セマンティックアーカイブ活用のための基礎技術としてのテキストマイニングとテキスト自動要約

増山繁，酒井浩之 (第７工学系, セマンティックアーカイブコア)

1　はじめに

セマンティックアーカイブ活用のための基礎技術として，テキストマイニングとテキスト自動要約を取り上げ，最近得られた成果を報告する．

テキストマイニングに関して，新聞記事コーパスから交通事故原因表現を抽出する手法を提案した．本手法は，事象の原因表現抽出にかなり汎用的に応用できるものと思われる．また，ヒューリスティックスによる方法や教師あり学習による方法と異なり，人手をほとんど要さないのが本手法の特徴である．

テキスト自動要約に関する研究では，検索結果から成る文書集合を対象として，ユーザの興味に適合するため，文書集合から抽出したキーワードをユーザに提示し選択させることでユーザとのインタラクションを導入した複数文書要約システムを開発した．更に，上記複数文書要約手法の応用として，論文からの発表用スライド自動生成の研究を行った．更に，箇条書きの自動生成を試みるなどの成果を得た．

2　交通事故事例に含まれる事故原因表現の新聞記事からの抽出 [1]

新聞記事の電子テキストデータに含まれる交通事故を扱った記事から，事故原因に関する情報として事故原因を表す表現（例えば,「ハンドル操作を誤った」）を自動的に抽出する手法を提案する．

テキストコーパスから所要の表現を抽出するタスクは，テキストマイニングにおける重要課題のひとつである。従来，個々の応用ごとに，人手で作成したルールに基づくか，または，学習データを人手で用意し，教師あり学習を用いて情報を抽出していた．本研究では，最初に極少数の手がかり表現を人手で用意するだけで，後は，自動的に実行されるブートストラップ的手法を提案した．

3　ユーザの要約要求を反映するためにユーザとのインタラクションを導入した複数文書要約システム [2]

本研究では，複数文書要約において，ユーザが知りたい情報を「要約要求」と定義し，要約要求を反映した要約 (すなわち，ユーザが知りたい情報を含む要約) を生成するために，ユーザとのインタラクション

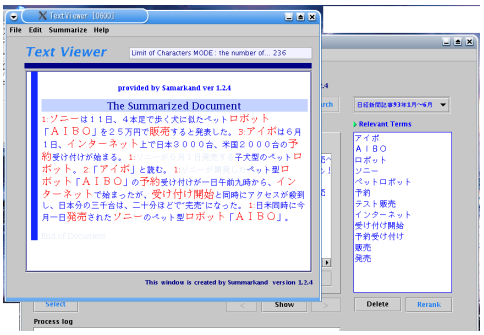

Fig. 1　ユーザ適合を考慮した複数文書自動要約システム

を導入した複数文書要約システムを提案する．具体的には，まず，要約対象となる文書集合から，検索要求の話題と関連した k 個のキーワードを自動的に抽出し，ユーザに提示する．ユーザは提示されたキーワードの中から自分の要約要求に適したキーワードを選択する．本複数文書要約システムは，ユーザが選択したキーワードを使用して要約を生成することで，ユーザの要約要求を反映した要約を生成する．

4　文章閲覧における読者の操作行動からの興味の推定 [3]

文章情報を対象とし,「興味の程度」を推定することを試みた．推定の手がかりとして，文章閲覧時のユーザの挙動データ、具体的には画面を送るためのスクロール操作の時間情報，を用い，ユーザの興味を推定する問題として設定した．従来，興味の程度と閲覧時間が相関することから，閲覧時間によって興味を推定していた（この特徴量を速度指標と呼ぶ）．しかし，文の難易度などによっても閲覧時間が影響されるため，速度指標のみでは推定精度と安定性の面で不十分と考え，それを補間する新たな指標として，文の言語情報から標準的な閲覧時間を推定し，それと実測閲覧時間の乖離の程度を指標化（乖離指標）した．その結果，興味の推定における正解率に関し，乖離指標単体で速度指標と同等の性能を示した．また，乖離指標と速度指標の相関が低いことも確認した．

5 LaTeX 原稿からのプレゼンテーションスライド自動生成 [4]

研究のプレゼンテーションでは，限られた時間の中で，聴衆に研究成果をよく理解してもらうためにプレゼンテーションスライドの作成が必要不可欠である．しかし，スライドの作成には多くの時間と手間を擁する．そのため，多くの研究者がスライド自動作成を望んでいる．本研究では，研究者の負担を軽減することを目的として，論文 LaTeX 原稿からスライドを自動生成する手法を提案する．本手法では，LaTeX ファイルの解析，スライドへの内容の割り当て，接続詞を利用した箇条書き生成を行う．LaTeX ファイルの解析では，スライド生成に必要な情報は残し，不要な情報の削除を行う．LaTeX ファイルの定型的な性質を利用すれば，必要な情報を特定することが可能である．スライド割り当てにおいては，論文中の名詞の出現頻度，エントロピー，idf 値に基づいて名詞の重要度を計算する．その重要度に基づいて，各セクションに対して，スライド枚数の割り当て，重要文の抽出を行う．接続詞を利用した箇条書き生成においては，並列関係表す接続詞を利用する．なぜなら，並列関係を表す接続詞を含む文には，その文と対になる文が存在することが多いからである．評価の結果，本手法は，論文に忠実なスライド生成に有効であることが分かった．

6 仕事文推敲支援に向けた連体修飾部不足に対する受容性判定法 [5]

文章推敲に関する従来研究では，主に，タイプミス，構文構造の複雑さ，表記の揺れを指摘する方法など，表記レベルと統語レベルの手法に重点が置かれていた．それに対して，本研究では，読みやすさを向上させるために，説明が不足していて論理展開が読みにくいと感じられる箇所を検出する技術を扱う．文章としては，情報を正確に伝達するための仕事文 (仕事用の文) を対象として，文単位での情報不足を推敲対象とする．この課題は意味処理に踏み込むため，これまで十分研究が行われてこなかった．なお，語用論の「協調の原理」によれば，量の格率と呼ばれる情報不足と情報過多に関する遵守すべき原則がある．このうち，情報過多を扱わない理由は，情報過多が，読者に，冗長な情報を無視するための負担を増やすだけであるのに対し，情報不足は理解困難という深刻な事態を招くことから，重要性が高いためである．

実験準備から解析に至る流れは以下のとおりである．まず，原文から連体修飾部を欠落させた課題文を生成し，次に被験者にその箇所に情報不足を感じるかどうかを判定させ正解判定データを作成した．その後，正解判定データの一部から機械学習を行い，残りのデータを機械判定させる．機械判定に用いる主な素性として，修飾部の欠落箇所におけるつながりの滑らかさに関係した語の連鎖に関する統計量を取り上げた．葯 1,000 箇所の判定課題に対し，SVM による機械学習アルゴリズムを用いた自動判定により正解率を測定した結果，機械判定の正解率として，ベースライン 50% ，上限 (人間の評価のバラツキから上限を定義)76%に対し，10-fold cross validation で 67% の正解率を得た．

7 まとめ

セマンティックアーカイブ活用のための基礎技術として，テキストマイニングと要約技術を取り上げ，最近得られた成果を報告した．セマンティックアーカイブ構築のための基礎技術として，書き言葉を対象とした言語処理のための音声ドキュメントの文境界推定に関しても予備的結果を得ている [6]．これについては，さらに研究を進めた上で，来年度報告書にて報告する予定である．

発表論文

[1] 酒井 浩之，梅村 祥之, 増山 繁, 交通事故事例に含まれる事故原因表現の新聞記事からの抽出. 自然言語処理, Vol.13, No.2, pp.99-123, 2006

[2] 酒井 浩之，増山 繁, ユーザの要約要求を反映するためにユーザとのインタラクションを導入した複数文書要約システム. 日本知能情報ファジィ学会誌, Vol.18, No.2, pp.265-279, 2006

[3] 梅村祥之, 増山 繁, 文章閲覧における読者の操作行動からの興味の推定. ヒューマンインターフェース学会誌, Vol.3, No.3, pp.435-444, 2006

[4] 宮本雅人, 酒井 浩之, 増山 繁, 論文 LaTeX 原稿からのプレゼンテーションスライド自動生成. 日本知能情報ファジィ学会誌, Vil.18, No.5, pp.752-760, 2006

[5] 梅村祥之, 増山 繁, 仕事文推敲支援に向けた連体修飾部不足に対する受容性判定法. 自然言語処理, 採録決定.

[6] 太田貴久, 酒井 浩之, 増山 繁, 書き言葉を対象とした言語処理のための音声ドキュメントの文境界推定. 第 1 回音声ドキュメント処理ワークショップ講演論文集, メディア科学リサーチセンター, pp.159-166, 2007