2V－9

# 自動抄録機能をもつ対話的文書検索システム
## ― 自動抄録機能 ―

小野 顕司　住田一男　三池誠司
(株) 東芝 研究開発センター

## 1. はじめに

自動抄録機能をもつ対話的文書検索システムを開発した[2][3]。文書検索では検索された文書が所望のものか効率的に判断できることが重要であり、そのためには文書の任意の部分を利用者が望む詳細度で提示するような機能が必要である。この機能を実現するため、検索システムに試作した自動抄録システムを組み込んだ。

抄録は、文章中の接続表現や話題表現を手掛かりとして解析抽出した修辞構造に基づいて生成される。利用者からの要求に応じて動的に抄録の長さを変えることができる。本稿ではこの自動抄録機能について述べる。

## 2. 自動抄録機能

自動抄録生成は **修辞構造解析**、**重要文評価**、**構造既約** の3ステップで行なわれる。

**修辞構造解析** ステップでは章節毎に、修辞構造を解析抽出する。修辞構造とは文章の各部分の間の修辞関係(順接や逆接、理由や例示など)の構造のことであり、この構造を文を単位とした木構造で表現する。この構造を接続詞などの修辞表現を手がかりとして求める。この処理の詳細は文献[1]に述べられている。

**重要文評価** ステップでは、原文の各文の重要度を、修辞関係毎に決められた相対重要度を元に計算する。相対重要度には、*RightNucleus*, *LeftNucleus*, *BothNuclei*の3つのタイプがある。*RightNucleus*は、修辞構造上右の枝のほうが左の枝より重要であることを示す。このタイプの修辞関係には、**順接**や**概括**などがある。これは、テキストの隣接する2つの部分がこれらの修辞関係を介してつながっている場合、後者の部分が前者の結論ないしまとめになっているので、前者より重要であるという判断を示している。

*LeftNucleus*は、修辞構造上左の枝のほうが右の枝より重要であることを示す。このタイプの修辞関係には、**例示**や**理由**などがある。これは、テキストの隣接する2つの部分がこれらの修辞関係を介してつながっている場合、後者の部分は前者の例示や理由にすぎないので、前者の方がより重要であるという判断を示している。

*BothNuclei*は、どちらの枝の重要度も等しいことを示している。このタイプの修辞関係には、**並列**や**継続**などがある。

実際の重要文評価は、減点方式で行なわれる。テキストの重要な部分を判定するため、システムは修辞構造上の修辞関係を付与されたすべてのノードに対して上述の相対評価を行ない、より重要でないと判定された枝を減点する。ルートノードの値を0として、減点値は、木構造の上位から下位へと加算されていく。原文の各文はこの構造の末端に位置し、上位の枝を介して伝達された減点値がその文の重要度(非重要度)を示すことになる。この重要文評価に応じて抄録に含まれる文が決定される。

**構造既約** ステップ では、ユーザの指定する抄録長に収まるよう重要度の低い文から削除してゆき、抄録を生成する。

生成される抄録は'大意的抄録'(原文の各段落の要旨を繋げた全文の縮小相似形的な抄録)であり、'要旨的抄録'(結論部分を1文程度にまとめたもの)ではない。

生成される抄録の長さや限界は、修辞構造に依存する。例えば多項目の箇条書きや列挙的な表現を文章が含む場合、各項目の間の修辞関係は**並列**であり*BothNuclei*タイプであるため各文は等しい重要度を持つ。従って、抄録中にはそのすべてが入るか入らないかのどちらかである。そのような場合、短い抄録が作れないか、あるいは作れた場合でも、抄録を僅かに長くするといったことはできない。

修辞構造解析は検索に先だって行なわれ、解析結果は原文とともに格納される。重要文評価、構造既約はユーザからの抄録長変更要求がある毎に、ユーザが指定した章節に対して行なわれ、抄録が動的に生成される。

## 3. 動作例

検索システム中の抄録提示ウインドウにおいて、任意の節をマウスで指定して、その節をより詳しく、あるいはより簡単な形で読むことが可能である。図1では検索された文書の第4節の抄録が表示されている。この節は原文では13文あるが、2文で表示されている。

## 4. 実験

2種類の文書に対して、生成された抄録文を重要文再現率の観点から評価した。

まず朝日新聞の社説30記事および東芝レビューの記事42に対し、節毎の重要文および最重要文を1人の被

---

A Document Retrieval System with an Interactive Abstract Generator — Details of the Abstract Generator —

Kenji ONO, Kazuo SUMITA, Seiji MIIKE
Research and Development Center, Toshiba Corp.
ono@isl.rdc.toshiba.co.jp

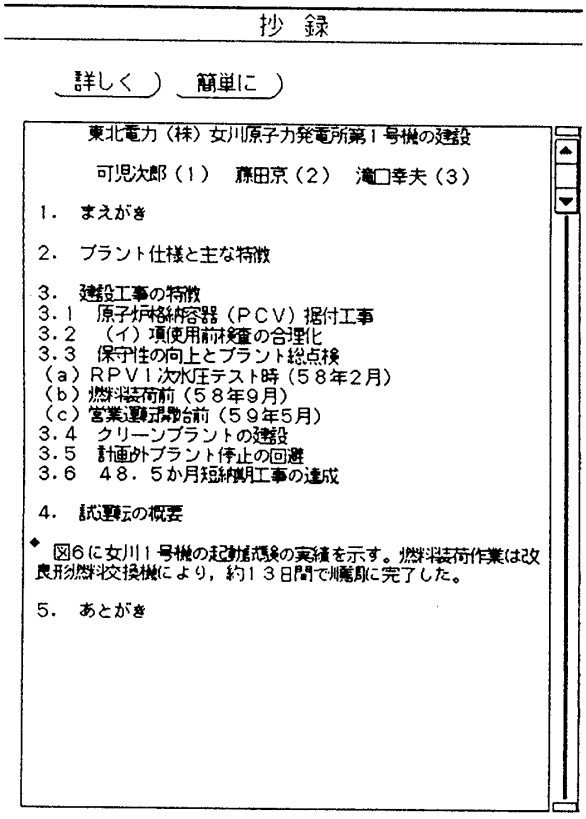


図 1: 抄録表示例

験者が判断した。次に各記事の抄録を生成し、被験者の選んだ重要文と比較した。

一般に大意的抄録の長さの目安は原文の 1/3 ～ 1/4 といわれている [5]。今回の実験では抄録の長さは、社説に関しては原文比 30%、東芝レビューに関しては原文比 25% にできるだけ近いものを生成するようにした。社説の方がより長い設定値になっているのは、羅列的な記述が多く、前述したように短い抄録の生成が困難であるからである。

結果を表 1に示す。社説に関しては、抄録の原文比の平均が 30%、最重要文再現率の平均が 60% である。東芝レビューに関しては、抄録の原文比の平均が 24%、最重要文再現率の平均が 74% である。

社説に対し東芝レビューの方が、低い原文比 (高い圧縮率) でありながら重要文再現率が高いという結果が得られている。これは以下の理由によるものと思われる。解説記事では一般に読者の知らない話題について述べるので、理解を助けるため、つまり述べられている事柄相互の関係を明確にするため接続詞などの修辞表現を多用する傾向がある。そのためシステムにとって手がかりとなる表層情報が多くなり、修辞構造解析の精度が高くなる。従って、重要文の絞り込みを高精度に行なうことができる。一方社説では読者に既知の話題を扱うことが多く、文間の関係は内容的に自明なのでにあえて明確に表現することが少ない。そのためシステムにとって解析の手がかりが少ないこととなり、修辞構造解析の精度がより悪く、それが結果に反映していると思われる。

表 1: 抄録の重要文再現率

| 文種 | 文書数 | 原文比 | 重要文再現率 | |
|---|---|---|---|---|
| | | | 重要文 | 最重要文 |
| 社説 (朝日新聞) | 30 | 0.3 | 0.41 | 0.60 |
| 論文 (東芝レビュー) | 42 | 0.24 | 0.51 | 0.74 |

## 5. おわりに

対話的文書検索システムの文書提示機能として試作した自動抄録システムを組み込んだ。また、社説および技報を対象に、重要文再現率の観点から抄録文を評価した。今後は修辞構造解析精度を高めるとともに、現在のセンテンスセレクション方式の抄録生成を句や節を単位としたより肌理細かい方式に改良する。また抄録文の評価に関して、理解性などの観点からも評価を行ないたい。

また文書検索の観点からの抄録提示の有効性の評価、および利用者の志向や視点を考慮した抄録の生成について検討を行ないたい。

## 参考文献

[1] Sumita, K., et al. : "A Discourse Structure Analyzer for Japanese Text", *Proc. Int. Conf. Fifth Generation Computer Systems 1992 (FGCS'92)*, pp.1133–1140, 1992.

[2] Sumita, K., et al., Document Structure Extraction for Interactive Document Retrieval Systems, Proc. SIGDOC '93 pp.301-310, 1993.

[3] 住田他, 自動抄録機能をもつ対話的文書検索システム — システムの機能と構成 —, 第４８回情処全大, 第３分冊, 2V-7, 1994.a

[4] 佐久間編, 文章構造と要約文の諸相, くろしお出版, 1989.