統計に基づく部分係り受け解析

乾健太郎　白井清昭　田中穂積　徳永健伸
東京工業大学大学院情報理工学研究科
{inui,kshirai,tanaka,take}@cs.titech.ac.jp

1 はじめに

近年，大規模コーパスの利用によって構文解析の精度を向上させる試みが盛んに行われており，報告されている性能も徐々に上がってきている．たとえば，Wall Street Journal による実験では labelled precision が 86%を越えたと報告されており [1]，日本語でも新聞記事を対象とした係り受け解析の実験で 85%～90%の精度が得られたという報告がある [5, 4]．構文解析で十分に高い精度が得られれば，格フレームなどの学習に必要な共起データをプレーンテキストから自動的にかつ大量に抽出することができるようになり，言語解析システムのブートストラップ的な洗練を実現する基礎を与えることができる．また，情報抽出や要約などの応用的タスクにおいても，高精度な構文解析は理解に基づく手法の開発を進める際に不可欠な要素技術である．しかしながら，解析精度が 90% 前後に留まっている現状では，現在の構文（係り受け）解析技術がこれらの需要に十分に応えているとは言いがたい．その一方で，構文解析という作業が実際には意味解釈や省略補完などの作業と不可分であることを考えると，現在盛んに研究されている統計的構文解析技術が解析精度を飛躍的に向上させることも期待しにくい．

このような背景から本稿では統計的部分係り受け解析方式を提案し，予備実験の結果を報告する．本方式では、確率言語モデルに基づいて順序づけされた文全体の係り受け構造の上位 n 個の候補から個々の部分的な係り受け関係の確信度を計算し，十分に高い確信度をもつ係り受け関係を選択することにより部分解析を実現する．本方式は次のような特長をもつ．

- パーザが持っている（語彙的・構文的）統計情報だけでは解消が困難な（語彙的・構文的）曖昧性については判断を保留し，曖昧性を部分的に解消する．
- パーザが行う部分的な曖昧性解消の精度をパーザの用途に応じて任意の高さに設定することができる．

部分構文解析という考え方は古くからあり，たとえば Jensen らの PEG パーザ [3] は構文的制約によって確実に決定できる部分構文構造だけを出力し，それ以外の部分の構文的曖昧性を保留することができる．しかしながら，構文的制約だけで解消できる構文的曖昧性は必ずしも多くないので，共起データの抽出という用途においても，情報抽出や要約などへの応用においても十分な貢献を期待するのは難しい．これに対し，本稿で提案する統計的部分係り受け解析では，統計情報を利用することにより，わずかな解析誤りのリスクを許容するだけで，規則ベースの部分構文解析に比べて判断を保留せざるをえない構文的曖昧性を大幅に減らすことができる．

表 1: 例文 (1) の解析結果

係り		b_1	b_2	b_3	b_4	b_5	b_6	b_7	$P(R_i)$
受け	R_1	b_8	b_8	b_4	b_5	b_8	b_7	b_8	5.11e-31
	R_2	b_5	b_3	b_4	b_5	b_8	b_7	b_8	1.32e-31
	R_2	b_5	b_4	b_4	b_5	b_8	b_7	b_8	1.01e-31
	R_4	b_5	b_5	b_4	b_5	b_8	b_7	b_8	7.36e-32
	R_5	b_2	b_8	b_4	b_5	b_8	b_7	b_8	7.35e-32
	R_6	b_8	b_8	b_4	b_5	b_6	b_7	b_8	9.76e-33
	R_7	b_8	b_8	b_4	b_5	b_7	b_7	b_8	8.95e-33
	R_8	b_2	b_3	b_4	b_5	b_8	b_7	b_8	4.71e-33
	R_9	b_2	b_4	b_4	b_5	b_8	b_7	b_8	3.61e-33
	R_{10}	b_2	b_5	b_4	b_5	b_8	b_7	b_8	3.48e-33

2 統計的部分係り受け解析

次のような文節切りされた品詞タグつき入力文に対し，その係り受け構造を決定するタスクを考える．

(1) [政界にも]$_1$ [二十代、]$_2$ [三十代の]$_3$ [若者が]$_4$ [飛び込み]$_5$ [「戦後政治」の]$_6$ [幕が]$_7$ [上がりました。]$_8$

i 番目の文節を b_i とすると，この文の正解の係り受け構造は，

係り	b_1	b_2	b_3	b_4	b_5	b_6	b_7
受け	b_5	b_3	b_4	b_5	b_8	b_7	b_8

のような係り文節から受け文節への写像関係として表現できる．この入力文に対し，たとえば，我々が開発中の統計的係り受け解析システム [4] は表 1 のような解析結果を出力する．ここで，R_i は終端記号列を入力文とする係り受け構造の i 番目の候補である．係り受け構造の候補は統計的係り受け解析システムが持つ確率言語モデル $P(R_i)$ によってランキングされている[1]．

係り受け解析（構文解析）システムを評価する際には文節ベースの係り先正解率やラベルつき再現率/適合率 (labelled recall/precision) など，1 位の候補と正解の重なりの割合を定量化するのが一般的である．文節ベースの係り先正解率によると，上の例では 1 番目の文節と

[1]実装上また効率性の都合で，すべての候補に共通する周辺分布は $P(R_i)$ の計算に含まれていないので，表 1 の $P(R_i)$ は正確な確率を表しているわけではない．

2番目の文節の係り先が誤りであり，他の文節の係り先は正しいので，文節ベースの正解率は $4/6 = 67\%$ と計算される．以下，これを部分係り受け解析に対比させて総係り受け解析と呼ぶ，総係り受け解析では1位の候補だけが評価の対象となるので，研究者の注意は1位の候補に集中しがちになるが，表1のように2位以下の候補も一緒に横に並べてみると，より多くの情報がそこから引き出せることがわかる．たとえば，b_1 の係り先の候補を見ると，b_2，b_5，b_8 の3つが有力で，いわばシステムがその判断に「迷っている」と言うことができる．これに対し，b_3 の係り先については，上位10位までの候補がいずれも b_4 で一致しており，システムがその判断に「自信を持っている」ことがわかる．このように，上位 n 位の候補を横並びで見ると，各文節の係り先の候補についてシステムがどの程度確信をもっているかを知ることができる．この「確信度」と呼べるような量をうまく見積ることができれば，確信度の高い係り受け関係だけを選択的に決定し，残りの部分の判断を保留することができるようになる．

このことは以下のように定式化できる．ある確率言語モデルが生成する係り受け構造の集合を $\mathcal{R}$ とし，その確率分布を $P : \mathcal{R} \mapsto [0,1]$ $\left(\sum_{R\in\mathcal{R}} P(R) = 1\right)$ とする．さらに，ある係り受け構造 $R \in \mathcal{R}$ の文節 b_i が文節 b_j に係ることを式 $R \models r(b_i, b_j)$，R の終端記号列が文 s であることを式 $R \models s$ で表すことにする．このとき，「入力文 s が係り受け関係 $r(b_i, b_j)$ をもつ」という命題に対する確信度を確率 $P(r(b_i,b_j)|s) = P(r|s) = \frac{P(s,r)}{P(s)}$ として定量化することにする（以下，誤解の恐れのない場合，$r(b_i, b_j)$ を r と略記する）．$P(r|s)$ は以下のように近似推定できる．まず，$\mathcal{R} = \mathcal{R}_H + \mathcal{R}_L$ を満たす任意の交わりをもたない2つの集合 $\mathcal{R}_H$，$\mathcal{R}_L$ について次式が成り立つ

$$P(s) = \sum_{R\in\mathcal{R}:R\models s} P(R) \quad (1)$$

$$= \sum_{R\in\mathcal{R}_H:R\models s} P(R) + \sum_{R\in\mathcal{R}_L:R\models s} P(R) \quad (2)$$

同様に，

$$P(s,r) = \sum_{R\in\mathcal{R}_H:R\models s\wedge r} P(R) + \sum_{R\in\mathcal{R}_L:R\models s\wedge r} P(R) \quad (3)$$

が成り立つので，$P(r|s)$ は次式の範囲内に抑えられる．

$$\frac{P^{s\wedge r}_{\mathcal{R}_H}}{P^{s}_{\mathcal{R}_H} + P^{s}_{\mathcal{R}_L}} \le P(r|s) \le \frac{P^{s\wedge r}_{\mathcal{R}_H} + P^{s}_{\mathcal{R}_L}}{P^{s}_{\mathcal{R}_H} + P^{s}_{\mathcal{R}_L}} \quad (4)$$

ただし，

$$P^{s}_{\mathcal{R}_H} = \sum_{R\in\mathcal{R}_H:R\models s} P(R) \quad (5)$$

表2: 表1から計算される係り受け確率と確率最大の係り受け構造

係り	b_1	b_2	b_3	b_4	b_5	b_6	b_7
受け正解	b_5	b_3	b_4	b_5	b_8	b_7	b_8
R^*	b_8	b_8	b_4	b_5	b_8	b_7	b_8
$P(r\|s)$	.57	.65	1.00	1.00	.98	1.00	1.00

$$P^{s\wedge r}_{\mathcal{R}_H} = \sum_{R\in\mathcal{R}_H:R\models s\wedge r} P(R) \quad (6)$$

$$P^{s}_{\mathcal{R}_L} = \sum_{R\in\mathcal{R}_L:R\models s} P(R) \quad (7)$$

式 (4) より，$P(r|s)$ を式 (8) で近似した場合の誤差 ϵ は高々 $\epsilon \le \frac{P^{s}_{\mathcal{R}_L}}{P^{s}_{\mathcal{R}_H} + P^{s}_{\mathcal{R}_L}}$ で抑えられる．

$$P(r|s) \approx \frac{P^{s\wedge r}_{\mathcal{R}_H}}{P^{s}_{\mathcal{R}_H}} \quad (8)$$

$P(r|s)$ の推定に $R \models s$ でない R が無関係であることは (4) からも明らかであるので，$R \models s$ を満たす R のうち確率の高い上位 n 個の集合をあらためて $\mathcal{R}_H$，$R \models s$ を満たす R のうち残りの確率の低いもの集合 $\mathcal{R}_L$ とすると，確率分布 P の偏りが大きい場合は，適当な大きさの n について $P^{s}_{\mathcal{R}_H} \gg P^{s}_{\mathcal{R}_L}$ が成り立つと期待できるので，誤差 ϵ を無視できる大きさに抑えることができる．式 (8) によって得られる係り受け関係 r の確率を以下では r の係り受け確率，あるいは簡単に r の確率と呼ぶ．

(8) にしたがって表1の例を解析すると，以下のようになる．R^* は各文節について係り受け確率を最大にする係り先を選択することによって得られる係り受け構造である．すなわち，

$$R^* \models r(b_i, b_j) \Leftrightarrow b_j = \arg\max_{b_j} P(r(b_i, b_j)|s) \quad (9)$$

$P(r|s)$ は R^* の個々の係り受け関係の確率である．R^* では，係り先確率の高い文節 b_3，b_4，b_5，b_6 の係り先はいずれも正解している．一方，文節 b_1 と b_2 の係り先が誤っているが，これらの係り受け確率はいずれも低いので，実際にはこれらの文節の係り先は判断が保留される．

3 予備実験

3.1 セッティング

京大コーパス [6] から無作為に抽出した4,800文のうち文節数が10以内の文，3,065文（13,692文節[2]）を対象に予備実験を行った．パーザへの入力は品詞タグつきの単語列である．実験には以下の3種類の確率言語モデルを用いた．

[2]文末の2文節は係り先が必ず一意に特定されるので評価の対象には含めない．

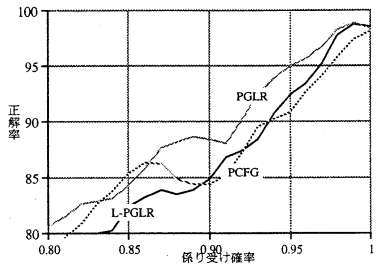

図 1: 各モデルの P-A 曲線

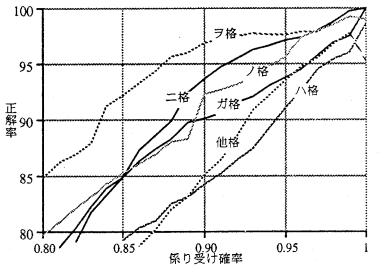

図 2: L-PGLR モデルにおける各文節種の P-A 曲線

L-PGLR モデル: 文献 [4] で提案したモデルで，確率 GLR (PGLR) モデル [2]（統語モデル）と単語間の確率的従属関係を表す統計量（語彙モデル）の組み合せからなる．PGLR モデルには終端記号数 59，非終端記号数 78，規則数 1312 の CFG を用いた．各終端記号は文節に対応するカテゴリを表し，文節の係り属性，受け属性，読点の有無などの素性の束として構成されている．文節の境界，文節の素性は入力に付加された品詞列から一意に定まる（詳細は文献 [7] を参照）．PGLR モデルの訓練には京大コーパスからテストセットを除いた約 15,000 文を用いた．語彙モデルは RWC コーパス・EDR 共起辞書から抽出した [名詞-格助詞-動詞] の共起事例（のべ 786 万）と EDR コーパスから抽出した [格助詞ベクタ-動詞] の共起事例（のべ 46 万）を用いて学習した．

PGLR モデル: 上述の PGLR モデル．単語間の確率的従属関係を無視したモデルである．

PCFG モデル: PGLR モデルに用いたものと同じ CFG から得られる確率文脈自由文法．訓練にも上述の PGLR モデルの訓練データと同じものを用いた．

3.2 P-A 曲線

表 2 の R^* と同様に，n-best の候補（以下の実験は $n = 50$ として行った）から各係り受け関係の係り受け確率を計算し，個々の文節について最大の係り受け確率を持つ係り先を特定する．このとき，係り受け確率と係り受けの正解率の相関関係を調べると図 1 のような確率-正解率曲線が得られた．以下，これを P-A 曲線と呼ぶ．

図 1 は，モデルの種類に関わらず，係り受け確率が高いほどその係り受け関係の正解率が高くなるという傾向をはっきりと示している．このことは，係り受け関係に対する確信度の尺度として係り受け確率を用いることが妥当であることを示唆している．さらに興味深いのは，いずれのモデルを用いた場合でも，係り受け関係の正解率が係り受け確率とだいたい一致している点である．この傾向は，図 2 のように係り文節の種類ごとに P-A 曲線をプロットした場合にもある程度見られる[3]．このことは，実験で使用した訓練データの量がこれらのモデルの複雑さに対してある程度十分であったことを示唆している．ただし，十分な訓練データがあっても，モデルが生成する係り受け構造の確率分布が真の分布に十分に近づくとは限らない．このことは次の C-A 曲線をプロットすることで明らかになる．

3.3 C-A 曲線

係り受け確率の閾値を σ として，σ 以上の確率をもつ係り受け関係だけを選択的に決定する作業を考える [4]．σ を $.5 \leq \sigma \leq 1$ の範囲で変化させると，図 3 のような被覆率-正解率曲線が得られた．以下，これを C-A 曲線と呼ぶ．ただし，被覆率，正解率は次式で与えられる．

$$\text{被覆率} = \frac{\text{係り先が決定された文節の数}}{\text{テストセット中の文節の数}} \quad (10)$$

$$\text{正解率} = \frac{\text{係り先が正解である文節の数}}{\text{係り先が決定された文節の数}} \quad (11)$$

図 3 は，確率の高い係り受け関係だけを選択すると文節ベースの正解率が上がることを示している．たとえば，L-PGLR モデルの場合，総係り受け解析の正解率は 84.5% に過ぎないが，被覆率を 75% に抑えるだけで正解率は 94% に上がる．逆に，2% のノイズ（正解率 98%）を許容するだけで全体の 50% の係り受け関係を抽出することができ，被覆率はある程度犠牲にしても高い精度を必要とする共起データ抽出作業のような場合に統計的部分解析が有効に働くことが期待できる．従来の総係り受け解析とは異なり，解析システムの

[3]図 2 は，格/係助詞を含む係り文節を対象にプロットした P-A 曲線である．「ハ格」は係助詞「ハ」を末尾に持つ文節，「他格」は「ハ，ガ，ヲ，ニ，ノ」以外の格/係助詞を末尾に持つ文節を表す．

[4]厳密に言えば，複数の文節の係り受け関係を同時に選択する場合はそれらの係り受け関係の結合確率を計算するべきである．しかしながら，実際には確率が 1 に近い係り受け関係しか選択されないので，同時に選択された 2 つの係り受け関係に強い負の従属関係があるということは起こりえない．すなわち，確率が 1 に近い係り受け関係だけを選択する限り，その判断の基準としては個々の係り受け関係の確率を参照だけで十分であると考えられる．

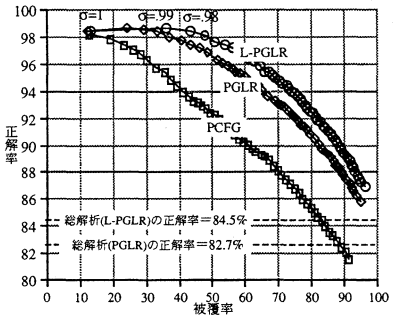

図 3: 各モデルの C-A 曲線

アプリケーションに応じて被覆率と正解率のどちらを重視するかをユーザが自由に選択できる点が重要である．また，$\sigma = 1$ のときの被覆率が 13% に過ぎない点にも注意したい．このことは，構文的制約で一意に決まる係り受け関係だけを選択的に出力する規則ベースの部分係り受け解析では大部分の文節の係り先が特定されないことを意味している．これに対し，統計的部分解析では，$\sigma = .99$，$\sigma = .98$ などの点を見てもわかるように，わずかなリスクを負うことによって判断を保留せざるをえない係り受け関係を大幅に減らすことができる．

3.4 言語モデルの評価尺度としての C-A 曲線

モデル間の優劣は C-A 曲線に顕著に現れる．従来の総係り受け解析では被覆率を 100% に固定し，正解率のみでモデルの性能を評価していたが，部分的係り受け解析では被覆率と正解率の組み合わせによってモデルを評価することになる．被覆率が同じなら正解率が高い方が望ましいし，その逆も同様である．たとえば，L-PGLR モデルと PGLR モデルを比較する場合，総係り受け解析の正解率はそれぞれ 84.5%，82.7% で，PGLR を L-PGLR に換えたときの誤り削減率は 10% に過ぎない．しかしながら，図 3 の C-A 曲線を見ると，被覆率が 75%，50% のときの誤り削減率はそれぞれ 23%，36% である．このように，総係り受け解析の正解率だけではモデルを過小評価，あるいは過大評価してしまう危険性があるが，C-A 曲線を調べることによって従来よりも多面的な評価が可能になる．

3.5 係り受け確率の近似推定の精度

2節で述べたように，係り受け確率を式 (8) で近似推定する場合，n-best の確率和 ($P^s_{R_H}$) が全候補の確率和 (P^s_R) に十分に近くなるように n を決める必要がある．図 3 は $n = 50$ として実験した結果だが，これと $n = 10$ の場合を比較した結果を 図 4 に示す．このグラフから，モデルによって n-best の確率和の収束の速さに差があることがわかる．たとえば，PCFG モデルは 3 つのモデルの中で最も単純なモデルであり，他のモデルに比べ確率分布に大きな偏りが生じにくいため，確率和の収束が最も遅い．n が大きくなると，解析途中のビーム幅なども広くする必要があり，一般に計算コストが高くなる．

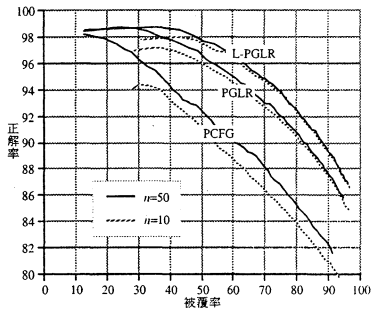

図 4: $n = 10$ と $n = 50$ の比較

4 おわりに

本稿では統計的部分係り受け解析の一手法を提案し，予備実験の結果を報告した．本手法は，統計的手法に基づいて選択された上位 n 個の候補による重みつき多数決によって部分解析の確信度を評価するというもので，言語モデルに依存しない汎用的な枠組である．また，本稿では問題を係り受け解析に限って議論したが，形態素解析を含む問題にも容易に拡張できる．今後は，実験を大規模化するとともに，確信度が高いにもかかわらず解析結果が誤りである例や確信度が低い係り受け関係の例を詳細に調べ，言語モデルの洗練をはかる予定である．

参考文献

[1] E. Charniak. Statistical parsing with a context-free grammar and word statistics. In *Proceedings of the 15th National Conference on Artificial Intelligence*, 1997.

[2] K. Inui, V. Sornlartlamvanich, H. Tanaka, and T. Tokunaga. A new formalization of probabilistic LR parsing. In *Proceedings of the 5th International Workshop on Parsing Technologies*, pp. 123–134, 1997. Available from `http://www.cs.titech.ac.jp/tr.html`.

[3] K Jensen, G. E. Heidorn, and S. D. Richardson, editors. *NATURAL LANGUAGE PROCESSING: The PLNLP Approach*. ©KAP, 1993.

[4] K. Shirai, K. Inui, H. Tanaka, and T. Tokunaga. An empirical study on statistical disambiguation of japanese dependency structures using a lexically sensitive language model. In *Proceedings of Natural Language Pacific-Rim Symposium*, pp. 215–220, 1997.

[5] 黒橋禎夫, 長尾眞. 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, Vol. 1, No. 1, pp. 35–57, 1994.

[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予稿集, pp. 58–61, 1997.

[7] 白井清昭. 統計情報を利用した統合的自然言語解析. 博士論文, 東京工業大学, 1998. `http://www.cs.titech.ac.jp/tr.html`.