203 帰納的学習により獲得された統語規則の記憶コストと曖昧性減少に基づく最適化手法

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

我々は，帰納的学習を用いて，統語規則（語彙と文脈自由文法の規則）を，いかなるタグも付けられていないコーパスだけから自動的に獲得する研究 [1, 2] を行っている．我々の手法は教師なし学習であり，以下の特徴を持っている．

- タグなしコーパスから学習できる．
- 統語規則の辞書が空の状態から学習できる．
- 対象となる文の種類を限定せず，対象に動的に適応できる．

しかしながら，その精度は充分なものではなく改善の余地があり，また，精度以外にも未解決の問題が残されている．

今回，我々は，獲得される統語規則を記憶コストと曖昧性減少に基づいて最適化することを中心に手法の改善を行った．記憶コストとは獲得される統語規則の総数を指している．一般に，統語規則の数が多くなるほど，適用できる統語規則の候補も増えるため，曖昧性は増大する．逆に，曖昧性がないようにするためには詳細に記述せねばならず，結果として，統語規則の数は多くなる．このことから，両者はトレードオフの関係にあると仮定し，その均衡点を見つけ出すことで統語規則の最適化を試みる．

2 統語規則の最適化

本手法の統語規則は，以下の流れで獲得される．

- ある文に対して，現在ある統語規則を用いて形態素解析と構文解析を行う．
- その結果が不完全なものである場合，その不完全な箇所を補うような統語規則の候補を獲得する．
- 候補の中から，頻度的に大きい候補を正しい統語規則と決定する．

この処理で獲得される統語規則は，ある特定の文を解析できるようにするための局所的な規則である．従って，この処理をコーパス中の全ての文に行っても，局所的な規則の集合しか得られない．局所的な規則は，他の規則との整合性をとることがないため，機能的に重複する部分や競合する部分を数多く残すことになる．こういった重複や競合する部分は，解析速度や解析精度を低下させる原因となる．我々は，局所的な規則全体を考慮することで，重複や競合する部分を取り除き，入力文全体を効率良く解析できる一般的な規則を作成することを試みる．一般的な規則を作成する指標として，我々は，入力された全ての文を，可能な限り曖昧なく解析できる統語規則の集合の中で，要素である統語規則の数が最も少ない集合が，一般的な統語規則の集合であると仮定した．可能な限り曖昧なく解析できる統語規則の集合とは，その集合を用いて文を解析したとき，計算上，求められる解の総数が最も少なくなる集合を指す．

2.1 統語規則数の減少

本手法では，統語規則をチョムスキー標準形で表現する．すると，右辺の長さが限られるため，作成可能な規則の総数は，規則を構成する統語範疇の総数に依存することになる．従って，統語範疇の数を減少させることで，統語規則の最大数を抑えることができると考えられる．本手法では，新たな統語規則を獲得するときに，既存の統語範疇で表わされることが不確実な統語範疇の部分は，システムによって新たに獲得される．それゆえ，獲得された統語範疇の中には，同じ統語範疇で表わすことが可能な統語範疇が多数存在する．

本手法は，統語範疇の同一化によって，統語範疇の減少を試みる．統語範疇が同一であるかの判断は，2つの解析木を比較することで行われる．解析木の形状が同一であるならば，解析木の同じ節点にあたる統語範疇は同一であると判断する．解析木の形状が同一とは，頂点から底辺に向かって分岐していく際に，全てが同じ箇所で分岐していることを指す．統語範疇の同一化により，同一の統語範疇で構成される統語規則が他にも存在することが判明した場合，その規則を統語規則辞書から取り除く．従って，この処理は，統語規則の最大数を抑えるだけではなく，機能的に重複する規則を取り除くこもできると考えられる．

2.2 解析結果の曖昧性の減少

ある一つの文に対して，複数の解析結果が導き出されるような統語規則の集合は，効率の面からも精度の面からも解析に不利益をもたらす．本手法によって獲得される統語規則は局所的な規則であるため，全体を通じて見た場合，統語規則の競合が多く起こっている．こうした競合する統語規則の大部分は，誤った統語規則を獲得したことが原因である．この誤った統語規則を取り除くことで，競合を抑えて精度を向上させることができると考えられる．ここで問題となるのは，言語が本来持っている曖昧性であり，誤った統語規則によるものか，言語本来のものかを区別しなくてはなら

ない．複数の解析結果が導き出される原因として，以下の３つが考えられる．

- 多品詞語による前終端記号の違い
- 解析木を組み上げる際に適用する統語規則の順序だけの違い
- 解析木を組み上げる際に適用する統語規則の種類と順序の違い

２番目は，日本語の「名詞の名詞の名詞」における曖昧性であり，３番目は，英語の「動詞 名詞 前置詞」における曖昧性である．この内，最初の２つについては言語本来の曖昧性であると考え，最後の原因だけを対象とする．

同じ統語範疇の並びに対して，異なる解析木が作成できる場合，その解析木を構成する統語規則は競合していると判断する．言語本来の曖昧性でなければ，競合している統語規則の一方を，偏差値情報を基に誤った統語規則であるとみなして取り除く．言語本来の曖昧性かどうかは，その統語規則を取り除いた場合に解析できなくなる文が存在する場合，言語本来の曖昧性であると判断する．また，競合するのは解析結果の一部であることから，底辺の統語範疇が３個となるような部分解析木に分解して曖昧性を判断する．こうすることで，競合している箇所だけを対象にして処理を行うことができ，処理効率やデータスパースネスなどの問題に対処できると考えられる．

3 その他の改良

その他の改良として，形態素の特定と偏差値情報の利用について述べる．

3.1 形態素の特定

本手法で獲得された統語規則を用いた解析では，文字を終端記号とする解析木を作成する．形態素に対応する部分解析木の頂点と終端記号の占める範囲を見ることで，形態素の品詞と文字列を判断することができる．それゆえ，形態素を表わす部分解析木を特定できれば良い．本手法の形態素解析では，今までに獲得した部分解析木を形態素と仮定して当てはめていく．このとき，１文字も残すことなく完全に部分解析木を当てはめることができたならば，当てはめた部分解析木は正しい形態素であると判断する．この判断は，常に正しいとは限らないが，正しい形態素を含む率を向上させることができると考えられる．しかしながら，この判断だけでは，１つの形態素を複数の形態素と判断する問題や，その逆に複数の形態素を一つの形態素と判断する問題を解決できない．この問題を解決するため，以下の仮定に基づいて，形態素を合成，および，分解する．

形態素の合成は，ある形態素と隣接して出現する形態素が予測可能かどうかに基づいて行われる．本手法では，予測可能かどうかを，隣接する形態素の相互情報量 [3] を用いて判断する．相互情報量の大きい形態素の組は１つの形態素と見なして合成する．

形態素の分解は，形態素文字列の共通部分と差異部分に着目して行われる．同一の統語範疇を持つ形態素の集合において，ある文字列が，常にその集合中の形態素の特定の位置に出現する場合，集合中の形態素は，その文字列から成る形態素とそれ以外の文字列から成る形態素に分解される．但し，チョムスキー標準形の統語規則を想定しているため，出現位置は先頭か末尾のどちらかに限定し，中央にある場合は考慮しない．

3.2 偏差値情報の利用

本手法では，記憶コストと曖昧性減少を基にした統語規則の最適化や，形態素の特定を行う．しかしながら，統語規則の最適化や形態素の特定における候補が一意に決定できることは少なく，数多くの候補が存在する．そうした候補の大部分は誤りであり，そういった候補の中から正しい候補を選び出さなくてはならない．しかしながら，本手法は，教師なし学習であるため，選択するための確実な根拠となるものを持たない．従って，本手法では，他の候補よりも多くの事例に適用できる候補は正しいという仮定に基づいて判断する．これまでは，この判断に頻度を使用していた．しかしながら，頻度を使用した場合，頻度の絶対値が大きくなると候補間の差の価値というのは，絶対値が小さいときと比べて相対的に小さくなる．この問題を解決するため，頻度の代わりに偏差値を使用する．偏差値は，その候補の集団内における位置付けを表わすので，他の候補よりも多くの事例に適用できる候補は正しいという仮定を頻度よりも正確に実現できると考えられる．

4 むすび

本稿では，タグなしコーパスから教師なしで獲得した統語規則を，記憶コストと曖昧性減少を基に教師なしで最適化する手法を提案した．また，以前の手法で問題となっていた形態素を特定する方法と，頻度の代わりに偏差値情報を使用することを述べた．今後は，本手法を実装したシステムによる実験を行って，本手法の有効性を確認する予定である．

参考文献

[1] 渋木英潔, 荒木健治, 桐内香次: "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", 情報処理学会第 60 回全国大会, 2000.

[2] 渋木 英潔, 荒木 健治, 桐内 香次: "平文を用いた統語規則の自動獲得", 電子情報通信学会 思考と言語 (TL) 研究会, 1999.

[3] 北 研二, 中村 哲, 永田 昌明, 音声言語処理, 森北出版, 1996.