7B−11

日本語論説文自動抄録システムの試作と評価*

知野 哲朗　浮田 輝彦
(株)東芝 関西研究所

小野 顕司　住田 一男
(株)東芝 研究開発センター
情報・通信システム研究所

1 はじめに

従来の抄録作成では、統計的な手法 [Luhn 58]、キーワードによる処理 [Isibasi 89]、スクリプトを用いた手法、要約に対する制約を用いる手法 [Yasuhara 89] などが提案されているが、領域を限定しない原文章 (の表層) から、文章の展開を把握し、文章としてのまとまりを持った抄録文章を生成することは不可能であった。これに対し、本研究では、(1) **文脈構造解析**によって得られる文間の修辞的 / 論理的関係に基づき重要文を判定し、(2) 抄録文章で文間の論理関係を正しく表わすよう、接続表現を適切に追加 / 変更する**文書自動抄録システム** (以下: **抄録システム**) を開発した。本稿では、システム構成、抄録生成、および出力である抄録文章の評価実験について報告する。

2 文書自動抄録システム

説明的原文章からその抄録文章を自動的に出力する**抄録システム**の構成を図1に示す。**単文解析部**では、入力文章の手がかり表現 (clue phrase) に注目して、**文脈構造**を抽出するための情報を抽出し、**文脈構造解析部**での**セグメンテーション処理**と、**思考制約規則**によって文脈構造を抽出する。そして、**重要文決定部**で、文脈構造に基づいて**重要文判定**を行い、**抄録生成部**で、**接続詞変更**と**文短縮化**を行なって抄録文章を生成する。なお、**抄録文章ブラウザ**は、抄録文章を文脈構造とともに表示するツールである。

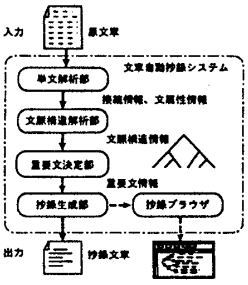


図 1: 抄録システムの構成

2.1 単文解析

単文解析部では、clue phrase の出現に注目した解析によって、文章中の各文から接続関係を抽出し、接続関係系列を得る (図2)。この接続関係系列において、数字は、対応する文 (センテンス) を表し、並列、逆接、同列、順接といった 34 種のラベルは、そのラベルに先行する部分と、後続の部分の間の接続関係を表している。

(1) 情報処理技術の日進月歩の発展によって、大量の情報が扱われるようになった。(2)「情報化社会」が到来したとも言われている。(3)だが、情報は、単に蓄積すれば良いという訳ではない。(4) 必要な時に、適切な情報を得られなければならないからだ。(5)つまり、不要な情報を捨てることも重要となってくるのである。(6)よって、日々発生する大量の情報の中から、本当に必要な情報を選択する目を養うことが、大きな課題となるのである。

⇓

[(1) 並列 (2) 逆接 (3) 理由 (4) 同列 (5) 順接 (6)]

図 2: 接続関係系列の抽出

*Implementation and Evaluation of Automatic Text Skimming System for Japanese Explanatory Texts,by T.Chino, T.Ukita(Toshiba Kansai Research Lab.), and K.Ono, K.Sumita(Toshiba Research and Development Center).

2.2 文脈構造解析

接続関係系列と clue phrase の出現パターンから、文脈構造に対する制約を与えるセグメンテーション規則と、接続関係の局所的な組合から、対応する部分文脈構造を規定する選好的知識とによって、接続関係系列から、図3の文脈構造を抽出する [Ono 93]。

2.3 重要文決定

文脈構造において接続関係を持つ文間の重要性の相対順序は、その両者の間に成り立つ接続関係によって規定する。**順接** (よって) や**逆接** (しかし) といった *Right−Nucleus* 型の接続関係を持つノードでは、(左部分木)<(右部分木) の重要性の順序が付けられ、**例示** (たとえば) や**補足** (ただし) といった *Left−Nucleus* 型では、逆の順序が付けられる。また、**並列** (また) や**対比** (一方) といった *Both−Nucleus* 型では、左右の部分木の重要性は同等である。(図3の破線が、図2の文章全体の文の重要性 (小 < 大) の順序を表している)。抄録文章は、この順序に基づき、段落ごとに、最も優先度の高い文 (のグループ) を抽出することによって生成される。

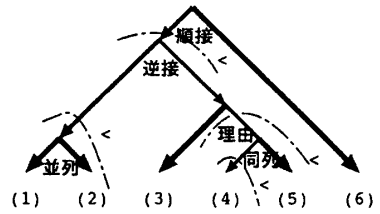


図 3: 文脈構造の例

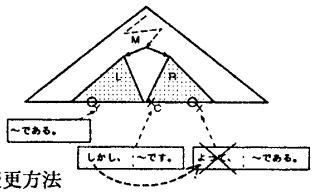
2.4 接続表現変更

重要文として抽出された文を、ただ単に原文章での出現順に並べるだけでは、正しい抄録文章を生成することは出来ない。これは、重要文抜出しのみによる抄録生成のような、従来技術における大きな問題点の一つであった。

本抄録システムでは、単文解析時に、接続表現辞書を参照して各文からの接続表現部分の分離の可否を判断しておき、図4のアルゴリズムに従って、文脈構造を Left-to-Right に遷移しながら、接続表現の変更を行なっている。

A 接続表現の変更条件

if	左部分木 L に重要文が存在
∧	右部分木 R に重要文が存在　∧ R の左端の文 C が非抽出
then	R の最初の重要文 X の接続表現を、C の接続表現に変更



B 接続表現の変更方法

if	文 C と文 X がともに接続表現分離可 then 文 C の接続表現と文 X の文内容 (接続表現を取り去った残りの部分) とを出力
if	文 C が分離不可 ∧ 文 X が分離可 then デフォルト接続表現テーブル[1]から得られる、文 C の接続関係に対応する表現と文 X をつなげて出力
if	文 C が分離可 ∧ 文 X の分離不可能 then 文 C の接続関係と文 X を出力
if	文 C と文 X がともに分離不可 then デフォルト接続表現テーブルから得られる文 C の接続関係に対応する表現と文 X をつなげて出力

図 4: 接続表現変更アルゴリズム

以上の接続表現変更処理によって、例えば、「A である。([しかし B である。つまり C である。)]」という文章から、第一文と第三文を抽出した場合に、「A である。つまり C である。」ではなく、「A である。しかし C である。」が生成され、原文章での文間の接続関係を正しく保存する抄録文章を生成できる。

3 評価実験

抄録文章は、読者が原文章より少ない分量を読むことで、その大体の内容を知ることを目的とする文章である。よって、理想の抄録文章とは、「短く、重要な情報を洩れなく伝える。」ものであるとし、抄録文章の評価基準として、以下の2つを設定した。

$$\text{抄録率} = \frac{\text{抄録文章の文字数}}{\text{原文章の文字数}} \quad (1)$$

$$\text{重要文密度比} = \frac{(\frac{\text{抄録文章に含まれるキーセンテンス数}}{\text{抄録文章の文数}})}{(\frac{\text{原文章に含まれるキーセンテンス数}}{\text{原文章の文字数}})} \quad (2)$$

抄録率は、原文章に対する抄録文章の文字数(サイズ)の比であり、この値が小さければより良い抄録となる。また、**原文章のキーセンテンス**とは、複数の被験者に対して原文章を提示し、キーセンテンス(重要文)を指摘させる実験によって抽出した文で、原文章内で相対的に重要な情報を含む文であると考えられる。よって、**重要文密度比**の式の分子および分母は、それぞれ抄録と原文の重要な情報を持つ文の比率を示していることとになり、式全体では、原文章から抽出されたキーセンテンスの占める比率が、抄録文章において、どの程度高まったかを示す指標となる。

3.1 抄録率の評価

さまざまな種類の文章での抄録率を示した表1から、本抄録システムによって、論説文や社説では1/6 ～ 1/5 程度、その他の文種でも 1/4 程度までの短さの抄録を生成出来ることが判る。

表 1: 各文種における抄録率

文種		社説	論説文	論文	コラム	Total
文書数		48	38	24	20	130
サイズ(平均)	文数	35.1	44.2	60.2	21.6	40.3
	文字数	327.0	382.2	991.3	183.2	443.7
抄録率		20.5%	16.5%	23.6%	27.4%	21.5%

3.2 重要文保存の評価

情報性の評価は、1. 複数の被験者に対する試験によってキーセンテンス(KS)を決定し、2. 生成された抄録について、各 KS の捕捉状況を調べることによって行った。なお、情報性の評価は、文章全体が1つの章からなる 20 文書(平均抄録率 41.6%)を対象とした。

3.2.1 被験者によるキーセンテンス抽出

各マテリアルを、それぞれ4 ～ 5名の被験者に提示して、以下の手順で KS を決定し、表2の結果を得た。

1. 段落内で、最も重要だと思われる文を1つ指摘させる(段落 KS 候補)
2. 段落 KS 候補から、文章全体の KS を 1/3 ～ 2/3 の比率で指摘させる(文章 KS 候補)
3. 文章 KS 候補から、文章全体を代表する文を1つ指摘させる(最重要文候補)
4. 各被験者が選んだ KS 候補の内、被験者の過半数の選択した文を、それぞれ、段落 KS、文章 KS、最重要文とする

表 2: 被験者によるキーセンテンス抽出結果

文種	原文数	抽出されたキーセンテンス		抽出された最重要文
		段落	文章	
社説文	187	55(29%)	28(15%)	5(50%)
論説文	218	27(12%)	22(10%)	5(50%)
TOTAL	405	82(20%)	50(12%)	10(50%)

3.2.2 従来方式との比較

下記の別方式を含め、抄録文章に抽出される文の文数が等しくなるようにして比較評価実験を行い、表3に示す結果を得た。

別方式： 各文に含まれるキーワード(名詞句)を抽出し、文章全体に渡る各キーワードの出現回数をそのキーワードの得点とし、各文中のキーワードの得点の合計の大小によって、各文の重要性を判定する[2]。

表 3: 各方式による情報性の評価実験結果

項目		段落	文章	最重要	Total
被験者に対する実験		82	50	10	142
方式	本方式	29 (1.43)	11 (1.26)	6 (1.38)	46
	別方式	26 (1.28)	12 (1.37)	2 (0.46)	40

表3は、各方式によって抽出されたそれぞれの KS の数と、重要文密度比(括弧内)[3]を表しており、本方式では、段落 KS、文章 KS および最重要文のすべてについて、重要文密度比が、26% ～ 43% 向上しており、全体として、別方式に対して優位性をもっていることがわかる。特に、最重要文に関し、別方式では、20 ～ 30% しか捕捉できなかったのに対し、本方式では、その 60% を捕捉している。これは、最重要文に、キーワードが必ず現れる訳ではないために、統計的処理だけでは不十分であることを表しており、本方式での文脈構造解析の利用の有効性を示している。

3.3 分析

抽出されなかった KS を分析し、その原因が、(1) 文脈構造解析の誤り (57.9%)、(2) 意味内容的に単独で重要な文 (14.8%)、(3) 重要性判定方式の誤り (14.8%)、(4) 単文解析の誤り (12.5%) であることが判明した。この内 (1) および (4) は、文脈構造 / 単文解析の高精度化によって対処する。

(2) は、例えば「その死によって、アルバニアにも必ず変化が起こるだろう。」といった文の場合に起こる。これは、原文章においても周囲の文に対し明示的な接続関係を持っていないが、文中に現れる「死」と言った重大な意味を持つ語句の出現や、「必ず～」といった断定的な表現によって、単独で重要な意味内容を持つ文となっている。本方式では、文間の論理関係に基づいた処理によって重要文判定を行なっているので、このようなケースの重要文は抽出出来ない。

(3) は、例えば「なぜなら、問題となっているのは、質の変化であるからだ。」の様な場合である。接続関係解析において、表現「～であるからだ」によって、先行する文に対し「理由」を与える文であることが正しく解析される。しかし、重要文判定において、「理由」部分は、先行する「結論」部分より重要性が低いと判定され、非抽出となる。被験者に対する調査では、この文もキーセンテンスとして抽出されており、「理由」部分がどのような場合にも「結論」部分より重要性が低くなるわけではないことが判明した。

以上の (2) および (3) を解決するには、文の重要性判定において、接続関係のみではなく、例えば、キーワードの情報や話題情報を併用する必要があると考えられる。

4 おわりに

本稿では、文間の論理的 / 修辞的関係に基づいて重要文を判定し、文間の接続関係を正しく表現するように接続詞を変更 / 補足して、まとまりのある抄録を生成する文書自動抄録システムついて報告し、評価実験によって、文章全体の最重要文の抽出に関して、従来方式に対して優位性を持つことを示した。

参考文献

[Luhn 58] Luhn, H.P.,The Automatic Creation of Literature Abstracts, *IBM Journal*, Vol.2, No.4, pp.159-165, 1958.

[Yasuhara 89] 安原,*et.al.*, 要約支援システム COGITO, 情報処理,Vol.30, No.10, pp.1258-1267, 1989.

[Isibasi 89] 石橋,*et.al.*, 英文要約システム「DIET」, 38th 情処春大, 1989.

[Ono 93] 小野, 住田, 知野, 浮田, 日本語論説文の自動抄録のための文脈構造解析, 情処春大, 7B-10, 1993.

[Mann 87] W.C.Mann, Rhetorical Structure Theory : Description and Construction of Text Structure, *Natural Language Generation*, G.Kempen(Ed.), pp.279-300, 1987.

[2]同一文章中に頻出する(名詞句)表現は、その文章の話題に関連するものであり、そのような表現を数多く含む文は、重要文である可能性が高いというヒューリスティックスによる

[3]重要文密度比が 1.0 である時、抄録の KS の割合が原文章と等しく、1.0 を越えている場合は、原文章よりキーセンテンスの割合が高く冗長性が低くなっていることを表す。