

特定表現の重点的解析による科学技術論文構造化手法

西村 健士 島津 秀雄

NEC C&C 情報研究所

科学技術論文の主題構成を計算機で認識する一手法を提案する。まず、表層的な解析のみによって主題判定が可能な文を処理し、次に、その文を核としてその近傍の文との主題の連続性を解析することにより、論文全体の主題分布を自動認識できることを示す。主題分類は論文から情報を抽出しようとする読者の観点で行なった。本手法の特徴は、章節構造を含む長い文書を迅速に処理できること、表層解析のみを行っているので読み手にとって処理内容が明快なことである。

AUTOMATIC TEXT STRUCTURING METHOD FOR SCIENTIFIC OR TECHNICAL PAPERS BY ANALYZING KEY SENTENCES

Kenshi Nishimura Hideo Shimazu

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

A text structuring method for scientific or technical papers is proposed. Topics are introduced as a reader's view who extracts information from the papers. Clue expressions analysis is used for judging the topic of sentence. Topic distribution of the whole text by topic continuance analysis between neighboring sentences is described. The proposed method is applicable to the fairly large-size texts and easy for the reader to understand the results.

1 はじめに

我々は数ページから数十ページの比較的長い科学技術論文を対象として、文書中から重要な情報を抽出する方式を検討している [1][2]。

従来の文章要約研究のほとんどは数文から数段落の比較的短い文章を対象としており、言語的知識や領域知識を駆使して深い文章解析を行なっているものが多い。一方、文章の長さを問わない情報抽出技術としてはキーワード自動抽出があげられる。これは形態素解析程度の比較的浅い処理で実現できるが、抽出される情報は単語の集合であり、文章中の情報の多くが捨て去られてしまう。

我々が対象とするサイズの科学技術論文には、既存技術の概要や問題点、主張したい結論やそれを裏付ける事実、あるいは論証の過程など様々な情報が含まれている。これらを「主題」と呼ぶことにする。

我々の情報抽出の目標は、具体的には、

- 単語レベルではなく主題レベルで情報を抽出する
- 使用する技術は表層的言語解析技術に限定する

ことである。解析手法を表層的なものに限定したのは、

- 領域への依存性を少なくして移植性を高める
- 解析部のブラックボックス化を避けて、ユーザ適応性を向上させる
- 全文の解析を現実的な時間内で行なう

ためである。

本稿では、論文全体を主題の列とする簡明な論文構造モデルを提案する。この論文構造モデルのもとでの論文構造化とは、論文の各文がどの主題について述べているのか、逆に言えば、ある主題に関する記述は論文のどこにあるのかを解析することである。この構造化が済めば、特定主題に関する情報を論文から容易に抽出することができる (図 1)。

以下、次の 2 節でそれぞれ「論文構造モデル」、
「論文構造化方式」を説明する。その後既存研究との比較を行なう。

2 論文構造モデル

2.1 主題の分類

我々は、情報処理学会全国大会論文集から論文 100 件を選び、先頭節と最終節でどのような主題が述べられているか調査した [1]。その結果、ほとんどの文が有限種類の主題のどれかに対応付けられることが分かった。表 1 にその様子を示す。

表 1 において、原則的には句点 (".") によって区切られた単位を 1 文として数えたが、テ形や連用中止によって実質的に複数の単文が連結されたものや改行、段付けされた列挙表現が 1 つの文として書かれているものは、その構成要素の一つ一つを文として数えた。また、1 つの文に複数の主題が読み取れる場合には、より重点的に表現されていると解釈できる方にポイントを与えた。従って、表中の文の数は形式的に数えた文の数とは必ずしも一致しない。

主題「論文全体の中心文」とは、例えば「本稿では...について述べる。」などの表現でその論文全体で筆者が記述する / したことを 1 文で簡潔に表現している部分である。主題「論文の構成」は「以下、次節で...について、第 3 節で...について述べる。」などの表現で、その論文の構成を予告するものである。以上の二つの主題は他の主題のように論文の実質的内容を担うものとは性質が異なる。

「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 (多くの場合その直前直後) その主題の中心的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。「その他」には、前章からの議論の続き、後続節で使う基本用語の定義、漫然とした感想が含まれる。「その他」に振り分けた文の中には、結局はある主題の導入の働きをするものもあり、「その他」か「補足説明」かの判断は微妙である。その文が存在していなくても主題の読み取りになら支障がなければ「その他」に分類した。

主題の出現傾向に関しては、流れにパターンがあること、共起しやすい主題があること (後述) も観察された。

同様の調査を他分野の論文に関しても行い、やはり各文のほとんどが表 1 のどれかに属することを確認した [2]。題材として選んだのは、「日本機械学会論文集」の A 編、B 編、C 編の各々から 5 件と「応用物理」からの 15 件、計 30 件の論文である。先頭節と最終節の主題構成にはやはり規則

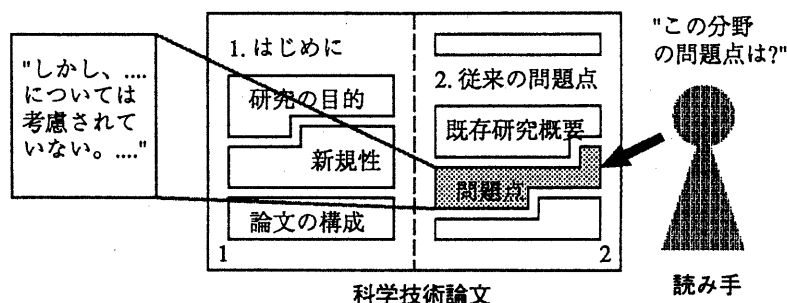


図 1: 主題レベルの情報抽出

性が見られた。その間の各節の文もいずれかの主題に分類することは可能だが、主題構成に階層構造、入れ子構造が見られ、先頭節、最終節よりも自動認識が難しい。

2.2 主題表現の特徴

我々はまた、表層レベルのマッチングで多くの文の主題判定が可能なことを見出した [1] [2]。表 2 にその例を示す。表 2 中の各表現は論文の分野に依存しないことが分かる。

3 論文構造化方式

3.1 処理手順

本節では計算機による論文構造化方式について説明する。処理は大きく 2 つのフェーズに分かれ、5 つの知識 (後述) が参照される。

1. 特定文の主題の判定

5 つの知識のうち最初の「主題を明示する表層パターン」を用いて、その文自身の主題を解析する処理 (ステップ 1)

2. その主題のスコープの解析

残りの 4 つの知識「主題の範囲を明示する表層パターン」、「主題構成に関する経験的知識」、「文の接続に関する言語的知識」、「文章の記述形式レベルの経験的知識」を順に適応して、その主題がどの範囲の隣接文まで共有されているか解析する処理 (ステップ 2 ~ 5)

主題の種類	頭	終
論文全体の中心文	92	78
論文の構成	12	
外部環境動向	6	
分野の課題・テーマ	47	
従来研究の概要	19	
従来研究の問題点	18	7
筆者等の従来研究の概要	47	
筆者等の従来研究の問題点	16	
研究内容の概略	54	15
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
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	196	50
その他	120	34
総計	628	350

「頭」…先頭節での文の数

「終」…最終説での文の数

表 1: 論文の主題分布

「本稿では、... について述べる。」(論文全体の中心文)
「次節で..., 次に..., さらに... 報告する」(論文の構成)
「近年... なってきている。」(外部環境動向)
「... が重要な課題となっている。」(分野の課題, 問題点)
「... らは、... であることを示している。」(既存研究の概要)
「先に著者らは... ことを示した。」(筆者等の従来研究の概要)
「しかし、... については考慮されていない。」((筆者等の)従来研究の問題点)
「本方式には、... 次のような利点がある。」(本研究の効果・特徴)
「以下では... 問題点について述べる。」(本研究の問題点)
「以下に今後の課題を示す。」(今後の課題)
「解析の大まかな手順を示す。」(具体的なアルゴリズム)

表 2: 表層表現の例

3.2 参照知識

読み手に対する処理の透明性, カスタマイズの容易性を重視する方針から, 特殊な知識は極力使わないようにした。5つの知識は以下の通りである。

1. 主題を明示する表層パターン

その文の属している主題を直接的に表現しているパターン(表 2)。

2. 主題の範囲を明示する表層パターン

論説調の文章には論旨を整理するために列挙表現が多用される。その列挙表現内はあるまとまった主題についての記述だと期待できるから, 列挙のスコープ判定はそのまま主題スコープの判定につながる [1] [3]。

列挙のサイズは様々で, 各項目が節や段落であるもの(レベル 1), 各項目が行であるいわゆる箇条書き(レベル 2), 各項目がベタに続いているもの(レベル 3)がある。我々は, 今後, LaTeX のように文書構造を示すコマンドが埋め込まれた電子化ドキュメントが一般的になってくると予想している。その中では, 上記のレベル 1, 2 の列挙のスコープは曖昧性無く解釈できるので, 今回は処理対象として扱わないことにした。実際, LaTeX の itemize 環境などを文書構造化の情報として利用する例が [4] の中に見られる。

我々が対処しなければならないのは, レベル 3 の列挙表現であり, 具体的には以下の

ような表現パターンである。

「その利点は次の 3 点である。1)...。2)...

まず, 後続する列挙表現全体の主題が予告され, その後各項目が (1), (2), ... などの数字や, 「まず」, 「次に」などの接続表現で続く。このような表層パターンに関する知識である。

3. 主題構成に関する経験的知識

技術論文に特有の主題構成に関する傾向を利用する。そのうちのいくつかを以下に紹介する。

● 共起する主題

例えば, 「従来技術の概要」の次に「従来技術の問題点」が述べられることが多く, 境界となる文(「従来技術の問題点」の最初の文)の先頭には逆接表現の来ることが多い。

● 章節構造との関連で一定の位置を占める主題

「論文の構成」は各章各節のはじめか終わりに来る場合が多く, それ自身の主題判定に寄与するばかりか, 後続する主題を暗示する場合もある。また, 「外部環境動向」は第 1 節の冒頭に出現する場合がほとんどである。

4. 文の連結に関する言語的知識

接続表現から2つの文の結合が強いと推測される場合は、双方の主題を同一化する。例えば、「具体的には...」、「これより...」などの表現がこれにあたる。また、提題表現などによって焦点をあてられた名詞句の連鎖も主題伝播の手掛かりにする。例えば「図5は...解析結果を示したものである。図5において、...」の2文では「図5」という名詞を頼りに後の文の主題を前の文の主題と同一と解釈する。これらの知識は絶対的なものではなく、前述の3つの知識を用いても主題が決まらない文に対して適応するものである。

5. 文章の記述形式レベルの経験的知識

以上の4種類の知識を用いても主題を決定できない文については、前の文の主題を継承するものとする(デフォルト・ルール)。前に文が無い場合は後ろの主題を継承させる。ただし、継承の範囲は形式段落内とする。段落内の全ての文の主題が不明の場合は、節の見出しから主題を判定する。

3.3 表層パターンの表現方法

これまで例を示してきたように、表層パターンの多くは文のはじめと終わりの数文節(連体指示詞、形式名詞などを数えずに2~3文節程度)で定義すれば十分である。各パターンは属性付けされた文節のリストで表現されており、処理対象論文の各文は形態素解析された後にこれらの表層パターンと照合される。例えば、主題「論文全体の中心文」の表層パターンの内部表現を簡略化して述べると以下ようになる。

最初の文節に関して、自立語概念が“報告”か“原稿”か“研究”であり、接頭辞「本」を持ち、提題助詞「は」を持つ、かつ、
主述部に関して、自立語概念が“報告する”か“説明する”か“提案する”か“述べる”か“考察する”か...であり、
テンスが過去でないもの。

3.4 解析例

表3に解析例を示す。題材は前述した「日本機械学会論文誌A編」から無作為に最初に抽出した論文である(尾野他、「多孔質体を伝ばする衝撃

応力波の減衰効果」Vol.58, No.551, pp.1055)。同論文の長さは図表も含めて5ページである。

表3において、左側は決定した主題の種類、右側の括弧内の数字はその文の主題を決定した知識の種類(つまりステップ番号)である。

ステップ番号の先頭に×がついているのは、主題を間違って判定したものである。決定ステップが45となっているものは、ステップ4で前文(あるいは前々文)と同じ主題と判定されたが前文の主題が未確定のため、ステップ5ではじめて具体的な主題が決まったものである。また、第4節で決定ステップが12となっているものは、ステップ1では独立して「得られた知見」と判定されたものがステップ2で前の文の主題のスコープに含まれると再判定されたものである。結果的にはこの文の主題は変わらない。決定ステップが14となっている文も同様である。

2.1節の「実験の何らかの説明」はステップ5で節の見出しから1文目の主題が決まり、それが後続の5つの文に伝播したものである。この部分以外では節の見出し情報は不要であった。

4 関連研究

最後に、本研究の特徴を明確にするために、特に説明的文章、論説文を対象とした他の要約研究との比較を行う。

対象文書のサイズ 前述したように、比較的短い文章を対象とした研究例は多いが、章節構造を持つような長い文書を対象とする研究は少ない。情報抽出の一種としてキーワード抽出も含めればその限りではないが、章節構造の下部構造としての文書の断片を抽出する研究としては[5]が挙げられる程度である。

抽出単位 抽出するデータの大きさは様々である。対象文書に述べられている内容をフレーム形式で定義しておき、そのスロットを埋めるというもの(例えば[6])や、何等かの基準で文単位に重要性を評価し、重要度の高い文を抽出する研究(例えば、

1. 緒言

筆者等の従来研究の概要	(1) 著者らは... 研究を行ってきた。 (4) これまで... 観察した。 (4) その結果... 説明してきた。
論文全体の中心文	(1) 本研究では... 明らかにすることにした。
論文の構成	(3) 具体的には... を行う。 (3)... 調べる。 (3) 次に... 説明を行うことにする。

2. 多孔質体モデルによる実験

2.1 供試片

実験の何らかの説明	(5)... (45) この供試片... (45) これは... (45)...dとして... (45) これより... (5)...
客観データ	(1) なお, 表 1 に... 機械的性質を示す。

2.2 実験装置と方法

実験装置の説明	(1) 図 2 に実験装置の概要を示す。 (5)... (45) この衝突棒の... (5)... (45) なお, この衝撃速度 V ...
---------	---

2.3 測定結果

実験結果の説明	(1) 図 3 に実験結果の一例を示す。 (5)... (5)... (45) なお, 持続時間 t は...
---------	---

この節は以降「客観データ」「得られた知見」の 3 回繰り返し (ほとんど第 1 ステップで決定)

3. 理論解析と考察

本研究の評価結果	(1)... その結果を評価することにする。
----------	------------------------

3.1 一次元伝播

モデルの説明	(1) 本実験に対応した... モデルを図 6 に示す。 (4) すなわち...
既知の知見	(1)... は次式で与えられることが知られている。 (4)...[1] ($\times 4$) ここで... ($\times 5$)...
客観データ	(1) その結果表 4 の各値が得られた。 (1) そこで表 1 と表 4 の値を用いて... 式 [1] の値を求めると表 5 となる。
得られた知見	(1)... ことがわかる。 (1)... ことがわかる。 (5)...

3.2 DFEM(動的有限要素法) 解析と評価

得られた知見	($\times 5$)... (1)... と考えられる。 ($\times 5$)...
--------	--

この後「モデルの説明」「客観データ」「得られた知見」が続く。(ほとんど第 1 ステップで決定)

3.3 考察

考察	(1)... 比較検討する。 (4)... 減少率については...
既知の知見	(1)... あることが知られている。 ($\times 5$)...
客観データ	(1) 結果を図 9 に示す。
得られた知見	(1)... ことがわかる。 (5)... (45) さらに... と思われる。 ($\times 45$)... (1) これより... こともわかる。

4. 結言

論文全体の中心文	(3)... 調べた。
得られた知見	(1) その結果, 以下のことが分かった。 (2)[1]... (2)[2]... (12)[3]... ことが分かった。 (14) また... ことが分かった。 ($\times 14$)...
謝辞	(1)... 謝意を表する。

表 3: 解析例

[7] [8] [9] [10])がある。後者では複数の文が抽出されてもそれらの間の関係は考慮されていない。論理の一貫性、結束性を保つように文を選択・修正する研究には、例えば、[5] [11] [12] などがある。[13] は表形式、文の抜粋、文章生成など様々な抽出・提示形態を提案している。我々のように各文を主題別に分類するのに近い研究例には [14] があるが、[14] では各文をドメイン特有の主題に分類しており、我々のシステムと分類基準が異なる。

使用する知識 表層レベルの解析技術(パターン照合中心)を基にした研究には [7] [8] [9] [10] などがあり、これらは「抽出単位」の部分で文単位に抽出を行なう研究として紹介したものと一致する。まとまりのある文章として抽出しようとする表層的な解析では済まず、例えば [12] [11] [3] などでは主に文章の構造に関する日本語の知識や論説文での論証構造に関する経験的知識を使って文章を構造化し、構造の各部分の重要性を評価して、抽出を行っている。抽出までは行っていないが、同様の文章構造化手法を検討しているものに、[15] [4] などがある。文章内容に関連した知識を用いて文の格解析にまで踏み込んだものには [14] [13] などがある。[5] では知識を詳細に記述したフレームを用意しておき、章節の見出しでそれを検索し、文章構造化処理にトップダウンに使っている。我々は、このように複雑な知識は、読者に対する処理の透明性を重視する立場から排除した。

次に我々の研究に近いものを紹介する。これらは、主題の集合としての論文構造モデルや表層パターンによる主題解析の有用性を示すものである。

神門は日本語の医学・物理学・経済学・国文学領域の原著論文 127 件を対象とした調査で、論文中の全ての文をある主題に分類できたことを報告している [16]。我々が「主題」と呼んでいる局所的な話題のことを神門は「構成要素カテゴリ」と呼んでいる。構成要素カテゴリの基本構造は「問題・問題に答える証左・解答」の 3 つ組であるが、階層構造の部分構造中に、例えば、「既知の事実

の要約」、「研究課題」、「操作・手法」などのカテゴリが定義されており、我々の主題分類と類似している。神門はカテゴリの出現パターンに規則性が見られること、接続関係や文末表現などが主題判定の手掛かりとして有効であることも報告している。

江原は講演論文約 1,300 件を対象として「論文全体の中心文」がどのような言葉で始まるか分析している [17]。「本稿では」、「本論文では」などの表現が多く、約 7 割の論文がこれで概要を把握できると報告されている。中本らは技術解説文章を対象として、「目的」、「方法」、「結果」、「背景」、「意見」、「特徴」、「課題」、「内容紹介」の表現パターンを抽出/評価し、構文木レベルでのマッチングで多くの文が抽出できることを報告している [18]。

表層パターンによる抽出方式を実際に計算機上に実装した研究例としては、[9] [10] のように、英語の論説文章の重要文を判定し、原文全体の中で該当部分を強調表示するものがある。抽出文章の一貫性の欠如はユーザインタフェースの工夫でカバーしている。特に、[10] では抽出規則のユーザカスタマイズについても論じている。

5 まとめ

本稿では科学技術論文の構造化手法について述べた。まず、実際の論文を対象とした分析により、論文全体が主題の列としてモデル化でき、しかも、論文各部の主題は表層的な解析によって自動認識できることを示した。次に、解析は、まず最も確からしい文の主題を決定し、その後その回りの文の主題を決めていくことによって進むことを実例を 1 つあげて説明した。最後に、本方式は、長い論文を対象としており、表層的な情報のみを使って、主題レベルの情報抽出が行なえる、という点で従来の研究より有利であることを論じた。

参考文献

- [1] 西村他. キー・センテンスの選択的解析によ

- る論文要約手法. In 情報処理学会第44回全国大会論文集 3, 1992.
- [2] 西村他. キー・センテンスの選択的解析による論文要約手法の分野 移行性評価. In 情報処理学会第45回全国大会論文集 3, 1992.
- [3] 小野他. 日本語論説文の自動抄録のための文脈構造解析. In 情報処理学会第46回全国大会論文集 3, 1993.
- [4] 中島他. 表層表現の解析に基づく論文改訂支援方法について. In 信学技法 NLC92-43, 1992.
- [5] 高松他. 見出し情報を用いたテキスト解析と情報抽出. In 情報処理学会論文誌 Vol.29 No.8, 1988.
- [6] 4th Message Understanding Conference, 1992.
- [7] 森本. 重要文抽出と不要語削除による抄録作成方式. In 情報処理学会第42回全国大会論文集 3, 1991.
- [8] 喜多. 説明文を要約するシステム. In 情処研報 自然言語処理 63-6, 1987.
- [9] 石橋他. 英文要約システム「diet」. In 情報処理学会第38回全国大会論文集, 1989.
- [10] 野崎他. 文章の特徴を抽出するための一手法. In 情報処理学会第45回全国大会論文集 3, 1992.
- [11] 知野他. 日本語論説文の自動抄録システムの試作と評価. In 情報処理学会第46回全国大会論文集 3, 1993.
- [12] 田村他. 文章の表現形式に基づいた要約文章の生成について. In 情処研報 自然言語処理 92-1, 1992.
- [13] 安原他. 要約支援システム cogito. In 情報処理 Vol.30, No.10, 1989.
- [14] 稲垣他. 事象型要約情報抽出システム. In 情報処理学会第44回全国大会論文集 3, 1992.
- [15] 福本他. 文の接続関係解析に基づく文章構造解析. In 情処研報 自然言語処理 88-2, 1992.
- [16] 神門. 構成要素カテゴリを用いた原著論文の内部構造分析. In 情処研報 情報学基礎 25-7, 1992.
- [17] 江原. 抄録化のためのトリガ語の分析. In 情報処理学会第42回全国大会論文集 3, 1988.
- [18] 中本他. 文書への意味属性付与のための意味辞書の拡張. In 情報処理学会第45回全国大会論文集 3, 1993.