主題・焦点ネットワークを利用した要約システム

太田剛史　　横山晶一　　西原典孝†
山形大学大学院　理工学研究科‡

1．はじめに

これまでにも要約の研究は、数多く行われている。その多くは、重要度を重視した文抽出をして要約文を生成するため、文間のつながりが悪くなっていることがあった。そこでネットワークを構築し談話の流れを考慮すれば、従来のシステムより全体を把握した要約ができると考えられる。すでに文のリンク関係を利用したアルゴリズム[1]はいくつかあるが、本研究ではネットワークの構築に有効なものとして、特に主題・焦点に着目しリンクを張った。主題・焦点の抽出[2]、またそれに関する研究[3]についてもいくつか行われている。

本研究では、それらの研究と語の重要度を組み合わせて作られた文脈的なパラメータ[4]を使い、要約アルゴリズムの構築に利用した。さらにアルゴリズムをシステム化し、より全体の内容を把握した要約を作成する。

2．研究概要

本研究のシステムの過程は以下のようになる。

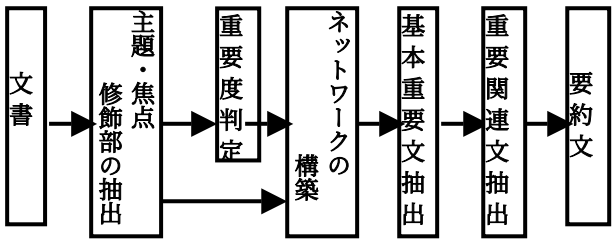


図1　システムの過程

まず入力された文書から主題・焦点・修飾部を抽出する。次に主題・焦点にスコア計算、ネットワーク化を行う。これらのパラメータをもとにユーザの指定する要約率に従って重要文及び、それ関連する文を判定して抽出する。抽出した文をまとめたものを要約文とする。

3．要約文の生成

3.1 日本語語彙大系[5]の利用

意味体系・単語体系を参照し、名詞に対して意味属性を付加する。意味属性は、その単語が属する全ての意味属性とその上位概念を付加する。これにより、これまで見つけることができなかった文間のリンクを結ぶことができた。

3.2 主題・焦点・修飾部の抽出

主題・焦点は前提条件を緩め、形態素解析[6]から抽出する。主題・焦点は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

また、文中の主題・焦点の前方部分に対して、次の並びとなっている名詞を修飾部として取り出す。

[a*]＋[a or b]＋[主題・焦点]

a:名詞＋助詞(の、と)　b:動詞の連体形

3.3 重要度の付加

段落ごとに処理し、ｎ文とｎ＋１文の主題・焦点・修飾部の関係、**表題**、接続詞、日本語語彙大系の親分類番号と文中の語の一致により主題・焦点に加点する。

1）ｎ文とｎ＋１文の主題・焦点その修飾部の一致により主題・焦点に加点する。
2）日本語語彙大系をもとにｎ文とｎ＋１文の主題・焦点・修飾部の親分類番号の一致により主題・焦点に加点する。
3）表題中の語が主題・焦点、また文中に含まれている場合加点する。
4）文中に特定の接続詞が存在した場合、主題・焦点に加点・減点する。

3.4 ネットワークの構築

段落ごとに処理し、ｎ文とｎ＋１文の主題・焦点・修飾部、および日本語語彙大系の親分類番号の一致により文にネットワーク番号を割り振る。

3.5 要約文抽出

主題・焦点のスコア、及びネットワーク番号をもとに文を抽出する。まず主題のスコアから重要文及びそれに関連文を判定する基準点を決定し、点数を比較して要約文を決定する。

要約文は次の二つの文で構成される。

・基本重要文

主題のスコアにより抽出されたネットワーク中で最重要の文である。この文の量Ｐを基本要約率ｐ(%)で指定する。

・重要関連文

基本重要文の内容を補う文として、ネットワーク関係を通して抽出された文である。この文の量Ｑを関連要約率ｑ(%)で指定する。

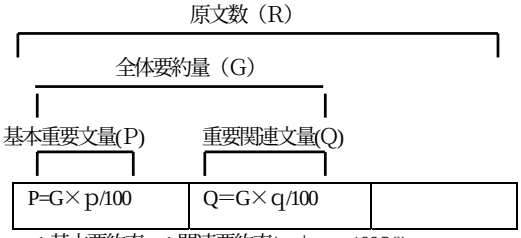
基準点と文のスコアを比較して基本重要文・重要関連文を抽出し、まとめたものを要約文とする。本システムでは要約率、基本要約率、関連要約率をユーザが指定できる。

A Summarization System using Themes and Focuses Network
†Tsuyoshi Ohta, Shoichi Yokoyama, Noritaka Nishihara
‡Graduate School of Science and Engineering,Yamagata University

・要約率

原文の量Rから出力する要約文数Gを要約率ｇ(%)で指定する。要約率ｇを指定した場合は次のようになり、その関係を図2に示す。

ｇ ＝ （Ｐ＋Q）/ Ｒ



p：基本要約率 q：関連要約率(p＋q＝100(%))

図2 要約率、基本要約率、関連要約率の比率

3.5.1 重要度基準

各ネットワークにおいて、主題または焦点のスコアの平均点 Da、最大・最少スコアの平均点 Dm をそれぞれ計算し、大きい方の値を各ネットワークの関連文基準 Di とする。さらに重要文基準 D を決定する。

Da ＝（ΣScore）/ n

Dm ＝ (Score-max + Score-mim) / 2

Di ＝ Da(Da > Dm) or Dm(Da< Dm)

D = (ΣDi / w)・ (n / N) or (≦)2

Score：主題または焦点のスコア

Score-max-mim：ネットワーク中で最大と最少のスコア

n：ネットワーク中の文の総数

w：ネットワークの総数

N：全ネットワーク中の文の最大総数

3.5.2 重要度の判定

1）基本重要文抽出

各ネットワークで一番スコア大きい文の主題スコアを重要文基準 D と比較し、基準以上なら基本重要文とする。基本重要文数Ｐ(G×p/100)を決定し、主題スコアの高かった順に文の数がＰを満たすまで抽出する。

2）重要関連文抽出

基本重要文のスコアの高い順に処理する。各ネットワーク中の次に主題スコアの高いものを関連文基準 Di と比較し、基準以上なら重要関連文としてQ(G×ｑ/100)を満たすまで抽出する。

抽出例

重要文 ― ｎ＋1文・ｎ＋2文

ｎ文　知能はますます低下し、廃人同様のまま 4 年後に死亡した。　－主題＋１　（計 3 点）

ｎ＋１文　この女性の脳は肉眼で見ても著しく萎縮していた。　－主題＋１＋１＋１　（計 5 点）

ｎ＋２文　顕微鏡で見ると脳の神経細胞は見る影もなく減っていた。　－主題＋１＋１　（計 4 点）

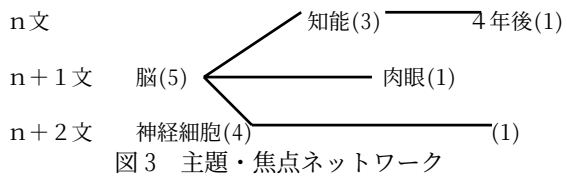


図3 主題・焦点ネットワーク

Di = 4　D = 3.2

日本語語彙大系を使用しない場合、ｎ文とｎ＋１文は別のネットワークになっており、ｎ＋１文とｎ＋２文は同点数でｎ＋２文のみが基本重要文として抽出されていた。

今回のシステムでは、ｎ文とｎ＋１文のリンクが発見でき、ｎ＋１文の主題に加点がされている。よってｎ＋１文が基本重要文として抽出できる。さらにｎ＋２文も主題の点数と Di と比較し重要関連文として抽出される。

4. 抽出結果

本システムで要約率 30%基本要約率 50%とし、従来システム[4]、及び Word2000 で要約率 30%として設定し文を抽出する。抽出結果を、人手によって抜き出された重要文と一致した文の数、また要約に不要と判断した文の数を表1に示す。

表1 抽出結果

	本システム	従来	Word2000
重要文	(9/27)	(8/27)	(9/27)
不要文	(1/10)	(3/10)	(3/10)

重要文については、従来研究や Word と同程度抽出できた。また不要と判断された文との一致も少なく、他のシステムより要約に必要な文を多く抽出することができた。

また本システムと Word との不一致は 4 文あった。それらの文や、お互いに抽出できなかった他の重要文の中には、主題・焦点の抽出精度を高める、また日本語語彙大系以外にもリンク関係を見つけることができるパラメータを使用することによって、抽出できるものがいくつかあった。

5. おわりに

本システムではネットワークに重点を置き文間のつながりを重視して文抽出を行った。また、日本語語彙大系を利用することにより、いままで見つけることができなかった文の関係をネットワークで繋ぐことができた。結果として、重要文以外にも要約に必要な文を多く抽出することに成功した。

今後の展望として、引き続き様々な文書でデータを取り、スコア、ネットワーク構築の判定を調整する。そして評価実験を繰り返し、新しいパラメータを取り入れるなどシステムの向上に努める。

参考文献

[1]Regina Barzilay,Michael Elhadad : Using Lexical Chains for Text Summarization,Summarization, pp111-120,The MIT Press (1998)

[2]廣町 潤,横山 晶一,西原典孝：形態素を用いた主題焦点抽出システム, 情報処理学会 第65回全国大会 講演論文集(2),pp11-13 (2003)

[3]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, 修士論文(2003)

[4]斉藤尚子,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 第六回年次大会 発表論文集 pp467-470(2000)

[5]池原 悟：日本語語彙大系,岩波書店(1997)

[6]形態素解析システム「茶筌」,奈良先端科学技術大学院大学