青山学院大学理工学研究科部経営工学専攻

反復語句と深層格に基づく要約システム ABISYS の開発

原田研究室　　矢　後　友　和（35501012）

近年，コンピュータネットワークに代表される計算機技術の急速な進歩により，膨大な量のテキスト情報が手に入るようになってきたため，自動要約に対する期待が高まっている．昨今の要約に関する研究では，文単位での要約から，語句単位で重要な語句を抽出するといった要約へと関心が移ってきている．しかし，既研究では表層情報や構文情報のみを用いており，文章の内容を意味理解していないので，意味的に重要な語句が多く削除されたり，不要な語句が多く要約文に残ったりする可能性がある．

そこで，本研究は，原文中の意味的に重要な語句のみからなる要約文を生成するシステム ABISYS を開発することを目的とした．意味的に重要な語句のみからなる要約を作成するために，本研究では「反復語句」に着目した．原文とその要約文の関係を調査した研究では，「反復語句」や「反復語句の前後のある範囲の語句」が要約文に残りやすいと述べている．反復語句とは，「文章中の異なる文に二度以上出現する同一語句ないしは同義･類義語句(ただし，付属語・感動詞・接続詞・連体詞・形式名詞・補助用言・指示語句は除く)」のことをいう．本研究では，この調査結果を基に，入力文から意味的に重要な語句のみを抽出して要約文を生成する要約システム ABISYS を開発した。

本システムの入力には日本語文章から意味解析システム SAGE によって生成された格フレーム群を用い，要約処理結果を，原文・要約処理で削除された語を示した原文・要約文の 3 形式で出力する．要約処理の具体的な手順は，(１)反復語句の抽出，(２)重要語句の抽出，(３)冗長な重要語句の削除，(４)文生成，の 4 ステップで行う．(１)反復語句の抽出では，まず，原文での表記が同一の同一語を抽出する．次に EDR 電子化辞書の概念体系辞書を用いて概念距離を計算し，「見出し」は違うが意味的に近い同意語を抽出する．最後に，ユーザが指定した要約強度(１～３)に従って要約文に残す反復語句を決定する．(２)の重要語句の抽出では，深層格を用いて，抽出された反復語句の隣接語句を抽出する．ただし，その隣接語句が用言の場合は，その用言の必須格や隣接用言も重要語句として抽出する．必須格は共起辞書を基に用言毎の各深層格の出現頻度を計算して決定する．(３)の冗長な重要語句の削除では，補足語修飾節と引用動詞の 2 点に着目し，意味的に冗長な語句を重要語句から削除していく．(４)の文生成では，(１)～(３)で抽出された要約要素語を文節番号でソートし，格フレーム中の助詞を補って要約文として生成する．ただし，同一文番号中のフレームに main 格を伴う用言がない場合は，不完全な文なのでその文を削除する．

中国茶に関する論説文の要約を本システムで生成したところ，原文の文数 16 文(422 文字)に対し，要約強度 1，2，3 に従って，4 文(53 字)，6 文(99 字)，9 文(157 字)となり，要約率は 0.25(0.126)，0.375(0.235)，0.5625(0.372)となった．生成された要約文の内容は，要約強度 1 では，要約文から原文の内容の一部を理解することはできない結果になったが，2 や 3 の場合では，大筋の内容を理解することができる要約文を生成した．従って，本システムは原文の意味内容をほぼもらすことなく文字数を 4 分の 1 程度まで短くした要約文を生成することができると言える．