様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　　元 年　　６ 月 ２４ 日現在

機関番号：６２６１５
研究種目：挑戦的萌芽研究
研究期間：2016～2018
課題番号：１６Ｋ１２５４６

研究課題名（和文）文書の俯瞰的理解を支援する対話的な要約作成システム

研究課題名（英文）Interactive Summarization System to Support Literature Survey

研究代表者
相澤　彰子（Aizawa, Akiko）

国立情報学研究所・コンテンツ科学研究系・教授

研究者番号：９０２２２４４７

交付決定額（研究期間全体）：（直接経費）　　2,700,000円

研究成果の概要（和文）：学術文献の俯瞰は研究者や技術者にとって、時間を要する困難な作業である。対象となる文献が大量にある上に、多くの場合、必要な情報は抄録には書かれておらず、論文全体を通読して捜す必要がある。本研究では、大量文書の俯瞰的な理解を助けるレビューマトリックスの作成を支援するために、与えられたクエリに対して複数論文を自動要約する手法の開発に取り組むとともに、評価用のデータセットを構築して有効性を検証した。

研究成果の学術的意義や社会的意義
ユーザの理解支援の研究においては、自動評価が容易に行えるタスクの設計が研究の重要な手段となる。本研究では、レビューマトリックスの作成というタスクを設定することで効率的にデータセットが作成できることを示した。また、支援に必要となる要素技術として自動要約や機械読解の先端技術の研究に取り組み、当該タスクにおける有効性を示した。

研究成果の概要（英文）：Scientific literature survey is time-consuming and difficult for researchers and engineers who need to grasp the trends of specific research topics. In many cases, necessary information is not written in the abstracts, and users are enforced to read through the entire papers. One of the conventional practices is to generate a review matrix that organizes the information-in-need in a table format. This research aims at supporting the users' creation of review matrices. In our study, we developed a method for automatically summarizing multiple articles for a given query. We also constructed a dataset for evaluation and verified the effectiveness of the proposed method.

研究分野：テキスト・言語メディア

キーワード：情報組織化　レビューマトリックス　文書要約　質問応答システム　文検索

１．研究開始当初の背景

学術文献の俯瞰は研究者や技術者にとって、時間を要する困難な作業である。対象となる文献が大量にある上に、多くの場合、必要な情報は抄録には書かれておらず、論文全体を通読して捜す必要がある。ここで近年、レビューマトリクスと呼ばれる一覧表形式を用いて、文献ごとにポイントをまとめる情報整理法が注目されている。これは読み手自身が、目的に応じたテンプレートを使って、文書ごとの要約を作成することに相当する。電子図書館では従来、目的とする論文に素早く到達するための検索機能の強化に焦点があてられてきたが、文献俯瞰の生産性を高めるためには、単に検索を支援するだけではなく、文書の記述全体を解析して目的に応じて要点を提示する技術が必要である。これは文書集合全体から確信度の高い事実を選別して抽出するテキストマイニングより難しい問題設定であり、システム側ではユーザに対して複数の可能性を提示してユーザが適切な候補を選択する対話型のシステム設計が求められる。

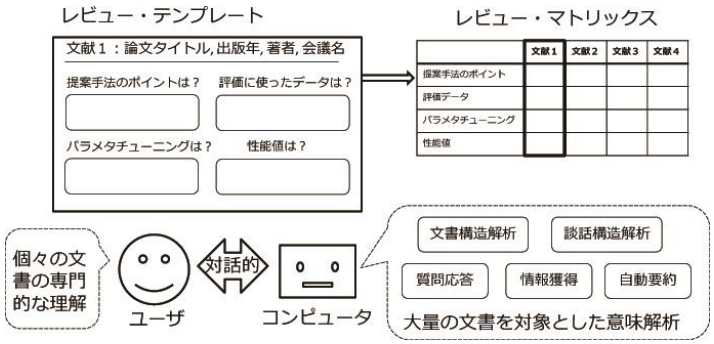

図 1　レビューマトリックスと関連要素技術

２．研究の目的

本研究の目的は、大量文書の俯瞰的な理解を助けるレビューマトリックスの作成支援である。その要素技術として、与えられたクエリに対して複数論文を自動要約する手法の検討に取り組み、以下の 3 つの研究テーマを設定した。

(1) 文章とその内容に関するクエリが与えられた場合に、文章中から回答を含む文を抽出する手法の検討。特に近年多くの自然言語処理タスクで優れた性能を上げている埋め込み表現を用いる手法を活用する。

(2) 情報の対比性を考慮した要約手法を検討する。複数文書の自動要約では文書間に共通する情報を重要情報として抽出することが通常行われるが、俯瞰作業においては文書間で対比するべき情報を抽出することから、新たな手法が必要となる。

(3) 対話的な情報獲得に向けた自然言語処理技術の可能性を検討する。

また、手法の有効性を評価するためには正解つきのデータセットが必要となる。このため本研究では、手法の検討を進めると同時に、学術論文を対象とした新たなデータセットの構築に取り組んだ。

３．研究の方法

本研究ではまず、レビューマトリックスの自動生成を「クエリ付き複数文書要約」として定式化した上で、自然言語処理分野において一般的な「共通タスク（shared task）のワークショップ」に注目し、タスクオーガナイザーの概要論文に掲載された参加システム一覧表をレビューマトリックスの正解データ、タスク参加者による投稿論文を要約対象文書とするデータセットを構築した。これに基づき 3 つの研究テーマに対して以下の手法の検討に取り組んだ。

(1) 文どうしの類似度計算に基づく重要文選択において、埋め込み表現の効果を調べるとともに、文類似度の計算を改善する手法を検討した。

(2) 要約として最適の文集合を得るための目的関数に、文書間で得られる要約の対比性を考慮する項を導入する手法を検討した。

(3) 文章と質問が与えられた際に文章中から適切な回答を選択する機械読解、および、文脈を考慮した対話文生成に関する手法の検討に取り組んだ。

４．研究成果

データセットの構築では、ACL Anthology から抽出した共通タスクの概要論文について、参加システム一覧表の見出し項目を調査し、Yes-No タイプや列挙タイプなどの質問タイプ別に分類の上、計算機可読な形式に変換した（発表文献②③）。また、比較表で参照されている参加システム報告論文を入手して全文テキストを抽出して、セクションや段落構成を考慮しながら文単位に分割した。さらに、共通タスクのワークショップ後に発表された新たな論文に対しても評価が行えるよう、人手で正解文を選択した拡張データセットを構築した（発表文献①）

(1) レビューマトリックスの項目の中で、特に自由記述による説明に焦点をあて検討に取り組んだ。具体的には、項目をクエリとみたてて対象文書中に含まれる文を重要度順にランキングした上で、さらに整数計画法を用いて必要な情報を含む文集合を抽出する。この重要文選択において、類語辞書によるクエリ拡張、および埋め込み表現に基づく類似度計算を適用する手法を提案し、構築したデータセットを用いて効果を検証した（発表文献②③）。また、埋め込み表現にグラフに基づく重要語指標を取り入れる手法を新たに提案し、構築したデータセットを用いて有効性を示した（発表文献①）。埋め込み表現を学習する際に、既存の語彙資源に登録された情報を活用する手法の有効性を検討し報告した（発表文献⑥）

(2) 目的関数に要約の対比性を考慮する項を導入する効果を検証して、得られた要約に対する影響を分析して国際ワークショップで発表した（発表文献②③）。

(3) 機械読解について、システムの性能プロファイルを作成するための手法を提案して複数のデータセットに対する詳細な分析結果を報告した（発表文献④）。また、文脈を考慮した対話文生成に隠れ変数を用いる深層学習モデルを適用し、有効性を報告した（発表文献⑤）。

５．主な発表論文等

〔学会発表〕（計 6 件）

① Kazutoshi Shinoda and Akiko Aizawa: “Query-focused Scientific Paper Summarization with Localized Sentence Representation.” The 3rd Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL 2018), workshop at ACM SIGIR 2018. ［査読有］（20180712）

② Hayato Hashimoto, Kazutoshi Shinoda, Hikaru Yokono, and Akiko Aizawa: “Automatic Generation of Review Matrices as Multi-document Summarization of Scientific Papers.” Second International Workshop on SCIentific DOCument Analysis (SCIDOCA 2017), associated with JSAI International Symposia on AI 2017 (IsAI-2017)［査読有］（20171115）

③ Hayato Hashimoto, Kazutoshi Shinoda, Hikaru Yokono, and Akiko Aizawa: “Automatic Generation of Review Matrices as Multi-document Summarization of Scientic Papers.” BIRNDL’17: Bibliometric-enhanced IR and NLP for Digital Libraries, workshop at SIGIR 2017［査読有］（20170811）

④ Saku Sugawara, Yusuke Kido, Hikaru Yokono, and Akiko Aizawa: “Evaluation Metrics for Machine Reading Comprehension: Prerequisite Skills and Readability. The 55th annual meeting of the Association for Computational Linguistics (ACL 2017). Vancouver,

Canada［査読有］（20170730-20170804）

⑤ Xiaoyu Shen, Hui Su, Yanran Li, Wenjie Li, Shuzi Niu, Yang Zhao, Akiko Aizawa, and Guoping Long: “A Conditional Variational Framework for Dialog Generation.” The 55th annual meeting of the Association for Computational Linguistics (ACL 2017) Vancouver, Canada［査読有］（20170730-20170804）

⑥ Thomas Perianin, Hajime Senuma, and Akiko Aizawa: “Exploiting Synonymy and Hypernymy to Learn Efficient Meaning Representations.” The 18th International Conference on Asia-Pacific Digital Libraries (ICADL 2016). Tsukuba, Japan. ［査読有］（20161207-20161209）

６．研究組織

(1) 研究分担者

研究分担者氏名：徳永　健伸
ローマ字氏名 TOKUNAGA, Takenobu
所属研究機関名：東京工業大学
部局名：情報理工学院
職名：教授
研究者番号（8 桁）：20197875

(2) 研究協力者

研究協力者氏名：
ローマ字氏名：