主題·焦点を用いた要約文の抽出

細梅久典† 横山晶一‡

†山形大学大学院 理工学研究科
‡山形大学 工学部

1 はじめに

近年、電子化されたテキストが世の中に満ち溢れ、自動要約技術などにより読み手が読むテキストの量を制御できることが求められている。

本研究では、要約のための手法として、文の主題·焦点を用いた方法を提案する。主題と焦点抽出アルゴリズムはすでに提案されている [1, 2, 3]。このアルゴリズムは新聞記事などの短い文に対しては効率的であるが、論説文などの長い文では、必ずしもうまくいくとは限らない。ここではこの方法を長い文に対しても適用できるように修正して用いる。

そして、主題·焦点の修飾関係や文間のつながりから、"ブロック" を提案する。文間の特徴を捉えた "ブロック" を決定しそのブロックのタイプに沿った重要な文を要約文として抽出する。本研究で提案する "ブロック" は、接続詞、文面上のサブジェクト、疑問文から決定される "大ブロック"、先に述べた "主題·焦点" に関連する修飾語、頻出語などにより決定する "中ブロック" がある。例えば、"大ブロック" 内の疑問文では、疑問文と、それに対する答えに相当する文を判定し、両者を要約文として抽出する。このように判定した文をまとめた形で、要約文を生成する [4]。この結果、ブロックごとの重要文抽出により、要約文の不自然さを解消する文抽出が可能になった。

2 要約文抽出アルゴリズム

ここでは、主題·焦点を用いてブロック化をはかる手法、重要文を特定する方法について述べる。要約を行う過程は図 1のようになっている。

2.1 主題·焦点抽出アルゴリズム

従来 [1] と同様以下のように主題·焦点を定義する。

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

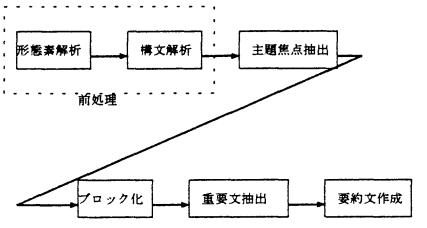

図 1: 要約文抽出過程

焦点 : その文で新しく導入された情報

この定義は、主格名詞がまだ知られていない新情報のときはその主格に「が」がつき、主格名詞が既に知られている旧情報のときはその主格に「は」がつくという原理である。

この定義に従って主題·焦点を抽出するアルゴリズムを概略以下のように定める。また、処理にあたっては、次のような前提条件を設ける。

1. 格関係をもたない文は対象外とする。
2. 曖昧性のない構文解析木が作られているものとする。

主題および主題の修飾語の抽出方法

以下では、文を述語の種類により名詞文、形容詞文、動詞文に分ける。

(1) 主題を表す「は」が存在する場合

「は」のつく直前の名詞を主題とする。それとともに、各々の修飾語を主題の修飾語として登録する。

(2) 主題を表す「は」が存在しない場合

名詞文では、「が」がある場合、述語名詞を主題とし、それに伴う修飾語を主題の修飾語として登録する。

「が」がない場合は、主題なしとする。形容詞文、動詞文については、いずれの場合も主題なしとする。

焦点および焦点の修飾語の抽出方法

名詞文では、「が」がある場合、「が」格要素を焦点とし、ない場合は、述語名詞を焦点とする。抽出した焦点の修飾部を登録する。

形容詞文、動詞文では、述語を修飾する「が」格要素、なければ、述語の直前の必須格要素を焦点とする。その格要素の修飾部を焦点の修飾部として登録する。

このアルゴリズムによって抽出される主題・焦点の例を以下に示す。太字下線は主題、下線は焦点を示す。

> 確かに、サンゴは環境ストレスに対して非常に敏感な生物である。
> (Newton 1999.6 Hot Topic 「地球温暖化がサンゴを襲う」)

2.2 ブロック化と重要文抽出

"ブロック"には、まとまりの大きさの順に大ブロック、中ブロック、リンク(小ブロック)がある。以下に各ブロックの特徴と、ブロックから抽出される重要文について述べる。

(1) 大ブロックと重要文

大ブロックは、「転換の接続詞(さて、次に、ところで、では)、疑問文、文面上のサブジェクト[1]」から決定する。以上のような手掛かり語により、著者は、明らかに意図的に話題を変更すると考えられる。

大ブロックからは、次のように、疑問文に関係する部分のみを重要文として抽出する。

大ブロック中の重要文抽出アルゴリズム

- 疑問文を抽出する
- 疑問文の主題を文の主題に挙げている文を要約文として抽出する。ない場合は、疑問文の次の文を抽出する。

(2) 中ブロックと重要文

中ブロックは、大ブロックの中のもう少し小さなまとまりとして定義される。中ブロック決定のために、主題・焦点、各修飾語、ブロック中に現れる単語の頻度情報[2]を用いる。

中ブロックは以下から決定する。

1. 高頻出語を主題の修飾語にもつもの
2. 主題、焦点、各修飾語に連続する数詞、時相名詞をもち同一主題の組合せ
3. 局所的頻出語 [10, 11, 12] を主題、焦点、各修飾語にもつブロック

1は、最も頻出度の高い語が主題の修飾語に現れた文から大ブロックの終わりまでを中ブロックと定める方法である。2は、「第1の○○は…、第2の○○は…」といった表現の文をブロック化する方法である。「第1の○○は…」という文章が始まり、連続する数詞を主題、焦点、各修飾語に現れる文が大ブロックが終わるまで複数出現するような場合、「第1の○○は…」から大ブロックの終わりまでを中ブロックと定める方法である。3は、局所的頻出語が大ブロック内で出現する文から出現し終わる文までを中ブロックと定める方法である。

1に相当するブロックの重要文抽出アルゴリズムとその重要文の例を以下に示す。

1に相当するブロックと重要文

- このブロックに対する重要文は主題の修飾語に高頻出語を含む文である。その文の次の文が主題を持たない文の場合、次文も抽出する。

> 16:ADHDの症状には、大きく分けて2つある。
> 17:不注意、多動および衝動的な行動の複合である。
> ⋮
> 22:男児のADHDの発症率は、女児に比べて少なくとも3倍であることが知られている。
> 23:ある研究では、男児の発症者は女児の発症者の9倍とされている。
> ⋮

この例では、

第16文 ADHDの症状には、大きく分けて2つある。

第17文 不注意、多動および衝動的な行動の複合である。

第22文 男児のADHDの発症率は、女児に比べて少なくとも3倍であることが知られている。

が抽出される。

[1] 文面上のサブジェクトとは、文の途中に挿入された表題などである。

[2] 頻度情報は形態素解析を行った後、ストップワードを削除し、その後出現する単語をカウントすることで得る。

(3) リンク (小ブロック) と重要文

文は、なんらかの単語によって関連をもっている場合が多い。キーワードの主題、焦点、もしくは各々の修飾語により、文が関連をもつ場合、以下に述べる焦点-主題のリンクから重要文を抽出する。前後2文からなる、主題、焦点間の関係は以下の図2の4通りである。図2のcは、n番目の文の焦点とn+1番目の文の主題

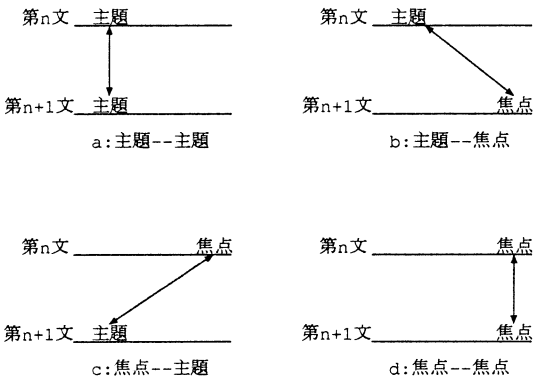

図 2: 主題-焦点関係図

が一致している場合である。このリンクは、話題が移行している場合であり、先頭文は話題提示の文で、その話題の説明が次の文から続く傾向が強い。

図2のcの場合を重要文として抽出する理由は「現象文」と「判断文」という観点[7][3]からである。

「現象文」中の「が」格は、本論文における「焦点」に相当し、「判断文」の「は」格が本論文における「主題」に相当する。この理論を本要約文抽出アルゴリズムに導入した。リンクによる重要文抽出アルゴリズムとその例を以下に示す。

- 第n文の焦点、もしくは焦点の修飾語が高頻出語で、かつ第n+1文の主題もしくは主題の修飾語がそれと一致する場合、第n文と第n+1文を重要文として抽出する。ただし、局所的頻出語をもつ中ブロック(3の場合)で、局所的頻出語が一致する場合のみ、第n文、第n+1文を重要文として抽出する。

第70文: おそらく、脳のドーパミンの使い方に関係するものであろう。

第71文: ドーパミンは、脳の特定の領域にあるニューロンから分泌され、他のニューロンの働きを抑制したり調節したりする。

第72文: たとえば、運動機能に障害が現れるパーキンソン病は、大脳基底核の下部にある黒質と呼ばれる部分のドーパミン放出ニューロンが死ぬことが原因で発病する。

⋮

上の例の第70文では、ありのままの現象を焦点(新情報)で伝える表現となっており、第71文以下では、"第70文の焦点"を「は」格で判断、解説するような文の流れになっている。上の例からは、第70文、第71文が重要文として抽出される。

現段階では、様々な語に関して、4通りのリンクがどの程度重要であるかは明らかになっていない。焦点-主題の関係のみ顕著であるため、このリンクによる要約文抽出を採用する。

以上のように、各々のブロックから重要文が抽出される。そして、それぞれのブロックからの重要文を全てまとめた形で要約文を作成する。要約文作成については、現段階では特別な処理は行われない。

このアルゴリズムで、本研究で対象とした論説文に対して抽出処理を行った結果を以下に示す。

3 要約文抽出結果

本アルゴリズムにより抽出した日経サイエンスの対象データに関する要約文の一部を図3に載せる。文数は本文の約25%程度に減少した。

7:キースは「注意欠陥多動性障害(ADHD)」だった。
16:ADHDの症状には、大きく分けて2つある。
17:不注意、多動および衝動的な行動の複合である(次ページの表)。
22:男児のADHDの発症率は、女児に比べて少なくとも3倍であることが知られている。
25:ADHDに典型的な行動パターンは、通常3歳から5歳の間に現れる。
44:この10年の画像処理による研究から、ADHDの子供でうまく働いていないと思われる脳の領域がわかってきており、それをもとにしてADHDの症状を説明できるようになった。

⋮

図 3: 要約文抽出例

要約文に関しては、それほど違和感のない文が抽出されていると考えられる。

[3]現象をありのまま判断の加工を施さないで心に移ったままをそのまま表現した現象文は「が」が使われる。それに対して、話し手の主観が判断を下しす「判断文」は、「は」が使われる。

4 おわりに

本研究は、従来 [1] の主題·焦点抽出アルゴリズムを改良し、論説文における主題·焦点の抽出を可能にした。そして、主題·焦点を用いた従来にない新しい手法である“ブロック”単位の要約文抽出方法による要約文を作成した。そのことで、文章表現の特質に合わせた文抽出が可能になり、要約文の不自然さを解消する文抽出ができるようになった。また、照応詞の補完を行わずに、主題の修飾語、主題·焦点中に現れる語の情報を利用して、文章をブロック化することができた。ブロックの中から不自然さを回避できるよう重要文を抽出しているため、ブロック内での文のつながりにほぼ違和感のないものを抽出できるものとなった。

ブロック化·重要文抽出の問題点としては次のようなものがある。

大ブロックの重要文抽出: 疑問文の主題が複数ある場合、重要文の抽出に工夫が必要

中ブロック 1の重要文抽出: 主題、主題の修飾語に高頻出語がともに現れた場合の処理

中ブロック 3の重要文抽出: 局所的頻出語としてどの程度のばらつきを許すか

リンクの重要文抽出: 現象文にあたる文の述語に関しタイプ分類のためさらなる研究が必要

アルゴリズム全般: アルゴリズム充実のため、さらに多くのテキストに対して実験と考察が必要

要約文抽出アルゴリズムは、話題の散在するテキストに対し大きな効力を発揮することがわかっている。したがって、今後は一般的な論説文データに対して応用できるように、さらに実験を重ねる予定である。

また、本抽出アルゴリズムはリンクを用いるため、ユーザーの知りたいキーワードに関する語のリンク情報を含んだ重要文を抽出することで情報検索としても利用価値のあるものである。今後は、要約文抽出のみならず、キーワード関連文抽出にも汎用性をもつアルゴリズムについて検討の予定である。

参考文献

[1] 吉田悦子： 主題·焦点の抽出と文脈構造ネットワークの構築. 山形大学大学院工学研究科修士学位論文 (1998).

[2] 吉田悦子. 横山晶一： 主題·焦点の抽出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).

[3] 吉田悦子. 横山晶一. 西原典孝： 主題間の関係を用いた文脈解析の一手法. 電子情報通信学会技術報告. 情報処理学会自然言語処理研究会報告NL124-3 (1998).

[4] 細梅久典： 主題·焦点を利用した要約文抽出に関する研究. 山形大学大学院理工学研究科修士学位論文. (2000).

[5] 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. Vol.6. No.6.pp.1–26 (1999).

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

[7] 三尾 砂: 國語法文章論. 三省堂(1948).

[8] 仁田 義雄: 現象描写文をめぐって「日本語学」. 5-2 p.56–p.69 明治書院(1986).

[9] 日本語形態素解析システム「茶筌 (ChaSen) version 1.51」.

[10] Marti A. Hearst： TextTiling Segmenting Text into Multi-paragraph Subtopic Passage. Computaional Linguistics.Volume23(1). pp.33–65 (1997).

[11] 仲尾 由雄、文書の意味的階層構造の自動認定に基づく要約作成. 言語処理学会第 4 回年次大会ワークショップ論文集、pp.72–79 (1998).

[12] 仲尾 由雄： 語彙的結束性に基づく話題の階層構成の認定. 自然言語処理. Vol.6. No.6.pp.83–112 (1999).