主題・焦点と表題との関連性を用いたキーワードの抽出

横山 晶一　小谷 郁夫
山形大学 工学部

1 はじめに

ネットワークと、その上を流れる情報量の爆発的な増大に伴い、必要な情報を的確に得ることがますます困難になりつつある。このために、流れてくる情報を取捨選択したり、求める情報を検索するためのキーワードの抽出や、文章の要約 [1. 6] が重要な問題になっている。

キーワード抽出や要約の手法としては、文章中に現れる語の頻度情報を利用したり、表題情報との関連性 [5] を用いたりする研究が行われている。しかしながら、文章の談話構造をある程度考慮しないと、関係のない語をキーワードとして抽出したり、要約文中から重要な部分が欠けてしまう場合がある。

一方、日本語の談話構造においては、主題・焦点の果たす役割が大きい。文の構文解析が行われていると仮定したときに、主題・焦点を抽出するアルゴリズムについてはすでに発表した [8]。このアルゴリズムを用いると、主題・焦点が的確に抽出でき、談話の流れを知るのに役立つ。

本研究では、談話構造や話題の推移に重要な役割を果たしていると考えられる主題・焦点を、文章のキーワードとしてどの程度まで用いることができるかという観点に基づいて実験を行った。その結果、表題との意味的な類似性を考慮すれば、抽出した主題・焦点が、キーワードとして有効であるという結果が得られた [4] ので報告する。具体的には、(人手で) 抽出した主題・焦点に、日本語語彙大系 [2] の分類番号を付し、その番号と、表題の分類番号とが一致するものをキーワード候補として残すという手法をとっている。以下にこの方法について詳しく述べる。

また、タイトルが付いていない場合や、タイトルに抽象語や動詞的な要素が含まれている場合には、この方法をうまく用いることができないが、その場合に、主題・焦点から抽出した語をグループ化することによって、キーワードが抽出できることを示す。

2 主題・焦点の抽出法

談話解析において、主題・焦点はさまざまな定義がなされてきた。ここでは、次のように定義する [8]。

主題： その文中で話題となっている要素であり、前述された既知の情報

焦点： その文中で新しく導入された情報

この定義に基づく主題・焦点について、第1文 (文章の最初に現れる文) と、第2文以下に対して別々の抽出を行う。また、各文を、述語の形から、動詞文 (「A は B する」のように、述語が動詞で終わるもの)、形容詞文、名詞文 (「A は B だ」) の3種類に区別し、それぞれ異なるアルゴリズムで処理を行う。

また、処理に当たっては、次のような前提条件を設ける。

(1) 「は」は主題の「は」に限り、対比の「は」は対象外とする。
(2) 格関係を持たない文は対象外とする。
(3) 曖昧性のない構文解析木が作られているものとする。

詳しいアルゴリズムについては、文献 [8] に譲るが、ここでは、第2文以降の処理の主要部分についてのみ述べる。

i) 主題を表す「は」が存在 (主題＝「は」の前の要素)
(a) 名詞文：焦点＝述語名詞
(b) 形容詞文：焦点＝述語または主題以外の格要素
(c) 動詞文：焦点＝主題を除いた先頭格要素

ii) 「は」はないが、必須格要素がすべて存在
(d) 名詞文：主題＝述語名詞、焦点＝主格名詞
(e) それ以外：主題＝「が」格より前の必須格要素、
焦点＝「が」格または先頭格要素

iii) 必須格要素が一つ欠けている場合
(f) 名詞文 (文照応)：主題＝前文の名詞化、
焦点＝述語名詞
(g) それ以外：主題＝素性一致なら前文の主題、
そうでないなら前文の焦点、
焦点＝存在する必須格の先頭要素

(a)〜(g) を適用して主題・焦点を抽出する。このアルゴリズムを応用すれば、要約 [1]、文脈構造 [7, 9] などに用いることができるが、それらは文献に譲る。

3 キーワード抽出方法

前節で述べた主題・焦点抽出アルゴリズムを用いて、次の手順でキーワードを抽出する。

1. 上記アルゴリズムに基づいて、人手で主題・焦点を抽出する。
2. タイトルから名詞を抽出する。このとき、日本語語彙大系 [2] を参照して、その名詞が含まれるグループの分類番号と、そのグループのすぐ上 (以下では「親」と称する) の番号を与える。複合名詞は、大系で参照できる最小単位に分解して分類番号を与える。タイトルが複数ある場合には、すべてのタイトルを参照する。
3. 抽出した主題・焦点の中に、タイトル中の名詞と一致する部分があれば、その単語をキーワードとして残す。
4. 抽出したその他の主題・焦点に対する分類番号を上と同じように調べ、タイトル中の分類番号と一致するもののみをキーワードとして残す。
5. 残されたキーワードをまとめる。

以下に、この手順に沿って実験した結果について述べる。

4 実験と評価

実験対象としては、この方法の有効性を検証するために、すでに著者によってキーワードが付された文献 [8]、もともとタイトルがなかったが、後に筆者とは別の人によってタイトルが付された文献 (天声人語) などを用いた。以下では、キーワードが付された文献 [8] を例として述べる。

4.1 実験手順

(1) 主題・焦点の抽出

前節の手順に従って、主題と焦点を抽出する。以下にその一部を示す。二重下線が主題、下線が焦点である。

これまでの談話解析は、名詞の出現頻度や照応関係、キーワード抽出の観点からの研究が主体であった。しかしながら、談話構造を正確に捉えるためには、主題を抽出して、そのつながりを見ることが必要である。また、照応関係の解析においても、主題が明らかであれば解析が容易になる。これまでは、主題の抽出方法が曖昧であったため、機械処理の段階までには至らなかった。

(2) タイトルからの名詞の抽出と分類番号の付与

タイトル「主題・焦点を用いた文脈解析の一手法」の中の名詞を抽出し、その語についている番号と、親の分類番号とをすべてつける。

主題 [1026, 1029, 1032, 1071 親 1019, 1070]、
焦点 [1030, 2616 親 1019, 2615]、文脈 [1085 親 1080]、
解析 [1946, 2068, 2075, 2498 親 1942, 2066, 2496]、
手法 [1035, 2506 親 1019, 2503]

(3) タイトルと一致する単語の抽出

上の、「主題」、「焦点」、「解析」などが主題・焦点として抽出された場合、これらを抜き出す。

(4) その他の語への番号の付与

たとえば、「抽出」という語が主題・焦点として抽出されたならば、その語に [2187 親 2186] という番号を付与する。

4.2 実験結果

表 1 抽出されたキーワードと著者のキーワードの比較

著者キーワード	主題、焦点、文脈解析、談話、話題
抽出キーワード	解析、方法、談話解析、主題、動詞文、名詞文、形容詞文、焦点、省略、補完、対象外、順序、文、課題、分類番号、ラベル、焦点「教室」、意味、入れ子構造、主話題、副話題、話題転換、話題、名詞、派生、流れ、文脈解析、問題点、話題推移形式
一致、部分一致	解析、談話解析、主題、焦点、文脈解析、話題推移形式、主話題、副話題、話題転換、話題

表1に、上記手順に従って抽出されたキーワードを、著者があげたものと比較して掲げる。

この表では、非常に多くのキーワードが抽出されているが、これは、対象とした文献が数ページにわたるものであり、その文のすべての主題・焦点から分類番号が一致するものを抜き出したためである。表から分かるように、5つのキーワードのうち、完全一致するものが3つ、部分一致するものが1つ抽出された。しかもそのうち、「話題」は、タイトルには入っていないものである。

4.3 評価と問題点

天声人語は、前述のように、もとのコラムにはタイトルがついておらず、英訳する際にタイトルを付けられたものと考えられるが、これについて上の手順で行った結果の一部を表2に示す。表で下線を引いた部分は、人間がキーワードとしてふさわしいと判定したものである。

表2 天声人語の結果

	タイトル	抽出されたキーワード
1	漫画による性教育	教育研究集会、性教育、授業、言葉、漫画、絵
2	名前と実体	ことば、名前、具体、意図、忘れがち
3	中国の行政改革	中国、改革案、行政改革
4	くさやの味	干物、香り、ムロアジ、海の幸

表から分かるように、タイトルの語がほとんどそのままキーワードとして抽出されているもの(3)や、逆に、タイトルから連想されるキーワードが並んでいるもの(4)がある。一般的に、タイトルから抽出された名詞が抽象的なものの場合には、キーワードとして適切でないものが抽出される傾向が見られた(2)。

たとえば、「西暦2000年問題に全力で取り組もう」(青山 幹雄、情報処理 Vol. 40, No. 5, p. 451)のような、タイトルが動詞中心で、サブタイトル(表の5〜8)も同様なコラムでは、下の表3に示すように、それほどよい結果は得られない。

この表で分かるように、タイトルが動詞中心で、しかも抽象的な場合には、この方法では、ほとんど適切なキーワードを取ることができない。そこで、予備的な段階であるが、タイトルがない場合の処理について次節で簡単に述べる。

表3 抽象的、動詞的な語を持つタイトルからのキーワード抽出

	サブタイトル	抽出されたキーワード
5	西暦2000年問題を客観的に理解しよう	特集、本特集、法的課題、危機意識、メディア情報
6	西暦2000年問題の背後にある意義を考えよう	西暦2000年問題、問題、難しさ
7	西暦2000年問題の教訓を活かそう	対象、ソフトウェア開発、西暦2000年問題
8	西暦2000年問題に全力で取り組もう	情報処理技術者、西暦2000年問題

5 タイトルがない場合の処理

表4 タイトルなしでのキーワード抽出の結果

	タイトル	抽出されたキーワード
1	漫画による性教育	性教育、授業、教育
		絵、漫画
		分析、研究
		女の子、男子、子供
		エイズ、不快
2	名前と実体	名、言葉
		言葉、名前
		言葉、意図
		事情、場合
		場合、例
		記録、辞書
		鮮明、具合

実際の文章では、タイトルがつけられていなかったり、前節に述べたように、タイトル情報をうまく用いることができなかったりする場合がある。そこで、タイトルがなくても、主題・焦点を用いれば、ある程度のキーワード抽出を可能にするために、次のような手順で実験を行った。

1. 主題・焦点を抽出する。
2. 抽出された主題・焦点に、日本語語彙大系[2]の分類番号をすべて(木構造の最上位からその語のところまで)与える。
3. 各単語間で一致する分類番号を求め、8割以上が一致するものを同じグループに入れる。複数

の語でグループを形成できたもののみをキーワードとする。

天声人語について、タイトルがないものとして行った実験結果を表4に示す。キーワードの欄のそれぞれの行が、上の手順でキーワードとして採用された語のグループを示している。

この表から、1のように具体的な内容はもちろん、2のように抽象的な内容であっても、比較的適切なキーワードが抽出されていることが分かる。また、同じ言葉が重複していくつかのグループに出てきているが、その場合でもキーワードとしての適切さは失われていない。

6 おわりに

主題・焦点とタイトルを用いてキーワードを抽出する方法について述べた。ここで述べたように、タイトルの名詞のみを用いたり、頻度情報のみを用いたりする場合と比べて、タイトルの名詞に近い意味を持つキーワードが適切に抽出されていることが分かる。

タイトルが抽象的であったり、タイトルに動詞的な要素が多く含まれている場合には、やや抽出の正確さが落ちるが、その場合には、最後に述べた、タイトルがない場合の手法を併用することによって、適切なキーワードが抽出できると考えられる。

タイトルがない場合の方法をタイトルのある場合の方法と併用すると、タイトルの意味に引きずられない、より適切なキーワードが抽出できる可能性があるので、今後はこの方向で検討する。

タイトルがない場合の処理で、意味分類番号が複数ある場合は、現在それらの中で最も一致度の高い数値を示したものを採用している。これは、意味のより近いものを採用するためである。その際の8割一致というのは、特に根拠があって設定した数値ではないので、この数値が適当であるかどうかや、比較をどのように行うかについてもなお検討の余地がある。

ここでは、比較的短いコラムを対象としたため、グループに入る語が複数あれば、すべてキーワードとしているが、もっと長い文章からキーワードを抽出する場合には、グループ化される語の数を多くしたり、頻度情報なども用いたりすることが必要であろう。

また、主題・焦点のアルゴリズムについても現在検討を行うとともに、自動化(前提条件を満たした場合にはほぼ完成している[3]が、前提条件まで入れたシステムはまだ未完成である)についても検討中である。さらに、他のアルゴリズムとの比較検討も行う予定である。

参考文献

[1] 細梅 久典・横山 晶一：主題・焦点を用いた要約文の抽出、言語処理学会第6回年次大会論文集 A5-1 (2000)

[2] 池原 悟他編：日本語語彙大系、岩波書店(1997)

[3] 井関 肇：日本語文の主題・焦点抽出システムの作成、山形大学卒業論文(1999)

[4] 小谷 郁夫：主題・焦点抽出システムの情報検索への応用、山形大学卒業論文(2000)

[5] 仲尾 由雄：見出しを利用した新聞・レポートからのダイジェスト情報の抽出、情報処理学会自然言語処理研究会資料 NL117-17 (1992)

[6] 奥村 学・難波 英嗣：テキスト自動要約に関する動向、自然言語処理 Vol. 6, No. 6 (1992) pp. 1-26

[7] 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回年次大会論文集 B6-1 (2000)

[8] 吉田 悦子・横山 晶一：主題・焦点を用いた文脈解析の一手法、電子情報通信学会技術報告 NLC97-29 (1997)

[9] 吉田 悦子・横山 晶一・西原 典孝：主題間の関係を用いた文脈構造ネットワークの構築、情報処理学会自然言語処理研究会資料 NL124-3 (1998)