20－3

統計知識と文脈情報を用いた一般化ＬＲ構文解析法の研究

Disambiguating Sentence by Using Generalized LR Parser with Statistical and Discourse Knowledge

中島 浩之* 白井 清昭** 田中 穂積**
Hiroyuki Nakajima Kiyoaki Shirai Hozumi Tanaka
ＮＴＴデータ通信*
NTT DATA COMMUNICATIONS SYSTEMS
東京工業大学**
Tokyo Institute of Technology

Abstract

Probabilistic LR(PLR) is a context-dependent probabilistic language model. According to our experiment with EDR corpora, stochastic context-free grammar(SCFG) performs better than PLR to disambiguate the results of parsing. We pay attention to conflict-actions in LR table and propose a new probabilistic method "PLR-conflict" which computes probability only when conflict-actions happen. Furthermore, in order to use contextual effects,we propose Context-Dependent Grammar(CDG), and Stochastic CDG(SCDG) which take care of conflict-actions with modified LR parser. The result of our experiment shows that PLR-conflict and SCDG performs better than SCFG and PLR to make disambiguation.

1 序論

文脈自由文法による自然言語の構文解析は、一般に一つの文に対し多数の構文木を生成する。この多数の構文木のうち、正しい構文構造を持つ構文木は数少ない。そのため自然言語理解の研究の初期の段階から、正しい構文木の抽出、つまり構文解析結果の曖昧性の解消が重要な問題の一つとなっている。

曖昧性を解消し、正しい構文木を抽出するため、多くの研究が行なわれてきた。語の概念に関する知識を使い意味処理を行なうものなど、文脈自由文法に他の知識を付け加えることで構文木の評価を行ない、曖昧性を解消しようとするものなどである [9]。本研究で取り上げる手法は、文法に知識を付け加えるものとは異なり、構文解析の動作についての統計的知識を利用するものである。統計知識を用いる手法の場合、構文木の評価は意味処理に比べはるかに小さい計算量で行なうことができ、意味処理による評価と合わせて曖昧性の解消に効果を発揮することが期待されている。

2 既存の研究

2.1 確率 LR 構文解析法 (PLR)

統計文脈自由文法 (Stochastic CFG,SCFG) は文脈自由文法規則の重要度を確率によって表現するものである。確率はそれぞれの規則に対して非終端記号を展開する回数に従って与えられ、構文木の評価値には使用した規則に与えられている確率の積を用いる。この評価値の高いものほどもっともらしいとすることで、構文解析結果の曖昧性の解消を図る。 SCFG は CFG 規則に確率を割り当てるため、それぞれの規則が持つ確率は文脈に依存しないものになるが、自然言語では文脈により非終端記号を展開する規則の使用頻度が異なることが知られており、この非文脈依存性を解消するため、いろいろな提案が行なわれている [2, 4]。Briscoe らによって提案された確率 LR 構文解析法 (Probabilistic LR、PLR)[1] は一般化 LR 構文解析器 [6] の動作の左文脈依存性を用いて SCFG の非文脈依存性を解消を図るものである。

*連絡先：〒210 川崎市幸区堀川町奥和川崎西口ビル ＮＴＴデータ通信 技術開発本部 情報科学研究所 知能情報研究担当 Tel 044(548)4606
Email:nakajima@rd.nttdata.jp(本研究は東京工業大学情報工学専攻田中研究室に在籍中に実施した)

一般化LR構文解析器の解析動作はLR構文解析表において左文脈（正確には構文解析の履歴）を反映する状態番号と先読み語により決定される。このため一般化LR構文解析法の構文解析動作はSCFGとは異なり左文脈に依存したものになる。PLRはLR構文解析器のある状態番号でどの動作が行なわれるかに着目し、解析表の各動作に確率を割り当てる。

State	a	b	...
0	*sh* 1(0.3)	*sh* 2(0.7)	
1	*sh* 2(0.1)	*re* 2(0.5)	
	re 2(3)(0.2)		
	re 2(8)(0.2)		

構文木の評価値は使用した動作の持つ確率の積を用い、SCFGと同様、ある文の複数の構文木の中で、この評価値の高いものほどもっともらしいと考える。

2.2 予備実験

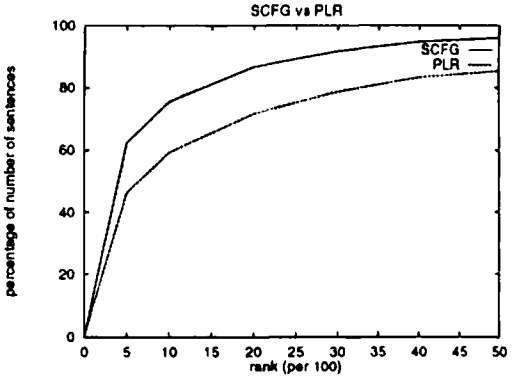
本研究では、EDR（日本電子化辞書研究所）の構文木つきコーパス[8]を用い、そこから抽出した日本語の文法を用いて実験を行なった。

本研究ではこのコーパス中の11474文（平均文長18.9）、およびコーパスと同じ構文木を生成可能な文脈自由文法（規則数109）を利用する。文法から構成されるLR(1)構文解析表は以下のようなものである（すべての競合動作がshift-reduce競合ないしreduce-reduce競合）。

状態数	184(終了状態含む)
reduceエントリ	1594個
shiftエントリ	968個
競合動作	676種1352個

コーパスの文をこの文法によって構文解析したところ、得られた平均構文木数は1378個であった。

実験では評価値の上位n%以内に正解が含まれている文の割合を求めた。（横軸が構文木の評価値の上位n%、縦軸が文の割合(%)）



PLRはSCFGに劣る結果を示している。LR構文解析表のreduce動作に確率を与えることでSCFGと等価な結果を獲得することができることは明らかであるので、PLRでの解析表への確率の与え方に問題があるといえる。そこで本研究ではLR表に対する確率の与え方を見直す。

3 PLR-conflict

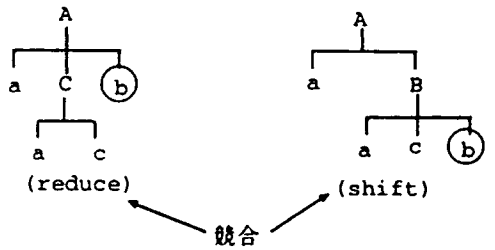
SCFGで規則に、PLRにおいて動作に確率を与えることで表現しようとしているのは、複数の候補動作の間でのpreferenceであると考えることができる。PLRではある状態からの動作の間でこのpreferenceを考えたが、辞書引きの曖昧性を考慮に入れない（先読み語が常に一つ）場合、動作の候補が複数あるのは競合動作を起こす場合のみである。そこで本研究では競合動作間のみに確率を与えることを提案する。

State	a	b	...
0	*sh* 1(1.0)	*sh* 2(1.0)	
1	*sh* 2(0.2)	*re* 2(1.0)	
	re 2(0.8)		

このようにLR構文解析表に確率を与えたものをPLR-conflictと呼ぶことにする。

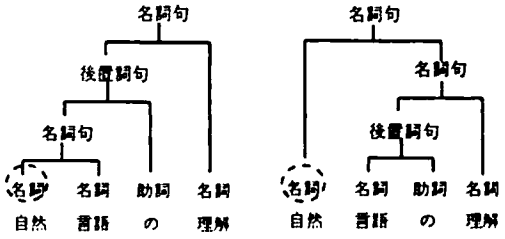
4 文脈従属文法による左文脈の細分化

PLRでは状態番号を利用することで文脈情報を利用しているが、本節ではさらに明確に文脈情報を利用することを考える。文法規則 $A \rightarrow aB$、$A \rightarrow aCb$、$B \rightarrow acb$、$C \rightarrow ac$ で、文"aacb"の"aac"まで解析を終了したとき、LR構文解析器はshift-reduce競合を起こす。



この場合、reduce 動作は $A \rightarrow aCb$ と $C \rightarrow ac$ の共起を、shift 動作は $A \rightarrow aB$ と $B \rightarrow acb$ の共起を表す。

これまで行なわれてきた非文脈依存性を解決する研究は、その多くが上例のような規則、記号の構文木における"縦方向での共起関係"を扱うものであるが、縦方向の共起関係だけでは説明しにくい事象がある。下図の場合、左側の構文木の方がもっともらしいと考えられる。



これは"自然言語"のような名詞の並びを途中で分割して後置詞句が生成されることがあまりないためにそう判断されるが、このことを縦方向の共起関係で扱うことは困難であり、むしろ"名詞、名詞、助詞"の並びが名詞の間で分割されて構文木を生成することは稀である、名詞—後置詞句の横方向の共起が稀にしかおきない、と表現するべきであろう。このように、規則・記号の共起に縦方向だけでなく、"横方向の共起関係"を扱うことは有益である。ここでは横方向の文脈、特に連接する記号／規則間の共起関係の利用について考える。

横方向の共起関係は、LR 構文解析器においてはグラフ構造化スタックのスタックトップと左文脈との共起にあらわれる。ここでは構文解析器に左文脈の最後に来る終端記号 (これを"LAST[1]"と呼ぶ) を保持し、利用することを考える。

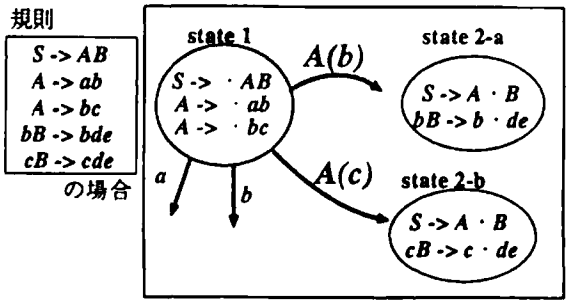
LR 構文解析表における状態番号は一般に複数の相異なる左文脈を同じ状態番号として扱っている。つまり LAST についての情報を状態番号は十分に反映してない。そこで LAST を状態番号に反映させるため、解析表の作成に非終端記号での遷移先を LAST によって区別し、構文解析時に LAST を参照するように LR 構文解析器を拡張することで左文脈の細分化を行なうことを考える。こうするとすべての LAST によって状態の分割を行なう場合、生成される解析表は膨大なものとなることが予想される。本研究ではどの場合に LAST による状態の分割を行なうか、LAST と CFG 規則の共起関係を文脈従属文法 (Context-Dependent Grammar,CDG[3]) で明示的に指定する。

まずあらかじめ CFG 規則の中で左連接終端記号によって使用頻度が異なると考えられるものを文脈従属文法規則で書き直し、LAST と CFG 規則の共起を記述する。

$$aA \rightarrow a\alpha$$

(a は LAST、A は非終端記号、α は終端／非終端記号列)
ここで a は LAST を表し、LAST が a である場合に限り CFG 規則 $A \rightarrow \alpha$ が適用可能であることを示す。

解析表の構成時、文脈従属規則が展開される場合は、文脈従属規則が展開されない場合と異なる状態に移ることになる。(構文解析時には LAST によって状態遷移先を決定する)



文脈従属規則が展開されない場合は、規則の使用頻度は LAST によって影響されないとし、状態の細分化を行なわない。このように文脈従属規則を状態の細分化の制限に用いることで、状態の不要な細分化を抑える。

以下、CDG を構文解析する拡張した LR 構文解析器を CDG 構文解析器と呼ぶ。

4.1 CDG 規則の作成

これまでの PLR、SCFG との比較を考え、これまで用いてきた CFG 規則と同じ言語を生成する CDG 規則を、CFG 規則を「文脈従属化」することで作成した。

1. 規則

$$A \rightarrow BC$$

に対して左文脈になり得るすべての終端記号 $p_1 \cdots p_n$ を獲得

[1]LAst Shifted Terminal の略

2. これら $p_1 \cdots p_n$ を用いて規則を

$$p_1 A \rightarrow p_1 BC$$
$$\vdots$$
$$p_n A \rightarrow p_n BC$$

と書き換える。

このようにして獲得した文法が元のCFG規則と同じ言語を生成することは明らかであろう。

本研究では先に上げた例、"自然言語の研究"で取り上げた"後置詞句"を展開する規則を文脈従属化した（文法規則数は109個から208個に増加）。文脈従属文法構文解析器においても、一般化LR構文解析法と同様、競合動作によってのみ曖昧性が生じる。そこでPLR-conflictと同様、競合動作のいずれが選択されるか、そのpreferenceを確率で表すことにする。

まず文脈従属文法規則に対して構文解析表を作成する（下表）。

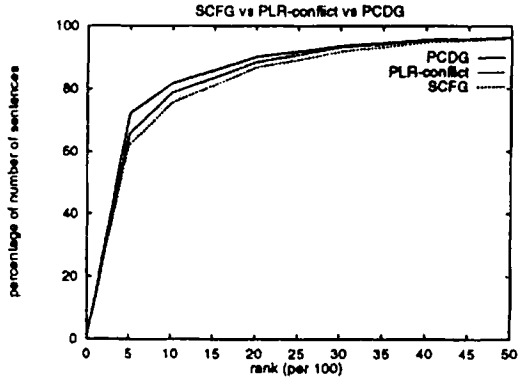
状態数	978(終了状態含む)
エントリ数	37818個
競合動作	7224種14448個

この解析表を用いて正しい構文木を作成するよう、CDG構文解析器を動作させ、正しい構文解析動作を獲得、競合動作間に確率を振り分けた。

得られた確率付き解析表を用いて解析を行ない、構文木には解析に使用した動作に振られた確率を掛け合わせた確率を評価値として与えた。こうして構文木に確率を与えた構文解析器を確率文脈従属解析器（Probabilistic CDG）と呼ぶ。

5 実験結果

PLR-conflict、PCDGの実験結果は下のグラフの通りである。



これまでのSCFG、PLRに比べ曖昧性の解消に大きな効果が認められる。複数の解析結果を生成する競合動作、並びに横方向の共起関係を利用することが有効であることを示しているといえよう。

6 将来の課題

今後の課題として、以下のものが挙げられる。

- 異なる言語、文法、コーパスでの実験：本研究で扱った言語は日本語のみ、文法は1種類のみであるので、異なる言語、文法、コーパスを用いて実験による検証を行なう必要がある。
- 辞書引きの曖昧性の取り扱い：一般化LR構文解析法上で形態素解析、構文解析を統一的に扱う手法であるMSLR [7]での単語間の連接可能性制約を文脈従属文法規則で記述することで、確率文脈従属解析器上で辞書引きの曖昧性と構文木の曖昧性を統一的に扱うことでより精度の良い結果が得られる可能性がある。
- 文脈従属文法構文解析器の拡張と音声認識への利用：確率文脈従属解析器の扱う文法規則を拡張することで[5]での音素環境クラスタの記述を行ない、音声認識に利用する。

参考文献

[1] Ted Briscoe and John Carroll. Generalized probablistic LR parsing of natural language(corpora) with unification-based grammars. *Computational Linguistics*, pp. 26–59, 1993.

[2] Mahesh Chitrao. *Statistical Technique for Parsing Message*. 博士論文, New York University, 1990.

[3] Robert F.Simmons and Yeong-Ho Yu. The acquisition and use of context-dependent grammars for english. *Computational Linguistics*, pp. 391–418, 1992.

[4] David M.Magerman and Mitchell P.Marcus. Pearl:a probablistic chart parser. In *EACL*, pp. 15–20, 1991.

[5] Akito NAGAI, Shigeki SAGAYAMA, Kenji KITA, and Hideaki KIKUCHI. Three different LR parsing algorithms for phoneme-context-dependent hmm-based continuous speech recognition. *IEICE TRANS. INF. & SYST.*, January 1993.

[6] Masaru Tomita and See-Kiong Ng. The generalized LR parsing algorithm. In *Generalized LR Parsing*, pp. 1–16. Kluwer Academic Publishers, 1991.

[7] 伴光昇. 形態素・統語解析の統合処理に関する研究. 修士論文, 東京工業大学, 1994.

[8] 日本電子化辞書研究所. ＥＤＲ電子化辞書仕様説明書, 3月１９９３年.

[9] 田中穂積. 自然言語理解, 知識工学講座, 8. オーム社, 昭和６３年.