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

博 士 論 文

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

令和 6 年 2 月

神戸大学大学院国際文化学研究科

前川 絵吏

論 文 要 旨

氏　　名　　: 前川　絵吏
所属専攻　　: グローバル文化専攻
コース名　　: 情報コミュニケーションコース
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

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であるが，日本語におけるテキスト平易化のための十分な規模のパラレル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. テキスト平易化による書き換え前後の文の類似度を測定するために相応しい評価指標を調査する．
3. 平易化パラレル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，難易度分類に影響を与える重要な特徴量が判明した．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，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．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ているという報告があったが，実験では層による解釈の違いを発見することができた．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，先行研究よりも本研究の BERT モデルが精度が高い結果が得られた．

2 つめに，平易化パラレルコーパスの類似性評価に相応しい指標を検証した．文同士の類似度を計測する手法はいくつか存在しているが，本研究では平易化前後のテキストの意味的類似度を対象とする．したがって，単語の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，コサイン類似度で文間類似度を求めた．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．実験に用いたデータセットは，JSICK および SNOW T23 である．類似度測定指標を説明変数，文の類似度を目的変数としたロジスティック回帰モデルを用いて説明力の高い指標を調べたところ，最も有効であるのは Sentence-BERT であることが明らかになった．

3 つめのに，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，実用的な書き換えで起こりうる編集操作である一つの文から複数文への分割，および複数文から一つの文への統合が考慮されていなかった．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，入力文と出力文の類似度を測定したところ Sentence-BERT のスコアで 0.86 となり，SNOW T23 の平均値 0.82 と比較しても高い値が得られた．

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．

目次

図目次

表目次

第 1 章

はじめに

1.1　研究の背景と目的

背景

総務省出入国在留管理庁の報道発表資料「令和 4 年末現在における在留外国人数について」[*1]によると，令和 4 年末の在留外国人数は，307 万 5,213 人（前年末に比べ 31 万 4,578 人 11.4% 増加）で過去最高となった．在留外国人数が増加している理由のひとつとして，日本の少子高齢化による労働力不足の問題を海外の労働力に頼る状況が続いていることが挙げられるが，少子化の解決には時間がかかるため今後も在留外国人の増加傾向が続くことが予想される．外国人へ情報を伝える際，理想的であるのは全ての外国人の母語に翻訳して伝えることであるが，世界には 7,000 を超える言語が存在するとも言われており，あらゆる言語への翻訳は現実的ではない．そこで，普通の日本語よりも簡単で，外国人にもわかりやすい「やさしい日本語」の活用が注目されている．

テキスト平易化（Text Simplification）の研究は，英語を中心としてニューラルネットワークや生成 AI の発展に伴い目まぐるしく進化している．テキスト平易化は，機械翻訳と同様にテキストからテキストへの言い換えタスクとして位置付けられる．機械翻訳タスクでは，膨大な学習データを用いたニューラル機械翻訳が主流となり Long Short-Term Memory（LSTM）に代表される Recurrent Neural Network（RNN）に基づく翻訳モデルを中心に研究が発展してきた．しかし，機械翻訳と異なりテキスト平易化のための学習データが充分にないことから，同じ手法を適用しても学習がすすまないという課題があった．英語においては Wikipedia 等から平易化のための学習データを構築して成果が見られたが，日本語においては同様の規模のコーパスはまだ構築されていない．

[*1] urlhttps://www.moj.go.jp/isa/index.html 最終閲覧日：2023 年 9 月 20 日

自動評価

テキストの言い換えタスクでは，参照文を用いた評価が一般的である．その代表的な手法として，一般的に BLEU[1] や SARI[2] のスコアが用いられる．図 1.1に自動評価のしくみを示す．評価に用いる参照文とは人手で書き換えた正解文のことであり，この文に近いほどスコアが高くなる．機械翻訳の指標である BLEU は，出力文と参照文の語彙一致率で評価する．例えば日英翻訳の場合，入力文「日本の桜は春に美しい花を咲かせます。」に対する参照文が“Japanese cherry blossoms bloom with beautiful flowers in spring.”とすると，出力文は参照文と語彙・語順が一致するほど評価が高い．BLEU は機械翻訳の評価を目的として設計されているため，テキスト平易化タスクにそのまま適用することが不適切であるという指摘がある [3]．先に述べた例文を平易化したときの参照文が「日本の桜は春にきれいな花を咲かせます.」だとすると，モデルが入力文をそのまま「日本の桜は春に美しい花を咲かせます。」と出力すれば書き換えを行なっていないにも関わらず BLEU の評価が高くなってしまう．そこで平易化に適切な評価方法として，文の編集操作をもとに評価する SARI が提案された．SARI は，入力文から参照文への編集操作（単語の削除，単語の追加，単語の保持）をもとに評価するため，モデルに積極的な書き換えを促す．具体的には，入力文に含まれていて参照文に含まれない単語が削除されたか，入力文に含まれず参照文に含まれる単語が追加されたか，入力文にも参照文にも含まれる単語が保持されたか，の三つの観点での評価の平均をとる．テキスト平易化の評価に利用される，入力文と意味的に一致する参照文のペアの集合をパラレルコーパス（並列コーパス）という．

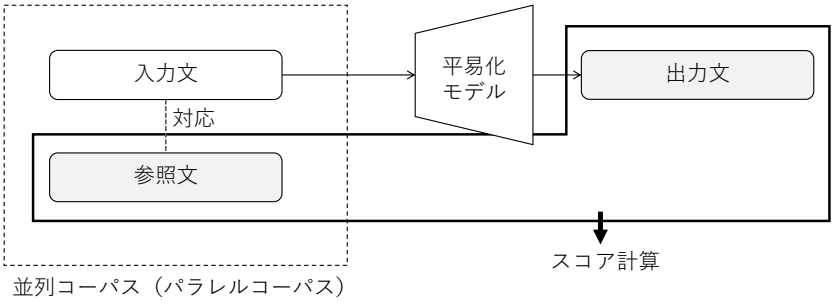


図 1.1　テキスト平易化のための自動評価のしくみ

平易化のためのコーパス

自動評価に利用できる既存のパラレルコーパスをいくつか挙げる．テキスト平易化に利用できるコーパスとしては，英語のニュース記事に対して学年ラベルが付与された Newsela データセット

[4] や，English Wikipedia と Simple English Wikipedia のテキストから構築したコーパス [5] が代表的である．日本語においては，難易度ラベルが付与されたデータセットや平易に言い換えたパラレルコーパスは少ない．代表的なコーパスである SNOW T23：やさしい日本語拡張コーパス [6] は，独自定義した UniDic 単語体系の 2,000 語に語彙制限し，人手で書き換えたコーパスである．難易度の高い単語を限定的な単語で書き換えるため，文によっては元の文より長くなり冗長な表現になってしまうものが含まれている．たとえば「政府は取締の一環として不法入国外国人を追放しています。」という文に対して「政府は悪い者を自由にしない流れの中で、法律に従わないで国へ入って来る外国人を、強い力で外へ出している。」[*2] という不自然な表現を含む平易文に書き換えられている．既知の語彙で伝わる一方で，可読性は落ちてしまうことがあるため，文長差が大きい文対や流暢性が低い文対をあらかじめ学習データから除去する研究がある [7]．また，平易化のためのパラレルコーパスを生成する先行研究には文の類似度から文ペアを作る研究報告 [8, 9] が主流である．

品質推定

自動評価にはパラレルコーパスが必要であるが，パラレルコーパスを構築するにはコストがかかるという課題があり，日本語のように資源が少ない言語では SARI などの自動評価はあまり活用できていない．そこで，図 1.2のように参照文を用いないテキスト平易化のための品質推定の研究が提案された [10, 11]．平易化モデルの出力文に対して，出力文そのものが文法的に正しいこと（文法性），入力文と出力文の意味が変わらないこと（類似性），出力文が平易であること（平易性）で評価する．これらの観点で品質評価のためのデータセットが発表されており [12, 13, 14]，品質推定のデファクトスタンダードになってきているが，日本語に関する品質推定の研究はまだ報告数が少ない．

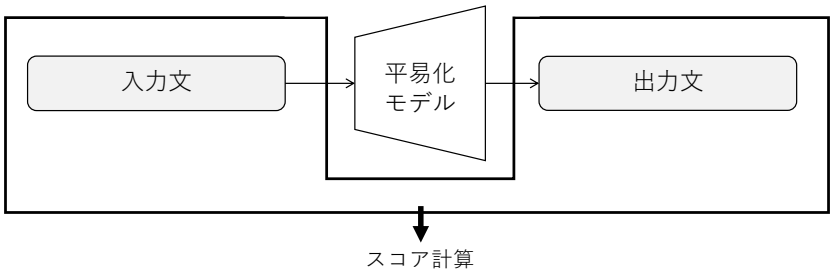


図 1.2　テキスト平易化のための品質推定のしくみ

[*2] SNOW T23: やさしい日本語拡張コーパス *Ab*_1281 より抜粋

目的

本研究でのテキスト平易化の目的は，日本語学習者のための読解支援である．他にも子どもの読解支援や機械翻訳の前処理を目的としてテキスト平易化が行われることがあるが，本研究の対象としない．

やさしい日本語は定義が明確に示されておらず，文化庁のガイドライン[*3]や地方自治体のガイドライン[*4][*5][*6]がそれぞれルールや手引きを作成している．各ガイドラインの目的としては災害時のコミュニケーションや日常生活における情報収集などさまざまで，やさしい日本語のルールが統一されているわけではない．本研究が目指すテキスト平易化とは，日本語教師がやさしい日本語へ書き換える操作を手本とし，それをモデル化することである．やさしい日本語の読み手には，初級から中級レベルの日本語学習者を設定している．

テキスト平易化を実現するには，先述のとおり学習データとしてのパラレルコーパスが充実していないということが大きな課題であった．本研究では平易化のためのパラレルコーパスを自動生成する手法を提案し，平易化モデルを構築することを 1 つめの目標とする．そして，平易化モデルの出力文を品質評価することを 2 つめの目標とする．

図 1.3に，本研究におけるテキスト平易化の実現のためのプロセスを示す．テキスト平易化モデルの学習に必要なパラレルコーパスを生成する手法を提案する（図 1.3の (1)）．テキスト平易化の学習データとして，NEWS WEB EASY[*7]のテキストを選ぶ．NEWS WEB EASY は，NHK NEWS WEB[*8]の記事を日本語教師と記者が共同編集でやさしい日本語に書き換えたものである．やさしい日本語に書き換えた文章と書き換える前の文章を使って，書き換え前後の対応するフレーズを抽出してパラレルコーパスを生成する．先行研究では文の類似度が高い文ペアを作る方法が主流であるが，本研究では文の区切りに依らない書き換え前後の意味的な一致を目指す．

作成した平易化パラレルコーパスを用いて平易化モデルを構築する（図 1.3の (2)）．本研究では事前学習済みモデル T5[*9]を用いてテキスト平易化を行う．Wikipedia，OSCAR，CC-100 の日本語データを用いて事前学習を行なったもので，語彙や文法を獲得済みである．平易化パラレルコーパスを用いてファインチューニングし，平易化モデルを構築する．

[*3] urlhttps://www.bunka.go.jp/ 最終閲覧日：2024 年 2 月 18 日
[*4] urlhttps://www.pref.shizuoka.jp/kurashikankyo/ 最終閲覧日：2024 年 2 月 18 日
[*5] urlhttps://tabunka.tokyo-tsunagari.or.jp 最終閲覧日：2024 年 2 月 18 日
[*6] urlhttps://www.sic-info.org/support/ 最終閲覧日：2024 年 2 月 18 日
[*7] https://www3.nhk.or.jp/news/easy/ 最終閲覧日：2024 年 2 月 18 日
[*8] https://www3.nhk.or.jp/news/ 最終閲覧日：2024 年 2 月 18 日
[*9] https://huggingface.co/sonoisa/t5-base-japanese 最終閲覧日：2024 年 2 月 18 日

平易化モデルの出力テキストを評価するには，品質推定の平易性および類似性に基づく方法で，正解文を必要としない評価を目指す．平易性の評価には，テキストを入力すると「平易文」か「通常文」かを推定するモデルを作る（図 1.3の (3)）．類似性の評価には，いくつかの類似度測定指標から，平易化によって書き換えたテキストの意味的類似度を測定するために相応しい指標を調査する（図 1.3の (4)）．

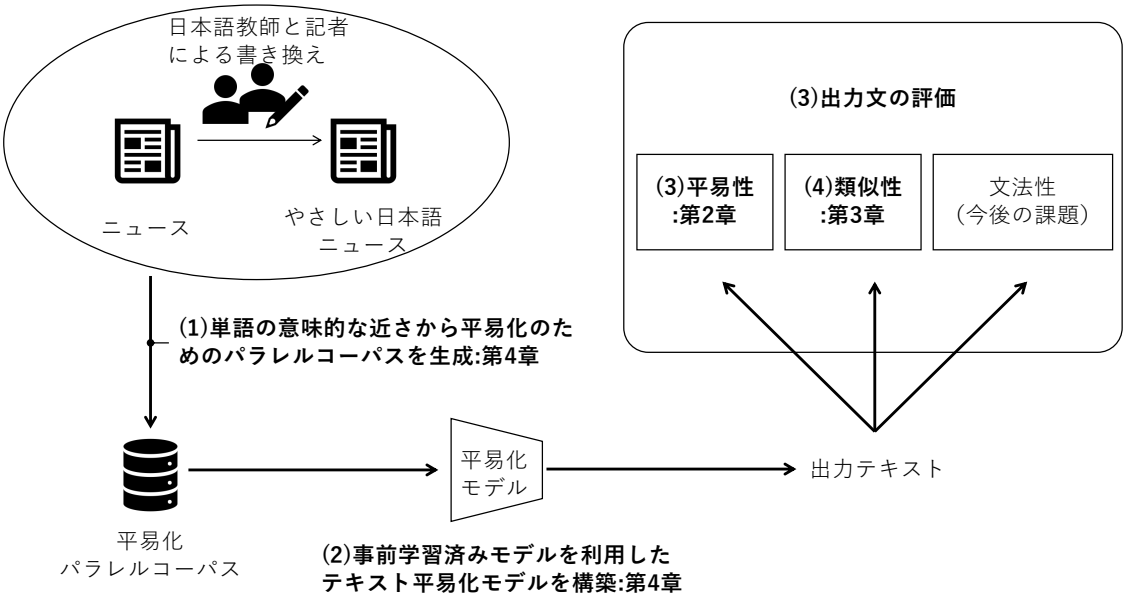


図 1.3　本研究において目指すテキスト平易化の枠組み

1.2　論文の構成

本論文の構成は，以下の通りである．第 2章では，テキスト平易化のための品質推定の一つであるテキストの難易度推定について述べる．テキストの難易度推定モデルを構築し，テキストの難易度に関わる重要な特徴量と深層学習の計算過程から，難易度推定モデルの解釈性を明らかにする．第 3章では，テキスト平易化のための品質推定の一つであるテキストの類似度測定について述べ

る．類似度評価のためのいくつかの指標のうち，平易化モデルの入力文と出力文の意味的類似度を測るために相応しい指標を検証する．第 4章では，テキスト平易化のためのパラレルコーパスを自動構築する手法を提案する．この手法で構築したパラレルコーパスを用いて生成モデルを学習し，モデルの出力テキストを難易度および類似度で評価する．最後に，第 5章では本研究での総括を行い，今後の課題について述べる．

第 2 章

日本語のテキスト難易度に関する分析

2.1 概要

機械学習には質の良いパラレルコーパスが必要であるが，テキスト平易化のような同一言語上で異なる難易度へ書き換えることは日常的に行う動作ではないため資源が豊富ではない．そこで本論文では，少量のパラレルコーパスよりテキスト平易化モデルを構築し，平易性・類似性で品質推定する方法を提案する．本章では品質推定における平易性を評価するための研究について述べる．

日本語の難易度評価やリーダビリティ推定に関する研究は語彙や文法から推定する手法が主流であったが，近年では深層学習を利用する手法がその精度の高さから注目されている．深層学習はブラックボックスで推定の根拠が明らかではないが推定精度は高い．テキストを処理する深層学習モデルのうち，BERT[15] は事前学習によって単語埋め込み表現を獲得する手法のひとつで，Transformer[16] のエンコーダを利用するためアテンションの重みに推定の注目度が表れると言われている．そこで，本研究では BERT モデルを用いたテキストの難易度推定において，先行研究と推定精度を比較し，さらに BERT モデルの説明可能性を調査する．

2.2 テキストの難易度推定に関する先行研究

2.2.1 日本語の難易度を表す特徴

日本語学習者にとっての文の難易度を判定する研究において，語彙の難易度や文法の複雑さに基づく方法が発表されている．川村らは日本語学習者のための文章の難易度判定システムの構築に関する研究 [17] において，単語難易度と文の長さを指標に，重回帰分析による難易度レベルの判定式を提案した．なお、単語の難易度については，旧日本語能力試験の出題基準を指標としている．日本語能力試験は 2010 年に改定され，出題基準が公開されなくなったため以前の出題基準が使われ

ることが多いことを付記しておく．一方，構文の複雑さについては，名詞修飾節や重文・複文のような構文が文の難易度に影響することが示されているが，どのような形で影響を及ぼしているかについては明確ではない．構文解析技術の精度の問題などもあるため，実験では文の長さのみを難易度の指標としている．

また，李 [18] も同様の目的で，文章の難易度を判定するリーダビリティ公式を作成した．難易度の指標として Hasebe ら [19] が構築した基準コーパスを使用している．基準コーパスとは日本語教科書と現代日本語書き言葉均衡コーパスから収集したテキストデータに 6 段階の難易度レベルを付与したものである．李らの推定式では，文の長さ，漢語率，和語率，動詞率，助詞率から重回帰分析で係数を求め，推定式の精度は 89.6% であった．

劉ら [20] は語彙レベルと構文の複雑さからテキストの難易度を判定する手法を提案した．語彙レベルは川村らと同様，旧日本語能力試験の出題基準をもとにしている．構文の複雑さを表す尺度には係り受け距離を採用し，これが長いほど難易度が高い文としている．実験では日本語能力試験の問題集をコーパスとして利用し，重回帰分析によって難易度算定公式を提案した．推定の精度は，72.2% であった．

張ら [21] は外国人の感覚に合った日本語の難易度を自動推定する手法を提案した．テキストの特徴量を抽出し，日本語学習者が付与した難易度スコアを指標として，特徴量が有効かどうかを調べた．難易度推定に有効な特徴量として，文の長さ，名詞の割合，動詞の数，動詞の割合，文節数，係り受けの距離，係り受けの距離，係り受けの回数，語彙平均レベル，レベル 0 からレベル 4 までの単語数，レベル 0 からレベル 4 までの単語の割合，外来語の数，外来語の割合，漢字の割合，ひらがなの数，ひらがなの割合，カタカナの数，カタカナの割合を挙げた．実験では難易度スコアを付与したのは母国語が漢字圏の出身者であったことから，漢字の含有率が少なければ難易度が高い傾向となり，学習者が漢字圏かどうかで難易度の指標が異なるという課題が明らかになった．

本研究では，これらの先行研究より特徴量の種類を検討する際に参考にした．

2.2.2 機械学習によるテキスト平易化

テキストの意味を保持しながら，わかりやすい表現に書き換えることをテキスト平易化という．テキストの自動生成や機械翻訳の技術が発展したことと言語資源が使えるようになってきたことから，テキスト平易化を自動で行う研究が活発になった．従来は統計的機械翻訳を用いた研究が中心であったが，深層学習の発展によってニューラル機械翻訳に移行している．機械翻訳は日本語と英語のような異なる言語間にも用いられるが，テキスト平易化を同一言語間の翻訳問題として解くこ

とができる．

日本語においては通常のテキストと平易なテキストのパラレルコーパスはほとんど無いため，機械翻訳モデルの学習精度が上がらない問題がある．梶原ら [9] は，大規模なコーパスの中から類似性の高い 2 文を抽出した擬似パラレルコーパスを構築する手法を提案した．この研究では統計的機械翻訳の手法を用いているが，パラレルコーパスを構築するために膨大なコストがかかるという問題を解決している．

機械学習でのテキスト平易化を評価する指標として，パラレルコーパスを利用した BLEU[1] や SARI[2] が代表的である．これらの手法では，通常のテキストを平易なテキストに書き換えたリファレンス文の単語と，モデルの出力がどれだけ一致しているかを測る．テキスト平易化において，パラレルコーパスやリファレンス文が少ないことが問題であるため，本論文ではモデルの出力を直接的に類似度・難易度で評価する手法を提案する．本章では，テキストを難易度で評価する手法について述べる．

2.2.3 分類モデルの解釈性

石井ら [22] は，BERT[15] におけるアテンションに着目して分類器の予測の解釈性を検討した．感情分析タスク用のデータセットで分類モデルを構築し，解釈性の高いアテンションが何層目にあるかを調べたところ，全 12 層のうち 10 から 12 層目のアテンションの解釈性が高いことがわかった．実験で使用したデータセットは，映画レビューにポジティブ，ネガティブのクラス分類と，その根拠となる箇所がアノテートされたもので，ERASER[23] の評価指標とともに公開されているものである．

仮に，テキスト平易化タスク用のデータセットに，難易度の根拠となる箇所がアノテートされたものが存在していたら，アテンションの解釈性を評価できる．現在はそのようなデータセットや指標が存在しないため，本研究では 2.2.1で触れた日本語の難易度に関する特徴量 [18, 20, 21, 24] に基づいて評価する．

2.3 本実験で用いる既存手法

2.3.1 BERT

BERT(Bidirectional Encoder Pepresentatoins from Transformers)[15] は，2018 年に Google の研究チームが発表した自然言語処理モデルである．質問応答，機械翻訳，文書分類などのさまざまなタスクで当時の最高スコアを達成した．自然言語処理は，単語を高次元のベクトルで表現し，

文章は単語データの並び（シーケンス）として扱う．BERT は事前学習と、ファインチューニングの 2 段階で構成される。BERT の事前学習に用いるデータは教師ラベルがついていない文章を用いるため，Wikipedia や新聞のデータなど比較的入手しやすいデータから単語のベクトル表現を得ることができる．また，モデルを少量の教師ありデータでファインチューニングすることでさまざまなタスクに適応することができる．

図 2.1に BERT の構造を示す．アーキテクチャの規模の違いによって $\mathrm{BERT_{BASE}}$ と $\mathrm{BERT_{LARGE}}$ に分類されるが，本研究では $\mathrm{BERT_{BASE}}$ を用いる．$\mathrm{BERT_{BASE}}$(以下，BERT と称する) は Transformer ブロックを 12 層もつ構造で，事前学習とファインチューニングを行う．

■事前学習　事前学習では，Musked Language Model という手法で、BERT へ入力する単語の一部をランダムにマスクし，マスクされた単語を予測する．シーケンス全体の 15% をマスクするが，そのうち 80% を [MASK] トークンに置き換え，10% をランダムな単語に置き換え，10% を置き換えずそのままとする。こうすることで，単語の意味を予測してベクトルを生成する際に，複数の意味を持つ単語を別のベクトルで表現することができる．

また，Next Sentence Prediction という手法では，次の文章を予測するタスクを事前学習で行う．2 文を [SEP] トークンで区切って入力し，それらが隣り合った文であれば [CLS] トークンに isNext を，隣り合っていない文であれば [CLS] トークンに notNext を出力する．こうすることで，2 つの文の関係性を学習する．

つまり，BERT は事前学習の過程で単語の穴埋め問題として前後の文脈を考慮した単語ベクトル表現を生成することと，2 つの文章が前後関係にあるかどうかを判断する能力を得ていることになる．

■ファインチューニング　BERT を既存のタスク処理モデルに接続し，ファインチューニングする．質問回答，同意文判定，感情分析，意味的類似性，含意関係など，さまざまなタスクにモデルをチューニングする．

■アテンション　初めに Transformer の構造について説明し，後に本論文で着目するアテンションについて述べる．

BERT は Transformer のエンコーダのみを使用する．Transformer のエンコーダの構造を図 2.2に示す．Transformer に入力する文は単語に分割してそれぞれ単語ベクトルに変換し，Input Embedding とする．Transformer は RNN や LSTM のような再帰構造ではないため，Positional Encoding で単語の位置情報を付与する．次の Multi-Head Attention ブロックでは入力単語列と

出力単語列の内積とその重みを計算する (計算の詳細は後述の Self-Attention で述べる)．関係性が強いほど重みは大きくなる．この処理は複数の Head に分割し並列で計算するため，次の Add & Norm ブロックでは，Multi-Head の計算結果を結合し，正則化処理を行う．Feed Forward ブロックは全結合層と ReLU 活性化関数を適用し，正則化処理を行う．これらの処理を通常 6 回繰り返す．

■Self-Attention　Transformer のアテンションは，同じ入力データ内で単語間の類似度や重要度を獲得する Self-Attention を用いる．従来の Attention は，例えば機械翻訳タスクにおいて同じ意味を示す異なる言語表現のシーケンスデータが与えられたとすると，異なるデータ間の類似度や照応関係を獲得するものであった．Self-Attention は，同じ入力データ間の類似度や照応関係を獲得する．同一文章内の類似度が獲得できることで，特に多義語や代名詞などが他のどの単語を示すのかを得られるようになった．

Transformer の Self-Attention の計算について説明する．アテンションは Query と Key-Value を出力にマッピングするもので，Query, Key, Value は全てベクトルである．Query, Key, Value は全て同じ入力データから生成する．Query と Key は次元 d_k を持ち，Value は次元 d_v を持つ．出力は Value の加重和として計算され，各 Value の加重和は対応する Key と Query の内積で計算する．内積によって計算する場合，d_k の値が大きいと内積が大きくなり，逆伝搬の softmax 関数が極端に小さくなりすぎてしまうことから，$\sqrt{d_k}$ で除算している ((2.1) 式)．

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2.1}$$

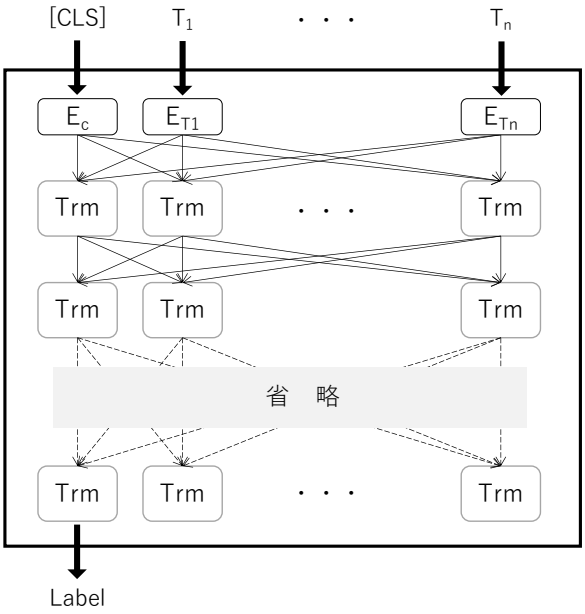


図 2.1　BERT の構造

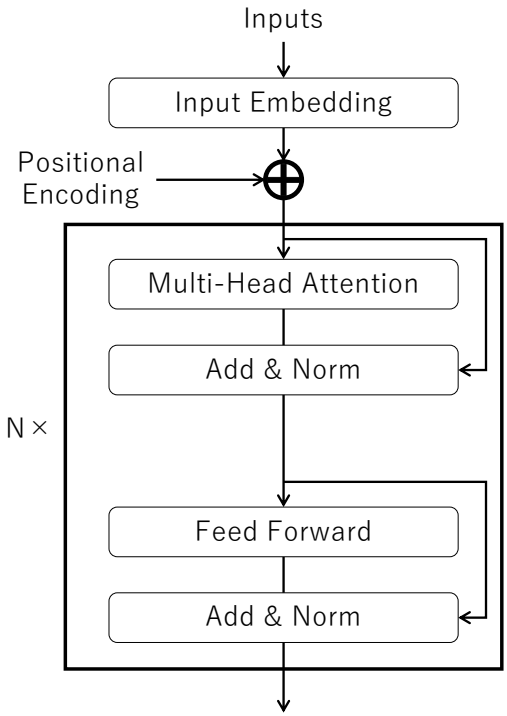


図 2.2　Transformer の構造

2.3.2　ランダムフォレスト

ランダムフォレストは複数の決定木で学習するため，はじめに決定木学習について触れておく．決定木はデータからツリー構造を作る分類モデルで，葉が分類を表し，枝がその分類に至るまでの条件を表す．葉の不純度が最も小さくなるように条件分岐を繰り返す．不純度とはクラス分類する際にどれだけ他のクラスのデータが混ざっているかを示す指標で，一般的にはジニ係数やエントロピーが使用される．例えば，「単語数は 20 より大きいかどうか」という条件（枝）で分岐したクラス（葉）が，難易度で分類できていれば精度が良いといえる．

ランダムフォレストは複数の決定木を使い，各予測の多数決を最終決定とする手法である．ランダムフォレストの手順を以下に示す．文中の下線部は，図 2.3に一致する．

(a) 全ての学習データからそれぞれの決定木を学習するための(b) 抽出データをランダムに選択する．それぞれの決定木を学習し，出力を決定する．学習後のランダムフォレストを分類モデルに利用する場合は，それぞれの決定木の出力から(c) 最も多い出力をランダムフォレストの出力とする．（回帰の場合は平均を取る．）

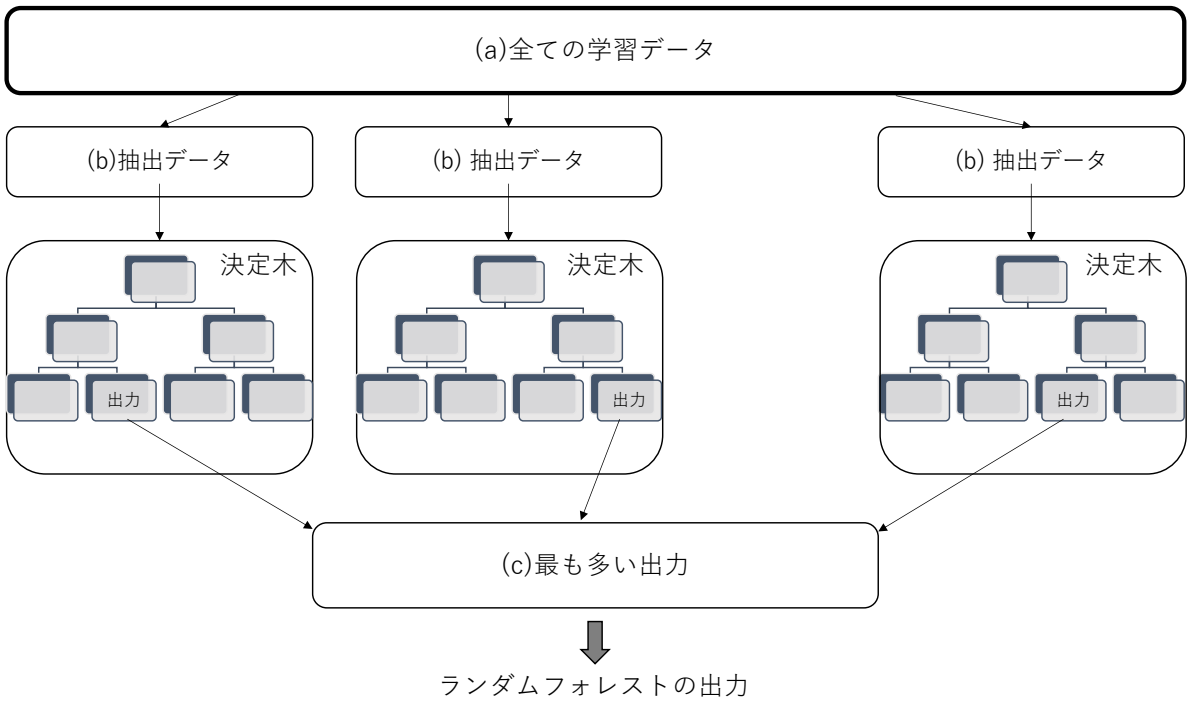


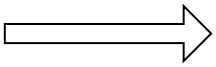
図 2.3　ランダムフォレストの仕組み

2.3.3 Permutation Importance

Permutation Importance[25] は機械学習モデルの特徴の重要性を図る手法の一つで，各特徴量がモデルの予測精度に貢献しているほどその特徴量の重要度が高いと判断する．事前にモデルを学習済みとし，正解率を求める．次に，1 つの特徴量についてデータをランダムに並べ替える．図 2.4は, 特徴量 B を並べ替えた例である．それ以外は全く同じ条件で学習し, 正解率を求める．全ての特徴量について 1 つずつ並べ替えて正解率を求め，正解率がどれくらい変化したかを比較する．

ランダムに並べ替えた特徴量はラベルを予測するための説明変数として機能しない．重要な特徴量である場合，その特徴量をランダムに並べ替えて学習すると正解率が低くなる．

特徴量A	特徴量B	特徴量C
3	500	TRUE
4	1000	TRUE
3	1200	FALSE
3	700	TRUE
4	550	FALSE
6	1100	FALSE
2	400	TRUE
5	1000	TRUE
3	1500	FALSE
4	900	FALSE



1つの特徴量を
ランダムに並べ替える

特徴量A	特徴量B	特徴量C
3	1200	TRUE
4	1500	TRUE
3	550	FALSE
3	1000	TRUE
4	500	FALSE
6	700	FALSE
2	1000	TRUE
5	400	TRUE
3	1100	FALSE
4	900	FALSE

図 2.4　Permutation Importance の仕組み

計算の手順を以下に示す．

1. データセット D を学習済みモデルで分類したときの正解率を s とする．
2. 特徴量 j ごとに以下を計算する．
   - 特徴量 j をランダムに並べ替えて，データセット D_j を生成する．
   - データセット D_j を，学習済みモデルで分類したときの正解率を s_j とする．
3. 特徴量が K 個あるとしたときの特徴量 j の重要度 PI_j を次の計算式で定義する．

$$PI_j = s - s_j (1 \leq j \leq K)$$

2.4　実験に使用するデータ

実験では，平易なテキストとして NEWS WEB EASY の記事から，通常のテキストとして NHK NEWS WEB の記事からテキストを抽出した．NEWS WEB EASY は外国人や子ども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．田中ら [26] によると，テキストの書き換えは日本語教師と記者が共同で行っている．日本語教師が日本語学習者の初中級者向けの語彙や文法に書き換えていることから，これらは日本語学習者向けのやさしい日本語として適していると判断した．

データを収集した期間は 2020 年 7 月から 2021 年 7 月までで，722 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．リンクがない記事は実験データの対象外とした．

以下に，Web ページのニュースをテキストデータに加工する手順を述べる．

スクレイピング

データ収集の対象である Web ページは，記事ごとにページが構成されており，それぞれのページに異なる URL が割り当てられている．データを効率よく収集するため NEWS WEB EASY の URL の一覧を作成し，HTTP リクエストを送信する．また，NEWS WEB EASY のページには「普通のニュースを読む」というリンクがあり，書き換え前の記事の URL を取得できる．書き換え前の記事の URL に対しても HTTP リクエストを送信する．

本文のみ抽出

HTTP レスポンスで得たデータは，本来ブラウザに表示するすべての情報を含んでいる．実験で使用するデータは本文のみであるため，HTML を解析して該当部分を抽出する．本文は HTML タグ内の class 名および id で識別している．

タグ・ルビの削除

スクレイピングによって取得したテキストには，図 2.5に示すように HTML タグ等が含まれる．HTML タグは不要でテキストのみを抽出するため，タグを除去する．また，NEWS WEB EASY には Web ブラウザで表示するための情報として漢字にルビが振られているが，解析に不要であるためルビを除去する．抽出したテキストは図 2.6のようになる．

```
<p><a href='javascript:void(0)' class='dicWin' id='id-0000'><ruby><span class="under">製品</span><rt>せいひん</rt></ruby></a>の<ruby>安全<rt>あんぜん</rt></ruby>などを<ruby>調<rt>しら</rt></ruby>べている<span class='colorC'>NITE</span>によると、ストーブや<a href='javascript:void(0)' class='dicWin' id='id-0001'><span class="under">ヒーター</span></a>で<ruby>火事<rt>かじ</rt></ruby>などになった<ruby>事故<rt>じこ</rt></ruby>が、<ruby>今年<rt>ことし</rt></ruby>3<ruby>月<rt>がつ</rt></ruby>までの5<ruby>年<rt>ねん</rt></ruby>に652ありました。77<ruby>人<rt>にん</rt></ruby><ruby>亡<rt>な</rt></ruby>くなりました。<ruby>事故<rt>じこ</rt></ruby>は<ruby>毎年<rt>まいとし</rt></ruby>11<ruby>月<rt>がつ</rt></ruby>ごろから<ruby>増<rt>ふ</rt></ruby>えて、1<ruby>月<rt>がつ</rt></ruby>にいちばん<ruby>多<rt>おお</rt></ruby>くなります。</p>
```

図 2.5　HTML レスポンスで取得するデータの形式

```
製品の安全などを調べているNITEによると、ストーブやヒーターで火事などになった事故が、今年3月までの5年に652ありました。77人亡くなりました。事故は毎年11月ごろから増えて、1月にいちばん多くなります。
```

図 2.6　HTML タグとルビを削除したテキスト

2.5　実験 1：BERT を用いたテキストのみによる難易度分類

2.5.1　概要

BERT は Transformer の構造を持ち，大規模コーパスを使用して事前学習されたモデルである．図 2.1で示した通り，BERT の構造は Transformer のエンコーダ部分のみを用いたモデルで，テキストを入力してラベルを出力する．本実験では，平易なテキストもしくは通常のテキストを入力すると，難易度を二値分類で出力するモデルを構築する．この実験ではモデルの精度だけでなく，Transformer の計算過程を分析することでどの単語に注目して難易度を推定しているのかを明らかにする．

2.5.2　実験に使用する計算機環境

実験に使用する計算機環境は以下の通りである．Web ニュースから収集したデータに難易度ラベルを付与し，BERT で難易度を推定する．データ収集と前処理については 2.4を参照のこと．

- CPU：Intel(R) Core(TM) i9-9900K CPU @ 3.60GHz
- GPU：GeForce RTX 2080 Ti
- メモリ：64GB
- OS：Ubuntu 18.04.5 LTS

2.5.3　手順

実験の手順は以下の通りである．

(1) テキストに難易度ラベルを付与

表 2.1に示すように，テキストを文に分割し，それぞれに難易度ラベルを付与する．難易度ラベルは，平易なテキストは「1」，通常のテキストは「0」とする．学習データとして 10,000 文（平易なテキスト：5,000 文，通常のテキスト：5,000 文），テストデータとして 1,200 文（平易なテキスト：600 文，通常のテキスト：600 文）を選択する．

表 2.1　難易度ラベルを付与したデータの例

例文	ラベル
通過した電車の車内では、事故現場にさしかかる前に、「本日で福知山線列車事故から 16 年を迎えます。	0
特に６０歳以上は、家からあまり出なくなって、人と話す時間が少なくなっています。	1
今後の対応については関係省庁と協議、検討をしていきたい」とするコメントを出しました。	0
１９日は、小学校の子どもたち５０人ぐらいと耳に障害がある子どもたち１０人ぐらいが集まりました。	1
ホーチミン市の幹部は「都市鉄道はホーチミンと日本をつなぐ懸け橋となる」と期待を述べました。	0
その結果、歳入全体のうち国債で賄う割合は 40.9 %と、当初予算の段階としては 7 年ぶりに 40 %を超え、財政状況は一段と深刻化しています。	0
男性の家族は「トヨタは会社の人が働きやすくなるように、真剣に考えてください。	1
外国から日本に来て、働きながら技術を習う技能実習生の中には、新しいコロナウイルスの問題で仕事をやめさせられた人がいます。	1
それだけ奄美大島には大変な宝物があるのだということを島民がいちばん気付かされたのではないか。	0
有料化前のことし 3 月に行った同様の調査に基づく推計では、レジ袋をもらわなかった人は 30.4 %だったということで、有料化のあと 2 倍以上に増えています。	0

(2) 日本語の事前学習モデルを取得

日本語 BERT 事前学習モデルは，東北大学知能情報科学講座の公開リソース[*1]を利用する．このモデルは日本語 Wikipedia をコーパスとして，事前学習したモデルである．

(3) データの前処理

BERT 用にデータを成形する．まず BERT Tokenizer で用いて単語をトークンへと分割し，ID へ変換する．文の先頭には [CLS] という特殊トークンを付与し，これは分類問題のラベルに利用される．文の最後に [SEP] という特殊トークンを追加し，文の区切りを示す．最後に入力シーケンスの長さを揃えるため，トークン数が指定の数に満たない場合 [PAD] という特殊トークンで埋める．

(4) ファインチューニング

(3) データの前処理を施したデータをモデルへ入力する．図 2.7に示すように，Transformer の最終層から出力された [CLS] を分類器へ入力する．BERT モデルのパラメータは計算コストを考

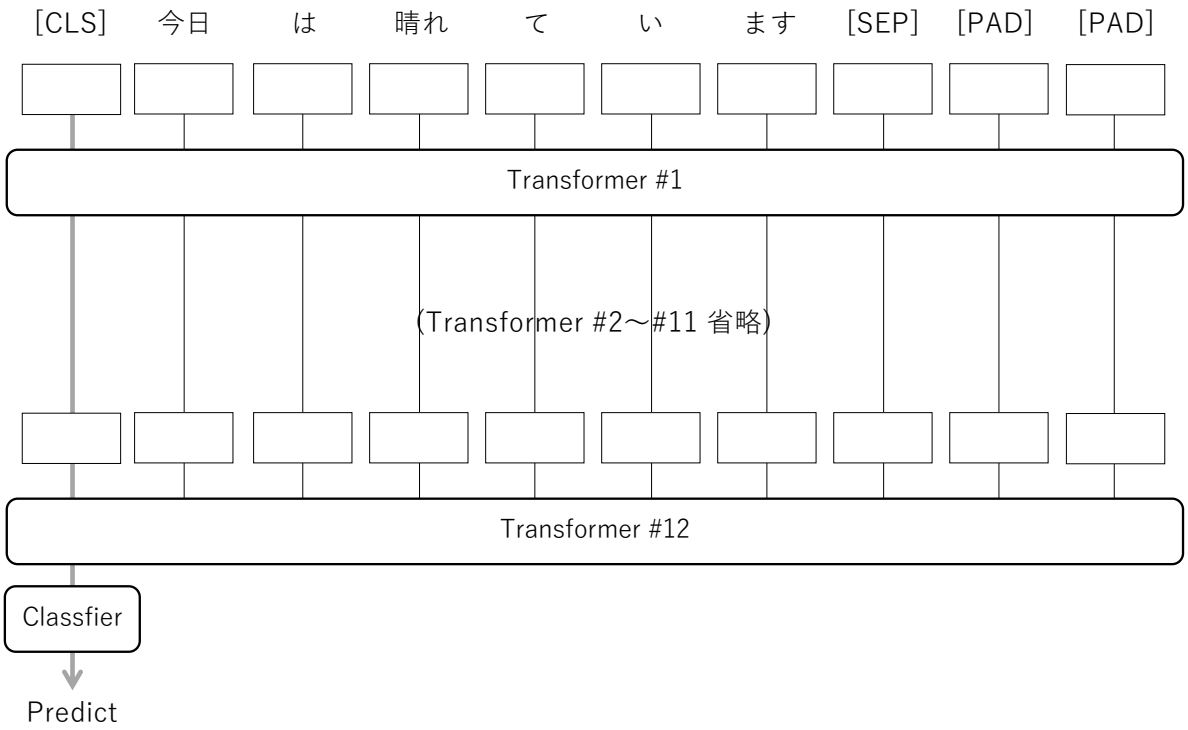


図 2.7　BERT へ入力するデータ形式

[*1] https://nlp.ist.i.kyoto-u.ac.jp/?ku_bert_japanese

慮して，入力の最大系列長を 256，ミニバッチサイズを 32，イテレーション数を 8，エポック数を 5 とした．

(5) アテンションを可視化

BERT へ入力したテキストのアテンションを HTML 形式で出力して可視化する．CSS の background-color をアテンションの重みに応じて計算し，アテンションの重みが強い単語ほどテキストの背景色が濃い赤で表現する．アテンションの重みを $attn$ とすると，RGB カラーは $(255, (1 - attn), (1 - attn))$ とする．アテンションの大きさは最大値によって規格化し，1 層目から 12 層目のそれぞれと，全ての層の平均を出力する．

2.6 実験 2：難易度分類に影響を及ぼす文法的特徴の抽出

2.6.1 概要

日本語のテキストから難易度に影響を与える文法的な特徴量を分析し，分類モデルを構築する．特徴量の選別は先行研究 [24] を参考とする．その後，Permutation Importance で特徴量の重要度を測り，難易度推定に重要な特徴量を明らかにする．

2.6.2 手順

実験の手順は以下の通りである．

(1) テキストの前処理

テキストから特徴量を抽出するために，形態素解析と係り受け解析を必要とする．ここでは形態素解析，および係り受け解析に使用するツールと特徴量の求め方を説明する．

形態素解析

形態素解析は，文法や単語の品詞情報に基づいて言語として意味を持つ最小の単位に分割し，それぞれの品詞や活用形などを識別することである．形態素解析にはオープンソースの MeCab を使用した．

次のテキストについて，MeCab で形態素解析した結果を図 2.8に示す．

新しく売る車は全部、電気を使う車にします。

MeCab にテキストを入力すると，最小単位の形態素と，その形態素の解析結果が表示される．

新しく売る車は全部、電気を使う車にします。
新しく　形容詞,自立,*,*,形容詞・イ段,連用テ接続,新しい,アタラシク,アタラシク
売る　動詞,自立,*,*,五段・ラ行,基本形,売る,ウル,ウル
車　名詞,一般,*,*,*,*,車,クルマ,クルマ
は　助詞,係助詞,*,*,*,*,は,ハ,ワ
全部　名詞,副詞可能,*,*,*,*,全部,ゼンブ,ゼンブ
、　記号,読点,*,*,*,*,、,、,、
電気　名詞,一般,*,*,*,*,電気,デンキ,デンキ
を　助詞,格助詞,一般,*,*,*,を,ヲ,ヲ
使う　動詞,自立,*,*,五段・ワ行促音便,基本形,使う,ツカウ,ツカウ
車　名詞,一般,*,*,*,*,車,クルマ,クルマ
に　助詞,格助詞,一般,*,*,*,に,ニ,ニ
し　動詞,自立,*,*,サ変・スル,連用形,する,シ,シ
ます　助動詞,*,*,*,特殊・マス,基本形,ます,マス,マス
。　記号,句点,*,*,*,*,。,。,。

図 2.8　MeCab による形態素解析の結果

ここで，出力形式を説明する．図 2.9に形態素「売る」の結果を示す．一つの形態素に対して，表層形，品詞，品詞細分類 1〜3，活用形，活用型，原形，読み，発音の解析結果が出力される．本研究の特徴量の抽出には，品詞として解析結果の「品詞」を参照する．活用形として「品詞細分類 1 (大分類)」を参照する．単語の頻度として「原形」の形態素を数える．

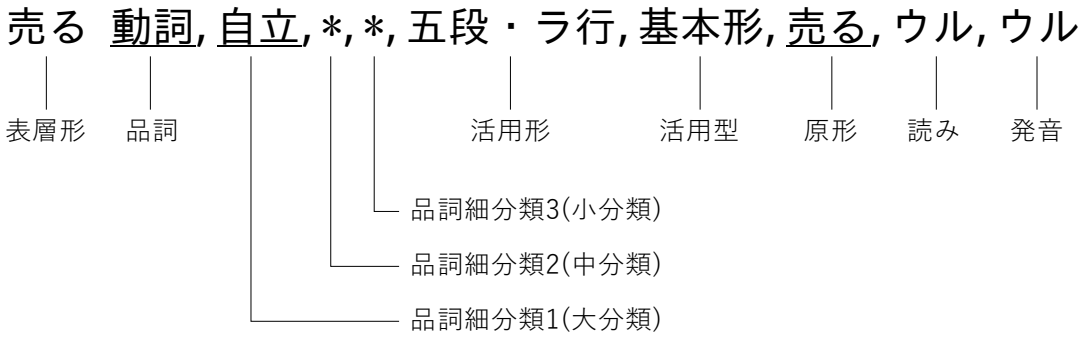


図 2.9　MeCab による形態素解析の意味

係り受け構造解析

先行研究 [20, 21] でも特徴量として挙げられていたのが，係り受け構造に関する特徴量である．係り受けとは，1 つの文章を文節にわけて文節間の「修飾する（係る)」「修飾される（受ける)」関係を示すことである．文節間の関係には，主語と述語，目的語と述語，修飾語と被修飾語がある．

次のテキストについて，係り受け構造を調べる．

来年の東京オリンピックとパラリンピックでは、試合の会場で８万人のボランティアが手伝う予定です。

この文を日本語係り受け解析器である CaboCha で解析し，係り受け関係を視覚的にわかりやすく出力した結果を図 2.10に，詳細な結果を図 2.11に示す．図 2.10より，「会場で」と「手伝う」，及び「パラリンピックでは」と「予定です。」に係り受け関係があり，後者の方が係り受け距離が長いことがわかる．

```
来年の東京オリンピックとパラリンピックでは、試合の会場で８万人のボランティアが手伝う予定です。
          来年の-D
  東京オリンピックと-D
  パラリンピックでは、----------D
                試合の-D       |
                  会場で-----D |
                  ８万人の-D | |
              ボランティアが-D |
                        手伝う-D
                      予定です。
```

図 2.10　CaboCha による係り受け構造解析の結果 (tree)

```
来年の東京オリンピックとパラリンピックでは、試合の会場で８万人のボランティアが手伝う予定です。
* 0 1D 0/1 0.047905
来年 名詞,副詞可能,*,*,*,*,来年,ライネン,ライネン
の 助詞,連体化,*,*,*,*,の,ノ,ノ
* 1 2D 1/2 1.201955
東京 名詞,固有名詞,地域,一般,*,*,東京,トウキョウ,トーキョー
オリンピック 名詞,一般,*,*,*,*,オリンピック,オリンピック,オリンピック
と 助詞,並立助詞,*,*,*,*,と,ト,ト
* 2 8D 0/2 -0.598596
パラリンピック 名詞,一般,*,*,*,*,パラリンピック,パラリンピック,パラリンピック
で 助詞,格助詞,一般,*,*,*,で,デ,デ
は 助詞,係助詞,*,*,*,*,は,ハ,ワ
、 記号,読点,*,*,*,*,、,、,、
* 3 4D 0/1 2.213136
試合 名詞,サ変接続,*,*,*,*,試合,シアイ,シアイ
の 助詞,連体化,*,*,*,*,の,ノ,ノ
* 4 7D 0/1 2.252695
会場 名詞,一般,*,*,*,*,会場,カイジョウ,カイジョー
で 助詞,格助詞,一般,*,*,*,で,デ,デ
* 5 6D 2/3 2.007041
８ 名詞,数,*,*,*,*,８,ハチ,ハチ
万 名詞,数,*,*,*,*,万,マン,マン
人 名詞,接尾,助数詞,*,*,*,人,ニン,ニン
の 助詞,連体化,*,*,*,*,の,ノ,ノ
* 6 7D 0/1 2.622306
ボランティア 名詞,一般,*,*,*,*,ボランティア,ボランティア,ボランティア
が 助詞,格助詞,一般,*,*,*,が,ガ,ガ
* 7 8D 0/0 -0.598596
手伝う 動詞,自立,*,*,五段・ワ行促音便,基本形,手伝う,テツダウ,テツダウ
* 8 -1D 0/1 0.000000
予定 名詞,サ変接続,*,*,*,*,予定,ヨテイ,ヨテイ
です 助動詞,*,*,*,特殊・デス,基本形,です,デス,デス
。 記号,句点,*,*,*,*,。,。,。
```

図 2.11　CaboCha による係り受け構造解析の結果 (lattice)

係り受けの距離を求めるには，図 2.11の出力結果から計算する必要がある．図 2.11の出力形式について説明する．

1 行目は解析するテキストである．2 行目「* 0 1D 0/1 0.047905」の解析結果は図 2.12の通りである．

3 行目，4 行目は文節に含まれる形態素の解析結果で，出力形式は MeCab と同様である．本研究では，係り受け最大距離，係り受け平均距離，係り受け被修飾数を特徴量とする．

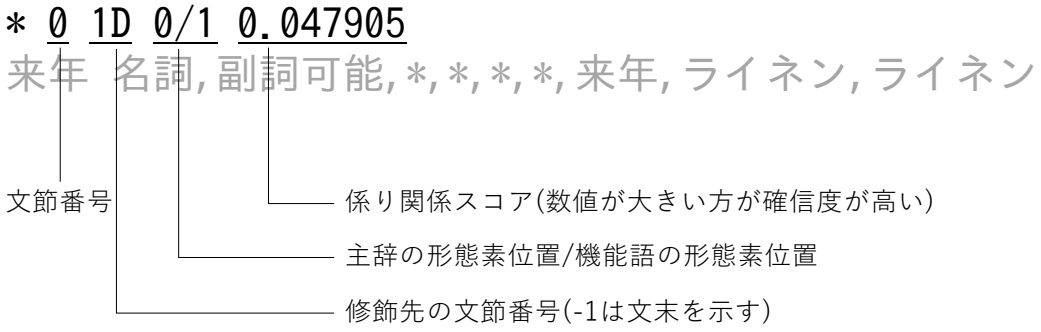


図 2.12　CaboCha による係り受け解析の意味

(2) 特徴量の計算

先行研究 [24] に基づき，テキストから特徴量を抽出する．特徴量は文ごとに求める．算出方法を以下に述べる．

単語数

テキストを MeCab で形態素に分割する．テキストに含まれる形態素の数を求める．形態素に分割する例を以下に示す．

> 来年/の/東京/オリンピック/と/パラリンピック/で/は/、/試合/の/会場/で/８/万/人/の/ボランティア/が/手伝う/予定/です/。

この場合，形態素数は 23 である．

$$\text{単語数} = \text{形態素数}$$

漢字率

テキストに含まれる漢字の数を文字数で割った値を求める．

$$\text{漢字率} = \frac{\text{漢字の個数}}{\text{文字数}}$$

外来語率

テキストを MeCab で形態素に分割し，全ての文字がカタカナである形態素の数を数える．これを単語数で割った値を求める．

$$\text{外来語率} = \frac{\text{全てカタカナの形態素の個数}}{\text{単語数}}$$

受身率

テキストを MeCab で形態素に分割し，接尾語の「れる」もしくは「られる」の数を単語数で割った値を求める．

$$\text{受身率} = \frac{\text{受身形である形態素の個数}}{\text{単語数}}$$

サ変接続名詞率

テキストを MeCab で形態素に分割し，品詞が「名詞」，かつ品詞細分類 1 が「サ変接続」と一致する形態素の数を単語数で割った値を求める．

$$\text{サ変接続名詞率} = \frac{\text{サ変接続名詞の個数}}{\text{単語数}}$$

副詞率

テキストを MeCab で形態素に分割し，品詞が「副詞」である形態素の数を単語数で割った値を求める．

$$\text{副詞率} = \frac{\text{品詞が副詞である形態素の個数}}{\text{単語数}}$$

読点率

テキストを MeCab で形態素に分割し，「、」の個数を単語数で割った値を求める．

$$\text{読点率} = \frac{\text{読点の数}}{\text{単語数}}$$

否定率

品詞が「助動詞」で，「ない」「ぬ」「ん」と一致する形態素の個数を単語数で割った値を求める．

$$\text{否定率} = \frac{\text{助動詞が否定形である形態素の個数}}{\text{単語数}}$$

出現頻度最大値

単語の出現頻度が頻繁であるほど，単語の難易度は低くなる傾向があると考え，単語の出現頻度を特徴量のひとつとして用いる．ただし，助詞の「は」「が」や助動詞の「です」「ます」等は難易度に関わらず頻繁に使用されるため，その形態素がテキストに含まれていると出現頻度最大値が同じ値になってしまう．そのため品詞が「名詞」「動詞」「形容動詞」「形容詞」である形態素の数のうちの最大値を求める．

出現頻度平均値

テキストを形態素に分割し，品詞が「名詞」「動詞」「形容動詞」「形容詞」である形態素の数の平均値を求める．

係り受け平均距離

テキストを CaboCha で係り受け解析する．テキストの先頭から文節ごとに文節番号と修飾先の文節番号との距離を求め，文節の数で割った値を求める．

ここで，係り受けに関する特徴量の求め方を示す．例文を係り受け構文解析した結果を表 2.2に示す．特徴量の分析に使用しないパラメータは表から除外した．

表 2.2　係り受け解析結果の一部

文節	文節番号	修飾先の文節番号	係り受け距離
来年の	0	1	1
東京オリンピックと	1	2	1
パラリンピックでは,	2	8	6
試合の	3	4	1
会場で	4	7	3
8 万人の	5	6	1
ボランティアが	6	7	1
手伝う	7	8	1
予定です.	8	-1	なし

係り受け距離は，修飾先の文節番号 − 文節番号 で求める．ただし，最後の文節は「修飾先の文節番号」が −1 であることから，距離は求めない．

係り受け平均距離は，係り受け距離の総和を文節の数で割って求める．表 2.2の場合，係り受け平均距離は $(1 + 1 + 6 + 1 + 3 + 1 + 1 + 1)/8 = 1.875$ である．

係り受け最大距離

テキストを CaboCha で係り受け解析する．文節単位での文節番号と修飾先の文節番号との距離を求め，最大値を求める．表 2.2の場合，係り受け最大距離は 6 である．

係り受け被修飾数

テキストを CaboCha で係り受け解析する．修飾先の文節番号を数え, 最大値を求める．表 2.2の場合，7 と 8 がそれぞれ 2 回ずつ修飾先に出現しているため，係り受け最大被修飾数は 2 である．

文ごとに特徴量を求め，1 つの特徴量ベクトルとする．特徴量ベクトルに難易度ラベルを付与する．データのサンプルを表 2.3に示す．

表 2.3　特徴量ベクトルのデータの例

単語数	漢字率	外来語率	受身率	サ変接続名詞率	副詞率	読点率	否定率	出現頻度 最大値	出現頻度 平均値	係り受け 平均距離	係り受け 最大距離	係り受け 被修飾数	ラベル
34	0.2295	0.0588	0	0.059	0	0.0588	0	25546	5175.375	2.8182	8	4	0
57	0.4314	0.0175	0	0.018	0	0.0526	0	34705	1611.1667	2.75	17	4	0
30	0.4035	0.0333	0	0.067	0.0333	0.1	0.0333	18821	1419.9333	2.8333	12	5	0
42	0.4359	0	0	0.095	0	0.0714	0	25546	2523.9048	2.7059	17	3	0
66	0.3636	0	0.0303	0.046	0	0.0606	0.0303	18821	1809.6667	2.6897	29	5	0
26	0.303	0	0	0.039	0.0385	0.0769	0	18821	4278.6364	2	7	3	1
16	0.2069	0	0	0.000	0	0	0.0625	11946	3456.375	1.25	2	2	1
16	0.2333	0.125	0	0.000	0.0625	0	0	34705	5145.1429	1.5	4	2	1
35	0.303	0.0571	0	0.000	0.0286	0.0857	0	1015	239.6923	2.8462	13	5	1
26	0.2292	0.0769	0	0.077	0	0.0385	0	18821	5306.9286	2	9	2	1

(3) ランダムフォレストによる学習

(2) の特徴量を用いてテキスト難易度（つまり表 2.3のラベル）を推定するモデルを構築する．分類モデルはランダムフォレストを使用する．ランダムフォレストは Python のパッケージである Scikit-learn を使用した．パラメータの設定は次の通りである．

- ツリーの数：100
- ツリーの深さ：最大 5
- ノードの最大値：制限なし

(4) 特徴量の重要度

2.3.3に示す Permutation Importance の手法で、(2) の特徴量のうち，難易度推定の精度に重要な特徴量を調べる．

2.7　結果

2.7.1　実験 1：BERT を用いたテキストのみによる難易度分類

BERT でテキストのみを使って難易度推定した結果，学習データの正解率は 95.7%，テストデータの正解率は 96.4% で，非常に良い結果が得られた．アテンションを可視化した結果は，実験 2 の結果をふまえて 2.7.3で示す．

2.7.2　実験 2：難易度分類に影響を及ぼす文法的特徴の抽出

ランダムフォレストで難易度推定した結果，正解率は 82.6% であった．次に，Permutation Importance でどの特徴量がモデルの正解率に影響を与えているかを調べた結果を図 2.13に示す．「サ変接続名詞率」，「単語数」，「受身率」，「漢字率」の 4 つの特徴量において，その特徴量をランダムに並べ替えると分類の精度が下がった．つまり，これらの特徴量が重要であることが分かる．これ以外の特徴量は，並べ替えても分類の精度が 0.01 しか変わらないことから，分類にはほとんど影響を与えていないと言える．

表 2.4に先行研究の推定精度と本研究による実験の推定精度を示す．推定精度を比較すると，実験 1 の BERT による難易度分類が最も精度が高いことがわかる．

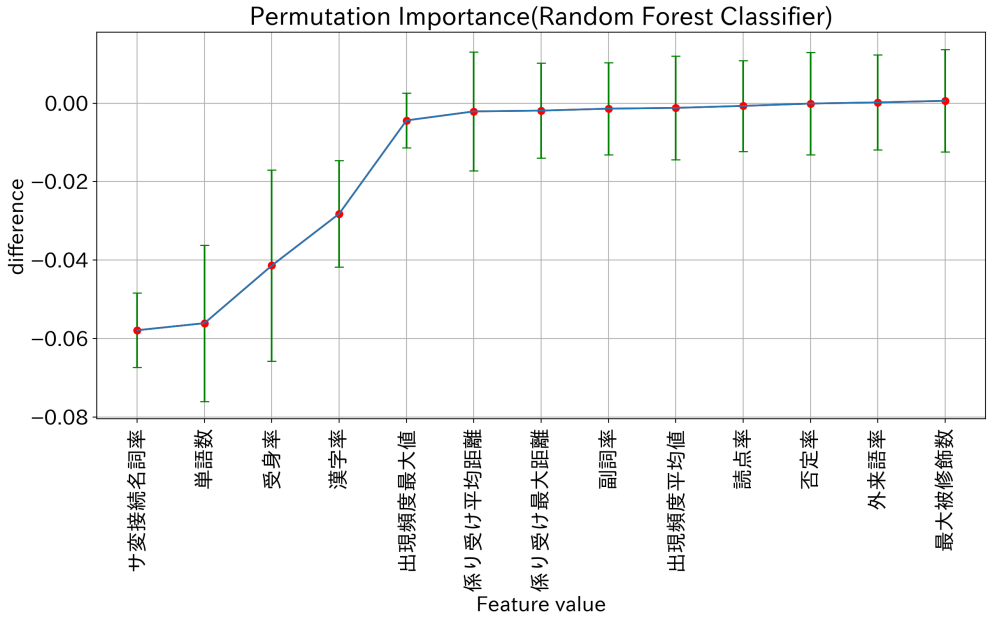


図 2.13　Permutation Importance で測定した特徴量の重要度

表 2.4　難易度推定精度の比較

先行研究	推定精度	手法
李 [18]	89.6%	重回帰分析
劉ら [20]	72.2%	重回帰分析
実験 1	**96.4%**	BERT
実験 2	82.6%	ランダムフォレスト

2.7.3　BERT モデルの文法的解釈に関する考察

実験 2 では難易度に影響を及ぼす文法的特徴が明らかになった．これらの特徴量が実験 1 の BERT モデルで解釈されているかを調べるために，Transformer のアテンションを可視化して考察する．

図 2.14から図 2.18までは，BERT によってテキストを難易度推定したときの，アテンションの重みを可視化したものである．アテンションが重いほど赤色が濃く表現され，推論に影響を与えた単語であるといえる．

図 2.14は，あるテキストをファインチューニング済みの BERT モデルで難易度推定したときの，アテンションを可視化したものである．BERT の構造上，Transformer を 12 層接続しており，上から 1 層目，2 層目，・・・，12 層目と順に示す．最後の「BERT の Attention を可視化_ALL」には全ての層のアテンションの総和を計算したものである．文中の [UNK] トークンは，BERT の事前学習でデータに含まれていなかった単語で，未知語として扱われる．未知語はベクトル表現が得られないため特殊なトークンで置き換える．また，アテンション可視化の際には単語ベクトルから単語に戻す処理を行うが，BERT に入力するテキストデータの前処理において活用形を原形にするため，単語の一部が原形のまま表現されている．

図 2.14の文中において，サ変接続名詞は「入院」および「活動」が該当する．全ての層のアテンションの総和を計算した「BERT の Attention を可視化_ALL」を見ると、「入院」も「活動」もアテンションが強いことがわかる．各層に着目すると，第 8 層目から第 10 層目のアテンションが特に強いことがみられる．BERT の難易度分類モデルは，サ変接続名詞に注目していると言える．図 2.16は BERT の第 11 層のアテンションの重みを可視化したもので，下線部は文末の単語を示す．第 1 層から第 12 層までのアテンションにはそれぞれ特徴があり，第 11 層のアテンションは最後の単語で強くなる傾向がある．このことは，アテンションがテキストに含まれる単語の数に注目している可能性を示唆している．図 2.15は，受身形を含むテキストのアテンションの重みを可視

化したもので、第 1 層目から第 12 層目までと全ての総和を示す．第 6 層目の受身「られる」「れる」に強いアテンションが見られる．他のテキストを難易度推定した結果も見てみる．図 2.17は他のテキストにおける第 6 層のアテンションの重みを可視化したもので，下線部は受身形を示す．受身形の単語はすべてアテンションの重みが強くなっている．アテンションは第 6 層目において受身形を捉えているといえる．図 2.18は 12 層の全てのアテンションの重みを累積したもので，下線部は漢字を示す．漢字は 1 文字単位であるが，BERT へ入力する際に単語に分割して単語埋め込み表現に変換するため，トークンの単位と一致しない．そのため，アテンションが漢字に注目しているかどうかを正確に調べることはできない．しかし，漢字を多く含むトークンに対してアテンションの重みが強くなっている傾向が見られた．

2.8 まとめ

本章では，テキストの難易度推定モデルを構築することと，深層学習による難易度推定モデルの説明可能性を調査することを目的として 2 つの実験を実施した．まず 1 つめの実験では，BERT による難易度分類モデルを構築し，先行研究と比較しても推定精度が高い結果を得ることができた．2 つめの実験では，先行研究で明らかになったテキストの文法的特徴に基づいて，文ごとに特徴量ベクトルを生成し，ランダムフォレストで難易度推定するモデルを構築した．その後，Permutation Importance で難易度推定に影響を及ぼす重要な特徴量を調べたところ，サ変接続名詞率，単語数，受身率，漢字率の 4 つの特徴量が該当した．ここで，一つめの実験で使用した BERT モデルの計算過程を可視化したところ，重要な特徴量に着目して推定を行なっている可能性を示唆することができた．また，アテンションの層によって解釈の違いがあることが明らかになった．

正解ラベル：Positive
推論ラベル：Positive

[BERTのAttentionを可視化_1]
[CLS] [UNK] 「 [UNK] 」 は [UNK] 病院 に 入院 する て いる 人 の ストレス や 不安 を 少ない する ため に 活動 する 犬 です [UNK] [UNK]

[BERTのAttentionを可視化_2]
[CLS] [UNK] 「 [UNK] 」 は [UNK] 病院 に 入院 する て いる 人 の ストレス や 不安 を 少ない する ため に 活動 する 犬 です [UNK] [UNK]

[BERTのAttentionを可視化_3]
[CLS] [UNK] 「 [UNK] 」 は [UNK] 病院 に 入院 する て いる 人 の ストレス や 不安 を 少ない する ため に 活動 する 犬 です [UNK] [UNK]

[BERTのAttentionを可視化_4]
[CLS] [UNK] 「 [UNK] 」 は [UNK] 病院 に 入院 する て いる 人 の ストレス や 不安 を 少ない する ため に 活動 する 犬 です [UNK] [UNK]

[BERTのAttentionを可視化_5]
[CLS] [UNK] 「 [UNK] 」 は [UNK] 病院 に 入院 する て いる 人 の ストレス や 不安 を 少ない する ため に 活動 する 犬 です [UNK] [UNK]

[BERTのAttentionを可視化_6]
[CLS] [UNK] 「 [UNK] 」 は [UNK] 病院 に 入院 する て いる 人 の ストレス や 不安 を 少ない する ため に 活動 する 犬 です [UNK] [UNK]

[BERTのAttentionを可視化_7]
[CLS] [UNK] 「 [UNK] 」 は [UNK] 病院 に 入院 する て いる 人 の ストレス や 不安 を 少ない する ため に 活動 する 犬 です [UNK] [UNK]

[BERTのAttentionを可視化_8]
[CLS] [UNK] 「 [UNK] 」 は [UNK] 病院 に 入院 する て いる 人 の ストレス や 不安 を 少ない する ため に 活動 する 犬 です [UNK] [UNK]

[BERTのAttentionを可視化_9]
[CLS] [UNK] 「 [UNK] 」 は [UNK] 病院 に 入院 する て いる 人 の ストレス や 不安 を 少ない する ため に 活動 する 犬 です [UNK] [UNK]

[BERTのAttentionを可視化_10]
[CLS] [UNK] 「 [UNK] 」 は [UNK] 病院 に 入院 する て いる 人 の ストレス や 不安 を 少ない する ため に 活動 する 犬 です [UNK] [UNK]

[BERTのAttentionを可視化_11]
[CLS] [UNK] 「 [UNK] 」 は [UNK] 病院 に 入院 する て いる 人 の ストレス や 不安 を 少ない する ため に 活動 する 犬 です [UNK] [UNK]

[BERTのAttentionを可視化_12]
[CLS] [UNK] 「 [UNK] 」 は [UNK] 病院 に 入院 する て いる 人 の ストレス や 不安 を 少ない する ため に 活動 する 犬 です [UNK] [UNK]

[BERTのAttentionを可視化_ALL]
[CLS] [UNK] 「 [UNK] 」 は [UNK] 病院 に 入院 する て いる 人 の ストレス や 不安 を 少ない する ため に 活動 する 犬 です [UNK] [UNK] [SEP]

図 2.14　第 1 層目から第 12 層目のアテンション (サ変接続名詞)

正解ラベル：Negative
推論ラベル：Negative

[BERTのAttentionを可視化_1]
[CLS] [UNK] 消費 者 庁 は 新しい 制度 の 周知 を 図る と ##とも ##に [UNK] 身 に 覚える が ない 商品 が 届く た 場合 に は 代金 引 ##換 を 求める られる て も 応じる ない こと や [UNK] 個人 情報 を 聞き ##出す れる お それ も ある ため 送り ##主 の 業者 に [UNK] て 連絡 する ない こと など [UNK] 引き ##続 ##く [UNK] 送る つける 商法 へ の 注意 を 呼びかける て いる ます [UNK] [UNK]

[BERTのAttentionを可視化_2]
[CLS] [UNK] 消費 者 庁 は 新しい 制度 の 周知 を 図る と ##とも ##に [UNK] 身 に 覚える が ない 商品 が 届く た 場合 に は 代金 引 ##換 を 求める られる て も 応じる ない こと や [UNK] 個人 情報 を 聞き ##出す れる お それ も ある ため 送り ##主 の 業者 に [UNK] て 連絡 する ない こと など [UNK] 引き ##続 ##く [UNK] 送る つける 商法 へ の 注意 を 呼びかける て いる ます [UNK] [UNK]

[BERTのAttentionを可視化_3]
[CLS] [UNK] 消費 者 庁 は 新しい 制度 の 周知 を 図る と ##とも ##に [UNK] 身 に 覚える が ない 商品 が 届く た 場合 に は 代金 引 ##換 を 求める られる て も 応じる ない こと や [UNK] 個人 情報 を 聞き ##出す れる お それ も ある ため 送り ##主 の 業者 に [UNK] て 連絡 する ない こと など [UNK] 引き ##続 ##く [UNK] 送る つける 商法 へ の 注意 を 呼びかける て いる ます [UNK] [UNK]

[BERTのAttentionを可視化_4]
[CLS] [UNK] 消費 者 庁 は 新しい 制度 の 周知 を 図る と ##とも ##に [UNK] 身 に 覚える が ない 商品 が 届く た 場合 に は 代金 引 ##換 を 求める られる て も 応じる ない こと や [UNK] 個人 情報 を 聞き ##出す れる お それ も ある ため 送り ##主 の 業者 に [UNK] て 連絡 する ない こと など [UNK] 引き ##続 ##く [UNK] 送る つける 商法 へ の 注意 を 呼びかける て いる ます [UNK] [UNK]

[BERTのAttentionを可視化_5]
[CLS] [UNK] 消費 者 庁 は 新しい 制度 の 周知 を 図る と ##とも ##に [UNK] 身 に 覚える が ない 商品 が 届く た 場合 に は 代金 引 ##換 を 求める られる て も 応じる ない こと や [UNK] 個人 情報 を 聞き ##出す れる お それ も ある ため 送り ##主 の 業者 に [UNK] て 連絡 する ない こと など [UNK] 引き ##続 ##く [UNK] 送る つける 商法 へ の 注意 を 呼びかける て いる ます [UNK] [UNK]

[BERTのAttentionを可視化_6]
[CLS] [UNK] 消費 者 庁 は 新しい 制度 の 周知 を 図る と ##とも ##に [UNK] 身 に 覚える が ない 商品 が 届く た 場合 に は 代金 引 ##換 を 求める られる て も 応じる ない こと や [UNK] 個人 情報 を 聞き ##出す れる お それ も ある ため 送り ##主 の 業者 に [UNK] て 連絡 する ない こと など [UNK] 引き ##続 ##く [UNK] 送る つける 商法 へ の 注意 を 呼びかける て いる ます [UNK] [UNK]

[BERTのAttentionを可視化_7]
[CLS] [UNK] 消費 者 庁 は 新しい 制度 の 周知 を 図る と ##とも ##に [UNK] 身 に 覚える が ない 商品 が 届く た 場合 に は 代金 引 ##換 を 求める られる て も 応じる ない こと や [UNK] 個人 情報 を 聞き ##出す れる お それ も ある ため 送り ##主 の 業者 に [UNK] て 連絡 する ない こと など [UNK] 引き ##続 ##く [UNK] 送る つける 商法 へ の 注意 を 呼びかける て いる ます [UNK] [UNK]

[BERTのAttentionを可視化_8]
[CLS] [UNK] 消費 者 庁 は 新しい 制度 の 周知 を 図る と ##とも ##に [UNK] 身 に 覚える が ない 商品 が 届く た 場合 に は 代金 引 ##換 を 求める られる て も 応じる ない こと や [UNK] 個人 情報 を 聞き ##出す れる お それ も ある ため 送り ##主 の 業者 に [UNK] て 連絡 する ない こと など [UNK] 引き ##続 ##く [UNK] 送る つける 商法 へ の 注意 を 呼びかける て いる ます [UNK] [UNK]

[BERTのAttentionを可視化_9]
[CLS] [UNK] 消費 者 庁 は 新しい 制度 の 周知 を 図る と ##とも ##に [UNK] 身 に 覚える が ない 商品 が 届く た 場合 に は 代金 引 ##換 を 求める られる て も 応じる ない こと や [UNK] 個人 情報 を 聞き ##出す れる お それ も ある ため 送り ##主 の 業者 に [UNK] て 連絡 する ない こと など [UNK] 引き ##続 ##く [UNK] 送る つける 商法 へ の 注意 を 呼びかける て いる ます [UNK] [UNK]

[BERTのAttentionを可視化_10]
[CLS] [UNK] 消費 者 庁 は 新しい 制度 の 周知 を 図る と ##とも ##に [UNK] 身 に 覚える が ない 商品 が 届く た 場合 に は 代金 引 ##換 を 求める られる て も 応じる ない こと や [UNK] 個人 情報 を 聞き ##出す れる お それ も ある ため 送り ##主 の 業者 に [UNK] て 連絡 する ない こと など [UNK] 引き ##続 ##く [UNK] 送る つける 商法 へ の 注意 を 呼びかける て いる ます [UNK] [UNK]

[BERTのAttentionを可視化_11]
[CLS] [UNK] 消費 者 庁 は 新しい 制度 の 周知 を 図る と ##とも ##に [UNK] 身 に 覚える が ない 商品 が 届く た 場合 に は 代金 引 ##換 を 求める られる て も 応じる ない こと や [UNK] 個人 情報 を 聞き ##出す れる お それ も ある ため 送り ##主 の 業者 に [UNK] て 連絡 する ない こと など [UNK] 引き ##続 ##く [UNK] 送る つける 商法 へ の 注意 を 呼びかける て いる ます [UNK] [UNK]

[BERTのAttentionを可視化_12]
[CLS] [UNK] 消費 者 庁 は 新しい 制度 の 周知 を 図る と ##とも ##に [UNK] 身 に 覚える が ない 商品 が 届く た 場合 に は 代金 引 ##換 を 求める られる て も 応じる ない こと や [UNK] 個人 情報 を 聞き ##出す れる お それ も ある ため 送り ##主 の 業者 に [UNK] て 連絡 する ない こと など [UNK] 引き ##続 ##く [UNK] 送る つける 商法 へ の 注意 を 呼びかける て いる ます [UNK] [UNK]

[BERTのAttentionを可視化_ALL]
[CLS] [UNK] 消費 者 庁 は 新しい 制度 の 周知 を 図る と ##とも ##に [UNK] 身 に 覚える が ない 商品 が 届く た 場合 に は 代金 引 ##換 を 求める られる て も 応じる ない こと や [UNK] 個人 情報 を 聞き ##出す れる お それ も ある ため 送り ##主 の 業者 に [UNK] て 連絡 する ない こと など [UNK] 引き ##続 ##く [UNK] 送る つける 商法 へ の 注意 を 呼びかける て いる ます [UNK] [UNK] [SEP]

図 2.15　第 1 層目から第 12 層目のアテンション (受身形)

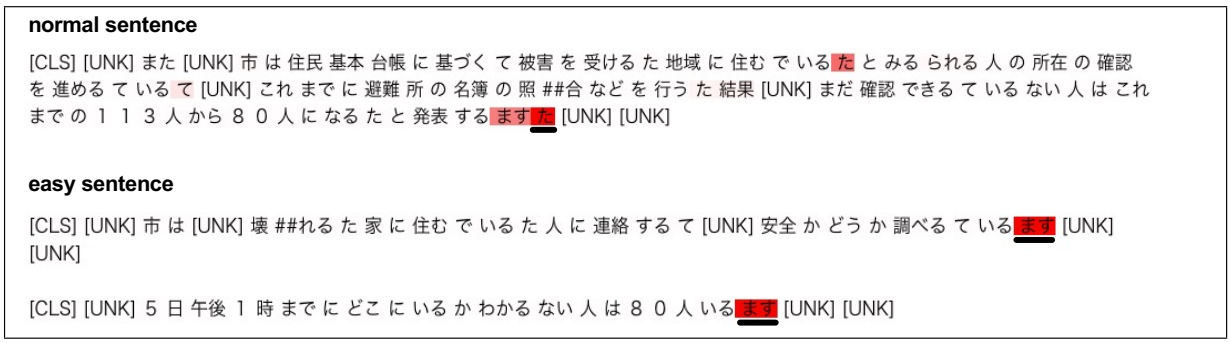


図 2.16　単語数とアテンション

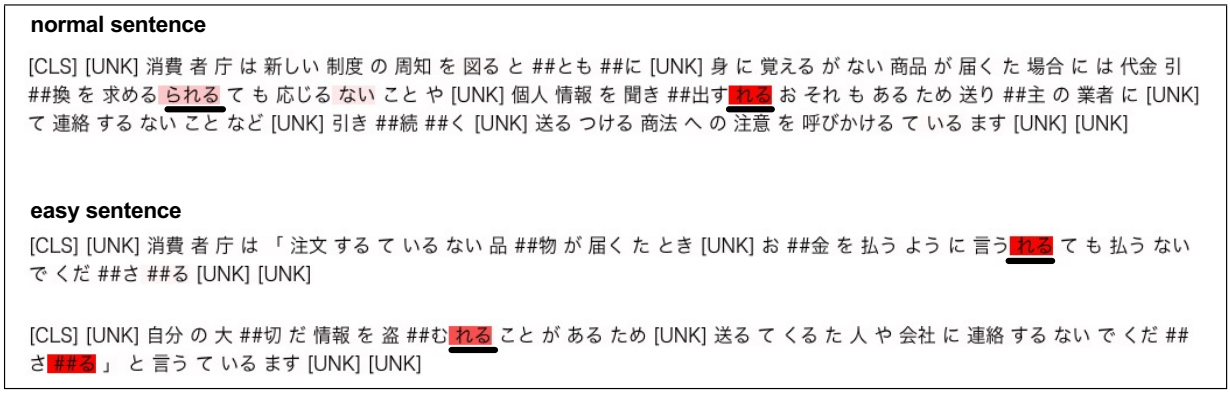


図 2.17　受身形と第 6 層目のアテンション

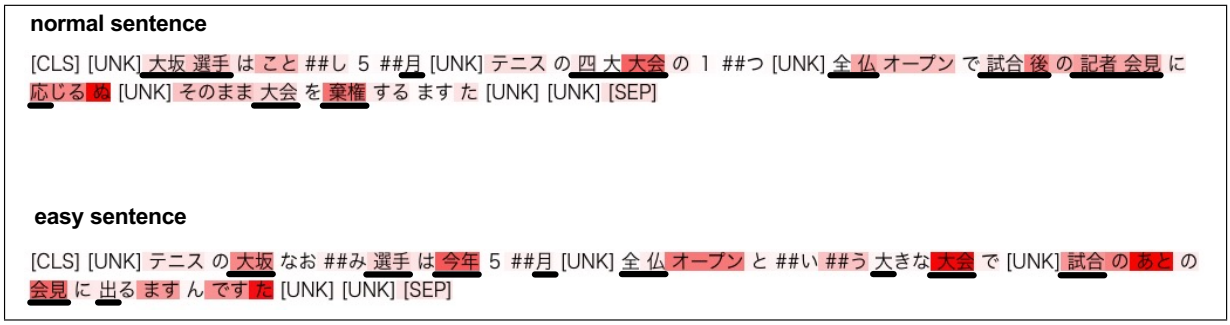


図 2.18　漢字とアテンションの関係

第 3 章

テキスト平易化のための類似度測定に関する研究

3.1 概要

テキスト平易化は，テキストの意味を保持したままテキストの難易度が低くなるよう書き換えることである．機械翻訳や要約のようなテキストからテキストへの書き換えタスクと同様に，テキスト平易化において一般的に使用されている評価指標は BLEU[1] や SARI[2] のような参照文と比較する手法が中心である．書き換えたテキストの評価として，BLEU は参照文と出力文の一致率で，SARI は入力文から参照文への編集操作に基づいて出力文を評価する．このような評価を自動評価をいうが，日本語においてはテキスト平易化のためのパラレルコーパス（入力文と参照文との意味的に一致する文ペア）が質的にも量的にも整備されていないことから，BLEU や SARI による評価が難しい．そこで，参照文を用いずにモデルの出力文を平易性，類似性，文法性で評価する品質推定が注目されている．本研究では出力文を平易性と類似性の観点で評価することを目的としており，本章では参照文を用いない品質推定において文の類似度を評価するための指標として相応しい手法を検証する．

3.2 関連研究

梶原ら [13] は，単語埋め込み表現を用いた文の類似度を測定する指標を検討したが，対象として深層学習（DL）ベースの指標は含まれていなかった．本研究でテキスト平易化のために書き換えられた文が，元の文との類似度評価をするために相応しい指標を調査することが目的であり，DL ベースの新しい指標も含めて検証する．

3.3　本研究で用いる手法

3.3.1　類似度測定の手順

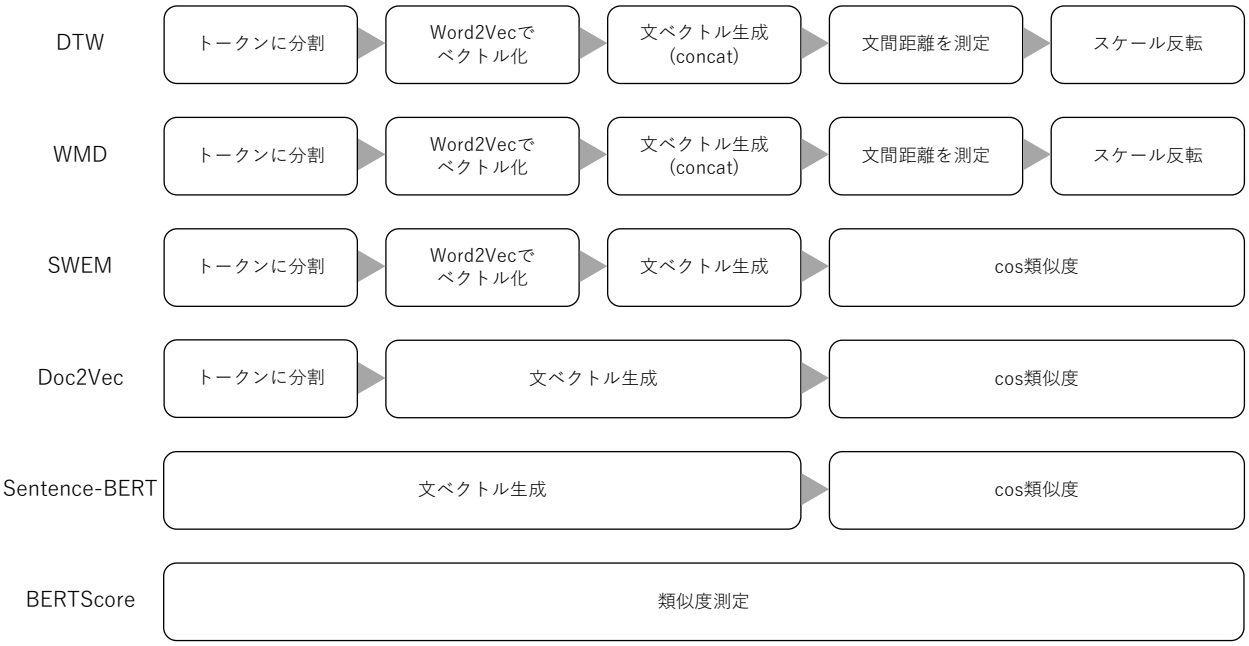


図 3.1　本実験で文間類似度を測定するための手法を比較

本研究では文間類似度を計測するために，DTW（Dynamic Time Warping）[27]，WMD（Word Mover's Distance）[28], SWEM（Simple Word-Embedding Model）[29], Doc2Vec[30], Sentence-BERT[31]，BERTScore[32] の 6 つの指標を試す．図 3.1は，それぞれの手順を比較したものである．DTW と WMD は文ベクトル同士の距離を測定する手法であるため，テキストをトークン（単語）の単位に分割し，Word2Vec を使ってトークンごとにベクトル化する．実際にはトークンは単語よりも小さい単位だが，本論文では便宜上「単語ベクトル」と表現する．単語ベクトルを連結して文ベクトルを生成して，文間距離を測定する．文間距離を類似度に置き換えるためにスケールを反転する．

SWEM はテキストをトークン単位に分割し，Word2Vec でベクトル化する．その後，SWEM 独自の手順で文ベクトルを生成する．文間類似度は，文ベクトルのコサイン類似度で求める．

Doc2Vec と Sentence-BERT は，テキストを入力すると学習済みモデルから文ベクトルを出力する．文間類似度は，文ベクトルのコサイン類似度で求める．

BERTScore は二つのテキストを入力すると，学習済みモデルから文間類似度を出力する．類似度として Precision，Recall，F1 を求めることができる．

3.3.2から 3.3.7では，詳細な手順を述べる．以降，原文を $X = (x_1, x_2, ..., x_{N_x})$，原文を平易化したものを $Y = (y_1, y_2, ..., y_{N_y})$ とする．

3.3.2 Dynamic Time Warping

Dynamic Time Warping(DTW) の詳細は 4.3.2で説明し，本章では文の類似度を測る手順のみを説明する．

DTW では二つの時系列データが最短となるときの各データの対応（パス）を得ることができる．本実験では各データの距離をユークリッド距離で計算する．テキストデータの場合，データの総距離は文間の距離と解釈できる．これは文と文の間の各単語の距離を計算し，総距離を最小化するワーピングパス経路を探索する．ワーピングパスは $W = \{(x, y)|x \in X, y \in Y, \forall x, y\}$ で与えられる．文間距離の求め方は以下の通りである．

$$\delta(x, y) = \| x - y \|, \tag{3.1}$$

$$DTW(X, Y) = \sum_{(x,y)\in W}^{|W|} \delta(x, y). \tag{3.2}$$

DTW を用いて類似度を求める方法は次のとおりである．テキストを分かち書きしてトークンに分割する．トークンをそれぞれ Word2Vec[33] で 200 次元のベクトルに変換し，文の整列順にベクトル列を作る．これを文ベクトルとする．二つの文ベクトルから文間距離を計算する．文の長さが長いほど文間距離は大きくなる．つまり類似度が低くなってしまうため，文間距離としては総距離をインデックスのペア数で割った値を使用する．また，文間距離が大きいほど類似度が低く，文間距離が小さいほど類似度が高いため，文間距離のスケールを反転させる．実験では，測定したデータの最大値から文間距離を差し引いた値を類似度とする．

3.3.3 Word Mover's Distance

Word Mover's Distance(WMD) の詳細は 4.3.3で説明し，本章では文の類似度を測る手順のみを説明する．

3.3.2の DTW がデータの始点から終点まで単調増加しながら距離を計算するのに対して，WMD はデータの順序を考慮しない．WMD を用いた文間距離の求め方は以下の通りである．

$$WMD(X,Y) = min \sum_{x,y \in D} T_{xy}\delta(x,y) \tag{3.3}$$

$subjectto$:

$$\sum_{y \in D} T_{xy} = \frac{1}{|X|} freq(x,X) \forall x \in X \tag{3.4}$$

$$\sum_{x \in D} T_{xy} = \frac{1}{|Y|} freq(y,Y) \forall y \in Y \tag{3.5}$$

ここで T_{xy} は単語から単語への重み行列で，一方の文の単語から他方の文の単語への配分比率を表す．

WMD を用いて類似度を求める方法は次のとおりである．テキストを分かち書きしてトークンに分割する．トークンをそれぞれ Word2Vec[33] で 200 次元のベクトルに変換し，文の整列順にベクトル列を作る．これを文ベクトルとする．二つの文ベクトルから文間距離を計算する．文間距離は，文の長さが長いほど文間距離は大きくなる．つまり類似度が低くなってしまうため，文間距離としては総距離をインデックスのペア数で割った値を使用する．また，文間距離が大きいほど類似度が低く，文間距離が小さいほど類似度が高いため，文間距離のスケールを反転させる．実験では，測定したデータの最大値から文間距離を差し引いた値を類似度とする．

3.3.4 Simple Word-Embedding Model

Simple Word-Embedding Model(SWEM) は単語ベクトルから文ベクトルを生成する手法である．文ベクトルを得る手法には，Doc2Vec や BERT のように大規模コーパスからニューラルネットワークで学習する方法が代表的であるが，データが大量に必要であることと計算に時間を要するという懸念がある．SWEM は単純な計算で文ベクトルを生成し，既存の CNN や LSTM モデルと同等以上の精度が得られるため，本研究では BERT モデルと比較するためのベースラインとして SWEM を採用した．

SWEM は以下の 4 つの方法が提案されている．ここで，L は単語ベクトルの長さで，単語ベクトルを $x = (v_1, ... , v_L)$ とする．

[1] SWEM-max：図 3.2のように，単語ベクトルの各次元に対して max pooling する．

$$SWEM\text{-}max = Max\ pooling\,(v_1, v_2, \ldots, v_L) \tag{3.6}$$

[2] SWEM-aver：図 3.2のように，単語ベクトルの各次元に対して average pooling する．

$$SWEM\text{-}aver = \frac{1}{L}\sum_{i=1}^{L} v_L \tag{3.7}$$

[3] SWEM-concat：SWEM-aver と SWEM-aver を結合する．

$$SWEM\text{-}concat \ = \ concat\ (SWEM\text{-}aver, SWEM\text{-}max) \tag{3.8}$$

[4] SWEM-hier：図 3.3のように，n-gram で average-pooling した結果を max-pooling する．SWEM-hier は，固定長の window で平均プーリング結果に対して最大プーリングを行う．本実験の window サイズは 5 とした．従ってトークンが 5 個以下の文は類似度が計算できないため，実験データから除外した．

SWEM を用いて類似度を求める方法は次のとおりである．テキストを分かち書きしてトークンに分割する．トークンをそれぞれ Word2Vec[33] で 200 次元のベクトルに変換し，上の [1] から [4] の方法で文ベクトルを生成する．文ベクトルからコサイン類似度を計算する．

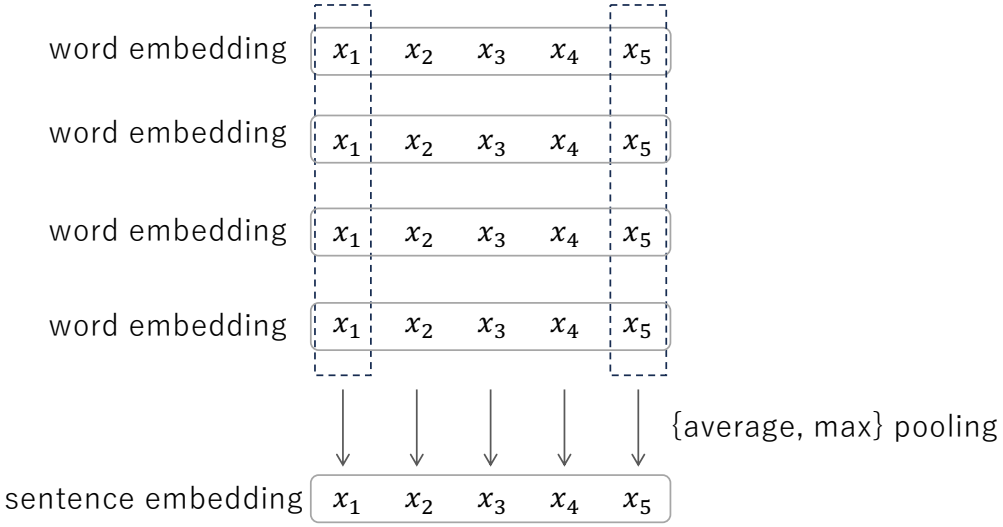


図 3.2　SWEM-aver and SWEM-max

3.3.5　Doc2Vec

Doc2Vec はテキストから固定長ベクトルを得る手法である．単語の埋め込み表現を獲得する Word2Vec[33] に対して，Doc2Vec は文（全ての単語）の埋め込み表現を獲得することに重点を置いている．Doc2Vec に入力する情報は単語群として扱うため，文としての単語の順序は考慮されない．Doc2Vec の埋め込み表現獲得のため，日本語 Wikipedia を用いてモデルの事前学習を行っ

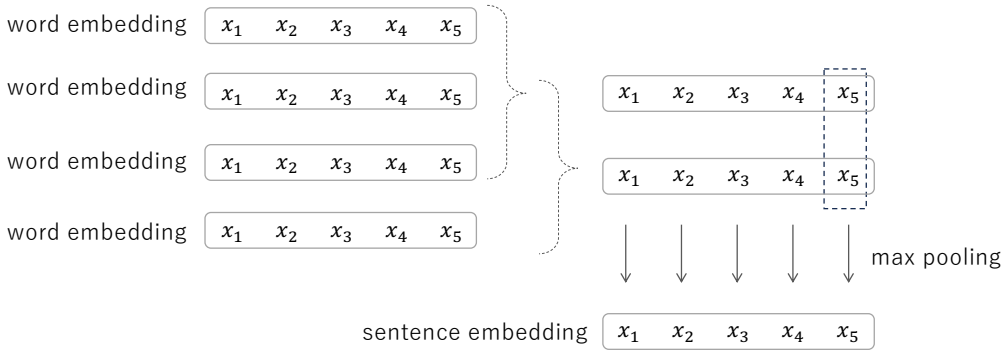


図 3.3　SWEM-hier

た．ベクトルの次元数は 200，window サイズは 15 とした．

Doc2Vec を用いて類似度を求める方法は次のとおりである．テキストを分かち書きしてトークンに分割する．トークンリストを Doc2Vec に渡して文ベクトルを生成する．文ベクトルからコサイン類似度を計算する．

3.3.6　Sentence-BERT

Sentence-BERT は BERT[15] をファインチューニングして文章ベクトルを生成する手法である．ファインチューニングには二つの類似する文章ペアを学習データとし，似た文章から生成されるベクトル同士は近くになるよう学習する．本実験で使用するモデルは東北大学が公開する事前学習済み BERT モデルをファインチューニングしたもので，huggingface で公開されている Sentence-BERT モデル[*1]を使用する．

Sentence-BERT を用いて類似度を求める方法は，Sentence-BERT モデルを用いてテキストから文ベクトルを獲得し，文ベクトルからコサイン類似度を計算する．

3.3.7　BERTScore

BERTScore は BERT[15] を利用してテキスト間の類似度を測定する．モデルに文対のテキストを入力すると，その類似度を Precision，Recall，F1 で出力するモデルである．内部的には，トークン（単語）のベクトル同士のコサイン類似度を合計して算出している．本来，BERTScore は生成文と参照文の類似度を求める手法であるため，モデルの出力をそのまま類似度とする．Recall，Precision，F1 スコアは次のように計算する．

[*1] urlhttps://huggingface.co/sonoisa/sentence-bert-base-ja-mean-tokens-v2

$$R_{BERT} = \frac{1}{|X|} \sum_{x_i \in X} \max_{y \in Y} x_i y_j \tag{3.9}$$

$$P_{BERT} = \frac{1}{|Y|} \sum_{y_j \in Y} \max_{x \in X} x_i y_j \tag{3.10}$$

$$F_{BERT} = 2\frac{P_{BERT} \cdot R_{BERT}}{P_{BERT} + R_{BERT}} \tag{3.11}$$

3.3.8 実験で使用するモデル，およびベクトルの次元数

表 3.1に実験で使用するハイパーパラメータ，および利用した事前学習済みモデルを示す．Word2vec の単語ベクトルは一般的に 100 次元から 300 次元に設定される．ベクトルの次元数は高ければ類似の単語の距離が遠くなってしまい，低ければ表現力が落ちてしまうため，本実験では 200 次元に設定する．また，単語の表現を文脈から推定する場合，前後の単語をいくつ考慮するかを window size として設定する必要がある．本実験では，単語ベクトルは window size を 5，文ベクトルは window size を 15 に設定する．また，事前学習によってベクトル表現を獲得する場合，事前学習に使用したコーパスの規模によって精度が変わる可能性があることから，本実験では全てのモデルで Japanese Wikipdeia を使用して事前学習を行なったモデルを使用する．公開されている事前学習済みモデルを使用した場合は，表 3.1の「事前学習モデル」の列に記載する．(*1) は GitHub[*2]から，(*2) は Huggingface[*3]から入手する．

3.4 実験

3.4.1 目的

3.3.2から 3.3.7までの類似度測定指標の中から，平易化によって書き換えた文の類似度を測定するために相応しい指標を検証する．人手で類似度ラベルを付与した JSICK コーパスと，人手でやさしい日本語に書き換えた SNOW T23 コーパスを用いて類似度を測定する．その後，ロジスティック回帰で 6 つの指標を説明変数，文の類似度を目的変数として推定モデルを作り，説明変数の中でモデルに影響を与えている指標を判断する．

[*2] urlhttps://github.com/singletongue/WikiEntVec
[*3] urlhttps://huggingface.co/cl-tohoku/bert-base-japanese-v2

表 3.1　各指標のベクトル次元数と事前学習に使用したコーパスサイズ

	ベクトル次元数	window size	事前学習用コーパス	事前学習モデル
DTW	200 (Word2vec)	5	Japanese Wikipedia	Wikipedia Entity Vectors(*1)
WMD	200 (Word2vec)	5	Japanese Wikipedia	Wikipedia Entity Vectors(*1)
SWEM	200 (Word2vec)	5	-	-
Doc2vec	200	15	Japanese Wikipedia	Original
Sentence-BERT	768	-	Japanese Wikipedia	cl-tohoku BERT(*2)
BERTScore	768	-	Japanese Wikipedia	cl-tohoku BERT(*2)

3.4.2　実験に使用するデータ

実験に使用するデータは，JSICK[2]，SNOW T23[6] である．JSICK は，含意関係と意味的類似度の正解ラベルのアノテーションを付与したデータセットである．英語版の SICK[34] というデータセットを人手で日本語に翻訳し，アノテーションを日本語に合わせて付与し直して構築された．このデータセットは平易化のためのデータセットではないが，本実験で使用する指標が類似度測定に相応しいことを確認できる．JSICK のデータセットのうち，今回の実験で使用する項目は日本語文 A，日本語文 B，日本語文 A と日本語文 B の関連度とする．関連度は 3 人のアノテーション結果の平均で求められ，最小値は 1，最大値は 5 である．JSICK の一部を表 3.2に示す．

表 3.2　JSICK データセットの一部抜粋

日本語文 A	日本語文 B	関連度
ある人が帽子をかぶって草原に座っている	ある人が野原で座っていて帽子をかぶっている	4.7
子供たちのグループが庭で遊んでいて、後ろの方には年を取った男性が立っている	庭にいる男の子たちのグループが遊んでいて、男性が後ろの方に立っている	3.7
二匹の犬が木のそばで遊んでいる	空中に跳んでいる犬は一匹もいない	1.7

SNOW T23 はやさしい日本語拡張コーパスと呼ばれ，リリースされている SNOW T15 に新た

なデータを追加したコーパスである．SNOW T15 は機械翻訳用の日英対訳コーパス[*4]から 5 万文をやさしい日本語に書き換えたものである．このデータセットにおけるやさしい日本語とは，独自に定義した UniDic 単語体系の 2,000 語に限定することである．SNOW T23 の一部を表 3.3に示す．

表 3.3　SNOW T23 データセットの一部抜粋

日本語	やさしい日本語
用紙の下部に名前を書きなさい。	準備された紙の下の方に名前を書きなさい。
彼は怒りに我を忘れた。	彼は自分のことを忘れるくらい腹を立てた。
新しいビルが私の窓からの眺めをさえぎった。	新しい高い建物で私の家の窓から外の景色が見えなくなった。

それぞれのデータセットにおける，1 文あたりの文字数と単語数を表 3.4に示す．source は書き換え前のテキスト，target は書き換え後のテキストを表す．JSICK は source と target に難易度の差はないが，SNOW T23 は source には通常文，target には平易文を割り当てる．SNOW T23 はやさしい日本語へ書き換えることで，文字数の平均は約 19 文字から約 21 文字へと増加しており，語彙を限定することで冗長になることが読み取れる．

表 3.4　JSICK，SNOW T23 における文字数と単語数

		num. of characters			num. of words		
		ave.	max.	min.	ave.	max.	min.
JSICK	source	21.38	52	10	12.42	28	6
	target	20.58	44	10	12.24	23	5
SNOW T23	source	19.47	47	9	12.11	34	6
	target	21.05	48	9	13.65	34	6

3.5　結果

類似度データセットである JSICK に対して，それぞれの指標で類似度測定した結果である．JSICK から 100 ペアの文対をランダムに抽出したものを実験の対象とする．JSICK には人手評価ラベルが付与されているため，人手評価ラベルとそれぞれの類似度測定値の相関係数を図 3.4に示す．人手評価ラベルは「score」と表記している．score と最も相関が強い評価指標は BERTScore

[*4] urlhttps://github.com/odashi/small_parallel_enja

の Recall 値で，相関係数は 0.81 であった．次いで，WMD と Sentence-BERT が 0.78 であった．複数の計算手法を持つ SWEM および BERTScore は，以降の実験では最も相関係数が高かった SWEM-aver と BERTScore の recall 値をそれぞれの代表値として利用する．

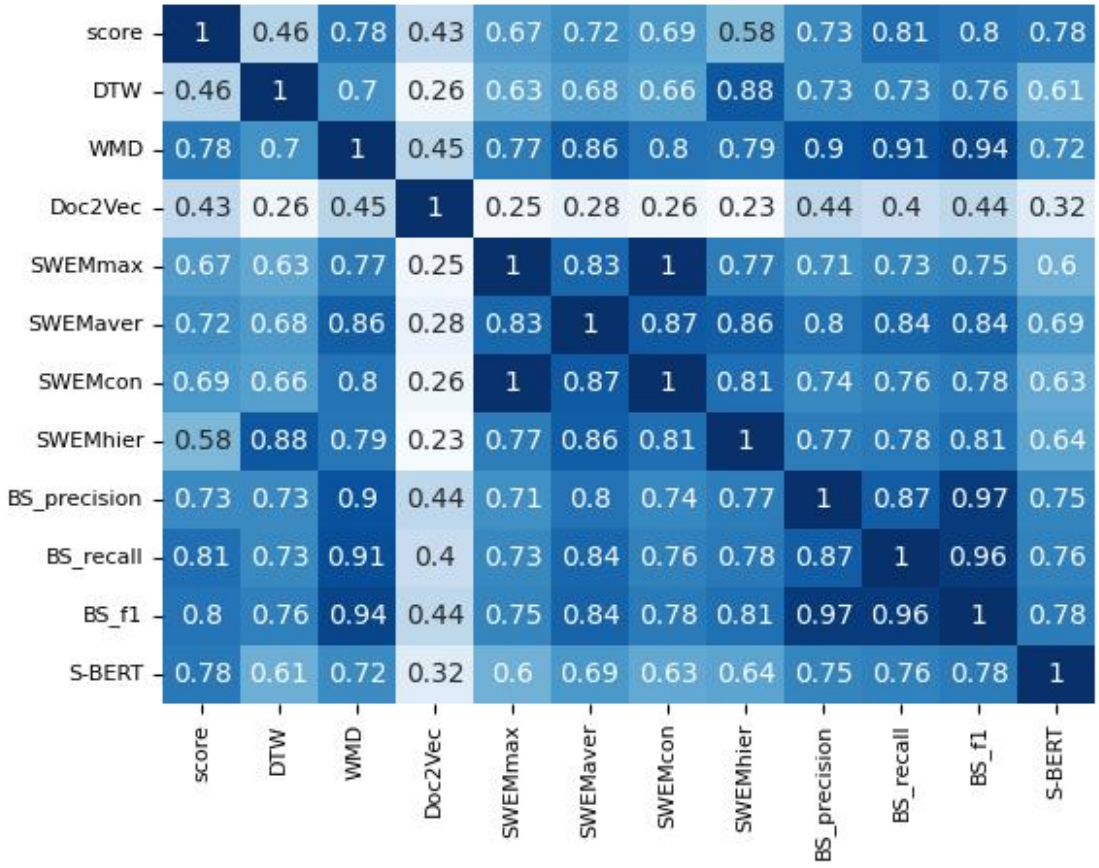


図 3.4　JSICK コーパスの類似度測定結果

次に，平易化のためのパラレルコーパスである SNOW T23 コーパスを使い，類似度測定値を比較する．SNOW T23 コーパスは類似度ラベルを持たないパラレルコーパスであるため，文対をランダムにシャッフルしたものをベースライン（BL）として類似度を比較する．実験ではコーパスから 100 ペアを抽出し，各類似度の標本平均を測定する．その調査を 100 回繰り返した結果の平均（標本平均の平均）を比較する．

図 3.5は標本平均の平均をエラーバーで表したものである．グラフの縦軸は類似度で，指標によってスケールが異なる．左側のバーが類似度，右側のバーがベースラインである．測定値の標準

誤差を赤い線で表している．ベースラインと類似度の平均の差が最も大きいのは Sentence-BERT であることがグラフから視覚的に読み取れる．表 3.5は標本平均の平均と標準誤差の値を示したものである．DTW は他の指標と比較すると標準誤差が大きく，標本平均のばらつきが大きいことがわかる．また，ベースラインの類似度とコーパスの類似度の平均値に差があるかを検証するために t 検定を実施した．表 3.5より，p 値は全ての指標において 0.05 を下回っており，統計的に有意差があると言える．

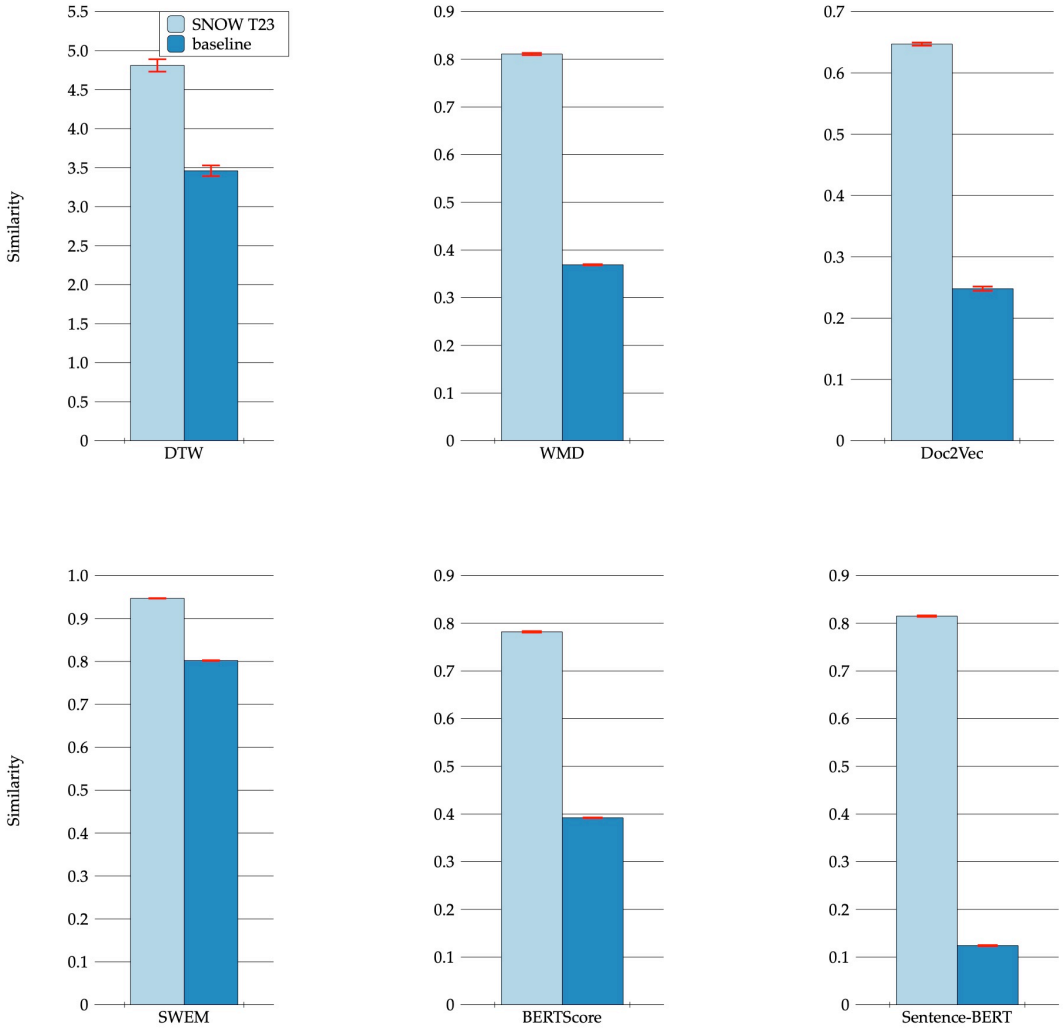


図 3.5　SNOW T23 コーパスの類似度測定結果

類似度測定に最も有効な指標を見つけるために，6 つの指標を説明変数としたロジスティック回帰を用いて類似度を推定するモデルを生成し，推定に重要な説明変数を明らかにする．ロジスティック回帰に用いるデータは SNOW T23 コーパスからランダムに抽出した文対 1,000 件（類似度が高いペア）と，シャッフルした文対 1,000 件（類似度が低いペア）である．シャッフルしてい

表 3.5　SNOW T23 コーパスの類似度とベースラインの比較

	DTW	WMD	Doc2Vec	SWEM	BERTScore	Sentence-BERT
類似度の標本平均	4.81E+00	8.11E-01	6.47E-01	9.47E-01	7.82E-01	8.15E-01
BL の標本平均	3.46E+00	3.69E-01	2.48E-01	8.02E-01	3.92E-01	1.24E-01
類似度の標準誤差	0.007929	0.002300	0.002485	0.000469	0.00150	0.001354
BL の標準誤差	0.006861	0.001138	0.003514	0.000790	0.00065	0.001141
p 値	1.78E-192	1.1E-217	3.61E-165	2.12E-210	3.42E-245	1.05E-287

ない文対には LABEL=1 を，シャッフルした文対には LABEL=0 を付与し，これを目的変数とする．データの 8 割（800 件）を学習データとして，データの 2 割（200 件）を検証データとして，モデルの正解率を測る．説明変数のうち，p 値が 0.05 を下回っているものが目的変数に対して影響を与えているものと判断できる．

ロジスティック回帰の結果を表 3.6に示す．p 値が 0.05 を下回っているのは Sentence-BERT のみである．また z 値は回帰係数を標準誤差で割った値で，この値が大きいほど誤差が小さいため，信頼できる値であると言える．説明変数の中で，z 値が最も大きいのは Sentence-BERT である．つまり，モデルに与える影響力が大きく，最も信頼できる説明変数が Sentence-BERT であることが明らかになった．このときのモデルの正解率は 97.3% である．

表 3.6　ロジスティック回帰の統計値

colname	回帰係数	z 値	p 値	significant
DTW	0.058235988	0.145217803	0.88453891	
WMD	1.570994076	0.579309427	0.562380406	
Doc2Vec	0.854304086	1.273847016	0.202717639	
SWEM	0.901826368	0.52882084	0.59692974	
BERTScore	2.41675449	0.643574413	0.519851447	
Sentence-BERT	6.710589564	5.724268067	1.04E-08	*

3.6　まとめ

本章では，テキスト平易化によって書き換えた文の意味的類似度を評価する方法を調査した．文の類似度を比較する実験では，日本語類似度データセット JSICK，やさしい日本語拡張コーパス SNOW T23 を使用した．評価指標として DTW，WMD，SWEM，Doc2vec，Sentence-BERT，BERTScore を用いて比較したところ，JSICK の類似度スコアと類似度指標には強い相関があった．SWEM の文ベクトル生成手法のうち単語ベクトルの各次元に対して平均をとる方法（SWEM-

aver）および BERTScore の Recall がそれぞれの指標の中で最も類似度ラベルと相関が高かったため，以降の実験では SWEM-aver と BERTScore の Recall を採用した．次にテキスト平易化のためのコーパスとして SNOW T23 において，それぞれの指標を計測したところ，全ての指標においてベースラインと比較すると $p<0.05$ で有意な差があることが明らかになった．つまり，全ての指標が平易化のための類似度測定に有効であると言える．さらに 6 つの指標を説明変数とするロジスティック回帰モデルにより，類似度推定に影響を与えている説明変数を調べたところ，Sentence-BERT の p 値のみが 0.05 を下回っており推定に影響を与えていることがわかった．これらのことから，Sentence-BERT がテキスト平易化のための類似度測定に相応しいと結論付けた．

第 4 章

テキスト平易化パラレルコーパスの構築と生成モデルを用いた言い換え

4.1 概要

4.1.1 背景

テキスト平易化は，テキストの意味を保持したまま難解なテキストを平易なテキストに書き換えるタスクである．機械翻訳やテキスト生成などの手法を使ってテキストを書き換える場合，モデルを学習するために教師データとなるパラレルコーパスが必要である．テキスト平易化のためのパラレルコーパスとは，意味が同じ文ペアで，一方が難解なテキスト，もう一方が平易なテキストであるデータのことを示す．同一言語において難易度の異なる書き換えを行うことは少ないため，自然にコーパスの資源が蓄積されることはないため，コーパスを構築するためには人手による書き換えが一般的ではあるがコストがかかるという問題がある．

近年，自然言語処理モデルでは Transformer[16] を用いた事前学習モデルが質疑応答，翻訳，テキスト生成などのタスクで活用されるようになった．事前学習フェーズでは，ラベルなしテキストを大量に学習して言語モデルを生成する．具体的には，単語の一部をマスクしてモデルに推測させることで単語の意味を習得し，また，二つの文が連続する文かどうかを推測させることで文の関係性を習得する．事前学習によって獲得した単語の埋め込み表現を各タスクへファインチューニングすることで，少ないデータセットであっても高い精度が得られる．本研究では，限られた資源から精度の高いパラレルコーパスを構築し，日本語による事前学習モデルをファインチューニングすることで平易化モデルを実現することを目指す．

先行研究によるとテキスト平易化のためのパラレルコーパスは，文と文を対応づけした文アライ

メントによるものが主流である．しかし，人手による平易なテキストへの書き換えでは，1 つの文を複数文へ分割したり，一部の文を省略したりすることがある．日本語の難易度に関する特徴によると文の長さが長いほど難易度が高くなるという研究報告 [18, 20, 17, 24] があり，前述の操作はこれに基づいた編集と言える．このことから，パラレルコーパスの構築において，文単位での対応付けに依らない，意味が一致する単語列を見つけ出すことが必要であると考えている．

本研究では，Dynamic Time Warping(DTW) を用いた単語アライメントによる，テキスト平易化のためのパラレルコーパスを構築する手法を提案する．テキスト平易化は同一言語における言い換えであるため，文法的に語順が入れ替わることが少ない．したがって，テキストを時系列データとして扱い，データの対応づけを求めることで言い換えを抽出する．書き換えによって文の分割・統合がある場合でも，意味的に一致する単語列として対応付けし，文の区切りに依らないパラレルコーパスを生成することを目指す．その後，提案手法によって構築した平易化パラレルコーパスを使用して平易化モデルにファインチューニングし，平易化モデルを構築する．モデルの出力文は，第 2章および第 3章の手法で品質推定する．

4.1.2　提案手法

パラレルコーパスの自動生成

テキストを単語に分割して単語を埋め込み表現（つまり単語ベクトル）に変換すると，文は単語ベクトルの並びとなる．単語ベクトルは文脈に影響を受けて算出し，意味が近い単語はベクトルの距離が小さい位置に配置されるため，書き換え前後の単語の対応はベクトル間の距離が小さい．図 4.1は単語ベクトルの並びをシーケンスデータと解釈し，シーケンスデータ同士のデータの対応を DTW で求めるときのイメージである．sequence 1 と seqence 2 は単語ベクトルを文における出現順に列挙したデータである．sequence 1 と seqence 2 の各単語ベクトル同士の距離をパス（path）という．DTW ではパスの総距離（warping path）が最短となるときの，単語ベクトルの対応を得ることができる．

本研究では，複数文を含むニュース記事を文に分割することなくテキスト全体をシーケンスデータとし，単語の対応を検出する．図 4.2のように，DTW によって単語の対応を検出し，あるテキストと意味的に一致するもう一方のテキストを抽出する．

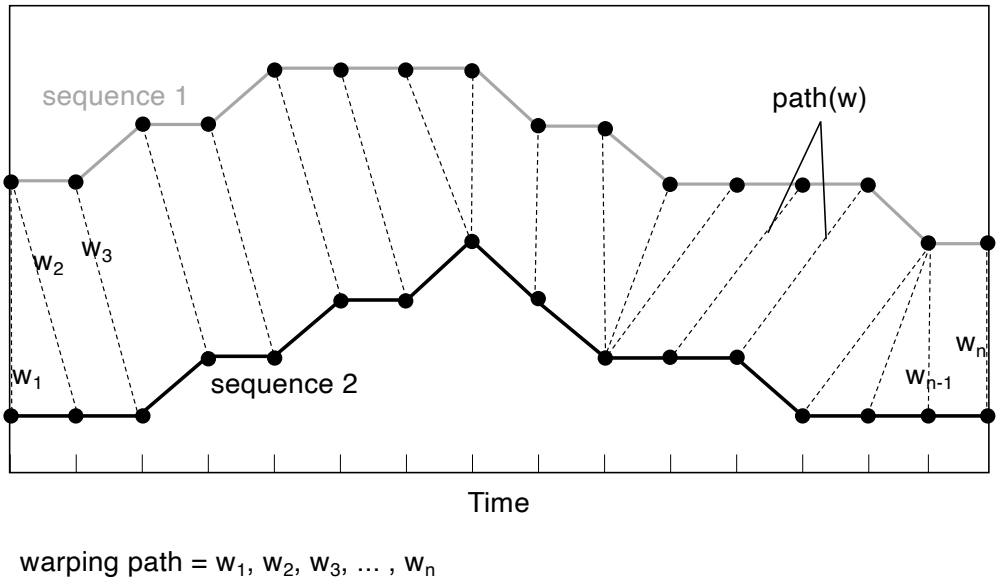


図 4.1　DTW を利用した文間距離

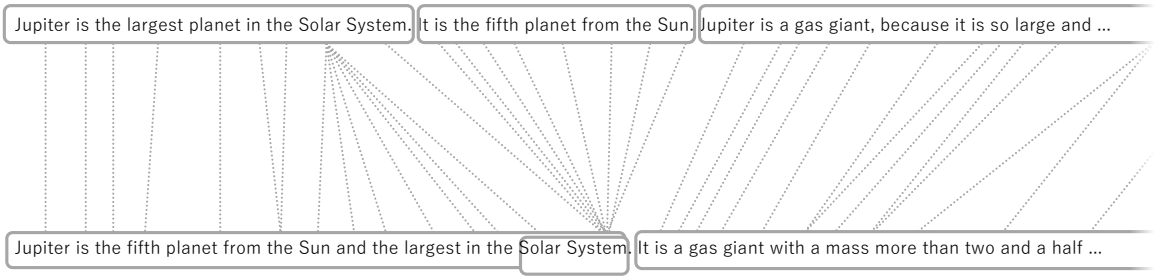


図 4.2　DTW による単語の対応

生成モデルを用いた平易化

生成したパラレルコーパスを用いて，テキスト平易化モデルを構築する．平易化モデルには事前学習済みの Text-to-Text Transfer Transformer(T5) を用いて，パラレルコーパスでファインチューニングする．モデルの評価には，平易性の品質推定として第 2章の手法を，類似性の品質推定として第 3章の手法を適用する．

4.2 関連研究

4.2.1 平易化のためのパラレルコーパス構築についての関連研究

テキスト平易化のためのコーパス自動構築に関する関連研究を紹介する．研究の対象が英語の場合，English Wikipedia と Simple English Wikipedia がコーパス資源として用いられることがある．Zhu ら [8] は，English Wikipedia と Simple Englieh Wikipedia から文の類似度が高い文ペアを選択することでパラレルコーパスを構築した．この研究では，次の 3 つの観点で文の類似度測定を行った．(1) 文レベルの類似度を TF-IDF で，(2) 単語の重複についての類似度測定を [35] の手法で，(3) 単語の編集による距離を [36] の手法でテストした．その結果，(1)TF-IDF の測定値が最も優れていたため，この手法を用いて文レベルの類似度を選択した．

また，日本語を対象としたコーパスの構築に関する研究としては，kajiwara ら [9] による大規模コーパスから難解な文と平易な文の同義な対を抽出し，擬似パラレルコーパスを構築する研究がある．大規模コーパスをリーダビリティ推定によって難解な文と平易な文に分類し，それぞれのコーパスから単語埋め込み表現に基づく文間類似度を用いて，難解な文と平易な文の対応を得る．大規模コーパスには現代日本語書き言葉均衡コーパス（BCCWJ）[37] を使用し，難易度測定にはリーダビリティ推定である [38] の単語難易度の平均値を用いた．

English Wikipedia と Simple English Wikipedia のパラレルコーパス [8] や，擬似パラレルコーパス [9] は，単文の類似度を測り，対応する文を探すアプローチでパラレルコーパスを構築した．しかし，平易な文は長さが短いという研究報告 [18, 17, 24] にしたがって，1 つの文を複数の文へ分割して書き換えられることがある．そのため，単文のペアを見つけるだけでは不十分で，単文と意味的に近い複数の文，もしくは文の一部を見つける必要があると考える．本研究では単語の対応によって意味的に近いテキストを見つける手法を提案する．

4.2.2 事前学習済みモデルを利用したテキスト平易化についての関連研究

丸山ら [39] は，異なる言語間の翻訳と比較すると，同一言語間のテキスト平易化はパラレルコーパス資源が非常に少ないことから，事前学習済みモデルを小規模なコーパスでファインチューニングする手法を試みた．充分なパラレルコーパスで学習したエンコーダ・デコーダモデルと比較すると，約 3,000 件の教師ありデータのみで学習した提案手法で同等の精度が得られた．事前学習済みモデルには Transformer[16] ベースのエンコーダ・デコーダモデルを採用した．実験の結果から，少ないパラレルコーパスで事前学習済みモデルをファインチューニングする手法が有効であると結

論づけた.

4.3 本章で用いる手法

4.3.1 Text-to-Text Transfer Transformer(T5)

自然言語処理では Transformer[16] の登場以降，テキストを自己教師あり学習で事前学習し，それをファインチューニングして解きたいタスクに適応する手法がスタンダードになってきた．事前学習の段階ではテキストにアノテーションを必要としないため，インターネット上にある膨大なテキストデータを利用してモデルを訓練することができる．例えば Wikipedia のような膨大なコーパスは事前学習に利用されることが多い．またファインチューニングでは少量のデータセットで様々なタスクに応用できるため，希少言語との翻訳をはじめデータセットが十分に用意できない分野では有意義な手法である．

T5 は Text-to-Text Transfer Transformer の略で，入力・出力が共にテキスト形式である様々なタスクを，同一のモデルで解くことができる．例えば，質疑応答，要約，翻訳などはそれぞれ別のモデルで学習する必要があったが，T5 は同一のモデルで各タスクの学習を行うことができる．T5 の基礎となっているのは Transformer[16] である．T5 はエンコーダとデコーダから成るモデルで，エンコーダに入力したテキストをもとにデコーダがテキストを生成する．最初に T5 の基盤である Transformer について述べた後，T5 の特徴である事前学習について述べる．

Transformer

Transformer モデルの構造を図 4.3に示す．左側のブロックはエンコーダで，2.3.1で説明する BERT と同様であるため説明を省略する．右側のブロックはデコーダで，エンコーダの構造に Masked Multi-Head Attention ブロックが追加されている．デコーダへの入力は出力テキストの教師データであるため，デコーダが予測する際に見るべきではない情報を含む．例えば翻訳タスクの場合，エンコーダへの入力が翻訳前テキスト，デコーダへの入力は翻訳後テキストである．デコーダは文の先頭から 1 単語ずつ出力していくため，デコーダの Masked Multi-Head Attention ブロックでは i 番目の予測の際は i 番目以降の単語をマスクして正解を見せないようにする．

T5

T5 のフレームワーク構成図を図 4.4に示す．T5 の特徴は，テキストを入力し，テキストを生成するあらゆるタスクに応用できることである．Text-to-Text のタスクには翻訳，要約，質疑応答な

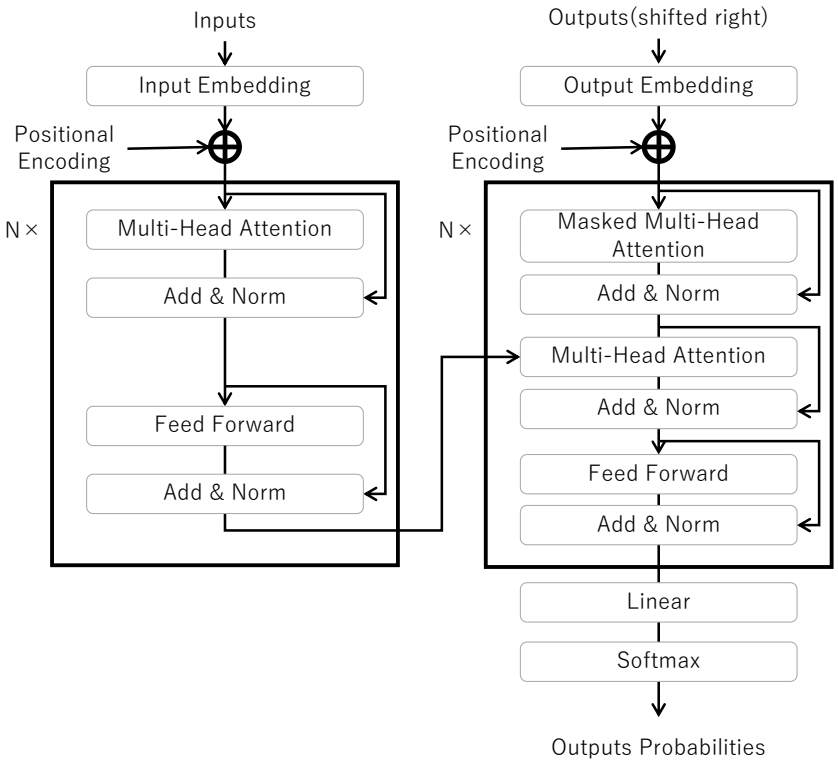


図 4.3　Transformer の構造

どがあり，従来はそれぞれに適したモデルを構築する必要があったが，T5 は 1 つのモデルで様々なタスクを解くことができる．図 4.4の入力テキストを見ると，テキストの前に":"で区切ったラベルを含むことがわかる．このラベルがタスクの種類を示している．

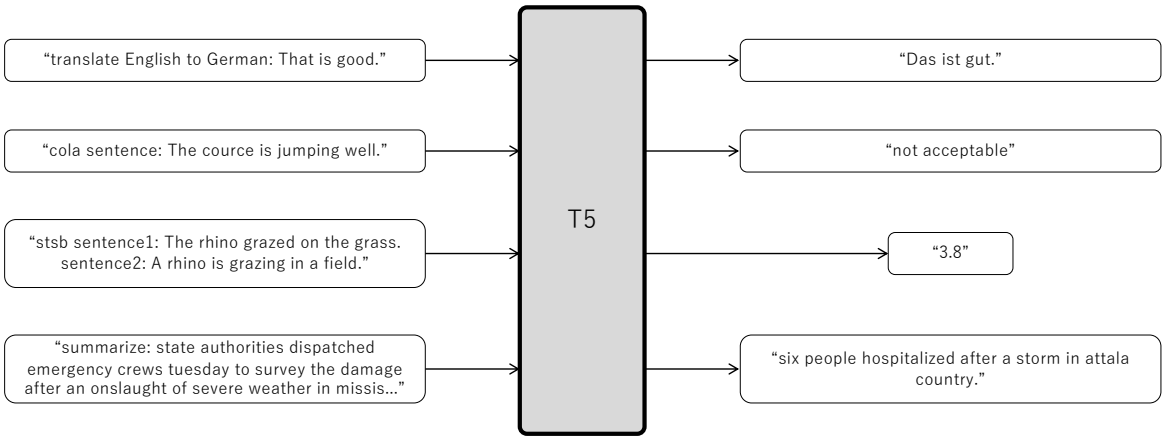


図 4.4　Text-to-Text フレームワーク：「Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer」の Fig.1 を元に筆者が手を加え作成した

モデルの事前学習では大量のテキストを読み込み，一部のマスクした単語を予測することで単語の特徴を捉える．図 4.5は事前学習におけるテキストのインプットとアウトプットの関係を示す．元のテキストから 15% のトークンをランダムにマスクしたものを入力とする．図 4.5の Original text のうち，“for”, “inviting”, “last” をマスクする例を考える．“for” と “inviting” は連続するため 1 つにまとめて<X>というマスクトークンとし，“last” は<Y>というマスクトークンとする．T5 は <X>，<Y>を予測し，それ以外の “Thank”, “you”, “me”, “to”, “your”, “party”, “week”, “.” は単語をそのまま出力するのではなく，ターゲットトークン<X>，<Y>，<Z>として置き換えて出力する．

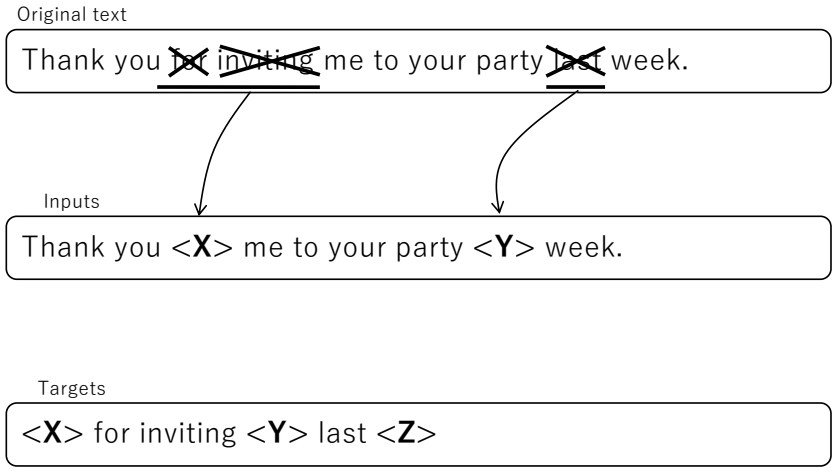


図 4.5　T5 における事前学習のマスクトークン：「Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer」の Fig.1 を元に筆者が手を加え作成した

4.3.2 DTW

本研究では，Web ニュースの記事においてテキストの単語アライメントをとるために Dynamic Time Warping(DTW) を使用する．DTW は，時系列データ同士の各点の距離を総当たりで求め，時系列データ間の最短ルートを求める．時系列データの対応する点を選ぶ際に重複を許すため，時系列データの長さが異なっても比較することが可能である．本研究ではテキストを単語区切りの時系列データとみなし，単語のベクトル表現を用いてデータの距離を測る．

本研究では Python ライブラリで提供されている fastDTW[27] を使用する．fastDTW は時系列同士の各点の距離をユークリッド距離で計算する．

長さが $|X|$ と $|Y|$ の二つの時系列データ X と Y があるとする．

$$X = x_1, x_2, ..., x_i, ..., x_{|X|} \tag{4.1}$$

$$Y = y_1, y_2, ..., y_j, ..., y_{|Y|} \tag{4.2}$$

W は X と Y の各点を結ぶパスである．

$$W = w_1, w_2, ..., w_K \qquad max(|X|, |Y|) \leq K < |X| + |Y| \tag{4.3}$$

K はパスの長さであり，パスの k 番目の要素を $w_k = (i, j)$ とする．i は時系列データ X のインデックス，j は時系列データ Y のインデックスである．パスは各時系列データの始点 $w_1 = (1, 1)$ から終点 $w_k = (|X|, |Y|)$ まで単調増加しながら距離を計算する．

$$w_k = (i, j), \quad w_{k+1} = (i', j') \quad i \leq i' \leq i + 1, \quad j \leq j' \leq j + 1 \tag{4.4}$$

時系列データを対応づける最短のパスは，

$$Dist(W) = \sum_{k=1}^{K} Dist(w_{ki}, w_{kj}) \tag{4.5}$$

$Dist(W)$ はパス W の距離で，$Dist(w_{k,i}, w_{k,j})$ はパスの k 番目の項目におけるデータ間の距離である．最短距離のパスを見つけるために，動的計画法が使われる．

4.3.3 Word Mover's Distance

Word Mover's Distance(WMD)[28] は単語埋め込み表現を用いて文書間の距離を測る手法である．一方の文のある単語からもう一方の文の別の単語への移動コストが最小となる距離を求める．

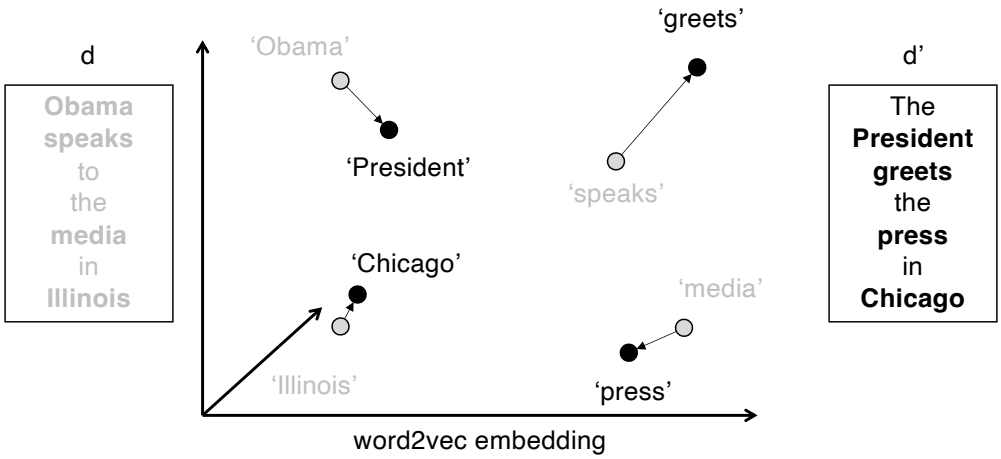


図 4.6　Word Mover’s Distance：「From Word Embeddings To Document Distances[28]」の Fig.1 を元に筆者が手を加え作成した

移動コストの計算は，単語ベクトルの距離を測る．図 4.6は文書 d（Obama speaks to the medhia in illinois.）から文書 d’（The President greets the press in Chicago.）へ変換するときの単語の対応づけを表す．ストップワードを除いた各文の単語同士を対応付けし，その移動コストの総和が最も小さくなるように文間距離を求める．

文間距離を測定する方法を説明する．Word2Vec モデルを用いて単語埋め込み表現を得る．

$$\mathbf{x}_w = (x_{w1}, x_{w2}, ..., x_{wn}) \tag{4.6}$$

ある単語 i と別の単語 j の距離は式 (4.7) となる．ここで，$c(i,j)$ を移動コストと呼ぶ．

$$c(i,j) = \|\mathbf{x}_i - \mathbf{x}_j\|_2 \tag{4.7}$$

次に，文の距離を測定する方法を説明する．2 つの文を nBOW(normalized bag-of-words) で表現したものを $\mathbf{d}$，$\mathbf{d}'$ とする．ベクトルの成分は正規化された出現頻度である．文書全体のコストは，それぞれの単語同士の移動コストの総和として計算する．

$$\sum_{i,j} \mathbf{T}_{ij}\, c(i,j) \tag{4.8}$$

ここで，$T_{ij} \geq 0$ は単語 i がどれだけ単語 j に移動したかを示す変換行列とする．行の成分の総和は，元の文の単語分布における各語の成分と一致する．

$$\sum_{j} \mathbf{T}_{ij} = d_i \tag{4.9}$$

同様に，各列の列の成分の総和は，移動先の文の単語分布における各語の成分と一致する．

$$\sum_{i} \mathbf{T}_{ij} = d'_j \tag{4.10}$$

2 つの文間の距離は，$\mathbf{d}$ から $\mathbf{d}'$ への全ての単語を移動するために必要な最小累積コストと定義することができる．

$$\min_{\mathbf{T}\geq 0} \sum_{i,j=1}^{n} \mathbf{T}_{ij}\, c(i,j) \tag{4.11}$$

$$\text{subject to:} \sum_{j=1}^{n} \mathbf{T}_{ij} = d_i \;\forall i \in \{1,...,n\} \tag{4.12}$$

$$\sum_{i=1}^{n} \mathbf{T}_{ij} = d';_j \;\forall j \in \{1,...,n\} \tag{4.13}$$

この最適化は，Earth Mover's Distance[40] という最適化問題の形式で解くことができる．

4.4 実験の概要

4.4.1 手順

本研究では，テキスト平易化モデルを学習するためのパラレルコーパスの構築方法を提案する．パラレルコーパスの資源として，ニュース記事の本文を使用する．平易なテキストとして NEWS WEB EASY[*1]の記事から，通常のテキストとして NHK NEWS WEB[*2]の記事からテキストを収集する．NEWS WEB EASY は，NHK NEWS WEB の記事を日本語学習者や小学生・中学生向けに，伝わりやすい表現に書き換えたものである．NEWS WEB EASY の Web ページから NHK NEWS WEB の Web ページへのリンクがあるため，記事の対応が取れている．通常のニュースの 1 つの文に一致する平易なニュースの単語列を抽出し，パラレルコーパスとすることを目的とする．4.4.2ではテキストのベクトル化について，4.4.3では DTW による単語の対応付けについて手順を述べる．

4.4.2 テキストのベクトル化

NHK NEWS WEB，NEWS WEB EASY のそれぞれの記事からテキストを抽出して分かち書きし，単語列を生成する．分かち書きには形態素解析ツールである MeCab を使用する．単語列を Word2Vec[33] で学習し単語ベクトルを得る．一般的に，Word2Vec モデルは単語を 100 次元から 300 次元のベクトルで表現する．この次元数はモデル学習時に設定することができ，次元数が高いほどベクトルの精度が上がるといわれている．本研究では，単語ベクトル同士の距離を測定する

[*1] https://www3.nhk.or.jp/news/easy/ 最終閲覧日：2023 年 10 月 10 日
[*2] https://www3.nhk.or.jp/news/ 最終閲覧日：2023 年 10 月 10 日

際に次元が高いほど計算量が増えてしまうことから，主成分分析（PCA）を適用して次元を削減する．

図 4.7は予備実験により単語ベクトルを主成分分析で次元圧縮したときの累積寄与率を示す．Word2Vec で 100 次元で表現したベクトルを，30 次元まで圧縮したとき累積寄与率は 94.8% となり，情報量として充分であるといえる．本研究では，単語ベクトルを 30 次元に圧縮したものを，単語間の距離を測定するために使用する．

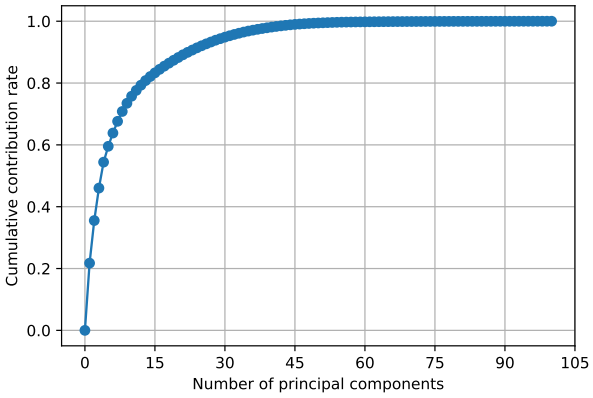


図 4.7　単語ベクトルを PCA で次元圧縮したときの累積寄与率

4.4.3　単語のアライメント

NEWS WEB EASY，NHK NEWS WEB から収集したテキストを分かち書きし，4.4.2の手法でベクトル化する．NEWS WEB EASY，NHK NEWS WEB の各ベクトルに対して，DTW を使って最短パスを求める．テキスト平易化が目的であるため，通常のニュース 1 文に対応する平易なニュースの単語列を抽出する．

4.5　実験 1: 平易化のためのパラレルコーパス構築

4.5.1　実験に使用するデータ

本実験では，2020 年 8 月から 2021 年 11 月の期間に収集したニュース記事をデータとして使用した．使用するテキストデータは本文のみとする．記事に含まれるタイトル，ルビ，記事の投稿日時，注目ワード，別の記事へのリンクテキストやバナー文字などはデータから除外する．収集した記事は 1,040 件で，記事に含まれる文の数は通常のニュースが平均 18.6 文，平易なニュースが平均 7.9 文であった．通常のニュース 1 文に対応する平易なニュースの部分的な単語列を抽出し，テ

キスト平易化のためのパラレルコーパスとした．

4.5.2 結果

生成したコーパスの文間類似度

自動生成したコーパスは 9,764 文対となった．パラレルコーパスの対応する文間類似度を DTW と WMD で測定した際の基本統計量を表に示す．

表 4.1 DTW と WMD の基本統計量

	DTW	WMD
mean	0.948236	0.776666
std	1.446837	0.194221
min	0.019100	0.071801
25%	**0.269925**	**0.638976**
50%	0.446100	0.762918
75%	0.908175	0.908817
max	13.393800	1.394344

ここで，DTW は単語アライメントを検出するためではなく，文間距離を測るために使用している．文間距離は式（4.5）で得られる．DTW の測定値は値の小さなデータが多く比較しにくく，値の大きなデータが突出していることで平均値を上げていたため，対数をとって WMD の文間距離と比較したところ相関係数は 0.83 であった．DTW の性質から，語順の変化が少ない書き換えの類似度測定に適していると考えられる．一方，WMD は語順を考慮しない類似度測定であることから DTW とは異なる指標であるが，構築したパラレルコーパスでは二つの指標に相関があるということがわかった．

文の分割・統合に対応したパラレルコーパスの抽出

表 4.2は，平易なニュースへの書き換えによって文の数が 5 から 6 に増加した例である．単語アライメントをとったところ，表 4.2の No.4 では通常文 1 文に対して平易文 2 文が対応している．書き換えによって文が分割されても，意味的に一致する単語列として 2 文を検出することに成功した．

表 4.3は，平易なニュースへの書き換えによって文の数が 5 から 2 に減少した例である．書き換えによって一部のテキストが削除され，対応する単語列が存在しない．実験の結果，表 4.3の No.1 から No.3 までの 3 つの通常文が，1 つの平易文に対応づけされた．また，No.4 と No.5 の 2 つの

通常文が，1 つの平易文に対応づけされている．複数文が統合され 1 文に書き換えられるケースでは，平易なニュース 1 文に対する通常のニュース複数文の対応を抽出すれば，文の統合を検出することが可能である．

表 4.4は，通常のニュースと平易なニュースのテキストの長さに差がある場合の結果である．テキストが大幅に削除された場合，通常のニュースの中盤の文に対応する単語列が見つからない結果になることが多い．表 4.4の No.6 から No.9 までは人手による平易化の際に文が削除されており，通常文に対応する平易文が存在しないが，実験結果ではいくつか部分的な単語列が対応づけられている．このことから，書き換え前後の文の長さに差があるとき，単語アライメントによる対応づけは有効ではないことがわかった．このときの DTW，WMD による計測値を確認すると，No.6 から No.9 までは文間距離が大きくなっており，パラレルコーパスとして相応しくないといえる．

表 4.2　生成したコーパス（文の分割を検出することに成功した例）

No.	通常のニュース	平易なニュース	DTW	WMD
1	アジア各地などとの往来を再び本格化させるのを前に、ビジネスで海外に渡航する人を対象としたＰＣＲ検査の予約システムが、9 月末にも導入される見通しになりました。	政府は中国やタイ、ベトナムなど１６の国や地域と「仕事でお互いに行ったり来たりできるようにしよう」と話しています。	0.25	0.78
2	政府は中国、タイ、ベトナムなど 16 の国や地域との間で、ビジネス関係者らの入国を相互に認めるための協議を進めていますが、こうした国や地域では、入国者に対して一定期間の隔離をしない代わりに、ＰＣＲ検査の結果が陰性である証明書の提出を求めるケースが多いということです。	１６の国と地域の多くでは、着いてからしばらくの間、決まった場所にいる代わりに、ＰＣＲ検査で新しいコロナウイルスがうつっていないことを証明する必要があります。	0.17	0.52
3	このため政府は、海外に渡航する前に速やかにＰＣＲ検査を受けられるよう、最寄りの医療機関を予約できるシステムを 9 月末にも導入する方向で準備を進めています。	このため政府は、仕事でこの１６の国や地域に行く前にＰＣＲ検査を受けることができるようにしたいと考えています。	0.21	0.47
4	経済産業省によりますと、ビジネスでの渡航者向けにＰＣＲ検査を実施できる人数は、現在全国で一日数百人程度にとどまり、本格的な往来の再開に向けて国内での検査体制の拡充が課題となります。	**家の近くの病院などで検査の予約ができるシステムを、９月の終わり頃までに作る予定です。政府は、たくさんの病院にこのシステムに参加してほしいと考えています。**	0.17	0.59
5	政府は予約システムに登録する医療機関を広く募集するとともに、空港内などに検査施設を設けることも検討することにしています。	空港などに検査ができる場所を作ることも考えています。	0.35	0.61

表 4.3　生成したコーパス（文の一部が省略され，対応する単語列が存在しない例）

No.	通常のニュース	平易なニュース	DTW	WMD
1	オーストラリア政府の研究機関は、温度が 20 度で暗い所にあるなど一定の限られた条件のもとでは、新型コロナウイルスは紙幣やガラスなどの表面で少なくとも 28 日間、感染力のある状態で存在するとする研究結果を発表しました。	オーストラリアの	3.14	1.02
2	この研究機関は、手洗いなどの徹底が重要だと指摘しています。	（対応する単語列が無い）	-	-
3	CSIRO ＝オーストラリア連邦科学産業研究機構は、一般的な物の表面に付いた新型コロナウイルスが、湿度 50 %の暗い環境でどれくらいの期間、残るのか実験で調べました。	研究所は、物に付いた新しいコロナウイルスが、湿度５０%の暗い所で、どのくらいの間うつる力があるか調べました。	0.14	0.44
4	その結果、温度が 20 度の場合、紙幣やガラスの表面では少なくとも 28 日間、ウイルスが感染力のある状態で存在することがわかったということです。	温度が２０℃の場合、	1.09	0.68
5	また温度が 30 度の場合は、紙幣では 21 日間、ステンレスやガラスでは 7 日間、綿では 3 日間、感染力のある状態でウイルスが存在したということです。	お札やガラスではいちばん短くて２８日の間、うつる力がありました。	0.32	0.57

表 4.4　生成したコーパス（文が大幅に削除され，対応する単語列が存在しない例）

No.	通常のニュース	平易なニュース	DTW	WMD
1	＜省略＞	＜省略＞		
2	しかし、受賞者が相次ぐ一方で、科学技術立国を支えると言われる日本の大学院の博士課程の学生の数は、修士課程から進学する学生が減り続け、文部科学省によりますと、ピーク時の平成 15 年度のおよそ 1 万 2 千人から、昨年度はほぼ半分の 5963 人まで減りました。	博士課程の学生は、いちばん多かった２００３年に約１万２０００人いましたが、去年は６０００人以下になりました。	0.18	0.62
3	また、人口 100 万人当たりの博士号取得者の数も、欧米が増加傾向にあるのに対し、日本は 2008 年度の 131 人から減少し、2017 年度には 119 人と、アメリカ、ドイツ、韓国の半分以下の水準にまで落ち込んでいます。	人口１００万人の中にいる博士の数は、アメリカやヨーロッパなどでは増えていますが、日本では減っています。いちばん多かった２００８年には１３１人でしたが、２０１７年には１１９人でした。	0.13	0.52
4	これについて、ノーベル化学賞を受賞した、大手化学メーカー旭化成の吉野彰さんは、博士号を取得しても将来のキャリアが不透明なままというのが重要な課題だと指摘しています。	これはアメリカやドイツ、韓国の半分以下です。	0.24	0.68
5	吉野さんは、欧米諸国などでは博士号を取得すると企業などでの就職が優位になる側面があるのに、日本では処遇がほぼ変わらないと指摘します。	吉野さんは「日本では博士になっても、会社の	0.42	0.68
6	そのうえで「企業は博士という学位を考慮し、それなりの待遇や給与で優遇することなどが必要ではないか。	給料など	*3.63*	*1.05*
7	産業界が博士課程を出た人をどう処遇するかが、これからの問題だ」と訴えます。	が	*8.92*	*1.12*
8	また大学の環境についても、若手が長期的に研究に打ち込めるようになっていないと指摘します。	ほとんど変わりません。	*1.73*	*1.04*
9	吉野さんがノーベル賞を受賞したリチウムイオン電池の研究を始めたのは吉野さんが 33 歳の時で、腰を据えて 1 つの研究を続けられたからこそ、30 年後に世界で評価される結果が出せたといいます。	大学でも	*2.72*	*1.14*
10	吉野さんは「大学の研究は、真理の探究、あるいは研究者自身の好奇心に基づきひたすら追い求めるもので、1 つのミッションとして絶対必要だと思います。	（対応する単語列がない）	-	-
11	そういった意味で、博士課程を経た人が 10 年間程度は安心して研究できる環境は、日本にとって非常に重要だ」と指摘しています。	（対応する単語列がない）	-	-
12	＜省略＞	＜省略＞		
13	そのため、専門知識を極めている博士を採用する動きが見られる」としたうえで、「海外では博士であればどこかの企業は絶対評価してくれるし、社会でも尊敬されるので、博士課程にも不安なく進める。	長い間研究することができません。	0.87	0.9
14	日本の博士についても、社会で価値を認識、評価することが重要で、その風潮を作っていきたい」と話していました。	博士になった人が１０年ぐらい安心して研究できるようにすることが大切です」と話しています。	0.28	0.68

4.6　実験 2: テキスト生成モデルへの応用

4.6.1　実験の概要

4.5で生成した平易化パラレルコーパスを用いて，テキスト平易化モデルを構築し学習する．生成した平易化パラレルコーパスの数が限定的であるため，あらかじめ語彙と文の関係性をとらえた事前学習済みモデルである T5 を使用し，平易化パラレルコーパスでファインチューニングする．T5 モデルで生成したテキストを 2章の手法で難易度評価，および 3章の手法で類似度評価して平易化モデルの有用性を測る．

4.6.2　実験に使用するデータおよび環境

平易化パラレルコーパス

4.5で生成したパラレルコーパスのうち，文対の類似度が高い上位 25% を抽出するため，表 4.1より閾値の設定を DTW が 0.26 以下，WMD が 0.63 以下とし，どちらも満たしている文対を選択する．1,426 件のコーパスが該当することから，学習用データを 1,327 件，検証用データを 70 件，テストデータを 29 件に割り当ててテキスト平易化モデルをファインチューニングする．

生成モデル

少ないコーパスから効率よくモデルを学習するため，事前学習済みモデルを使用する．事前学習は計算量が多く，非常にコストがかかるため，本研究では日本語 T5 事前学習済みモデルを使用する．モデルは Hugging Face で公開されている t5-base-japanese[*3]で，事前学習に使用した日本語テキストは次のとおりで約 100GB の規模である．

- Wikipedia の日本語ダンプデータ[*4](2020 年 7 月 6 日時点のもの)
- OSCAR の日本語コーパス[*5]
- CC-100 の日本語コーパス[*6]

難易度推定モデル

2.5の実験で構築した BERT による難易度推定モデルを使用する．BERT モデルはテキストを

[*3] https://huggingface.co/sonoisa/t5-base-japanese 最終閲覧日：2024 年 2 月 18 日
[*4] https://ja.wikipedia.org/wiki/ 最終閲覧日：2024 年 2 月 18 日
[*5] https://oscar-project.org 最終閲覧日：2024 年 2 月 18 日
[*6] https://data.statmt.org/cc-100/ 最終閲覧日：2024 年 2 月 18 日

入力すると難易度レベルが「平易」か「通常」かを推測する．

類似度測定に用いる指標

3章で検証した類似度指標のうち，Sentence-BERT を用いて生成モデルの出力文を評価する．生成モデルに入力する文と，モデルの出力文との類似度が高ければ，出力文の評価が高いと判断する．

4.6.3 結果

自動生成したパラレルコーパスを用いて平易化モデルをファインチューニングした結果を付録A.1に記載する．

難易度評価

平易化モデルの出力文を平易性について品質推定する．2章の難易度推定モデルで，生成モデルの出力文を難易度推定した結果を表 4.5に示す．出力文の 89.7%（29 文のうち 26 文）が「平易文」であるという結果になった．

表 4.5　難易度推定の結果

		実際の分類	
		入力文（通常）	出力文（平易）
推定値	平易	0	26
	通常	29	3

類似度評価

平易化モデルの出力文を類似性について品質推定する．3章では，平易化モデルのための類似度評価に最も適切な指標は Sentence-BERT であると結論づけたので，本章では類似度測定に Sentence-BERT を使用する．図 4.8は Sentence-BERT で入力文と出力文の類似度を測定した結果である．EJC0.7 コーパスの平易文と難解文の類似度測定結果のヒストグラムを図の上部に，T5 生成モデルの出力文と，元の入力文との類似度測定結果を図の下部に示す．類似度の分布を見ると，EJC0.7 コーパスより生成モデルの出力文の方が数値が高い．類似度の値は EJC0.7 の平均は 0.38，生成モデルの出力の平均は 0.86 で，2 標本による t 検定の結果，p 値は 0.05 以下で有意差があると言える．

平易化モデルの出力文を Sentence-BERT を用いて難易度推定したところ, 類似度の平均は 0.85, 最大値が 0.995，最小値が 0.692 となった．表 3.5より，SNOW T23 コーパスを Sentence-BERT

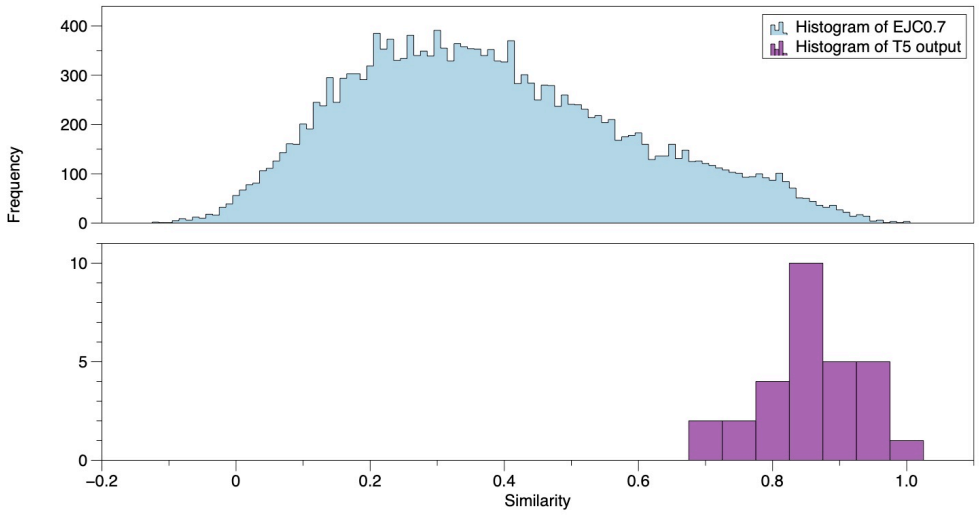


図 4.8　生成モデルの出力文評価（Sentence-BERT）

で類似度測定したときの平均が 0.815 であり，この値と比較しても平易化モデルの出力文が類似度が高い結果であった．

4.7　まとめ

DTW で単語アライメントを考慮した通常のニュースと平易なニュースからパラレルコーパスを自動生成する方法を提案し，ニュース記事を用いて平易化コーパスを生成した．先行研究では文と文の類似度が高い文対をコーパスとする研究が一般的であったが，本研究では書き換えによる文の統合・分割を考慮し，意味的に一致する単語列からコーパスを作ることに成功した．また，生成した平易化パラレルコーパスの中から類似度が高い上位 25% のコーパスを利用して，事前学習済みのテキスト生成モデルをファインチューニングして平易化モデルを構築した．平易化モデルが書き換えた出力文を難易度推定および類似度測定したところ，出力テキストのうち 89.7% が平易であると推定された．また，出力文と元のテキストとの類似度を Sentence-BERT で測定すると 0.856 であり，既存コーパスの SNOW T23 の平均値を上回る結果が得られた．

ただしコーパスの自動生成において，人手で書き換えたテキストの長さに差が大きい場合は上手く対応をとれず，生成したパラレルコーパスの文間距離が大きくなってしまう．今回の実験では文間距離が大きくなった場合は閾値を設定して平易化パラレルコーパスから除去しているが，テキストの長さに差がある場合や意味的に対応する部分が欠如している場合については今後の課題とする．

第 5 章

おわりに

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であるが，日本語におけるテキスト平易化のための十分な規模のパラレル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した．

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. テキスト平易化による書き換え前後の文の類似度を測定するために相応しい評価指標を調査する．
3. 平易化パラレル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した．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．その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．また，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，サ変接続名詞率，単語数，受身率，漢字率が重要な特徴量であることがわかった．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，モデルが重要な特徴量をとらえていることを示唆される結果が得られた．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，本実験で得られた新しい見解としては，テキスト平易化タスクにおいてはアテンションの第 6 層に受

身，第 11 層に単語数（文の長さ）が表現され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，重回帰分析を用いた先行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，平易化パラレル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いくつか存在しているが，本研究では平易化前後のテキストの意味的類似度を対象とする．したがって，単語や文法の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，文の類似度を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．実験に用いたデータセットは，JSICK，SNOW T23 である．類似度測定指標を説明変数，文の類似度を目的変数としたロジスティック回帰モデルを用いて説明力の高い指標を調べたところ，どの指標も類似度測定に有効であるが，最も有効なのは Sentence-BERT であることが明らかになった．

3 つめに，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ため，単語ベクトルを用いた．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という性質を利用して，文全体をベクトル列とみなし Dynamic Time Warping で時系列データ同士の最短経路を求め，対応する単語を見つける．つまり，文間距離が最も小さくなるときの単語の意味的な対応をもとに，一方の文に対応する単語列を抽出することで，文の区切りに依らないパラレルコーパスの生成を目指した．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．難易度推定では，出力文の 89.7% が平易であるという結果になった．また，入力文と出力文の類似度を測定したところ 0.86 となり，SNOW T23 の平均値 0.82 と比較しても高い値が得られた．

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．

今後の課題を述べる．一つめに，実用化に向けた被験者実験を行うことが挙げられる．テキスト平易化の目的は日本語学習者の初級者に対して伝わりやすい文章に書き換えることである．これにより日本語学習者が情報や教材をより効果的に吸収できるようになることを目指している．本研究の手法を実際に導入し，効果を検証するためにはさらなる実験が必要である．具体的には，システムを構築しテキスト平易化の手法を適用したアプリケーションを開発する．これを実際の日本語学習者に使用してもらい，彼らのフィードバックを収集することで，システムの効果や改善点を明らかにすることができる．

二つめに，日本語以外の言語への応用可能性を調査することを挙げる．同一言語上の言い換えにおいては文法的に語順の変化が少ないため，言い換えのための手法やモデルが異なる言語においても適用できる可能性がある．本研究では日本語のニュースからテキストを収集したが，他言語においてコーパス資源がある場合に手法の適用可能性が期待できる．

本研究の成果が言語学習やコミュニケーションの向上に貢献できることを目指して，実現に向けて研究を継続していきたい．

謝辞

この学位論文を完成させるにあたり，多くの方々に支えられ，ご協力いただきました．心から感謝申し上げます．

はじめに，指導教員である村尾 元 教授には，熱心で的確なご指導を賜りました．私が未熟な状態で研究の第一歩を踏み出した時，快く相談に乗ってくださいました．研究テーマの決定から具体的なアプローチの立て方まで，的確なアドバイスが頼りになりました．先生のおかげで自らの研究に自信を持ち，継続することができました．心から感謝申し上げます．

また，情報コミュニケーションコースの大月 一弘 教授，康 敏 教授，清光 英成 教授，西田 健志 准教授，大山 牧子 准教授には，日常の議論や助言を通じて，大きな助けをいただきました．先生方のご尽力に感謝いたします．

キャンパスライフ支援センターのみなさまにも感謝いたします．センターで様々な活動に携われたことが私の大学生活をより豊かなものにしました．また，研究に対する温かい応援も励みとなりました．

大学教育推進機構 大学教育研究センターのみなさまにも心から感謝いたします．リサーチ・アシスタントとして従事した経験は，大学教育において新たな視点を得る貴重な機会でした．

研究室で一緒に過ごした仲間にも感謝します．研究室での 5 年間，みなさまと楽しい時間を過ごせたおかげで充実した学生生活を送ることができました．

私の研究活動を支えてくれた家族にも心から感謝します．家族の理解と協力があってこそ，この論文を完成させることができました．

最後に，応援してくださった全ての方々に心から感謝いたします．皆様のおかげでこの研究が進行でき，論文が完成に至りました．

心からの感謝の意をこめて，謝辞とさせていただきます．

研究業績

学術誌（査読付き）

[1] Eri Maekawa, Hajime Murao, “Interpreting BERT Attention Trained for Japanese Difficulty Classification from the Viewpoint of Grammatical Features”, ICIC Express Letters, Part B: Applications, 13 7, pp. 697-703, 2022.

[2] Eri Maekawa, Hajime Murao, “Measuring Sentence Similarity for Text Simplification: Identifying Effective Metrics”, ICIC Express Letters, Part B: Applications (掲載予定)

口頭発表（査読付き）

[3] Eri Maekawa, Hajime Murao, “A Proposal to Create a Pseudo-Parallel Text Corpus for Simplifying Japanese using DTW”, Proceedings of 17th International Technology, Education and Development(INTED2023) Conference, pp.6542-6550, Valencia, Spain, 2023.

[4] Eri Maekawa, Hajime Murao, “Analysis of the Behavior of Foreign Tourists Using Mobile Translation Devices” Proceedings of SICE2020, pp. 1038-1042, Online, 2020.

[5] Eri Maekawa, Hajime Murao, “Measuring Sentence Similarity for Text Simplification: Identifying Effective Metrics”, Proceedings of The 17th International Conference on Innovative Computing, Information and Control (ICICIC2023), Kumamoto, 2023.

[6] Eri Maekawa, Hajime Murao, “The Comparison of Word Embeddings and Feature Vectors in Text Classification by Difficulty Level” The 15th International Conference on Innovative Computing, Information and Control (ICICIC2021), Online, September 2021.

口頭発表（査読なし）

[7] 前川 絵吏，村尾 元，“日本語の難易度に関する特徴分析”，言語処理学会第 27 回年次大会発表論文集，言語処理学会，北九州国際会議場（オンライン開催），3 月，2021 年.

参考文献

[1] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a method for automatic evaluation of machine translation. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*, pp. 311–318, 2002.

[2] Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. Optimizing Statistical Machine Translation for Text Simplification. *Transactions of the Association for Computational Linguistics*, Vol. 4, pp. 401–415, 2016.

[3] Elior Sulem, Omri Abend, and Ari Rappoport. BLEU is not suitable for the evaluation of text simplification. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 738–744, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[4] Wei Xu, Chris Callison-Burch, and Courtney Napoles. Problems in current text simplification research: New data can help. *Transactions of the Association for Computational Linguistics*, Vol. 3, pp. 283–297, 2015.

[5] William Coster and David Kauchak. Simple English Wikipedia: A new text simplification task. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 665–669, Portland, Oregon, USA, June 2011. Association for Computational Linguistics.

[6] A. Katsuta and K. Yamamoto. Crowdsourced Corpus of Sentence Simplification with Core Vocabulary. In *The 11th International Conference on Language Resources and Evaluation (LREC 2018)*, pp. pp.461–466, 2018.

[7] 畠垣光希, 梶原智之, 二宮崇. やさしい日本語へのテキスト平易化のための訓練データの精選. 第 21 巻, pp. 293–300, 2022.

[8] Zhemin Zhu, Delphine Bernhard, and Iryna Gurevych. A monolingual tree-based translation model for sentence simplification. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, p. 1353–1361, USA, 2010. Association for Computational Linguistics.

[9] Tomoyuki Kajiwara and Mamoru Komachi. 平易なコーパスを用いないテキスト平易化. 自然言語処理, Vol. 25, No. 2, pp. 223–249, 2018.

[10] Xingxing Zhang and Mirella Lapata. Sentence simplification with deep reinforcement learning. In Martha Palmer, Rebecca Hwa, and Sebastian Riedel, editors, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 584–594, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[11] Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. Data-Driven Sentence Simplification: Survey and Benchmark. *Computational Linguistics*, Vol. 46, No. 1, pp. 135–187, 03 2020.

[12] Sanja Stajner, Maja Popovic, Horacio Saggion, Lucia Specia, and Mark Fishel. Shared task on quality assessment for text simplification. pp. 22–31, 05 2016.

[13] Tomoyuki Kajiwara and Atsushi Fujita. Semantic features based on word alignments for estimating quality of text simplification. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pp. 109–115, Taipei, Taiwan, November 2017. Asian Federation of Natural Language Processing.

[14] 廣中勇希, 井川朋樹, 梶原智之, 二宮崇. Qents：テキスト平易化の品質推定のためのデータセット. pp. 516–520, 2022.

[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *NAACL-HLT*, 2019.

[16] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. *In Advances in neural information processing systems*, pp. 5998–6008, 2017.

[17] 川村よし子, 北村達也. 日本語学習者のための文章の難易度判定システムの構築と運用実験. *Journal CAJLE*, pp. 18–30, 2013.

[18] 李在縞. 日本語教育のための文章難易度に関する研究. 早稲田日本語教育学 第 21 号, Vol. 21,

pp. 1–16, 2016.

[19] Hasebe Yoichiro and Lee Jae-Ho. Introducing a readability evaluation system for Japanese language education. *Proceedings of the 6th international conference on computer assisted systems for teaching & learning Japanese*, pp. 19—22, 2015.

[20] 劉志宇, 内田理. 日本語を学習する外国人を対象とした日本語テキスト難易度推定手法. Technical Report 11, 東海大学大学院工学研究科情報理工学専攻, 東海大学情報理工学部情報科学科, jan 2012.

[21] 張萌, 伊藤彰則, 佐藤和之. 「やさしい日本語」 作成支援のための日本語の難易度自動推定の検討. 研究報告自然言語処理 (NL), Vol. 2012, pp. 1—6, 2012.

[22] 石井愛, 小松祐城, 脇森浩志. 予測根拠として解釈性の高いアテンションの選択. 言語処理学会 第 26 回年次大会 発表論文集, 2020.

[23] Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C Wallace. ERASER: A Benchmark to Evaluate Rationalized NLP Models. *arXiv*, 2019.

[24] 前川絵吏, 村尾元. 日本語の難易度に関する分析. 言語処理学会年次大会発表論文集, 第 27 巻, 2021.

[25] Aaron Fisher, Cynthia Rudin, and Francesca Dominici. All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously. *Journal of Machine Learning Research 20 (177), 1-81, 2019*, 2018.

[26] Hideki Tanaka, Tadashi Kumano, Isao Goto, and Hideya Mino. やさしい日本語ニュースの制作支援システム. 自然言語処理, Vol. 25, No. 1, pp. 81–117, 2018.

[27] Stan Salvador and Philip K. Chan. FastDTW: Toward Accurate Dynamic Time Warping in Linear Time and Space. *Intelligent Data Analysis*, Vol. 11, No. 5, pp. 561–580, 2006.

[28] Matt J. Kusner, Yu Sun, Nicholas I. Kolkin, and Kilian Q. Weinberger. From Word Embeddings to Document Distances. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, p. 957–966. JMLR.org, 2015.

[29] Dinghan Shen, Guoyin Wang, Wenlin Wang, Martin Renqiang Min, Qinliang Su, Yizhe Zhang, Chunyuan Li, Ricardo Henao, and Lawrence Carin. Baseline needs more love:

On simple word-embedding-based models and associated pooling mechanisms. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 440–450, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[30] Quoc V. Le and Tomas Mikolov. Distributed representations of sentences and documents, 2014.

[31] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3982–3992, 2019.

[32] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. BERTScore: Evaluating Text Generation with BERT. *arXiv*, 2019.

[33] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space. *InProceedings of Workshop at the International Conferenceon Learning Representations*, 2013.

[34] Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. A SICK cure for the evaluation of compositional distributional semantic models. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pp. 216–223, Reykjavik, Iceland, May 2014. European Language Resources Association (ELRA).

[35] Regina Barzilay and Noemie Elhadad. Sentence alignment for monolingual comparable corpora. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, pp. 25–32, 2003.

[36] Vladimir Iosifovich Levenshtein. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, Vol. 10, No. 8, pp. 707–710, 2 1966. Doklady Akademii Nauk SSSR, V163 No4 845-848 1965.

[37] 国立国語研究所. 現代日本語書き言葉均衡コーパス (bccwj), 2011.

[38] 梶原智之, 小町守. Simple ppdb: Japanese. 言語処理学会第 23 回年次大会発表論文集, pp.

529–53–2, 2017.

[39] Takumi Maruyama and Kazuhide Yamamoto. Extremely Low Resource Text simplification with Pre-trained Transformer Language Model. *2019 International Conference on Asian Language Processing (IALP)*, Vol. 00, pp. 53–58, 2019.

[40] Yossi Rubner, Carlo Tomasi, and Leonidas J. Guibas. The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, Vol. 40, No. 2, pp. 99–121, 2000.

付録 A

A.1　生成モデルの出力テキスト

表 A.1: テキスト生成の評価

No.	上段から平易なテキスト，生成文，通常のテキスト
1	団体は「ニューヨークでもアジア系の住民に対するヘイトクライムが続いています。同じアジア系として、みんなで助け合いたいです」と話しました
	ニューヨークではアジアの人たちがたたかれたりする事件が起こっています。イベントを行った団体の人は「アジアの人たちと一緒に頑張ろうというメッセージを伝えたいです」と話しています
	主催した団体では、ニューヨークでもアジア系住民に対するヘイトクライムとみられる事件が相次いでいることから、「同じアジア系として連帯するメッセージを伝えたい」としています。
2	新しいコロナウイルスが終息することを願って、山梨県甲州市で桃の収穫が始まりました
	フルーツ山梨は、新しいコロナウイルスが早くなくなることを願って、甲州市で「ヨゲンノトリ」という鳥の絵をつけた桃を作りました
	新型コロナウイルスの終息の願いを込め、江戸時代にコレラを予言したとしてなどで話題となっている「ヨゲンノトリ」の図柄を入れた桃の収穫が、山梨県甲州市で行われています。
3	新しいコロナウイルスが広がっているため、官房長官は 24 日、イギリスから出ることができないようにします。また、新しいコロナウイルスが広がらないようにするために、新しいコロナウイルスが広がらないようにすると言いました
	イギリスでは、今までよりうつる力が強くなった新しいコロナウイルスが広がっています。このため、ヨーロッパなど多くの国がイギリスからの飛行機を止めています。日本も 24 日から、イギリスから日本
	イギリスで感染力が強いとされる変異した新型コロナウイルスが広がっていることを受けて、○○官房長官は、24 日以降、当分の間、イギリスからの新規の入国を拒否するなど、入国制限を強化する新たな措置を実施すると明らかにしました。
4	によると、今月 3 日から 7 日までにウイルスがうつった人は全国で 1100 人ぐらいいました。アプリに登録した
	入れた数は 610 万ありました。しかし、3 日から 7 日までにコロナウイルスがうつったことがわかった人は 1100 人以上いたのに、
	のまとめによりますと、今月 3 日から 7 日までに感染が確認された人は全国で 1100 人余りに上っていて、感染をアプリに登録した人の割合は 0.3% 以下となっています。
5	ました。避難所に食料やマスクなどを届けた女性は「まだ、マスクなど足りないので、早く物資を届けたいです」と話しました
	」と話しました。避難してきた人にマスクなどを届けた女性は「避難してきた人のほとんどがマスクを持っていないし、役所にも十分

No.	上段から平易なテキスト，生成文，通常のテキスト
	また、避難所に食料やマスクなどの物資を届けた女性スタッフは、「感染対策のハンドブックを持参し配布しているが、被災者の多くがマスクを持っておらず行政の備蓄も不足しているため、一刻も早く物資を届ける必要がある」と述べました。
6	○○さんには大変申し訳なかったです」と話しました
	○○さんには本当に申し訳なかったと話しました
	○○さんには大変申し訳なかった」と話しました。
7	駅のホームには、農家や農協の人がたくさん集まって、約 2000 個の干し柿をつるしました
	作っています。7 日、農家の人たちが駅に集まって、ホームに柿を
	7 日はかみのやま温泉駅の下りホームに農家や農協の職員などが集まり、およそ 2000 個の干し柿をつるす作業を行いました。
8	地球と火星は 2 年 2 か月に 1 回、近づきます。地球の周りを回る周期が違うため、地球と火星は 2 年 2 か月に 1 回、近づきます
	回るのにかかる時間が違うため、地球と火星は近くなったり遠くなったりします。火星は、6 日午後 11 時すぎに地球にいちばん
	火星は地球の隣にある惑星で、太陽のまわりを回る周期の違いから地球と火星はおよそ 2 年 2 か月に 1 回、接近します。
9	生徒たちは去年卒業しています。こん棒を作った人は「選手のことを考えて、使いやすいように作りました
	持つところの形を 4 種類作りました。こん棒を作った○○○○さんは「選手のことをいちばんに考えて使いやすいように
	生徒たちの多くは去年、卒業していて、こん棒を製作した 1 人の○○○○さんは「選手のことをいちばんに考え使いやすいように作りました。
10	た。今年 4 月以降仕事がなくなった人の中で、今月の時点で再就職している人は男性で 24.1%、女性で 38
	で、女性は男性の 1.4 倍になりました。このうち 4000 人の 10 月の給料を調べました。新しいウイルスが広がる前と
	また、ことし 4 月以降仕事を失った人のうち先月の時点で再就職していない人は男性で 24.1% 女性で 38.5% と、女性が男性のおよそ 1.6 倍になっています。
11	県によると、今シーズンは 2 トンの輸出を見込んでいて、来月からベトナムのスーパーなどで売る予定です
	和歌山県によると、全部で 2t のみかんを輸出する予定で、12 月からホーチミンのスーパーなどで売ります。インターネットなどを使って、ベトナムの人にみかんを紹介します
	県によりますと、今シーズンは 2 トンの輸出を見込んでいて、来月からベトナム南部の都市、ホーチミンのスーパーなどで販売されるほか、やフリーペーパーを使って活動を行うということです。
12	アメリカと中国の貿易がうまくいっていないため、アメリカが中国からお金をもらうときの手数料が低くなっています。このため、アメリカが中国からお金をもらうときの手数料が高くなっています
	考えて決められています。今のルールでは、先進国に払う手数料が安くなっています。しかし、中国からアメリカに払う手数料が安く
	しかし、米中の貿易摩擦を背景にアメリカが中国から支払われる手数料の水準が低く問題だと指摘したことから国際ルールが見直され、各国の手数料が引き上げられることになりました。
13	ました。病院は、手術を受けた人の 69% で、ほかの人の 6.7% より 12 倍も亡くなっていました。研究グループは、手術の日を 69% にして、ほかの人の 56%
	て、約 5 万 5000 人が亡くなっていました。手術を受けた人の年齢や病気などを
	患者の年齢や持病などを踏まえて死亡率を解析したところ、外科医の誕生日に手術を受けた患者では 69% で、それ以外の患者の 5.6% と比べ、およそ 12 倍になっていたとしています。
14	北海道では、札幌市でウイルスがうつった人が 100 人以上見つかりました。このほかの県でも、クラスター=ウイルスがうつった人の集団がたくさんできています
	北海道は 17 日、札幌市に住んでいる人に、ウイルスがうつる危険がある場合は、できるだけ出かけないでくださいと言い
	北海道では、札幌市での感染確認が連日 100 人を超え、高止まりとなっているほか、札幌市以外の地域でもクラスター=感染者の集団が相次いで発生するなど、感染拡大に歯止めがかからない状況が続いています。

No.	上段から平易なテキスト，生成文，通常のテキスト
15	千葉市消防局は「ドアや窓を閉めた車の中などで、可燃性ガスが入ったスプレーを使うと、火事になる危険があります
	火が高く上がりました。千葉市消防局は「窓を閉めたままの車の中で、スプレーを使ったあとライターなどをつけると、スプレー
	千葉市消防局は「ドアや窓を閉めきった車の中などで可燃性ガスを含むスプレーを使用し、その後ライターなどを使うと、爆発が起きるおそれがある。
16	厚生労働省は、妊婦にワクチンを注射する計画を立てたとき、胎児や本人への影響を調べるデータが足りなかったことから、優先的に注射していました
	厚生労働省は、妊娠している女性に先に、ワクチンを注射するように、市や町などに言いました
	妊婦へのワクチン接種について、厚生労働省は国内で接種を始めた当初、胎児や本人への影響に関するデータが不足していたことなどから、接種を受けることを努力義務とせず、優先接種の対象にしていませんでした。
17	仕事を続けることができないので、国の支援が必要です」と話しています
	は少なくなっています。みんなのために一生懸命働こうという気持ちだけでは、仕事を続けることが難しくなって
	責任感や使命感だけでは働き続けることが難しい状況になっていて、国の支援が必要だ」と話しています。
18	○○選手はで「死ね」や「死ね」などのことばや、差別的なことばが書かれたメッセージを書きました。○○選手は「日本には人種差別がないと言っている人がいるけど、こうやって差別的なことを言っている人がいます
	2 人の父はアフリカのベナン出身で、母は日本人です。○○さんは「日本では人種で差別することはないと言う人がいます
	○○選手はで「死ね」などのことばのほか、差別的なことばが記されたメッセージを掲載したうえで、「日本には人種差別が無いと言っている人がいるけどこうやって人種差別発言をする人がいます。
19	岐阜県は 25 日、10 年前に県の小学生からもらったはがきのうち 600 枚をなくしました。岐阜県は 25 日、10 年後の自分と環境のために頑張ると誓ったはがきのうち 600 枚をなくしたと言いました
	岐阜県は 2011 年、小学校 5 年生の子ども 8650 人に、10 年後の自分に送るはがきを書いてもらいました
	岐阜県は 25 日、10 年前に県内の小学生から募集し、環境に優しい取り組みなどを 10 年後の自分に誓ったはがき、およそ 600 枚を紛失したと発表しました。
20	ノーベル賞を取るか、日本の大学院でノーベル賞を取るかが注目されています。日本の大学院で博士課程に入る人は、ピークの平成 15 年度から減って、去年度はほとんど半分になっていました
	日本人は去年まで 2 年続けてノーベル賞をもらっています。しかし、日本では大学院の博士課程で研究する学生が少なくなっています。博士課程の学生は、いちばん多かった 2003 年に約
	3 年連続での日本人の受賞となるか注目されますが、科学技術立国を支えると言われる日本の大学院の博士課程の学生の数は、修士課程から進学する人の数がピーク時の平成 15 年度から減り続け、昨年度はほぼ半分となっていて、ノーベル賞の受賞者からも対策を求める声があがっています。
21	救急や心臓の専門家が作る蘇生ガイドラインでは、パッドの表記を「小学生大人用」、「未就学児用」にしました
	このため、今までの書き方を変えて「未就学児用」と「小学生大人用」と書くことになりました
	これを受けて、救急や心臓の専門家が作る蘇生ガイドラインでは、パッドの表記を「小学生大人用」、「未就学児用」と変更し、を販売するメーカーも順次対応することにしています。
22	前線は来週にかけて日本付近に停滞します。西日本では雨がもっと多くなりそうです。東日本や北日本でも大雨になるかもしれません
	と、来週ごろまで日本の近くに前線があって、日本中で雨がとてもたくさん降る可能性があります
	前線は来週にかけて日本付近に停滞し、西日本ではさらに雨量が増えるほか、東日本や北日本でも大雨となるおそれがあります。
23	研究グループは、隣のケージにいるハムスターにどれだけ飛まつがうつるか調べました。2 日後の朝、隣のケージで飛まつがうつっていたのは野生では 1 匹もいませんでした。しかし、ヨーロッパ系統では 8 匹中 5 匹
	実験をしました。今世界中で広がっているのは、変化したウイルスです。グループは、2 つのウイルスが、隣のおりにいるハムスターにどの

No.	上段から平易なテキスト，生成文，通常のテキスト
	研究グループが飛まつを通じて隣接する別のケージにいるハムスターにどれだけ感染させるか調べたところ、2日後の時点で、隣のケージで感染していたのは野生型では1匹もいなかったのに対し、ヨーロッパ系統では8匹中5匹に上り、飛まつ感染しやすい性質があると考えられるとしています。
24	パッドなどには、小学校入学前までの子どもが使います。小学生は「成人用」を、小学生は「小児用」を使います
	1つは「成人用」です。「成人」は大人という意味ですが、小学生には「成人用」を使う必要があります
	パッドなどには「成人用」と電気ショックのエネルギー量を減らした「小児用」の2種類がありますが「小児用」の対象は小学校入学前までの子どもで、小学生には「成人用」を使う必要があります。
25	このボックスを作ったのは、派遣社員として働きながら副業で学生服のリユースショップを始めた○○○○さんです
	箱を置いた○○○○さんは、3人の子どものお母さんで、派遣社員で働きながら制服をリサイクルする店の仕事をしています
	回収ボックスを設けたのは派遣社員として働きながら副業で学生服のリユースショップを始めた○○○○さんです。
26	回収ボックスを設けたのは派遣社員として働きながら副業で学生服のリユースショップを始めた木村彩子さんです。
	で、最も少なくなりました。環境省は、工場などで仕事が少なくなってエネルギーをあまり使わなくなったことや、太陽の光などのエネルギーを利用することが増え
	これについて環境省は、鉄鋼や機械などの製造業で生産量が減少し、エネルギーの消費量が減ったことや太陽光や地熱といった再生可能エネルギーの利用が拡大したことなどを理由に挙げています。
27	ロシアの北極では、数年前から大きな穴がたくさん見つかっています。周りの人に迷惑にならないか心配です
	ロシアのヤマル半島やシベリアなど凍った土が広がる所では、2014年から大きな穴が
	ロシアの北極圏の永久凍土では、数年前から巨大な穴が相次いで確認され、周囲への影響が懸念されています。
28	ウイルスについて質問された○○会長は「変異したウイルスは、新しい株に取って代わられる準備ができているので、早晩、変異株が主流になると思っておいてください」と答えました
	わかりました。政府の専門家の会議の○○会長は、10日の国会で「変化したウイルスが広がり始めています。今までのウイルスの代わりに、変化したウイルスが多くなると考えたほうがいい
	この中で、政府の分科会の○○会長は変異ウイルスについて「間違いなく既存株に取って代わるプロセスが始まっていて、早晩、変異株が主流になると考えておいたほうがいい」と指摘しました。
29	成田空港からの移動手段についての議論が続いています。空港に乗り入れる鉄道の会社は、外国人が使うことができるようにしたいと考えています
	ています。成田空港の会社は、外国から着いた人が電車に乗ることができるように、特別な車両を用意することも
	成田空港からの移動手段をめぐっては、このほか、空港に乗り入れる鉄道に入国者専用の車両を設けることも検討されています。