語の共起確率に基づく係り受け解析とその評価

藤尾正和† 松本裕治†

本論文では，粗い日本語係り受け解析手法として，語の共起確率に基づく係り受け解析手法を提案し，その評価を行う．学習および評価用コーパスとしてEDRコーパスを使用し，文節および文単位の係り受け精度を調べる．またどのような係り受け関係名において誤りが多いのか調べるため，関係名ごとの解析精度も調べる．英語において，比較的近いモデルおよび情報を用いたCollins (1996)のモデルと文節単位の係り受け精度を比較した結果，EDRコーパスを使用した日本語解析に関しては，我々のモデルの精度がCollinsのモデルを上まわった．また，現状の統計モデルのもとでさらに解析精度を上げるため，再現率を犠牲にして適合率を上げる手法（部分解析），および適合率を犠牲にして再現率を上げる手法（冗長解析手法）についても提案する．“確信度”（乾ら，1998）を使用した**Global**のほか，**Local/norm**，**Ratio/next**の3つの手法について評価を行った結果，少くとも我々の統計モデルを使用する場合，解析精度，速度などを考慮すると**Ratio/next**が優れているということが分かった．

Dependency Analysis Based on Lexical Collocation Probability and Its Evaluation

Masakazu Fujio† and Yuji Matsumoto†

We present statistical models of Japanese dependency analysis based on lexical collocation probability. We use the EDR corpus for both training and evaluation, and evaluate the precision of the models in terms of correct dependency pairs and correct sentences. We measure the correct rate of dependency pairs for each type of dependency relation. To achieve higher performance under the current statistical parsing model, we propose a method that intend to acquire higher precision rate at the cost of recall rate (partial parse), and the method to acquire higher recall rate at the cost of precision rate (redundant parse). We propose and compare three partial (redundant) parse methods, **Global**, **Local/norm**, **Ratio/next**, and find that **Ratio/next** is superior to others among our methods.

1. はじめに

近年大規模なコーパスが利用可能となり，構文解析の分野でも，さかんに統計的解析手法が研究されてきた．現在さまざまな統計的手法および規則主導型の係り受け解析システムが存在するが，精度のうえでは規則主導の解析手法が，係り受け単位でおよそ9割の正解率を示し，最も高精度を実現している．しかし，規則主導型には，適用分野を変更したり拡大したりするたび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．

最も基本的な統計的構文解析モデルとしては，文脈自由文法（CFG）の各文法規則に確率を付与した確率文脈自由文法（PCFG）があげられる．しかしPCFGは，そのままのモデルでは，語彙的優先度が各規則の適用確率に反映されないため（語に関する確率が出現するのは品詞から語を生成するときの確率だけである），曖昧性の解消にはあまり役に立たない．そこで非終端記号に主辞情報を取り入れたり，係り受け関係にある句の主辞間距離を取り入れたモデル[4),8)]が提案されてきた．語彙情報の重要性は，PCFG以外の枠組みでも認識され，Tree Adjoining GrammarやLink Grammarなどでも，語彙情報を取り入れた統計モデルが提案されている[7),12)]．また，L-PGLR（Lexicalized PGLR）を使用した日本語の統計的係り受けモデルとしては，Shiraiらの研究[14)]がある．Shiraiらのモデルでは，確率モデルを構文モデルと語彙的優先度に分離し，パラメータ学習のための便宜を図っている．

ほかに，CFGをベースとしないモデルとして，確率決定木により解析木を伸張していくMagermanのモデル[9)]や，依存関係にある句間の依存確率のみを使用したCollinsのモデル[2)]などがある．Collinsのモデルでは，句の主辞や距離その他の統語的特徴を使用

† 奈良先端科学技術大学院大学
Graduate School of Information Science, Nara Institute of Science and Technology

して，2つの句が特定の係り受けタイプの依存関係となる確率を推定する．

現在英語の分野では，Charniak[1] に報告されているとおり，先にあげた Magerman[9]，Collins[2] および Charniak[1] によるモデルが，係り受けの精度で約86～87%の精度を実現している．

ところで，これまであげた統計的解析の研究は，いずれも完全に係り受け解析したときの精度のみを議論しており，再現率を犠牲にして適合率を上げる場合（以降，部分解析と呼ぶ）や，逆に適合率を犠牲にして再現率を上げる場合（以降，冗長解析と呼ぶ）については，特に述べられていない．

むしろ応用場面では，語彙知識を獲得する場合や情報検索をする場合，あるいはより詳細な言語解析の前処理として利用する場合など，部分解析や冗長解析が必要な場面も考えられる．

そこで我々は，まず語の共起確率に基づく我々の日本語係り受け解析モデルを提案し，解析精度を評価した．そして現在の統計モデルのもとで，さらなる解析精度を実現するため，統計的部分解析手法と冗長解析手法を提案し，その評価を行った．

本研究の統計的係り受け解析モデルでは，多くの解析システムで用いられている CFG などの文法規則を使用せず，各係り受け関係の確率を，単語，品詞，付属語，距離などの表層レベルの情報を用いて推定し，統計的係り受け解析を行う．これは，先にあげた Collins のモデルに近い．

我々のモデルは，これまで提案されてきたモデルのうち，解析履歴を確率の前件部に持つ統計モデルなどと比べると，記述力は弱い（他の文節から係られているか，あるいは係り先に他に何が係っているか，などの情報を扱えない）．しかし，条件付き確率の推定に使用する表層的，統語的属性を工夫することで，ある程度の精度を実現できると考えている．したがって我々はまず，使用属性をいくつか変えて，我々のモデルの性能を調べた．

日本語の構文解析で，統計的部分解析に着目した研究としては，乾ら[5] の研究がある．乾らの方法は，まず上位複数解を決定したあと，それぞれの解に付与された確率を用いて，各係り受け関係の確信度を計算する．確信度とは，各係り受け関係を含む解析結果の確率を足しあわせ，解析結果全体の確率の和で正規化したものである．そして，この確信度の値がある閾値以上であれば，部分解析結果として出力する．確率モデルとしては，Shirai ら[13] で提案されているモデルを使用している．

我々の統計モデルにおいても，乾らと同様の手法およびほかに2つの手法を提案し，解析精度の評価を行った．また，部分解析手法で使用した尺度を冗長解析手法にも適用し，解析精度の評価を行った．冗長解析とは，適合率を犠牲にして再現率を上げる方法で，各文節に対し複数の係り先候補の出力を許すものである．

以下の章では，まず2章で今回使用した統計モデルについて述べたあと，3章で，システムの処理の全体の流れについて説明する．そして4章で，まず係り受け構造をすべて決定するときの評価を行い，5章と6章で，部分解析アルゴリズムと冗長解析アルゴリズムの説明と評価を行う．そして7章でまとめを述べる．最後の8章に，他の研究との比較について述べる．

2. 統計モデル

本研究で用いる，文節属性☆に基づく係り受け解析モデルについて説明する．文節属性については，2.1節で詳しく説明する．

初めに，使用する記法について説明する．

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 $\langle w_1,t_1\rangle,\ldots,\langle w_n,t_n\rangle$ を T，文節にまとめられ属性付けされた文節列 $\langle b_1,\mathbf{f_1}\rangle,\ldots,\langle b_m,\mathbf{f_m}\rangle$ 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．ただし $Dep(i)$ は文節 b_i の係り先の文節番号を表す．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．

ここでは，係り受けの構造とは，対象にしている文において，次の2つの制約を満す係り受け関係の組合せをいう．

(1) 文末を除き各文節は文末側に必ず1つの係り先を持つ

(2) 係り受けは非交差

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．これは次のような式で書ける．

$$P(L,F,T|S)=P(L|F,T,S)P(F|T,S)P(T|S)$$

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るとすると，上式の右辺は次のように簡単化できる．

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

☆ 具体的には，文節の主辞，読点の有無，係り関係など．

表 1 係り受け関係の例（括弧は文節を表す）
Table 1 Examples of "bunsetsu" features.

[まっすぐに]$_1$[のびた]$_2$[道のように]$_3$[思えた.]$_4$						
	係り側		文節間属性		受け側	
	主辞	関係名	文節数	句読点数	主辞	関係名
1 → 2	まっすぐだ	形容詞/連用	0	0	のびる	動詞/タ形
2 → 3	のびる	動詞/タ形	0	0	道	の-ようだ（連用）
3 → 4	道	の-ようだ（連用）	0	0	思う	動詞/ 基本

したがって，$P(L|F)\ P(F|T)\ P(T|S)$ の 3 項の積が最大になるように，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，形態素解析システム「茶筌」[10] の最適解出力を使用し，文節区切りと属性の付与は，人手で記述した文節区切りルールにあてはめることで決定的に行った．すなわち $P(T|S)=1$，$P(F|T)=1$ となる．よって以下の式で表される係り関係の組合せ，L_{best} を求めればよい．

$$L_{best} = \underset{L}{\operatorname{argmax}}\, P(L|F)$$

さらにここで，それぞれの係り受けは独立であると仮定すると，以下のように展開できる．

$$P(L|F) = \prod_{i=1}^{m-1} P(i \overset{\text{rel}}{\rightarrow} j \mid \mathbf{f_1} \ldots \mathbf{f_m}) \qquad (2)$$

$P(i \overset{\text{rel}}{\rightarrow} j \mid \mathbf{f_1} \ldots \mathbf{f_m})$ は，文節区切りが実行され，属性集合 $\mathbf{f_1} \ldots \mathbf{f_m}$ が与えられたときに，文節 $\mathbf{b_i}$，$\mathbf{b_j}$ が，係り受け関係にある確率を表している．

我々のモデルでは，確率 $P(i \overset{\text{rel}}{\rightarrow} j \mid \mathbf{f_1} \ldots \mathbf{f_m})$ を，語彙共起確率と距離確率の積で定義する．比較のために，この 2 つの確率を分離しないモデル（Collins 96 [2] のモデルに相当）についても述べる．

各確率を定義するため，以下の節では，まず今回使用した属性について説明する．

2.1 使用文節属性

文節属性は，各文節に対して定義されるものと，2 つの文節により定義されるものがある．前者を文節属性，後者を文節間属性と呼ぶ．

文節属性

- 文節の主辞 h_i
- 関係名 r_i
- 句読点 p_i

これらの属性は，各文節に対して定義される．文節の主辞は，各文節の主要な語で，何を主辞にするかは，文節区切りルールの中で記述することができる．我々は主辞属性として，見出し語（活用語の場合はその原形），品詞，意味クラスのいずれかを用いた．意味クラスとして，主辞の分類語彙表 [6] における分類番号を使用した．

表 2 実験に使用した文節間属性の種類
Table 2 Variations of distance features.

文節間属性の種類	
dst1	文節対間の文節数と読点数
dst2	文節対間の文節数と読点数．ただし，2 より多い文節数は区別しない
dst3	文節対間の文節数と読点数．ただし，2 より多い文節数，読点数は区別しない
dst4	文節対間の文節数と読点数．ただし 2〜5 の文節数は "B"，5 以上は "C" というように表現する．

関係名とは，係り受けの種類を表すもので，付属語をともなう文節の場合はその付属語の語彙/品詞/活用形の 3 つ組（複数の付属語から成る場合は，すべてを使用），そうでない場合は，一番最後の内容語の品詞と活用形で定義される．**表 1** に例をあげる．

文節間属性

- 文節間文節数 d_{ij}
- 文節間句読点数 p_{ij}

文節間属性とは，2 つの文節により定義される属性で，おもに文節間の距離概念を導入するのに使用する．上にあげた文節間の文節数や文節間の句読点数などがその例である．表 1 に例をあげる．

文節間属性に関しては，文節数などの値をそのまま使用する場合，あるいは "2 以上" を 1 つのグループと見なす方法など，さまざまな方法が考えられる．2 つ以上を区別しないのは，隣とその次までは実際の距離が意味を持つが☆，それ以上だと，たとえば 5 であるか 6 であるかにそれほど意味がないと考える場合である．我々は，**表 2** にあげた距離属性のいずれかを用いた．

上で述べていない文節間属性として，受け側文節との間に存在する格助詞の数や，述語の数などが考えられる．これらは係り側文節が，格助詞を含む場合など

☆ 助詞 "の" の係り受けの場合など．

に有効と考えられる．しかしいくつかの実験を繰り返した結果，d_{ij}，p_{ij} を使用した場合より精度が悪かった（理由については 7 章で検討する）．したがって，4 章では d_{ij}，p_{ij} についての実験結果のみを示す．

2.2 主辞共起確率と距離確率

上でも述べたように，各係り受け関係の確率を，主辞共起確率と距離確率の積で定義する．これらの確率は，前節で説明した属性集合を使用して，以下のように定義される．

$$P_h^{ij} = P(i \xrightarrow{\text{rel}} j \mid h_i, r_i, p_i, h_j, r_j, p_j) \quad (3)$$

$$P_d^{ij} = P(i \xrightarrow{\text{rel}} j \mid r_i, p_i, d_{ij}, p_{ij}) \quad (4)$$

P_h^{ij} は文節 i と j の組合せに関する主辞共起確率を表し，P_d^{ij} は距離確率を表す．

ここで，距離確率の定義に，属性 r_i，p_i が使用されているが，これは，係り受け距離の性質は関係名や，読点の有無に依存すると考えられるからである．また，逆にこの 2 つの属性を考えることで係り受けの距離的な性質はほぼ決まり，主辞共起確率と独立になると考えた．

これらの確率を用いて，各係り受け確率は，以下のように定義される．

$$P(i \xrightarrow{\text{rel}} j \mid \mathbf{f_1}, \ldots, \mathbf{f_m}) = P_h^{ij} \times P_d^{ij}$$

係り受け確率を，2 つの独立な確率に分離して考えることで，データの過疎性を軽減することができる．

それぞれの確率は，最尤推定により，解析済みコーパスから以下の式で求めることができる．$C(...)$ はコーパス中の出現頻度を表す．

$$P_h^{ij} = \frac{C(i \xrightarrow{\text{rel}} j,\ h_i, r_i, p_i, h_j, r_j, p_j)}{C(h_i, r_i, p_i, h_j, r_j, p_j)} \quad (5)$$

$$P_d^{ij} = \frac{C(i \xrightarrow{\text{rel}} j,\ r_i, p_i, d_{ij}, p_{ij})}{C(r_i, p_i, d_{ij}, p_{ij})} \quad (6)$$

2.2.1 主辞共起確率の smoothing

主辞共起確率の中で使われている h_i という属性は，もともと主辞の見出し語を表すが，語レベルの共起頻度をそのまま使用する場合，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，以下の 3 つの主辞共起確率モデルを考えた．

(1) **POS モデル** 主辞情報として，品詞のみを使用．
(2) **LEX モデル** 主辞情報として，語と品詞を使用．
(3) **BGH モデル** 主辞情報として，語と語の意味クラスと品詞を使用．

BGH モデルの意味クラスとして，分類語彙表[6] における分類番号を使用した．LEX モデルでは，語による共起確率がデータ中に存在すればそれを使用し，存在しなければ品詞に関する共起確率を使用した．BGH モデルでは，語による共起確率がデータに存在すればそれを使用し，存在しなければ意味クラスによる共起確率を調べ，存在すればそれを使用し，存在しなければ品詞に関する共起確率を使用した．

表 3 主辞共起確率の推定に使用する属性の組
Table 3 Features used for estimating Head-Collocation probability.

係り側		受け側	
関係名	主辞	主辞	関係名
のびる	動詞	名詞	の-ようだ
のびる	動詞	名詞	の
のびる	動詞	名詞	-

2.2.2 関係名の曖昧性

文節中に付属語が連続して複数現れた場合，関係名の定義は複数考えられる．例として，表 1 の 3 → 4 という係り受けを考えてみる．POS Model においては，関係名の長さにより，主辞共起確率推定用の属性として，**表 3** の組合せのどれかを使用することが可能である．

学習段階では，これらの組合せそれぞれに対し，頻度 1 と数える．関係名を省略しないもの，したものそれぞれの頻度を 1 として数えるのは，解析段階で，バックオフ用のデータとして使用するためである．すなわち，最初に関係名を省略しないで学習データを検索し，データが存在すればその確率を使用する．データが存在しなければ，データが見つかるまで，関係名を順々に短かくしていく．係り側の関係名は，関係名を構成する形態素のうち左端のもから 1 ずつ削り，受け側の関係名は右端のものから 1 つずつ削っていく．ただし係り側の関係名は，最低 1 つの形態素が残るようにする．受け側の関係名については，関係名を構成する形態素がなくなるまで削る．削る順序は受け側の関係名が最初で，受け側の関係名を構成する形態素がなくなったら，それでもデータが見つからなければ係り側の関係名を構成する形態素を 1 つ削り，受け側の関係名は最初の長さに戻した後，同じ操作を繰り返す．

2.3 Collins のモデル

Collins[2] のモデルの対象は英語であり，日本語と異なり係り受けタイプ☆の推定も必要であるという点で，基本的なモデルが異なるが，我々の枠組みの中でモデル化すると，主辞共起確率と距離確率を分離しな

☆ 木構造を考えたときの，係り側ノードの非終端記号，親ノードの非終端記号，受け側ノード非終端記号 3 つの非終端記号で定義される．

いモデルに相当する．

$$P(i \xrightarrow{rel} j \mid \mathbf{f_1}, \ldots, \mathbf{f_m}) \stackrel{\text{def}}{=} P(i \xrightarrow{rel} j \mid h_i, r_i, p_i, h_j, r_j, p_j, d_{ij}, p_{ij})$$

コーパスからの最尤推定も同様にして行える．

$$P(i \xrightarrow{rel} j \mid \mathbf{f_1}, \ldots, \mathbf{f_m}) \approx \frac{C(i \xrightarrow{\text{rel}} j, h_i, r_i, p_i, h_j, r_j, p_j, d_{ij}, p_{ij})}{C(h_i, r_i, p_i, h_j, r_j, p_j, d_{ij}, p_{ij})}$$

3. 解析アルゴリズム

基本的な解析の流れは，次のとおりである．

(1) 形態素解析システム「茶筌」による，入力文の形態素解析（最適解のみの出力）
(2) 品詞タグ，見出し語などをもとにした，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うえで，確率最大となる係り受けの組合せの決定

文節区切り

基本的に自立語列+付属語列を文節とした．文節属性のうち，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の違いに対処するため，文節および属性の定義を，見出し語や品詞による正規表現で記述できるようにした．

各文節間の係り受け確率の計算

解析段階では，すべての付属語を用いた関係名から使用し，学習データ中に存在しなかった場合，統計データが得られるまで，関係名として使用する付属語を順々に減らしていくことで，バックオフ的なスムージングを行った（2.2.2 項参照）．

解析アルゴリズム

2 章で説明した統計をもとに，各文節間の係り受け関係の確率を求める．この中から，次の制約の下で，確率最大となる係り受けパターンを決定する．

(1) 文末を除き，各文節は文末側に必ず 1 つの係り先を持つ
(2) 係り受けは非交差

CYK アルゴリズムを使用することで，以上の制約を満たす解を効率的に計算できる．

4. 提案モデルによる解析精度の評価

学習および評価用のデータとして，EDR[17] コーパスを用いた．EDR の係り受けの単位と，我々のシステムの出力する文節文字列とは一致しないので，文節文字列と EDR の単語とのマッチングを行いながら，EDR コーパスの係り受け情報を，文節単位での係り受け情報に解釈しなおしたコーパスを作成した．作成したコーパスを，20 文おきに 20 個のファイルに分割し，1 つを評価用に，残りを学習に用いた．

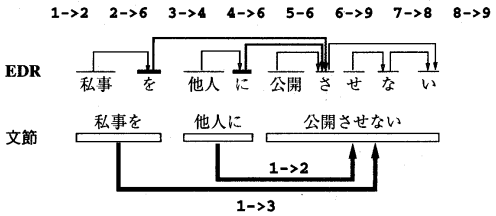

図 1 正解係り先の決定手順
Fig. 1 Process to decide correct dependency relation.

4.1 文節単位の係り受けコーパス作成手順

EDR コーパス中に現れる各文に対し，「茶筌」による形態素解析と，文節まとめあげの段階まで解析を行う．各文節文字列と EDR 括弧付けコーパス中の要素を文字列照合しながら，各文節の正解係り先を特定する．文節の区切りと，EDR コーパスにおける括弧付け要素の区切りが一致する場合は，文節の最後部と一致する EDR コーパスの要素の係り先を含む文節を正解係り先文節とする．区切りが交差する場合は隣へ係るとした．図 1 の例でいうと，1 番目の文節の末尾と照合する EDR コーパスの要素は 2 番目の要素である．今ここで，EDR コーパスの 2 番目の要素は，EDR コーパスの括弧付けから 6 番目の要素に係っていることが分かったとする．この 6 番目の EDR コーパスの要素は，文節でいうと 3 番目の文節に含まれるので，1 番目の文節の係り先は 3 番目の文節であることが分かる．

コーパスを作成する段階で，文節区切りルールにあてはまらず，文節区切りに失敗した文（形態素解析が誤り，文節ルールにあてはまらなかった場合など）や，EDR コーパスの間違いや係り受けの交差により，係り先がおかしくなる文は取り除いた☆．

4.2 評価方法

我々のモデルと Collins のモデルを使用した場合について，主辞属性と距離属性を変えて，解析精度の比較を行った．主辞属性として LEX，POS，BGH，距離属性として dst1～dst4 の組合せについて，実験を行った．

☆ 残った文，つまり使用した文は 206381 となった．

各モデルにおけるシステムの精度は，正解した係り受け文節数の割合（適合率）で計算する．

$$\text{適合率} = \frac{\text{係り先の正しい文節数}}{\text{出力係り受け数 (総係り受け数)}}$$

すべての文節の係り先をそれぞれの文節の隣とした場合を Base Line と呼び，最低基準とした．

各モデルにおいて，出現数 5 以下の属性の確率値は使用しなかった．これは，少ない出現頻度による確率推定は信頼性が低いのと，実際予備実験において，獲得したデータをそのまま用いるより，適当な出現頻度数で足切りした方が精度が良かったからである．

関係名が付属語から成る場合，品詞を使用せず，見出し語のみにした実験も行った．“格助詞/が” と “接続助詞/が” の区別がなくなるなどの問題点があるが，形態素解析システム「茶筌」による形態素解析誤りの影響を減らせる可能性があるので試してみたが，結果には優意な差が認められなかった．

4.3 文節単位の精度

表 4，**表 5** に 19 万文で学習を行い，残りの 1 万文で解析を行ったときの，各モデルの解析精度を示す．表 4 は，我々のモデルを使用した場合の解析精度である．距離属性の欄が “-” となっている部分は，距離確率を使用しなかった場合の結果を示している．表 5 は，2.3 節で説明した Collins のモデルを使用して解析した場合の解析精度である．

分類語彙表の意味クラスは，トップノードからの深さが 2 番目の階層から 6 番目の階層までを使用し，それぞれに対し実験を行った．しかし，いずれも 6 番目の階層までを使用した場合が最も精度が高かったため，以降 6 番目までの階層を使用した結果を示している．

まず我々のモデルを使用した場合であるが，LEX モデルで dst2 を使用した場合が最も精度が良かった．ただし POS モデルとの差はそれほど見られない．また，dst1 を使用した場合の精度が著しく悪い．むしろ，距離確率を使用しない方が，LEX，POS，BGH のいずれの場合でも精度が良かった．これは，距離属性として文節間距離をそのまま使用した場合，必要以上にデータが分散し，確率の推定精度が落ちるためと思われる．実際，出現頻度 10 以下の距離確率パラメータが使用された回数（1 万文を解を析した場合）は，dst2 を使用した場合 398 回なのに対し，dst1 を使用した場合は 4009 回であった☆．

語彙情報を使用した場合のデータの過疎性を緩和する目的で，意味クラスを使用した BGH モデルについても実験を行っているが，使用したいずれの距離属性においても，BGH モデルは POS モデルよりも精度が低かった．この結果は学習量を変えた実験でも同様であった．理由の 1 つとして，意味クラスとして使用した分類語彙表の分類が，係り受け関係の違いを反映していない，ということが考えられる．

Collins のモデルを使用した解析では，smoothing の方法として，Collins[2] にあるような出現頻度による重みづけや，我々のモデルと同様の back-off 手法など，いくつか手法を試した．その中で，最も精度の良かったものについての結果を示している．Collins のモデルを使用した場合も，主辞共起確率モデルとして LEX，距離属性として dst2 を使用した場合の精度が最も高かった．

我々のモデルと Collins のモデルの結果を見比べると，主辞共起確率と距離確率を分離した我々のモデルの結果の方が良い．日本語の場合，関係名☆☆も属性の一部となるため，英語での学習の場合と比べてよりデータの過疎性が強まり，主辞共起確率と距離確率を分離した我々のモデルの精度が良かったと思われる．

以下に，POS モデルでは間違えたが，LEX モデルでは正解した例をあげる．“○” は，ボックスで囲まれた文節の，EDR での正解係り先を示し，“×” は，シ

表 4 文節単位の係り受け精度（我々の統計モデル）
Table 4 Precision of correct dependency relations under our statistical model.

	主辞属性			
距離属性	POS	LEX	BGH	Base Line
-	0.7615	0.8001	0.7710	
dst1	0.6670	0.6836	0.6633	
dst2	0.8649	0.8689	0.8524	0.6237
dst3	0.8637	0.8675	0.8524	
dst4	0.8626	0.8674	0.8534	

表 5 文節単位の係り受け精度（Collins のモデル）
Table 5 Precision of correct dependency relations. Equal model to Collins' model.

	主辞属性		
距離属性	POS	LEX	BGH
dst1	0.7954	0.8146	0.7666
dst2	0.8131	0.8189	0.8094
dst3	0.8021	0.8169	0.7775
dst4	0.7722	0.7987	0.7483

☆ 出現頻度 10 以下の距離確率パラメータを使用しないようにしても，推定できない例が増えるため，精度は下がる．

☆☆ 付属語で定義する係り受け関係のこと．Collins の英語のモデルでは，係り受け関係にある文節の非終端記号と親ノード（構文木を考えた場合）の非終端記号の 3 つ組で定義し，解析段階で，確率最大となる係り受け構造と係り受け関係を推定する．

ステムの出力した係り先を示す.

- ダライ・ラマは, [ラサを] 中心に$_{\bigcirc}$ 東チベットに 影響力を 持ってきた.$_{\times}$
- [炉の] 優れた$_{\times}$ 治金特性が$_{\bigcirc}$ 明らかにされた.

最初の例では, 主辞として品詞を使用した場合, “格助詞を” と “格助詞に” の共起が低いため, 動詞 “持ってきた” に係る解を出力してしまう. しかし, 語情報まで使用すると, “名詞+を中心に” という表現は, 頻出する言い回しなので, 係り受け関係になる確率も高く, 正解出力が得られている.

2番目の例では, “名詞+の 優れた” という係り受けは, “性能の 優れた” などの表現に見られるように自然な表現であるが, 語情報まで使用すると, “炉の優れた” という表現は不自然となる. 逆に “炉の 特性” という係り受けの方が確率が高くなっている.

逆に, LEX モデルで間違えたが, POS モデルでは正解となった例も見られた.

- [決まったことは] 周知の 通りだが$_{\bigcirc}$ 新渡戸には 議論が あった.$_{\times}$

この場合は, “ある” という単語の出現頻度が非常に多く, 名詞あるいは名詞相当句が非常に係りやすいためであると考えられる.

このように, 品詞の代わりに語を使用することで間違える例もあるが, 平均すると, 語まで使用することで, よりもっともらしい確率の推定ができ, 文節単位の解析精度の差にも表れていると考えられる.

4.4 文単位の精度

学習に使用しなかった1万文を解析して, 文単位の正解率を求めた. 文中の係り受けすべてが正しい場合に, その文が正解であると定義する. 各文について上位 n 解を求め, その中に係り受け構造がすべて正しい解が存在するかどうかを調べる. この数 n のことを, 解析候補数と呼ぶ.

解析に使用したモデルでは, 主辞共起確率モデルとして LEX モデル, 文節間属性として dst2 を使用した.

図2 は, 横軸が解析候補数を表し, 縦軸が正解率を表す. たとえば, 解析候補数が5のときの正解率とは, 上位5番目までの解に正解係り受け構造が含まれる割合を示している. 文中の文節数により, 結果を分けて示す. 評価した1万文の中では, 最大の文節数は28であった.

文節数に応じて解析候補の数が著しく変化するので, 正解を得る困難さが異なる. 文節数6のときの可能な解析結果数は42通りで, 文節数7から9のときは, 163～1430の解析候補が存在する☆.

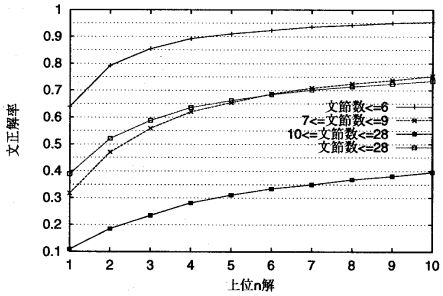

図2 上位 n 解を用いた文単位の正解率
Fig. 2 Precision of correct parses within n-best parses.

文節数7から9の文に関しては, 上位10解で, 75%近い精度を実現できている. また, 文節数6以下のものに関していえば, 上位4解で, ほぼ90%の文正解率になることが分かる. 文節数10から28の場合の文正解率は低いが, この場合の可能な解析結果数が, 4862～約70兆であることを考えると, 仕方ないと思われる.

4.5 関係名ごとの評価

表6 は, 学習に使用しなかった1万文を解析して, 各関係名ごとに解析精度を調べたものである. 一部の関係名のみ示している. 副助詞 “は”, 時相名詞, 普通名詞, 動詞の連用形は, 出現頻度が多く, 精度も低い. 時相名詞の精度が悪いのは, 季節や日時のような表現で, 離れた位置にある動詞に係る場合と, すぐ隣の名詞と複合名詞句を形成する場合とがあるためと思われる. 普通名詞の精度が悪いのは, 並列句を形成する場合と動詞に係る場合との曖昧性が原因と思われる.

係り受け精度が70%以下のものを調べると, 学習コーパス中の出現頻度が少ないものがほとんどである☆☆. それらの関係名を構成する付属語の品詞の種類は, 名詞性名詞助数辞, 名詞接続助詞, 副詞的名詞, 形容詞性述語接尾辞, 述語接続助詞, 動詞性接尾辞などで, 我々の現在の設定では, これらの機能語を持つ文節の関係名は, 語により細分化されている. 出現頻度の少ない付属語に関しては, 語による関係名の区別を使用しないなどの工夫が有効かもしれない.

係り受け解析で十分な精度が実現できれば, 動詞の項構造などの言語知識の獲得や, 翻訳, 要約などへ応用することが考えられる. しかし実現した精度は文節単位で約87%の精度であり, その要求に十分応えられ

☆ $k(n) = \sum_{i=2}^{n} k(i-1) * k(n-i+1); k(1) = k(2) = 1$
☆☆ 3けた (111回) の出現頻度のものはわずかに1種類 (など/名詞性名詞接尾辞/) だけで, 1けたのものも多い.

表 6 関係名ごとの解析精度の分類．学習に使用しなかった 1 万文で解析
（主辞共起モデル=LEX，文節間属性=dst2）
Table 6 Precision for each dependency type.

関係名（語彙/品詞/活用形）	適合率	正しい係り受け数	全体の係り受け数
/形容詞/連体	0.9558	1039	1087
を/格助詞/	0.9415	7178	7624
の/名詞接続助詞/	0.9251	11210	12118
に/格助詞/	0.9197	5901	6416
する/動詞/基本	0.9047	503	556
形容詞/連用	0.8923	978	1096
と/述語接続助詞/	0.8878	696	784
が/格助詞/	0.8856	5115	5776
/動詞/基本	0.8828	1379	1562
/動詞/タ形	0.8594	721	839
が/述語接続助詞/	0.8529	580	680
と/格助詞/	0.8450	1586	1877
も/副助詞/	0.8412	1706	2028
で/格助詞/	0.8306	1015	1222
だ/判定詞/テ形	0.8263	980	1186
は/副助詞/	0.8037	6276	7809
/動詞/テ形	0.8033	960	1195
/時相名詞/	0.7973	1192	1495
/普通名詞/	0.7599	1190	1566
/動詞/連用	0.7516	838	1115

る精度とはいえない．そこで以下の章では，再現率を犠牲にして適合率を上げる部分解析と，適合率を犠牲にして再現率を上げる冗長解析手法を提案し，解析精度を調べた．

5. 部分解析手法

部分解析アルゴリズムとは，一部の係り受けしか出力しないことによって，再現率を犠牲にする一方，高い適合率を得る手法である．部分解析を行うには，各文節ごとにいずれかの係り先を選択し，出力するか否か，決定する必要がある．

この章では，部分解析手法（あるいは係り受けの出力判定のための尺度）として以下にあげる 3 つの尺度を提案する．以下で提案する尺度は，次章で述べる冗長解析においても使用する．

- **Global** n-best 解を求めたあと，各係り受けについて，出現した回数だけ出現した解の確率を足しあわせ，解析結果全体の確率の和で正規化した値〔⇒ 確信度（乾ら[5]）〕
- **Ratio/next** 各文節の係り受け候補を確率で順序付けしたあとの，第 1 候補と第 2 候補の確率の比の値
- **Local/norm** 係り受け確率を係り側文節の係り先候補の確率全体で正規化した値

これらの尺度を使用し，以下の手順で部分解析を行う．

(1) 各尺度による値を計算する（この値を信頼値と呼ぶことにする）．

(2) 各文節ごとに，信頼値の最大となる係り先を選ぶ．

(3) 各係り受け関係が，あらかじめ決めた閾値以上であれば出力し，そうでなければ出力しない．

Global は，乾ら[5]で使用されていた確信度という尺度を我々のモデルにおいて計算したものである．今回の実験では，n の値を 50 にして行った．

Ratio/next は，係り受け確率が最大のもののうち，次候補との違いが大きいものを出力するために考えた尺度である．

Local/norm において，係り受け確率を係り先候補の確率全体で正規化しているのは，係り受け確率が等しい係り受け関係（ただし係り側文節が異なるとする）があった場合，係り側文節の位置が文末側の方が，競合する係り先候補が少ないという点で，より信頼できるのではないかと考えたからである☆．

5.1 解析精度の評価

部分解析アルゴリズムでは，出力される係り受けは，文節総数に対して一定ではない．よって，以下に定義する適合率と被覆率で，精度を評価する．

$$\text{適合率} = \frac{\text{出力した係り受けのうち正解した数}}{\text{出力した係り受け数}}$$

☆ 確率が同じ 0.5 であった場合でも，残りの係り先候補の係り受け確率の和が大きいと，正規化した値は小さくなる．

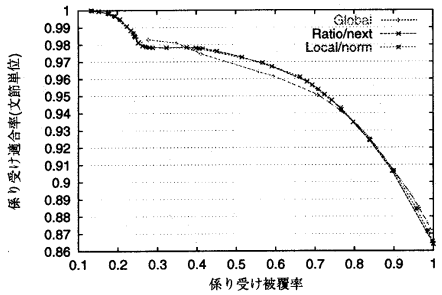

図 3 適合率と被覆率の関係
Fig. 3 Relationship between precision and coverage.

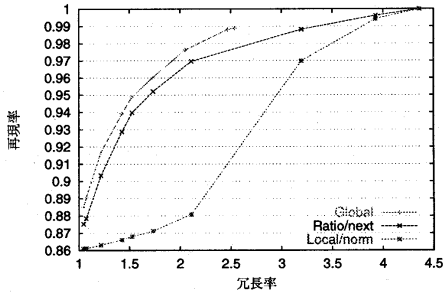

図 4 冗長解析における再現率と冗長率の関係
Fig. 4 Relationship between recall and coverage.

$$\text{被覆率} = \frac{\text{出力した係り受けの数}}{\text{総係り受け数}}$$

主辞共起確率モデルとして LEX モデル，距離属性として dst2 を使用し，前節で提案した 3 つの尺度それぞれについて部分解析を行った．

図 3 は，それぞれの部分解析手法の閾値を変えたときの，係り受け出力の総係り受け数に対する被覆率と係り受け精度（適合率）の関係を表す．横軸が被覆率で，縦軸が適合率を表す．

図を見ると，いずれの尺度を用いても，被覆率を犠牲にすることで適合率が上昇しており，統計的部分解析手法が有効に働いていることが分かる．

それぞれの尺度を比べると，**Local/norm** と，**Ratio/next** は，ほぼ同じ適合率の振舞いを見せている．**Global** も，それほど違わない適合率を示しているが，被覆率が 0.4 から 0.9 の間では若干精度が落ちていて，被覆率がその範囲より高いか低いときには，逆に精度が若干上がっている．また **Global** を使用した場合，被覆率が 0.28 以下の部分がない．これは，上位 50 解のなかに，1 種類の係り先しか出現しなかった文節がそれだけあったことを意味している．その文節の係り受けについては確信度は最高の 1 となるので，閾値をいくら厳しくしてもそれ以上被覆率が下らない．いいかえると精度は上がらない．

解析速度を考えると，**Local/norm** と **Ratio/next** は最適解のみ計算すればよいのに対し．**Global** は上位 50 解を計算する必要上，解析速度が低下する☆．**Global** において，解析候補数 n の値を 100 にした実験も行ったが，それほど精度は向上しないうえ（被覆率 0.55 のときの差が最大で，約 0.5%程度），解析速度が低下する．したがって，少なくとも我々のモデルに関しては，**Local/norm** あるいは **Ratio/next** を使用するのが適当であるといえる．

☆ 1000 文を解析した結果，**Global** を使用した場合 41 分 17 秒，**Ratio/next** で 7 分 13 秒，**Local/norm** で 6 分 54 秒の時間がかかった．計算機のスペックは，cpu UltraSPARC, memory 512 M.

6. 冗長解析手法

冗長解析とは，部分解析の場合と逆で，1 つの文節に対し複数の係り受け関係の出力を許すことによって，高い適合率を得る手法である．

部分解析に使用した同じ尺度を，冗長解析に対しても使用することができる．その場合の手順は，以下のようになる．

(1) 各係り受け関係の，信頼値（部分解析の章で定義）を計算する．
(2) 適当に設定した閾値を基に，信頼値が閾値以上の係り受け関係をすべて出力する．

6.1 解析精度の評価

冗長解析アルゴリズムでは，出力される係り受けは，文節総数に対して一定ではない．よって，以下に定義する適合率と冗長率で，精度を評価する．

$$\text{再現率} = \frac{\text{出力した係り受けのうち正解した数}}{\text{総係り受け数}}$$

$$\text{冗長率} = \frac{\text{出力した係り受けの数}}{\text{総係り受け数}}$$

式のうえでは，冗長率は被覆率と同様に定義されるが，通常冗長解析における出力係り受け数は，総係り受け数より多くなるため，ここでは冗長率というように名前を変えた．

5 章で提案した尺度を使用し，閾値を変化させて，冗長率と再現率の関係について調べた．**図 4** に結果を示す．

図 4 の横軸は，冗長解析における冗長率を表し，縦軸は再現率を表す．

いずれの尺度を使用した場合も，閾値を適当に変えることで，適合率を犠牲にして，再現率を高めることができているのが分かる．

各尺度による解析精度を比べると，**Global** は，**Ratio/next** との再現率の差は 1%ほどであるが，最も精度が高く，**Local/norm** は，**Global** や **Ratio/next** に比べて著しく精度が低い（冗長率が高くない部分において）．

解析速度を考えると，5.1 節でも述べたのと同じ理由で，**Global** は **Ratio/next** に比べて遅い．

また **Global** を使用した場合，再現率が 99%以上になっていないことが分かる．これは，上位 50 解の中に正解係り受けが含まれていない場合，再現率 100%を達成することはできないからである．ここで，解析候補数を 50 より増すことはあまり意味がない．なぜなら，4.4 節でも述べたように，文節数が増すと解析候補数は指数関数的に増えるため，解析候補数を 50 から 100 にしたところで，達成可能な冗長率の上限はそれほど変わらない☆．実際，確信度の閾値 0 にしたときの，$n = 50$ のときと $n = 100$ のときの再現率の差を調べたところ 0.6%ほどの差しかなく，速度が重視される場面では有効とはいえない．

したがって，少なくとも我々の統計モデルを使用して冗長解析を行う場合，**Ratio/next** を使用するか，必要とされる再現率に応じて **Global** と **Ratio/next** を使い分けるのが適当であるといえる．たとえば，速度が重視されない場合は，再現率 99%以上必要とするのでなければ，**Global** を使うのが適当であろう．

7. おわりに

語の共起確率と距離確率による統計的係り受け手法について提案した．共起確率を使用して統計的に係り受け解析を行う場合，語の共起が有効であることを示した．語彙情報を使用した場合，文節単位の係り受け精度でおよそ 87%となった．ただし品詞の共起を利用した場合とそれほど解析精度の差がないため（0.4%程度），使用できる資源（データのサイズ）に限りがある場合は，品詞モデルで十分と思われる．

同様の情報を使用し，英語係り受け共起確率を推定した Collins[2] のモデルについても，我々の枠組みで Collins のモデルに相当するモデルを作成し実験を行ったが，日本語係り受け解析については我々のモデルによる解析精度の方が高かった．

☆ 1 万文を解析した場合，$n = 50$ のとき，2.55356 に対し，$n = 100$ のとき，2.94996 であった．

またより高い解析精度を実現するため，統計モデルを使用した部分解析手法と冗長解析手法を提案し，評価を行った．部分解析とは，再現率を犠牲にして適合率を上げる手法であり，冗長解析とは，適合率を犠牲にして再現率を上げる手法である．

日本語において，統計的モデルを使用した部分解析を行った研究として，乾ら[5] の研究がある．我々は，彼らの使用した “確信度” の尺度以外に 2 つの尺度を使用して部分解析を行い，解析精度の評価を行った．その結果，我々の統計モデルにおいても部分解析手法が有効に働くことを示すとともに．我々の統計モデルの下では，“確信度” を使用した手法（本論文中では，**Global** と呼んでいる）よりも，**Local/norm** と **Ratio/next** による手法の方が，大部分の被覆率において精度が良いことも分かった．解析速度も，**Local/norm** および **Ratio/next** を使用する場合が速いことも分かったので，少なくとも我々のモデルにおいては，**Local/norm** あるいは **Ratio/next** を使用する方が適当であるといえる．

冗長解析についても，部分解析に使用した 3 つの尺度を使用して，解析精度の評価を行った．その結果，冗長解析においてもこの 3 つの尺度は有効に働くことが分かった．また “確信度” を利用したもの（**Global**）の精度が最も高く，約 1%の差で，係り受け確率値の比を利用したもの（**Ratio/next**）がそれに続いた．ただし **Global** の場合，解析候補の中に正解係り受けが存在せず再現率に限界が出ること，また部分解析で述べたのと同様の理由で解析速度が **Ratio/next** に比べて著しく遅いことから，必要とする適合率に応じて両者を使い分けるのが適当であるということも分かった．

本論文で示した部分解析，冗長解析についての結果は，我々の統計モデルを使用した場合の結果である．他の統計モデルを使用して冗長解析を行った場合にもいえるかどうかは，別に検証する必要がある．

部分解析，冗長解析の精度をさらに向上させるためには，使用する統計モデルそのものの解析精度を向上させることも重要である．

各関係名ごとの精度の評価でも明らかになったように，出現頻度が少く，精度の悪い関係名については，関係名に語彙情報を使用するのをやめて学習データが過疎になるのを防ぐ必要がある．逆に頻度が多くて精度の悪いものに関しては，より詳細な情報を使用する必要がある．前者については，関係名の決定ルールをより詳細にし，特定の付属語（列）ごとに，関係名の指定方法を変えられるようにする予定である．後者に

ついては，決定木や決定リストなど，属性を自動選択する方法に何らかの工夫を加えることで実現しようと考えている．また，2.1 節でも少し触れたが，文節間属性として，文節間の格助詞の数や述語の数を使用した場合は，精度が良くなかった．これは，これらの属性は有効ではないというより，使用の仕方に問題があると思われる．これらの属性は，係り受けの曖昧性解消への有効性は関係名ごとに異なると考えられ，各関係名ごとに使用する文節間属性を設定する必要があると考えられる．

参 考 文 献

1) Charniak, E.: Statistical Parsing with a Context-free Grammar and Word Statistics, *Proc. 14th AAAI*, pp.598–603 (1997).
2) Collins, M.J.: A New statistical Parser Based on Bigram Lexical Dependencies, *Proc. 34th Annual Meeting of the Association for Computational Linguistics*, pp.184–191 (1996).
3) Haruno, M., Shirai, S. and Oyama, Y.: Using Decision Trees to Construct a Practical Parser, *Proc. 17th COLING and the 36th Annual Meeting of ACL*, pp.505–511 (1998).
4) Hogenhout, W.R. and Matsumoto, Y.: *Training Stochastical Grammars on Semantical Categories, Connectionist, Statistical and Symbolic Approaches to Learning for Natural Language Processing*, Wermter, S., Riloff, E. and Scheler, G. (Eds.), pp.160–172, Springer-Verlag (1996).
5) 乾健太郎，白井清昭，田中穂積，徳永健伸：統計に基づく部分係り受け解析，言語処理学会第 4 回年次大会 (1998).
6) 国立国語研究所：「分類語彙表」形式による語彙分類表（増補版）(1996).
7) Lafferty, J., Sleator, D. and Temperley, D.: Grammatical Trigrams: A Probabilistic Model of Link Grammar, *Proc. the AAAI Conference on Probabilistic Approaches to Natural Language*, pp.89–97 (1992).
8) Li, H.: A Probabilistic Disambiguation Method based on Psycholinguistic Principles, *Proc. 4th Workshop on Very Large Corpora*, pp.141–154 (1996).
9) Magerman, D.M.: Statistical Decision-Tree Models for Parsing, *Proc. 33rd Annual Meeting of ACL*, pp.276–283 (1995).
10) Matsumoto, Y., Kitauchi, A., Yamashita, T. and Hirano, Y.: Japanese Morphological Analyzer ChaSen2.0 Users Manual, Information Science Technical Report NAIST-IS-TR99009, Nara Institute of Science and Technology (1999).
11) 南不二男：現代日本語文法の輪郭，大修館書店 (1993).
12) Schabes, Y.: Stochastic Lexicalized Tree-Adjoining Grammars, *Proc. 14th COLING*, pp.425–432 (1992).
13) Shirai, K., Inui, K., Tanaka, H. and Tokunaga, T.: An emprical study on statistical disambiguation of Japanese dependency structure using a lexically sensitive language model., *Proc. 4th Natural Language Processing Pacific Rim Symposium*, pp.215–220 (1997).
14) Shirai, K., Inui, K., Tokunaga, T. and Tanaka, H.: An Emprical Evaluation on Statistical parsing of Japanese Sentences using Lexical Association Statistics., *Proc. 3rd Conference on Empirical Methods in Natural Language Processing*, Granada, Spain, pp.80–87 (1998).
15) 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間の係り受け解析の方法とその精度，情報処理学会論文誌，Vol.36, No.10, pp.2353–2361 (1995).
16) 宇津呂武仁，藤尾正和，西岡山滋之，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文係り受け解析における評価，言語処理学会第 5 回年次大会併設ワークショップ「構文解析—現状の分析と今後の展望」，pp.79–86，言語処理学会 (1999).
17) 日本電子化辞書研究所：EDR 電子化辞書仕様説明書 (1995).

(平成 10 年 10 月 15 日受付)
(平成 11 年 10 月 7 日採録)

藤尾 正和

1972 年生．1995 年京都大学理学部生物学科卒業．同年奈良先端科学技術大学院大学情報科学研究科博士前期課程入学．専門は自然言語処理．統計的手法，機械学習等に興味を持つ．

松本 裕治（正会員）

1955年生．1977年京都大学工学部情報工学科卒．1979年同大学大学院工学研究科修士課程情報工学専攻修了．同年電子技術総合研究所入所．1984～85年英国インペリアルカレッジ客員研究員．1985～87年（財）新世代コンピュータ技術開発機構に出向．京都大学助教授を経て，1993年より奈良先端科学技術大学院大学教授，現在に至る．工学博士．専門は自然言語処理．人工知能学会，日本ソフトウェア科学会，言語処理学会，認知科学会，AAAI，ACL，ACM各会員．