話題の推移に着目した談話構造解析システム

2M−01

沼野 元伸　西原 典孝　横山 晶一
（山形大学 理工学研究科）

1. はじめに

膨大な文書データから必要な情報を抽出するための研究が数多く行われてきた。その一つとして、談話の構造を捉えることで効率の良い情報抽出が可能と考え、主題・焦点を用いた談話構造表現の抽出の手法[1]や、主題とその修飾部に着目した談話構造の解析[2]についての研究が行われ、一定の成果をあげている。

談話構造を捉えるにあたっては、主題・焦点、それらの修飾部が重要な役割を持つことがわかっている。しかしながら、[1][2]では各文間の談話の推移を捉えているものの、文書全体の大きな談話構造を捉えるまでに至っていない。またこれらは、曖昧性のない構文木を前提条件として必要とし、現状での機械化は困難である。

本研究では、人手による前処理を行わずに主題・焦点及び各々の修飾部をパラメータとして抽出し、文書全体の話題の推移を抽出するような談話構造解析システムの構築を行った。そして、このシステムの有用性について考察した。

2. 談話構造解析システム

文書には、話題の推移が存在する。時として、これは章分けや小題と言った形で実際の文書中に現れてくる。文書を構成する最小単位を 1 文と考えると、その文の話題は主題に現れる。よって、主題やその修飾部の推移を捉えることで、話題の推移を捉えることが可能である。

しかしながら、機械処理のみで、形態素解析から的確な主題・焦点を捉えることは現状では困難である。

A System Analyzing Discourse Structures using Topic Transition
Motonobu Numano,Noritaka Nishihara ,Syouichi Yokoyama
Yamagata University

単純な文章からの主題・焦点の抽出は可能であるものの、実際には、複文、埋め込み文、会話文などの取り扱いなど、いくつかの問題が存在する。

そこで、本研究では文書を文単位で捉えるのではなく段落単位で捉え、段落中から主題となりうる主題の候補を抽出し、その主題の候補と、他の段落との関係性を見ることで話題の推移を掴むことにした。また、段落単位で文書を捉えることで、文書全体の話題の推移と関係性の低い冗長な主題を省くことができる。

3. システムの手順

まず、茶筌により形態素解析を行い、次に、主題・焦点の候補、各々の修飾部となる名詞の抽出をはじめに行う。さらに、抽出した名詞と他段落との繋がりを基に話題の推移を捉え、話題の推移を示すデータを出力する。

3.1 各種パラメータ(話題名詞)の抽出

主題・焦点の候補の抽出に関しては、[1]で提案された手法を基に、機械的な処理によって行う。次に、抽出された主題・焦点の候補に助詞「の」と、これに順ずる接続語によって接続している名詞を、修飾部として抽出する。

抽出された各名詞中の複合名詞を、文書中に出現する名詞との比較から分割を行い、抽出された名詞、分割された名詞を話題名詞と呼ぶこととする。

例 1)

n 段落　たこ焼き　専門店　は、・・・
　主題　+3[5]　+3[4]
n+1 段落　・・・を　カレー　専門店が　、・・・
　焦点　+2[2]→[6]　+2[6]
話題名詞　たこ焼き、カレー、専門店

3.2 各段落における話題名詞への加点

話題名詞の文書中での利用のされ方から、話題の

範囲を特定できると考えられる。よって本研究では、抽出された話題名詞の各段落での利用のされ方から、段落ごとに話題名詞に点数を与える。この点数の推移に話題の推移が表れる。

表 1. 各段落での基本となる得点規則

話題名詞の出現形態	話題名詞への得点
主題(「A は」で出現)	+3
主題(その他)	+2
焦点	+2
主題・焦点以外	0
サ変名詞の動作主	0
出現しない	-1

表 1 は、各段落における話題名詞への得点の基本となる規則である。ただし、話題名詞は 1 点以上の得点を得るまで点数を持たないこととし、点数を持っている状態から点数が 0 点を割ったときにはその話題名詞は再び点数を持たない状態になるものとする。また、それぞれの修飾部中に出現する話題名詞に対する得点も表 1 に準ずる。

3.3 点数付けの特殊規則

話題名詞 A が初めて 0 点より大きい点数を加点され、A が修飾関係、または複合名詞を構成要素同士の関係にある話題名詞 B を持つとき、A への得点よりその段落での B の点数が大きいのであれば、A へは B の点数を与えるものとする。つまり、例 1 においては、n+1 段落での「カレー」の得点が 2 点であり、「専門店」が点数 6 点を持っていることから「カレー」の点数を 6 点とする。

また、点数を持った話題名詞 A が点数を失うまで、以降の段落での得点が常に-1 点である時は、A が点数を持った話題名詞 B と修飾関係、複合名詞の関係になければ、以降の段落において A は点数を失うものとする。例 1 において、n+2 段落以降、n+7 段落まで「カレー」の得点が-1 点であったとする。このとき、n+5 段落において「ピザ専門店」という複合名詞が出現し「カレー」と「専門店」との関係が失われたとすれば、n+2 段落から n+4 段落間の「カレー」の点数は「5,4,3」と推移し、n+5 段落以降は点数を持たない状態になる。

4. システムの出力と有用性の考察

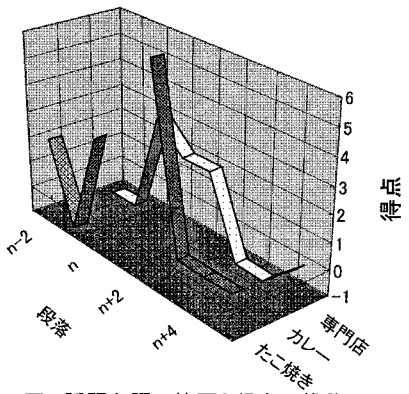

図1 話題名詞の範囲と得点の推移

本システムでは話題名詞の各段落での点数を出力する。この出力において、数段落連続して話題名詞の点数を持つとき、この段落の範囲がその話題名詞の範囲となる。また、前段落との点数の差(段落での得点)は、話題名詞の段落での話題の強さと考えられる。図 1 は、システムの出力を基に作成した、話題名詞の範囲と、各段落における話題名詞の話題としての強さを示すグラフである。

おわりに

システムの出力から図 1 のようなグラフを作成し、グラフとシステム中で抽出した話題名詞間の関係性から、複数の文書の談話構造を抽出した。複数の段落に大きく跨る話題に関しては比較的良好な抽出を行えたが、話題の範囲の狭いものに関しては的確に捉えることのできなかったものも多数存在する。これは、話題名詞間の関係性の強化からある程度の改善が見られるのではないかと考えられる。

謝辞

形態素解析システムとして、奈良先端科学技術大学院大学松本研究室の「茶筌」を使用させていただきました。ここに謝意を示します。

参考文献

[1]斎藤尚子、横山晶一：語の重要度を考慮した談話構造表現の抽出、言語処理学会第 6 回年次大会 発表論文集 pp.467-470(2000)

[2]杉本孝太：主題修飾部に着目した談話分析に関する研究、山形大学工学部卒業論文(2001)