



テキスト平易化のためのコーパス自動生成と評価に関する研究

前川, 絵吏

(Degree)

博士 (学術)

(Date of Degree)

2024-03-25

(Date of Publication)

2025-03-01

(Resource Type)

doctoral thesis

(Report Number)

甲第8804号

(URL)

<https://hdl.handle.net/20.500.14094/0100490029>

※ 当コンテンツは神戸大学の学術成果です。無断複製・不正使用等を禁じます。著作権法で認められている範囲内で、適切にご利用ください。



論文要旨

氏名 前川 絵吏
専攻 グローバル文化専攻
指導教員氏名 村尾 元

テキスト平易化のためのコーパス自動生成と評価に関する研究

論文要旨

総務省出入国管理庁の報道発表資料によると、令和 4 年末の在留外国人数は初めて 300 万人を超えて過去最高を更新した。外国人が日本で生活する上で、言語の壁により不便を余儀なくされている生活環境を改善するため、多言語翻訳に加えて「やさしい日本語」の活用が提唱されている。テキストをやさしい表現へ自動変換することをテキスト平易化 (Text Simplification) といい、機械翻訳やテキスト生成の発展に伴いめまぐるしく進化している研究分野の一つである。機械翻訳モデルのように学習データを用いた手法は対訳コーパスが大量に必要なが、日本語におけるテキスト平易化のための十分な規模の平行コーパスが存在しないため、機械翻訳と同様の手法で成果を得ることができない。そこで、次の 3 つのアプローチでテキスト平易化を提案する。

1. 日本語の難易度推定に影響する特徴量を調査し、難易度分類モデルを構築する。
2. テキストの意味的類似性を測定するために相応しい評価指標を調査する。
3. 平易化平行コーパスの自動構築手法を提案し、事前学習済みモデルを使ってテキスト平易化を実現する。

1 つめの日本語の難易度推定では、日本語のテキストを難易度によって二値分類するモデルを構築する。具体的には、関連研究から日本語の難易度に影響を及ぼす特徴量を列挙し、テキストを特徴量ベクトルに変換した。特徴量ベクトルを用いてランダムフォレストで難易度を推定するモデルを構築した。その後、分類に影響を与える主要な特徴量を調べるために Permutation Importance で特徴量の重要度を調べたところ、特定の特徴量が難易度分類に影響することがわかった。これらの重要な特徴量が、深層学習モデルの計算過程でどのように解釈されているかを分析するために、BERT モデルのアテンションを可視化したところ、先行研究では 12 層から成る BERT のアテンション層は 10 層目から 12 層目で分類の特徴を捉えるという報告があったが、本実験では各層によってそれぞれ異なる特徴量を捉えていることが明らかになった。難易度推定モデルの正解率はランダムフォレストが 82.6%、BERT モデルが 96.4% となり、日本語の難易度推定に関する関連研究よりも本研究の BERT モデルが精度が高い結果となった。

2 つめに、平易化平行コーパスの類似性評価に相応しい指標を検討した。文同士の類似度を計測する手法は数多く存在しているが、本研究で対象とするのは平易化前後のテキストの意味的類似度のことを示す。したがって、単語や文法的一致率を評価する既存手法は用いず、単語ベクトル列から文ベクトルを生成して類似度を測定する方法、

または文の類似性を深層学習モデルで学習する手法を選択する。単語をベクトル化して文をベクトル列とし、ベクトル間の距離を測定する手法として Dynamic Time Warping(DTW), Word Mover's Distance(WMD)を用いた。テキストから文ベクトルを生成する手法として Simple Word-Embedding Model(SWEM), Doc2Vec, Sentence-BERT を使用し、コサイン類似度で求めた。深層学習を利用する手法として 2 つの文の類似度を学習したモデルである BERTScore を用いた。実験に用いたテキスト対は、JSICK, SNOW T23, およびオリジナルコーパスである。実験の結果から、WMD, Sentence-BERT および BERTScore が意味的類似度を測定するために有効な指標であることが明らかになった。

3 つめのテキスト生成について、本研究ではテキスト平易化パラレルコーパスを自動構築する手法を提案した。既存手法としては、Simple English Wikipedia と English Wikipedia の対応するテキストから、文の類似度が近いものを文対とする研究がある。他にも一つの文ともう一方の文の対応を調べる研究がいくつかあるが、これらの手法では書き換えの際に一つの文を複数に分割したり、複数の文を一つの文に統合したりすることが考慮されていない。そこで、人手によって書き換えられたニュース記事から、文で区切ることなく記事全体を単語単位で対応付けすることで、平易化パラレルコーパスを抽出する手法を提案した。書き換えによって別の単語に置き換わることを想定して、表面的な単語の一致ではなく単語の意味的な距離が近いことを測る手法を提案する。単語をベクトル化すると、意味が似ているベクトルは近くに配置される性質を利用して、文全体をベクトル列とみなし、文間距離が最短となるときの意味的に近い単語のアライメントをとった。提案手法によって構築した平易化パラレルコーパスは、文の統合や分割にも対応できており、一つの文と複数の文を対応づけることに成功した。さらに、事前学習済みのテキスト生成モデルを平易化パラレルコーパスでファインチューニングし、平易化モデルを構築して書き換え後のテキストを 1 つめの難易度推定モデル、および 2 つめの類似度測定指標で確認した。難易度推定モデルでは出力文の 89.7%が平易文であるという結果となり、類似度測定では平易化パラレルコーパスの平均よりも類似度が高い結果となった。

以上より、日本語におけるテキスト平易化ではパラレルコーパスが少ないという問題に対して、パラレルコーパスを自動構築する手法を提案し、テキスト生成モデルをファインチューニングしてテキスト平易化を実現した。テキスト平易化の品質推定においては平易性を測る分類モデルを構築し、類似性を測る最適な指標を明らかにすることができた。これまでに提案した手法や分析についてさらに課題を検討し、テキスト平易化システムの実用に向けて研究を継続していきたい。

論文審査の結果の要旨

氏 名	前川 絵吏			
論文題目	テキスト平易化のためのコーパス自動生成と評価に関する研究			
判 定	合 格 ・ 不 合 格			
論文チェックソフトによる確認	☒ 確認（1回目 1/29, 2回目 2/20） ☐ 未確認 理由：			
審 査 委 員	区 分	職 名	氏 名	論文審査結果について
	委 員 長	教 授	康 敏	☒ 確認
	委 員	教 授	村尾 元	☒ 確認
	委 員	准教授	齊藤 美穂	☒ 確認
要 旨				
本論文は、日本語テキスト平易化作業の自動化を目的として、平易化日本語対訳コーパス（以下「対訳コーパス」）の生成に関するコンピューティング手法を提案しようとするものである。対訳コーパスは、通常難易度の日本語文に対して、同じ意味の、平易化された日本語文を対応付けした並列コーパスである。具体的に、本論文では、通常難易度の日本語テキストとそれを手動で平易化した日本語テキストから、対応する文の対を自動的に抽出する手法を提案、評価している。こうして生成された対訳コーパスを、機械翻訳に利用される機械学習モデルに学習させることで、通常難易度の日本語テキストを、平易化された日本語に“翻訳”することが可能となる。 本論文は 5 章から構成されている。第 1 章は研究の目的と背景および提案手法の概要、さらに論文の構成について書かれている。第 2 章と第 3 章では対訳コーパスの質評価手法の検討を行っており、第 4 章では対訳コーパスの自動生成手法の提案と生成した対訳コーパスの質評価を行っている。いずれの章においても、先行研究をはじめとする関連研究の現状と本論文の位置付けについて、十分な量の既発行の文献を参照しながら説明されている。第 5 章はまとめと結論である。 対訳コーパスの質評価に関しては、難易度と類似度という 2 つの指標について手法を検討している。まず、第 2 章では、難易度の推定手法として、BERT とランダムフォレストという 2 種類の機械学習モデルを検討している。いずれの手法についても、NHK が提供する NEWS WEB か				

ら収集した日本語文を「通常難易度」、同じく NHK が提供する NEWS WEB EASY から収集した日本語文を「平易」として、2 クラス分類を行うように学習し、その後、未学習の日本語文に対してクラス分類テストを行っている。その結果、ランダムフォレストの 82.6% に対して、BERT で 96.4% という正解率を得ている。これは先行研究における 89.6% という重回帰分析の結果と比べても高い精度である。

第 3 章では、類似度の推定について、機械学習により日本語文を密ベクトル化して比較する 6 種類の手法を検討している。先行研究において手動で作成された 2 種類の対訳コーパス JSICK と SNOW T23 を対象として、それらのコーパスに含まれる文対について、そのままの文対（これは類似）と、ランダムに相手を入れ替えた文対（これは類似していない）を用意する。全ての文対について、6 種類の手法を用いて計算した類似度を説明変数、「類似」または「非類似」の分類を目的変数としてロジスティック回帰分析を行い、説明力について比較している。その結果、Sentence-BERT の有効性が高いことが示されている。

第 4 章では、対訳コーパスの自動生成について、DTW (Dynamic Time Warping) という動的計画法に基づく手法の利用を提案している。DTW は、信号処理において、2 つの信号間の類似性を発見するために利用される手法である。本論文では、2 つのベクトル化された文対を 2 つの信号とみなす。まず、実験 1 で、NEWS WEB の記事を通常難易度の日本語テキスト、それに対応する NEWS WEB EASY の記事を平易化した日本語テキストとして、対応する文対の抽出を試みている。こうして生成された対訳コーパスに含まれる文対について、第 3 章で検討した Sentence-BERT を用いて類似度の計算を行ったところ、SNOW T23 の平均 0.81 (表 3.5) に比べて、提案手法で生成した対訳コーパスでは平均 0.38 となっており、手動で生成した対訳コーパスに比べて類似度は低い。しかし、通常難易度の日本語 1 文に対して、複数の平易化された日本語文を対応させることができるなど、1 文対 1 文ではない対応を抜き出すことが可能であることが示されている。

最後に、実験 2 として、生成された対訳コーパスを T5 という機械翻訳に利用される深層学習モデルに学習させ、そののち、未学習の通常難易度の日本語文を入力、出力された日本語文の質評価をしている。その結果、89.7% の出力文は平易であり、類似度平均は 0.86 となった。この類似度は SNOW T23 の 0.81 に比べると高く、文の意味を保ったまま通常文の平易化が行えていると言える。

本論文では、通常難易度の日本語テキストと、それに対応する平易化された日本語テキストから、自動的に対訳コーパスを生成する新規な手法を提案している。また、対訳コーパスの質評価に関する手法を検討し、提案手法によって生成された対訳コーパスが、手動で作成された対訳コーパスと比べても遜色ないこと、つまり提案手法の有効性を示した。

本論文の一部は 1 件の国際論文誌に掲載済みであり、また 1 件の国際論文誌に掲載が確定している。また 4 件の査読付き国際会議での口頭発表が行われており、国内会議 1 件と合わせると合計 7 件の研究発表が行われている。いずれも情報分野に関する学術会議である。

本研究は情報分野、特に自然言語処理の分野において学術的に価値があるだけでなく、日本語教育など平易化された日本語が必要となる様々な場面で有用であり、学術的のみならず、実際の場面で今後貢献できると期待される。

よって、学位申請者の前川 絵吏氏には博士（学術）の学位を得る資格があると判定する。