帰納的学習を用いたタグなし文からの統語規則の自動的かつ動的な獲得手法

6K－4　渋木 英潔　荒木 健治　栃内 香次
北海道大学大学院工学研究科

1．はじめに

ルールベースの統語解析手法において、解析の精度は、解析に適切な統語規則が存在するかどうかに依存している。適切な統語規則を全て事前に与えておくことは困難である。また、人手で統語規則の不備を補う場合には、莫大な労力が必要となる。それゆえ、我々は、帰納的学習を用いて動的に統語規則を自動獲得する手法を提案する。

本手法で扱う統語規則は、チョムスキー標準形[1]で表わされる文脈自由文法である。文脈自由文法とした理由は、基礎として広く一般に使用されているからである。チョムスキー標準形を使用した場合、解析結果は二分木となるが、二文節間の係り受け関係を求める係り受け解析[2]の結果とすれば問題はない。

2．従来研究

コーパスから統語規則を獲得する手法として、文献[3][4][5]がある。文献[3]は、学習用コーパスとして括弧付きコーパスを必要とした。しかしながら、学習対象に適した、括弧付きコーパスのようなタグが付与されたコーパスが常に手に入るとは限らない。また、対象となるタグなしコーパスを学習できるようにタグ付けする場合の労力は、統語規則を与えるのと同じくらいの労力を必要とする。従って、タグなし文を処理できないような手法は頑健性に欠けると考えられる。文献[4][5]は、学習の指標として十分な量の統語規則を事前に与える必要があった。事前に与えられた知識を活用することは否定しないが、事前に与えられなければ学習できないような手法も頑健性に欠けると考えられる。それゆえ、本手法はタグなし文から学習することができ、事前に与えられる統語規則に依存しない。その結果，本手法は頑健なものとなった。

3．概要

本手法は、解析部と学習部からなる。タグなし文を処理するために、解析部では、現時点までに獲得された統語規則を参照してタグなし文を解析する。その文を完全に解析することができなかった場合、現在の統語規則に不備があると判断され、解析途中の不完全な解析結果を学習部に受け渡す。不完全な解析結果とは、解析木のルートノードの数が1でないものか、解析木の終端記号（単語）と前終端記号（品詞）が結びついていないものである。前者の場合、図1中のAのように、解析木のルートノードをまとめ上げる統語規則を獲得し、後者の場合、図1中のBのように、終端記号と前終端記号を結び付ける統語規則を獲得する。このとき、ノードして必要な統語範疇は、システムが任意に獲得し、番号で識別する。獲得された統語範疇は、二つの解析木を比較することで同じ統語

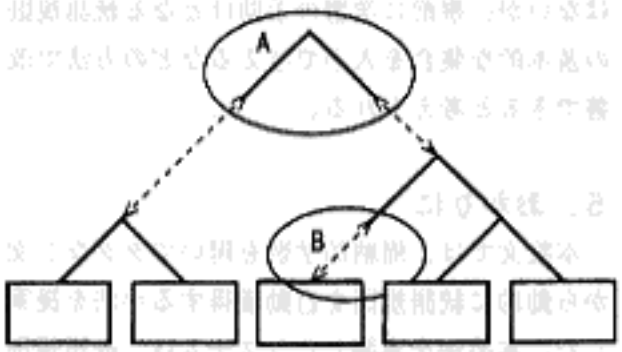

図1：不完全な解析木からの学習

範疇にまとめられる。それぞれの解析木に含まれる統語範疇は、その統語範疇以外の統語範疇が同じであった場合に同一の統語範疇と見なされる。獲得される統語規則の候補が複数あり、一意に決定できない場合には、全ての候補を一度獲得し、その後の解析結果から、統計的に頻度の高い統語規則を正しい統語規則と見なして統語規則の選別を行う。以上のようにして、入力文を解析できるような統語規則を動的に自動獲得していく。

4．実験

本手法の頑健性を証明するために、事前に一切の統語規則を与えないという最も本手法に困難な条件で実験を行った。外国人のための日本語学習用テキストから 860 文をとり、繰り返し学習させた。その 860 文中、何文を解析できるようになったかを解析成功率で示す。図2は、入力回数が 30 回までの解析成功率の推移を表わしている。30 回目の段階で 85.3%の解析成功率が得られた。また、その時の解析成功結果を人手で分析した結果、42.8%が正解（文節以上の係り受け関係に誤りがないもの）であった。本手法は、何も解析できない状態から、42.8%の精度で 85.3%を解析できるまでに学習でき、その有効性が確認された。

42.8%の解析正解率は、実用レベルの精度ではないが、事前に学習の手助けとなる統語規則の基本的な集合を人手で与えるなどの方法で改善できると考えられる。

5．おわりに

本論文では、帰納的学習を用いてタグなし文から動的に統語規則を自動獲得する手法を提案した。本手法を実装したシステムは、統語規則の初期状態が空という最も困難な条件で、85.3%の解析成功率と 42.8%の解析正解率を得られる統語規則を獲得し、本手法の有効性を立証した。

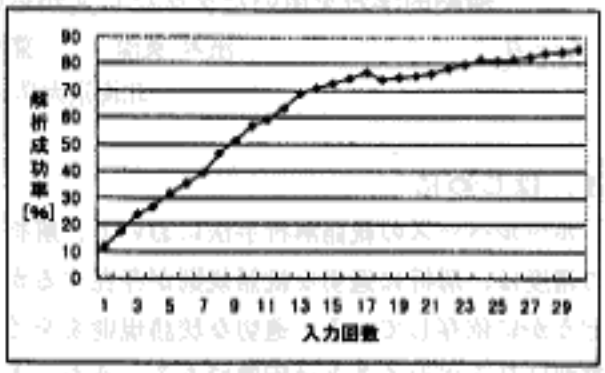

図2：解析正解率の推移

今後の課題として、事前に統語規則の初期集合を与えた場合の、本手法の有効性を確認したい。

参考文献

1) 長尾真編：自然言語処理，岩波講座ソフトウェア科学 15，岩波書店，(1996).
2) 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，(1998).
3) 白井清昭，徳永健伸，田中穂積：括弧付きコーパスからの日本語確率文脈自由文法の自動抽出，自然言語処理，vol.4，no.1，pp.125--146，(1997).
4) M. Kiyono，J. Tsujii：Hypothesis selection in grammar acquisition，Proc. of the 15th COLING，vol.2，pp.837--841，(1994).
5) M.Kiyono，J.Tsujii，Combination of Symbolic and Statistical Approaches for Grammatical Knowledge Acquisition, Proc. of the 4th Conference on Applied Natural Language Processing，pp.72--77，(1994).