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　２ 年　 ６ 月 １５ 日現在

機関番号：１２６０８
研究種目：若手研究(B)
研究期間：2017～2019
課題番号：１７Ｋ１２７３８

研究課題名（和文）要約技術と簡約技術を融合したテキストの個人化

研究課題名（英文）Text Personalization with Automatic Summarization and Text Simplification

研究代表者
西川　仁（NISHIKAWA, HITOSHI）

東京工業大学・情報理工学院・助教

研究者番号：００７６５０２６

交付決定額（研究期間全体）：（直接経費）　　3,100,000円

研究成果の概要（和文）：簡約コーパスの構築が完了し，また入手した大規模な新聞記事自動要約データとあわせて要約および簡約を実行するモデルを構築した．具体的には，新聞記事に対し要約及び平易化の両方を行った．その過程を記事全体に対する処理と各文に対する処理に分けてモデルを実装し，それぞれのモデルの学習には，構築した簡約コーパスを用いた．生成された記事に対する定量評価の結果，記事処理及び文処理モデルの両方を適用することで，元記事や一方のモデルのみを適用する場合に比べ生成されたテキストの品質を測定する自動評価尺度であるBLEU，ROUGE，SARIの値が向上し，より正解テキストに近い，短く平易な記事を生成できることが示された．

研究成果の学術的意義や社会的意義
情報化社会の進展に伴い，自動要約および平易化といった，テキストの読解を支援する技術への需要が高まっている．長いテキストから重要箇所を抽出し短くまとめる「要約」は読み手の迅速な内容把握を可能にし，大量の文献の読解や調査などを必要とする知識労働者の生産性を大幅に向上せしめることが期待される．また，専門用語などの難解な表現に対し削除及び易しい表現への置換を行う「平易化」は外国人や子供など語彙知識が不足している読み手の読解を補助する．これらを組み合わせ読み手に合わせてテキストを柔軟に変化させることによって電子化されたテキストを読解する幅広い層に対して読解支援を行うことが可能となる．

研究成果の概要（英文）：We built a text-simplification corpus, and developed a model summarizing and simplifying texts with a large-scale automatic summarization corpus. We summarized and simplified newswire articles with that model. We implemented two models: a model which processed one article sentence-by-sentence, and a model which processed one whole article all at once. Those models were learned with the above corpora. Our quantitative experiments showed that a model that combined the above two models could generate better outputs than the single model did in terms of automatic quantitative evaluation methods such as BLEU, ROUGE, and SARI.

研究分野：知能情報学

キーワード：自動要約　文簡約

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

テキストの主たる流通経路はインターネットに移りつつある．現在では多くの人々がインターネットを経由して，生活に必要な様々な情報を入手している．一方，インターネット上のテキストは必ずしも広く一般の読み手に向けて書かれたものばかりとは言えない．インターネット上のテキストには，知識を持たない読み手には内容を理解することが難しい専門的な内容を含むものや，また，生活に必要であるのにもかかわらず，使われている語彙の難しさなどが問題となって読み手のスムーズな読解を妨げるものが存在する．

本研究の目標は，この問題の解決である．すなわち，読み手が読解しやすいように，テキストを読み手に合わせて個人化するシステムを構築する基盤を確立することが本研究の目標である．

個人化システムは，語彙力や分野知識などからなる読み手の属性と，テキストをどの程度の長さで，どのような点に焦点をあてて読みたいのかを指示する読み手の情報要求，そして読解の対象となるテキストを受け取る．その上で，読み手の情報要求に合わせてテキストを要約するテキスト要約手法と，語彙力に合わせて難しい語彙を言い換えるテキスト簡約手法を用いてテキストを読解しやすく書き換える．

本研究に存在する主な課題は，(1) 個別の要素技術を調和させることと，(2) ユーザ情報を要約および簡約に適切に利用することである．本研究ではこれらの課題を解決し，テキストを個人化する技術の基盤を確立する．

２．研究の目的

項目(1) 読み手の情報要求に合わせた要約の生成読み手が，どの程度の長さのテキストで，どのような情報について読みたいのかをシステムが把握した上で，それに基づいて適切な要約を出力する手法を確立する．正しい要約を人手で用意し，これを再現できる要約手法を確立する．

項目(2) 語の言い換えのための言語資源の構築と簡約の生成テキストに含まれる難しい語を言い換えるには，難しい語の表記とその語の簡約された表記の組を辞書（簡約辞書）として保持する必要がある．そのため，簡約辞書を構築するとともに，それを用いてテキストを簡約する手法を確立する．特に，語が，元のテキストの文脈とそぐわないものに置き換えられてしまうとテキストが不自然なものとなるため，そのような誤った置き換えが生じないように工夫する．

項目(3) 要約技術と簡約技術の出力の最適化要約技術が出力する要約と，簡約技術が出力する簡約とを，適切に組み合わせて読み手にとって最良の要約を生成する方法を確立する．具体的には，入力テキストに含まれる難しい語の言い換え候補を複数生成し，それを要約に組み入れた上で，要約の内容性（要約が入力テキストの重要な情報を含んでいるか）と可読性（要約がテキストとして自然なものとなっているか）の最適化を行う．

項目(4) データとシステムの公開本研究で構築するデータは公開し，後に続く研究を奨励する．同様にシステムも無料で公開し，本技術を広く利用できるようにする．

３．研究の方法

簡約コーパスを構築し，また大規模な新聞記事自動要約データを入手し，これらを組み合わせ要約および簡約を実行するモデルを構築した．具体的には，新聞記事に対し要約及び平易化の両方を行った．その過程を記事全体に対する処理と各文に対する処理に分けてモデルを実装し，それぞれのモデルの学習には，構築した簡約コーパスを用いた．

４．研究成果

生成された記事に対する定量評価の結果，記事処理及び文処理モデルの両方を適用することで，元記事や一方のモデルのみを適用する場合に比べ，生成されたテキストの品質を測定する自動評価尺度であるBLEU，ROUGE，SARIの値が向上し，より正解テキストに近い，短く平易な記事を生成できることが示された．また，実際の生成例から，文処理モデルの学習の際に異なるコーパスを組み合わせることで，生成時の文法的な誤りを抑制できる可能性や，文処理モデルを先に適用することで，記事処理モデルを適用する際により多くの重要文を目標要約長内に収められる可能性が示唆された．

当初計画には含めていなかったが，本研究課題の遂行にあたって自動要約におけるゼロ代名詞の問題があらためて浮上してきたため，ゼロ代名詞照応解析の研究を実施した．これは大規模日本語均衡コーパス BCCWJ を訓練事例と用い，深層学習を利用してゼロ代名詞照応解析を行うもので，これについては十分な成果が得られたものと考えている．

今後の課題としては，記事処理及び文処理モデルの学習の際に，単語の汎化や簡約コーパスに含まれる誤った対応づけの除去など改善の余地があり，これらに対する有効な処理方法について検証する必要がある．また，より質の高い簡約コーパスを構築するために，文単位でなく文節単位で自動対応づけを行う手法も検討する必要がある．

５．主な発表論文等

〔雑誌論文〕　計2件（うち査読付論文　2件／うち国際共著　0件／うちオープンアクセス　2件）

１．著者名 山城颯太，西川仁，徳永健伸	４．巻 26(2)
２．論文標題 大規模格フレームによる解候補削減を用いたニューラルネットゼロ照応解析	５．発行年 2019年
３．雑誌名 自然言語処理	６．最初と最後の頁 509-536
掲載論文のDOI（デジタルオブジェクト識別子） https://doi.org/10.5715/jnlp.26.509	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 珊瑚彩主紀，西川仁，徳永健伸	４．巻 26(2)
２．論文標題 外界一人称と二人称を考慮する日本語述語項構造解析の分野適応	５．発行年 2019年
３．雑誌名 自然言語処理	６．最初と最後の頁 483-508
掲載論文のDOI（デジタルオブジェクト識別子） https://doi.org/10.5715/jnlp.26.483	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

〔学会発表〕　計3件（うち招待講演　0件／うち国際学会　2件）

１．発表者名 Mizuki Sango, Hitoshi Nishikawa, and Takenobu Tokunaga
２．発表標題 Effectiveness of Domain Adaptation in Japanese Predicate-Argument Structure Analysis
３．学会等名 The 32nd Pacific Asia Conference on Language, Information and Computation（国際学会）
４．発表年 2018年

１．発表者名 Souta Yamashiro, Hitoshi Nishikawa, and Takenobu Tokunaga
２．発表標題 Neural Japanese Zero Anaphora Resolution using Smoothed Large-scale Case Frames with Word Embedding
３．学会等名 The 32nd Pacific Asia Conference on Language, Information and Computation（国際学会）
４．発表年 2018年

１．発表者名 菅井内音，西川仁，徳永健伸
２．発表標題 ニューステキストの要約及び平易化
３．学会等名 言語処理学会第26回年次大会
４．発表年 2020年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考