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　５ 年　 ６ 月　 ５ 日現在

機関番号：１４５０１
研究種目：基盤研究(C)（一般）
研究期間：2019～2022
課題番号：１９Ｋ１２２４７

研究課題名（和文）表面的特徴に基づいた「やさしい日本語」の自動生成への深層学習の適用

研究課題名（英文）An Application of Deep Learning to generate Simplified Japanese by using "Surfece Charancteristics" of text.

研究代表者
村尾　元（Hajime, Murao）

神戸大学・国際文化学研究科・教授

研究者番号：７０２７３７６１

交付決定額（研究期間全体）：（直接経費）　　3,200,000円

研究成果の概要（和文）：本研究では，表面的な特徴に基づいて，通常の日本語テキストをやさしい日本語に変換するシステムの構築を試みた。主な成果は次の通りである。1. 通常の日本語テキストに比べて，やさしい日本語が得意的に有する表面的な特徴を明らかにした。2. 表面的な特徴に基づいて日本語テキストの「やさしさ」を90%以上という高精度で判定するシステムを構築した。3. T5（Text-to-Text Transfer Transformer）を用いた通常の日本語テキストをやさしい日本語に変換するシステムを試作し，その有用性を示した。

研究成果の学術的意義や社会的意義
本研究により，日本語テキストの「やさしさ」を自動的に判定することが可能となり，また，通常文をやさしい日本語に変換する可能性が示された。これにより，やさしい日本語に関する特別な知識がなくとも，子どもや日本に滞在する外国人に対して情報提供を行うことができる。また，従来より研究されてきた，意味・内容に基づく文章変換と組み合わせることにより，さらに精度を高めることが可能となり，より広範に適用できる可能性がある。

研究成果の概要（英文）：We tried to construct a system to translate ordinary Japanese texts into simplified ones by using the "surface characteristics" of texts, which are not semantic features but lexical features such as how separators like commas, periods, and spaces were used. As a result, we achieved the following: 1. we clarified the difference in the surface characteristics between regular and simplified texts. 2. We constructed a system to evaluate the difficulty of Japanese texts, showing a pretty good result of over 90% accuracy. 3. We employed T5 to convert standard Japanese texts into simplified ones.

研究分野：社会システム科学

キーワード：やさしい日本語　機械学習　機械翻訳　学習支援　深層学習　BERT　Transformer

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

就労・就学を目的として来日する非日本語母語話者の増加と多様化に伴い，教育現場においては，日本語による専門教育が困難な場面が生じている。彼らの多くは日本語を多少なりとも学んでいるが，高度な内容の理解は困難な場合もある。一部の技術系専門学校などでは，専門教育以前に，さながら日本語学校のようになっているという実態も聞く。

これに対して，学生の母語での教育も考えられるが，専門的な内容については，教員の負担も大きい。そもそも，将来的に日本での就労・就学を目指す非日本語母語話者にとっては，日本語での教育が望ましい。したがって，教員の負担を軽減しつつ，「やさしい日本語」による専門教育を実施するためにも，コンピュータによる支援は必要不可欠である。

一方，社会的な要請も大きいことから，コンピュータによる「やさしい日本語」生成支援については，これまでも様々な研究がなされている。それらは大きく，「やさしい日本語」の文章作成支援（田中ら，伊藤らなど）と難解な文章の「やさしい日本語」への自動変換（乾ら，川村ら，梶原ら）に大別される。特に後者の自動変換は本研究とも関連が深いが，それらの多くは，難易度付きの同意語辞書を用いた，難解な用語の平易な表現への置き換えに留まっている。これは文章の内容を維持しつつ，用語の難易度を下げるというアプローチである。

文章の持つ特徴を，「スタイル」（形式や書式）と「コンテンツ」（内容）に分けるならば，これらは「コンテンツ」に主眼を置いた「やさしい日本語」変換と言えよう。教育の観点から言えば，文章の意味が重要であるため，同一の意味を保持しつつ，難解な用語の平易な表現への置き換えが有用であることは疑いようもない。しかし，緊急情報のように短い文章では，難解な用語は少数で，形式的な文体が理解の障壁となる場合も多い。また，専門的知識においては，専門用語の難解さはもちろんだが，独特の言い回しが文章全体の難易度を上げている。したがって，用語の置き換えとは異なるアプローチによる「やさしい日本語」の自動生成もまた必要と考えられる。

２．研究の目的

本研究では，句読点や空白，仮名と漢字の使い分け，改行の使用法といった，文章の表面上の特徴，いわば文章の「スタイル」に注目した「やさしい日本語」への自動変換手法の提案を目的とする。研究の中心的な課題は次のようなものである：

(1)「やさしい日本語」の「スタイル」を明らかにすること。

(2)「やさしい日本語」の「スタイル」に基づいて日本語テキストの「やさしさ」を評価すること。

(3) 日本語テキストの「やさしさ」を評価として，通常の日本語を「やさしい日本語」に変換する手法を構築すること。

従来の「やさしい日本語」への自動変換手法では，変換前後の意味の同一性の維持に主眼が置かれているため，句点や空白，仮名と漢字の使い分け，改行の使用法などの表面上の特徴は，むしろ記述上のゆらぎまたはノイズとして，無視されるのが一般的である。本研究では，従来法においては，そのように扱われていたような情報に着目し、むしろ積極的に用いようとしている点で独自性がある。

とはいえ，本研究で提案する手法は，けっして従来法と相反する物ではない。むしろ，相補的であり，意味を重視した従来法と同時に利用することで，自動変換の精度を高めることが出来ると考えられる。このように，従来法との組合せで新しい手法や結果を生み出すことが期待できる点で創造的といえる。

また，提案手法は従来法で必要とされてきた難易度付きの同意語辞書やコーパスを必要としない。このため，自動変換の際に必要となるコンピュータのリソースはそれほど多くなく，また高速に適用できる点は特徴的である。

３．研究の方法

３.１ 概要

本研究では，第 2 章で挙げた 3 つの課題に従って研究を進める。文頭の数字はそれぞれ課題に対応している。

(1) まず，通常の日本語テキストと，やさしい日本語テキストの特徴量を比較し，やさしい日本語に特徴的なスタイルを明らかにする。その目的で，機械学習による分類手法であるランダムフォレストと，分類に有用な特徴量を分析する Permutation Importance を用いる。

(2) 続いて，得られた特徴量を用いて，日本語テキストの難易度を推定する手法を構築する。ここでは機械学習手法の BERT を用いる。

(3) 最終的に通常の日本語テキストをやさしい日本語に変換する手法を提案する。この目的で同じく機械学習手法の Transformer を用いる。Transformer の学習のためには入力テキストと変換後のテキストが必要であるため，まず，通常の日本語テキストとそれをやさしい日本語に変換したテキストから，通常の日本語文とやさしい日本語文の対訳コーパスを生成する。このために，DTW（Dynamic Time Warping）を用いた手法を提案する。これを用いて Transformer を学習する

ことで通常の日本語テキストをやさしい日本語に変換するシステムが実現される。
図 1 に全体のシステム構成を示す。上記(1)及び(2)が STEP2 に，上記(3)が STEP1 に対応する。

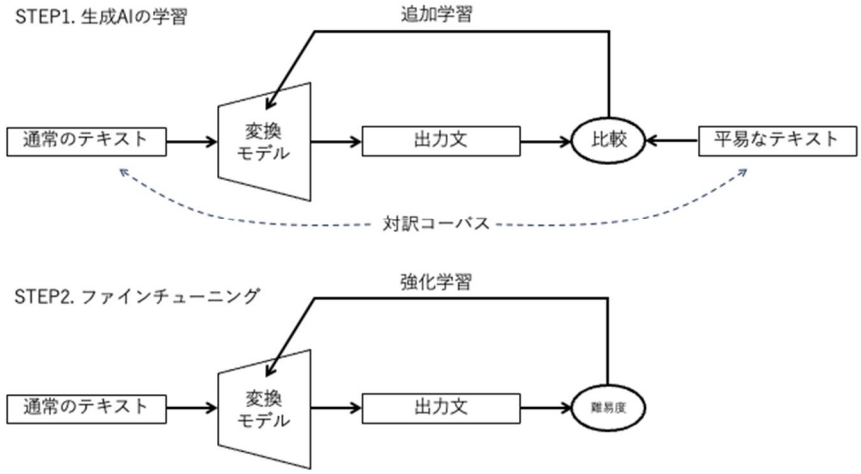

図 1 提案システムの全体像

３.２ やさしい日本語のスタイル

３.２.１ 検討した特徴量

前処理の結果から，以下の特徴量を計算した。
・単語数：テキストに含まれる形態素の数。
・漢字率：テキストに含まれる漢字数を文字数で割った値。
・外来語率：全ての文字がカタカナである形態素数を単語数で割った値。
・受身率：「れる」もしくは「られる」である接尾語を単語数で割った値。
・サ変接続名詞率：品詞が「名詞」かつ品詞細分類が「サ変接続」と一致する形態素数を単語数で割った値。
・副詞率：品詞が「副詞」である形態素数を単語数で割った値。
・読点率：「、」または全角の「，」である形態素数を単語数で割った値。
・否定率：品詞が「助動詞」で，「ない」「ぬ」「ん」と一致する形態素数を単語数で割った値。
・出現頻度最大値：単語の出現頻度が高いほど，単語の難易度は低くなる傾向がある。従って，品詞が「名詞」「動詞」「形容動詞」「形容詞」である形態素の出現頻度を特徴量のひとつとして用いる。
・出現頻度平均値：品詞が「名詞」「動詞」「形容動詞」「形容詞」である形態素数の平均値。
・係り受け平均距離：テキストの先頭から文節ごとに係り受け距離を求め，これを文節の数で割った値。
・係り受け最大距離：文節単位での係り受け距離の最大値。
・係り受け被修飾数：文節単位での係り受けの個数。

３.２.２ やさしい日本語の特徴量の抽出

まず，各テキストを「通常文」または「やさしい日本語」のどちらかに分類し，ラベルを付与する。各テキストの特徴量を入力，そのテキストのラベルを教師としてランダムフォレストにより学習をする。十分に学習し，テキストを「通常文」と「やさしい日本語」に分類できるモデルが得られたら，次に Permutation Importance を計算する。

Permutation Importance は入力の各次元（ここでは特徴量）をランダムにシャッフルすることで，その次元の重要性を検出する。重要な次元をシャッフルした場合は，分類性能が大きく下がる。

３.３ 日本語テキストの難易度推定

ランダムフォレストは 80%を超える高い性能で「通常文」と「やさしい日本語」を分類することができるが，より高い性能を実現するために深層学習の一つである BERT を用いる。すなわち，形態素列を入力とし，３.２.２で付したラベルを教師として，事前学習済みの BERT を追加学習する。

３.４ やさしい日本語への変換システムの構築

３.４.１ 概要

図 1 の STEP1 にあるように，Transformer を学習するためには対訳コーパスが必要となる。そのためにまず，通常テキストと，対応するやさしい日本語テキストから，対訳コーパスを自動生成する。次に，生成された対訳コーパスを用いて，Transformer を学習する。

３．４．２ 対訳コーパスの自動生成

やさしい日本語対訳コーパスの自動生成に DTW（Dynamic Time Warping）を用いる。手順は以下の通り。

1. 通常テキストとやさしい日本語テキストをトークンに分解し，さらに，それぞれのトークンを埋め込み表現に変換する。
2. テキストをベクトル列とみなし，DTW を適用。通常テキストとやさしい日本語テキスト間のトークンごとの対応を求める。
3. 通常テキストの１文に対応するやさしい日本語テキストの部分は抽出し，コーパスに登録する。

(1) DTW による対応トークンの発見

DTW は異なる時系列データの間で，各パスの距離を総当たりで計算し，時系列データ間の最短距離を求める手法である。これを各トークンを埋め込み表現に変換したテキストに適用することで，トークンの意味が近いほどパスの距離が近くなる。また最短ルート距離はテキストデータ間の距離，即ち類似度を表すことになる。

またテキスト間の距離を計測する別の指標である Word Mover's Distance(WMD)を用いて，DTW によるコーパスの妥当性を検証する。WMD は一方の文のあるトークンから，他方の文のトークンへの移動コストを距離とするもので，一方の文全体を他方の文全体に変換する際の距離によって文の類似度を計算する。本研究では，移動コストをトークンベクトル間の距離とする。

３．４．３ T5（Text-to-Text Trasfer Transfomer）の学習

日本語大規模データによって事前学習済みの T5 を，上記手法によって生成したやさしい日本語コーパスにより追加学習を行う。即ち対訳コーパスの「通常文」を入力として変換後の「出力文」を得る。これがやさしい日本語になっているかどうかを評価し，学習する。

４．研究成果

４．１ データ

本研究では，通常テキストとして，NHK NEWS WEB の記事からテキストを抽出し，やさしい日本語のテキストとして，対応する NEWS WEB EASY の記事のテキストを抽出した。抽出した段階では文同士の対応はない。NEWS WEB EASY は外国人や子供向けに分かりやすい言葉でニュースを書き換えたものである。書き換えは日本語教師と記者が共同で行なっている。

データを収集した期間は 2020 年 8 月から 2021 年 11 月の期間に収集した，1,040 本の記事を対象とし た。NEWS WEB EASY の Web ページには「普通のニュースを読む」というリンクがあり，NHK NEWS WEB の記事と対応が取れている。

４．２ やさしい日本語のスタイル

まず，３．２．１に挙げた特徴量を抽出し，これをツリー数 100，ツリーの深さ最大 5，ノード数の制限なしというパラメータのランダムフォレストで学習した。その結果，分類の正解率は 82.6%であった。これに Permutation Importance を適用した。図 2 に特徴量の重要度，即ち，正解率の下降の大きいものから順に並べて示す。「サ変接続名詞率」，「単語数」，「受身率」，「漢字率」の 4 つの特徴料において，分類の精度が有意に下降しており，これらの特徴量が通常文とやさしい日本語の分類に重要であることが示唆された。

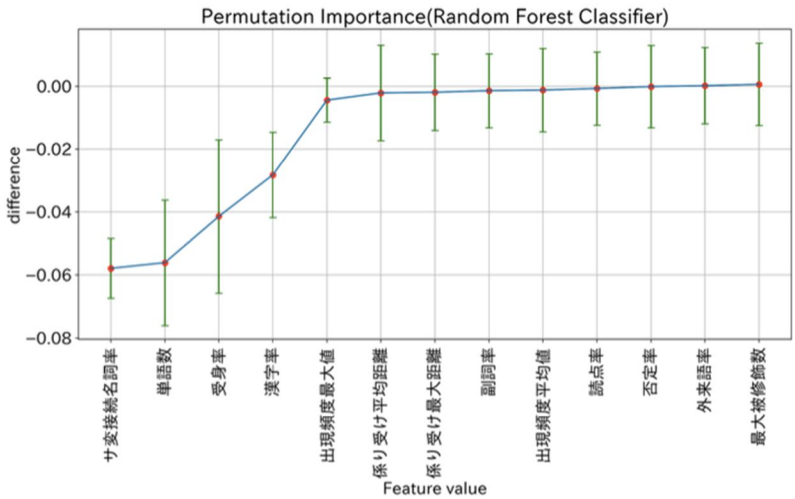

図 2 Permutation Importance の結果

４.３ テキストのやさしさの推定

３.３の実験結果，すなわち，BERT による難易度推定については，学習（エポック）が進むごとに精度が改善されており，最終的には，学習用データに対する正解率が 96.4%，未学習の検証用データに対して 95.7%と良好な結果が得られた。

BERT による推定については，テキストのどのような特徴を用いて推定しているのか検証を行い，ランダムフォレストと同様の特徴量に注目して推定が行われていることが明らかとなっている。

４.４ T5 によるやさしい日本語生成

３.４の実験結果，すなわち，T5 に入力した通常文と，やさしい日本語への変換後の出力文の例を図 3 に示す。定量的な評価は行なっていないが，テストデータ 29 件に対して BLEU の平均評価は 0.34 となった。うまく生成されている場合もあれば，矛盾や相反する出力文を生成する場合もある。

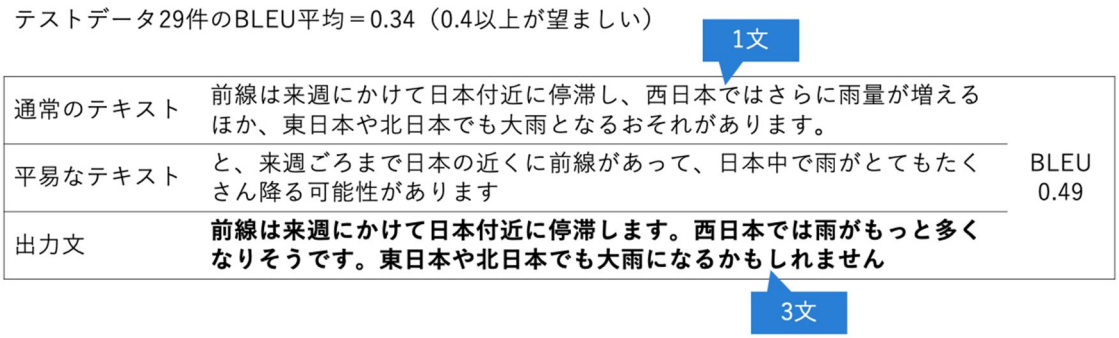

図 3 やさしい日本語生成の例

４.５ まとめ

本研究では通常の日本語テキストを，やさしい日本語テキストに自動変換するシステムを構築した。大きなシステムであり，それぞれの箇所について精緻な検証はできていない部分もあるが，概ね以下のような成果が得られた。

1. スタイル（表面的な記述）における，やさしい日本語の特徴を明らかにした。
2. スタイルに基づいて日本語テキストの「やさしさ」を評価するシステムを構成した。正解率は 90%を超える精度である。これを用いることで，日本語の文章がやさしい日本語と言えるかどうかの判定を自動的に行うことができる。
3. 通常文をやさしい日本語に変換するシステムのプロトタイプを作成した。定量的な評価は行なっていないが，ある程度，やさしい日本語を生成することができ，可能性が示された。

５．主な発表論文等

〔雑誌論文〕　計5件（うち査読付論文　5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 Eri Maekawa, Hajime Murao	４．巻 Online
２．論文標題 The Comparison of Word Embeddings and Feature Vectors in Text Classification by Difficulty Level	５．発行年 2021年
３．雑誌名 Proceedings of the 15th International Conference on Innovative Computing, Information and Control (ICICIC2021)	６．最初と最後の頁 -
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Hajime Murao	４．巻 ICICIC2019-065
２．論文標題 A Study on Finding Differences in Movement of Expert and Novice Darts Players by Using a Kinect-Like 3D Image Sensor	５．発行年 2019年
３．雑誌名 Proc. of the 14th International Conf. on Innovative Computing, Information and Control	６．最初と最後の頁 1-6
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Megumi Kawata, Hajime Murao	４．巻 ICICIC2019-159
２．論文標題 Estimating Desk Work Status from Video Stream Using a Deep Neural Network	５．発行年 2019年
３．雑誌名 Proc. of the 14th International Conf. on Innovative Computing, Information and Control	６．最初と最後の頁 1-4
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Eri Maekawa, Hajime Murao	４．巻 13
２．論文標題 Interpreting BERT Attention Trained for Japanese Difficulty Classification from the Viewpoint of Grammatical Features	５．発行年 2022年
３．雑誌名 ICIC Express Letters, Part B: Applications	６．最初と最後の頁 697-703
掲載論文のDOI（デジタルオブジェクト識別子） 10.24507/icicelb.13.07.697	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 Eri Maekawa, Hajime Murao	4．巻 -
2．論文標題 A PROPOSAL TO CREATE A PSEUDO-PARALLEL TEXT CORPUS FOR SIMPLIFYING JAPANESE USING DTW	5．発行年 2023年
3．雑誌名 INTED2023 Proceedings (The Proc. of the 17th Int. Technology, Education and Development Conf.)	6．最初と最後の頁 6542-6550
掲載論文のDOI（デジタルオブジェクト識別子） 10.21125/inted.2023.1745	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 前川 絵吏，村尾　元
2．発表標題 日本語の難易度に関する特徴分析
3．学会等名 言語処理学会 第27回年次大会
4．発表年 2021年

1．発表者名 Eri Maekawa, Hajime Murao
2．発表標題 Analysis of the Behavior of Foreign Tourists Using Mobile Translation devices
3．学会等名 The SICE Annual Conference 2020（国際学会）
4．発表年 2020年

1．発表者名 Megumi Kawata, Hajime Murao
2．発表標題 Study on the Effect of Appearance of Personified Agents in Persuation
3．学会等名 The SICE Annual Conference 2020（国際学会）
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

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関