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*

5C−3

西村 健士　島津 秀雄†

日本電気(株) C&C情報研究所‡

1 はじめに

われわれは，科学技術論文中に記述されている種々の情報 —分野の課題，研究手法，新規性，応用のヒントなど(これらを“主題”と呼ぶことにする) — を予め設定しておき，各主題が論文のどの部分に書かれているか自動的に判定し，節のタイトルと各主題とをメニューとして提示する要約システムを提案した[1]。同システムでは，読み手がメニューの一項目を選択すると，その主題に対応する原文部分が反転表示される(図1)。

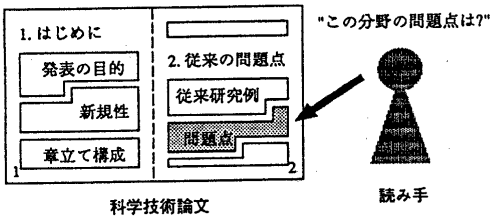

図1: 要約システムの概念図

図1のシステムを構築するに当たっては，「情報処理学会第42回全国大会講演論文集3」中の論文100件(以降，題材Aと呼ぶ)の序節と最終節を対象に，主題の分類作業と主題自動認識手法の検討を行なった。その結果，「以下に手順を説明する」など，主題の記述範囲を明示的に示す表現(“キー・センテンス”と呼ぶ)を集中的に解析するのが有効であることを報告した[1]。図2にシステムの主題認識部の処理フローを示す。図2において，表現パターンの照合は形態素解析後に文節に付与された各種属性(主に主題表現や主述語の自立語概念)をもとに行なわれる。

題材Aの論文は情報工学分野に限られていた。本稿では，

1. 主題分類とその論文中での出現パターン
2. 主題自動認識におけるキー・センテンス利用の有効性

の2つに関して本手法の分野移行性を評価したので報告する。

新たな題材として，「日本機械学会論文集58巻551号」(1992-7)のA編，B編，C編のそれぞれから5件(以降，題材Bと呼ぶ)，「応用物理第61巻」(1992)の第1号から第7号までの「研究紹介」欄から15件(以降，題材Cと呼ぶ)，計30件の論文を無作為に抽出した。題材C中には，最新研究成果の発表を目的とするものの他に，最近の技術動向の概説記事も含まれている。

以下，序節と最終節に分けて検討結果を説明する。

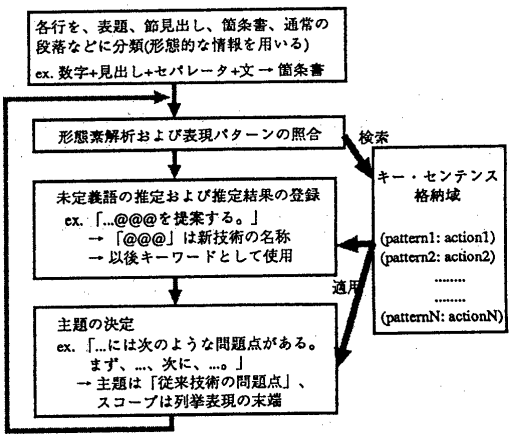

図2: 主題認識部の処理フロー

2 序節の主題構造とキー・センテンス

題材Aでは，序節中に特定の主題が特定の順序で現われる傾向があったが[1]，題材B，Cについても同様の傾向が見られた。また，キー・センテンスとして扱える表現も多く発見できた。以下，主題の出現パターンに沿って例を示す。

1. 動向，課題

「近年，... 必要性が高まってきた。」
「... が重要な課題となっている。」

2. 従来研究の概要

「... らは，... であることを示している。」
「先に著者らは... ことを報告した。」

3. 従来研究の問題点

「しかし，... が問題になる場合が多く，...」
「... については考慮されていない。」

4. 全体の主題

*The Evaluation of Portability of Text Summarization Method by Analyzing Key Sentences

†Kenshi NISHIMURA, Hideo SHIMAZU

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

「本稿 / 本研究 / 本報では，...。」

5. 本研究の概要，成果

「この方式の特徴は... である。」
「その利点は次の 3 点である。 1...。 2...」

6. 論文の構成

「... 検証した。また，... 検討を加えた。さらに... 報告する。」

主題の種類と出現パターン　分析の結果，序節中の主題は [1] と同じように分類でき，主題の出現順序も 30 件全部がほぼ上記のパターンに収まった。若干見られた例外を以下に示す。

最後に「従来研究」について再び言及したもの (1 件)

「論文の構成」が 2ヶ所以上に分散して記述されているもの (2 件)

後半の半分以上を実験試料の説明にさいているもの (1 件)

キー・センテンスの有効性　キー・センテンスは，口語的表現を多用した随想的な文章 (題材 C に 1 つ) を除けば，題材 B，C のどの題材にもよく現れた。既にいくつか例を示した通りである。しかし，主題表現や主述語の自立語概念をもとにした表現パターン照合では主題を推定できない例もあった。例えば，「従来研究の問題点」について，

「... できない。」
「... 計算量が膨大となる。」

という表現が唯一主題決定の情報となる場合があった。最初の例は可能の否定表現で，それをもとにこの部分が何らかの問題点を述べていると推定することも無理ではない。しかし，後の例では，「計算量が膨大となる」ことは数値計算を行なっている研究者の間では問題とみなされるという一般的な推論が要求される。

3　最終節の主題構造とキー・センテンス

最終節中においても主題の出現パターンに題材 A との共通性が見られた。キー・センテンスも多く発見できた。以下に例を示す。

1. 全体の主題

「... 本研究では，... を構築した。」
「... について検討を行なった。」

2. 本研究で得られた新たな知見，効果，特徴

「... ことができ，ことも確かめられた。」
「... 以下のことが分かった。 1...。 2...」

3. 本研究の問題点

「もちろんなお... 問題がある。」
「... にはいまのところ対応できない。」

4. 今後の課題

「以下に今後の課題を示す。 (1)...。 (2)...」
「今後は... 実験が必要である。」

主題の種類と出現パターン　最終節については，題材 B と題材 C に関して顕著な違いが見られた。

題材 B は表現が簡潔で節全体の文章量が少なく，主題も明確であった。15 件のほとんどは「全体の主題」と「本研究で得られた新たな知見」のみから構成され，後者に箇条書きが用いられているものが 9 件もあった。「本研究の問題点」，「今後の課題」に関する記述のあるものは，それぞれ 1 件と 2 件にすぎない。また，「全体の主題」の直後に「論文の構成」ついて述べているものが 2 件あった。

題材 C の 15 件のうち，論文全体をまとめる機能をもつ最終節があったのは 13 件だった。上記の 4 つの各主題はおおよそ 13 件の全てに見られたが，それ以外の主題に該当する記述も見られた。その内訳は以下の通りである。

「従来研究」についての記述を繰り返すもの (1 件)

該当分野への今後の期待など漠然とした感想を最後に述べているもの (3 件)

その他，成果の産業分野への応用可能性など研究内容と直接関係の無いことを述べているもの (2 件)

題材 C には技術紹介記事が多いので，このような結果になったのであろう。

キー・センテンスの有効性　題材 B ではキー・センテンスは有効に働く。書かれている主題の数が少ないので，むしろ主題自動認識の必要性は小さいともいえる。

題材 C でも上記の 4 つの主題に関してはキー・センテンスによる解析が有効であることが分かった。しかし，例えば，「漠然とした感想」という主題に関しては，

「... 期待したい。」

など，あまり意味の明確でない述語表現に着目するくらいしか認識手法が見い出せなかった。

4　おわりに

序節と最終節を対象にして，複数分野の科学技術論文間で，主題構造に共通性が見られるか検討した。また，特定表現パターンに着目した主題自動認識手法の有効性についても同様の検討を行なった。その結果，おおよそ主題構造に共通性を見い出すことができ，主題認識手法も有効に働く場合の多いことが分かった。

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，情処第 44 回全国大会 3, pp.313, 1992

[2] 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，情報処理学会情報学基礎研究会資料 25-7, 1992