キー・センテンスの選択的解析による論文要約手法*

5C−6

西村 健士　島津 秀雄　高島 洋典†

日本電気(株) C&C 情報研究所‡

1 はじめに

文書自動要約の一ステップとして，言語学的な情報をボトムアップに使ったり予め文書内容の枠組を用意しておきこれをトップダウンに使用したりして文書を構造化することがよく行なわれる。残念ながら，科学技術論文など章節構造を持った比較的長い文書をこの手法で構造化するのは現在の技術レベルでは困難である。

しかし，科学技術論文では論理の明解性が重要なので，論文の組み立てを示唆する表現が数多く出現する。科学技術論文の要約システム開発にあたって，我々は，このようなキー・センテンスを選択的に解析することによって，文書を近似的に構造化する方針をとった。すなわち，各文を葉とした精密な木構造に文書を変換するのではなく，文書を話題毎に分割することを目標とした。

まず，科学技術論文における話題(以降，“主題”と呼ぶ)を分類する必要がある。100件の技術論文の序章と最終章を対象として実際に主題の分類調査を行なった。続いて，論文100件の全ての章を対象にして論文の構造化に役立ちそうな文を抽出し，大きく3種類に分類した。続く2章でこれらを解説し，最後に，試作システムの処理の流れと動作例を説明する。

2 論文主題の分類

「情報処理学会第42回全国大会講演論文集3」中の日本語論文100件(3-1 〜 3-201)を対象とし，その序章と最終章の内容を人手で調べて主題の分類を行なった。序章と最終章を選んだのは，論文に書かれるべき主題はこの2つの章に凝縮されているものと期待したからである。

2.1 最終章中の主題

100件中97件の論文が，(「謝辞」，「参考文献」，「付録」を除く)最後の章で論文全体をまとめる記述を行なっていた。章見出しには，「おわりに」，「まとめ」などが多い。

主題の種類	文の数
論文全体の内容を最もよく代表する文	78
従来研究の問題点	7
本研究の効果,特徴	37
本研究の評価結果	11
本研究の問題点	33
本研究のその他の説明	15
研究の進行状況	10
今後の課題	34
今後の予定	45
(上記各主題の)補足説明	50
その他	34
総計	350

上表は主題毎に最終章中の各文をカウントしたものである。一つの文に複数の主題が読みとれる場合には，より表現に重点が置かれていると思われる方へ振り分けた。ただし，テ形や連用中止によって実質的に2つ以上の単文に分割できるような重文については，双方の主題に1ポイントずつ与えた。(このような文は少数。)インデントを用いた列挙表現では各行を1文と数えた。

表中，「補足説明」とは，主題を述べる中心の文ではないが，近くにある(多くはその直前直後)中心文を補足することにより主題の叙述に貢献している文のことである。「補足説明」文は全ての主題分をまとめて50文であった。

上表より，9割の文が上記のいずれかの主題に分類可能なのが分かる。「その他」は，前章からの議論の続き，考察，漫然とした感想などから構成される。

多くの論文には，「内容を代表する文」があること，主題の流れに傾向があること(おおよそ表の上から順に並べられる)，同じ主題の文の表層パターンは非常に似ていること，が分かった。例えば，「内容を代表する文」の多くは「本稿では，....について述べた。」という形をしている[2],[3]。

断っておくが，分類作業者(筆者)の主観が上表の分類に混入していることを否定することはできない。例えば「今後の課題」と「今後の予定」の2つに関してはどちらにでもとれる文もあり，上表の分類は明確に線引きのできるものではない。

2.2 序章中の主題

続いて序章の主題分類を行なった。目次や抄録のある論文もあったが，それらを除く最初の章を対象とした。序章のない論文は1件もなかった。多くは，「はじめに」という章見出しが付けられている。文の数え方は前節と同じである。

主題の種類	文の数
外部環境動向	6
当該分野の課題	34
当該分野で注目されているテーマ	13
従来研究の概要	19
従来研究の問題点	18
筆者らのこれまでの研究の概要	47
筆者らのこれまでの研究の問題点	16
論文全体の内容を最もよく代表する文	92
研究内容の概略	54
論文の構成	12
(上記各主題の)補足説明	196
その他	120
総計	628

最終章に比べると文章の量も多く，主題の構成も複雑である。「補足説明」，「その他」の文の全体に対する割合が高くなっている。「その他」に振り分けた文でも結局はある主題への導入の働きをしているので，「その他」か「補足説明」かの判断は微妙である。その文が存在しなくても主題情報の読み取りになんら支障がなければ「その他」に分類した。

序章に関しても，「内容を代表する文」はほとんどの論文に存在した。主題の出現順にも傾向があり，おおよそ上の表の順である。「内容を代表する文」は章の最後にくることが多い。また，研究概要と問題点はペアで現れることが多い。

表層的なパターンで主題を自動認識するのは最終章に比べるとやや困難である。ただし，「内容を代表する文」に関しては例外である[4]。従来研究や筆者たちの研究の概要/問題点を述べる文についても表現パターンに傾向が見られた。しかし，主題の中心文のみを抽出するのではなく補足的な文を含めた主題の範囲を認識するのは容易ではない。

---

*Text Summarization by Analyzing Key Sentences

†Kenshi NISHIMURA, Hideo SHIMAZU, Yosuke TAKASHIMA

‡C&C Information Technology Research Labs., NEC Corp.

3 主題判定のキー・センテンス

続いて，前章で定めた各主題の記述位置を同定する鍵となる文を論文全体から抜き出し，分類する。

主題を明示的に示すもの

「解析の大まかな手順を示す。」(具体的方式)
「本方式には，...次のような利点がある。」(特徴)
「このようにして解析した結果を...に示す。」(評価)
「以下では...問題点について述べる。」(問題点)

主題の範囲を明示的に示すもの

「...は以下の3種類に分類することができる。」
「...大まかな手順を示す。1...。2...。」

未定義語の定義を与えるもの

「以下，...とは後者を指すことにする。」
「...マシンXXXを開発した。」(固有名詞)
「...とは...のことである。」

科学技術論文にはシステム名などの固有名詞や新概念の名称が頻出するので，未定義語の定義を与える文の解析は重要である。これらの未定義語は章節の見出し中にもよく現れるため，章節構造のみで主題判定を行なうアプローチには無理がある。

論文全体を緻密に解析しなくてもキー・センテンスのみに注目することにより主題の把握が可能である例を図1に示す。この文書は前述100論文中から選んだ(4C-3, 3-65頁)。なお，箱の部分は未定義語を表している。

4 試作システム

最後に試作要約システムについて簡単に説明する。

4.1 処理の流れ

図2に処理の流れを示す。入力はべた書きの論文である。まず，形態的な特徴をもとに，表題，著者名/所属，章節見出しなど認識を行ない，続いて序章，最終章，本論部分の順に以下の処理を行なう。

まず，キーワード照合によって各文がキー・センテンスであるか否かを判定する。多くの文はヒットしない。この最初の処理によって処理速度向上を図っている。続いて，未定義語を定義するキー・センテンスを解析し，未定義語の属性を推定して推定結果をルールベースに挿入する。最後に，主題決定用のプロダクション・ルールを参照して各キー・センテンスを解析し，論文各所の主題決定を行なう。

4.2 処理例

図3に処理例を示す。左下のウインドウが初期ウインドウである。このウインドウで論文ファイルを読み込み解析ボタンをクリックすると，論文の主題構成が分析され，右上のメニューウインドウが生成される。メニューウインドウ上で"**"に囲まれた文字列が主題を，それ以外の文字列が章節の見出しを表している。ユーザが主題文字列をクリックすると，その主題に対応した原文部分が左下のウインドウ上に反転表示される。

本試作システムで抽出可能な主題の種類はわずかで，ルール規模もまだ小さい。今後サンプル文書を増やしてルール規模が収束するか確認する作業を行なう。また，主題に対応した原文部分を指し示すだけでは正当な意味での「要約」とはいえない。ユーザの望む量まで原文を圧縮し，適当な形態に変形加工するという課題が残っている。

5 おわりに

科学技術論文の要約システム作成を目標として，論文の主題分類を行ない，キー・センテンスを選択的に解析することにより主題分布を認識する方式を提案した。

参考文献

[1] 西村他，科学技術論文要約システムの開発環境，情処第43回全国大会3, pp.181, 1991

[2] 矢島他，文書への意味属性付与のための意味辞書の開発，情処第43回全国大会3, pp.325, 1991

[3] 岩井他，意味解析を用いた文書構造化手法，情処第43回全国大会3, pp.327, 1991

[4] 江原，抄録化のためのトリガ語の分析，情処理第42回全国大会3, pp.180, 1991

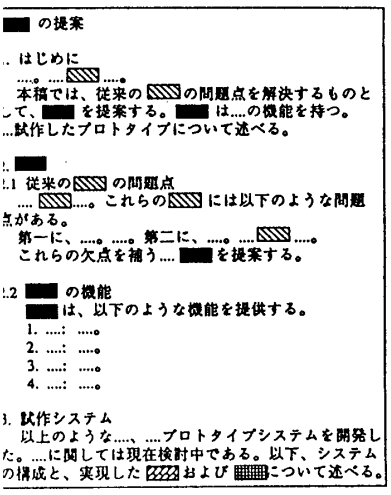

図1 サンプル文書

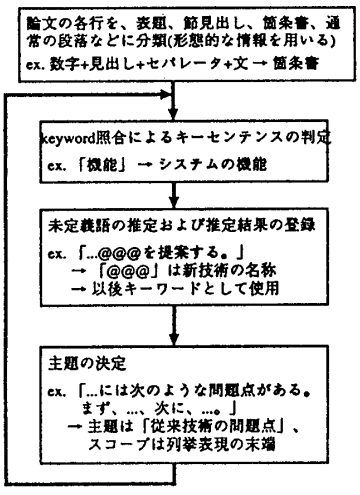

図2 試作システムの処理の流れ

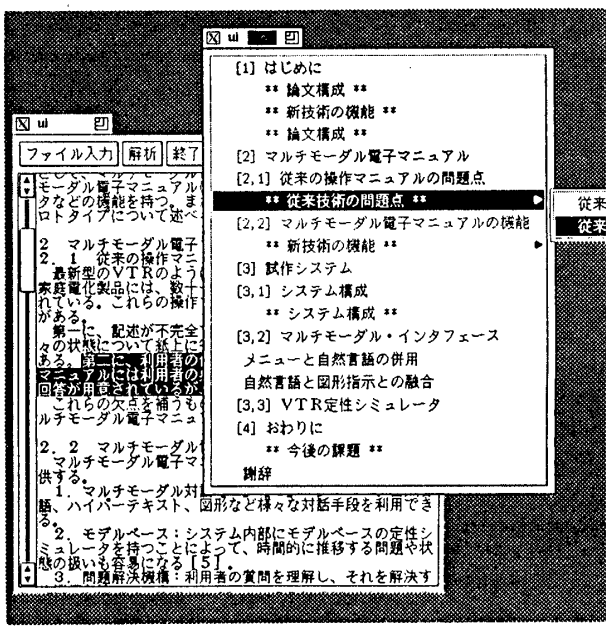

図3 試作システムの動作例