1ZA-7

複数文の結合を用いた要約システム

近藤憲司　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

自動要約手法は、多くが重要文抽出法・文圧縮法である。どの手法でも機械処理に際して文単位の抽出は必ず行われている。この抽出法では出現頻度は低いがその文章特有の情報をあらわす単語が欠落することが多い。また、文のつながりが不自然になることがある。

本研究では、日本語の談話構造において重要な役割を果たす主題・焦点パラメータ及び係り受け解析を用いて文圧縮を行い、圧縮した文章に対して主題・焦点が含まれているパターンを解析し、perl を用いて 2 文間の接続を行うシステムを作成した。

２．主題・焦点を基礎とした文圧縮

２．１．主題・焦点の定義

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：その文で新しく導入された情報

以上より、「主題+焦点」の構成になっている文章と「主題なし+焦点」の構成になっている文章が連続して出現している場合、2 文は連結できる可能性が高い。

A Summarization System using the Concatenation of Sentences

Kenji KONDO, Shoichi YOKOYAMA,

Noritaka NISHIHARA

Yamagata University

２．２．主題・焦点の文圧縮への応用

本研究では、文章を細かく分割しその組み合わせにより文を生成する処理を含む。そのため鈴木[1]の単位文分割処理を適用する。以下にそのアルゴリズムを示す。

(1)単位文の分割処理：各単語を「接頭語」「自立語」「付属語」の 3 種類に分類し、1 つの自立語を中心に条件に従い結合する。

(2)主部及び述部の設定条件：

主部⇒主題・焦点の含まれている単位文節

述部⇒各単位文の末尾の句読点及びその直前の単位文節

(3)範囲の設定：南瓜[2]を用いて係り受け解析を行い、主部及び述部に係る部分を抽出する。

３．複数文の連結

システムの流れは以下のようになる。

(1) 原文の入力

(2) 茶筌[3]を用いた形態素解析

(3) 廣町[4]による主題・焦点抽出処理

(4) 単位文への分割処理

(5)原文に対して、一文単位での圧縮処理

(6)主題・焦点のパターンを調べる

(7)パターンに応じた文連結処理

⇒主題+焦点の場合（一文中に主題と焦点が両方含まれている場合）・・・そのまま出力

⇒主題+焦点なし　主題なし+焦点の場合

- 前文の主題と、それに続く文の焦点を連結させていく形にする

・ 現段階において「主題+は+焦点以下」となるように出力する。

出力例

・圧縮文

「既に自分の名前や誕生日を思い出せない子供が出ている」と国連難民高等弁務官事務所。

基金二百二十万ドルで、コンピュータ・ディスクを検索センターに配布する。

・文生成後

国連難民高等弁務官事務所　は　コンピュータ・ディスクを検索センターに配布する。

４．要約文の評価

(1)TF-IDF 法を用いた場合の評価

本システムで作成した要約文の評価には単語の正規頻度の TF 法に意味的に重要箇所を含むかどうかの評価に用いられる IDF 法を加味した TF-IDF 法を用いる。以下、総単語数がそれぞれ 76, 101, 207 の文章を本システム及び MS-WORD を用いて要約文を作成した。そしてそれぞれの要約文に対し TF, IDF, TF-IDF 法により算出したスコアの比較を表１にまとめる

表１　TF-IDF 法を用いたスコア

	単語数 76 の文	単語数 101 の文	単語数 207 の文
TF	34.14(-1.6)	43.38(+1.84)	98.06(-9.64)
IDF	40.08(+6.82)	41.24(+8.46)	81.24(+12.04)
TF-IDF	76.8(+1.14)	80.46(+2.6)	128.66(+4.64)

カッコ内は MS-WORD の自動要約との比較

・TF-IDF スコアは自立動詞・名詞を元に算出するがサ行動詞「する」などの対処が必要であるため自立名詞の直後にサ行動詞「する」が存在する場合前述の名詞と一緒にして算出する

(2)人間の目からみた客観的な要約文として以下の点で評価する

・読みやすさ

・どれだけの内容を把握できるか

・文同士の連続性

以上の 3 点について、重要文抽出法で作成した要約文と共にそれぞれ 5 段階で第三者に評価をしてもらう。５が最も良い場合とする。

表２　第三者による客観的評価

	単語数 76 の文	単語数 101 の文	単語数 207 の文
読みやすさ	2.2 (-0.2)	2.4 (-1.4)	2.2 (-1.6)
内容の把握	3.8 (+0.6)	3.6 (+1.0)	3.2 (+0.4)
文の連続性	3.6 (+0.8)	3.6 (+1.2)	3.8 (+1.4)

５．まとめ

本研究において、一般の単語の生起頻度を元にした要約システムと比較した場合、文の長短に関係なく TF-IDF のスコアにおいて上回る要約文を作成するシステムを構築することができた。また、客観評価においても、内容の把握及び文のつながりに優れた要約文を出力することが可能となった。

しかし、現段階では以下の条件に該当する文章はすべて出力しないとしている。

・ 主題及び焦点のどちらも含まない文

・ 鍵括弧などカッコで括られた文

・その他、主題及び焦点のどちらかもしくは両方が含まれているが出力では文としてなりたたないもの

今後の研究で、以下の文章に対する更に適切な対応を考える必要がある。

参考文献

[1]鈴木史子：山形大学卒業論文(2002)

[2]南瓜：奈良先端科学技術大学院大学

[3]茶筌：奈良先端科学技術大学院大学

[4]廣町他：情報全大(2)(2002)　pp.11-13