主題・焦点のスコアを用いたキーワードの抽出

横山 晶一　菅野 崇
山形大学 工学部

1　はじめに

日本語の談話構造から得られる重要な情報である主題・焦点を用いてキーワードを抽出すると有効であることはすでに報告した [1]。この報告では、主題・焦点と表題との関連性に着目して、抽出された主題・焦点と、表題とが意味的に類似している場合に、それらをキーワードの候補として選択するという手法が有効であることを示した。また、表題がない場合でも、主題・焦点の中の意味的に類似している語をグループ化することによって、キーワードを得る手法についても言及した。

これらの手法は、比較的短い文章に対しては有効であるが、やや長い文章からキーワードを抽出しようとすると、キーワードの数が多くなり過ぎるという欠点があった。また、主題・焦点の抽出を、構文解析が完全にできた文を前提として人手で行っていたために、人によって抽出結果が異なるという問題もあった。

本報告では、抽出された主題・焦点およびそれらの修飾部に対し分類番号を付加するとともに、その現れ方や表題との関係を考慮してスコアを与え、各々の語のスコアを合計することによって重要なものを選択し、それらをキーワード候補として選定する試みについて述べる [2]。この手法を用いれば、表題の有無にかかわらず、キーワードとして有効な語を高い点で抽出することができる。また、人手で行う部分を現在自動化しているが、その試みについても触れる。

2　抽出法とスコアの計算法

2.1　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、すでに報告したが、次のように定義する [3]。

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文と、第2文以下に対して別々の抽出を、各文の述語の形で3種類（動詞文、形容詞文、名詞文（「AはBだ」））に区別し、それぞれ異なるアルゴリズムで処理を行うのは前回と同じである。また、構文解析が正しく行われているなどの前提条件（自動化にあたっては、この前提条件を緩めている）も同様である。

2.2　キーワード抽出方法

キーワード抽出の手順は、前半部（主題・焦点を抽出してそれらを候補とする）は、ほとんど前回に述べた手法と同じである（異なる部分は明記する）が、後半部（スコアを計算する）が前回とは異なっている。

1. アルゴリズムに基づいて、主題・焦点を抽出する。この部分は、現在形態素解析システム [4] と結合価を用いて自動化を行っている [5]。このときに、前のシステムにはなかった試みとして、主題・焦点を直接修飾している名詞も抽出する。
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 [6] を参照して、その名詞が含まれるグループの分類番号と、そのグループの上位概念の番号を与える。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。タイトルがない場合にはこの処理は行わない。この部分も前回の発表時には手作業であったが、現在自動化を試みている。
3. 抽出した主題・焦点、修飾部の名詞について、上と同様に分類番号を付加する。
4. 表題との関連やグループ化の程度を考慮しながら抽出した名詞に対してスコアを計算する。計算のしかたについては以下に述べる。
5. 合計点の高いものを選んでキーワード候補とする。

2.3 スコアの計算法

分類番号の一致の判定やスコアの計算を表１にまとめて示す。

表１　スコアの計算法

スコア計算部分	点
主題・焦点部	
主題・焦点	+2
主題・焦点修飾部	+0
これらの頻度	+1
語彙大系分類番号	
具体名詞	+5
事	+3
抽象物	+2
抽象的関係	+1
(固有名詞)	(+6)
グループ化	
表題の語とグループ化	+10
表題の語と一致	+10
一致した固有名詞	+20
グループ化の程度	+n

(1) 主題・焦点部のスコア計算

前節のアルゴリズムと、上の表１に示すように、抽出された主題・焦点、また、それらの修飾部に対して加点する。具体的には、主題と焦点には、抽出された時点で自動的に２点を加える。修飾部自体には加点しない。ただし、主題・焦点、それらの修飾部に当たる語が１回出現するごとに１点を加える。これは、頻度情報を考慮することに相当する。

(2) 語彙大系によるスコア計算

日本語語彙大系 [6] では、分類番号の上から２段目で、「具体」と「抽象」に二分される。「抽象」は、さらに、人間や自然界の行為・活動を表す「事」、その行為の結果を表す「抽象物」、種々の「抽象的関係」に分けられる。これらの中で「具体」の重要度が高いと考えて、表１に示すように、具体名詞に５点を加え、以下表のような点数を加える。

固有名詞（形態素解析システムでは多くの場合未知語として出現する）は、文書を特徴づける重要なキーワードとなる可能性が高い。しかしながら、文書中に余り出現しない固有名詞を選択すると、誤るおそれがある。そこで、主題・焦点、それらの修飾部中で２回以上出現した場合に限って、表１に示す６点を加点する。表１でカッコで示したのはその意味である。

(3) 分類番号に基づくグループ化

前回の発表においては、表題のある場合には、表題中の名詞と一致するものを無条件でキーワードとして選択し、表題がない場合には、語彙大系の分類番号のうち、上から 80%が一致するものを選んでグループ化していた。この方法では、分類番号の段数が異なる場合や、分類番号で比較的浅いところにある場合に、グループ化が困難であるという欠点があった。

今回のシステムでは、分類番号を下位から比較し、下から２段のうちで１つでも一致するものがあれば、それらの語をグループ化する。

例：交通網 [/1 名詞/2 具体/388 場所/389 施設/390 公共施/417 交通路]
　　幹線道路 [/1 名詞/2 具体/388 場所/389 施設/390 公共施/417 交通路/418 道路]

この例では、「交通網」の下から２段は、「公共施」、「交通路」である。「幹線道路」は「交通路」、「道路」であるので、「交通路」が一致する。したがって、この２語はグループ化される。

グループ化された語のスコアは次のように計算する。まず、主題・焦点が表題の語とおなじグループにまとめられたときには、その語はきわめて重要度が高いと判定し、点数を 10 点加点する。表題の語と全く同じ語の場合にはさらに 10 点を加える。すなわち、表題中の語が主題・焦点となったときには、自動的に 20 点が与えられることになる。また、グループ化できない固有名詞の場合には、表題の語と全く同じ語であったときに、20 点を加える。

次に、主題・焦点同士でグループ化された場合には、類似の意味の語がどのくらい出現しているかで加点する。具体的には、一つの主題・焦点に対してその語とグループ化された他の主題・焦点の数＝点数とする。たとえば、「消防」という語に「消火」と「救急活動」がグループ化された場合には、「消防」に２点が加えられる。

3 実験と評価

実験対象としては、社説や論説文（天声人語など）を用いた。以下に社説の例を前回発表した手法と比較しながら述べる。なお、以下の表では前回発表した手法を従来手法、今回発表する手法を本手法と呼ぶ。

3.1 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（毎日新聞の1994年社説）を示す。二重下線が主題、下線が焦点である。

出初め式は気持ちがいい。伝統行事は人の心を暖める。

その消防職員に団結権を認めていないのは、先進国といわれる国では日本だけである。フランスの場合、一部の大都市で軍隊が消防にかかわっているが、団結権を全く無視している先進国はない。なぜ日本だけがそうでなければならないのか。

実は国際労働機関(ILO)からマイヤー次長ら幹部二人が年明け早々来日し、細川護熙首相とも会談した。首相は団結権について「早く方向性を打ち出せるようにしたい」と答えはしたが、本当に動き出す気持ちがあるのだろうか。

現行の地方公務員法は消防職員が、勤務条件の維持改善と自治体当局に対する交渉を目的に団体を結成したり、加入することを禁じている。だが、一方で日本政府は一九六五年、結社の自由及び団結権の保護に関する条約（ILO87号条約）を批准した。この際政府は、警察とともに消防職員は国内法により団結権を禁止できるとし、今日に至った。（以下略）

上の文章で、字体の異なる部分（団結権、大都市など）は、主題・焦点の修飾部を示している。現在は、主題・焦点の直前の名詞をとるという単純なアルゴリズムで抽出しているので、副詞的な語句や、動詞の目的語などが誤って抽出される場合がある。

3.2 キーワードの抽出結果の比較

上記の社説に対するキーワード抽出結果を、従来手法（表題を用いる場合と、表題を用いないで意味分類番号の一致のみでグループ化したものをキーワードとした場合）と本手法について表2に示す。

表2 キーワード抽出結果の比較

従来手法 (表題使用)	消防、消火、幹部二人、首相、 団結権
従来方法 (表題不使用)	出初め式、九一年初夏、気持ち、 心、姿勢、経緯、禁止方針、 先進国、地方公務員法、場、 幹部二人、首相、日本政府、 政府、状況、事情、交渉、 関係、理由、団結権、災害、 地震、大火、火災件数、 木造家屋、救急活動、消防、 消火、事態、必要、肝心
本手法 (上位15語)	団結権、首相、消防、次長、 消防業務、幹部二人、救急活動、 消火、消防団、任務遂行、 地方公務員法、政府、日本、 場、都市

従来手法で表題との関連性を用いた場合には、重要な語は抽出されているが、抽出される語の量が少なく、すべてのキーワードが網羅されていないという欠点がある。

一方、表題を用いないで、類似の分類番号を持つ語をグループ化した場合には、多くの語句を抽出することができるが、表からも分かるように、抽象的な語が比較的多いし、キーワードとしては数が多過ぎる。また、いずれの場合にも、どのキーワードが重要度が高いかが不明である。たとえば、人間がキーワードを抽出した場合に上位に来ることが予想される「消防」という語と、余り重要ではない（ストップワードになることもある）「場」とでは、この手法ではどちらが重要かは決定できない。

本手法では、点数を計算してキーワード候補を抽出する。ここでは10点以上の15語を示してある。「団結権」が最も高い得点（28点）を示し、以下「首相」(21)、「消防」(20)と続く。さらに主題・焦点の修飾部も考慮することにより、「消防業務」、「消防団」などの語が抽出されている。なお、この表で「次長」は本来「マイヤー次長」となるべきであるが、本手法では、「固有名詞＋一般名詞」という形の複合名詞を分離しているので、表に示すような結果になっている。

4 おわりに

主題・焦点と、それらの修飾部を用いてキーワード候補を抽出し、さらにそれらに対して分類番号を付与してグループ化し、スコアを計算することによってキーワードが抽出できることを示した。表題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、表題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。また、この手法を用いれば、長い文章に対しても、余り多くのキーワードを取り出さずに設定ができる。現在、いろいろな長さ、スタイルの文章についてこの手法を適用することを計画している。

上で述べたアルゴリズムの部分は自動化されているが、プログラムのつなぎの関係上、現在のシステムでは、主題・焦点とそれらの修飾部の抽出、分類番号の付与の部分までをすでに抽出されたものとしてシステムを構築している。主題・焦点およびそれらの修飾部の抽出については、前提条件である「正しい構文解析が行われていること」を緩めて、形態素解析と、動詞の結合価、付随する名詞の意味素性を用いた、より簡便な手法で自動化することを試みている [5]。また、意味分類番号の付与も、単語のデータベースのようなものを作って付けることを検討中である。これらが自動化されれば、システムとしてキーワードが自動的に抽出できるようになる。ただし、条件を緩和したために、主題・焦点や修飾部の抽出が不完全になるおそれがあり、この改良も課題である。修飾部については、ごく簡単なアルゴリズムしか用いていないので、上の例でも誤りが多い。これも今後改良の予定である。

上の例では上位 15 語句をキーワードとして採用したが、スコアの点数配分の検討、スコア計算の閾値設定の有無、キーワードの数の設定の検討なども今後に残された課題である。さらに研究を続けたい。

参考文献

[1] 横山 晶一・小谷 郁夫：主題・焦点と表題との関連性を用いたキーワードの抽出、言語処理学会第 6 回年次大会論文集 A3-3 (2000) pp. 245–248

[2] 菅野 崇：主題・焦点のスコアを用いたキーワード抽出、山形大学卒業論文 (2001)

[3] 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)

[4] 形態素解析システム 茶筌 Ver.2.02、奈良先端科学技術大学院大学 (2000)

[5] 廣町 潤：形態素解析を用いた主題・焦点解析システム、山形大学卒業論文 (2001)

[6] 池原 悟他編：日本語語彙大系、岩波書店 (1997)