主題・焦点リンクを用いた重要文抽出システム

横山晶一†　　菅野崇‡　　西原典孝†
†山形大学工学部情報科学科
‡山形大学大学院理工学研究科（現・福島コンピュータセンター）
{yokoyama, nisihara@yz.yamagata-u.ac.jp}

数多くの文章の中から情報を抽出したり、文章要約のために重要文を抽出したりする手法は、自動化が望まれ、また、さまざまな研究手法が提案されている分野である。我々は、すでに、主題・焦点を用いたキーワード抽出システムを作成して、抽出されたキーワードが有効であることを確認した。

本研究では、この手法で抽出されたキーワードを用いた重要文抽出システムについて述べる。抽出されたキーワードを文の重み付けに用いる手法は、従来類似の研究がいくつかある。このシステムでは、要約する文章の中での重み付けされた文の割合を基本要約率として指定し、それらの文と、主題・焦点リンクを通じて抽出される文とで相補って重要文を抽出するところが従来研究とは異なる。本稿では、重要文抽出結果を示すとともに、小規模な評価も行ってこの手法の有効性を示す。

An Extraction System of Important Sentences using Theme-Focus Link

YOKOYAMA Shoichi†　KANNO Takashi‡　NISHIHARA Noritaka†
†Department of Informatics, Faculty of Engineering, Yamagata University
‡Graduate School of Science and Engineering, Yamagata University
(Now, Fukushima Computer Center)

Information extraction of some sentences and extraction of important sentences are requested for automatic processing, and proposed with many methods and systems. We have already proposed a keyword extraction system using themes and focuses, and confirmed the effectiveness of them.

This paper proposes an extraction system of important sentences using the keywords extracted from the above method. There are some studies with keyword weighting for important sentences. Uniqueness of our system is that weighted sentences are specified as the fundamental summarization rate, and that related sentences with theme-focus link from these sentences are derived. We show the results of important sentence extraction with different types, and also show the effectiveness of this method with results of human evaluation.

１．はじめに

ネットワーク上における情報氾濫は、すでに人間が手作業で対処できる限界を超えている。必要な情報を的確に選択するためには、情報検索や抽出のためのキーワードを自動的に抽出することや、文書の要約、文書からの重要文抽出[1]を適切に行う必要がある。

我々はすでに、与えられた文章からキーワードを抽出するシステムを作成した[2,3,4]。このシステムの特徴は、単語の頻度情報のみならず、文章の談話情報を表す主題・焦点と、文章に付された表題との意味的関連性を重視して、適切なキーワードを抽出しようとするものであった。

本稿では、抽出された主題・焦点、キーワードを用いて、文の重要度を定め、重要文を抽出するシステム[2]について述べる。その際、重要文の抽出の割合と、重要文と主題・焦点を通じて関連のある文の割合とを指定することにより、使用者側のさまざまな要求に合致した、性格の異なる重要文が抽出できることを示す。

２．システムの概要

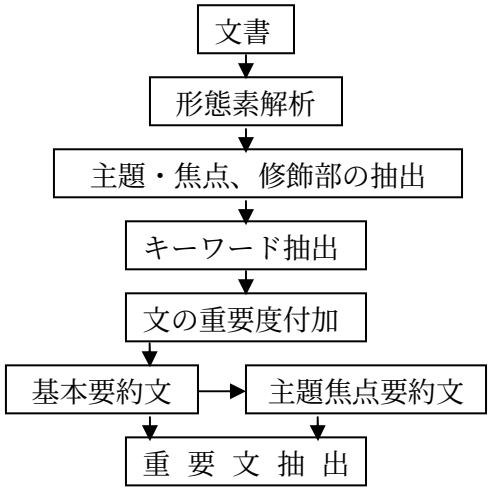

図１　システムの概要

図１にシステムの概要を示す。入力された文章は、まず形態素解析され、そこから、主題・焦点とその修飾部が取り出される。これをもとにキーワード抽出処理が行われる。抽出された主題・焦点とキーワードを用いて、基本要約文が抽出され、それらの文との主題・焦点の関連性を考慮した主題・焦点関連文がさらに抽出される。これらの文を並べることによって、重要文抽出が行われる。

３．キーワード抽出

キーワード抽出については、すでに発表している[3]ので、ここでは簡単に解説する。

まず、与えられた文書を形態素解析（茶筌[5]を用いる）して、その結果から主題・焦点を抽出する[6,7]。主題・焦点の定義は、従来我々の用いている定義[8]と同じで、主題は、「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」、焦点は、「その文中で新しく導入された情報」である。さらに、主題・焦点に対する修飾部を抽出する。これらの主題・焦点、修飾部に対して、初期点数を与える。主題の初期点数は 1.5 点、焦点は 1.0 点、修飾部は、これらに接続する格助詞に応じて、係数をかける。

抽出された語に対して、日本語語彙大系[9]による意味分類番号とすべての上位概念分類番号を、シソーラス自動獲得システム[10]を用いて付与する。この番号をもとに、同じ分類番号や、同じ上位分類番号を持つ語をグループ化する。グループ化された語群には、単独の語よりも高いスコアが与えられる。

また、これらの語は、表題や小題とのつながりも大きいので、表題、小題の語句ともグループ化を行って、これらに高いスコ

アを与える。

主題・焦点１語当たりの重要度は、ある語句 xが、他の主題・焦点語句 $w_1,w_2,\ldots,w_n$ とグループ化し、さらに表題・小題 $t_1,t_2,\ldots,t_m$ とグループ化したときに次のように計算される。

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

ここで、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による点数、β：主題・焦点の増減による補正係数である。

４．要約の生成

抽出されたキーワードをもとに、重要文抽出を行う。キーワードは文の重み付けに用いる。ここでは、主題・焦点リンクを活用することで、従来とは異なる重要文抽出法を行う。

4.1 用語の定義と要約率の関係

・基本要約文

キーワードによって重み付けされた文である。この文の量を**基本要約率 p(%)**で指定する。

・主題・焦点要約文

基本要約文の内容を補う文として、基本要約文から主題・焦点リンクを通じて抽出される。この量は、**主題・焦点要約率 q(%)**として指定する。

・全体要約率 g(%)

出力する要約の全体量である。原文の何%に要約するかを指定する。

これらの要約率は、システム開始時にユーザが指定する。入力文書の文の数を N とすると、全体要約率 g%で指定される文の量 G は、

G = N×g / 100

となる。また、基本要約文の量 P は、

P = G×p / 100

また、p+q=100 であるから、主題・焦点要約文の量 Q は、

Q = G ×q / 100

G = P + Q

となる。

たとえば、原文が 60 文、ユーザ指定で、全体要約率 30%、基本要約率 30%（すなわち主題・焦点要約率 70%）の場合、全体の文の量 G は 18 文、そのうち基本要約文が５文、主題・焦点要約文が 13 文となる。

基本要約率を小さく（主題・焦点要約率を大きく）すると、抽出される重要文は、最も重要な文を抽出して、それと関連のある文を並べるという形になる。一方、基本要約率を大きく（主題・焦点要約率を小さく）すると、基本要約文が重視される。つまり、原文のいろいろな部分から重要文を集めてきて全体の要約をするという形になる。

4.2 要約生成方法

4.2.1 基本要約文の抽出

まず、入力された原文に対して重要度を与える。原文１文内に出現したキーワードのスコアの総和を文の重要度とする。文の重要度 D は次式で示される。

$$D = (1.0 - \frac{N}{MaxN}) \cdot \sum_{i=1}^{N} Score_i$$

N：文中に含まれる主題・焦点および修飾部の総数

$MaxN$：全文中での主題・焦点および修飾部の最大数

$Scorei$：主題・焦点および修飾部のスコア

すべての文に対してこの重要度を与え、重要度の高い文を基本要約率で指定される文の数だけ上位から選択する。

4.2.2 主題・焦点要約文の抽出

取り出した基本要約文に基づいて、主題・焦点要約文を抽出する。そのために、基本要約文と他の文との間で、主題・焦点リンクを調査して、リンクが結べる文を抽出する。主題・焦点リンクは以下のように分類される。

表１　主題・焦点リンクの分類（一部）

＜前部＞　＜基本要約文＞　＜後部＞
焦点　→主題 焦点　→主題
焦点　→主題修飾部 焦点修飾部→主題 焦点修飾部→主題 焦点　→主題修飾部
焦点修飾部→主題修飾部 焦点修飾部→主題修飾部

上表では、前の方にある文の焦点が、基本要約文の主題になったり、基本要約文の焦点が、後ろの方にある文の主題や主題修飾部になったりすることを表現している。表の中では上段にあるものほど優先される。リンクの強い文を、基本要約文の前後１～２文の中から探し出して、強い順に、指定された数だけの文を抽出する。

もし、この操作で、文の数が足りない場合には、抽出された主題・焦点要約文から、さらに主題・焦点リンクを調査して、つながりの強い文を抽出し、指定の数の文が得られるまでこの操作を繰り返して抽出を続ける。

4.3 重要文抽出結果

上に述べた手法によって、実際の文章に対して重要文抽出した例を付録に示す。システムの出力は、原文 45 文に対して全体要約率 30%（13 文）、基本要約率 30%（主題・焦点要約率 70%）（以下 30:70）のものと、同じく全体要約率 30%、基本要約率 50%（主題・焦点要約率 50%）（以下 50:50）のもの、２種類である。

30:70 では、基本要約文４文を中心として、それらとリンクで結ばれた文とで重要文抽出がなされている。一方、50:50 においては、基本要約文の数は７文で、それらとつながりの深い各１文が付加されたかっこうになっている。

このように、本システムでは、要約率を指定することによって、立場の異なる要約を生成することができる。

５．評価実験

複数の要約に対して、人間の被験者を用いて比較評価を行った。比較方法としては、３名の被験者に原文を提示し、全体要約率（この実験では、上記の重要文抽出と同じ 30%）に相当する数の文を人手で抽出してもらい、システムの出力との一致数を比較するという方法をとった。また、被験者に、システムが出力した重要文を提示して、話題が異なる箇所、つまり要約文の前後で話の内容が異なる箇所に印をつけてもらい、その指定箇所の数の変化を比較した。

以下の表２に、上記の 30:70, 50:50 で出力した重要文と、参考出力として基本要約率 100%、すなわち重要文のみを抽出した結果との比較を示す。また、参考として、市販の、Microsoft Word2000 の 30%の要約率との比較も示す。表３には、話題が異な

る箇所の数を被験者ごとに示す。

表２　被験者の選択した重要文との一致数

	被験者１	被験者２	被験者３
30:70	4(/13)	6(/13)	4(/13)
50:50	8(/13)	7(/13)	8(/13)
100	8(/13)	9(/13)	8(/13)
Word	7(/13)	5(/13)	8(/13)

表３　話題が異なる箇所の数

	被験者１	被験者２	被験者３
30:70	2	2	2
50:50	3	4	5
100	4	6	6

表２、表３から分かるように、被験者により多少の差はあるものの、30:70 の場合には、重要文の一致数が少なく、話題が異なる箇所も少ない。50:50 では、一致数は多いが、話題の異なる箇所も多くなっている。これは、システムの要約率変化を反映していると思われる。本システムと、Word との一致結果を比べてみても、30:70 では一致数が劣るが、50:50 では余り差が見られず、重要文が適切に抽出されていることが分かる。

参考として出力した基本要約率 100%の出力結果は、被験者との一致数は、50:50 よりわずかに多いという結果になった。これは、50:50 の要約に含まれる主題・焦点要約文でも、基本要約文と関連した情報として重要性を持つため、これらが被験者との一致率向上につながったのだと思われる。また、基本要約率 100%では、話題の異なる場所はやや多くなっている。すなわち、話のつながりがやや悪いという結果である。

６．おわりに

本研究では、文章の主題・焦点とキーワードを用いて重要文抽出を行うシステムについて述べた。このシステムは完全に自動化されており、文書入力によって、重要文が抽出される。基本要約率と、主題・焦点リンクを用いることにより、ユーザの意図するさまざまな重要文抽出が行えるようになった。

今後は、システムの改良を行うとともに、重要文を適切につないで、人間に近い要約を出力するシステムにしたいと考えている。

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、Vol.6, No.6 (1999) pp.1-26

[2] 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(2003)

[3] 菅野崇、横山晶一、西原典孝：主題・焦点の意味グループ化によるキーワード抽出、言語処理学会第９回年次大会論文集(2003) pp.393-396

[4] 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(2001) pp.177-180

[5] 形態素解析システム「茶筌」、奈良先端科学技術大学院大学

[6] 廣町潤、横山晶一、西原典孝：形態素解析を用いた主題・焦点抽出システム、情処全大(2003.3)

[7] 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(2003)

[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、情報処理学会自然言語処理研究会資料 NLC97-29(1997)

[9] 池原悟他編：日本語語彙大系、岩波書店(1997)

[10] 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)

付録　システムによる要約例

「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）

下線部は基本要約文、文頭の数字は原文中の文番号。

・全体要約率 30%（13 文）、基本要約率 30%（4 文）、主題・焦点要約率 70%（9 文）

6 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。

7 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。

9 トヨタはここで、都市近郊林の再生策を探るためのフィールドテストを粘り強く繰り返す。

10 実験林はうっそうとした自然林が残る放置試験区、風通しや日当たりが良くなるように低密度に植林された整備区に大きく分かれる。

11 CO2 や酸素(O2)の濃度分布はどうなり、環境の違いで幹が太くなるスピードはいかに変化するのか。

21 トヨタはなぜ、これほど熱心に樹木の研究に取り組むのか。

22 トヨタが新規事業を担当する部署の中にバイオテクノロジーを研究するグループを置いたのは九〇年五月。

23 その数年前から産業界はバイオブームに沸いており、トヨタも微生物や遺伝子組み換え、脳の情報処理システムの解明などの基礎研究を続けた。

25 しかし、意気込みとは裏腹に、つい数年前までトヨタの環境バイオは社会貢献活動の一環に過ぎなかった。

38 植林事業の成否はトヨタのコスト構造にも影響を与える。

39 トヨタはオーストラリアである試みを検討している。

41 トヨタは実験林で遺伝子レベルの解析を続けてきた。

43 「トヨタの森」での地道な取り組みと大規模な海外植林が一本の糸でつながった。

・全体要約率 30%（13 文）、基本要約率 50%（7 文）、主題・焦点要約率 50%（6 文）

2 地球温暖化防止ブエノスアイレス会議(COP4)が始まった。

3 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。

6 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。

7 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。

9 トヨタはここで、都市近郊林の再生策を探るためのフィールドテストを粘り強く繰り返す。

10 実験林はうっそうとした自然林が残る放置試験区、風通しや日当たりが良くなるように低密度に植林された整備区に大きく分かれる。

18 トヨタは遺伝子の集合体である染色体を増加させる技術を応用、様々な「四倍体樹林」を作り出した実績を持つ。

21 トヨタはなぜ、これほど熱心に樹木の研究に取り組むのか。

22 トヨタが新規事業を担当する部署の中にバイオテクノロジーを研究するグループを置いたのは九〇年五月。

29 トヨタは日本製紙、三井物産と共同でオーストラリアでの植林事業に乗り出す。

30 植林用地が手当てできれば、成長が早いユーカリを毎年五百ヘクタール、十年で五千ヘクタール植林する計画。

39 トヨタはオーストラリアである試みを検討している。

41 トヨタは実験林で遺伝子レベルの解析を続けてきた。