによって、更に解析性能が向上することが期待される。

表 10 実験 3 の結果

学習データの量 (文数)	50	100	150	200	250	300	350	400	450
53 文の文検出数	43	43	42	41	42	41	40	43	43
一意正解数	16	24	25	25	28	29	29	29	31
平均候補数	3.8	2.2	2.0	1.9	1.7	1.5	1.5	1.5	1.5
曖昧度減少率	0.53	0.61	0.62	0.60	0.63	0.63	0.60	0.67	0.68

5 むすび

本研究の結論は次のようにまとめることができる。

- (1) 係り受け距離の頻度分布は係り受け解析におけるヒューリスティクスとして有効である。
- (2) 係り受け距離の頻度分布は、係り文節の種類別に求めた方がよい。
- (3) 頻度分布の補間により、係り受け規則の不完全さとコーパスの希薄性のある程度補うことができる。
- (4) 頻度分布を抽出する学習データ量を増加することにより解析性能が向上する。ここで用いた学習データよりもっと多くのデータを使用することにより、解析性能がさらに向上する可能性がある。

問題点としては、以下のことが挙げられる。

- (a) 係り文節の分類は、自立語が非活用語の場合にはその品詞属性に基づいて行っている。しかし、これには次のような問題がある。コーパス中に高頻度で出現する「名詞」は、その半数程度が「連体詞」のような働きをして名詞に係る。また、「前」、「ため」などの「名詞」は「副詞」のような働きをして動詞に係ることがある。「連体詞」と「副詞」では、それを受ける文節との距離の頻度分布が大きく異なる。このような問題があるので、係り受け距離の頻度分布を求めるときの係り文節の分類基準をもっと工夫する必要がある。
- (b) ここで用いた係り受け規則はまだ不完全である。 P_1 を用いたときの文検出率がこの係り受け規則のカバー率を示しているが、表 7 に見られるように 100% には達していない。したがって、実際のコーパスをよりよくカバーする係り受け規則を設定する必要がある。

- (c) 文中の部分文節列に対して、係り受け構造に曖昧さがあっても、それによる意味の曖昧さがほとんど生じない場合もある。例えば、「両手の 十本の 指を」に対して、図2の二つの係り受け構造から得られる意味の違いが実際にどれ程問題となるであろうか。

図 2

本研究で行った評価法では、出力される係り受け構造がコーパスのラベルと完全に一致しない限り、文としては検出されなかったことになる。しかし、このような曖昧性を解消する意義がどれくらいあるかを考える必要があると思われる。

- (d) 「は」のような係り助詞で終る文節がどの文節に係るとするかは、微妙な問題を含んでいる。例えば、図3は、「インタビューは」が主文の述語「及んだ」に係るような係り受け構造を示しているが、これに対して、図4のように「インタビューは」が「始まり」に係るとする係り受け構造を考えることもできる。図3は、文を文末まで見てから係り受け構造を定めたものであり、図4は、文を左から見て行き、可能な係り受けを定めて行ったものと言えるであろう。

図 3

図 4

いずれか正しいかは別にして、使用したコーパスにはこの2種類の考え方に基づくラベル付けが混在しており、統一性を欠いている。これが、係り受け距離の頻度分布の推定や、解析結果の評価に影響を与えていると思われる。

以上の結果と問題点を踏まえて、今後は文節間の係り受けに関する言語現象を詳しく調査し、係り受け解析に対するより優れたペナルティ関数の設定法について研究を進める予定である。

参考文献

- 阿部匡伸, 匂坂芳典, 梅田哲夫, 桑原尚夫 (1990). 研究用日本語音声データベース利用解説書. ATR 自動翻訳電話研究所.
- Hindle, D. and Rooth, M. (1993). "Structural Ambiguity and Lexical Relations." *Computational Linguistics*, 19(1), pp. 103-120.
- 伊理正夫他 (1987). 現代応用数学の基礎 1. 日本評論社, pp. 163-164.
- 黒橋禎夫・長尾真 (1994). "並列構造の検出に基づく長い日本語文の構文解析." 自然言語処理, 1 (1), pp. 35-57.
- 丸山宏・荻野紫穂 (1992). "日本語における文節間係り受け関係の統計的性質." 第 45 回情報処理学会全国大会論文集, 3 巻, pp. 173-174.
- 松永昭一・好田正紀 (1986). "係り受けの整合性に基づく文節認識の後処理." 日本音響学会昭和 61 年度春季研究発表会講演論文集, pp. 1-1-23.
- 尾関和彦 (1987). "係り受けの整合度に基づき最適文節列を選択する多段決定アルゴリズム." 信学論 (D), 70-D(3), pp. 601-609.
- 吉田将 (1972). "二文節間の係り受けを基礎とした日本語文の構文分析." 信学論 (D), 55-D(4), pp. 238-244.
- 張玉潔・尾関和彦 (1995). "文節間係り受け距離の統計的性質とその係り受け解析への応用." 電子情報通信学会技術研究報告, NLC 95-58, pp. 61-68.

略歴

張 玉潔: 1983 年北方交通大学情報工学科卒業. 1986 年中国科学技術大学大学院修士課程修了. 同年中国科学院計算技術研究所に勤務. 機械翻訳の研究開発に従事. 国家科学技術進歩一等賞受賞. 現在, 電気通信大学情報工学専攻博士課程在学中. 自然言語処理に興味がある.

尾関 和彦: 1965 年東京大学工学部電気工学科卒業. 同年 NHK 入社. 1968 年より 1 年間エジンバラ大学客員研究員. 音声言語処理, パターン認識などの研究に従事. 電子通信学会 第 41 回論文賞受賞. 現在, 電気通信大学情報工学科教授. 工学博士. 電子情報通信学会, 情報処理学会, 日本音響学会, IEEE 各会員.

(1996 年 4 月 24 日 受付)

(1996 年 8 月 20 日 再受付)

(1996 年 9 月 19 日 採録)