JAIST Repository
https://dspace.jaist.ac.jp/

Title	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
Author(s)	白井，清昭；乾，健太郎；徳永，健伸；田中，穂積
Citation	自然言語処理，5(3)：85-106
Issue Date	1998-07-10
Type	Journal Article
Text version	publisher
URL	http://hdl.handle.net/10119/7928
Rights	Copyright (C) 1998 言語処理学会．白井清昭，乾健太郎，徳永健伸，田中穂積，自然言語処理，5(3)，1998，85-106．
Description	

JAIST
JAPAN
ADVANCED INSTITUTE OF
SCIENCE AND TECHNOLOGY

Japan Advanced Institute of Science and Technology

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

白井　清昭[†]　乾　健太郎[††]
徳永　健伸[†]　田中　穂積[†]

本論文では，構文解析の曖昧性解消を行うために，構文的な統計情報と語彙的な統計情報を統合する手法を提案する．我々が提案する統合的確率言語モデルは，構文的優先度などの構文的な統計情報を反映する構文モデルと，単語の出現頻度や単語の共起関係などの語彙的な統計情報を反映する語彙モデルの 2 つの下位モデルから成る．この統合的確率言語モデルは，構文的な統計情報と語彙的な統計情報を同時に学習する過去の多くのモデルと異なり，両者を個別に学習する点に特徴がある．構文的な統計情報と語彙的な統計情報を独立に取り扱うことにより，それぞれの統計情報を異なる言語資源から独立に学習することができるだけでなく，それぞれの統計情報が曖昧性解消においてどのような効果を果たすのかを容易に分析することができる．この統合的確率言語モデルを評価するために，日本語文の文節の係り受け解析を行った．構文モデルを用いたときの文節の正解率は 73.38%となり，ベースラインに比べて 11.70%向上した．また，構文モデルと語彙モデルを組み合わせることにより，文節の正解率はさらに 10.96%向上し 84.34%となった．この結果，本研究で提案する枠組において，語彙的な統計情報は構文的な統計情報と同程度に曖昧性解消に貢献することを確認した．

キーワード:　統計的構文解析，構文的統計情報，語彙的統計情報，統合的確率言語モデル

A Framework of Integrating Syntactic and Lexical Statistics in Statistical Parsing

Kiyoaki Shirai[†] , Kentaro Inui[††] , Takenobu Tokunaga[†]
and Hozumi Tanaka[†]

In this paper, we propose a new framework of statistical language modeling integrating syntactic statistics and lexical statistics. Our model consists of two submodels, the syntactic model and lexical model. The syntactic model reflects syntactic statistics, such as structural preferences, whereas the lexical model reflects lexical statistics, such as the occurrence of each word and word collocations. One of the characteristics of our model is that it learns both types of statistics separately, although many previous models learn them simultaneously. Learning each submodel separately enables us to use a different language source for different submodels, and to make understanding of each submodel's behavior much easier. We conducted a preliminary experiment, where our model was applied to the disambiguation of dependency structures of Japanese sentences. The syntactic model achieved 73.38% in *Bunsetu* phrase accuracy, which is 11.70 points above the baseline, and when incorporating the lexical model with the syntactic model, further 10.96 point gain was achieved, to 84.34%. Thus the contribution of lexical statistics for disambiguation is

as great as that of syntactic statistics in our framework.

KeyWords: *statistical parsing, syntactic statistics, lexical statistics, integrated probabilistic language model*

1 はじめに

コーパス，辞書，シソーラスなどの機械可読な言語データの整備が進んだことから，自然言語処理における様々な問題の解決に何らかの統計情報を利用した研究が盛んに行われている．特に構文解析の分野においては，構文的な統計情報だけでなく，単語の出現頻度や単語の共起関係といった語彙的な統計情報を利用して解析精度を向上させた研究例が数多く報告されている (Schabes 1992; Magerman 1995; Hogenout and Matsumoto 1996; Li 1996; Charniak 1997; Collins 1997)．ここで問題となるのは，このような語彙的な統計情報を構文的な統計情報とどのように組み合わせるかということである．このとき，我々は以下の2つの点が重要であると考える．

- 解析結果の候補に与えるスコアが，構文的な統計情報のみを反映したスコアと語彙的な統計情報のみを反映したスコアから構成的に計算できること

  このことによる利点を以下に挙げる．

  - 個々の統計情報を個別に学習できる

    構文的な統計情報を学習する際には，学習用言語資源として比較的作成コストの高い構文構造が付加されたコーパスが必要となる[1]．しかしながら，推定パラメタの数はそれほど多くはないので，比較的少ないデータ量で学習することができる．これに対して，語彙的な統計情報は，単語の共起に関する統計情報を学習しなければならないために大量の学習用データを必要とするが，構文構造付きコーパスに比べて作成コストの低い品詞付きコーパスを用いても学習することが十分可能である．このように，統計情報の種類によって学習に要する言語資源の質・量は大きく異なる．そこで，構文的な統計情報と語彙的な統計情報を異なる言語資源を用いて個別に学習できるように，それぞれの統計情報の独立性を保持しておくことが望ましい．

  - 曖昧性解消時における個々の統計情報の働きを容易に理解することができる

    例えば，曖昧性解消に失敗した場合には，構文的な統計情報と語彙的な統計情報を独立に取り扱うことにより，どちらの統計情報が不適切であるかを容易に判断

† 東京工業大学大学院 情報理工学研究科 計算工学専攻, Department of Computer Science, Graduate School of Information Science and Engineering, Tokyo Institute of Technology

†† 九州工業大学 情報工学部 知識情報工学科, Department of Artificial Intelligence, Kyushu Institute of Technology

1 Inside-Outside アルゴリズム (Lari and Young 1990) に代表されるような EM アルゴリズムを用いて，構文構造が付加されていないコーパスから構文的な統計情報を学習する研究も行われている．しかしながら，このような教師なしの学習は一般に精度が悪く，現時点では構文構造が付加されたコーパスを利用した方が品質の良い統計情報を学習できると考えられる．

することができる．

- 個々の統計情報を反映したスコアが確率的意味を持っていること
  構文的な統計情報を反映したスコアと語彙的な統計情報を反映したスコアを組み合わせて全体のスコアとする場合，両者のスコアの和を計算すればいいのか，積を計算すればいいのか，またどちらか片方に重みを置かなければならないのかなど，その最適な組み合わせ方は自明ではない．このとき，個々のスコアが確率的意味を持つように学習することにより，確率の積としてそれらを自然に組み合わせることができる．

ところが，語彙的な統計情報を利用して構文解析の精度を向上させる過去の研究の多くは以上の条件を満たしていない．例えば田辺らは，確率文脈自由文法 (Probabilistic Context Free Grammar, 以下 PCFG) における書き換え規則の非終端記号に，その非終端記号が支配する句の主辞となる単語を付加すること (以下，これを PCFG の語彙化と呼ぶ) によって語彙的従属関係を PCFG の確率モデルに反映させる方法を提案している (田辺, 冨浦, 日高 1995)．一方，英語を対象に PCFG を語彙化した研究としては Hogenout ら (Hogenout and Matsumoto 1996), Charniak (Charniak 1997), Collins (Collins 1997) によるものがある．しかしながら，PCFG の語彙化によって構文的な統計情報と語彙的な統計情報を組み合わせる方法は，非終端記号に単語を付加することによって規則数が組み合わせ的に増大し，推定するパラメタ数も非常に多くなるといった問題点がある．また，構文的な統計情報と語彙的な統計情報を同時に学習するモデルとなっているが，先ほど述べたように両者は独立に学習できることが望ましい．PCFG をベースとしない SPATTER パーザ (Magerman 1995) や SLTAG (Schabes 1992; Resnik 1992) にも同様の問題が存在する．これらの研究は語彙的な統計情報を利用して解析精度の向上を図ってはいるが，構文的な統計情報と独立に学習する枠組にはなっていない．

構文的な統計情報と語彙的な統計情報を独立に学習する枠組としては Li によるものが挙げられる (Li 1996; 李 1996)．Li は，解析結果の候補 I に対して，構文的な統計情報を反映させた確率モデル $P_{syn}(I)$ と単語の共起関係を反映させた確率モデル $P_{lex}(I)$ を別々に学習する方法を提案している．そして，語彙的な制約は構文的な制約に優先するといった心理言語学原理に基づき，まず $P_{lex}(I)$ を I のスコアとして用い，一位とそれ以外の候補のスコアの差が十分に大きくなかった場合に限り $P_{syn}(I)$ をスコアとして用いている．すなわち，構文的な統計情報と語彙的な統計情報をそれぞれ独立に学習してはいるが，これらを同時に利用して曖昧性解消を行っているわけではない．また，この 2 つのスコアの持つ確率的意味が不明確であり[2]，その最適な組み合わせ方は自明ではない．

本研究では，構文的な統計情報と語彙的な統計情報を組み合わせる一方法として，統合的確率言語モデルを提案する (Inui, Shirai, Tanaka, and Tokunaga 1997a; 乾, 白井, 徳永, 田中 1997; 白井, 乾, 徳永, 田中 1996)．この統合的確率言語モデルの特徴は，単語の出現頻度，およ

2 $P_{syn}(I)$, $P_{lex}(I)$ は確率と呼ばれてはいるが，どのような事象に対する確率なのかは不明である．

び単語の共起関係といった2つの語彙的な統計情報を局所化し，構文的な統計情報と独立に取り扱う点にある．また，構文的な統計情報を構文構造の生成確率として，語彙的な統計情報を単語列の生成確率としてそれぞれ学習し，これらの積を解析結果の候補に対するスコアとすることにより，曖昧性解消に両者を同時に利用することができる．この統合的確率言語モデルの詳細については2節で述べる．3節ではこの統合的確率言語モデルの学習，およびそれを用いた日本語文の文節の係り受け解析実験について述べる．最後に4節で結論と今後の課題について述べる．

2　統合的確率言語モデル

まず，本論文で一貫して用いる記号について説明する．

- 入力文字列 $A = a_1, \cdots, a_m$
- A を生成する単語列 $W = w_1, \cdots, w_n$
- W を生成する品詞列 $L = l_1, \cdots, l_n$
- L を生成する構文構造 R

本研究では，形態素解析と構文解析を同時に取り扱うことを仮定する．すなわち，入力文字列 A が与えられたときに，その正しい単語列 W，正しい品詞列 L，正しい構文構造 R を求めることを目的とする．例えば,「彼女がパイを食べた」という入力文に対する解析結果の候補の例を図1に示す．

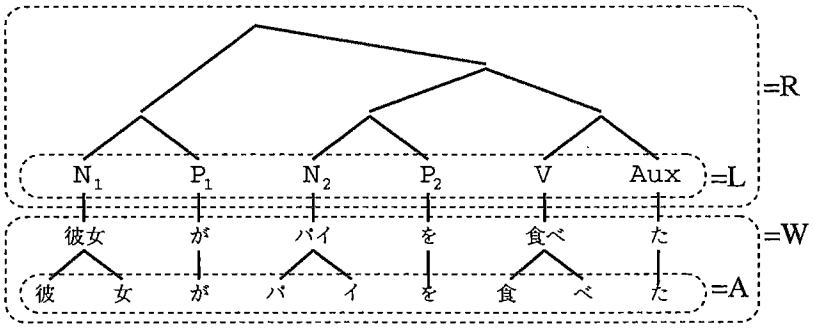

図1　例文 “彼女がパイを食べた” とその解析結果

各解析結果の候補に対してその生成確率 $P(R, L, W, A)$ を計算し，これが最大の解析結果を選択することによって曖昧性解消を行う．さらに，確率モデル $P(R, L, W, A)$ を以下のように分解する．

$$P(R, L, W, A) = P(R) \cdot P(L|R) \cdot P(W|L, R) \cdot P(A|W, L, R) \tag{1}$$

ここで，構文構造 R は最終的に品詞列 L を生成するものと仮定すると，$P(L|R)=1$ となる (図 1 参照)．また，単語列 W が決まれば入力文字列 A は一意に決まるので，$P(A|W)=1$ となる．したがって，式 (1) は以下のように簡略化できる．

$$P(R,L,W,A)=P(R)\cdot P(W|R) \tag{2}$$

本研究では，式 (2) に示した通り，解析結果の生成確率を以下の 2 つの確率モデルの積として計算する．

(1)　構文モデル $P(R)$

構文構造 R の生成確率である．この確率モデルには構文的な統計情報を反映させる．

(2)　語彙モデル $P(W|R)$

構文構造 R が与えられたときに，それから単語列 W を生成する確率である．この語彙モデルには語彙的な統計情報を反映させる．

2.1 構文モデル $P(R)$

構文モデルとしては，構文的な統計情報を反映し，かつ構文構造 R の生成確率を高い精度で推定するものであれば，どのような確率モデルを利用してもよい．構文モデルに利用できる確率モデルとしては，PCFG や確率一般化 LR 法 (Probabilistic Generalized LR Method，以下 PGLR) などが挙げられる．

我々は，PGLR を構文モデルの有力な候補として考えている．PGLR とは，構文解析手法のひとつである一般化 LR 法を拡張したものである．PGLR は，LR 表に記述された各状態遷移の遷移確率を推定し，その遷移確率の積によって 1 つの状態遷移列，すなわちそれに対応する構文構造の生成確率を与えるモデルである[3]．この PGLR は PCFG に比べて，次のような特長を持つ (Inui et al. 1997b)．

- 文脈依存性を取り扱うことができる．
- 隣接する品詞間の共起関係を取り扱うことができる．
- 距離に関する優先度を取り扱うことができる．

ここで，隣接する品詞間の共起関係とは，品詞 bi-gram のような品詞列の出現に関する統計情報であり，形態素解析の曖昧性解消に有効であると考えられる．また，距離に関する優先度とは，単語はなるべく近い単語に係りやすいといった，係り受け関係にある単語間の距離に関する統計情報である．

3 一般化 LR 法に確率を組み込む試みには様々なものがあるが (Wright 1990; Ng and Tomita 1991; Briscoe, Carroll 1993)，本研究における PGLR とは Inui らによるモデル (Inui, Sornlertlamvanich, Tanaka, and Tokunaga 1997b; Sornlertlamvanich, Inui, Shirai, Tanaka, Tokunaga, and Takezawa 1997) を指す．

2.2 語彙モデル $P(W|R)$

語彙モデルは，品詞列 L を末端とする構文構造 R が与えられたときに，それから単語列 W を生成する確率である．この語彙モデルは，式 (3) のような各単語 w_i の生成確率の積として計算することができる．

$$P(W|R) = \prod_{w_i} P(w_i|R, w_1, \cdots, w_{i-1}) \tag{3}$$

例えば，図 1 の例において，単語を文の後ろから順番に生成していくと仮定すると，語彙モデル $P(W|R)$ は以下のような単語の生成確率の積として計算できる．

$$\begin{aligned} P(W|R) &= P(\text{彼女}, \text{が}, \text{パイ}, \text{を}, \text{食べ}, \text{た}\,|R) && (4)\\ &= P(\text{た}\,|R)\cdot && (5)\\ & \quad P(\text{食べ}\,|R, \text{た})\cdot && (6)\\ & \quad P(\text{を}\,|R, \text{食べ}, \text{た})\cdot && (7)\\ & \quad P(\text{パイ}\,|R, \text{を}, \text{食べ}, \text{た})\cdot && (8)\\ & \quad P(\text{が}\,|R, \text{パイ}, \text{を}, \text{食べ}, \text{た})\cdot && (9)\\ & \quad P(\text{彼女}\,|R, \text{が}, \text{パイ}, \text{を}, \text{食べ}, \text{た}) && (10) \end{aligned}$$

2.2.1 単語生成文脈

式 (3) の各項 (図 1 の例では式 (5)〜(10)) のパラメタ空間は非常に大きく，これを直接学習することは一般に不可能である．ところが，各単語 w_i の生成に強く影響するのは各項の確率の前件 $R, w_1, \cdots, w_{i-1}$ 全てではなく，その一部のみであると考えられる．例えば，図 1 の例文において，“パイ” は動詞 “食べ” のヲ格の格要素となっている．このとき，“パイ” という単語を生成する際には，式 (8) の前件 “R, を, 食べ, た” (図 2 の斜線部) のうち，品詞 N と単語 “を”,“食べ” (図 2 の丸で囲まれた部分) によって十分近似できると期待できる (式 (11)).

$$P(\text{パイ}\,|R, \text{を}, \text{食べ}, \text{た}) \simeq P(\text{パイ}\,|N[s(\text{食べ}, \text{を})]) \tag{11}$$

式 (11) において，$N[s(\text{食べ}, \text{を})]$ は，“食べ” という動詞のヲ格の格要素となっている名詞を表わしている．すなわち，$P(\text{パイ}\,|N[s(\text{食べ}, \text{を})])$ は，“食べ” という動詞のヲ格の格要素となっている名詞から “パイ” という単語が生成される確率を表わしている．したがって，式 (11) には，“パイ” という単語そのものがどれくらい出現しやすいかといった単語の出現頻度と，“パイ” と “食べ” がどの程度共起しやすいかといった単語の共起関係が反映されている．

ここで，単語生成文脈 c_i を以下のように定義する．

> 単語 w_i の単語生成文脈 c_i とは，w_i の生成確率の前件 $R, w_1, \cdots, w_{i-1}$ から w_i の生成に強く影響する部分のみを取り出したものである．

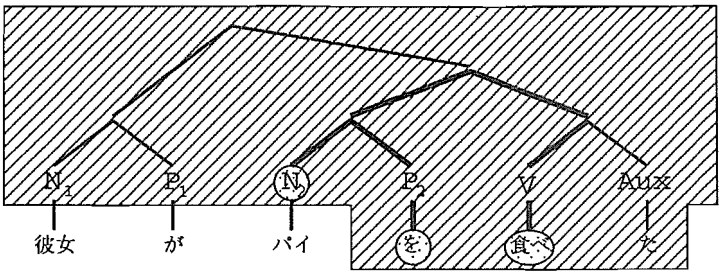

図 2　“パイ” を生成するときの単語生成文脈

先ほどの例においては，単語 “パイ” の単語生成文脈は “s(食べ, を)” である．そして，各単語 w_i の生成確率の前件 “$R, w_1, \cdots, w_{i-1}$” を，その単語の品詞 l_i と単語生成文脈 c_i に縮退することにより，語彙モデル $P(W|R)$ を以下のように近似する．

$$\begin{aligned} P(W|R) &= \prod_{w_i} P(w_i|R, w_1, \cdots, w_{i-1}) \\ &\simeq \prod_{w_i} P(w_i|l_i[c_i]) \end{aligned} \tag{12}$$

2.2.2　単語生成文脈決定規則

単語生成文脈を導入する際に問題となるのは，どのような単語に対してどのような単語生成文脈を選べばよいのかということである．我々は，これを人手で作成した規則によって記述する．以下，単語 w_i の単語生成文脈 c_i を決定する規則を単語生成文脈決定規則と呼ぶ．

単語生成文脈決定規則の例を以下に挙げる．

- 単語の共起関係を全く考慮しない場合

  単語 w_i について，周囲の単語との従属関係を考慮しない場合には，その単語の生成確率はその単語の品詞 l_i のみに依存するとみなす．例えば，図 1 の例において，助動詞 “た” と動詞 “食べ” を生成する際に他の単語との語彙的従属関係を考えない場合には，それぞれの生成確率 (5),(6) は以下のように近似すればよい．

$$P(\text{た}|R) \simeq P(\text{た}|AUX) \tag{13}$$

$$P(\text{食べ}|R, \text{た}) \simeq P(\text{食べ}|V) \tag{14}$$

  これに対応した単語生成文脈決定規則を以下に示す．この規則は単語生成文脈を決定する際のデフォルト規則でもある．

  【単語生成文脈決定規則 #1】

  単語 w_i を生成する際に他の単語との従属関係を考慮しない場合には，単語 w_i の単語生成文脈 c_i を空とする．

- 格要素となる名詞が助詞を介して動詞に係る際の従属関係を考慮する場合

前述のように，格要素となる名詞が助詞を介して動詞に係る際には，動詞・助詞の組と名詞との間には語彙的従属関係が存在する．このような語彙的従属関係を確率モデルに反映させるために単語生成文脈決定規則 #2 を定義する．

【単語生成文脈決定規則 #2】

> 単語 w_i の品詞 l_i が N(名詞) であり，かつ助詞 p を介して動詞 v に係っているとき，単語 w_i の単語生成文脈 c_i を $s(v,p)$ とする．このとき，w_i の生成確率 $P(w_i|N[s(v,p)])$ は動詞 v の格 p の格要素となる名詞 N から単語 w_i が生成される確率を表わす．

例えば，図 1 の例において，名詞 “パイ” は動詞 “食べ” のヲ格の格要素であり，名詞 “彼女” は動詞 “食べ” のガ格の格要素となっている．したがって，これらの単語を生成する際にはこの規則が適用され，それぞれの生成確率 (8),(10) は以下のように近似される．

$$P(\text{パイ}\,|R,\text{を},\text{食べ},\text{た}) \simeq P(\text{パイ}\,|N[s(\text{食べ},\text{を})]) \tag{15}$$

$$P(\text{彼女}\,|R,\text{が},\text{パイ},\text{を},\text{食べ},\text{た}) \simeq P(\text{彼女}\,|N[s(\text{食べ},\text{が})]) \tag{16}$$

- 助詞とその係り先用言の従属関係，格間の従属関係を考慮する場合

図 1 の例文においては，2 つの助詞 “が” と “を” が動詞 “食べ” に係っている．このとき，これらの生成確率 (7),(9) を以下のように近似しても，助詞とその係り先用言との間の語彙的従属関係，および同じ用言に係る助詞同士の従属関係 (以下，これを格間の従属関係と呼ぶ) を語彙モデルに反映させることができる．

$$P(\text{を}\,|R,\text{食べ},\text{た}) \simeq P(\text{を}\,|P[m(\text{食べ},\{\phi_1,\phi_2\})]) \tag{17}$$

$$P(\text{が}\,|R,\text{パイ},\text{を},\text{食べ},\text{た}) \simeq P(\text{が}\,|P[m(\text{食べ},\{\phi_1,\text{を}\})]) \tag{18}$$

式 (17) は，助詞 P が 2 つの助詞の係り先となっている動詞 “食べ” に係っているときに，品詞 P から単語 “を” が生成される確率を表わしている．一方式 (18) は，助詞 P が 2 つの助詞の係り先となりかつそのうちの 1 つは “を” である動詞 “食べ” に係っているときに，品詞 P から単語 “が” が生成される確率を表わしている．

助詞とその係り先用言の従属関係，および格間の従属関係を語彙モデルに導入するために，単語生成文脈決定規則 #3 を以下のように定義する．

【単語生成文脈決定規則 #3】

> 単語 w_i の品詞 l_i が P(助詞) でありかつ用言 h に係っているとき，単語 w_i の単語生成文脈 c_i を $m(h,\{\phi_1,\cdots,\phi_j,p_{j+1},\cdots,p_n\})$ とする．このとき，w_i の生成確率 $P(w_i|P[m(h,\{\phi_1,\cdots,\phi_j,p_{j+1},\cdots,p_n\})])$ は，用言 h が n 個の助詞の係り先となりかつ用言に近い $p_{j+1},\cdots,p_n$ の助詞が既に生成されているときに，ϕ_j として w_i が生成される確率を表わす．

単語生成文脈決定規則 #3 において，同じ用言に係る助詞は用言に近いものから順番に生成されると仮定している．すなわち，助詞が出現する順序も考慮されている．

- 助詞の係り先が用言か体言かを考慮する場合

  助詞の係り先が用言である場合と体言である場合とでは，助詞の生成確率 $P(w_i|P)$ の分布は著しく異なると考えられる．例えば，係り先が用言の場合には“が”，“を”などの助詞は出現しやすいが，助詞“の”は出現しにくい．これに対して，係り先が体言の場合，すなわちその助詞を含む文節が連体修飾節となっている場合には，助詞“の”が出現する場合が圧倒的に多いと予想される．したがって，助詞の生成確率 $P(w_i|P)$ を学習する際に，その助詞の係り先が用言もしくは体言であるかを区別しないで学習するのは望ましいことではない．これに対応するには，以下のような単語生成文脈決定規則 #4 を定義すればよい．

【単語生成文脈決定規則 #4】

> 単語 w_i の品詞 l_i が P(助詞) であり，かつその助詞の係り先が体言であるとき，単語 w_i の単語生成文脈 c_i を nd とする．nd はその助詞の係り先が体言であることを表わすシンボルである．このとき，w_i の生成確率 $P(w_i|P[nd])$ は，体言を係り先とする助詞から単語 w_i が生成される確率を表わす．

助詞の単語生成文脈を決定する際には，助詞の係り先が用言である場合には単語生成文脈決定規則 #3 が，助詞の係り先が体言である場合には単語生成文脈決定規則 #4 が適用される．

ここに挙げた単語生成文脈決定規則 #1～#4 が単語生成文脈を決定するための全ての規則というわけではない．本節では，特に用言の格関係に注目して語彙モデルに反映させるべき語彙的従属関係 (単語の共起関係) の例を挙げたが，他の種類の語彙的従属関係を語彙モデルに反映させるように単語生成文脈決定規則を拡張・洗練することもできる．すなわち，語彙モデルにおいてどのような語彙的従属関係を考慮するかは，単語生成文脈決定規則の追加・変更によって柔軟に調整することが可能である．

単語生成文脈として何を選択するかを自動的に学習することも考えられる[4]が，我々は言語学的知見に基づくヒューリスティクス規則によって単語生成文脈を選択する方向で研究をすすめている．なぜなら，語彙モデルにどのような種類の語彙的従属関係を反映させるかを単語生成文脈決定規則によって明確に記述することにより，モデルに反映された統計情報が曖昧性解消に有効であるかどうかなど，モデルの特性の分析を容易に行うことができるからである．

4 例えば，Magerman は確率の前件としてどのような素性を選択すればよいのかを決定木を用いて自動学習している (Magerman 1995).

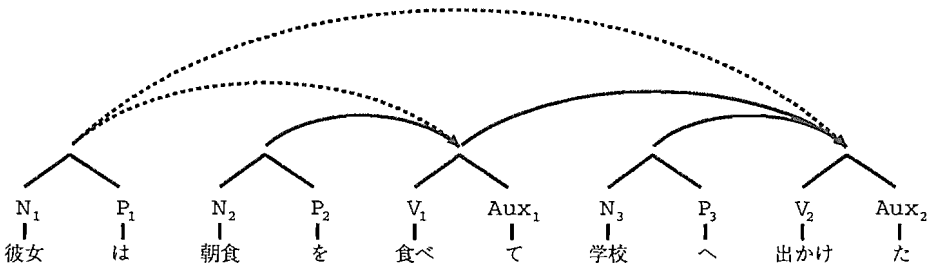

図 3　並列構造を持つ例文

2.2.3　従属係数

これまでは単語を生成する際に考える単語生成文脈は常に一つであると仮定していた．しかしながら，一般には，一つの単語を生成する際に複数の単語生成文脈を考慮しなければならない場合もある．例えば，図 3 の例文において，2 つの文節 “食べ-て”[5] と “出かけ-た” は並列の関係にある．したがって，この例文中の名詞 “彼女” は動詞 “食べ” のハ格の格要素であり[6]，同時に動詞 “出かけ” のハ格の格要素でもある．したがって，単語生成文脈決定規則 #2 に従えば，“彼女” という単語を生成する際の単語生成文脈としては s(食べ, は) と s(出かけ, は) の 2 つがある．このとき，“彼女” の生成確率は次のように推定することが望ましい．

$$P(\text{彼女}\,|N[s(\text{食べ}, \text{は}), s(\text{出かけ}, \text{は})]) \tag{19}$$

同様に，この例文中の助詞 “は” は動詞 “食べ” と “出かけ” の両方に係っているとみなすことができる．したがって，単語生成文脈決定規則 #3 に従えば，“は” という単語を生成する際の単語生成文脈として m(食べ, $\{\phi_1$, を $\}$) と m(出かけ, $\{\phi_1$, へ $\}$) の 2 つがあると考えられ，“は” の生成確率も次のように推定することが望ましい．

$$P(\text{は}\,|P[m(\text{食べ}, \{\phi_1, \text{を}\}), m(\text{出かけ}, \{\phi_1, \text{へ}\})]) \tag{20}$$

ところが，式 (19) や (20) のように複数の単語生成文脈を前件に持つ確率モデルは，推定するパラメタの数が爆発的に増大する可能性がある．そこで本研究では，複数の単語生成文脈を以下のように取り扱う．まず，説明を簡略化するために，単語 w_i が 2 つの単語生成文脈 c_1 と c_2 を持つとする．このとき，単語 w_i の生成確率 $P(w_i|l_i[c_1, c_2])$ を以下のように近似する．

$$P(w_i|l_i[c_1, c_2]) = \frac{P(l_i[c_1, c_2]|w_i) \cdot P(w_i)}{P(l_i[c_1, c_2])} \tag{21}$$

$$= \frac{P(l_i[c_1]|w_i) \cdot P(l_i[c_2]|l_i[c_1], w_i) \cdot P(w_i)}{P(l_i[c_1]) \cdot P(l_i[c_2]|l_i[c_1])} \tag{22}$$

5 “-” は単語の区切りを表わす．

6 本研究では，名詞が助詞を介して用言に係る場合は常に，その名詞を用言の表層格の格要素として取り扱う．

$$
\begin{aligned}
&\simeq \frac{P(l_i[c_1]|w_i)\cdot P(l_i[c_2]|l_i,w_i)\cdot P(w_i)}{P(l_i[c_1])\cdot P(l_i[c_2]|l_i)} && (23)\\
&= \frac{P(l_i[c_1]|w_i)}{P(l_i[c_1])}\cdot\frac{P(l_i[c_2]|l_i,w_i)}{P(l_i[c_2]|l_i)}\cdot P(w_i) && (24)\\
&= \frac{P(w_i|l_i[c_1])}{P(w_i)}\cdot\frac{P(w_i,|l_i,l_i[c_2])}{P(w_i|l_i)}\cdot P(w_i) && (25)\\
&= P(w_i|l_i)\cdot\frac{P(w_i|l_i[c_1])}{P(w_i|l_i)}\cdot\frac{P(w_i|l_i[c_2])}{P(w_i|l_i)} && (26)
\end{aligned}
$$

式 (22) から式 (23) の変形において，2 つの単語生成文脈 c_1 と c_2 は互いに独立であると仮定している．

$$
\begin{aligned}
P(l_i[c_2]|l_i[c_1]) &\simeq P(l_i[c_2]|l_i) && (27)\\
P(l_i[c_2]|l_i[c_1],w_i) &\simeq P(l_i[c_2]|l_i,w_i) && (28)
\end{aligned}
$$

ここで，従属係数 $D(w_i|l_i[c_i])$ を式 (29) のように定義する．

$$
D(w_i|l_i[c_i]) = \frac{P(w_i|l_i[c_i])}{P(w_i|l_i)} \tag{29}
$$

この従属係数を用いれば，式 (26) から式 (30) が導かれる．

$$
P(w_i|l_i[c_1,c_2]) \simeq P(w_i|l_i)\cdot D(w_i|l_i[c_1])\cdot D(w_i|l_i[c_2]) \tag{30}
$$

以上では単語 w_i が 2 つの単語生成文脈を持つ場合を考えていたが，単語 w_i が n 個の単語生成文脈 $c_1,\cdots,c_n$ を持つ場合にも同様の近似が可能であり，最終的に以下の式が得られる．

$$
P(w_i|l_i[c_1,\cdots,c_n]) \simeq P(w_i|l_i)\cdot\prod_{c_i} D(w_i|l_i[c_i]) \tag{31}
$$

式 (29) で定義した従属係数 $D(w_i|l_i[c_i])$ は単語 w_i と単語生成文脈 c_i の相関関係を評価する統計量である．例えば，w_i と c_i に相関関係がない場合，すなわち w_i と c_i が互いに独立である場合には，式 (29) の分子 $P(w_i|l_i[c_i])$ は分母 $P(w_i|l_i)$ にほぼ等しくなり，従属係数 $D(w_i|l_i[c_i])$ は 1 に近い値を取る．これに対し，w_i と c_i に正の相関関係がある場合には，単語生成文脈 c_i を前件に加えた確率 $P(w_i|l_i[c_i])$ は単語生成文脈 c_i を無視した確率 $P(w_i|l_i)$ よりも大きくなるので，その従属係数は 1 より大きい値を取る．同様に，w_i と c_i に負の相関関係がある場合には従属係数は 1 より小さい値を取る．

複数の単語生成文脈 $c_1,\cdots,c_n$ の下での単語 w_i の生成確率は，単語生成文脈を無視した単語の生成確率 $P(w_i|l_i)$ と，w_i と c_i の相関関係を他の単語生成文脈とは独立に評価した従属係数 $D(w_i|l_i[c_i])$ の積によって計算できることを式 (31) は示している．従属係数 $D(w_i|l_i[c_i])$ を他の単語生成文脈と独立に推定・学習することにより，確率モデルのパラメタ空間を推定可能な大きさに抑制することができる．例えば，図 3 の例において，“彼女” の生成確率 (19) と “は”

の生成確率 (20) はそれぞれ以下のように推定される．

$$
\begin{aligned}
& P(\text{彼女}\,|N[s(\text{食べ},\text{は}),s(\text{出かけ},\text{は})]) \\
\simeq\ & P(\text{彼女}\,|N)\cdot D(\text{彼女}\,|N[s(\text{食べ},\text{は}))\cdot D(\text{彼女}\,|N[s(\text{出かけ},\text{は})]) \qquad (32)
\end{aligned}
$$

$$
\begin{aligned}
& P(\text{は}\,|P[m(\text{食べ},\{\phi_1,\text{を}\}),m(\text{出かけ},\{\phi_1,\text{へ}\})]) \\
\simeq\ & P(\text{は}\,|P)\cdot D(\text{は}\,|P[m(\text{食べ},\{\phi_1,\text{を}\})])\cdot D(\text{は}\,|P[m(\text{出かけ},\{\phi_1,\text{へ}\})]) \qquad (33)
\end{aligned}
$$

従属係数を導入する利点として，単語生成文脈を複数取り扱うことができるという点の他に，式 (36) に示すように，語彙モデル $P(W|R)$ を単語の出現頻度のみを反映した $P_{cf}(W|L)$ と単語の共起関係のみを反映した $D(W|R)$ との積に分解できるという点が挙げられる．

$$
\begin{aligned}
P(W|R) &\simeq \prod_i P(w_i|l_i[C_{w_i}]) && (34)\\
&\simeq \prod_{w_i} P(w_i|l_i)\cdot \prod_{c_{ij}\in C_{w_i}} D(w_i|l_i[c_{ij}]) && (35)\\
&= P_{cf}(W|L)\cdot D(W|R) && (36)\\
P_{cf}(W|L) &= \prod_{w_i} P(w_i|l_i) && (37)\\
D(W|R) &= \prod_{w_i}\prod_{c_{ij}\in C_{w_i}} D(w_i|l_i[c_{ij}]) && (38)
\end{aligned}
$$

上式において，C_{w_i} は単語 w_i の単語生成文脈の集合を表わしている．

式 (37) の統計量 $P_{cf}(W|L)$ は，単語生成文脈を無視したときに品詞 l_i から単語 w_i が生成される確率の積であり，単語の出現頻度に関する優先度が反映される．これに対し，式 (38) の統計量 $D(W|R)$ は各単語 w_i とその単語生成文脈 c_{ij} の従属係数の積を表わしており，w_i と c_{ij} の相関関係に関する優先度 (すなわち単語の共起関係) が反映される．このように，語彙モデルを単語の出現頻度，および単語の共起関係のみを反映させた 2 つの統計量の積として分解することにより，1 節で述べたように，曖昧性解消時におけるそれぞれの統計情報の働きを容易に理解することができる．

3　評価実験

本節では，前節で提案した統合的確率言語モデルの評価実験について述べる．統合的確率言語モデルは本来形態素解析，構文解析を同時に行うことを前提としているが，そのような大規模な実験を行う前の予備実験として，まずは文節列を入力とする文節間の係り受け解析のみを行った．

3.1 構文モデルの学習

本節の実験では，入力として単語列，品詞列，文節区切りが与えられたときに，それぞれの文節の係り先となる文節を決定する．このような文節の係り受け解析を CFG(文脈自由文法) を用いて行った．

まず，CFG 規則の終端記号として，文節の統語的特性を反映した文節ラベルを用いる．この文節ラベルの定義を (39) に示す．

$$\text{文節ラベル} \stackrel{def}{=} (\text{受け属性}, \text{係り属性}, \text{読点の有無}, \text{用言種別}) \tag{39}$$

ここで，“受け属性”，“係り属性” はそれぞれ文節の受け属性と係り属性であり，“連用”，“連体”，“格関係” の組によって表わされる．例えば，“パイ-を” や “彼女-の” など,「名詞 助詞」といった品詞並びによって構成される文節は，他の文節から連体修飾を受ける可能性があるので受け属性は “(連体)” となり，他の文節を連体修飾したり用言を修飾してその格要素および表層格を表わす可能性があるので係り属性は “(連用, 格関係)” となる[7]．また “読点の有無” は，その文節の末尾が読点であれば “1”，そうでなければ “0” といった値を取る．これは，読点を末尾に持つ文節は直後の文節には係りにくく，読点を末尾に持たない文節よりも遠くに係る傾向があるので，この違いを構文モデルに反映させるためである．一方 “用言種別” は，“格関係” を受け属性に含む文節タイプを細分化するための属性であり，文節の主辞が自動詞，他動詞，形容詞，名詞述語のときにはそれぞれ “自動詞”，“他動詞”，“形容詞”，“名詞述語” といった値を取る．また，“格関係” を受け属性に持たない文節のときにはその値は常に “ϕ” である．2.2 節で例示した単語生成文脈決定規則は，単語の共起関係の中でも特に用言の格関係に注目している．用言を主辞とする文節の文節ラベルを細分化したのはこのためである．この文節ラベルは，文節を構成する単語列の品詞情報をもとに一意に決定されるものとする．また，これらの文節ラベルの整合性[8]をチェックする規則を作成し，その集合を文節の係り受け解析に用いる CFG とした．この CFG の概要を表 1 に示す．

表 1　CFG の概要

規則数	961
非終端記号数	51
終端記号数 (文節ラベル数)	42

本実験では，構文モデル $P(R)$ として PGLR を利用した．また，この構文モデルの学習には京大コーパス (黒橋, 長尾 1997) を使用した．京大コーパスの各例文には，単語区切り，単語の品詞，文節区切りと文節の係り受け解析の結果 (構文構造) が付加されている．京大コーパス

7 ここでの “格関係” とは，用言を受け側とした格関係のみを指す．
8 例えば，“連体” を係り属性に含む文節は “連体” を受け属性に含まない文節には係らない．

の 9,944 例文に対して，コーパスの各例文とそれに付加された構文構造を作り出すような LR 表における状態遷移列を求め，また状態遷移が行われた回数を数え上げる．このようにして得られた状態遷移回数を状態遷移確率に変換することにより，PGLR のパラメタ推定を行った．

3.2 語彙モデルの学習

本実験では，式 (36) に示した語彙モデル $P(W|R) = P_{cf}(W|L) \cdot D(W|R)$ のうち，$P_{cf}(W|L)$ の計算を省略できる．なぜなら，単語列及び品詞列はすでに入力として与えられているため，全ての解析結果の候補について品詞から単語への生成確率の積 $P_{cf}(W|L)$ は等しいからである．したがって，語彙モデルとして学習するのは従属係数の積 $D(W|R)$ のみでよい．今回の実験では，単語生成文脈決定規則 #2〜#4 によって定められる従属係数 (40),(47),(50) を $D(W|R)$ の要素とし，これらの学習を行った．

まず，格要素の従属係数 (40) の学習について説明する．

$$D(n|N[s(v,p)]) = \frac{P(n|N[s(v,p)])}{P(n|N)} \tag{40}$$

RWC コーパス (Real World Computing Partnership 1995) と EDR 共起辞書 (日本電子化辞書研究所 1995) から，名詞 n が助詞 p を介して動詞 v に係る事例 (n,p,v) をそれぞれのべ 6,888,849 組，975,510 組収集した．式 (40) の分子および分母の確率モデルはこれらの訓練事例から最尤推定した．

さらに，分子の確率モデル $P(n|N[s(v,p)])$ を推定する際に以下のような近似を行った．

- 名詞 n の意味クラスによる抽象化

  名詞 n の意味クラスの集合を $C_n = \{c_{n_1}, \cdots, c_{n_m}\}$ として，$P(n|N[s(v,p)])$ を以下のように推定した．

  $$P(n|N[s(v,p)]) \simeq \sum_j P(n|c_{n_j}) P(c_{n_j}|N[s(v,p)]) \tag{41}$$

  今回の実験では，名詞意味クラス c_n として，日本語語彙体系 (池原, 宮崎, 横尾 1993; 池原, 宮崎, 白井, 横尾, 中岩, 小倉, 大山, 林 1997) の名詞シソーラスのルートから深さ 3 に位置する 151 個の意味クラスの集合を用いた．これらの意味クラスは互いに排他的である．

  $D(n|N[s(v,p)])$ を推定する場合，名詞 n が日本語語彙体系に登録されておらず，その名詞意味クラスが不明な場合には，その従属係数は学習不可能であるとして $D(n|N[s(v,p)]) \simeq 1$ とした．これは，n と $s(v,p)$ との間の従属関係を無視することに相当する．

- バックオフ方式によるスムージング

  確率モデル $P(c_n|N[s(v,p)])$ を推定する際，この確率の分母となる事例 $(*,p,v)$ ($*$ は任意の名詞意味クラスを表わす) の出現回数がある閾値 λ よりも小さい場合には，v を動

詞意味クラス c_v を用いて抽象化した確率モデル $P(c_n|N[s(c_v,p)])$ によって近似した．

$$P(c_n|N[s(v,p)]) \simeq P(c_n|N[s(c_v,p)]) \tag{42}$$

また，事例 $(*,p,c_v)$ の出現回数が λ を越えない場合には，動詞意味クラス c_v の抽象度を段階的に上げていき，必ずλ 個以上の訓練事例から確率モデルを推定するようにした．本実験においては，動詞意味クラス c_v として分類語彙表 (国立国語研究所 1996) の 5 桁および 2 桁の分類コードを動詞意味クラスとして利用した．動詞を分類語彙表の 2 桁の分類コードに抽象化しても学習事例数が λ を越えなかったとき，もしくは $(*,p,v)$ の事例数が λ 以下でありかつ動詞 v が分類語彙表に登録されていなかった場合には，十分信頼度の高い確率モデルが学習できなかったとして，従属係数 $D(n|N[s(v,p)]) \simeq 1$ とした．なお，今回は$\lambda = 100$ として実験を行った．

次に，用言に係る助詞に関する従属係数 (43) の学習について説明する．

$$D(p_i|P[m(h,\{\phi_1,\cdots,\phi_i,p_{i+1},\cdots,p_n\})]) = \frac{P(p_i|P[m(h,\{\phi_1,\cdots,\phi_i,p_{i+1},\cdots,p_n\})])}{P(p_i|P)} \tag{43}$$

n 個の助詞 $p_1,\cdots,p_n$ が同じ用言 h に係っている場合には，それぞれの p_i に対応する従属係数 (43) の積を計算すれば良い．この従属係数の積は式 (46) のように変形できる．

$$\begin{aligned}
& \prod_i D(p_i|P[m(h,\{\phi_1,\cdots,\phi_i,p_{i+1},\cdots,p_n\})]) && (44)\\
= & \prod_i \frac{P(p_i|P[m(h,\{\phi_1,\cdots,\phi_i,p_{i+1},\cdots,p_n\})])}{P(p_i|P)} && (45)\\
= & \frac{P(p_1,\cdots,p_n|P_1,\cdots,P_n[m(h,\{\phi_1,\cdots,\phi_n\})])}{\prod_i P(p_i|P)} && (46)\\
\stackrel{def}{=} & D(p_1,\cdots,p_n|P_1,\cdots,P_n[m(h,\{\phi_1,\cdots,\phi_n\})]) && (47)
\end{aligned}$$

したがって，学習しなければならないのは，ある用言 h が $P_1,\cdots,P_n$ の n 個の助詞の係り先となっているときに単語 $p_1,\cdots,p_n$ を同時に生成する確率モデル $P(p_1,\cdots,p_n|P_1,\cdots,P_n[m(h,\{\phi_1,\cdots,\phi_n\})])$ と，品詞 P(助詞) から単語 p_i が生成される確率 $P(p_i|P)$ である．以降，簡単のため，前者の確率モデルを以下のように記述する．

$$\begin{aligned}
P(\vec{p}\mid h,n) \stackrel{def}{=} & P(p_1,\cdots,p_n|P_1,\cdots,P_n[m(h,\{\phi_1,\cdots,\phi_n\})]) && (48)\\
& \text{但し，}\vec{p}=(p_1,\cdots,p_n)
\end{aligned}$$

確率モデル $P(\vec{p}|h,n)$ を学習するために，n 個の助詞 $\vec{p}$ が同じ用言 h に係るという事例 $(\vec{p},h)$ を EDR コーパスから収集した．今回の実験では，用言 h として動詞，形容詞，名詞述語の 3 つを考えた．用言 h が動詞，形容詞，名詞述語であるときの，また h に係る助詞の数 n が 1,

表 2　EDR コーパスから収集した事例 $(\vec{p}, h)$ ののべ数

h	$n=1$	$n=2$	$n=3$	$n \geq 4$
動詞	231,730	123,915	30,375	3,961
形容詞	19,266	7,686	1,292	154
名詞述語	28,636	9,327	1,238	98

2，3，4 以上であるときの事例 $(\vec{p}, h)$ ののべ数を表 2 にまとめる．

n が 4 以上のときには学習に十分な事例を収集することができなかった．そこで，n が 4 以上のときには，従属係数を 1，すなわち助詞とその係り先用言との語彙的従属関係や格間の従属関係を無視することにした．

$$n \geq 4 \text{ のとき} \qquad D(p_1, \cdots, p_n | P_1, \cdots, P_n[m(h, \{\phi_1, \cdots, \phi_n\})]) \simeq 1 \tag{49}$$

$n = 1$ のときの式 (48) の分子の確率モデル $P(\vec{p}|h, n)$ は表 2 に示した事例から最尤推定した．また，$n = 2, 3$ のときの確率モデル $P(\vec{p}|h, n)$ は最大エントロピー法を用いて推定した[9]．

最後に，体言に係る助詞に関する従属係数 (50) の学習について説明する．

$$D(p|P[nd]) = \frac{P(p|P[nd])}{P(p|P)} \tag{50}$$

この従属係数を学習するために，EDR コーパスから体言に係る助詞 p をのべ 273,062 個収集した．式 (50) の分子はこの訓練データから最尤推定した．また，式 (50) の分母 $P(p|P)$ は，ここで収集した体言に係る助詞の事例と，表 2 に示した用言に係る助詞の事例から，同様に最尤推定した．尚，式 (46) の分母の各項 $P(p_i|P)$ も式 (50) の分母の確率モデルと同じものを使用した．

3.3　実験結果

3.1 節にて学習した構文モデル $P(R)$，および 3.2 節にて学習した語彙モデル $P(W|R)$ を用いて，文節の係り受け解析を行った．まず，テスト文として，京大コーパスの中から文節数 7〜9 の文をランダムに 500 文選び，これをテスト文とした．構文モデル $P(R)$ を学習する際に用いた訓練用例文にはこれらのテスト文は含まれていない．文節数 7〜9 という比較的文長の短い例文をテスト文として選んだのは，本実験で用いた PGLR パーザがまだ開発の途中であり，長い文長の例文の解析に非常に多くの時間を要するためである．

テスト文の係り受け解析結果の評価尺度として，文節の正解率を以下のように定義する．

$$\text{文節の正解率} = \frac{\text{係り先の正しい文節の数}}{\text{テスト文に含まれる文節の数}} \tag{51}$$

この文節の正解率は生成確率が一位である解析結果の候補について計算する．また，文の最後

9 この詳細については (白井, 乾, 徳永, 田中 1997) を参照．

表 3　文節の正解率

	後置詞節	全ての文節
BL	62.92%	61.68%
Syn	69.63%	73.38%
F	71.36%	74.69%
M	78.19%	78.55%
P	84.06%	82.22%
all	86.30%	84.34%

に位置する 2 つの文節は評価の対象から除外する．これは，文の一番最後にある文節は係り先がなく，また文の最後から 2 番目にある文節は常に文の一番最後の文節に係るからである．

2.2 節に述べたように，語彙モデルにおいてはいくつかの種類の統計情報を取り扱う．ここでは，構文的な統計情報，および語彙モデルにおいて考慮された語彙的な統計情報のそれぞれの曖昧性解消における効果を調べるために，以下に述べる 6 種類のモデルを用意し，それらを比較した．結果を表 3 に示す．

BL:　ベースライン

全ての文節の係り先を，(1) 全ての文節は係り得る文節の中でできるだけ近いものに係る，(2) 一文中における文節の係り受け関係は互いに交差しない，として決定するモデルである．

Syn:　従属係数を無視したモデル

$D(W|R) = 1$ としたモデルである．すなわち，構文モデル $P(R)$ で学習した統計情報のみを用いて曖昧性解消を行う．

F:　格要素となる名詞に関する従属係数のみを用いたモデル

$D(W|R)$ として，式 (40) によって与えられる従属係数のみを考慮したモデルである．

M:　用言に係る助詞に関する従属係数のみを用いたモデル

$D(W|R)$ として，式 (47) によって与えられる従属係数のみを考慮したモデルである．

P:　体言に係る助詞に関する従属係数のみを用いたモデル

$D(W|R)$ として，式 (50) によって与えられる従属係数のみを考慮したモデルである．

all:　全ての従属係数を用いたモデル

上記全ての従属係数を考慮したモデルである．

表 3 から，語彙モデルにおいて考慮した語彙的な統計情報のうち，体言に係る助詞に関する従属係数 (モデル P) が正解率の向上に一番大きく貢献することがわかる．すなわち，助詞が用言に係っているか否かの違いがその生成確率に大きく影響し，その違いを考慮することによって曖昧性解消の精度を大きく向上させることができた．また，表 3 における “後置詞節” とは，

“彼女-が”，“パイ-を”など，用言の格要素および表層格を表わす可能性のある文節を指す[10]．テスト文全体における 2,975 個の文節のうち，1,788 個がこの後置詞節に相当する．この後置詞節のみで評価した場合，全ての文節で評価した場合に比べて，語彙的な統計情報を考慮したモデル (F,M,P,all) と構文的な統計情報のみを考慮したモデル (Syn) との文節の正解率の差が大きくなっている．これは，今回の実験で用いた語彙モデルにおいては，語彙的な統計情報の中でも用言の格関係に注目しているため，語彙モデルが“後置詞節”の係り先の曖昧性解消に特に有効に働いているためと考えられる．

構文モデルと全ての語彙的従属関係を考慮した語彙モデルを組み合わせて曖昧性解消に用いた場合 (all)，構文モデルのみを用いた場合 (Syn) と比べて文節の正解率が 10.96%向上し，また構文モデルのみを曖昧性解消に用いたときのベースラインとの文節の正解率の差が 11.70%であることから，文節の係り受け解析の精度向上において，語彙モデルは構文モデルと同程度の貢献をしていると考えられる．本研究で提案した統合的確率言語モデルにおいては，語彙的な統計情報を局所化し構文的な統計情報とは独立に学習しているが，このようなアプローチにおいても，語彙的な統計情報は曖昧性解消の精度向上に十分大きく貢献すると期待できる．

最後に，本研究で提案する統合的確率言語モデルを用いた解析結果と KNP パーザ (黒橋, 長尾 1994) による解析結果との比較を行った．KNP パーザは形態素解析システム JUMAN(松本, 黒橋, 宇津呂, 妙木, 長尾 1994) の形態素解析結果を入力とし，文節の区切りを認定してから文節の係り受け解析を行う．そこで，3.3 節の実験で用いた 500 個のテスト文のうち，JUMAN の形態素解析結果による形態素区切りおよび KNP パーザによる文節区切りの結果がコーパスと一致した 388 文を対象に，両者の係り受け解析結果の比較を行った．結果を表 4 に示す．

表 4　KNP パーザとの比較

	後置詞節	全ての文節
本手法	86.57%	84.53%
KNP パーザ	86.79%	85.71%

本手法は KNP パーザよりも文節の正解率で 1%程度劣っている．今回の実験では，統合的確率言語モデルに組み込む語彙的従属関係として，格要素と動詞との従属関係，助詞と係り先用言との従属関係，格間の従属関係などを考慮した．しかしながら，これ以外にも，曖昧性解消に有効であると考えられる語彙的従属関係が数多く存在する．特に，今回の実験では連体修飾に関しては語彙的従属関係を何も考慮していないので，そのことによる解析誤りが多かった．例えば,「彼女の紫色の帽子が風に飛ばされた」という文においては，文節“彼女-の”が (a)“紫色-の”に係る，(b)“帽子-が”に係るという 2 つの解釈がある．ところが，連体修飾する“彼女”については語彙的従属関係を考慮していないので，より近い文節に係る解釈 (a) に高い確率が

10 この後置詞節には，“太郎-の”など，実際には体言を修飾する文節も含まれる．

与えられてしまう．これを回避するためには，以下のような従属係数を学習し語彙モデルに加えればよい．

$$D(n_1|N[n_2]) = \frac{P(n_1|N[n_2])}{P(n_1|N)} \tag{52}$$

式 (52) の分子 $P(n_1|N[n_2])$ は，ある名詞 N が n_2 を連体修飾しているとき，その名詞として単語 n_1 が生成される確率を表わしている．このような従属係数を考慮することにより，“彼女” は “紫色” よりも “帽子” を連体修飾することが多い，すなわち D(彼女 $|N$[紫色]) $<<$ D(彼女 $|N$[帽子]) であると考えられるので，正しい解釈 (b) に高い確率を与えると期待できる．このように，統合的確率言語モデルに新たな種類の語彙的従属関係を反映させるときには，それに対応した従属係数を新たに語彙モデルに加えるという形で容易に対処できる．これは，語彙的従属関係を局所化して構文的優先度などの他の統計情報と独立に学習するように，また異なる種類の語彙的従属関係は異なる従属係数として独立に学習するようにモデルを設計したことに依る．

一方，後置詞節のみで評価した場合には，本手法と KNP パーザの文節の正解率はほぼ等しい．とはいえ，後置詞節の係り先の特定に失敗する場合も少なくない．我々は現在その原因を調査中であり，その一部については既に報告している (Shirai, Inui, Hozumi, and Tokunaga 1997)．今後，曖昧性解消に有効な統計情報を新たに組み込んだり，また解析誤りの原因を調査しそれらに対処することにより，係り受け解析の精度向上を図っていきたい．

4　おわりに

本研究では，形態素解析・構文解析を同時に行う際に，構文的な統計情報と語彙的な統計情報を組み合わせて曖昧性を解消するひとつの手法を提案した．我々の手法の特徴は，構文的優先度，隣接する品詞間の共起関係，距離に関する優先度といった構文的な統計情報を構文モデル $P(R)$ として，単語の出現頻度および単語の共起関係を語彙モデル $P(W|R)$ として，それぞれ独立に学習する点にある．このことは，個々の統計情報を異なる言語資源から学習できるだけでなく，曖昧性解消時における個々の統計情報の働きを容易に分析することができる．実際に，京大コーパスを用いて構文モデルを，RWC コーパスや EDR コーパスを用いて語彙モデルを学習した．また，これらの確率モデルを用いた日本語文の文節の係り受け解析実験の結果，構文的な統計情報と語彙的な統計情報のそれぞれが曖昧性解消に大きく貢献することを確認した．

最後に今後の課題について述べる．まず，統合的確率言語モデルが本来想定している形態素解析と構文解析を同時に行い，その有効性を実験的に確認することが挙げられる．また，今回の実験では文長の比較的短い文を対象にしたが，文長の長い文の係り受け解析を行うことにより，統合的確率言語モデルの特性をさらに調査する必要がある．文長の長い文においては，二重格を取りにくいなどの格間の従属関係がさらに有効に働くのではないかと予想される．最後に，統合的確率言語モデルと他の統計的構文解析に関する研究とを実験的に比較することが挙

げられる．特に今回の実験は日本語を対象にしたが，構文的な統計情報と語彙的な統計情報を独立に学習するアプローチが英語などの他の言語においても本当に有効であるのかどうかは今後調査していく必要があると思われる．

謝辞

本研究にあたり，日本語語彙体系を提供して下さいましたNTTコミュニケーション科学研究所知識処理研究部翻訳処理研究グループに感謝いたします．

参考文献

Briscoe, T. and Carroll, J. (1993). "Generalized Probabilistic LR Parsing of Natural Language (Corpora) with Unification-Based Grammars." *Computational Linguistics*, **19** (1), 25–59.

Charniak, E. (1997). "Statistical Parsing with a Context-free Grammar and Word Statistics." In *Proceedings of the National Conference on Artificial Intelligence.*

Collins, M. (1997). "Three Generative, Lexicalised Models for Statistical Parsing." In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 16–23.

Hogenout, W. R. and Matsumoto, Y. (1996). "Experiments with Using Semantical Categories in Parsing Systems." 言語処理学会第2回年次大会発表論文集, pp. 381–384.

池原悟, 宮崎正弘, 横尾昭男 (1993). "日英機械翻訳のための意味解析用の知識とその分解能." 情報処理学会論文誌, **34** (8), 1692–1704.

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 (1997). 日本語語彙体系 — 全5巻 —. 岩波書店.

乾健太郎, 白井清昭, 徳永健伸, 田中穂積 (1997). "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, **96** (114), 35–42.

Inui, K., Shirai, K., Tanaka, H., and Tokunaga, T. (1997a). "Integrated Probabilistic Language Modeling for Statistical Parsing." Tech. rep. TR97-0005, Dept. of Computer Science, Tokyo Institute of Technology.

Inui, K., Sornlertlamvanich, V., Tanaka, H., and Tokunaga, T. (1997b). "A New Formalization of Probabilistic GLR Parsing." In *Proceedings of the International Workshop on Parsing Technologies.*

黒橋禎夫, 長尾眞 (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), 35–57.

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.” 人工知能学会全国大会論文集, pp. 58–61.

Lari, K. and Young, S. (1990). “The Estimation of Stochastic Context-free Grammars Using the Inside-Outside Algorithm.” *Computer speech and languages*, **4**, 35–56.

李航 (1996). “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ェア, **13** (6), 489–501.

Li, H. (1996). “A Probabilistic Disambiguation Method Based on Psycholinguistic Principles.” In *Proceedings of the Workshop on Very Large Corpora*, pp. 141–154.

Magerman, D. M. (1995). “Statistical Decision-Tree Models for Parsing.” In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 276–283.

松本裕治, 黒橋禎夫, 宇津呂武仁, 妙木裕, 長尾眞 (1994). “日本語形態素解析システム JUMAN 使用説明書 version 2.0.” テクニカル・レポート, 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.

Ng, S. K. and Tomita, M. (1991). “Probabilistic LR Parsing for General Context-Free Grammars.” In *Proceedings of the International Workshop on Parsing Technologies*, pp. 154–163.

日本電子化辞書研究所 (1995). “EDR 電子化辞書仕様説明書第 2 版.” テクニカル・レポート TR–045.

Real World Computing Partnership (1995). “RWC text database.” `http://www.rwcp.or.jp/wswg.html`.

Resnik, P. (1992). “Probabilistic Tree-Adjoining Grammar as a Framework for Statistical Natural Language Processing.” In *Proceedings of the 14th International Conference on Computational Linguistics*, Vol. 2, pp. 418–424. COLING ’92.

Schabes, Y. (1992). “Stochastic Lexicalized Tree-Adjoining Grammars.” In *Proceedings of the 14th International Conference on Computational Linguistics*, Vol. 2, pp. 425–432. COLING ’92.

Shirai, K., Inui, K., Hozumi, T., and Tokunaga, T. (1997). “An Empirical Study on Statistical Disambiguation of Japanese Dependency Structures Using a Lexically Sensitive Language Model.” In *Proceedings of the Natural Language Processing Pacific Rim Symposium*, pp. 215–220.

白井清昭, 乾健太郎, 徳永健伸, 田中穂積 (1996). “統計的日本語文解析における種々の統計量の扱いについて.” 自然言語処理シンポジウム「大規模資源と自然言語処理」. http://www.etl.go.jp/etl/nl/nlsympo/96/.

白井清昭, 乾健太郎, 徳永健伸, 田中穂積 (1997). “最大エントロピー法による格の従属関係の学

習." 言語処理学会第 3 回年次大会発表論文集, pp. 337–340.

Sornlertlamvanich, V., Inui, K., Shirai, K., Tanaka, H., Tokunaga, T., and Takezawa, T. (1997). "Empirical Evaluation of Probabilistic GLR Parsing." In *Proceedings of the Natural Language Processing Pacific Rim Symposium.*

田辺利文, 冨浦洋一, 日高達 (1995). "語の共起関係の文脈自由文法への取り込み法." EDR 電子化辞書利用シンポジウム論文集, pp. 25–31.

Wright, J. H. (1990). "LR Parsing of Probabilistic Grammars with Input Uncertainty for Speech Recognition." *Computer Speech and Language*, 4, 297–323.

国立国語研究所 (1996). 分類語彙表 増補版.

略歴

白井 清昭: 1993 年東京工業大学工学部情報工学科卒業. 1995 年同大学院理工学研究科修士課程修了. 1998 年同大学院情報理工学研究科博士課程修了. 同年同大学院情報理工学研究科計算工学専攻助手, 現在に至る. 博士 (工学). 統計的自然言語解析に関する研究に従事. 情報処理学会会員.

乾 健太郎: 1990 年東京工業大学工学部情報工学科卒業. 1992 年同大学大学院理工学研究科修士課程修了. 1995 年同大学大学院理工学研究科博士課程修了. 同年同大学院情報理工学研究科計算工学専攻助手. 1998 年九州工業大学情報工学部知識情報工学科助教授, 現在に至る. 博士 (工学). 自然言語処理に関する研究に従事. 人工知能学会, 体系機能言語学会, 各会員.

徳永 健伸: 1983 年東京工業大学工学部情報工学科卒業. 1985 年同大学院理工学研究科修士課程修了. 同年 (株) 三菱総合研究所入社. 1986 年東京工業大学大学院博士課程入学. 現在, 同大学大学院情報理工学研究科計算工学専攻助教授. 博士 (工学). 自然言語処理, 計算言語学に関する研究に従事. 情報処理学会, 認知科学会, 人工知能学会, 計量国語学会, Association for Computational Linguistics, 各会員.

田中 穂積: 1964 年東京工業大学工学部情報工学科卒業. 1966 年同大学院理工学研究科修士課程修了. 同年電気試験所 (現電子技術総合研究所) 入所. 1980 年東京工業大学助教授. 1983 年東京工業大学教授. 現在, 同大学大学院情報理工学研究科計算工学専攻教授. 博士 (工学). 人工知能, 自然言語処理に関する研究に従事. 情報処理学会, 電子情報通信学会, 認知科学会, 人工知能学会, 計量国語学会, Association for Computational Linguistics, 各会員.

(1997 年 12 月 1 日 受付)
(1998 年 2 月 6 日 再受付)
(1998 年 4 月 10 日 採録)