主題・焦点の意味グループ化によるキーワード抽出

菅野 崇　横山 晶一　西原 典孝
山形大学工学部

1．はじめに

世界は様々な文書で満ち溢れており、我々がその中から必要なものを取り出すには非常な困難を伴う。こうした傾向はネットワークが発達した現在さらに顕著なものとなっており、必要な情報を取り出す手段がますます重要となっている。情報を取捨選択し、必要な情報を選び出す手段には、求める情報の検索のためのキーワード抽出や文章の自動要約[1]などがある。キーワード抽出や自動要約の方法としては、語の頻度情報を用いたり、表題との関連性[2]を用いるものがあるが、文章の談話構造をある程度考慮しないと、関係のない語が抽出されたり、重要な語句が欠けてしまう場合がある。

我々はすでに日本語の文章の談話推移に重要な役割を果たしている主題・焦点を用いてキーワード抽出をおこなった[3]。その結果、表題との意味的類似性をみることで主題・焦点がキーワードとして有効であることが分かった。しかしながら、主題・焦点以外の重要情報の欠落や、意味的類似性をみる際に抽象的な語句がとりだされやすいという問題点があった。

本報告では、これらの問題点を改善した新しいキーワード抽出方法を提案する。また、従来方法では、処理の一部に人手による作業をはさむため十分な検証がおこなえなかったが、本報告では全ての処理を自動化したシステムとなっており、種々の応用が可能である。

2．抽出方法

本報告によるキーワード抽出方法について述べる。キーワード抽出の流れは図1のようになる。以下ではこの処理の流れに従い、抽出方法を解説する。

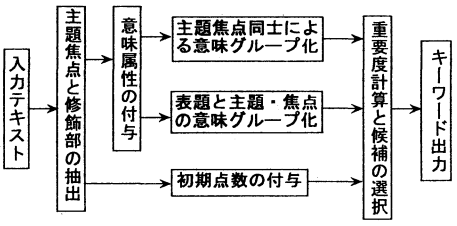

図1　キーワード抽出の流れ

2．1　主題・焦点と修飾部の抽出

1）主題・焦点の抽出

主題・焦点に関しては、今まで様々な定義[4]が提案されてきたが、従来方法と同様に次のように定義する。

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

この定義に基づき、抽出アルゴリズム[5]を用いて主題・焦点を抽出する。これは、文の述語の形から、動詞文、形容詞文、名詞文の3つに分類し、それらと文中の助詞の情報（「は」格、「が」格など）から主題・焦点を抽出するというものである。しかしながら、構文解析が正しくおこなわれているという重い前提条件があったため、従来方法ではこの部分を手作業でおこなっていた。今回この条件を緩め、抽出アルゴリズムに基づいて形態素解析と動詞の分類から抽出する方法[6]を用いた。形態素解析には茶筌[7]を使用している。

2）修飾部の抽出

次に茶筌の形態素解析結果を用いて主題・焦点の修飾部を抽出する。従来方法では主題・焦点の助詞「の」格の直前の名詞のみを修飾部とするものであった。今回、助詞との関係から再帰的に処理することで、より柔軟に修飾部を抽出する。

修飾部の抽出は、文内における主題・焦点の前方に対して次の並びの部分に対しておこなわれる。

[—a—] [対応助詞] [主題・焦点]

a:「名詞」もしくは「動詞の連体形＋名詞」

このとき対応助詞が一致すれば、a を修飾部として抽出する。対応助詞は連体化「の」、並列助詞「と」「や」など（茶筌の分類による）である。主題・焦点が「動詞の連体形＋名詞」の場合には、格助詞「が」、「に」なども対応助詞とする。また、このとき後述の重要度付加で用いる初期点数を付与する。まず、主題・焦点に初期点数（主題：1.5 点、焦点：1.0 点）を与える。その次に修飾部語句 a の初期点数を「主題・焦点の初期点数×対応助詞の係数」で与える。対応助詞の係数は「の」：×0.6、「と」：×0.7、「が」：×0.5、「で」：×0.4

などである。

また、a が修飾部語句として抜き出された場合、さらにその前方に対して次の並びをみて修飾部抽出を試みる。

[—b—][対応する助詞][—a—]

a, b:「名詞」もしくは「動詞の連体形＋名詞」

対応助詞は上記と同じである。対応助詞が一致すればbを修飾部語句として抜き出す。この際、初期点数は、修飾部語句 a の点数に対応助詞の係数を掛けたものとなる。b が抜き出された場合、同様にさらに前の並びを判定して修飾部を抽出していく。このように主題・焦点の前方に対して再帰的に処理が行われ、修飾部が抽出される。修飾部の初期点数は主題・焦点から遠ざかるほど少なくなる。

以下の例では、二重下線が主題、下線が焦点、波線が修飾部でその下に初期点数が示してある。なお、対応助詞が格助詞で抽出された修飾部の初期点数は、直前にある「動詞の連体形＋名詞」（以下では「共通して…」）の点数を参照するため、以下のようになっている。

DNAによる親子鑑定の専門会社がパンフレットで共通して
0.14　0.45　0.75　0.6　1.5
述べているのは「真実を知る権利」だ。
0.4　1.0

（日本経済新聞 1998　DNA 親子鑑定、個人の利用広がる、遺産相続からむ）

２．２　意味属性の付与

主題・焦点と修飾部（以下、主題・焦点語句）及びタイトルの名詞に対して日本語語彙大系[8]を参照して意味属性を付与する。従来方法では人手により意味属性を付与していたが、今回、自動処理[9]によって付与する。意味属性は付与する名詞が直接含まれる意味分類番号と全ての上位概念分類番号を付与する。また１つの名詞に複数の意味属性が存在する場合にはその全てを付与する。複合名詞では、語彙大系で参照できる最小単位の名詞に分割して付与する。2 つ以上に分割された場合、2 つまでは全ての名詞の意味属性を付与するが、３つ以上に分割された場合には、複合名詞の先端と終端の名詞に付与する。

都市郊外林
[1 名詞/2 具体/388 場所/458 地域/464 行政区画]
[1 名詞/2 具体/388 場所/458 地域/465 都市]
[1 名詞/2 具体/388 場所/468 自然/469 地勢/510 景観/511 地相/512 地相(自然)/513 森林]

この例では「都市」と「林」の２つに意味属性が付与されている。

２．３　意味のグループ化

意味属性の付与された主題・焦点語句に意味のグループ化をおこない、これをもとに点数を付与する。意味のグループ化は、主題・焦点語句同士によるグループ化、表題と主題・焦点語句とのグループ化よっておこなわれる。

１）主題・焦点語句同士による意味グループ化

主題・焦点語句に対して付与された意味属性をもとに意味のグループを形成する。全ての主題・焦点語句から、任意に２つの語句を取り出し、意味分類番号を比較する。意味分類番号が一致した場合、それらを同じ意味のグループとする。また、意味分類番号が一致しない場合でも、２つの語句の上位分類番号が＜具体＞である場合にのみ、親分類番号を参照しグループ化を試みる。これを全ての語句で総当り的におこなう。２語の一致で個々にグループを形成するのではなく、一致したものを以前のグループに追加してより大きなグループを形成していく。また、上位分類番号＜抽象＞に属する語は意味範囲が比較的広く分類されているため、親分類番号によるグループ化形成はおこなわない。図２に「新幹線」「地下鉄」「高速道路」の語句でのグループ化例を示す。

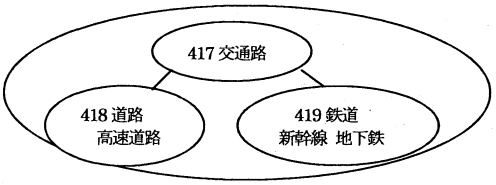

図２　意味のグループ化

まず、「新幹線」と「地下鉄」に関して意味分類番号＜419 鉄道＞が一致しグループを形成する。また、これらと「高速道路」の３語は、共に上位分類番号＜具体＞に属しているので親分類番号＜417 交通路＞を参照して、グループを形成する。

次に、形成されたグループをもとに、グループ内の語句に点数を与える。グループ内の語句への点数の与え方は、親分類番号・意味分類番号による一致と、属する上位分類番号によって変化する。点数の与え方は表１のようになる。

表1　主題・焦点によるグループ化点数

	親番号	意味分類番号
具体	+1.0	+2.0
抽象物		+1.2
事		+1.0
抽象的関係		+0.8

意味分類番号の一致による点数では、上位分類番号<具体>、<抽象物>などの具体性・抽象性の判断から点数を細分化する。また、<具体>に属する語句では、親分類番号が一致するよりも意味分類番号が一致する方が重要性が高いとして、以上のように点数を与える。

たとえば上記の「新幹線」「地下鉄」「高速道路」のグループでの点数は、「新幹線」では「地下鉄」と意味分類番号の一致で2.0加点、さらに「地下鉄」と親分類番号が一致して1.0加点され、計3.0点となる。「地下鉄」に関しても同様に計3.0点となる。これに対し「高速道路」では「新幹線」「地下鉄」と親分類番号での一致のみとなり、計2.0点となる。このように同じグループ内の語句でも意味の近接関係で点数が変化する。

また、このグループ化の処理の際に、同一語句の出現を考慮して点数を与える。同じ語句が主題・焦点として出現するごとに1.0加点する。グループ化されない固有名詞に関しても、同一語句の出現を考慮して主題・焦点として出現するごとに2.0加点する。

2）表題と主題・焦点語句とのグループ化

主題・焦点語句同士のグループ化と同様に、表題の語句ともグループを形成し点数を付与する。全ての主題・焦点及び修飾部と表題・小題の語句で親分類番号と意味分類番号を比較してグループ化する。形成された表題・小題グループ内の語句に対して点数を表2のように付与する。

表2　表題とのグループ化点数

	親番号	意味分類番号
表題	+5.0	+8.0
小題	+2.0	+3.0

表題とのグループ化は全ての主題・焦点語句に対しておこなわれるが、小題とのグループ化はその小題が適用される範囲内の語句とでおこなう。また、グループ内の語句が表題・小題語句と完全一致した場合には、より重要度が高いとしてさらに2.0加点する。

2．4　重要度計算とキーワードの選択

主題・焦点1語当たりの重要度を計算する。主題・焦点語句xが、他の主題・焦点語句 $w_1,w_2,\ldots,w_n$ とグループ化し、表題・小題 $t_1,t_2,\ldots,t_m$ とグループ化したときの重要度は次の式より計算される。

$$Score_x = I_x + \beta \cdot \sum_{i=1}^{n} S_i + \sum_{j=1}^{m} T_j$$

I_x：xの初期点数
S_i：xと w_i の主題・焦点グループ化の点数
T_j：xと t_j との表題グループ化の点数
β：補正係数

βは、文書によって主題・焦点数が変化するため、主題・焦点グループ化の点数が増加しすぎるのを抑える係数である。

この式を用いて全ての主題・焦点語句に重要度を与え、重要度の高い上位語句をキーワードとして選択する。キーワードの選択数は、キーワードの利用目的によって異なるが今回上位15語を選択した。

3．実験と評価

実験対象には、新聞記事や雑誌記事を用いた。以下に新聞記事での抽出例を従来法と比較しながら述べる。

3．1主題・焦点と修飾部の抽出

前節の方法により主題・焦点と修飾部を抽出する。以下にその一部を示す。二重下線が主題、下線が焦点、波線が修飾部である。

『CO2削減を競う温暖化防止ブエノスアイレス会議（1）戦略資源生む“トヨタの森”』（日本経済新聞1998年記事）

地球温暖化防止ブエノスアイレス会議(ＣＯＰ４)が始まった。昨年の京都会議(ＣＯＰ３)では二酸化炭素(ＣＯ２)など温暖化ガスの排出量削減の国別目標が決まり、ブエノスアイレス会議では削減のための具体的なルール作りが焦点になる。新たな国際的な枠組みづくりをにらみながら、企業は地道な省エネ努力を重ね、排出権取引など新制度への参画を模索する動きも出てきた。生き残りをかけてＣＯ２削減を競う企業の姿を追う。

愛知県豊田市の郊外に、トヨタ自動車が所有する実験林「トヨタの森」がある。九三年に三ヘクタールでスタートした実験林は九六年には十五ヘクタールに拡大された。小高い山の傾斜地や湿地帯を順路通りに散策すると優に一時間はかかる。(以下省略)

3．2　キーワード抽出結果の比較

先ほどの記事に対するキーワードの抽出結果を以下に示す。また、比較として従来法によるものを以下に示す。

本報告により抽出されたキーワード

トヨタ、実験林、戦略資源、ブエノスアイレス会議、森、計画、ＣＯ２削減、都市郊外林、樹木 、森林、京都会議、再生策、植林事業、海外植林、四倍体樹木

従来法により抽出されたキーワード

実験林、ブエノスアイレス会議、戦略資源、森、京都会議 、計画 、森林、トヨタ 、国別目標 、区域 、実績 、影響 、条件 、植林事業 、海外植林

キーワードは本報告、従来法ともに重要度の高い順に並べてある。従来法の結果では、キーワード上位の語句に表題関連の語句が抽出されているが、その他の語句では抽象的な語句の出現が多くなっている。これは従来法が意味属性の親分類番号のみを用いて、比較的広い意味範囲からグループ化していたことによる。たとえば「実績」、「影響」などでは意味的に抽象的な割合が多く、文章を特徴づける語句にはあまりふさわしくないと思われる。一方、本報告の結果では、従来法と同様に上位語句では表題と関連性のある語句が抽出されている。しかしその他の語句には不要な抽象語句の出現が押さえられている。表題語句であり、主題・焦点として最も多く出現した「トヨタ」(固有名詞) が最上位となっており、その次に表題語句「森」とそれに関連して「森林」 樹木」など意味グループ化によって抽出されている。従来法と比べ、適切に意味のグループ化がおこなわれていることが分かる。

4．おわりに

本報告では、文章から主題・焦点と修飾部を抽出し、それらを意味グループ化することでキーワードが抽出できる方法を示した。表題語句との一致を用いたり、頻度情報を用いたり場合と比べて、表題語句と意味的に関連性があるキーワードが適切に抽出されているのが分かる。

今後の展望として、様々な条件での評価実験をおこなっていく予定である。被験者による正解キーワードとの比較、文章の種類を変更した上での評価など、実験を重ねていきたい。また、本報告によるキーワード抽出の応用として、キーワードを用いた重要文抽出[10]の実験もおこなっている。これとの結果を踏まえた上でキーワードを評価していきたい。

今回、設定キーワード数、キーワードを抽出するためのグループ化点数などの各種パラメータはいくつかの試行からこのように決定したが、今後、文書の種類・長さで動的に決定することを考えている。評価実験の結果と併せて検討していきたい。

参考文献

[1] 奥村　学・難波英嗣:テキスト自動要約に関する研究動向, 自然言語処理, Vol6, No6,(1999) pp. 1-26.

[2] 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)

[3] 横山晶一・菅野 崇:主題・焦点のスコアを用いたキーワードの抽出、言語処理学会第７回年次大会論文集 B3-2 (2001) pp.177-180

[4] 野田尚史:新日本語文法選書１「は」と「が」、くろしお出版(1996)

[5] 吉田悦子・横山晶一:主題・焦点を用いた文脈解析の一手法、情報処理学会自然言語処理研究会資料 NLC97-29(1997)

[6] 廣町 潤:形態素を用いた主題・焦点抽出システム、山形大学修士学位論文(2002)

[7] 形態素解析システム「茶筌」、奈良先端科学技術大学院大学

[8] 池原 悟他編:日本語語彙大系、岩波書店(1997)

[9] 阿部亮介:シソーラス自動獲得システムの構築に関する研究、山形大学卒業論文(2001)

[10] 菅野崇、主題・焦点によるキーワード抽出とそれを用いた自動要約、山形大学修士学位論文(2002)