全文情報からの意味的情報の抽出と加工

2D-8

岩淵　保　　荒井幹夫　　藍沢　実

(株) テレマティーク国際研究所

1．はじめに

近年、データベース・サービスで全文を提供する機運が醸成されつつある．これらのシステムで、利用者の利便を考えると全文からの意味的情報の抽出とそれの文章表現化の加工作成システムのサポートが必要になってくる．
今回我々は、原文献を読まなくても、内容理解が得られる報知的抄録の生成システムを実験的に作成した．
これに情報処理に関する論文１０編を入力し結果をＪＩＣＳＴ提供抄録と比較し評価した．
この結果、一部内容が詳しくなり文字数も多くなることが明かとなった。今後のシステム展開を考えシンプルを第一とするシステムでまとめあげた．
以降、このシステムの構成と処理手順について述べる．
なお、ここで言う「文」とは、文頭から「。」：「．」により区切られる一つのまとまった意味を終りまであらわした一つづきの言葉列を言う．
また、「単語」とは、自立語のことで複合語はそのままの形で存在する．

2．システムの構成と処理手順

本システムの処理手順を以下に、流れを図１に示す．

1．原文献ファイルの入力：ファイル名を指示する．
2．前処理：本文中にある図や表などの挿入によって発生した不用なスペースや、復改コードを削除し、文の区切りを確実にする．
3．重要語の抽出：下記の手順で抽出する．
　①標題、副標題の行数を指示する．指示された行の中で漢字、カナ文字で構成される単語を抽出する．
　②本文中の「は」、「には」、「とは」の前の漢字、カナ文字で構成される単語を抽出する．
　③　②で抽出された単語の前に「と」、「や」があると、その前の単語も、漢字、カナ文字で構成されていれば抽出する．
　④抽出された単語が、一文字であれば、経験則で重要語になりえないので外す．
　そのほかは、重用語とする．
4．最重要語の確定：抽出された重要語の使用頻度をカウントし、最も多い単語を最重要語として確定する．
原文献内の参照されている図や表の標題やその内容の単語、または参考文献も頻度チェックの範囲とする．
単語の頭から５文字目までが同一であれば、同一単語としてカウントする．
5．重要文の抽出：最重要語が必ず含まれており且つ、重要語が一つ以上含まれている文を抽出し、重要文とする．
例外処理として「結果」、「今後」が含まれている文は、無条件で重要文とする．
6．整形作業：抽出された重要文の文章は、前後の文との間で脈絡の欠けたものとなる．これを整形する．
整形は、コンピュータによる自動整形と人手による整形作業に分かれる．
下記手順で自動整形を行う．
①重要文の中で下記単語を削除する．
「上記」、「上記と」、「前述の」、「前述のように」、「ここでは」、「これらの」、「この」
②文中に下記単語が含まれている場合、その文をそっくり削除する．
　１）「表」と次の語が数字の場合
　２）「図」と次の語が数字の場合
　３）文頭に数字がある場合
下記事項が現在人手作業として残る．
①意味が重複している文の削除
②前後の関係で意味不明となる文の削除
人手作業は、自動整形後の文章を読むだけで上記２項の条件による削除だけであり、文や単語の追加修正は不用である．
7．出力：作成された文章を意味的情報として出力する．

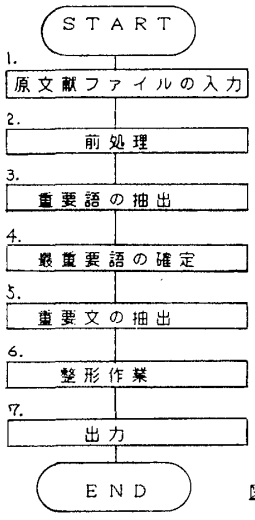

図１　処理の流れ

3．おわりに

シンプルなシステムではあるが報知的抄録が作成できた．
今後、実験結果を生かし下記事項への進展を計りたい．

1．自然な文章整形への人手排除
2．文字数限定抄録作成システム
3．箇条書等も含む要旨の抽出、加工技術
4．利用者の抽出事項指定による（例えば、「実験方法だけを知りたい」）文の抽出、加工技術
5．自動索引システムへの展開

参考文献

1）日本科学技術情報センター情報部：ＪＩＣＳＴ抄録作成テキスト
2）喜多壮太郎：説明文を要約するシステム、情報処理学会、自然言語処理、６３－６（１９８７）

Automatic detection and arrangement of semantic information on full-text documents
Tamotu IWABUCHI , Mikio ARAI , Minoru AIZAWA
Telematique International Research Laboratories