文書構造に基づく自動抄録生成

Abstract Generator Based on Document Structure Analysis

住田　一男　K.Sumita
小野　顕司　K.Ono
三池　誠司　S.Miike

日本語文書の構造解析に基づいた自動抄録生成システムを開発した。文書構造は，“例示”や“補足”といった文と文の間の関係(修辞関係と呼ぶ)や文章としてのまとまりを，文を単位とした2分木で表現するものであり，文書中の接続詞などの表層的な言回しを手がかりにして抽出する。この文書構造で表現される修辞関係に基づき文の重要度を評価し，抄録文から削除する文を決定するとともに，原文との論旨の一貫性を保つため接続詞の付換えを行い，抄録文を生成する。技術論文や新聞社説などの論説文を対象として，文書検索システムの提示機能として実現した。読まなければならない文書の量が減らせるため，利用者の検索作業の効率向上が図れる。

This paper describes an automatic abstract generation system incorporating a document structure analyzer. From a document, the system extracts a document structure representing rhetorical relationships among sentences such as “exemplification” and “supplementation” as well as sentence chunks. The system evaluates the importance of sentences based on the analyzed structure and decides which sentences should be discarded from the abstract. It also tries to generate an abstract consistent with the original text by replacing conjunctive expressions.

The system has been realized as a document presentation function of a document retrieval system, and has been evaluated using expository writings such as technical papers and newspaper editorials. The experimental results obtained confirmed the validity of the generated abstracts.

1 まえがき

ワープロやパソコンの普及，インターネットなどのコンピュータネットワーキングの発展に対応して，文書の電子化が加速的に進んでいる。扱わなければならない文書情報が増大するにつれ，文書情報への効率的なアクセスを可能にするための技術が求められている。自動抄録生成は，あふれかえる情報への効率的なアクセスを実現するうえで，実現が望まれている重要な要素機能である[1]。

従来試みられた代表的な手法として，単語頻度に基づく方法と世界知識を利用した方法がある。前者は，重要語を含む文を重要文として選び抄録文を生成する(重要語の選定は語の使用頻度を手がかりにする)[2]。この手法には，生成する抄録文が文の羅列になってしまうという問題がある。後者は，ある分野における物事のありかたやそれらの間の関係を世界知識として記述し，その知識に基づいて抄録や要約を生成する[3]。この手法は，扱う文書の内容に依存した知識をあらかじめ記述しておく必要があり，内容を限定しない用途では現実的な手法ではない。

今回開発した方式は，接続詞や文末の言回しなどを手がかりに文書の構造化を行い，その構造に基づき重要文の選定をして抄録文を生成する[4]。分野や内容によらない表層的な情報を手がかりにした処理により，原文との整合性を保った抄録文生成を図った。

ここでは，対話型文書検索システム“BREVIDOC”(Broad-catching System with an Essence Viewer for Retrieved Documents)[5]の文書提示機能として実現した自動抄録システムの構成と機能について述べる。

2 システムの構成

開発した自動抄録システムの構成を図1に示す。システム

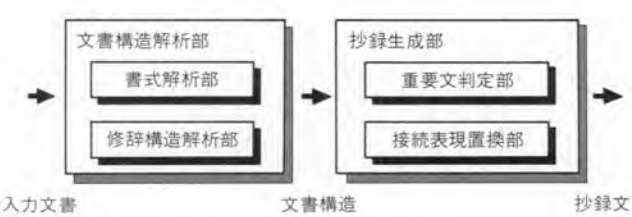

図1．自動抄録システムの構成　入力文書から文書構造を抽出した後，その構造に基づいて抄録文を生成する。
Configuration of abstract generation system

は，文書構造解析部と抄録生成部からなる。文書構造解析部は書式解析部と修辞構造解析部から，また，抄録生成部は重要文判定部，接続表現置換部からなる。

文書構造解析部では，章節の階層構造の認識や，文や段落の切れ目の解析(書式解析)を行った後，文章本体について文間の修辞関係の解析と，文章のまとまりの抽出を行う(修辞構造解析)。

一方，抄録生成部では，抽出された文書構造に基づき文の重要度を判定した後(重要文判定)，原文を表現する文書構造での文間の修辞関係が，抄録文を表現する文書構造においても成立するように接続表現の付換えを行い，抄録文を生成する(接続表現置換)。

次の章では，上述の処理部の各処理を詳述する。

3 文書の構造化

書式解析部は，文書の章や節の階層関係，文や段落の切れ目などを解析する。解析にあたっては，章節見出しに付与される章節番号，句点や空白などを手がかりにしている。この処理は，文書自動レイアウトシステムDarwin[6]における書式構造の解析と同様の処理を行っている。

一方，修辞構造解析部では，章や節の見出しを除いた文章本体を対象として，文と文の間の修辞関係を抽出するとともに，文章のまとまりを検出し，修辞構造として表現する。

3.1 文章の修辞構造

ある文章例に対する修辞構造を**図2**に図示する。数字は，各文の文番号を示している。また，各部分木の間に付与されている“逆接”や“順接”などは修辞関係である。この修辞構造では，第5文が第4文に対して逆接という関係に，第3文から第5文のまとまりが第2文に対して定義という関係に，そして，第2文から第5文のまとまりが第1文に対して順接という関係にあることなどを表している。

修辞関係は，接続詞や文末表現などの文間の接続的な関係を明示する表現を34種類に分類したものである。**表1**に修辞関係の例と，その修辞関係に相当する表層表現の例を示す。

1 区間分割相関方式は，収束が保証されているという利点をもっている反面，全区間をひとわたり修正するのに，トレーニング信号をN回受信しなければならず，収束に時間がかかるという欠点をもつ。
2 そこで，収束を速くするための便法として，区間の仕切を取り払い，トレーニング信号受信のたびに（21）式を全区間［L，M］に一斉に適用する方式を考えた。
3 これを，“部分相関方式”と呼ぶ。
4 部分相関方式では，当該サブ区間以外のタップ利得が一時的に凍結されているという条件は成り立っていないので，収束は理論的には保証されていはい。
5 しかし，修正係数αを十分小さく選べば，近似的にこの条件が満たされているから，実用上は収束が期待しうる。

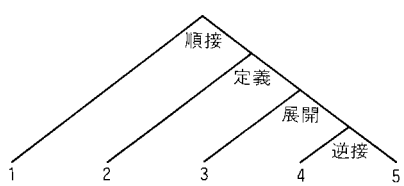

図2. 文章例およびその修辞構造　文章の左に付されている数字は文番号を意味する。また，文章中の下線は説明のため付与した。
Example of text and its text structure

表1. 修辞関係の例
Examples of rhetorical relationships

関係名	表　現　例
順　接	したがって，そうすると
理　由	なぜなら，というのも，…だからである。
対　比	一方，他方
例　示	例えば，このほか
背　景	従来，…されつつある。
逆　接	しかし，にもかかわらず
話　題	この特集では，ここでは
定　義	これを…と呼ぶ。
換　言	つまり，すなわち
展　開	この，その

3.2 修辞構造の解析

修辞構造の解析は，以下の三つのステップからなる。

(1) 修辞関係抽出　　接続詞や文末の表現から修辞関係を取り出し，文書中の各文に対する同定子と修辞関係からなる系列(修辞関係系列と呼ぶ)を生成する。例えば，図2の文章例に対しては，第2文の“そこで”という表現から“順接”が，第3文の“これを…と呼ぶ。”という表現から“定義”が，第5文の“しかし”という表現から“逆接”がそれぞれ抽出される。また，第4文には接続表現が明示されていないので，“展開”という関係が付与される。その結果，次の修辞関係系列が抽出されることになる。

(1　順接　2　定義　3　展開　4　逆接　5)

表1で例示した表現と修辞関係との対応情報を，修辞関係抽出規則と呼ぶif-then型の規則で記述している。この規則の各条件部には，表層表現や形態素列に関する照合パターンを記述するようにした。照合した規則に対応する修辞関係が，その文の修辞関係として選ばれる。

(2) 分断処理　　修辞関係抽出から出力される修辞関係系列内の各修辞関係は，文単位に求めた関係にすぎない。一方，文書中には，複数の文にわたる構造を規定する修辞的な表現が存在する。このステップでは，このような複数文を範囲として依存関係をもつ修辞表現から，構造化に対する制約情報を取り出す。

この処理も，修辞関係抽出処理と同様にルールベースな処理である。記述されている規則としては，「“確かに”という表現が文頭で用いられている文の何文か後に，逆接の関係

をもつ文が現れた場合，その文の直前でまとめる」という規則が例としてあげられる。このステップで検出した文章のまとまりに関する情報は，制約として修辞関係系列に埋め込まれる。

(3) 構造候補生成　分断処理で付与された制約を破らない2分木構造を生成するとともに，生成された構造の優先度を求め，優先度の高い構造候補を出力する。優先度判定は，隣接する修辞関係の間で成立する論旨の流れの開始／終了に着目して行われる。例えば，(1　例示　2　順接　3)という修辞関係系列を考えた場合，例を述べるという議論が第2文で終了していると判断できる。つまり，

((1　例示　2)　順接　3)の構造のほうが，

(1　例示　(2　順接　3))より自然性が高い。

このような構造の選好性を，あらゆる修辞関係の組合せ(34×34)に関して，あらかじめ判定し規則として整備した。この規則を用いて構造の優先度づけを行う。

抽出した修辞構造は，書式解析で得られた章や節の階層構造と関係づけて文書構造として出力する。

4 抄録生成

章や節ごとに解析した修辞構造に基づいて抄録文を生成する。この処理は，以下の二つのステップから成る。

(1) 重要文判定　修辞関係は，その関係によって結び付けられる左右のノードに関する相対的な重要度により，次の三つのタイプに分類することができる。

右核型　右ノードが重要(例　順接，換言，逆接)

左核型　左ノードが重要(例　例示，定義，補足)

等価型　左右ノードの重要度が等価(例　並列，対比，展開)

例えば，修辞関係“順接”(表現例としては“したがって”など)は，右ノードが左ノードの結論と位置づけられる修辞関係である。そこで，このように右ノードが相対的に重要となる修辞関係を，右核型と分類した。同様に，左ノードが重要である修辞関係は左核型に，左右ノードの重要度が変わらない修辞関係は等価型に，それぞれ分類するようにした。

重要文の判定では，上述の相対的重要度の分類に基づいて文の重要度を判定する。例えば，図2に示した構造を考えると，“順接”は右核型であるから右側の部分木(文2，3，4，5)の重要度が高く，“定義”は左核型であるから左側の部分木(文2)の重要度が高いというように判定できる。修辞構造中のすべての修辞関係について，このような重要度を判定することにより，文の順序づけが可能となった。

図2の構造に対して各文の重要度で順序づけた結果を**図3**に示す。影の濃い部分は，その部分に位置する文の重要度が高いことを表している。この構造では第2文の重要度がもっとも高く，最短の抄録は第2文ということになる。

指定された圧縮率(抄録文と原文の文数比)となる文数になるように重要度の上位の文を選び，これらの重要文を出現順に並べることで抄録文を生成する。

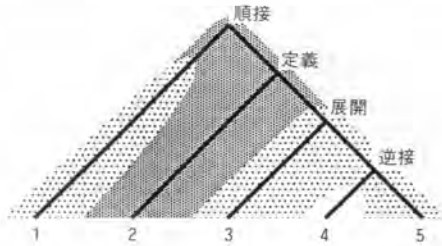

図3．重要文評価　影は文章中の重要度を表す。影の濃い部分は重要度が高いことを示す。

Sentence importance evaluation

(2) 接続表現置換　文間の修辞関係の一貫性が保たれるように，接続表現の付換えを行う。例えば，次のような文章を考えた場合，重要文を並べるだけでは原文と論旨の異なる抄録文が生成されてしまう。

1　家を建てるにはいろいろなものが必要です。

2　例えば，くぎや柱が必要です。

3　すなわち，建材といったものが必要です。

上記の例では，(1　例示　(2　換言　3))という修辞構造が得られる(ただし，1，2，3は各文の接続表現以外の部分を表現している)。この場合，三つの文の重要度は，文1，3，2の順で順序づけられるので，第1文と第3文が重要文として選ばれる。

単に文を並べるだけでは，次のような原文と論旨の異なる抄録文が生成される。

1　家を建てるにはいろいろなものが必要です。

3　すなわち，建材といったものが必要です。

第3文の接続表現を第2文の“例えば”に置き換えることにより，次のように原文で成立する関係を保つことができる。

1　家を建てるにはいろいろなものが必要です。

3　例えば，建材といったものが必要です。

接続表現置換処理では，修辞構造上で次の条件が成立した場合，接続表現の付け換えを行う。

if　左部分木Lに重要文が存在
　　and　右部分木Rに重要文が存在
　　and　Rの最左端の文Cが重要文でない

then　R内の最左端の重要文Xの接続表現を文Cの接続表現に置き換える

上記の例では，Lが1に，Rが(2　換言　3)に，Cが2に，Xが3に，それぞれ対応することになる。

5 文書検索システムの提示機能としての利用

開発した自動抄録システムをBREVIDOCの文書提示部に適用した。このシステムでは，検索した文書の抄録文を生成し，原文と抄録文をペアにして利用者に提示する。抄録文と原文との文数比は，対話的に指示することが可能である。

図4は，文数比を変えた場合に，提示される抄録文の詳しさが変化する様子を示している(例示の原文は東芝レビューから引用した)。

提示した抄録画面のうち(a)の画面は，原文中の著者要約部分から生成した抄録と原文の章節見出しを表示している。(b)は，(a)の抄録部分についてより詳しい抄録を提示した画面である。また，(c)は4章を指定してその部分の抄録を生成し提示した画面である。

図4. 抄録文の提示例　抄録と提示文書の各章の見出しが提示されている。(b)は(a)に対して抄録部分を詳しく，(c)は4章の部分を詳しく表示した例。

Examples of abstracts

6 あとがき

文書構造に基づく自動抄録生成システムについて述べた。接続表現や文末の表現などの修辞表現を手がかりにすることで，処理対象の文章の内容や分野に依存しない方式を実現した。しかし，現在のところ修辞表現が多用される社説や技術論文といった論説文に処理対象が限定されている。論説文以外に，抄録を必要とする文書としては，新聞記事のように事実を伝えるタイプの文書がある。今後，このタイプの文書も扱えるようにしていく。

文　献

(1) オンラインデータベースディレクトリ'91，東洋経済新報社，p.20(1991)
(2) H.P.Luhn：The Automatic Creation of Literature Abstracts. IBM Jounal,pp.159-165 (Apr.1958)
(3) 稲垣博人："事象解析による要約情報の抽出"，情報処理研究会資料，NL 84-3 pp.17-24 (1991)
(4) K.Ono, et al：Abstract Generation Based on Rhetorical Structure Extraction,Proc.COLING'94,pp.343-348 (1994)
(5) S.Miike,et al:A Full-Text Retrieval System with a Dynamic Abstract Generation Function,Proc.SIGIR'94,pp.152-161 (1994)
(6) I.Iwai,et al:A Documet Layout System Using Automatic Document Architecture Extraction,Proc.CHI'89,pp.369-374 (1989)

住田　一男　Kazuo Sumita

1982年入社。自然言語処理，情報検索などの研究・開発に従事。現在，研究開発センター 情報・通信システム研究所主務。

Communication & Information Systems Research Labs.

小野　顕司　Kenji Ono

1987年入社。自動抄録生成，情報検索の研究・開発に従事。現在，研究開発センター 情報・通信システム研究所。

Communication & Information Systems Research Labs.

三池　誠司　Seiji Miike

1984年入社。自然言語処理，情報検索の研究・開発に従事。現在，研究開発センター 情報・通信システム研究所主務。

Communication & Information Systems Research Labs.