句、節、文の接続関係を考慮したパラグラフの自動要約

5N－3

小堀 誠　田村 直良
横浜国立大学大学院 工学研究科 電子情報工学専攻
{kobori,tam}@tamlab.dnj.ynu.ac.jp

1　はじめに

本研究[1]では、パラグラフを構成する節、句、文間の関係を考慮して構造化を行い、構造上の特徴をパラメタとして用いた判定法によって、文章の種類に依らずに、パラグラフ単位で自動要約する手法について論じる。

論説文、新聞社説を対象に、照応、省略、語彙連鎖、文末表現など多くの談話要素から重要文を選択していく抄録手法を用いた要約システムは扱いやすいため、数多く研究されてきた。[2]

しかし、抄録手法を用いて作られた要約文は原文をそのまま用いるため、一文が長くなりがちで、冗長な表現を含むことがある。そして、離れた場所から文を選び、一つの文章とするため、原文章の結束性が崩れ、文間の隣接関係が不自然になる場合がある。

一般の文章を対象に自動要約を考えた場合、文章の特徴、形式が一様でないため、従来の抄録手法による要約は有効でない。この場合、意味的なまとまりである段落を処理単位とした要約が基本となる。この手法では、文章の展開を追って要約を行なうので、原文章の結束性を保ちながら要約文を生成できる。

本研究では、一般文章に対応した要約を実現するために、200字前後のパラグラフを対象とした要約システムを提案する。

2　要約システムの概要

2.1　パラグラフの構成

本研究ではパラグラフの要約の立場から、パラグラフを以下のようにモデル化する。

パラグラフは一つの意味的なまとまりとなっている。そして一つ以上の文によって、構成されている。

パラグラフ内の文は修辞関係によって、二分木で構造化できる。この階層を文レベルとする。

文は単文と複文の2種類に分けられ、複文は文の中心的な役割を担う主節と、それに特定の関係で結び付く接続節によって構成される。[1]本研究では、条件、理由、逆接、順接で結ばれる節を二分木によって構造化し、修辞関係と同等に扱う。主節と従属節がこれらの修辞関係によって構造化されている階層を節レベルとする。

単文、主節、従属節は、文節によって構成される。ここで文節とは、「用言」、「用言＋助動詞」、「体言＋助詞」のパターンである。文節の中で用言は述部の役割をし、文中では名詞にかかる連体修飾、文末ではその文の意味の中心的な役割を担う。本研究では、述部を右端とした部分木を作成することにより、一文の文節の係関係の構造化を行う。(図1)

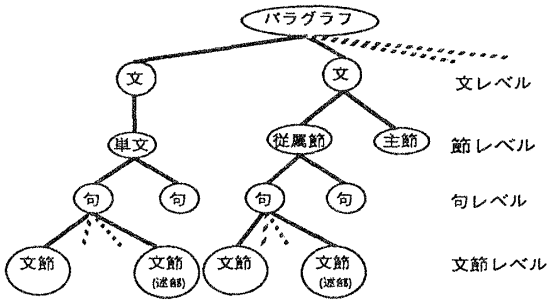

図 1: パラグラフの構造とレベル

2.2　システムの構成

本研究では、パラグラフを入力とし、一文ごとに必要な節を選び、冗長な文節を削除することによって、要約文を出力するシステムの開発を目指す。システムの構成は図2のようになる。

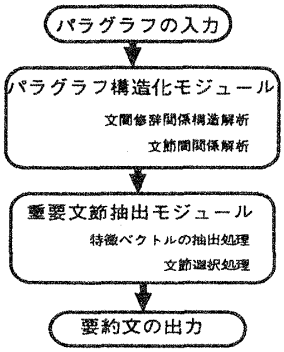

図 2: システムの構成

[1]Paragraph Summarization Based on the Relation between Phrases, Clauses and Sentences
Makoto Kobori　Naoyoshi Tamura
Department of Electrical and Computer Engineering, Yokohama National University

2.3 パラグラフ構造化モジュール

パラグラフ構造化モジュールは、パラグラフを構成する文を修辞関係によって構造化を行なう「文間修辞関係構造解析」と、特定の関係で結ばれた節をもつ複文を分割する「複文分割解析」と、文を後置詞句と述語に分け、後置詞句が述語にかかるように関係づける「複文の分解及び文節間関係解析」からなる。

2.3.1 文間修辞関係構造解析

パラグラフ内の文の構造化は、接続表現から同定される修辞関係と、隣合う2文の文末表現等から同定される修辞関係を用いて行なう。構造化は下記のルールを用いる。

1. スタックの先頭が文である。かつ、接続表現がある場合、スタックの先頭要素S1と二番目の要素S2をポップする。[S2, 修辞関係,S1]をプッシュする。
2. スタックの先頭が文の時、スタックの二番目の要素の核と文末表現等の関係から修辞関係を同定し、同様にしてスタックの先頭の要素と次にプッシュされる文との修辞関係を同定する。スタックの先頭要素と二番目の要素との結び付きが強い場合、スタックの先頭要素S1と二番目の要素S2をポップし、[S2, 修辞関係,S1]をプッシュする。
3. それ以外はプッシュのみ行なう。

2.3.2 複文分解解析及び文節間関係解析

前節のルールによって構造木が生成された後、葉である一文が修辞関係をもつ複文の場合、接続節と主節にわけ、修辞関係によって構造化を行なう。(複文分解解析)

文節間関係解析ではパターンにより後置詞句、述部を抽出する簡易をパーザ用いて文を文節に分け、後置詞句が述部にかかるように関係づける。

3 重要文節抽出モジュール

3.1 特徴ベクトルの抽出処理

構造上の特徴を検出するために、各レベルに特徴ベクトルを導入し、機械学習による要約の手がかりとする。

文レベルの特徴として、次の要素を用いる。

- 文末表現 (意見、断定、叙述、時制)
- 参照表現を含む
- 助詞（は、が）の出現
- 接続表現の有無
- 語彙連鎖を形成する重要語の数
- パラグラフの先頭または末尾までの距離
- 文の長さ
- 特定の修辞関係

句レベルの特徴ベクトルには次の要素を用いる。

- 主節である
- 語彙連鎖を形成する重要語の数
- 句の長さ
- 共起関係をもつ語が含まれる

文節レベルの特徴ベクトルには次の要素を用いる。

- 動詞の必須格である
- 普通名詞、抽象名詞を含む
- 述語である
- 文節の位置
- 語彙連鎖中の重要語を含む
- 抽象名詞を修飾する文節である
- 複合名詞、固有名詞である
- 程度を表す形容詞を含む
- 共起関係をもつ語が存在する。
- 取り立て詞 (強調) を含む
- 慣用句を含む

3.2 文節選択処理

文節の要不要について、訓練データを用意し、C4.5による機械学習により決定木を作成し、文節の選択の判定を行なう。

4 実験と評価法の方針

実験にあたって、まず訓練と評価のために要約結果による正解を定める。要約は、学生10人を被験者として、新聞記事、論文、web上にある情報技術関連記事からランダムに選んだ1000個のパラグラフについて、長さが5割程度になり、かつ意味が伝わるようにするために、重要であると思う文節を選ばせる。5分割交差検定 (Five fold cross validation) を用いて、再現率・適合率を求める。

5 まとめ

一般文章に対応した要約を実現するために、200字前後のパラグラフ内の文を対象にし、文中から重要な文節を抜き出して要約文章を生成する要約システムを提案した。

パラグラフ内の文と、条件、理由、逆節、順接の関係で結ばれた複文内の節を修辞関係と同等に扱い、さらに一文を文節に分割することにより、構造化を行なった。抽出の判断は機械学習による。被験者が要約した1000パラグラフにより、5分割交差検定を行なう。

参考文献

[1] 益岡隆志, 田窪行則. 基礎日本語文法 - 改定版 -. くろしお出版, 1995.

[2] 田村直良, 山下卓規, 奈良雅雄. パラメータの学習による文章構造解析と自動抄録. 言語処理学会 第4回年次大会ワークショップ論文集, pp. 64–70, 1998.