関連語に基づく文の関係を用いたコラムの自動要約

大木環美† 木村昌臣†
芝浦工業大学工学部情報工学科†

1. 研究背景と目的

現在自動要約における研究が多数行われている．それらの研究では新聞記事や論文を対象としていることが多く，これらの文章は接続詞や照応関係など文章構造が明確であるため，要約が比較的容易である．しかし，これらの情報が不足しており文章構造が明確でない文書に対してはこのような手法の適用は難しい．そのような文書に対する要約手法としては，GDA[1]を用いた要約の研究が行われている．これは，文書中の文章や語句の意味的な構造を明示する XML タグである GDA を付加しこれをもとに要約を行う方法である．この研究の問題点は GDA のタグ付けがされていない文章は要約できない点にあり，本研究では GDA を利用せずに文章構造が明確でない文書の自動要約を目的とする．

2. 研究内容

接続詞や照応関係による文章構造が明確でない場合にはそれら以外の関係を用いて要約を行う必要があるため，それらの情報によらず文と文の関係を捉えた要約方法を考案する．本研究では，短い文章で筆者の意見を示し，接続詞や照応関係などによらない書かれ方が多いと考えられるコラムの1つである朝日新聞の天声人語[2]を対象とする．

2-1. 予備実験（アンケート）

コラムにおける重要な文は筆者の意見を表す文であると考えられる．そこで，そのような文を抽出するためにアンケートを行い，被験者が考える筆者の意見を示す文を集計した．これにより，1つのコラムに筆者の意見を示す文が複数存在し，1 文では筆者の意見を示しきれないという結果が得られた．

この結果を踏まえたうえで，3 つのコラム中のそれぞれで筆者の意見を示している文を 3 つ選択し順位をつける，というアンケートを改めて行った．被験者は 20 代の男女 12 人であった．この結果，被験者によらず筆者の意見を示すとして選択された文が存在したため，読者に共通して捉えられる重要な文が存在するという知見が得られた．

2-2. 関連語によるネットワークの構築

次に，文と文の関係を抽出する方法を考える．コラムでは主題を象徴する語が存在し，そのキーワードに絡めて筆者の意見が語られると考えられるため

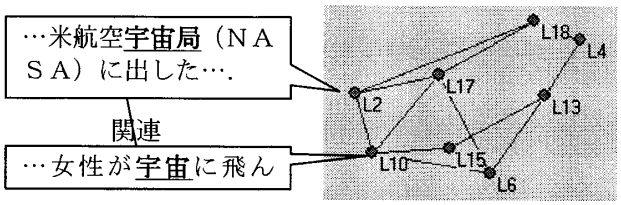

図 1：関連語による文のグラフ
（L*N*は *N*番目の文を表す）

キーワードは文章の構成を取得する上で重要な指標であると考えられる．また，キーワードには多数の関連する語が存在し，関連する語が存在する文同士には関係があると考えられる．そこで，図1のように文をノードとし，文間に関連語が存在する場合にエッジを張りグラフ構造（文群ネットワーク）を構築する．ここで関連語とは「宇宙局」と「宇宙」といったような概念として関係がある語，類義語，同一の語を対象としている．このネットワークで文と関連語の関係を明確にし，重要な文を抽出する．

2-3. 媒介中心性の利用

2-2 で関連語により取得した文との関係には，キーワードとはならない単語同士の関連によるものも存在すると考えられる．また，話題の中で終結し他の話題との関係がないと考えられる文は要約文に適さない．これらのエッジを省くために媒介中心性[3]を用いる．媒介中心性とは，あるノードv_iが他の 2 つのノードv_jからv_kへの最短経路上に含まれる割合である．なお，クリーク内のノードや，エッジが 0 ないし 1 本のノードは他の文間の関係に重要な役割を持たない文を表すノードであると考えられる．

本研究では，媒介中心性を，ある文が他の文をどれだけ媒介しているかを示している指標として考え，その値が0でない文はコラムにおいて話題を展開し，文章の流れを作る役目を担っていると考える．この考え方にもとづき，媒介中心性を持つ文を抽出することにより，要約文に必要な文を求める．

2-4. 関連語群の構築

2-3 までの方法のみの適用だけでは，抽出される文が多すぎてしまい，充分に要約できていないと考えざるを得ない場合がある。そこで，2-3 で得られた文のみで新たにネットワークを作り，そのネットワークにおける重要な文を抽出することで，全体の重要な文を求めることとする．ここで，2-3 で求めたネットワークで関連語を求め，その関連語で関連語のグループ（関連語群ネットワーク）をつくり，主要

Automatic Summarization Of Columns Using Connection Of Sentences Based On Related Words
† Megumi Ohki, Masaomi Kimura
† ShibauraInstituteOfTecnorogy

な関連語のまとまりを抽出することで，キーワード群を求める．関連語群ネットワークは，単語をノード，それらの関連の有無をエッジとしたネットワークであり，そこに含まれるノード数が多いほど関連語が多く，エッジ数が多いほど関連が密であると考えられる．そこで，抽出した各関連語群ネットワークに含まれる総エッジ数を比較し，その数が平均より高いものをキーワード群ネットワークとして取得する．（これに含まれる語群をキーワード群と呼ぶ）

2-5. 文の評価式

キーワードは筆者が最も言いたいことを記述するための主題として扱われているため，文群ネットワークにおけるキーワードとなる語を含む文が文章内で重要な文に隣接していると考えられることから，キーワードとなる語が関連先に多く存在する文のほうがより筆者の意見を示していると考えられる．2-4で求めたキーワード群中の単語で関連する語が群内に多いものはキーワード群中でも重要であると考えられる．そこで，関連語ごとにキーワード群ネットワークでのエッジ数をそれぞれ重みとし，関連先の文内に存在するキーワード群中の各関連語の重みの合計が最も高い文が重要であると考えられる．重要な文はキーワードごとに異なると考えられるため，関連先に存在するキーワード群ごとに，重みを計算し，評価式を用いて重要な文を抽出する．

ある文について関連先の重みの合計を文の重要度として考えると，その文の次数が多くなるほど重みの合計が高くなり主要なキーワード群によって張られたエッジでないものも含まれてしまい，必ずしも本当に重要な文が得られるとは限らない．そこで，関連先の重みの合計を文に張られているエッジ数で割って得た数を重要度とするとよりよい結果が得られると考えられる．さらに，ネットワークの中で媒介中心性の高いノードは文脈の中で関連語の集まる文もしくは文のつながりを作る文であり，エッジ数が同じ文でも重要度に差が出ると考えられるため，重みの合計に媒介中心性の情報を反映させる．

これらを考慮して本研究では評価式を

$$T(i,g)=\frac{\text{(2-3)で求めた関連語群 g の文 i の重みの合計}}{\frac{1}{1+\text{文 i の媒介中心性}}+\text{文 i のエッジ数}}$$

とした．以下、この評価式の値を得点と呼ぶ．

求めた得点から重要な文を求めるために，キーワード群ごとの最も高い得点から重要な文の抽出数を求める．評価式から得られる得点は，ほぼエッジ数あたりのキーワード群中の関連語数となるため，あるキーワード群での文の最も高い得点をtとすると，そのキーワード群中に重要な語がt個含まれていると考えられる．また，重要である語が多く含まれているほど，重要な文を持つ割合も多くなると考えられることから，得点の大きい順にt個の文を重要な文として抽出することとする．

2-6. プロトタイプシステム作成

これまで考案した手法をアルゴリズムとして，プロトタイプシステムを作成した．

前処理として，市販辞書をもとに関連語を構築するためのデータベース（以下，辞書データベース）を作成した．辞書は広辞苑第5版を使用した．

システム実行時には、まずコラムを読み込み，辞書データベースと単語を比較して関連語の組を取得する．その際，辞書データベースのある見出し語の意味部分に別の見出し語が存在している場合と，双方の意味部分に同じ単語が存在する場合にそれらの見出し語同士を関連語とする．以降は考案手法と同じ流れで処理を行い，重要文を抽出する．

3. 評価

20代男女12人を対象に，予備実験と異なる3つのコラムから最も筆者の意見を代表すると思われる文を選んでもらうアンケートを行い，その結果いずれかの被験者が選んだ文を正答として，プログラム結果を精度・再現率の2つの観点から評価した．重要文抽出については2つのコラムでの平均精度が0.71，平均再現率が0.46となった．この結果，本手法は重要な文のうち7割程は取れているが，余分な文も半数ほど抽出してしまっていることがわかった．さらに1つのコラムでは精度，再現率ともに0であった．この原因を目視で確認したところ，重要なキーワードの関連語が抽出できず，媒介中心性を求めたのち，改めてネットワークを作る時点で省かれてしまっていたことがわかった．このことから関連語の組の構築方法について改善を行えばよりよい結果が得られると考えられる．

4. まとめ

関連語がある文を結んで出来るネットワーク構造のうち、他の文間の関係に寄与し、重要なキーワードに関係する文を抽出する手法に基づきコラムの自動要約手法を考案し，そのプロトタイプシステムを構築した．この手法を検証するため，3つのコラムを対象としてコラムの筆者の主張が最も含まれると被験者が考えた文と本手法で得られる要約文を比較したところ2つのコラムで7割前後の精度が得られた．

しかし，関連語の組の生成精度不足が原因で正しく要約されないコラムもあったため，今後はその原因である関連語の生成方法を改善する予定である．

参考文献

[1]Global Document Annotation (2008) http://i-content. org/gda/

[2] 朝日新聞オンライン記事データベース「聞蔵」: http://database.asahi.com/library/

[3]大平雅雄・柗本真佑・前島弘敬・亀井靖高・松本健一：OSS コミュニティにおける共同作業プロセス理解のための中心性分析，グループウェアとネットワークサービス・ワークショップ 2007 論文集，Vol.2007, No.11, pp.7-12,