形態素解析を用いた主題・焦点抽出システム

廣町　潤　横山　晶一　西原　典孝
（山形大学　理工学研究科）

１．はじめに

談話解析において、談話構造を正確に捉えるためには、主題・焦点を抽出し、そのつながりを見ることが必要となる。主題・焦点の情報は、情報抽出、文章の要約などにも有用である[1]。我々は主題・焦点の抽出アルゴリズム[2]を確立しているが、それには「構文解析が完成している」という重い前提条件があった。そこで、この前提条件をはずし、形態素解析のみから主題・焦点を抽出するシステムを構築した[3.4]。

このシステムでは、主題を「は」の名詞から、焦点を「が格」から抽出する処理を基本に、埋め込み文の処理、「は」「が」が複数あったときの処理などを追加し、抽出を行っていた。しかし、従属節、埋め込み文など、複雑な文章に対しては抽出精度に問題があり、ある程度の結果はでたが、実用化段階には至らなかった。

本研究では、このシステムに従属節の判定などを加えて改良することにより、さらに抽出精度を上げた。また、要約などに用いるためには、主題・焦点のみでは情報が少なすぎるので、修飾情報などを同時に抽出した。これによって実用化可能なシステムを構築した。

２．システムの概要

システムの流れを図１に示す。

形態素解析は茶筌[5]を使用する。

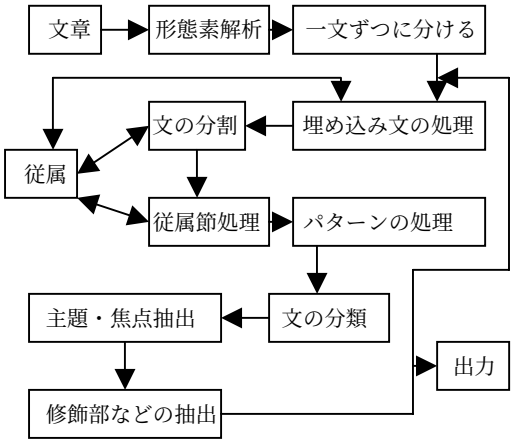

図 1.抽出システムの流れ図

2.1　埋め込み文の判定

「～という[名詞]」、引用の「と」により、その前方読点までを「埋め込み文」とする。次に従属節の判定を行い、読点の前が従属節であるなら、つながりがあるとし、「埋め込み文」をさらに前方の読点まで伸ばす。

2.2　従属節の判定

野田[6]を参考に、表１の分類に従って、従属節の判定をする。判定は読点ごとに行われる。

表 1.従属節の分類

種類	代表例	主題	焦点	文を分ける
強い従属節	ながら、と、たら、(れ)ば、ほど、とき、「連用形」、て	×	×	
理由節	ため、から、ので、のに	○	×	
並列節	し、けれど、が、「連用形」			○
引用節	と、って	×	×	

2.3　文の分割

従属節の処理を参照し並列節で文を分割する。

2.4　従属節の処理

従属節の判定を参照し、従属節部分を主題・焦

A System Extracting Themes and Focuses using Morphological Analysis

Jun Hiromachi,Shoichi Yokoyama,Noritaka Nishihara
Yamagata University

点として抽出しないようにする。

2.5　文の分類

述語の形から名詞文、動詞文、形容詞文を分類する。

2.6　主題・焦点の抽出

名詞文：「は」があるときは、主題「は」の名詞、焦点は述語名詞。「が格」があるときは、主題述語名詞、焦点「が格」。「が格」がないとき、主題は補完処理、焦点述語名詞。

動詞文：主題「は」の名詞、ないとき主題補完。焦点は「が格」ないときは、動詞を辞書で調べ必須格を抽出。

形容詞文：主題は「は」の名詞、ないときは主題補完。焦点は、「が格」、ないときは辞書で調べて必須格を抽出。

焦点は、主題が述語名詞以外は、主題の後方から抽出する。

2.7　「が格」の処理

「が格」と述語の間に動詞が存在した場合、その「が格」はその動詞に係るものとみて、焦点にならないようにする。

2.8　主題の補完

名詞文：前文が名詞文なら、前文の述語名詞を補完する。動詞文なら焦点、あるいは前文を名詞化して補完。

動詞文：動詞を辞書で調べ、マッチするものを、前文の主題・焦点、前々文の主題と調べる。前方の主題を調べる範囲は同段落内まで。段落先頭文の場合は、前段落内。適切なものがない場合は「一般」を補完。

形容詞文：動詞文と同様。

文を分割した文で、後部分の文に省略があるときは、前部分の主題を補完。

2.9　修飾部などの抽出

主題・焦点の前が動詞、助詞「の」などの場合修飾部として抽出する。

3.抽出例

例は、主題=二重下線、焦点=下線、埋め込み=囲み線、文を分ける=網掛け、補完をφで示す。

文）n文：首相は、その実現のために政府機構の改革が必要だと述べて、[φ=首相]改革案を示した。

n+1 文：[φ=改革案]事務機構などを減らし、非常設機構も削減する、という思い切った計画である。

n文：主題=首相、焦点なし、
　　主題=「首相」を補完、焦点=改革案

n+1 文：主題=「改革案」を補完　焦点：計画、
　　焦点修飾=思い切った

文）国連難民高等弁務官事務所はコンピュータによる初の子供検索システムをスタートさせる。

主題：国連難民高等弁務官事務所

焦点：子供検索システム、焦点修飾=コンピュータによる初の

4.おわりに

従属節を考慮に入れることにより、抽出精度が向上した。しかし、複雑な文章や、省略された主題の補完など、抽出を誤ることがある。今後は、さらに抽出アルゴリズムの改良を進め、より正確な抽出を目指す。

参考文献

[1]横山晶一・菅野崇：言語処理学会第 7 回大会論文集

[2]吉田悦子・横山晶一：信学技報、NLC97-29(1997)

[3]廣町潤:形態素解析を用いた主題･焦点抽出システム、山形大学卒業論文(2001)

[4]S.Yokoyama,J.Hiromachi,N.Nishihara.:Proc.IEEE SMC conf. (2001)pp.882-886

[5]形態素解析システム「茶筌」、奈良先端科学技術大学院大学

[6]野田尚史：新日本語文法選書 1　「は」と「が」、くろしお出版(1992)