205 帰納的学習を用いた長文分割手法

長島康人　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

ある程度以上の長さの文を対象とした場合、その文に対する翻訳や解析などの処理が失敗することがある。また、長文は人間が読む場合においても、理解しにくい悪文であることが多いため、長文の短文分割の研究が成されてきた。[1][2]しかし、従来手法では解析規則は人手により設計されたものが用いられてきたため、実際の多様な言語表現に対して規則の無矛盾性を保ちながら規則系を拡大することは困難である[3]。この観点から、規則の獲得に帰納的学習を導入し、それにより矛盾することなく規則の獲得を行う手法を提案する。その際、どのような情報を用いるかということを考えねばならないが、今回対象とする文が長文であるため、先に述べたように、構文解析や意味解析において誤ることが予想される。そこで、長い文に対しても比較的良好な精度が見込める形態素解析から得られる情報のみを用いて分割規則を獲得する。また、誤ったルールの繰り返し適用による誤りの拡大を防ぐため、フィードバックを行いルールごとに付加されている制約を変化させる。

2 処理過程

2.1 概要

本手法の処理過程をFig.1に示す。まずシステムに入力される品詞タグ付けされた英文に対して、獲得した分割ルールを用いて分割を行う。次に、分割結果に対して人手により校正を行い、それを用いて新たなルールの獲得を行う。その後、使用されたルールに対するフィードバック処理が行われる。

2.2 学習部

学習部では品詞タグ付けされた英文とその正分割例から、帰納的学習により分割ルールを獲得する。獲得されるルールのかたちは、分割点直後の単語とその前後の形態素列を一組として構成され、各ルールごとにいくつ以上の形態素の一致でルールが適用されるかの制約がある。

2.3 フィードバック部

まず人手による正誤の判定を行い、ルールごとの正解数、誤り数をカウントする。これらを用い分割ルールにフィードバック処理を行い、各ルールごとの制約を変化させることによりルールの誤適用を防ぐ。

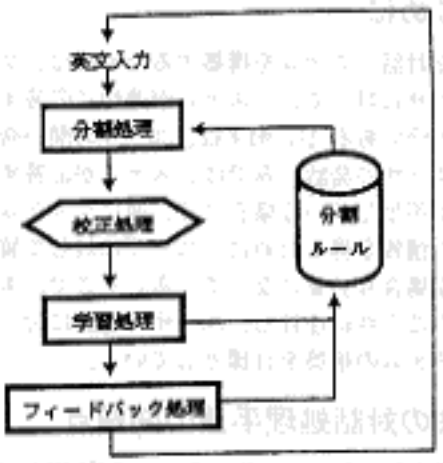

Fig. 1 処理過程

3 評価実験の方法

実験はルールごとの制約を固定したものと、フィードバックを用い制約を変化させた場合の二種類で行い比較検証する。分割点は「等位節及び従属節の接続点」と定義する。また、本実験ではその性質上ある程度一つ一つの文が長いものが適当であると考え、言語処理関係の英語論文を対象として用いる。実験は分割ルールが空の状態から行う。

4 おわりに

本稿では、長文分割のルール獲得の際に帰納的学習を導入する手法を提案した。これにより従来手法の問題であった矛盾の無い規則の獲得が可能となるものと考えられる。今後は、今回提案した手法に基づくシステムを作成し、その有効性を確認していく予定である。

参考文献

[1] 張玉潔, 尾関和彦：分類木を用いた日本語長文の自動分割, 言語処理学会第4回年次大会発表論文集,pp390-393(1998)

[2] 張玉潔, 尾関和彦：分類木を用いた日本語文の自動文節分割 情報処理学会研究報告,vol.97,no.85,pp1-8(1997)

[3] 森英悟, 荒木健治, 富永喜一, 栃内香次：帰納的学習による表層文から意味表現への変換規則の自動獲得と適用, 電子情報通信学会論文誌,Vol.J81-D-No.7,pp1621-1630(1998)