5 言語コーパスをより有効に使うために

宇津呂 武仁 utsuro@ics.tut.ac.jp
豊橋技術科学大学工学部情報工学系

近年の構文解析研究においては，本特集「例文を使って文の解析をしよう」で解説してきたように，人手で文法規則を開発するというアプローチの困難さを解消するために，コーパスを利用した構文解析の方法の研究が盛んとなっている．また，その中でも特に，パラメタ推定法などの理論的基盤の確立された統計モデルに基づくアプローチの研究が主流となっている．しかし，構文構造付コーパスを人手で作成するには莫大なコストがかかることから，現在利用可能な構文構造付コーパスはわずかに数種類にとどまっている．また，現在利用可能な構文構造付コーパスは量的にも限られているのが現状である．そのため，これまでの統計的構文解析モデルの研究においては，現存する構文構造付コーパスを最大限利用して最適な解析性能を達成するという点に研究の重点が置かれてきた．それらの研究において最も重要になるのは，統計的な観点から自然言語の構文構造をどのようにモデル化するかという点であり，より具体的には，最適な解析性能を達成するために，統計モデルの種類として何を用い，モデル内に取り込む言語情報の種類として，何をどのように選べばよいかという点である．

このような現状をふまえて，本稿では，統計的構文解析モデルをどのように調整し，どのような言語情報をモデルに取り込むと，現存する構文構造付コーパスからどのような解析性能が得られるのかについて，いくつかの研究事例を通して説明する．さらに，解析性能を最適化することを目的として，モデルに取り込む情報を自動的に選択するなどの方法によりモデルを自動学習する試みについて，その代表的なものを紹介する．最後に，現状のまとめを簡単に行い，今後必要と思われる研究の方向を展望する．

どのような情報をモデルに取り込むべきか?

本章では，統計的構文解析モデルに取り込まれる情報とその解析性能との相関関係について，英語文・日本語文が対象の場合についてそれぞれ説明する．

英語の場合——

英語文を対象とした統計的構文解析モデルの研究においては，本特集「例文を使って文の解析をしよう」で紹介された，確率文脈自由文法に基づくモデル・依存構造に基づくモデルのいずれも研究されているが，共通の訓練・評価データを用いた実験結果では，確率文脈自由文法を拡張したモデルの一種[3]が現在最も高性能であるとされている．以下では，確率文脈自由文法にさまざまな情報を取り込んだモデルを考え，モデルに取り込まれる情報と解析性能との相関について考察する．

まず，例として，**図-1**に，“Corporate profits rose.”という英文の句構造を示す．この句構造の各構成素のうち，葉節点以外は，“句の文法カテゴリ：主辞（句の中心的意味を担う語彙）”という形式で，また葉節点は，各語彙によって記述されている．たとえば，この文の主語である“Corporate profits”という名詞句に対応する‘**np**：profits’という構成素は，

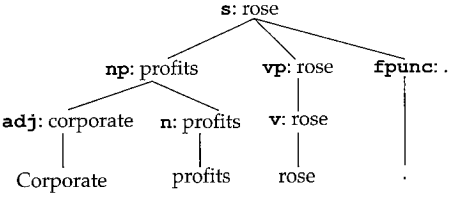

図-1 句構造の例（全体）

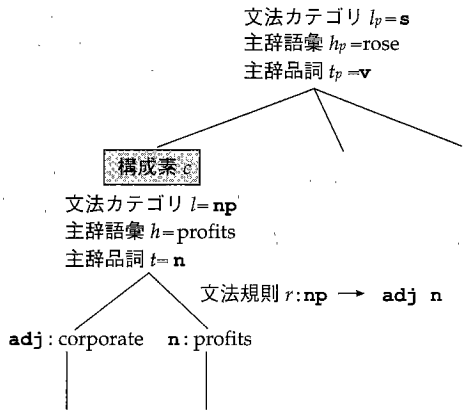

図-2　句構造の例（部分）

モデル		精度（%）
(A)	確率文脈自由文法（PCFG）	73.75
(B)	(A)＋構成素の主辞単語＋親節点の文法カテゴリ	83.75
(C)	(B)＋構成素・親節点の主辞単語の単語クラス	84.45
(D)	(B)＋親節点の主辞単語	86.60
(E)	(D)＋構成素の主辞単語の品詞	88.05
(F)	(E)＋最大エントロピーInspiredモデル＋3重マルコフ文法	89.70

表-1　統計的構文解析モデルの解析精度
（確率文脈自由文法とその拡張，英語の場合）

句の文法カテゴリが名詞句を表す‘**np**’であり，句の中心的語彙が‘profits’である．

ここで，この構成素をcとし，この構成素を中心として図-1の句構造の部分構造（**図-2**）を考え，統計的句構造モデルに取り込まれる情報の記法を説明する．図-2に示すように，構成素cの文法カテゴリをl（ここでは‘**np**’），主辞となる語彙をh（ここでは‘profits’），主辞となる語彙の品詞をt（ここでは名詞を表す‘**n**’）とする．また，構成素cの親節点‘**s**:rose’に関しても，その文法カテゴリをl_p（ここでは‘**s**’），主辞となる語彙をh_p（ここでは‘rose’），主辞となる語彙の品詞をt_p（ここでは‘**v**’）とする．さらに，構成素cを展開する際に用いられる文法規則は，

np→adj n

であるがこれをrとする．以下では，これらの表現を用いて，モデルに取り込まれる情報と各モデルの性能の相関をみていく．

主だったモデル[2), 3)]とそれぞれの解析精度☆1の一覧を**表-1**に示す．他のすべてのモデルの基本となる最も簡略化されたモデルが，確率文脈自由文法（Probabilistic Context Free Grammar; PCFG）である．これは，与えられた文に対する句構造をπとすると，以下の式によって句構造πの確率値を計算するというモデルである．

$$p(\pi) = \prod_{c \in \pi} p(r \mid l) \qquad (1)$$

具体的には，句構造π中の各構成素cについて，その構成素cの文法カテゴリlを展開する際に，左辺をlとする文法規則rの相対的な重みに相当する条件付確率$p(r \mid l)$の積によって句構造πの確率値を計算する．このモデルは，主辞の語彙などの情報を用いず，文法カテゴリだけを用いたモデルで，表-1のモデル（A）に示すように，他の語彙を利用するモデルと比べても解析精度は10%ほど低い．なお，表-1のモデル（A）では，文脈自由文法の規則として，訓練データのPenn Treebankから自動的に抽出したものを用いている．

次に，上記の確率文脈自由文法に，構成素の主辞単語の情報を追加したモデル（表-1のモデル（B））を考える（この基本的な考え方は本特集「例文を使って文の解析をしよう」参照）．このモデルは，式（1）の各条件付確率$p(r \mid l)$の条件部分に，構成素cの主辞となる語彙h，およびcの親節点の文法カテゴリl_pを追加した以下のモデルで表現される．

$$p(\pi) = \prod_{c \in \pi} p(r \mid h, l, l_p) \qquad (2)$$

これは，各文法規則rの重みを考える際に，これらの追加された情報を考慮して，より細分化して各規則の重みを考えることに相当する．つまり，その意図は，構成素cの文法カテゴリlが同じであっても，その句の中心的な語彙hが何であるかなどによって，適用される規則の重みを細分化して考えようということである☆2．このモデルの解析精度は，確率文脈自由文法（表-1のモデル（A））と比較して約10%向上する．

続いて，式（2）のモデルに，構成素cの主辞となる語彙hを推定する条件付確率$p(h \mid l, h_p, l_p)$を追加したモデルとして，次式を考える．

$$p(\pi) = \prod_{c \in \pi} p(h \mid l, h_p, l_p)\, p(r \mid h, l, l_p) \qquad (3)$$

このモデルの最大の特徴は，条件付確率$p(h \mid l, h_p, l_p)$が2つの単語hおよびh_pの間の共起の強さを表現している点である．これは，図-2の例では，名詞‘profits’と動詞‘rose’の間の共起の強さをこの条件付確率によって表現することに相当し，これによって，2単語の間の共起の強さの度合いをモデルの中に取り込むことを実現している．上式の2つの条件付確率の推定法としてはいくつかの方法が考えられるが，ここでは，単語クラスタリング法により得られた単語

☆1　いずれも，Penn Treebank（本特集「例文を使って文の解析をしよう」参照）において，共通の訓練・評価データを用いた場合の比較．評価データは単語数40以下の文約2,000文．解析精度は，出力した構文木の各構成素が，i）正解の構文木中の構成素を再現する率，と，ii）正解の構文木中の構成素に矛盾しない率，の平均によって測定される[4)]．

☆2　ただし，実際に各条件付確率$p(r \mid h, l, l_p)$を求める際には，データの過疎性の問題を避けるために，以下の平滑化（smoothing）を行い，実際の推定値$\hat{p}(r \mid h, l, l_p)$および$\hat{p}(r \mid l)$の2種類の確率値の重み付き線形和を用いる．

$$p(r \mid h, l, l_p) = \lambda_1 \hat{p}(r \mid h, l, l_p) + \lambda_2 \hat{p}(r \mid l)$$

その際の重みλ_1およびλ_2は，削除補間法（deleted interporation）と呼ばれる手法の一手法により最適化される．

クラスを用いる場合（表-1のモデル（C））と単語そのものを用いる場合（表-1のモデル（D））の2通りの結果を示す．表-1のモデル（C）および（D）にあるように，この情報の追加により約1～3%程度精度が向上する．また，クラスタリングによる単語クラスを用いる場合よりも，単語そのものを用いた場合の方が2%以上精度がよいことが分かる．

さらに，式（3）のモデルにおいて，構成素cの文法カテゴリlから主辞となる語彙hを推定する過程の中間段階として，構成素cの主辞となる語彙の品詞tを推定する過程を追加し，条件付確率$p(t|l, h_p, l_p)$を追加したモデルとして，次式を考える．

$$p(\pi) = \prod_{c \in \pi} (p(t|l, h_p, l_p)\, p(h|t, l, h_p, l_p)\, p(r|h, t, l, h_p, l_p))$$

このモデルの解析精度は表-1のモデル（E）に示したとおりであり，モデル（D）と比べても約1.5%精度が向上している．このモデルがこのようなよい性能を示すことは，直観的には自明ではなく，実際にこのモデルは，巧妙にモデルを調整し実験的評価を繰り返す過程で発見されたといってよい．重要な点は，現在利用可能な量のコーパスの範囲ではデータの過疎性の問題が起きやすいことから，いきなり詳細な情報（この場合は主辞となる語彙）を推定するのではなく中間的な段階の情報（この場合は主辞となる語彙の品詞）を推定する過程を追加した方が，適当な平滑化の効果が得られ解析性能も向上するということである．

さらに，現時点で最高の性能を示すといわれているモデル[3]は，表-1のモデル（E）に，最大エントロピモデル（下記の「最適なモデルの自動学習の試み」の節参照）と類似の細工を施し，また訓練データのPenn Treebankから自動的に抽出した文法規則を用いるのではなく，3重マルコフ文法[4]と呼ばれる確率モデルを用いたモデルで，表-1のモデル（F）に示すように，90%近い解析精度を示す．このモデルの場合も，これらの巧妙なモデルの調整により適当な平滑化の効果が得られ，さらなる解析性能の向上が実現されたといってよい．

日本語の場合 ——

日本語文を対象とした統計的構文解析モデルの研究においても，英語文の場合と同様，確率文脈自由文法に基づくモデル・依存構造に基づくモデルのいずれも研究されてはいるが，実際の評価実験の範囲・規模の点からいえば，依存構造に基づくモデルの方がよく研究されているといえる．ただし，英語文の場合と違い，共通の訓練・評価データを用いて異なったモデルの性能を客観的に比較することは定着していない．以下では，依存構造に基づく統計的構文解析モデルにおいて，モデルに取り込まれる情報・その取り込まれ方と解析性能との相関について考察する．

本特集「例文を使って文の解析をしよう」で説明されたように，一般に，日本語の統計的依存構造解析のモデルでは，文中の個々の文節b_i（以下，係り元文節）が文節b_j（以下，係り先文節）に係る確率$p(b_i \rightarrow b_j)$の積によって1文の依存構造Dの確率を定義する．

$$\prod_{b_i \rightarrow b_j \in D} p(b_i \rightarrow b_j) \qquad (4)$$

この際，各文節b_iおよびb_jやそれらに関連した情報としてどのような情報を考慮して確率値$p(b_i \rightarrow b_j)$を推定するかが問題となるが，通常は，**表-2**に示すような情報を確率変数として統計モデルが定式化される[5),6),10)]．ただし，これらの情報がすべて同等に有用というわけではなく，各情報がモデルの解析性能に寄与する度合いは大きく異なる．一例として，EDRコーパス（本特集「例文を使って文の解析をしよう」参照）を訓練・評価コーパスとして，決定木学習により確率変数（属性と呼ばれる）の自動選択を行った研究事例[6)]において，各情報を削除した場合の解析精度の増減を表-2の右欄に示す☆3．これから分かるように，モデルから削除した場合に多少なりとも解析精度が減少する情報が，解析性能に寄与している情報であり，これらの情報としては，係り元文節のタイプ，二文節間距離，係り先文節の主辞の品詞，二文節間の助詞「は」の有無，係り元・先文節の句読点，などが挙げられる．最大エントロピモデルにより確率変数の重み付けを行っている他の研究事例[10)]に

☆3　解析精度は，各文節の係り先が正しく出力される割合によって測定されており，その最高値は約84%程度である．

情報（確率変数）	とり得る値	削除時の解析精度の増減（%）
係り元文節の主辞の語彙情報 係り先文節の主辞の語彙情報	頻出100/200単語， シソーラス上位クラス	+1～3
係り元文節の主辞の品詞	32種類	+0.1
係り先文節の主辞の品詞	32種類	−2.1
係り元文節のタイプ	114種類	−9.3
係り先文節のタイプ	114種類	−0.5
係り元文節の句読点	なし・読点	−1.2
係り先文節の句読点	なし・読点・句点	−1.6
係り元文節の括弧	9種類	0
係り先文節の括弧	9種類	0
二文節間距離	3段階	−5.2
二文節間の助詞「は」の有無	有・無	−1.8
二文節間の読点の有無	有・無	0

表-2　統計的依存構造解析において用いられる情報（日本語の場合）

おいても同様の結果が得られており，日本語の統計的依存構造解析のモデルにおいて，各情報が解析性能に寄与する度合いに一定の特徴があるといえる．一方，係り元・先文節の主辞の語彙情報については，削除した方が解析性能が上がるという結果が得られている．これについては，頻出語だけでなく全語彙をモデルに取り込むことを試みた他の研究事例[5]では若干の性能向上が確認されているので，現時点では，主辞の語彙情報が解析性能に寄与するか否かについて統一的な帰結を出すのは難しい．ただし，日本語の統計的依存構造解析のモデルにおいては，各文節の主辞の語彙情報以外にも多様な情報が取り込まれているので，相対的に各文節の主辞の語彙情報の重要性が低くなっているといえる．また，一般には，分野の狭いコーパスほど構文解析における語彙情報の効果が大きいとされているのに対して，現在日本語で利用可能な構文構造付コーパスは，いずれも比較的広範囲の分野にわたって文を収集したものであるという点も考慮すべきであろう．

最適なモデルの自動学習の試み

前章では，現存する構文構造付コーパスを最大限利用して最適な解析性能を達成することを目的として，統計的構文解析モデルに取り込まれる情報とその解析性能の相関関係について説明した．一方，本章では，モデルに取り込む情報を自動的に選択するなどの方法により最適な解析性能のモデルを自動学習する試みについて，その代表的なものを紹介する．

決定木学習 ——

決定木学習は，属性とクラスの組で表現された事例に対して，属性の情報を用いてクラスを決定する規則もしくは確率分布を学習する手法の1つで，機械学習の手法の1つであるが，コーパスを利用した自然言語処理においても広く利用されている．決定木学習の目的は，調べる必要のある属性をなるべく少なくし，しかもなるべく決定的に目的となるクラスを決定する規則もしくは確率分布を学習することである．しかし，これを満たす最適解を求める問題はNP完全であり，通常は，決定木全体でクラスに関する条件付エントロピをなるべく小さくするように属性が選ばれ，決定木が構成される．なお，クラスの集合をC，決定木の葉節点の集合をLとすると，決定木全体でのクラスに関する条件付エントロピ$H(C|L)$は次式で与えられる．

$$H(C|L) = \sum_{l \in L} p(l) H(C|l)$$
$$= -\sum_{c \in C} \sum_{l \in L} p(l) p(c|l) \log p(c|l)$$

決定木学習を統計的構文解析の問題に適用した研究事例としてはいくつかのものが存在するが，そのいずれにおいても，構文解析における個々の部分問題をクラス決定問題として定式化して，その部分問題の確率分布の学習において決定木学習を適用している．具体例としては，英語においては，句構造モデルに基づく統計的構文解析において，文法規則を適用する際の確率分布，句構造を組み上げる際の確率分布，句の文法カテゴリを決定する際の確率分布などを，決定木学習を用いて推定しているものがある[1), 8)]．一方，日本語においては，統計的依存構造解析において，文節間の係り受けの確率分布を，決定木学習を用いて推定しているものがある[6)]．この場合の文節間の係り受けの確率分布の決定木学習においては，表-2に示した情報を属性の候補として，上記のエントロピ基準に基づいて属性を1つずつ選択していくことにより**図-3**に示すような決定木が構成され，最終的な確率分布が学習される．この決定木では，まず最初に調べる属性として“二文節間距離”が選ばれ，距離が1文節の場合には次に“係り元文節のタイプ”を，また距離が2～5文節の場合は次に“係り元文節のタイプ”を調べるという形で決定木が構成されている．“二文節間距離=1文節”かつ“係り元文節のタイプ=助詞「の」”の場合はp（係る）$=0.6$，p（係らない）$=0.4$となるという確率分布が学習され，その場合以外は「係る」か「係らない」かが一意に決定する．

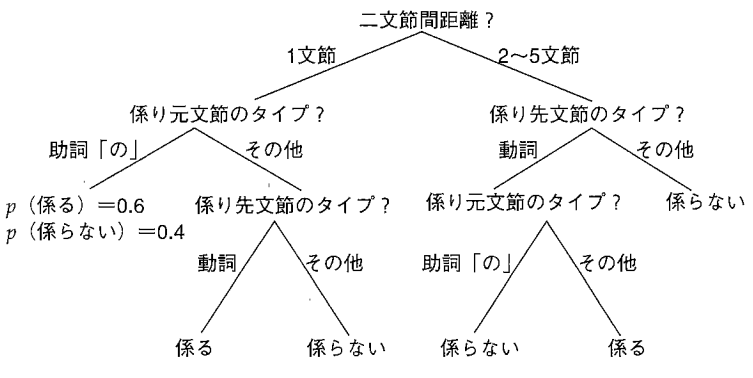

図-3　統計的依存構造解析のための決定木の例

決定リスト学習 ——

決定リスト学習は，決定木学習と同様に，クラス決定問題における学習手法の1つであり，Yarowskyによって提案され，単語の語義・アクセントの決定問題に適用された学習法[13)]がよく知られている．決

定リストは，決定木のような木構造でなく，一次元のリスト構造の形式で属性とクラスの対を列挙した規則である．Yarowskyによって提案された学習法[13]は，属性fのもとでのクラスcの条件付確率$p(c|f)$（厳密には，各確率値に平滑化を施した後の尤度比の対数値 $log\frac{p(c|f)}{p(\neg c|f)}$ ）の大きい順にクラス決定規則を並べるという単純なものであるが，語義決定問題のように単一の属性がクラス決定の有力な手がかりとなりやすい問題には適しているといわれている．一方，決定木学習は，単一の属性ではクラス決定の有力な手がかりとはなりにくいが，複数の属性を組み合わせればクラスの決定が容易になるような問題に適している．また，両者は，属性の選択の戦略においても異なっており，決定木学習が，クラスの偏りがなるべく大きくなり，しかも分割後の個々の部分集合もある程度大きくなるような属性を選択するのに対して，決定リスト学習は，分割後の部分集合の大きさはまったく考慮せず，小規模でもよいからなるべく精度の高い属性が得られるように属性選択を行う．

統計的構文解析において決定リスト学習を利用した研究事例はあまりないが，数少ない例としては，筆者ら[12]が，日本語従属節のスコープの包含関係を学習する問題に焦点を当てて決定リスト学習と決定木学習を比較した事例がある．日本語従属節のスコープの包含関係の学習においては，2つの従属節の間で相対的にどちらのスコープが広いかを決定する必要がある．したがって，属性選択の方式としては，表-2に挙げたような文節の属性の有用性を独立に評価する決定木学習方式ではなく，前従属節と後従属節の情報を組にして属性選択を行う方式が望ましい．筆者らは，決定リスト学習において候補となる属性を，前後の従属節の双方の属性の組として属性選択を行い，**表-3**のような決定リストの形で日本語従属節のスコープの包含関係を決定する確率分布の学習を行った．また，決定木学習との比較においては，決定リスト学習の方が確率モデルとして約10倍細かいモデルを学習し，モデルの被覆率において決定木学習を若干上回る結果が得られた．

最大エントロピモデル

前節で説明した決定木学習および決定リスト学習は，モデルに取り込む属性を自動選択することにより最適な解析性能のモデルを学習するという方法であった．これに対して，モデル自身の特徴として，厳密に最適な属性の選択を行わなくても，頻度の下限などある一定の条件を満たす属性集合を用意しておくだけで，ある程度の性能が達成できるモデルとして最大エントロピモデルと呼ばれるものが知られており，統計的言語処理の分野においても近年よく研究されている．

最大エントロピモデルは，与えられた制約のもとでエントロピを最大化するという条件によって求められるモデルであるが，統計的言語処理の分野では，履歴xの条件のもとでのyの出現確率を表す，以下の条件付確率分布の形式のモデルがよく用いられる．

$$p_\lambda(y|x) = \frac{exp\left(\sum_i \lambda_i f_i(x,y)\right)}{\sum_y exp\left(\sum_i \lambda_i f_i(x,y)\right)}$$

ここで，関数f_iは，素性関数（feature function）と呼ばれる二値関数で，与えられた事象(x, y)において素性関数f_iに相当する制約が成り立つか否かを表す（決定木学習・決定リスト学習における属性に対応）．またλ_iは，素性f_iのパラメタを表し，各素性f_iがモデルの中でどれだけ重要な役割を果たすかの重みを表す．最大エントロピモデルは，与えられた制約のもとでエントロピが最大となるモデルであるので，本来，未知データに対して確率値をなるべく一様に配分するようなモデルの形式になっており，データの過疎性に強いという特徴を持つ．具体的には，上式のモデルの形式において，与えられた事象(x, y)において各制約f_iが成り立つ場合にのみ，モデルがパラメタλ_iの値に依存する．したがって，未知データにおいて適用範囲が比較的狭い制約が満たされない場合でも，適用範囲の広いより一般的な素性を用意しておくだけで，その素性に対するパラメタが適用され，平滑化と同等の効果が得られるという利点がある．さらに，モデルの詳細な形式を手動で調整しなくても，必要と思われる制約を素性として用意しておくだけで，各素性f_iのパラメタλ_iを最尤推定することによりある程度の性能を持ったモデルが得られる．特に，素性間に依存関係があり，そのうちの最適な素性を簡単には選択できな

従属節の付属語部分の属性f		クラスc（後従属節が前従属節を包含するか？）	確率値 $p(c\|f)$	属性fの頻度
前従属節	後従属節			
連用形	判定詞（┐節末）	しない	1	548
連用形	“では”	しない	1	536
⋮	⋮	⋮	⋮	⋮
読点無	読点有，“のが”	する	1	123
⋮	⋮	⋮	⋮	⋮
読点無，副詞（┐節末）	読点有，“が”	する	1	10
読点有	読点無，判定詞（┐節末）	しない	0.997	1541
⋮	⋮	⋮	⋮	⋮
副詞的名詞	連用形	しない	0.538	1280
（無条件）	（無条件）	しない	0.5378	87366

表-3　日本語従属節の包含関係を判定する決定リストの例

いような場合でも，そのままモデルを推定することが可能である．

以上のような利点から，最大エントロピモデルは，統計的言語処理においてもいくつかの問題に適用され，その有用性が確認されている．そのうち，統計的構文解析においては，前述の決定木学習と同様に，構文解析における個々の部分問題の確率分布の推定において最大エントロピモデルが用いられることが多い．たとえば，英語においては，句構造を組み上げる際の確率分布を最大エントロピモデルによって推定するものがある[9]．また，日本語においては，統計的依存構造解析において，式（4）の文節間の係り受け確率 $p(b_i \rightarrow b_j)$ を最大エントロピモデルによって推定するものがある[10]．その他には，構文解析における統語的曖昧性解消での利用を目的として，動詞の下位範疇化の確率モデルを最大エントロピモデルによって推定するものがある[11]．

複数モデルの融合 ——

これまで説明した3つの方法は，いずれも，モデルに取り込む情報を自動選択するなどして1つのモデルを最適化しようという方法であった．これに対して，まったく挙動の異なる複数の構文解析モデルの出力を受け取って，各モデルの長所をうまく生かすことにより，全体の解析性能が上がるように複数モデルの出力を融合するという方法がある[7]．これは，基本的には多数決の原則に基づくものである．具体的な方法としては，多数派のモデルによって支持された構成素を集めて句構造木を構成する融合法や，各文ごとに，他のどのモデルの出力する句構造木ともなるべく似通った句構造木を出力するモデルにスイッチする融合法などを試している．解析精度の格差が数%程度の3つのモデル[2),4),9)]の融合を行った結果では，最良のモデルよりも1～2%の精度向上を達成した．

現状のまとめと今後の展望

統計的構文解析の研究においては，現在利用可能な限られた言語コーパス（ここでは構文構造付コーパス）を最大限利用して最適な解析性能を達成することが重要であり，本稿では，これまで得られた研究の成果について説明してきた．最後に，現状を簡単にまとめるとともに，今後必要と思われる研究の方向を展望してみたい．

英語・日本語の各言語を対象とするモデルにおいて，それぞれ一定の結果が明らかになってきてはいるが，基本的に，英語と日本語ではお互いに異なったモデルについての研究・評価が行われており，モデルの性能の言語依存性についてはまだまだ不明な点が多い．特に，片方の言語で一定の性能が確認されたモデルでも，もう一方の言語ではほとんどその性能が確認されていないものが多い．たとえば，現在，英語で最高の性能を達成しているモデル[3)]が，果たして日本語でどのような性能を示すのかは非常に興味深い点であるといえる☆4．

また，決定木学習など，モデルに取り込む情報を自動選択する手法が，統計的構文解析モデルの学習において実際にどこまでの効果があるのかについても，まだ不明な点も多い．たとえば，現在，英語で最高の性能を達成しているモデル[3)]は，人間の手によって確率文脈自由文法を巧妙に拡張することにより実現されたもので，決定木学習などの機械学習による属性の自動選択によって実現されたものではない．しかし，このように人間の手によって巧妙に調整されたモデルと，モデルに取り込む情報の自動選択等の手法により学習され得るモデルの間の優劣が網羅的に調べられたわけでもない．たとえば，モデルに取り込む情報の自動選択などの考え方によって，人間が巧妙に調整したモデルをさらに改良する余地があるのか否かなどは非常に興味深い点であり，今後の研究の成果が期待される．

☆4　ただし，現在，日本語で利用可能な構文構造付コーパスには，構成素の文法カテゴリの情報が付与されていないため，英語で一定の性能が確認されたモデルをそのまますぐに日本語に適用できるというわけではない．モデルの性能の言語依存性を厳密に調べるには，構文構造付コーパスの整備まで含めて検討する必要があるといえる．

参考文献

1) Black, E.: Towards History-Based Grammars: Using Richer Models for Probabilistic Parsing, In Proceedings of the 31st Annual Meeting of ACL, pp.31-37 (1993).
2) Charniak, E.: Statistical Parsing with a Context-Free Grammar and Word Statistics, In Proceedings of the 14th AAAI, pp.598-603 (1997).
3) Charniak, E.: A Maximum-Entropy-Inspired Parser, In Proceedings of the 1st Conference of NAACL, pp.132-139 (2000).
4) Collins, M.: Three Generative, Lexicalized Models for Statistical Parsing, In Proceedings of the 35th Annual Meeting of ACL and the 8th Conference of EACL, pp.16-23 (1997).
5) 藤尾正和, 松本裕治: 語の共起確率に基づく係り受け解析とその評価, 情報処理学会論文誌, Vol.40, No.12, pp.4201-4212 (Dec. 1999).
6) 春野雅彦, 白井　諭, 大山芳史: 決定木を用いた日本語係受け解析, 情報処理学会論文誌, Vol.39, No.12, pp.3177-3186 (Dec. 1998).
7) Henderson, J. C. and Brill, E.: Exploiting Diversity in Natural Language Processing: Combining Parsers, In Proceedings of the 1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora, pp.187-194 (1999).
8) Magerman, D. M.: Statistical Decision-Tree Models for Parsing, In Proceedings of the 33rd Annual Meeting of ACL, pp.276-283 (1995).
9) Ratnaparkhi, A.: Learning to Parse Natural Language with Maximum Entropy Models, Machine Learning, Vol.34, No.1/2/3, pp.151-176 (1999).
10) 内元清貴, 関根　聡, 井佐原均: 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, Vol.40, No.9, pp.3397-3407 (Sep. 1999).
11) Utsuro, T., Miyata, T. and Matsumoto, Y.: General-to-Specific Model Selection for Subcategorization Preference, In Proceedings of the 17th COLING and the 36th Annual Meeting of ACL, pp.1314-1320 (1998).
12) 宇津呂武仁, 西岡山滋之, 藤尾正和, 松本裕治: コーパスからの日本語従属節係り受け選好情報の抽出およびその評価, 自然言語処理, Vol.6, No.7, pp.29-60 (1999).
13) Yarowsky, D.: Decision Lists for Lexical Ambiguity Resolution: Application to Accent Restoration in Spanish and French, In Proceedings of the 32nd Annual Meeting of ACL, pp.88-95 (1994).

（平成12年5月24日受付）