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 １４ 日現在

機関番号：１５１０１
研究種目：基盤研究(C)（一般）
研究期間：2014～2018
課題番号：２６３３０２５２

研究課題名（和文）記述必要項目特定技術を利用した内容欠落文書の改善

研究課題名（英文）Improvement of content missing documents using description required item identification technology

研究代表者
村田　真樹（MURATA, Masaki）

鳥取大学・工学研究科・教授

研究者番号：５０３５８８８４

交付決定額（研究期間全体）：（直接経費）　　3,700,000円

研究成果の概要（和文）：ルールベースの技術により論文データでは「比較」「問題点」「目的」「例」を記載必要項目として0.6から0.8のF値(性能)で記載の欠如を検出できた。技術を汎用化した。修正したい文書の類似文書を入力して、その文書での高頻度出現事項を重要事項として、その重要事項の記載の有無を発見する技術を構築した。単語レベルの情報抽出に基づき、8割程度のF値で欠落を検出できた。記載が欠落している個所に関わる情報をウェブから取得してユーザに提示する技術を構築した。文に基づいて情報抽出する技術を構築できた。新聞や製品情報の文書で、文レベルで記載の欠落を発見する実験を遂行できた。5割から8割程度のF値で欠落を発見できた。

研究成果の学術的意義や社会的意義
従来の推敲システムであまり扱われていなかった内容欠落文の修正の問題を扱ったため、推敲システムの扱える範囲が増え、推敲システムの発展に寄与する。さらに社会的に広く見れば、本技術が会話・通信・会議にも応用されることにより、種々のコミュニケーションにおける内容欠落文の減少による情報伝達率の上昇により、人間活動の大幅な効率化につながると期待される。

研究成果の概要（英文）：Rule-based techniques using items "comparison", "problem", "purpose" and "example" in article data could detect the absence of description with an F measure of 0.6 to 0.8. The technology was generalized. We made the technology to discover the presence or absence of the description of the important matters by using the frequent occurrence items in the documents similar to the documents to be modified as the important matters. Based on word level information extraction, we could detect absence of the description at F measure (performance) of about 0.8. We built a technology to obtain information related to the missing part from the web and present it to the user. We could build a technology to extract information based on sentences. In the newspaper and product information documents, we were able to carry out an experiment to find absence of the description at the sentence level. We found absence of the description in the F measure of about 0.5 to 0.8.

研究分野：自然言語処理

キーワード：内容欠落文書の改善　文書推敲　記載必要項目　情報抽出　文生成　機械学習

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９、ＣＫ－１９（共通）

１．研究開始当初の背景

文章の改善としては以下のことが考えられる。

問題１．誤字の修正・適切な語の選択・冗長な表現の修正支援

問題２．語順や文の順番の修正・語と語の係り受けの誤り修正

問題３．内容欠落文（書くべき内容が書かれていない文）の改善支援（本課題で扱う問題）

上記のうち、問題１と問題２は既に先行研究が多数ある。一方、問題３を自動処理で扱う研究はほとんどないため、本課題で扱う。

提案者は、誤った日本語文を抽出する技術、適切な英語表現に変換する文パターンを抽出する技術、冗長な表現を検出し修正を支援する技術、語順や文の順序を推定する技術、係り受けの複雑さを計量する技術を構築し、問題１と問題２を解決した。

これに対して、問題３は上記の技術に加えて、文章の種類ごとに存在する書くべき項目を明らかにする記載必要項目特定技術、その書くべき項目が文章中に記述されているかを確かめる記載確認技術、執筆者が書き逃していた書くべき項目を好適に執筆できるように支援する記載支援技術が必要と想定している。書くべき項目の列挙には、文章群からその文章群において書くべき重要となる情報を抜き出す情報抽出の技術が必要である。書くべき内容が文章中にあるかの判定には、機械学習の技術が必要である。記載支援技術には、文章の言い換え技術と可視化技術が必要である。

２．研究の目的

ある文書群において書くべき記載必要項目を情報抽出技術等を利用して特定し、その記載必要項目が記載されていない内容の欠落した文章を改善するのに必要となる技術を明らかにする。文章の種類ごとに記載必要項目が決まっている。新聞であれば誰がいつ何をしたかなどの 5W1H の情報などが、論文であれば研究対象・研究成果・有効性などが、記載必要項目となる。ウィキペディアであれば記載するページの内容ごとに記載必要項目が決まる。法則のページならば発見年や発見者やどの概念から派生して生まれたものか等が記載必要項目となる。記載必要項目が記載されていない場合、内容が不明瞭となり可読性が低下する。

３．研究の方法

本課題では、以下の 3 つのステップにより内容欠落文の改善を支援する技術を構築する。

ステップ 1. 記載必要項目を明らかにする

ステップ 2. 記載必要項目が記載されているかを確認する

ステップ 3. 書き逃していた記載必要項目の執筆を支援する

本課題では、新聞、ウィキペディア、QA サイトの文章、論文、学生の文章を実験のデータとして用いる。文章の種類ごとに、その文章群において出現頻度の高い記載項目を記載必要項目として抜き出す（記載必要項目特定技術）。実験データの各文章において記載必要項目が記載されているかを人手で確認し、記載必要項目が記載されていない場合は記載するように人手で修正する。このとき、記載内容自体は執筆者でないとわからない場合は、実際とは異なるかもしれないが、想定可能な文章を記載してとにかく埋める。上記作業により、記載必要項目が記載されているか否かを記した文章を収録した記載必要項目有無データベースと、記載必要項目を人手で書き足して修正した記載必要項目修正データベースをあわせて１万文を超える規模で構築する。記載必要項目有無データベースを機械学習することで、記載必要項目が記載されてい

ない内容の欠落した文章を検出できる。機械学習は内容欠落文の検出のみならず、検出の際の学習により内容欠落文の言語的特徴も出力できる。これは、内容欠落文の改善のための有益な知見となる。記載必要項目修正データベースに言い換え技術を利用することで、内容欠落文の改善のための文パターンや規則を獲得できる。これらのパターンや規則も内容欠落文の改善に役立つ知見・支援技術の一部である。

４．研究成果

(1) 論文データにおける記載必要項目を検討した。自然言語処理分野の論文データにおいて頻出する単語をもとに記載必要項目を決定した。「必要性」「新規性」「比較」「問題点」「目的」「例」を記載必要項目とした。またこれらの記載必要項目の有効性を確かめるために、記載必要項目の記載の欠如を検出する研究も行った。「比較」「問題点」「目的」「例」では、0.6 から 0.8 の F 値で記載の欠如を検出できた。

(2)機械学習に基づく記載必要項目の記載の欠如を検出する研究を行った。昨年度に行ったルールベースに基づく記載必要項目の記載の欠如を検出する手法よりも性能が低かった。ルールベース手法の有効性を確認できた。また、記載必要項目の記載が不十分な文書の人手による分析を行った。記載が不十分な文書を、その不十分さに基づき 5 レベルに分けて分析した。

(3)言い換えと可視化の技術を用いて内容欠落文の記載を改善する記載支援技術の研究を行った。記載必要項目修正データベースに言い換え技術を利用することで、内容欠落文の改善のための文パターンや規則を獲得した。具体的には、論文における問題点と目的の記載が不十分な文章について修正するのに役立つ文パターンを獲得できた。また、内容欠落箇所に関わる情報をウェブなどから検索して執筆者に情報を提示し、執筆者が内容を考える参考とできる技術の開発を行った。具体的には、情報検索や質問応答技術を利用してこの研究を推進した。あるウィキペディアの記事において記載が欠落している個所に関わる情報をウェブから取得してユーザに提示する技術を構築した。

(4)技術の汎用化を行った。具体的には、修正したい対象の分野がどのようなものであっても処理できるようにした。修正したい文書の類似文書を入力として、その文書での高頻度出現事項を重要事項として、その重要事項の記載の有無を発見する技術を構築した。従来行っていた、ウィキペディアや論文での実験以外に、新聞や小説の文書でも、記載の欠落を発見する実験を遂行できた。単語レベルの情報抽出に基づき、8 割程度の F 値で欠落を検出できた。

(5)機械学習に基づく言語解析・情報抽出・文章作成支援など、文章の修正と情報欠落文書の分析に資する文章処理技術の検討も行った。例えば、文章修正技術、言い換え技術、意味解析、情報抽出に関わる研究を行った。情報抽出の研究では、従来単語に基づいて情報抽出をしていたが、新たに文に基づいて情報抽出する技術も新たに構築できた。実験を行ったところ、7,8 割の F 値で情報抽出できた。

(6)単語レベルで行っていたものを文レベルに拡張した。従来では、単語レベルで情報を抽出し、情報の洩れを指摘していたが、本年度は、文レベルで情報を抽出し、情報の洩れを指摘する研究を行った。新聞や製品情報の文書で、記載の欠落を発見する実験を遂行できた。5 割から 8

割程度の F 値で欠落を発見できた。

５．主な発表論文等
〔雑誌論文〕（計 7 件）

① Masaki Murata, Yuki Abe, Using Machine Learning for Automatic Estimation of Emphases in Japanese Documents, IEICE Transactions on Information and Systems, 査読あり, Vol.E100-D, No.10, 2017, pp.2669-2772
DOI:10.1587/transinf.2016EDL8247

② Masaki Murata, Shunsuke Tsudo, Masato Tokuhisa, Qing Ma, Correcting Redundant Japanese Sentences Using Patterns and Machine Learning for the Development of Writing Support Systems, International Journal of Computational Linguistics and Applications, 査読あり, Vol. 7 No. 2, 2016, pp.183-199

③ Masaki Murata, Satoshi Ito, Masato Tokuhisa, Qing Ma, Order Estimation of Japanee Paragraphs by Supervised Machine Learning and Various Textual Features, Journal of Artificial Intelligence and Soft Computing Research, 査読あり, Vol. 5, No. 4, 2015, pp.247-255
DOI:10.1515/jaiscr-2015-0033

〔学会発表〕（計 12 件）

① 岡崎健介，村田真樹，馬青，複数文書からの文レベルの情報の書き漏らしの検出，言語処理学会第 25 回年次大会，2019

② Masaki Murata, Naoya Nonami and Qing Ma, Using Information Extraction and Search Engines for Automatic Detection of Inadequate Descriptions and Information Supplements in Japanese Wikipedia, 19th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing 2018), 2018

③ 岡崎健介，村田真樹，馬青，複数文書からの重要情報の抽出と表の生成，言語処理学会第 24 回年次大会，2018

④ Hokuto Akano, Masaki Murata, Qing Ma, Detection of Inadequate Descriptions in Wikipedia using Information Extraction based on Word Clustering, Joint 17th World Congress ofInternational Fuzzy Systems Association and 9th International Conference on Soft Computing and Intelligent Systems (IFSA-SCIS 2017), 2017

⑤ Takuma Okada, Masaki Murata and Qing Ma, Automatic Detection and Manual Analysis of Inadequate Descriptions in a Thesis, 2016 Joint 8th International Conference on Soft Computing and Intelligent Systems and 2016 17th International Symposium on Advanced Intelligent Systems, 2016.