正しい構文解析木による禁止パターンの学習

2E－4

黒岩　眞吾　　松本　一則　　榊　博史
国際電信電話株式会社　上福岡研究所

1．はじめに

一般に文法書に書かれている構文解析規則を文脈自由形解析規則としてインプリメントし、文章を解析した場合、多数のあいまいな構文解析木が生成されてしまう。これは、人手によって作成される文法が、注目している中心的な現象のみを説明し、その文法を含む全現象を説明していないことに起因すると考えられる。筆者らは、文法書の例文は、このあいまい性を減少させるための一つの指針として掲載されているとの考えに基づき、例文とその正しい構文解析木をもとに文法の精密化を続けてきた。本報告ではこれら例文と正しい構文解析木から、計算機が文法規則の精密化のためのデータを自動生成する手法について提案する。具体的には、本報告で用いるパーザのフィルタリングデータである禁止パターンを学習する。パーザーは拡張LINGOLにフィルタリング機構等の改良を加えた機械翻訳システムＫＡＴＥの解析処理部[1]を用いた。

2．禁止パターンを用いるパージング手法

禁止パターンは文脈自由形解析規則（以下書き換え規則と呼ぶ）の適用制限を行なうための知識として各ルールごと、および単語ごとに記されている。書き換え規則は一般にA→B1 B2 ・・・Bn（式1）のように表される。しかし、これらの規則の集合で自由に組み上げられた構文解析木は1節で述べた理由により、各書き換え規則が示す局所的部分で正しいに過ぎず、上記木のより広域な部分での正しさは保証されていない。そこで正しくない部分を除去するために、（式1）の適用の直後に得られた木について、適用された書き換え規則を含みしかもより広域な部分で正しくない部分があるかどうかを調べ、無ければ木の組立を継続し、あるならばその木全体を消去するという方法を用いる。この正しくない部分を禁止パターンとし、以下禁止木と呼ぶ。禁止木の一例を図1に示す。この禁止木は文法的には「関係節(WHRELCL)に後方より修飾された名詞句(NP)を更に後方より動詞過去分詞(EDP)により修飾することはできない」という禁止規則に対応するものである。

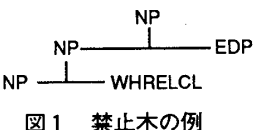


図1　禁止木の例

3．正しい解析木による学習

構文解析時の広範囲な構造の正しさを保証して、書き換え規則を作成することは人間にとっては困難なことである。一方、例文について正しい構文構造を選択することは比較的容易に行うことができる。以上の考えに基づき、例文とその正しい構文解析木（以下正解木と呼ぶ）を用い2節で述べた禁止木の自動生成を行ない、構文解析データとする。例文を現在の書き換え規則で解析し複数の解析木を得、同一例文で作成された正解木とそれ以外に生成される多くの構文解析木（以下不正解木と呼ぶ）を比較して、誤った部分構造を見つけ、それをもとに禁止木を構成し学習する。

4．禁止木の決定

不正解木と言ってもその構造のすべてが誤っているわけではない。ある文例の不正解木中の部分構造 x→x1　x2 ・・・ xn（図2）および各々のカテゴリー（NP等の非終端記号をカテゴリーとよぶ）のスパン（支配する単語列：図2ではwmからwm+iがx1に支配される単語列である）に着目したとき、同一の例文の正解木中にも同一の構造 x→x1　x2 ・・・ xn が存在して、かつ各々対応するカテゴリーの支配するスパンが等しい（図中T1~n部の構造は問わない）とき、部分構造 x→x1　x2 ・・・ xn は正しいと判定する。不正解木中の全ての部分構造について以上の判定を行い正しいと判定されなかった部分構造を誤構造とし禁止木の候補とする。このうち本報告では、隣接する誤構造はそれら全てをまとめて一つの禁止木とする。すなわち、得られた構造の内もっとも特殊なものを着目した不正解木による禁止木とする。このようにして得られた文章ごとの禁止木のうち、学習データ中の他の文章の正解木の生成を阻止するものは矛盾する禁止木と呼び、本報告においては無効な禁止木としパージングには用いない。例として

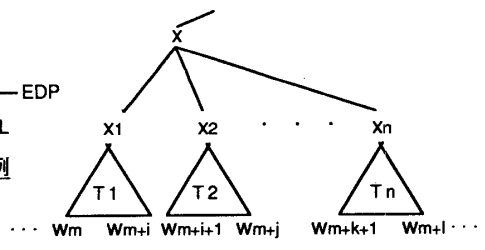


図2．不正解木中の部分構造の判定

Learning of Forbidden Pattern from Correct Syntactic Structure

Shingo Kuroiwa, Kazunori Matsumoto, Hiroshi Sakaki
KDD Kamifukuoka R & D Labs.

「The boy who has cars painted blue is smart.」の正解木と不正解木から生成された禁止木を図3に示す。

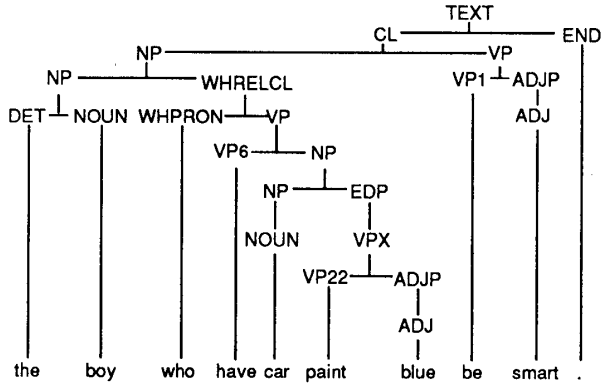


図3-a 正解木

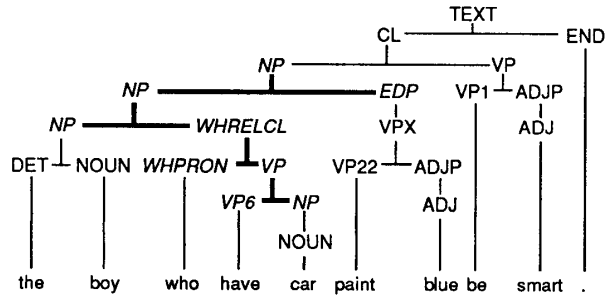


図3-b 不正解木

例文: The boy who has cars painted bluse is smart.

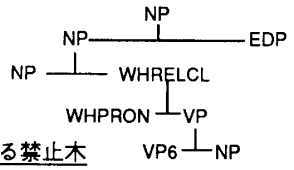


図3-c a,bより得られる禁止木

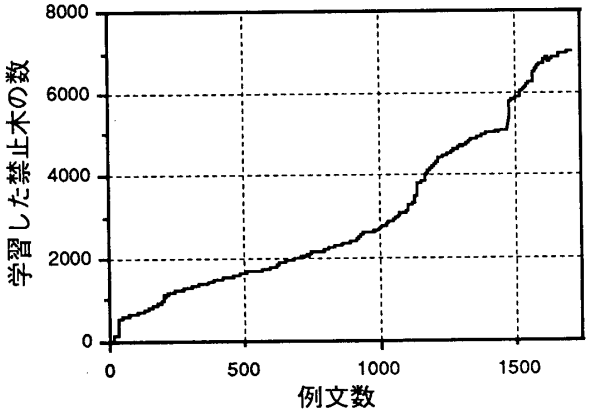
5. 実験および評価

禁止木の学習と、学習結果として得られた禁止木を用いたパージングの実験を行った。

5.1 正解木による禁止木学習実験

［実験および結果］例文を学習するに従って有効な禁止木の数がどのように変化するか測定を行なった。学習のためのデータとしは、本パーザーの書き換え規則作成の基礎となったホーンビーの文法書[2]から得た1697例文とそれらに対応する正解木を用いた。結果をグラフ1に示す。

グラフ1 例文数と禁止木の学習数



［考察］十分学習が進んだ段階で禁止木の増加がなくなることが予測されたが今回の学習例ではそのような傾向が現れなかった。またデータ中にいくつか不連続な点があるがこれはその不連続点の文が数千オーダーのあいまいな解析木を生成するためである。

5.2 学習した禁止木による解析実験

［実験と結果］5.1で得られた禁止木(7021本)を用いてパージング実験を行なった。対象としては学習に用いた文および英検4級テキストからの660例文を用いた。生成される解析木の数を表1に、解析された木中に正解木の含まれている率を表2に示す。比較のために、禁止木を全く使用しなかった場合、および人手による禁止木（現在機械翻訳システムで使用されている：130621本）を用いた結果も示す。

［考察］学習を行ったデータについては良好な結果が現れている。また、非学習データに関しても長年に渡って人手で作成したデータには及ばないものの比較的少数の学習禁止木で良好な結果が得られた。

表1． あいまいな木の数表

	禁止木（本数）		
テキスト	用いない	人手(130621)	学習(7021)
Hornby	21.2	2.9	1.6
英検4級	7.3	2.0	3.1

表2．正しい解析木を含んでいる率

テキスト	用いない	人手(130621)	学習(7021)
Hornby	100.0％	99.9％	100.0％
英検4級	100.0％	99.8％	92.3％

6. おわりに

規則の数が多くなるに従って、人手による文法の精密化が困難になってきている。人間が得意とする部分と計算機が得意とする部分をうまく融合することによってより良いルール体系を作成していく手法の提案を行なった。現在、学習によって得られた禁止木データを解析すると同時に、矛盾した禁止木にも着目して解析を進めている。学習により得られる禁止木が人手で作成した禁止木より特殊なものとなる傾向があるが、この傾向への対応を含めて今後以上の解析結果をもとにより精密な学習方式を目指す。

［参考文献］

[1]H.Sakaki et.al : "A Parsing Method of Natural Language by Filtering Procedure", Tran. of the IECE, Vol.E 69, No.10 (1986)

[2]A S Hornby : "Guide to Patterns and Usage in English, second edition", （英語の型と語法：伊藤健三訳注）Oxford University Press (1975)