形態素・構文解析の曖昧性解消に対する種々の統計情報の貢献度に関する考察

白井 清昭　徳永 健伸　田中 穂積

東京工業大学 大学院情報理工学研究科

1　はじめに

形態素解析や構文解析の曖昧性解消に統計情報を利用した研究は近年盛んに行われている．日本語を対象にした構文解析についても，統計情報の利用を試みた研究が数多く行われ，その成果が報告されている [1, 2, 3, 14]．このような統計的構文解析の結果を評価する際には，例えば第一位の候補が正解となる文の割合や係り先が正しく認定された文節の割合など，主に定量的な評価が行われている．しかしながら，統計的手法を改善し，曖昧性解消の精度を向上させるためには，このような定量的な評価だけでは不十分である．統計情報が構文解析の精度向上にどのように寄与するのか，また構文解析に失敗した場合にその原因は何であるのか，といったいわゆる定性的な評価を行うことが必要不可欠である．このような定性的な評価は，取り扱う統計情報の種類が増えれば増えるほど，複数の要因が複雑に絡み合って解析結果の候補の優先順位を決定するために，一般に難しくなる．本研究は，品詞並びの優先度，構文的優先度，単語の出現頻度，単語の共起関係などの統計情報を同時に利用して形態素および構文解析を行い，その解析結果を詳細に分析することにより，これらの統計情報，特に語彙的な統計情報 (単語の出現頻度，単語の共起関係) が解析精度にどのような影響を与えるかについて考察することを目的とする．

本研究では，複数の統計情報を同時に取り扱うモデルとして統合的確率言語モデル [10, 11] を利用する．統合的確率言語モデルは，複数の統計情報を別々のモデルとして個別に学習するため，個々の統計情報の働きを比較的容易に分析できると考えられる．2節ではこのモデルの概要を述べ，3節ではこれを用いた形態素・構文解析実験について説明する．4節では，分かち書き，品詞付け，文節切り，文節の係り受け解析の精度向上に対する各種統計情報の貢献度を調べ，また解析事例を調査することによって明らかになった各種統計情報の働きや問題点についてまとめる．最後に，5節で本研究のまとめと今後の課題について述べる．

2　統合的確率言語モデル

統合的確率言語モデル [10, 11] は，形態素解析と構文解析を同時に行うことを前提に，その解析結果の生成確率を与えるモデルである．以下，この統合的確率言語モデルの概要を簡単に説明する．

図1は，入力文「彼女が哲学の本を読んだ」に対する形態素解析・構文解析結果の一例である．図1において，A は入力文字の集合，W は単語集合，L は品詞集合，R は入力文の構文構造 (構文木) を表わす．

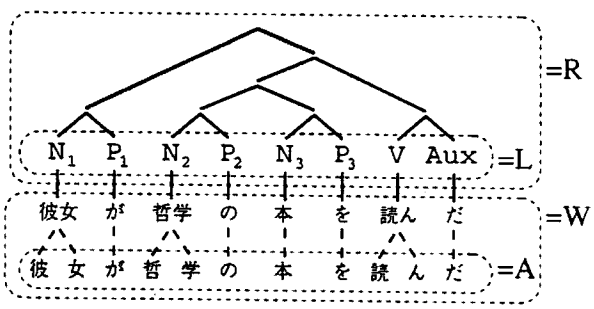

図 1: 解析結果の例

形態素・構文解析を行った場合，図1のような解析結果の候補は一般に多数存在する．ここで，同時分布 $P(R, L, W, A)$ を推定し，この値によって最も尤もらしい解析結果の候補を選択することにより，曖昧性を解消することを考える．統合的確率言語モデルでは，$P(R, L, W, A)$ を以下のように推定する．

$$
\begin{aligned}
& P(R, L, W, A) \\
&= P(R) \cdot P(L|R) \cdot P(W|R, L) \cdot P(A|R, L, W) \quad (1) \\
&= P(R) \cdot P(W|R, L) \quad (2) \\
&\simeq P(R) \cdot P(W|L) \cdot D(W|R) \quad (3)
\end{aligned}
$$

式 (1) から式 (2) への変形は，R から L は一意に決まるので $P(L|R)=1$ となり[1]，W から A は一意に決まるので $P(A|R,L,W)=1$ となることから導かれる．式 (3) に示すように，$P(R,L,W,A)$ は以下の 3 つのサブモデルの積により推定される[2]．

1. 構文モデル $P(R)$
   構文木 R の生成確率である．

2. 単語導出モデル $P(W|L)$
   単語集合 W の生成確率であり，各単語 w_i の文脈自由な導出確率の積として推定する．

$$P(W|L) \simeq \prod_{w_i \in W} P(w_i|l_i) \qquad (4)$$

   図 1 の例では，単語導出モデルは，品詞 N_1(名詞) から単語 "彼女" が生成される確率 $P(\text{彼女}|N_1)$ や品詞 P_1(助詞) から単語 "が" が生成される確率 $P(\text{が}|P_1)$ などの積によって推定される．

3. 従属係数モデル $D(W|R)$
   単語の共起関係を反映した統計量であり，各単語 w_i と単語生成文脈 c_{ij} の従属係数の積として推定される．

$$D(W|R) = \prod_{c_{ij} \in C} D(w_i|l_i[c_{ij}]) \qquad (5)$$

   ここで，単語生成文脈 c_{ij} とは，単語 w_i の導出に深く関わりがあると思われる入力文中の他の単語を指す．また，式 (5) の各項 $D(w_i|l_i[c_{ij}])$ は従属係数と呼ばれ，c_{ij} と共起するという条件の下で品詞 l_i から単語 w_i が導出される確率 $P(w_i|l_i[c_{ij}])$ と，そのような制約なしに品詞 l_i から単語 w_i が導出される確率 $P(w_i|l_i)$ の比として定義される (式 (6))．

$$D(w_i|l_i[c_{ij}]) \stackrel{def}{=} \frac{P(w_i|l_i[c_{ij}])}{P(w_i|l_i)} \qquad (6)$$

   したがって，$D(w_i|l_i[c_{ij}])$ は，w_i と c_{ij} に正の相関関係があれば 1 より大きい値を取り，負の相関関係があれば 1 より小さい値を取り，何も相関関係がなければ 1 に近い値を取る．
   例えば，図 1 において，"本" という単語は "読む" という動詞のヲ格の格要素であることから，"本" の単語生成文脈が「を，読む」であるとする．このとき，$D(\text{本}|N_2[\text{を},\text{読む}])$ という従属係数が $D(W|R)$ の要素となるが，これは "本" という単語が "読む" という動詞のヲ格の格要素としてどの程度現われやすいか，言い替えればヲ格の主辞と格要素として "読む" と "本" がどの程度共起しやすいかを表わしている．
   単語生成文脈は一つの単語について一般に複数存在する．例えば，図 1 の "本" という単語は，"哲学" という単語から連体修飾を受けているので，「の，哲学」という単語生成文脈も持つと考えられる．このとき，$D(\text{本}|N_2[\text{を},\text{読む}])$ と $D(\text{本}|N_2[\text{の},\text{哲学}])$ という 2 つの従属係数が考慮される (後者は "哲学" と "本" の共起のしやすさを表わす)．また，単語 w_i に対する単語生成文脈 c_{ij} をどのように決定するかについては，自動的に学習する手法なども考えられるが，本研究では人手で作成した規則によって一意に決定する．

統合的確率モデルの特長を以下に挙げる．

- 各モデルを別々の言語資源から独立に学習できる

  例えば，構文モデル $P(R)$ を構文木付きコーパスから，単語導出モデル $P(W|L)$ を品詞付きコーパスから独立に学習することができる．一般に，統計情報の種類によって，学習に必要な言語資源の質や量が異なるため，個々の統計情報は別々な言語資源から個別に学習できることが望ましい．

- 解析精度における各種統計情報の貢献度を容易に見積ることができる

  ある一つの統計情報を反映したサブモデルを全体のモデルに加えた場合と加えない場合とで解析精度を比較することにより，その統計情報が形態素・構文解析の精度向上にどの程度貢献しているかを見積ることができる．また，種々の統計情報がそれぞれ別々のモデルに反映されているため，例えば解析がうまくいかなかった場合に，どの統計情報に原因があるかを容易に調査することができる．

3 実験

本節では，2 節で説明した統合的言語モデルを実際に言語データから学習し，それを用いた形態素・構文解析実験について述べる．

[1] ここでは，構文木 R は品詞を葉とすることを前提にしている．
[2] 式 (2) から式 (3) への変形の詳細については文献 [11] を参照．

3.1 統合的確率言語モデルの学習

3.1.1 構文モデル $P(R)$ の学習

今回の実験では，構文モデルとして確率一般化LRモデル [5, 12, 13] (以下PGLRモデル) を使用した．PGLRモデルとは，構文解析アルゴリズムのひとつである一般化LR法において，構文解析時に実行されるLR表の各アクションの実行確率を推定し，R を生成するために用いられたアクションの実行確率の積によって構文木 R の生成確率 $P(R)$ を推定する確率モデルである．

まず，京大コーパス [6] から，文長が50以下の例文164,950文をランダムに選択し，これを訓練データとした．この訓練データから，品詞を前終端記号とし，文節の区切りと文節の係り受け関係を表わす構文木を生成する文脈自由文法を自動獲得した．さらに，得られた文脈自由文法からLR表を作成し，これと訓練データに付加された構文木からPGLRモデルを学習した．

3.1.2 単語導出モデル $P(W|L)$ の学習

単語導出モデルの学習にはRWCコーパス [9] を使用した．RWCコーパスは，毎日新聞の91年から95年の新聞記事に対して品詞タグを付与した品詞付きコーパスである．しかしながら，構文モデルの学習に用いた京大コーパスでは形態素解析ツールJUMAN [7] の品詞体系が用いられており，RWCコーパスで使われている品詞体系とは異なる．そのため，JUMANを用いてRWCコーパスの例文を形態素解析し，その結果から単語の導出確率 $P(w_i|l_i)$ を最尤推定した．但し，導出確率の前件 l_i には，JUMANの品詞体系の大分類 (動詞，名詞，形容詞など) を使用した．

3.1.3 従属係数モデル $D(W|R)$ の学習

従属係数モデルに反映させる単語の共起関係として以下の5種類を考慮し，それぞれに対応した従属係数を学習した．

1. $D(n|N[p,v])$
   格要素となる名詞 n と，格 p 及びその主辞となる動詞 v との間の共起関係を反映した従属係数である．今回の実験では，RWCコーパスから，名詞 n が助詞 p を伴って動詞 v に係る共起事例 (n,p,v) を収集し，この従属係数を学習した．また，動詞については分類語彙表 [15]，名詞については日本語語彙体系 [4] の意味クラスを用いたスムージングを行った．
2. $D(p_1,\cdots,p_n|P_1\cdots P_n[v])$
   n 個の助詞 $p_1\cdots p_n$ が同じ動詞の格としてどれだけ共起しやすいか (格間の従属関係)，及び個々の助詞 p_i と動詞 v がどれだけ共起しやすいか (格と主辞との間の従属関係) を表わす従属係数である．図1の例では，助詞 “が” と “を” が同じ動詞 “読んだ” の格となっているので，$D(\text{が},\text{を}|P_1P_2[\text{読む}])$ という従属係数が全体のモデルに加えられる．
   今回の実験では，EDRコーパス [8] から，n 個の助詞 $p_1\cdots p_n$ が同じ動詞 v に係るという共起データを収集し，この従属係数を学習した．
3. $D(p_1,\cdots,p_n|P_1\cdots P_n[Adj/NounPred])$
   主辞が形容詞または名詞述語である場合の格間の従属関係を表わす従属係数である．Adj と $NounPred$ は，それぞれ主辞が形容詞，名詞述語であることを表わすシンボルである．すなわち，主辞が動詞の場合と異なり，格と主辞との間の従属関係は無視している．これは，格と主辞との間の従属関係を学習するための十分な訓練データが得られなかったためである．この従属係数は，主辞が動詞の場合と同様に，EDRコーパスから収集された共起事例をもとに学習した．
4. $D(n_2|N[\text{の},n_1])$
   「n_1 の n_2」というように，2つの名詞 n_1, n_2 が，前者が後者を連体修飾するという形でどの程度共起しやすいかを表わす従属係数である．今回の実験では，RWCコーパスから，名詞 n_1 が助詞「の」を伴って名詞 n_2 を連体修飾する共起事例を収集し，この従属係数を学習した．
5. $D(n|N[touten, mod\_type])$
   名詞の係り先に関する従属係数である．ここで，$touten$ は名詞 n の直後に読点があるか否かを表わし，mod_type は名詞 n が連用修飾するか連体修飾するかを表わす．この従属係数は，例えば時を表わす名詞が読点を伴なう場合には副詞的名詞として用言に係りやすく (ex.「三月、東京を訪れた」)，読点を伴わない場合には他の名詞を連体修飾して複合語を構成しやすい (ex.「三月三日はひなまつりだ」) ことを考慮している．今回の実験では，EDRコーパスから，名詞が連用または連体修飾するか，その際読点を伴うか伴わないかといった共起事例 $(n, touten, mod\_type)$ を収集し，この従属係数を学習した．
   この従属係数は単語間の共起関係を反映したもので

表 1: 構文モデルに関する実験結果

(+D +L)	−S	+S	
係り受け正解	3.50%	14.81%	(+11.31)
文節切り正解	15.43%	29.22%	(+13.79)
品詞付け正解	23.05%	36.01%	(+12.96)
単語切り正解	77.98%	83.13%	(+5.15)

表 2: 単語導出モデルに関する実験結果

(+S +L)	−D	+D	
係り受け正解	8.85%	14.81%	(+5.96)
文節切り正解	18.52%	29.22%	(+10.7)
品詞付け正解	21.81%	36.01%	(+14.2)
単語切り正解	76.54%	83.13%	(+6.59)

表 3: 従属係数モデルに関する実験結果

(+S +D)	−L	+L	
係り受け正解	8.85%	14.81%	(+5.96)
文節切り正解	27.37%	29.22%	(+1.85)
品詞付け正解	35.39%	36.01%	(+0.62)
単語切り正解	82.72%	83.13%	(+0.41)

はないが，構文モデルや単語導出モデルでは考慮されない統計情報であるので，従属係数モデルの要素としてモデル全体に反映させた．

3.2 形態素・構文解析実験

京大コーパスから文長が 50 以下の例文 500 文をランダムに選択し，これをテスト文とした．これらのテスト文の中には，3.1.1 の構文モデルの学習に用いた例文は含まれていない．テスト文の平均文長は 32.74 である．

MSLR パーザを用いて，これらのテスト文の形態素解析および構文解析を行った．MSLR パーザは，形態素・構文解析を同時に行えるように一般化 LR パーザを拡張した解析ツールである[3]．文法は，3.1.1 の構文モデルの学習の際に作成した文脈自由文法を使用した．したがって，今回の実験での構文解析は文節間の係り受け解析を行っていることに等しい．一方，辞書は，3.1.2 の単語導出モデルの学習の際に用いた形態素・品詞対の集合をそのまま用いた．

テスト文 500 文のうち，全体の 2.8%にあたる 14 文が文法や辞書の不備により解析に失敗した．残りの 486 文 (97.2%) については，1 つ以上の形態素・構文解析結果が得られた．次節では，これらの解析結果の評価を行う．

4 考察

4.1 各サブモデルの貢献度

まず，統合的確率言語モデルを構成する 3 つのサブモデル (構文モデル，単語導出モデル，従属係数モデル) が形態素・構文解析の精度向上にどの程度貢献するかを調査した．ここでは，これらのサブモデルを組み合わせたいくつかの確率モデルについて，確率モデルが 1 位の生成確率を与える構文木と京大コーパスに付加された形態素・構文情報を比較し，形態素解析及び構文解析の正解率を調べた．結果を表 1, 2, 3 に示す．

[3]以下の URL にてフリーで公開されている．
http://tanaka-www.cs.titech.ac.jp/pub/mslr/index.html

表 1, 2, 3 において,「単語切り正解」は単語区切り (分かち書き) が正しい文の割合を,「品詞付け正解」は単語区切りと単語に付加された全ての品詞が正しい文の割合を表わす.「文節切り」は，形態素解析に成功し，さらに文節切りが正しい文の割合を表わし,「係り受け正解」はさらに文節間の係り受け関係も正しい文の割合を表わしている．以上 4 つの評価基準について，これらを満たす文の集合は以下のような包含関係を持つことに注意していただきたい．

$$\text{係り受け正解} \subset \text{文節切り正解} \subset \text{品詞付け正解} \subset \text{単語切り正解}$$

表 1 において，“−S” の列は解析結果の生成確率を計算する確率モデルとして $P(W|L) \cdot D(W|R)$ を用いた場合 (構文モデルを考慮しない場合)，“+S” の列は $P(R) \cdot P(W|L) \cdot D(W|R)$ を用いた場合 (構文モデルを考慮した場合) の正解率を示している．また，( ) 内は両者の差を表わしている．この結果から，全ての正解基準について，構文モデルをモデル全体に加えることによって正解率が大きく向上することがわかる．構文モデルには，品詞の連接関係，構文的優先度，距離に関する優先度などの統計情報が反映されていると考えられるが，これらの統計情報が形態素・構文解析に大きく貢献することがわかる．

表 2 は，表 1 と同様に，解析結果の生成確率を計算する確率モデルとして，単語導出モデルを加えない場合 (“−D”) と加えた場合 (“+D”) の正解率を比較している．この結果から，単語導出モデルは特に文節切りや品詞付けの正解率の向上に大きく貢献することがわかる．単語

導出モデルには単語の出現頻度に関する統計情報が反映されていることから，単語の出現頻度が文節切りや品詞の特定に有効に働くことがわかる．

表3は，解析結果の生成確率を計算する確率モデルとして，従属係数モデルを加えない場合 ("−L") と加えた場合 ("+L") の正解率を比較している．この結果から，従属係数モデルは係り受け解析の正解率の向上には貢献するが，他の評価基準による正解率にはあまり貢献していないことがわかる．従属係数モデルには単語共起に関する統計情報が反映されていることから，単語の共起情報が文節間の係り受け解析に特に有効であると結論できる．

直観的には，構文モデルは構文解析の曖昧性解消に，単語導出モデルは形態素解析の曖昧性解消に有効であると考えられる．しかし，実際には構文モデルは形態素解析の曖昧性解消にも有効であるし，単語導出モデルは構文解析の曖昧性解消にも有効である．したがって，統計情報を用いて曖昧性を解消する場合には，形態素解析と構文解析をそれぞれ独立に行うよりも，両者を同時に行った方が望ましいと言える．

4.2 解析事例の分析

本項では，個々の文の解析結果を調べ，統合的確率言語モデルが取り扱う各種統計情報が形態素・構文解析の曖昧性解消にどのように働くか，また有効に働かない要因として何があるのかを考察する．ここでは，3つのサブモデルのうち，特に単語導出モデルと従属係数モデルの2つに注目した．

4.2.1 単語導出モデルの分析

テスト文500文のうち，単語導出モデルをモデル全体に加えた場合 (表2の "+D") と加えない場合 (表2の "−D") とで1位の構文木が異なる文に注目し，解析結果の詳細な分析を行った．その考察を以下にまとめる．尚，これ以降に挙げる例文において，T_n^{+D} は単語導出モデルを考慮した場合，T_n^{-D} は単語導出モデルを考慮しなかった場合に1位となった解析結果を表わす．但し，n は例文番号を表わす．

単語導出モデルが分かち書きに影響を与える場合

単語導出モデルをモデル全体に加えることにより，分かち書きが正しくなった例がいくつか見られた．特に，単語数の少ない解析結果の候補に対して高い生成確率を与える傾向が見られた．例えば，以下の例文1において，単語導出モデルは，「である」という文字列が助詞 "で" と動詞 "ある" という2つの単語から構成されるという解析結果 T_1^{-D} より，判定詞 "である" という1つの単語から構成されるという解析結果 T_1^{+D} に対して高い確率を与えている．

[例文1] 政・官・業のトライアングルの再構築である。

$< T_1^{-D} >$	$< T_1^{+D} >$
で (助詞), ある (動詞)	である (判定詞)
P(で | 助詞) = 0.0675 P(ある | 動詞) = 0.0212	P(である | 判定詞) = 0.112

また，単語数が同じ場合でも，分かち書きが正しい解析結果に高い生成確率を与える場合もある．以下の例文2において，「第一次分」という文字列の生成確率は，正しい分かち書きの場合 (T_2^{+D}) とそうでない場合 (T_2^{-D}) とを比べると，約140倍の差がある．

[例文2] 一万八千ドルを第一次分として届けた。

$< T_2^{-D} >$	$< T_2^{+D} >$
第一 (副詞), 次 (名詞) 分 (名詞)	第 (接頭辞), 一次 (名詞) 分 (接尾辞)
P(第一 | 副詞) = 0.0112 P(次 | 名詞) = 0.000296 P(分 | 名詞) = 0.000357	P(第 | 接頭辞) = 0.101 P(一次 | 名詞) = 0.000114 P(分 | 接尾辞) = 0.0142

逆に，単語導出モデルを考慮することによって分かち書きが正しくない解析結果に対して高い生成確率を与えた例も見られたが，今回の実験で調べた限りにおいては数はあまり多くなかった．また，以下の例のように，どのように分かち書きするべきかが明確でない場合もあった．

[例文3] 医学部全体では例がないという。

$< T_3^{-D} >$	$< T_3^{+D} >$
医 (名詞), 学部 (名詞)	医学 (名詞), 部 (名詞)
P(医 | 名詞) = 0.000221 P(学部 | 名詞) = 0.000284	P(医部 | 名詞) = 0.000097 P(学 | 名詞) = 0.00121

単語導出モデルが品詞付けに影響を与える場合

ある単語が「名詞」と「接尾語」という2つの品詞を持つ場合，単語導出モデルを全体のモデルに加えない場合には「接尾語」となる単語が誤って「名詞」と認識されていても，単語導出モデルを加えることにより品詞が正しく認識される場合が多い．以下に例を示す．

[例文4] 批准書は、同年十二月十八日に交換された。

$< T_4^{-D} >$	$< T_4^{+D} >$
日 (名詞)	日 (接尾辞)
P(日 | 名詞) = 0.00284	P(日 | 接尾辞) = 0.0714

T_4^{-D} と T_4^{+D} では，ともに “十八”,“日”,“に” の 3 つの単語が 1 つの文節を構成し，T_4^{-D} では「名詞 名詞 助詞」，T_4^{+D} では「名詞 接尾辞 助詞」という品詞列になっている．しかしながら，今回使用した構文モデルは，後者のような名詞と接尾辞によって構成される文節を含む構文木よりも，前者のような名詞と助詞だけで構成される文節を含む構文木に対してより高い確率を与える傾向が見られた．これは，訓練コーパスにおいて，文節を構成する品詞並びとして後者より前者のパタンの方が多く出現したためと考えられる．したがって，単語導出モデルをモデル全体に加えない場合には，T_4^{-D} に対して高い生成確率が与えられる．この誤りは，単語導出モデルをモデル全体に加えることによって補正することができる．一般に,「名詞」を品詞として持つ単語の異り数は「接尾辞」を品詞として持つ単語の異り数よりも多いので，$P(w|\text{名詞}) \ll P(w|\text{接尾辞})$ となり，品詞を接尾辞とする解析結果が優先されるからである．

逆に，単語導出モデルを用いて品詞付けを間違えた以下のような例もある．

[例文 5] 医師の応診を受けた後、休養した。

$< T_5^{-D} >$	$< T_5^{+D} >$
後 (名詞)	後 (接尾辞)
P(後 | 名詞) = 0.000453	P(後 | 接尾辞) = 0.00793

T_5^{+D} は,「受けた後」で一つの文節を作るという不自然な解析結果であるが，P(後 | 名詞) $\ll P$(後 | 接尾辞) であるために，この解析結果に最も高い生成確率を与えている．しかしながら，このような事例は，品詞付けを正しく修正した例に比べてはるかに数が少なかった．

単語導出モデルが文節切り，文節の係り受けに影響を与える場合

単語導出モデルをモデル全体に反映させることによって文節切りを正しく認識することができた例を以下に示す．

[例文 6] 飲み干すと隣の家に移る。

$< T_6^{-D} >$	$< T_6^{+D} >$
飲み干す┐ と/隣/の┐ 家/に┐ 移る/。	飲み干す/と──┐ 隣/の┐ | 家/に┤ 移る/。
P(と | 名詞) = 0.0000148	P(と | 助詞) = 0.0905

T_6^{-D} は，“と” と “隣” がともに名詞であり，これらが複合名詞となって 1 つの文節を構成するという誤った解析結果であるが，単語導出モデルをモデル全体に加えることにより，“と” が助詞であり，かつ “隣” とは別の文節を構成するという正しい解析結果 T_6^{+D} に高い生成確率が与えられる．

単語導出モデルをモデル全体に反映させることにより，文節間の係り受け解析が正しく行われた例もある．その主な要因は，品詞付けや文節切りが正しく行われた要因と同じであり，品詞付けや文節切りが正しくなると同時に文節間の係り受け解析も正しくなる場合が多かった．

4.2.2 従属係数モデルの分析

4.2.1 の単語導出モデルの分析と同様に，従属係数モデルをモデル全体に加えた場合 (表 3 の “+L”) と加えない場合 (表 3 の “−L”) とで 1 位の構文木が異なる文に注目し，解析結果の詳細な分析を行った．その考察を以下にまとめる．尚，これ以降に挙げる例文において，T_n^{+L} は従属係数モデルをモデル全体に加えた場合，T_n^{-L} は従属係数モデルを加えなかった場合に 1 位となった解析結果を表わす．

従属係数モデルが分かち書き，品詞付けに影響を与える場合

表 3 からわかるように，従属係数モデルは分かち書きや品詞付けの精度向上にはあまり影響しない．解析例を調べてみても，顕著な特徴は見られなかった．

従属係数モデルが文節切りに影響を与える場合

3.1.3 で説明したように，今回の実験で使用した従属係数モデルには異なる種類の統計情報が反映されている．これらのうち，特に名詞に関する従属係数 $D(n|N[touten, mod\_type])$ が文節切りに影響を与える例が多く見られた．

例文 7 は，従属係数モデルを全体のモデルに加えることにより文節切りが正しく行われた例である．

[例文 7] 九一年二月、転倒事故で介助者の必要な車いす生活となった。

$< T_7^{-L} >$	$< T_7^{+L} >$
九一/年/二/月/、──┐ 転倒/事故/で──┤ 介助/者/の──┤ 必要な┐ | 車いす/生活/と┤ なった/。	九一/年┐ 二/月/、┐ 転倒/事故/で──┐ 介助/者/の┐ | 必要な┐ | 車いす/生活/と┤ なった/。
	D(年 | $N[\phi$, 連体]) = 15.5

従属係数モデルをモデル全体に加えない場合，一文内に含まれる文節の数が少ない解析結果 T_7^{-L} に高い生成確率が与えられ，「九一年二月、」は1つの文節とみなされる．しかしながら，今回の実験に用いた京大コーパスにおいては，「九一年二月、」は「九一年」と「二月、」の2つの文節に分けることが正しいとされている．従属係数モデルをモデル全体に加えた場合，"年"は読点が直後にない場合には連体修飾しやすいという統計情報が従属係数 $D(\text{年}\,|N[\phi,\text{連体}])$ によって与えられ，その結果「九一年二月、」を2つの文節に分ける解析結果 T_7^{+L} に高い生成確率が与えられる[4]．

例文8は，従属係数をモデル全体に加えることにより文節切りが誤って行われた例である．

[例文 8] 木村は引退後、十四世名人に推挙された。

$<T_8^{-L}>$	$<T_8^{+L}>$
木村/は 引退/後/、 十四/世/名人/に 推挙/さ/れた/。	木村/は 引退/後/、 十四/世 名人/に 推挙/さ/れた/。
	$D(\text{世}\,|N[\phi,\text{連体}]) = 8.61$

例文7の場合とは逆に，"世"という名詞が，読点が直後にない場合は連体修飾しやすいという従属係数 $D(\text{世}\,|N[\phi,\text{連体}])$ が考慮されたため，1つの文節とするべき「十四世名人」という文字列を2つの文節に分けた解析結果 T_8^{+L} に対して高い確率が与えられた．

以上の2つの例は，複合名詞内で文節を切るか否かという問題に多いに関係がある．複合名詞は全て1つの文節とする場合には，上記の例のような違いは無視される．

従属係数モデルが係り受けに影響を与える場合

従属係数モデルで考慮される統計情報のうち，文節の係り受け解析の精度向上に最も貢献するのは格間の従属関係を反映した従属係数 $D(p_1,\cdots,p_n|P_1\cdots P_n[v])$ である．この従属係数が有効に働いた例が例文9である．

[例文 9] 徹底的な弾圧で半数が死んだ。

$<T_9^{-L}>$	$<T_9^{+L}>$
徹底的な 弾圧/で 半数/が 死んだ/。	徹底的な 弾圧/で 半数/が 死んだ/。
$D(\text{が}\,|P[\text{死ぬ}]) = 2.84$	$D(\text{で,が}\,|PP[\text{死ぬ}]) = 6.04$

[4]本実験では，同一文節内の単語間の係り受け関係は考慮していないので，$D(\text{年}\,|N[\phi,\text{連体}])$ は T_7^{-L} の生成確率の計算には使われない．

この場合，「死ぬ」という動詞にはデ格とガ格が共起しやすいという統計情報が従属係数 $D(\text{で,が}\,|PP[\text{死ぬ}])$ によって反映され，その結果「弾圧で」という文節の係り先が正しく認識されている．

一方，今回学習した従属係数モデルが文節間の係り受け解析に有効に働かない例もいくつか見られた．

[例文 10] 画面上で欲しい商品を選び、注文から決済まで行う。

$<T_{10}^{+L}>$ 画面/上/で
欲しい
商品/を
選び/、
注文/から
決済/まで
行う/。

T_{10}^{+L} は，「注文から」という文節が「決済まで」に係ることを表わしているが，正しい係り先文節は「行う。」である．しかしながら，「注文から」という文節が「行う。」に係るときには以下のような従属係数が考慮されるため，解析結果の生成確率が低く推定される．

$$D(\text{注文}\,|N[\text{から},\text{行う}]) = 0.0417 \qquad (7)$$

この従属係数は人間の直観によれば低すぎるように思われるが，これは訓練データ量の不足が原因であると考えられる．式(7)のような格要素に関する従属係数はRWCコーパスから取り出された (n,p,v) という共起事例から学習しているが，p="から"である共起事例はのべ163,247個であり，p="が"のときの1,221,120個や"を"のときの2,961,010個と比べてかなり少ない．

例文10のように，「AからBまで行う」とある場合には，2つの格要素AとBには共起関係があると考えられる．このような格要素間の従属関係も構文解析の曖昧性解消に有効であると考えられるが，今回の実験では考慮されていない．

従属係数モデルが有効に働かなかった例をもうひとつ挙げる．

[例文 11] 対局室前の廊下に、終局を待つ取材陣が息を殺して待機していた。

$<T_{11}^{+L}>$ 対局/室/前/の
廊下/に/、
終局/を
待つ
取材/陣/が
息/を
殺して
待機/して/いた/。

T_{11}^{+L} は,「取材陣が」という文節が「殺して」に係ることを表わしているが，正しい係り先文節は「待機していた。」である．実際には，"取材陣" は "殺す" と "待機する" の両方の主語となっているので,「取材陣が」という文節の係り先を「殺して」か「待機していた。」のどちらかに決めようとするのはあまり意味がない．このような並列構造を正しく認識するように，現在のモデルを改良することが今後の重要な課題のひとつとして挙げられる．

5 おわりに

本論文では，統合的確率言語モデルを利用して，複数の統計情報を同時に用いて形態素・構文解析の曖昧性を解消し，個々の統計情報が解析精度の向上にどのように貢献するかを考察した．特に，単語導出モデルに反映されている単語の出現頻度，従属係数モデルに反映されている単語の共起関係について，分かち書き，品詞付け，文節切り，係り受け解析に与える影響を実際の解析例を観察して調べた．

単語の出現頻度については，間違った解析結果を優先させる事例がいくつか見られたものの，分かち書き，品詞付け，文節切りに関して有効に働くことを確認した．しかしながら，今回の実験では未知語は存在しないことを仮定していたが，全ての単語を辞書に登録することはほとんど不可能であることを考えると，現実的な設定とは言えない．今後は，未知語処理も含めて形態素・構文解析を行ったとき，単語導出モデルがどのように働くかを調べる必要があるだろう．特に，単語導出モデルの各項 $P(w_i|l_i)$ を推定する際に，w_i が未知語の場合にその確率をどのように推定するかについては十分検討しなければならない．

単語の共起情報については，特に係り受け解析の精度向上に貢献することが確認できたが，データスパースネス問題や並列構造の取り扱い方に問題を残していることがわかった．また，今回の実験では取り扱わなかった単語の共起情報としては，4.2.2 で触れた格要素間の従属関係の他に，副詞と動詞の間の共起関係や述語間の共起関係などがある．これらの統計情報も従属係数モデルに組み込むべきであろう．

今回の実験による形態素・構文解析の精度は決して高いとは言えず，改善の余地が残されている．今後は，考察で明らかになった問題点を克服し，統合的確率言語モデルの改良を進める予定である．また，今回の考察では，統計情報に関する問題点だけでなく，文法や辞書の不備により解析がうまくいかなかった例もいくつか見つかった．統計情報は曖昧性解消に有効であるが，統計情報だけを利用するのではなく，文法や辞書の整備も同時にすすめることにより解析精度の向上を図っていきたい．

参考文献

[1] Eduardo de Paiva Alves, Haodong Wu, and Teiji Furugori. A method for estimating strength of association and its application to structural disambiguation. 自然言語処理, Vol. 5, No. 3, pp. 53–65, 1998.

[2] 江原暉将. 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, 1998.

[3] Masakazu Fujio and Matsumoto Yuji. Japanese dependency structure analysis based on lexicalized statistics. In *Proceedings of the EMNLP*, pp. 87–95, 1998.

[4] 池原悟, 宮崎正弘, 白井諭, 横尾昭男, 中岩浩己, 小倉健太郎, 大山芳史, 林良彦. 日本語語彙体系 — 全 5 巻 —. 岩波書店, 1997.

[5] Kentaro Inui, Virach Sornlertlamvanich, Hozumi Tanaka, and Takenobu Tokunaga. A new formalization of probabilistic GLR parsing. In *Proceedings of the IWPT*, 1997.

[6] 黒橋禎夫, 長尾眞. 京都大学テキストコーパス・プロジェクト. 人工知能学会全国大会論文集, pp. 58–61, 1997.

[7] 松本裕治, 黒橋禎夫, 宇津呂武仁, 妙木裕, 長尾眞. 日本語形態素解析システム JUMAN 使用説明書 version 2.0. Technical report, 京都大学工学部長尾研究室, 奈良先端科学技術大学院大学 松本研究室, 1994.

[8] 日本電子化辞書研究所. EDR 電子化辞書仕様説明書第 2 版. Technical Report TR-045, 1995.

[9] Real World Computing Partnership. RWC text database. `http://www.rwcp.or.jp/wswg.html`, 1995.

[10] Kiyoaki Shirai, Kentaro Inui, Takenobu Tokunaga, and Hozumi Tanaka. An empirical evaluation on statistical parsing of japanese sentences using lexical association statistics. In *Proceedings of the EMNLP*, pp. 80–87, 1998.

[11] 白井清昭, 乾健太郎, 徳永健伸, 田中穂積. 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, Vol. 5, No. 3, pp. 85–106, 1998.

[12] Virach Sornlertlamvanich, Kentaro Inui, Kiyoaki Shirai, Hozumi Tanaka, and Takenobu Tokunaga. Empirical evaluation of probabilistic GLR parsing. In *Proceedings of the NLPRS*, pp. 169–174, 1997.

[13] 田中穂積, 今井宏樹, 白井清昭. 文脈自由文法の制約と異なるレベルの接続制約を同時に用いた glr ベースの構文解析法. 言語処理学会第 5 回年次大会併設ワークショップ論文集, 1999.

[14] 内元清貴, 関根聡, 井佐原均. Me による日本語係り受け解析. 情報処理学会情報処理学会自然言語処理研究会, Vol. 98, No. 99, pp. 31–38, 1998.

[15] 国立国語研究所. 分類語彙表, 増補版, 1996.