自然言語処理 117－12
（1997. 1. 20）

統計的手法を用いた係り受け解析

藤尾 正和，松本 裕治
奈良先端科学技術大学院大学 情報科学研究科
{masaka-h,matsu}@is.aist-nara.ac.jp

本研究では、ルールベースの係り受けシステムにつきものの保守管理の不便さ、ユーザが内部を変更できないことに起因する不自由さに対処する目的で、統計的手法を用いた日本語係り受け解析システムの作成を目指す。
統計情報を用いることにより、確率値付きの係り受け結果を出力できるので、係り受け正解コーパスの作成支援や、より詳細な情報を用いた処理と組み合わせて曖昧性爆発の回避などにも使用できる。
システムの特徴をまとめると、1. 文節の定義、属性の決定はユーザ側で指定可能、2. 統計量を用いているので適用分野をかえて学習し直すことが容易 3. 正解例をためていくことで解析の精度の向上を図ることができる、などが挙げられる。EDR[3] の構文情報付きコーパスを用いて学習およびシステムの評価を行った。

[キーワード] 係り受け解析，括弧付きコーパス，共起確率，スムージング

Japanese Dependency Structure Analysis based on Statistics

FUJIO Masakazu and MATSUMOTO Yuji
Graduate School of Information Science, Nara Institute of Science and Technology

The purpose of this paper is to present an adaptable Japanese dependency analysis system based on a statistical model. We only used surface information like morphemes, part-of-speech tags, distances, and don't use any grammar rules written by grammarians. Because of the adoption of statistical method, the system's outputs have a natural measure of priority. The system is usable for acquisition of correct bracketted sentences with less human interaction, for an efficient natural language processing(NLP) combined with other NLP systems.
We conducted some experiments for outside sentences taken from EDR corpus.

[keyword] statistical model, dependency structure, bracketted corpus, collocation probability, smoothing

1 はじめに

これまでの日本語係り受け解析の研究は、現象の一部を扱った研究か、網羅的なものでも研究者の手で係り側受け側の文節を細かく分類し、分類間の係り受けを記述するルールベースの研究がほとんどであった。

しかしルールベースの係り受け解析システムは、適用分野を変更したり拡大する度に研究者の手でルールを書き直さねばならず、保守管理が煩雑であるという問題点がある。これは基礎技術を様々な応用分野に適用しようという目的からは望ましくない。

そこで本研究では、統語解析済みコーパスから学習した統計量をもとに、構文規則、書き換え規則などの文法規則を仮定せずに日本語の係り受け解析を行うシステムの作成を目指す。

統計情報を用いた統語解析の研究には、単語の共起頻度を用いた研究が多いが、共起情報を解析に利用する場合、これまでは非終端記号や品詞タグレベルでの共起頻度を扱った研究が主流であった。しかし近年それだけでは構文的な曖昧性の解消能力が不十分であることが指摘されてきたため、語彙情報を積極的に利用する方向に進んできた [2, 6, 8]。

日本語の統語解析で、単語レベルの情報を使用した例としては、[1, 4, 7, 13, 11, 12] などがあげられる。しかしこれらの多くは CFG などの書き換え規則を基にしたものであるか、従来の係り受け規則が残っていて共起情報はあくまで優先づけにすぎないものが多い。共起関係のみから解析システムを構築した例としては、安原 [13] があげられるが、そこ

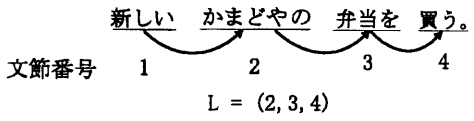


図 1: L の具体例

では主辞に関しては品詞に縮退させた共起情報を用いている。また共起関係の出現を確率としてではなく、正解事例として蓄積して、次回の解析に利用している。

本研究では、多くの解析システムで用いられているCFGなどの構文規則を使用せず、日本語の係り受けを係り側文節が受け側文節を選択する過程と考え、各係り関係の確率を単語レベルの情報を直接用いて推定し、統計的係り受け解析を行った。

本来ならば、形態素解析、文節区切り(属性付与を含む)、係り受け解析のすべてを考慮して、確率が最大となる組み合わせを求めるなくてはならないが、本研究では形態素解析、文節決定(文節の切り出しと属性の付与)はJUMAN (ChaSen)[9]および自作の文節区切りプログラムを用いて決定的に行い、属性付与された文節列から係り受けパターンを決定する段階についてのみ統計処理を行った。係り受け解析の前提となる文節決定部分は、定義ファイルに記述することでユーザが自由に指定できる。

以下の章では、まず第2章で今回使用した統計モデルについて述べたあと、第3章でシステムの処理の全体の流れについて説明する。ここで文節区切りの手法についても述べる。そして第4章で実験方法と結果、第5章でまとめを述べる。

2 統計モデル

本研究では日本語の係り受けを、係り側文節が受け側文節を選択する過程であると考える。

入力文字列をS、分かち書きされてタグ付けされた単語列$< w_1,t_1 >,\ldots,< w_n,t_n >$を$T$、文節にまとめられ属性付けされた文節列$< b_1,\mathbf{f_1} >,\ldots,< b_m,\mathbf{f_m} >$を$F$、文節区切りに対する係り受けパターンの組$\{Dep(1),Dep(2)\ldots Dep(m-1)\}$を$L$とする。但しDep(i)は文節$b_i$の係り先の文番号を表す。$w_i,t_i,b_i,m$はそれぞれ単語、タグ、文節、文節数を表す。また$\mathbf{f_i}$は文節$b_i$の持つ属性の集合を表すものとする[1]。

ここでは係り受けのパターンとは、対象にしている文において

1. 文末を除き各文節は文末側に必ず一つの係り先を持つ
2. 係り受けは非交差

という制約を満たす係り受けの組み合わせを言う。

最も一般的な形では、係り受け解析とは条件付き確率$P(L,F,T|S)$が最大になるL,F,Tを求めることである。これを式で書くと

$$\begin{aligned} L_{best} &= \operatorname*{argmax}_L P(L,F,T|S) \\ &= \operatorname*{argmax}_L P(L|F,T,S)\ P(F|T,S)\ P(T|S) \end{aligned}$$

となる。

文節区切り(属性決定を含む)は分かち書きとタグから決定でき、係り受けは文節区切りのみで決定できると考えられるので、上式は

$$P(L,F,T|S) = P(L|F)\ P(F|T)\ P(T|S) \qquad (1)$$

と書ける。従って$P(L|F),P(F|T),P(T|S)$の3項の積が最大になるよう、分かち書きとタグ付け、文節区切り、係り受けを決定すればよい。

本研究では話を単純化するため、分かち書きとタグ付けはJUMAN (ChaSen)の最適解出力を使用し、文節決定は今回自作したプログラムを用いて決定的に行った[2]。すなわち$P(T|S)=1,P(F|T)=1$となる。よって

$$P(L|F) \qquad (2)$$

を最大化する係り関係の組合せLを求めればよい。

多くの先行研究と同様、それぞれの係り受けは独立であると仮定すると

$$P(L|F) = \prod_{i=1}^{m} P(Dep(i){=}j \mid \mathbf{f_1}\ldots\mathbf{f_m}) \qquad (3)$$

(ただしLは非交差)

と表される。

(3)式の各項の考え方にはいくつかある。例えばCollins[2]の場合、$F_c(R \mid \langle w_i,t_i\rangle,\langle w_j,t_j\rangle)$[3]という確率を考えている。これは$\langle w_i,t_i\rangle$と$\langle w_j,t_j\rangle$が同一文中に現れたときに関係Rで係り関係にある確率を表している。この確率をコーパス中のデータを基に最尤推定して、

$$\hat{F}_c(R \mid \langle w_i,t_i\rangle,\langle w_j,t_j\rangle) = \frac{C(R,\langle w_i,t_i\rangle,\langle w_j,t_j\rangle)}{C(\langle w_i,t_i\rangle,\langle w_j,t_j\rangle)}$$

としている。

ここで$C(\langle w_i,t_i\rangle,\langle w_j,t_j\rangle)$とはコーパス中で、$\langle w_i,t_i\rangle,\langle w_j,t_j\rangle$が同一文中に現れた頻度、$C(R,\langle w_i,t_i\rangle,\langle w_j,t_j\rangle)$はコーパス中で、同一文中に

[1] 具体的には、文節の主辞、読点の有無、係り関係など

[2] 文節決定については第3章を参照

[3] 本研究での属性集合f_iに当たるのが$\langle w_i,t_i\rangle$と考えればよい

$\langle w_i, t_i \rangle, \langle w_j, t_j \rangle$ が現れ、関係 R で係り関係にあった頻度を表す。

この最尤推定した確率を基に、可能な係り先候補と係り関係について正規化して

$$P_c(Dep(i)=j \mid S, B) \approx \frac{F_c(R \mid \langle w_i, t_i \rangle, \langle w_j, t_j \rangle)}{\Sigma_{k=1\ldots m, k \neq j, p \in P} F_c(p \mid \langle w_i, t_i \rangle, \langle w_k, t_k \rangle)}$$

としている。

S は入力文を表す。B は BaseNP と呼ばれるもので、いくつかの名詞をまとめて、ひとかたまりにしたものである[4]。すなわち係り受けは、BaseNP およびその他の単語を要素とする関係となる。

これに対し本研究では、まず $F(\Delta, \mathbf{f_j} \mid \mathbf{f_i})$ という確率を考える。Δというのは、生成される係り受けの距離属性であり、$\mathbf{f_j}$は受け側属性である。この確率は、係り受けの係り側属性が$\mathbf{f_i}$の時に、その係り受けの距離属性がΔで受け側属性が$\mathbf{f_j}$である確率を表す。この確率を用いて、式 (3) の各項を近似する。係り受け候補の確率の和が 1 となるように正規化して、

$$P(Dep(i)=j \mid \mathbf{f_1} \ldots \mathbf{f_m}) \approx \frac{F(\Delta, \mathbf{f_j} \mid \mathbf{f_i})}{\Sigma_{k=1\ldots m} F(\Delta, \mathbf{f_k} \mid \mathbf{f_i})}$$

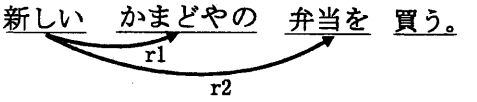
となる。

Collins のモデルと本研究のモデルの一番の違いは、確率 F を定義する時の前件部分であり、前者は同一文中に出現した文節組合わせごとに正規化されているが、後者では係り側文節ごとに正規化してある。

具体的にモデルの違いを見るため、今ここで (3) 式の各項を計算する際に競合関係となる 2 つの係り受けの尤度を比べることを考える。例として図2を考える。r1 と r2 の確率を比べるとき、Collins のモデルでは係り受け r1 の組合わせがコーパス中の同一文に出現したときに、実際に係り受け関係にある確率を推定することになり、確率を推定するときの分母の頻度が r2 の場合と異なってしまう。つまり本当にこの 2 つ係り受けの尤度を比較するならば、同一文中に" r1,r2 両方" の組合わせが存在したときに、どちらかが係り受け関係にある頻度を基に確率を推定しなくてはならない[5]。これに対し本研究のモデルでは係り側文節ごとに組合わせの出現頻度を計上しているので、ある文節の係り先を推定する際の確率の信頼度に、違いは生じてこない。ここで

[4]文節区切りの段階に当たると考えれば良い

[5]実際には比較候補があるだけ組合わせを考える必要があるが、これは現実的でない



係り受け	関係	主辞1	主辞2	距離
r1	形容詞/基本形	新しい	かまとや	1 文節
r2	形容詞/基本形	新しい	弁当	2 文節

図 2: 係り受け確率の比較対象

は、属性として各文節の主辞、係り関係 (係り側の文節が持つ属性とする)、距離、読点の有無を考え、確率 Fを以下のように展開する。

$$\begin{aligned}
& F(\Delta, \mathbf{f_j} \mid \mathbf{f_i}) && (4)\\
=\ & F(\Delta, h_j, t_j, r_j \mid \mathbf{f_i}) && (5)\\
=\ & F(h_j \mid \mathbf{f_i})\, F(r_j \mid \mathbf{f_i}) && (6)\\
& \times F(t_j, t_{ij} \mid \mathbf{f_i})\, F(\Delta \mid \mathbf{f_i}) \\
=\ & F(h_j \mid h_i, r_i) && (7)\\
& \times F(t_j, t_{ij} \mid r_i, t_i) && (8)\\
& \times F(\Delta \mid r_i, t_i) && (9)\\
& \times F(r_j \mid r_i) && (10)
\end{aligned}$$

$$t_i = \begin{cases} 1 & h_i, r_i, t_i \in \mathbf{f_i} \text{ 文節 } b_i \text{に読点が存在する時} \\ 0 & \text{それ以外} \end{cases}$$

$$t_{ij} = \begin{cases} 1 & \text{文節 } b_i, b_j \text{間に読点が存在する時} \\ 0 & \text{それ以外} \end{cases}$$

b_iとb_jの間の距離情報であるΔとしては、文節数、読点の有無、係り関係と同じ関係が文節間に現れる回数、などが尺度として考えられる。本研究では、文節間距離、間に存在する読点を距離の尺度として考えた。

(5) から (6) への変形では、生成される係り受けの属性の選択のうち距離、受け側の主辞、受け側の読点、受け側の (係り側としての) 関係は、それぞれ独立と仮定している。さらに各項に関して、依存すると考えられる属性のみを考え、展開している。(7) では、主辞の共起は、係り側の属性のうちの主辞と関係名にのみ依存すると仮定している。(8) では、受け側の属性の読点の有無は、係り側の属性のうちの関係名と係り側の読点の有無にしか依存しないと仮定している。(9) では、距離属性は係り側の属性のうち関係名と読点の有無にしか依存しないと仮定している。距離確率の条件部に読点の属性を残しているのは、距離分布は係り側の文節の読点

```
(((((((1 "1"))(4 "に"))((2 "日"))(3 "置き"))
(((((5 "ボランティア"))(6 "が"))(((((7 "食事"
))(8 "を"))((((9 "届け"))(10 "る"))))))))))(11
"。"))
```

図 3: EDR 中の例文 (bracket 情報のみを表示)

の存在によっても影響を受けると考えられるからである。(10) では、付属語間の共起 (例えば「格助詞を」と「格助詞を」の共起など) は関係それ自身で決まると仮定している。

それぞれの確率を、コーパス中での出現頻度で最尤推定し、

$$\hat{F}(h_j \mid h_i, r_i) = \frac{C(h_j, h_i, r_i)}{C(r_i, h_i)} \quad (a)$$

$$\hat{F}(t_j, t_{ij} \mid r_i, t_i) = \frac{C(t_j, t_{ij}, r_i, t_i)}{C(r_i, t_i)} \quad (b)$$

$$\hat{F}(\Delta \mid r_i, t_i) = \frac{C(\Delta, r_i, t_i)}{C(r_i, t_i)} \quad (c)$$

$$\hat{F}(r_j \mid r_i) = \frac{C(r_j, r_i)}{C(r_i)} \quad (d)$$

で近似する。ここで $C(h_j, h_i, r_i)$ はコーパス中の係り関係のうち、係り側の主辞が h_i、係り関係が r_iで、受け側の主辞が h_jであったものの頻度であり、そのほかの C も同様である。

(7), (8), (9), (10) の確率をそれぞれ、主辞共起確率、読点確率、距離確率、関係共起確率と呼ぶことにする。

例として "新しいかまどやの弁当を買う。" という文に対して属性の一部を図 2に示す。

以下この章の残りで、それぞれの確率を推定について述べる。

2.1 主辞共起確率

2.1.1 共起頻度の数え上げ

括弧付けされた EDR コーパス中の文を係り受けの正解例として利用し、

(係り関係, 係り側主辞, 受け側主辞)
(係り関係, 係り側主辞)

のそれぞれの組合わせの共起頻度をカウントした。EDR の括弧付けも完全ではなく誤りを含む。図 3 の例では格助詞「に」の位置がおかしく、このまま使用すると係り先がおかしくなる。学習データからは、このような機械的に除去可能な誤りは取り除いた。

利用可能な誤り文を除いた 207386 文のうち 1 割をテストサンプルとして残しておき[6]、残りを学習データとして用いた。

2.1.2 分類語彙表を用いたデータ量の過疎性への対処

語彙レベルの共起頻度を使用する場合、どうしてもデータ量の過疎性の問題は避けられない。そこで主辞の代りに分類語彙表 [5] を用いたスムージングを行うため、各分類語彙表 ID ごとの共起頻度もカウントした。分類語彙表 ID の上位 5 桁目までのすべての ID に対し、係り側主辞、受け側主辞それぞれ、あるいは両方に対し分類語彙表 ID を用いた時の共起頻度をカウントした。単語に対し複数の分類語彙表 ID があるときは、すべての分類語彙表 ID に 1 ずつ頻度を割り当てた。分類語彙表 ID のことを以下では BID と略す。

解析時の使用法

共起頻度の種類には、どの主辞に対して BID を用いるかによって (単語, 関係, 単語)、(BID, 関係, 単語)、(単語, 関係, BID)、(BID, 関係, BID) の 4 通りがある。以降それぞれ type1、type2、type3、type4 の共起とよぶことにする。

それぞれの type の頻度と確率の計算について、以下に示す。

$$P_{type1} = \frac{C(\text{係側単語}, \text{関係}, \text{受側単語})}{C(\text{係側単語}, \text{関係})}$$

$$P_{type2} = \frac{C(\text{係側 } BID, \text{関係}, \text{受側単語})}{C(\text{係側 } BID, \text{関係})}$$

$$P_{type3} = \frac{C(\text{係側単語}, \text{関係}, \text{受側 } BID)}{C(\text{係側単語}, \text{関係})} \times \frac{1}{D(\text{係側単語}, \text{関係}, \text{受側 } BID)}$$

$$P_{type4} = \frac{C(\text{係側 } BID, \text{関係}, \text{受側 } BID)}{C(\text{係側 } BID, \text{関係})} \times \frac{1}{D(\text{係側 } BID, \text{関係}, \text{受側 } BID)}$$

ここで D(係側単語 (BID), 関係, 受側 BID) とは、(係側単語 (BID)、関係、受側 BID) の共起データ中の、受け側単語の異なり語数である。type3 と type4 に関して Dで割っているのは、単語の頻度に直して確率を計算するためである。両タイプの共起カウントは BID を使用しているので、このままでは単語での共起頻度とは単純に比較できない。

[6] 10 文おきにテストサンプル用に確保した

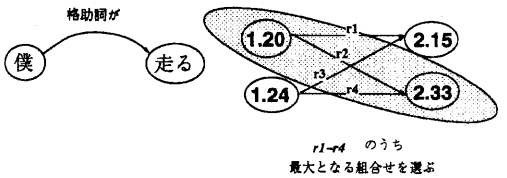


図 4: 分類語彙表 ID の選択

また type2、3、4 に関しては、一つの語彙に対して BID が複数存在する場合がある。ここでは、主辞に対し BID が複数存在する場合、それぞれの type ごとに、最も出現頻度の高い BID を選び、その BID の共起頻度を使用する (図 4)。これは適切な意味クラスに基づく意味のある共起は、自然に出現頻度が高くなると考えたからである。

分類語彙表のクラスは上位から 5 桁目までのすべての組合わせを利用した[7]。基本的には分類が細かいものから統計データを探索し、データ中に存在した場合、それを使用する。分類の細かさの尺度では順序がつけられない組み合わせに関しては、統計データ中に存在するものをすべて探索し、そのなかで確率がもっとも高いものを採用した。

2.2 距離確率の推定

各関係、係り先までの距離、候補数 (文末までの文節数) および文節間の読点の数の組み合わせに対し、共起頻度を求めた。候補数ごとに分けたのは、候補数が 2 の時に係り受け距離が 1 であるのと、候補数が 10 ある時に 1 であるのとは意味合いが異なると考えられるからである。候補数ごとに分ける別のメリットとして、文末属性を特別に考える必要がないことがあげられる。

解析時には、こうして獲得したデータのうち、同質と見なせる統計量を組み合わせて使用した。例えば候補数 10 の時の係り受け距離 1 と、候補数 11 の時の係り受け距離 1 とは同じデータとして考えられる、といった予想が成り立つ。

今回採用した距離情報使用のアルゴリズムを、以下に示す。

if 候補数 $> min$

$$\hat{C}(\text{距離} = D, \text{候補数} = K)$$

$$= \frac{\sum_{i=min}^{s} C(\text{距離} = \frac{D(K+i)}{K}, \text{候補数} = K+i)}{s}$$

$else$

$$\hat{C}(\text{距離} = D, \text{候補数} = K)$$

$$= C(\text{距離} = D, \text{候補数} = K)$$

このアルゴリズムの min および s に対して適当と思われる値を手動で求めた。具体的には、それぞれ整数値なので、0 から適当な数まで値を変えた実験を行って、適当と思われる値を採用した。

$$F(\Delta \mid r_i, t_i) \approx \frac{\hat{C}(j-i, tn_{ij}, r_i, t_i)}{C(r_i, t_i)}$$

tn_{ij} : 係り受け文節間に存在する読点の数を表す。

2.3 読点確率、関係共起確率

距離確率を求めたのと同様に、EDR 全文から除去可能なエラー文を除いた文に対して頻度を数え上げて確率を計算した。

$$F(t_j, t_{ij} \mid r_i, t_i) \approx \frac{C(t_j, t_{ij} r_i, t_i)}{C(r_i, t_i)}$$

$$F(r_j \mid r_i) \approx \frac{C(r_j, r_i)}{C(r_i)}$$

3 システム概要

システムの概要を図 5に示す。基本的な解析の流れは、

1. JUMAN (ChaSen) による形態素解析
2. 品詞タグ見出し語等をもとにした、文節区切りと属性の決定
3. 各文節間の係り受けの確率の計算
4. 係り受けの制約を満たした上で、確率最大となる係り受けの組合わせを決定

となる。

以下この章の残りで、2、4番目の文節区切りと解析アルゴリズムについて説明する。

3.1 文節区切り

基本的に自立語列+付属語列を文節とし、文節属性は、主辞は最後の自立語、関係名は付属語列の見出し語と品詞 (活用があるものは活用形も含む) とした。ただし例外や使用者による文節単位の考え方

[7] 20 万文の学習データの場合は、データ量の問題で今回は 4 桁目までしか使用してない

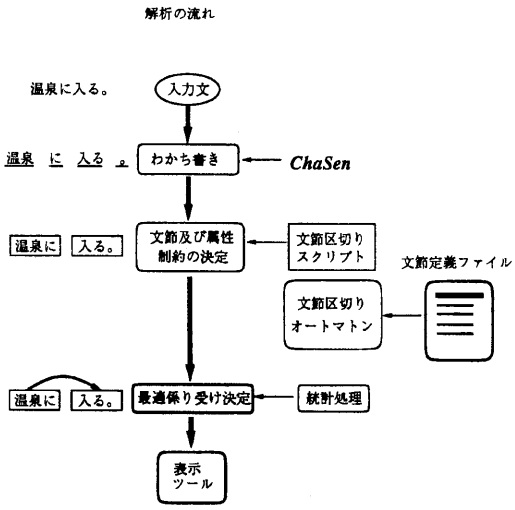


図 5: システム概要

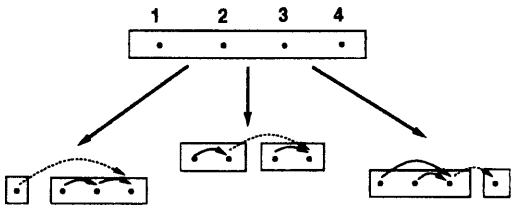


図 6: CYK アルゴリズム

の違いに対処するため、本システムでは、見出し語や品詞を使用した正規表現を用いた文節および属性の定義を定義ファイルに記述できるようにした。これにより、名詞性名詞接尾辞は主辞候補であるとか、ある付属語は関係名とはならない、あるいはこの正規表現中ではこの自立語は主辞となり得ない、といった設定をすることが可能となっている。

3.2 解析アルゴリズム

第2章で説明した統計をもとに、各文節間の係り関係の確率を求める。この中から、

1. 文末を除き各文節は文末側に必ず一つの係り先を持つ
2. 係り受けは非交差

という制約の元で確率最大となる係り受けパターンを決定する。これはCYKアルゴリズムを使用して効率的に決定できる。

図6に、1文節から4文節までの係り受けの最適組合わせを求める場合の例を示す。決定済みの2つの組み合わせることは、その主辞間に係り関係を作ることに対応する。

4 実験

4.1 実験方法

隣にかけた場合、共起情報に品詞情報のみを使用した場合、語彙情報と分類語彙表を組み合わせた場合について実験を行った。学習サイズは1万文と20万文で行い、学習に使用しなかったEDR中の1000文,10000文についてそれぞれ実験を行った。

4.2 定量的評価

表1と表2に、1万文で学習した場合と、20万文で学習した場合の実験結果を示す。EDRから自動で抽出した正解例をもとに正解率を求めた。表中括弧内の数字が正解率を表す。表中の記号の意味は以下のとおりである。

- **b**: 共起情報として、品詞のみを使用したもの
- **s**: 単語レベルの共起頻度と、その分類語彙表によるスムージングを利用したもの
- **t**: すべての文節を隣りにかけた場合
- **-i,-r,-d,-t**: それぞれ主辞共起、関係共起、距離、読点確率を使わないことを表す

テストデータ量	1000文	10000文
係り受け総数	7254	74886
実験b	5664(0.781)	57900(0.7744)
実験s	5846(0.806)	57958(0.7739)

正解数 (正解率)

表 1: 学習データ1万文による実験結果

どの属性がどの程度有用なのかわかるように、各情報を省いて行った実験を、表2の下段に示した。距離情報がかなり有用であることがわかる。いずれの属性も、他の属性を阻害してはいないことがわかる。

4.2.1 各係り関係ごとの評価

各係り関係およびごとの正解率の一部を図7,8に示す。一番左が関係名を表す。関係名に'0'や'1'がついているものは、それぞれ係り側文節に読点がある場合と、ない場合を区別して正解率を求めたものである。全般的には、格助詞の一部や活用語など、

テストデータ量	1000 文	10000 文
係り受け総数	7254	74886
実験 b	5881(0.811)	60173(0.8035)
実験 s	5947(0.820)	60267(0.8048)
実験 t	4474(0.617)	45081(0.6020)

正解数 (正解率)

使用情報を変えた実験結果 (1000 文)			
実験 s-td	5031(0.694)	実験 s-r	5735(0.790)
実験 s-ti	5752(0.793)	実験 s-i	5859(0.808)
実験 s-tr	5656(0.780)	実験 s-d	5620(0.775)
実験 s-di	5346(0.737)	実験 s-t	5848(0.806)
実験 s-dr	5318(0.733)	実験 s-rid	4981(0.687)
実験 s-ir	5595(0.771)	実験 s-rit	5543(0.764)
		実験 s-idt	4465(0.616)

正解数 (正解率)

表 2: 学習データ 20 万文による実験結果

読点が存在すると精度が著しく下がっていた。副助詞「は」の場合は逆であった。

一部の関係については 97%以上の正解率が出ている。「格助詞が」「動詞の連用形」「副助詞は」などは頻度が多いので、解析精度を大きく左右すると思われるが、正解率はあまり高くない。したがって、これらの係り関係の精度をあげれば全体の精度も上がることが、容易に予想される。

4.3 定性的評価

解析失敗例の一部を、図 9に示す。正解率においても見られたように、読点を含む例 (特に並列構造) の解析に失敗することが多かった。JUMAN の出力間違い (引用助詞の「と」が格助詞の「と」と解析された) によるものもみられた。また EDR そのものの誤りから不正解となったものもあった。EDR の誤りには、例文ごとに方針が違っていることによる" 揺らぎ" が原因の場合がある。たとえば「～から... へ」という例文の場合、EDR の正解例が「から」の係り先が主動詞の場合と「～へ」に係っている場合の両方みられた。ただ「～へ」に係っている場合、「～へと」や「～への」などのように他の格助詞を伴っていることが多いので、関係共起頻度を求める際の受け側属性をより細かくすることで、ある程度対処できる。「を」格が本来「中心に」に係るべきところが、主動詞に係っている例があるが、これは関係共起頻度が邪魔をした例と言える (-r で解析すると、正しくなるので)。"自分と同じ関係が現れる回数" という距離尺度も有用かも知れない。

```
キロ/名詞性名詞助数辞/0 総数:29 正解数:29 正解率:1
時/名詞性名詞助数辞/0 総数:58 正解数:58 正解率:1
じゃ/副助詞/ 総数:26 正解数:26 正解率:1
な/終助詞/ 総数:83 正解数:81 正解率:0.97
ぬ/助動詞/条件 総数:36 正解数:35 正解率:0.97
的/名詞性名詞接尾辞/ 総数:201 正解数:198 正解率:0.99
%/名詞性名詞助数辞/0 総数:67 正解数:67 正解率:1
人/名詞性名詞接尾辞/ 総数:52 正解数:52 正解率:1
円/名詞性名詞助数辞/0 総数:88 正解数:87 正解率:0.99
用/名詞性名詞接尾辞/0 総数:71 正解数:70 正解率:0.98
な/終助詞/0 総数:82 正解数:80 正解率:0.98
割/名詞性名詞助数辞/0 総数:23 正解数:23 正解率:1
化/名詞性名詞接尾辞/0 総数:206 正解数:203 正解率:0.99
```

図 7: 頻度 20 以上で正解率 97 %以上だったもの (outside10000 文より)

```
から/格助詞/ 総数:1844 正解数:1238 正解率:0.671
が/格助詞/ 総数:11820 正解数:9039 正解率:0.765
が/格助詞/0 総数:11077 正解数:8553 正解率:0.772
が/格助詞/1 総数:743 正解数:486 正解率:0.65
で/格助詞/ 総数:2466 正解数:1794 正解率:0.727
で/格助詞/0 総数:2445 正解数:1776 正解率:0.726
で/格助詞/1 総数:21 正解数:18 正解率:0.9
と/格助詞/ 総数:4289 正解数:3478 正解率:0.811
と/格助詞/0 総数:4133 正解数:3429 正解率:0.830
と/格助詞/1 総数:156 正解数:49 正解率:0.3
に/格助詞/ 総数:13160 正解数:10961 正解率:0.833
に/格助詞/0 総数:12525 正解数:10538 正解率:0.8414
に/格助詞/1 総数:635 正解数:423 正解率:0.67
へ/格助詞/ 総数:268 正解数:228 正解率:0.85
へ/格助詞/0 総数:258 正解数:224 正解率:0.87
へ/格助詞/1 総数:10 正解数:4 正解率:0.4
まで/格助詞/ 総数:342 正解数:258 正解率:0.75
より/格助詞/ 総数:172 正解数:101 正解率:0.59
を/格助詞/ 総数:15626 正解数:13399 正解率:0.8575
を/格助詞/0 総数:15402 正解数:13253 正解率:0.8605
を/格助詞/1 総数:224 正解数:146 正解率:0.65
のみ/副助詞/ 総数:11 正解数:10 正解率:0.91
は/副助詞/ 総数:16060 正解数:11708 正解率:0.729
は/副助詞/0 総数:10975 正解数:7739 正解率:0.705
は/副助詞/1 総数:5085 正解数:3969 正解率:0.781
/動詞/連用 総数:7441 正解数:5137 正解率:0.690
```

図 8: 主な係り関係ごとの正解率

5 おわりに

距離属性一つとっても、係り受け関係間の読点の存在を属性とするか、あるいは実際に現れた個数を属性にするかなど考えられる属性はいくつもある。また定性的評価のところでも述べた通り、考慮すべき付属語列をどこまでとするか (今回の実験では、関係共起確率を計算するときの受け側の付属語列は、一番内側のものだけみている) は定かではない。

今までは有効と思われる属性を直感で決定し、システムの構築と評価を行っていたが、考慮する属性が増えるに従い、効率が悪くなる。例えば、他に有用と思われる尺度として、白井 [10] で述べられている付属語列間の順序関係などがあげられるが、今回は付属語列間の関係については、共起のみに着目している。

今後の予定として、考え得る様々な属性のうち、

なにが有効かを決定するため(各関係ごとに違っていてかまわない)、複数の属性の集合から有効な属性を自動抽出できないかと考えている。

謝辞

統計モデルに関して貴重な指摘をいただいた、東京工業大学の乾健太郎氏、奈良先端大学の宇津呂武仁氏に感謝致します。またCYKの実装をしてくれた奈良先端大学の平野善隆君、パトリシアツリーについて色々教えてくれた米沢恵司君に感謝致します。

参考文献

[1] 荒川直哉, 竹澤寿幸, 森元逞. 統計的手法による部分木併合. 信学技報, May 1996.

[2] Michael John Collins. A new statistical parser based on bigram lexical dependencies. *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*, pp. 184–191, June 1996.

[3] 日本電子化辞書研究所,EDR電子化辞書仕様説明書. 1995.

[4] 乾健太郎, 白井清昭, 徳永健伸, 田中穂積. 種々の制約を統合した統計的日本語文解析. 情報処理学会自然言語処理研究会, Nov 1996.

[5] 国立国語研究所: 分類語彙表, 秀英出版, 1964,1993.

[6] John Lafferty, Daniel Sleator, and Davy Temperley. Grammatical trigrams: A probabilistic model of link grammar. *Proceedings of theAAAIConference on Probabilistic Approaches to Natural Language*, pp. 89–97, October 1992.

[7] 李航. 心理言語学原理に基づいた確率的曖昧性解消法. コンピュータソフトウエア Vol.13 No.6, Nov 1996.

[8] Eva Wai man Fong and Dekai Wu. Learning restricted probabilistic link grammars. *IJCAI-95 Workshop on New Approaches to Learning Natural Language Processing*, pp. 49–56, Aug 1995.

[9] 松本裕治, 今一修, 山下達雄, 北内啓, 今村友明. 日本語形態素解析システム『茶筌』version 1.0b5 使用説明書. 奈良先端科学技術大学院大学 松本研究室, 1996.

[10] 白井諭, 池原悟, 横尾昭男, 木村淳子. 階層的認識構造に着目した日本語従属節の係り受け解析の方法とその精度. 情報処理学会論文誌, Oct 1995.

[11] 田上敦士, 田辺利文, 冨浦洋一, 日高達. 限定頻度を考慮した確率文脈自由文法. 信学技報, Oct 1996.

[12] W.R.Hogenhout and Y.Matsumoto. Experiments with using semantical categories in parsing systems. 言語処理学会第2回年次大会, March 1996.

[13] 安原宏. 縮退型共起関係を用いた学習機能付き係り受け解析システム. 自然言語処理, Oct 1996.

```
高めた。
├賃金を
├利潤と＊                    (JUMAN の間違い)
├まわし、
│ ├産業に
│ │ └他の
│ └１万２０００人を
│   └２年間に＊
├合理化で＊
└鉄道局では、
  ├ある
  └抱えた
    └人員を
      └１０万人の
わかった。
├高床式宮殿と
├構造などから、
│ └柱の
│   └呼ばれる
│     └総柱と
│       └焼け具合、＊          (並列構造)
│         └基壇前面の
│           └見つかったことや、＊
│             ├木材が
│             │ └炭化した
│             └中央階段付近から

持ってきた。
├影響力を
├東チベットに
├中心に
├歴代ラサを＊                  (関係共起頻度)
└ダライ・ラマは、
  └誕生した
    └１６世紀後半に
する。
├仲間入り
│ └構造不況業種の＊            (EDR 誤り)
├大学は、
└１０年後、
ある。
├急変しつつ
├仁川は
└１５０万都市、＊              (並列句)
順調という。
├経過は
└移るなど、
  ├一般病室へ
  │ └集中治療室から＊
  └なって
    └１４日に
```

図 9: 解析失敗例 (＊は係り先が間違っていることを表す