論文の内容の要旨

論文題目　**Studies on Thematic Hierarchy Detection and Its Application for Text Summarization**

話題階層の検出とテキスト要約への適用に関する研究

氏　　名　仲尾 由雄

本論文では、テキスト中の話題階層を検出するアルゴリズムと、テキスト要約における話題階層の利用について論じる。本研究で取り扱う話題階層とは、テキストの各部分が何について書かれたものであるかを階層的に表現したテキスト構造である。本稿では、このような話題階層を、語彙の反復による結束性（語彙的結束性）を手がかりとして、自動的に検出する新しいアルゴリズムを提案する。そして、話題階層の 2 つの利用方法について論じる。一つは、単一テキストに含まれる主要な話題の検出であり、もう一つは、複数テキストに含まれる関連話題の抽出である。

テキスト要約においては、適切な粒度の話題を柔軟に検出する技術が求められる。例えば、計算機上で電子書籍を拾い読みしようとしている利用者に対して、最初に提示する要約としては、主要な話題を幅広く含む簡潔な要約が適当と考えられる。その要約で、利用者が読みたい話題を見つけた場合には、その話題の記述箇所を、より詳細に要約して提示することが適当であろう。このような場合、話題階層を利用すれば、主要な話題の検出や、より高度な処理の前処理として、指定された話題に対応する適切な大きさの箇所を切り出すことが実現できる可能性がある。

本論文では、テキストから適切な粒度の話題を検出する基礎技術として、語彙的結束性に基づく話題階層検出手法を提案する。提案手法の特徴は、語彙の反復だけを手がかりに、テキストをほぼ同じ大きさの区画に分割する点にある。これにより、テキストに含まれる様々な話題のまとまりを、テキスト全体より少し小さい程度の話題のまとまりから、段落程度の話題のまとまりまで、体系的に検出することができる。提案手法の評価として、3 種類の長めの文書を対象に、文書の論理構造と検出結果とを比較したところ、検出した話題階層は、文書の論理構造とよく一致していることが観察された。この結果は、本手法が、様々な粒度の話題を正しく検出できることを示唆していると解釈できる。また、情報検索のテストコレクション用いて評価実験を行ったところ、話題階層に基づき検出した重要語は、少なくとも、新聞記事のリードパラグラフに含まれる重要語と同等以上に、検索結果の関連性判定作業を支援する上で有用であることが示された。これらの結果は、話題階層の利用により、適切な話題を幅広く抽出する上で有効なことを示唆すると解釈できる。

本論文では、また、複数の関連文書から関連箇所を抽出する手法を提案する。比較する文書対のそれぞれについて検出した話題階層を、各層を構成するテキスト区画を単位に比較し、関連度の高い区画の対を抽出する手法である。この手法は、抽出区画対の対応関係の正しさ、抽出話題の網羅性・簡潔性、および、主要な話題のカバー率という 3 つの観点から評価した。国会における代表質問と答弁を使った実験では、抽出区画対の約 8 割が正しく同一の話題に対応し、また、新聞に要旨として掲載された内容の約 6 割は抽出された関連箇所の対から読み取れることがわかった。この結果は、複数の話題が混在する文書同士を比較し、話題の関連する箇所を見いだす上で、話題階層の利用が有効なことを示唆すると解釈できる。