様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成 ２７ 年　 ５ 月 ２８ 日現在

機関番号：１５１０１
研究種目：基盤研究(C)
研究期間：2011～2014
課題番号：２３５００１７８

研究課題名（和文）冗長な文の改善に役立つ言語的特徴の機械的発見と作文支援

研究課題名（英文）Automatic acquisition of linguistic characteristics and writing support useful for modification of redundant sentences

研究代表者
村田　真樹（MURATA, Masaki）

鳥取大学・工学（系）研究科（研究院）・教授

研究者番号：５０３５８８８４

交付決定額（研究期間全体）：（直接経費）　　4,000,000円

研究成果の概要（和文）：本研究では、冗長な文章の改善に役立つ技術を構築した。本課題は、日本語文章を対象とした。本課題はコミュニケーションの効率化に寄与する。手法としては機械学習を利用し、冗長な文の検出は約７、８割、冗長な文の修正は、「可能」「という」「すること」が原因となって冗長となった文において６割以上の正解率であった。文章レベルの冗長な表現を収集し、収集したデータを人手で分析し、文章レベルの冗長な表現における種々の特徴を明らかにした。

研究成果の概要（英文）：In this study, we constructed systems useful for improving redundant sentences. We handled Japanese sentences. Our study contributes to efficiency of communication. We obtained accuracies of about 0.7 to 0.8 in detection of redundant sentences using machine learning. We obtained accuracies of more than 0.6 in correction of redundant sentences including the Japanese words, "kanou" (possible), "toiu" (that is) and "surukoto" (perform), using machine learning. By gathering and examining redundant expressions in long texts including long sentences and long paragraphs, we clarified various kinds of characteristics in redundant expressions of long texts.

研究分野：自然言語処理

キーワード：文生成　冗長な表現　文の修正　機械学習

１．研究開始当初の背景
(1) 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
問題２．語順の修正・語と語の係り受けの誤りおよび複雑性の修正
問題３．冗長な表現の改善（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。これに対して問題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。

(2)問題３は、新たに、冗長な個所の候補を取り出すために文章中の類似箇所を発見する技術、抽出した冗長な個所の候補が真に冗長かを判定する技術、冗長な個所を簡素な表現への言い換える技術が必要と想定している。類似箇所の発見には情報検索の技術が必要であり、真に冗長かを判定するためには、機械学習の技術が必要であり、簡明な表現への言い換えには言い換えの技術が必要である。提案者は、言い換え、情報検索、機械学習の研究分野で多くの業績をあげており、本課題を遂行するのに適している。

２．研究の目的
機械的な手法により冗長な文章の改善に役立つ言語的特徴を明らかにするとともに、その知見により冗長な文章の改善に役立つ支援技術を構築する。例えば、「まず初めに円高の解決に向けた解決策の検討を考えたい。」の文のように、文内で同じ単語や同義語が複数回出現する文は冗長でわかりにくい。この文は冗長な表現を削除することで「まず円高の解決策を検討したい。」と修正可能である。また、同じ内容の段落を複数含む文章も冗長である。本課題では、文レベルの問題から長い文章レベルの問題まで含めて、冗長な文章の改善に役立つ支援技術を構築する。本課題は、日本語文章を対象として研究を行う。本課題はコミュニケーションの効率化に寄与する。

３．研究の方法
(1) 一文内の冗長な表現の検出に関する支援技術を研究する。まず、大雑把に冗長な文を収集するために、文章の冗長度を定義する。簡単には、ある一定の語数における語の異なり数の逆数を、文章の冗長度（冗長性の度合い）と定義する。上記で定義した冗長度を計算し、冗長度の高いデータを収集する。収集したデータを人手で分析して、冗長性判定用データベースと冗長性修正文集合データベースを作成する。冗長性判定用データベースには、冗長な文と冗長でない文が格納される。冗長性修正文集合データベースには冗長な文とそれを修正した文の対が格納される。次に、一文内の冗長な表現の検出に関する研究を行う。冗長性判定用データベースを教師データとして利用して、教師あり機械学習により、与えられた文が真に冗長であるか否かを自動判定する冗長性判定技術を構築する。

(2) 一文内の冗長な表現の修正支援の研究を行う。構築した冗長性修正文集合データベース中の冗長な文とその修正文の対を照合し、修正用の文パターンや語順変更の規則を取得する。これら文パターンや規則を利用して、冗長な文の修正案を提示する簡素な表現への言い換え技術を構築する。また、冗長性修正文集合データベースを学習データとした教師あり機械学習による修正も試みる。

(3)構築した文レベルの冗長性の検出・修正の研究の方法を段落・文書レベルのものに拡張し、その手法を利用して、段落・文書レベルの冗長性判定用データベースと冗長性修正文集合データベースを作成するとともに、一段落内及び一文書内における冗長性の改善支援の研究をする。段落、文書レベルになると自動修正はかなり困難と思われる。冗長性の検出に重点をおきながら可能な範囲で研究を進める。

４．研究成果
(1)一文内の冗長な表現の検出に関する支援技術を研究した。まず、大雑把に冗長な文を収集するために、文章の冗長度を定義した。簡単には、ある一定の語数における語の異なり数の逆数を、文章の冗長度（冗長性の度合い）と定義した。上記で定義した冗長度を計算し、冗長度の高いデータを収集した。収集したデータを人手で分析して、実際に冗長な文、冗長でない文を収集した。これを通じて、以下に示す、冗長性判定用データベースと冗長性修正文集合データベースを作成した。冗長性判定用データベースには、冗長な文と冗長でない文が格納される。冗長性修正文集合データベースには冗長な文とそれを修正した文の対が格納される。

(2)一文内の冗長な表現の検出に関する研究を行った。冗長性判定用データベースを教師データとして利用して、教師あり機械学習により、与えられた文が真に冗長であるか否かを自動判定する冗長性判定技術を構築した。その技術の性能評価を行った。種々の冗長な表現をまとめて教師あり機械学習で学習した場合は、冗長な表現の検出はF値0.5程度で行うことができた。冗長な表現を形成する理由となる表現ごとに機械学習で学習した場合は、冗長な表現の検出はF値0.7～0.8程度で行う

ことができた。高い性能で冗長な表現を検出できるという重要な意義のある事実を確認できた。機械学習にはSVMを用い、素性には文中に出現する単語や文字列を用いた。

(3)表現の順序、文体、表現など、冗長な文章の改善に資する文章処理技術の検討も行った。例えば、表現の変化の検出と考察を行った。また、文章の改善のため、文章における文の順序を推定する研究を行った。二つの文のうち、どちらの文を先に記述すべきかの推定において、教師あり機械学習を利用した手法を提案し7割から8割の性能を得た。これについても、高い性能で文の順序を推定できるという重要な意義のある事実を確認できた。

(4)一文内の冗長な表現の修正に関する支援技術を研究した。冗長な文を修正する方法として、パターンを用いた手法と機械学習を用いた手法を提案した。「可能」「という」「すること」が原因となって冗長となった文の修正の実験を行った。パターンを用いる手法と機械学習を用いる手法のいずれかが、最も頻度の高いものを出力とするベースライン手法よりも同等以上の正解率であった。パターンを用いる手法と機械学習を用いる手法のいずれかで 6 割以上の正解率で冗長な文を修正できた。修正後の表現のみの推定(修正前の表現の範囲を特定できなくてよい) では、パターンを用いる手法と機械学習を用いる手法のいずれかで 7 割以上の正解率を得た。以上により、「可能」「という」「すること」については、パターンを用いる手法と機械学習を用いる手法がある程度冗長な表現の修正に役立つことがわかった。

(5)実際の文書の推敲での冗長な文の修正ではもっと確実な手法を用いる必要がある場合も考えられる。そのため修正をするのではなく、修正箇所の検出を自動で行い、さらに検出した冗長箇所の修正候補を頻度の高い順に並べ、ユーザーに提示するという方式を検討した。この方式では、冗長な個所とその修正候補が提示されるため、文書作成者の修正作業の負担が軽減されると思われる。この方式で必要となるデータの構築も試みた。
表現の順序、文体、表現など、冗長な文章の改善に資する文章処理技術の検討も行った。例えば、表現の変化の検出と考察を行った。

(6)文章の改善のため、文章における段落の順序を推定する研究を行った。二つの段落のうち、どちらの段落を先に記述すべきかの推定において、教師あり機械学習を利用した手法を提案し 6 割から 8 割の性能を得た。先に行った文の順序推定と、この段落の順序推定の比較も行い、順序推定において重要な事柄を分析できた。

(7)文章レベルで冗長な文章とそれを修正したデータベースを作成した。そのデータベースのデータを分析し、典型的な 3 種類の分類を明らかにした。その３種類は、分類１「文単位の修正で十分なもの」、分類２「補足または説明をする文を先頭文にまとめる形で短く簡潔な文章に修正されるもの」、分類３「長い文を箇条書きにまとめる形で修正するもの」である。機械学習を用いる手法と、冗長度を用いる手法により冗長な文章を検出した。機械学習を用いた実験では機械学習の素性として「冗長度」を利用した際の正解率が最も高かった。機械学習を用いた手法の正解率(0.66) が、冗長度を用いる手法の正解率(0.65) と同程度の正解率であった。

(8)文章レベルの冗長な表現を冗長度などを用いて自動で収集し、収集したデータを人手で分析した。文章レベルの冗長な表現における種々の特徴を分析できた。表現の順序、文体、表現など、冗長な文章の改善に資する文章処理技術の検討も行った。文章レベルでの冗長な表現の分析では、文章レベルの冗長な表現を修正する際には、大きくは、「文章の分割」「内容のある表現の削除」「文章の順序変更」「同じ表現の繰り返しの削除」「装飾・表現の変更」により修正をするとよい場合があることがわかった。またそれぞれの修正の仕方の下位の項目としては、「文章の分割」には「1 文を複数の文に分割する」「箇条書にする」「1 段落を複数の段落に分割」という修正方法があることがわかった。また、「内容のある表現の削除」には、「補足表現の削除」「修飾語の削除」「接続詞の削除」「条件節の削除」があることがわかった。また、「文章の順序変更」には、「結論を前に補足を後ろに」「主語と述語を近づける」「理由を後ろに」「理由を前に」「主題を前方に」「順位の順に」「時系列の順に」という修正方法があることがわかった。また、「装飾・表現の変更」には、「引用符・括弧の利用」「読点の挿入，場所の変更」「読点の削除」「平仮名を漢字に」「漢字を平仮名に」「ある表現を片仮名に」「複数の文を１ 文にまとめる」という修正方法があることがわかった。

(9)冗長な表現に関わる先行研究はほとんどなく冗長な表現に関わる部分は本研究は新規である。文、段落の順序推定については先行研究がいくつかあるが、本研究には豊富な言語表現も利用して教師あり機械学習で順序推定を行っているという特徴がある。

５．主な発表論文等

〔雑誌論文〕（計４件）

① Yuya Hayashi, Masaki Murata, Liangliang Fan, Masato Tokuhisa, Japanese Sentence Order Estimation using Supervised Machine Learning with

Rich Linguistic Clues, International Journal of Computational Linguistics and Applications, 査読有, Vol.4, No.2, pp.153-167, 2013.
http://www.gelbukh.com/ijcla/2013-2/IJCLA-2013-2-pp-153-167-Japanese.pdf

② Masaki Murata, Masahiro Kojima, Takuya Minamiguchi and Yasuhiko Watanabe, Automatic Selection and Analysis of Japanese Notational Variants on the Basis of Machine Learning, International Journal of Innovative Computing, Information and Control, 査読有, Vol.9, No.10, pp.4231-4246, 2013.
http://www.ijicic.org/ijicic-12-10041.pdf

③ Masaki Murata and Masao Utiyama, Compound Word Segmentation Using Dictionary Definitions -- Extracting and Examining of Word Constituent Information --, ICIC Express Letters Part B: Applications, 査読有, Vol.3, No.3, pp.667-672, 2012.
http://www.ijicic.org/elb-3(3).htm

〔学会発表〕（計１５件）

① 村田真樹，徳久雅人，馬青，冗長な文章の人手による分析，言語処理学会第 21 回年次大会，P4-24，pp.984-987，2015 年 3 月 19 日，京都大学（京都府京都市）．

② Satoshi Ito, Masaki Murata, Masato Tokuhisa, Qing Ma, Order Estimation of Japanese Paragraphs by Supervised Machine Learning, Proceedings of Joint 7th International Conference on Soft Computing and Intelligent Systems and 15th International Symposium on Advanced Intelligent Systems (SCIS & ISIS 2014), pp.1096-1101, 2014 年 12 月 5 日，北九州国際会議場（福岡県北九州市）．

③ 都藤俊輔，村田真樹，徳久雅人，馬青，機械学習と冗長度を用いた冗長な文章の検出，言語処理学会第 20 回年次大会，P8-17, pp.939-942, 2014年3月20日，北海道大学（北海道札幌市）．

④ 都藤俊輔，村田真樹，徳久雅人，馬青，パターンと機械学習による冗長な文の修正と修正のヒント出力，言語処理学会第 19 回年次大会，P3-4，pp.588-591，2013 年 3 月 15 日，名古屋大学（愛知県名古屋市）．

⑤ Shunsuke Tsudo, Masaki Murata, Masato Tokuhisa, and Qing Ma, Machine Learning for Analysis and Detection of Redundant Sentences Toward Development of Writing Support Systems, The 6th International Conference on Soft Computing and Intelligent Systems, and The 13th International Symposium on Advanced Intelligent Systems (SCIS-ISIS 2012), p.2225-2228, 2012 年 11 月 22 日，神戸コンベンションセンター（兵庫県神戸市）．

〔図書〕（計０件）

〔産業財産権〕
○出願状況（計０件）

○取得状況（計０件）

〔その他〕
特になし。

６．研究組織
(1)研究代表者
村田　真樹（MURATA, Masaki）
鳥取大学・大学院工学研究科・教授
研究者番号：５０３５８８８４