BERTを用いたグレーディングモデル開発のための予備調査

岡 翼[1,a)]　吉良 伊織[1,b)]　中野 明[1,c)]

概要：日本国内における英文リーダビリティ研究は，教育利用の目的の観点から，英語学習者にとっての英文の理解のしやすさに注目して研究が進められている．代表的な研究に，OFYL(Ozasa-Fukui Year Level)があり，このリーダビリティ指標では，「単語の数」や「音節数」の他に，英語を母国語とする人を対象とするリーダビリティ指標には無い「単語の難易度」，「熟語の難易度」を説明変数に含めている．このリーダビリティを用いることで，例えば，日本の英語学習者が自分のレベルに合った英文を自動的に選択できるようになる．しかしながら，このリーダビリティ指標では不定詞や関係代名詞などの文法的な観点を考慮できていない．本研究では，Attention 機構を有する Transformer モデルである BERT を用いて，より文法的な視点を含めて英文の学年を予測するモデルである「グレーディングモデル」を開発するための予備調査を行う．この調査によって，BERT を用いて関係代名詞の用法の that が高精度で抽出できることが明らかとなった．

1. はじめに

英語は世界中で使用されている言語であり，日本でも近年のグローバル化を背景に，英語教育のスタートを年少化するなど，英語教育に対する注力が盛んとなっている．英語のスキルの中でもとりわけ重要なリーディング能力を養う学習方法として，多読と呼ばれる方法がある．多読は，文章を分析しないで大意を把握する読書法である [1]．この多読においては，自分のレベルに適した文章を選ぶことが重要となる．しかし，学習者自身がその判断を行うのは容易ではないといった問題を抱えている．この問題の解決方法として，計算機による文章の自動判定を行う研究が進められている．

この文書分類には英語の読みやすさを表すリーダビリティが役立つ．リーダビリティ研究は国内外を問わず行われており，国外の研究であれば ARI(Automated Readability Index) や Fog Count，Flesch Reading Ease Formula が挙げられ [2]，国内の研究では OFYL(Ozasa-Fukui Year Level)[3][4] や YL(Yomiyasusa Level)[5] がある．これらの研究を比較すると，第一言語や英語の教育環境によって，リーダビリティを算出する際に検討する部分が異なったものとなっている．本研究では，日本語を母語とする英語学習者（以下，「日本の EFL[*1]学習者」と呼ぶ）に適したリーダビリティ指標である，OFYL に注目する．

OFYL は日本の EFL 学習者向けのリーダビリティ指標である．OFYL の Ver. 3.2nhnc によると，文あたりの単語数，単語あたりの音節数，単語の難易度，熟語の難易度を説明変数とし，文が教科書で登場する学年を目的変数として重回帰分析を行っている．このリーダビリティ指標を作るために，事前に文中に出現する単語の難易度，熟語の難易度を英語の専門家によって吟味する必要がある．しかし，英語のカリキュラムが年々変化していく中，その都度英語の専門家によって単語や熟語，英文の難易度を判定しなおすのは容易でないといった問題点を抱えている．

2. 英文リーダビリティ算出に関する先行研究

日本における英語教育の状況が考慮されたリーダビリティに関する先行研究について論じる．

2.1 OFYL

OFYL(Ozasa-Fukui Year Level) は日本国内における英語教育の状況を考慮されたリーダビリティ指標であり，単語数，音節数の他に単語の難易度や熟語の難易度を説明変数とし，教師データとなる教科書の文の出現学年を目的変数として重回帰分析を行っている．単語の難易度や熟語の難易度の評価を行うために，外部の大学受験用の単語帳

[1] 久留米工業高等専門学校
National Institute of Technology, Kurume College, Kurume, Fukuoka 830–8555, Japan
[a)] s59208to@kurume.kosen-ac.jp
[b)] s59211ik@kurume.kosen-ac.jp
[c)] nakano@kurume-nct.ac.jp

[*1] EFL は「English as a Foreign Language」の略であり，日本人にとっての英語のような，「外国語としての英語」を指す．なお，「第二言語としての英語」は ESL(English as a Second Language) と呼ばれる．

や熟語のリストに依存している．(Ozasa, T et al., 2014) における英文の OFYL の定義式は次のようになっている．式 (1) は式 (2) の指数 Diff の式である．式中の Words/S は文あたりの単語数，Syllables/W は単語あたりの音節数，WordDiff/W は単語の難易度，IdiomDiff/S は熟語の難易度を表す．

$$\begin{aligned}\text{Diff} = {} & 0.0863\text{Words/S} + 0.2943\text{Syllables/W} \\ & + 0.6332\text{WordDiff/W} + 0.0665\text{IdiomDiff/S} \\ & + 0.5366\end{aligned} \quad (1)$$

$$\text{OFYL} = 3.8593/(1 + 766.9372\exp(-2.5709\text{Diff})) + 0.9 \quad (2)$$

それまで，音節数や単語数などのような単純な要素のみによって求められていたリーダビリティ指標で，英語を母語とする人に向けた指標であり，日本の EFL 学習者に適合するリーダビリティ指標ではなかった．OFYL では音節数や単語数に加えて，更に単語の難易度や熟語の難易度を説明変数に加えることで，日本における英語教育の状況を反映してより日本の EFL 学習者に適合したリーダビリティ指標となっていることが示されている．

2.2 グレーディングモジュール

OFYL が持つ上記した問題点を改善するために，中野（2022）では，AI 技術を用いたリーダビリティ評価モジュール作成の為の試作として，OFYL の説明変数に加えて，「to 不定詞の数」「関係代名詞の数」などを更なる説明変数とし，文の OFYL を教師データとして SVM(Support Vector Machine) を用いることでグレーディングモジュールを作成した．OFYL の分類結果と異なる学年の分類となった英文はどれも to 不定詞を含んだ文であり，OFYL とは異なる新たな視点からの分類も確認している．

3. 研究目的

OFYL を算出するためには，英語の専門家による多大な労力が必要である．本研究では，専門家による手動的な判断を要しないリーダビリティ算出器を作成することを目標としている．

4. BERT の概要

BERT（Bidirectional Encoder Representations from Transformers）[6] は，Google が 2018 年に発表した Transformer [7] のアーキテクチャを用いた大規模言語モデルであり，名前が示すように，双方向的な文脈を考慮した単語の理解が可能である．BERT は，Transformer の構造の中で Encoder 部分のみを用いている．

本研究で用いる BERT の事前学習モデルには，Huggingface 社が提供するオープンソースのライブラリである Transformers で実装・公開されている「bert-base-uncased」を採用する [8]．この事前学習モデルは, (Devlin et al., 2018) 中の $\text{BERT}_{\text{BASE}}$ のアーキテクチャのモデルであり，層の数は $L = 12$, self-attention head の数は $A = 12$，また，隠れ層の次元は $H = 768$ である．以降も，この変数を用いるとする．

BERT は，多段の Encoder により構成されているため，入力された各トークンに対し，埋め込みベクトルを出力として得ることができる．BERT を具体的なタスクに適応させる為には，タスクに応じて BERT の出力層を取り換え教師データを用いて追加の学習（Fine-tuning）を行う必要がある．この学習のことを，以下「fine-tuning」と呼ぶ．

5. グレーディングモデル

本研究で開発を目指す英文に対応する学年を予測するモデルのことを以下「グレーディングモデル（GM）」と呼ぶ．本章では，この GM の根幹をなす BERT についての解説と，その学習に用いる教師データのセットについて詳述する．

5.1 BERT による学年推定

Transformer で用いられている Attention 機構では，遠く離れた単語間の関係を捉えることができる．これによって，著者らは，関係代名詞や不定詞などの文法の構造や係り受け関係も捉えることができると考えている為，Transformer モデルである BERT を採用した．

本研究で検証するタスクにおいては，各トークンの情報がより重要となる．その為，1 つの文を代表するベクトルを構成するにあたって，各トークンで得られた出力を用いた構成をとるのが望ましいと考えられる．したがって，BERT に入力した各トークンに対する BERT による出力を用いて，最終的な出力として英文のリーダビリティの予測値を得るような構成にする必要がある．その構成の一例として，各トークンの出力に対し，それらの平均を取る average-pooling を行い，その出力を線形層に入力してリーダビリティの予測値を得る構成がある．

5.2 教師データ

OFYL では，高校で学ぶ英文についても検討されているが，本研究では，中学校の英文のみ検討する．本研究では，中学校英語教科書の本文を収集し，4 単語未満の文は前処理として取り除く（この前処理された本文のみのデータを以降「文データ」と呼ぶ）．文データの各文に学年のラベル付けを行い，それを教師データとして利用する．教師データの作成に用いた教科書は，「NEW CROWN（2020 年版）」の 1, 2, 3 年生版である．教師データの学年あたりの数は表 1 の通りとなった．

表 1: 教師データの学年あたりの数

-	1 年生	2 年生	3 年生	合計
文の数	313	435	480	1228

5.3 データ拡張

教師データにより幅広い表現力を持たせるための手法としてデータ拡張と呼ばれるものがある．本研究でも，教科書データのうち，訓練データと検証データにそれぞれ拡張を施す.[*2]具体的な方法は次のとおりである．なお，この節において，訓練データないし検証データを単にデータと呼ぶ.

データ拡張の手順

(1) データを 1 学年，2 学年，3 学年の文に分割
(2) 分割したデータを基に，各学年で初登場する英単語をリスト化する．以下，これを「初出単語リスト」と呼ぶ.
(3) データに対して形態素解析を行い，名詞と動詞，代名詞のみを抽出
(4) 3. で抽出した単語を個別に `[MASK]` に置き換え，それぞれを新しい文とする.
(5) BERT の事前学習モデルを用いて Masked Language Model(MLM) を解き，上位 10 件の単語のうち，そのデータに対応する学年と同じ学年の初出単語リストに属する単語で `[MASK]` を置き換え，それを拡張後の新たなデータとする.

拡張後のデータの件数は，訓練データが 18980 件に，検証データが 6350 件である.

6. 予備調査

本予備調査では，BERT を用いたリーダビリティ算出の実現にあたって必要となる 2 項目について検討を行う．ひとつは，BERT によって文法的な観点をどの程度捉えることが可能であるかの調査，二つ目は，fine-tuning による用いるアーキテクチャによる精度の差の調査である.

6.1 調査内容

本研究で行う 2 つの調査それぞれについて，「調査 1: BERT を用いた主格の関係代名詞 that を含む文の類似度計算」と「調査 2: 2 種類の fine-tuning に用いるアーキテクチャによる精度比較」として以下に調査内容を詳述する.

6.1.1 調査 1: BERT を用いた主格の関係代名詞 that を含む文の類似度計算

BERT を用いることで，入力トークン（英文の単語列）に対して，出力として「各トークンの埋め込みベクトル」を得ることができる．本調査では，”that”の埋め込みベクトルに焦点を当てて，このベクトルの比較による特定の構文の抽出を行う.

比較元の英文は以下とする.

”I bought a notebook that was made in America.”

この英文に含まれる文法についての特徴は以下の通りである.

- 主節の主語は代名詞
- 主節の動詞は一般動詞（不規則変化）
- 主節の動詞は過去形
- 主節の目的語は単数
- 主節の目的語に対する主格の関係代名詞 that
- 関係詞節は受動態
- 関係詞節内の時制は過去
- 関係詞節内に前置詞句
- 前置詞句内に固有名詞

この英文は，3 年生の教科書を参考にし，著者らが独自に作成した文である.

次に，具体的な処理の手順は以下である.

手順

(1) 比較元の英文における，BERT を用いて文脈化された中での「”that”の埋め込みベクトル」を取得する.
(2) 教師データ（1228 文）中の各データの文に対し，その文中の「各トークン（単語）の埋め込みベクトル」を得る.
(3) 得られた埋め込みベクトル群の中で，最も比較元の英文における「”that”の埋め込みベクトル」との Cosine 類似度が最大となるベクトルを調べる.
(4) 教師データの各データの文に対して，Cosine 類似度が最大となるベクトルに対応するトークン（単語）と，その Cosine 類似度を得る.

以上より，全ての教師データ（1228 文）の Cosine 類似度の結果を 0.5 以上 0.6 未満, 0.6 以上 0.7 未満, 0.7 以上の 3 通りに整理し，考察する.

BERT は，トークン（単語）ごとに，前後のトークンとの関係を考慮したベクトルを作成しており，文脈化された中での「トークン（単語）の埋め込みベクトル」のように言い表される．この調査では，文脈化された中での「”that”の埋め込みベクトル」を比較元とし，Cosine 類似度を全ての単語に対して調べている.

このことから著者らは，ここで計算される Cosine 類似度が高い単語は，比較元の英文での”that”の使い方と似た使い方をした単語であることを意味していると仮定しており，本調査で検証を行う．結果は次章に示す.

[*2] テスト用データにはデータ拡張を施さない．また，訓練データの情報が検証データに影響を与えないように，データ拡張は訓練データと検証データに対し別々に行う.

Cosine 類似度 $\cos(x, y)$ の定義は次のとおりである．ただし，$x = (x_1, x_2, ..., x_H)^\top \in \mathbb{R}^H$ と $y = (y_1, y_2, ..., y_H)^\top \in \mathbb{R}^H$ である．

$$\cos(x, y) = \sum_{i=1}^{H} x_i y_i \Big/ \sqrt{\sum_{i=1}^{H} x_i^2} \sqrt{\sum_{i=1}^{H} y_i^2}$$

6.1.2 調査 2: 2 種類の fine-tuning に用いるアーキテクチャによる精度比較

fine-tuning に関する調査として，5.2 節で前述した教師データを用い, 事前学習モデルを fine-tuning して, 分類性能を検証する．なお，教師データを，訓練データ: 検証データ: テストデータ=6:2:2 となるよう分割を行って fine-tuning に利用する．fine-tuning を行う際は，BERT の出力から更に，最終的な出力が実数値となるような新たな層を接続する形で行う．そこで，用いる層による accuracy の差異についても検証する．用いる層は次のものである．

- 畳み込み層
- 線形層

ただし，線形層は，BERT に入力した各トークンに対応する BERT の出力ベクトルを平均して，それを線形層に入力するものとする．[*3]それぞれの層を用いた時のグレーディングモデルの全体像を図 1 と図 2 に示す．

また，同様の検証を訓練データと検証データを拡張して新たに得られた教師データを用いても行う．訓練データ，検証データ，テストデータ，拡張された訓練データ，拡張された検証データの数は表 2 の通りである．

表 2: 教師データの用途別の件数と，拡張後の訓練データと検証データの件数

-	拡張前の件数	拡張後の件数
訓練データ	736	18980
検証データ	245	6350
テストデータ	247	-

最適化アルゴリズムには AdamW を用いる．また，本調査においては，ハイパーパラメータ自動最適化フレームワークである”Optuna”[9] を用い，ハイパーパラメーターである学習率とバッチサイズを調整している．ただし，探索範囲は，学習率は 1e-6 から 1e-4 の範囲，バッチサイズは 1, 2, 4, 8, 16, 32 のいずれかとし，試行回数は 100 回としている．

6.2 結果と考察

6.2.1 調査 1 の結果と考察

”that”の埋め込み表現に関する調査の結果は表 3 の通り

[*3] 正確には，対応する Attention Mask が 1 となるトークンについての平均

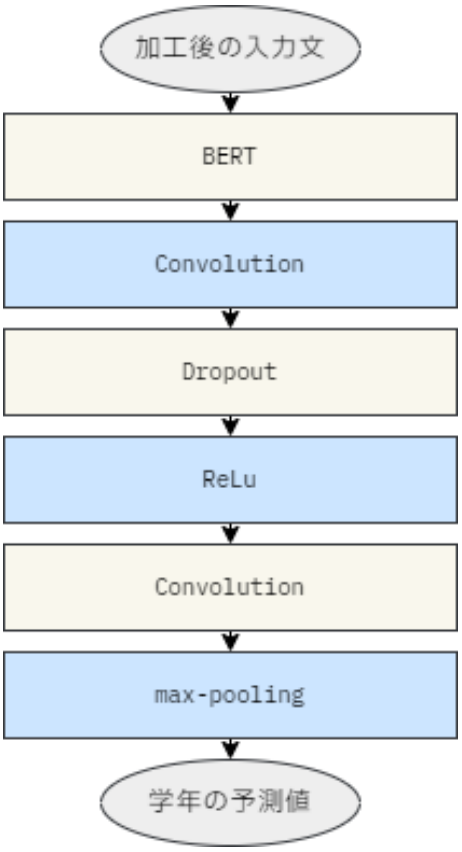

図 1: 畳み込み層を用いた場合のモデルの全体図

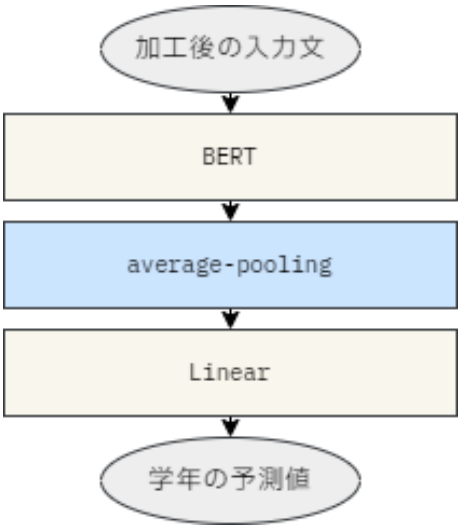

図 2: 線形層を用いた場合のモデルの全体図

となった．

表 3: 比較元の文中の”that”と類似したトークンを含む文の数．括弧内の数字は，表の該当箇所に属する文の中で真に関係代名詞を含んでいる文の数である．

範囲	1 年生	2 年生	3 年生
0.5〜0.6	42(0)	104(0)	124(28)
0.6〜0.7	0	0	11(7)
0.7〜	0	0	3(3)

まず，”that”との類似度が 0.7 以上である文についての特徴を述べる．表 4 に示すように，全てが関係代名詞の用法であることがわかり，that 節の中には，比較元の文と同様に，場所を示す修飾語や出自を示す修飾語が含まれている．

表 4: ”that”との類似度が 0.7 以上の範囲に属する文．比較元の”that”に最も類似している単語をダブルクォーテーションで示す．

本文	類似度
This is a book ”that” shows houses in Asia.	0.77
This is a postcard ”that” I get from Kenya.	0.74
It shows some animals ”that” you can see there.	0.70

次に，”that”との類似度が 0.6 以上 0.7 未満の範囲に属する文の例の範囲に属する文についての特徴は，表 5 から分かる通り，0.7〜の文よりも，より幅広い幅の関係代名詞 that が含まれている点である．また，他の that の用法や，前置詞も含まれるようになった．

表 5: ”that”との類似度が 0.6〜0.7 の範囲に属する文の例．比較元の”that”に最も類似している単語をダブルクォーテーションで示す．

本文	類似度
There are three things ”that” I want to do there.	0.62
In Japan, I was so busy ”that” I could not...(略)	0.60
Then I saw a program on TV ”about” refugee...(略)	0.60

最後に，”that”との類似度が 0.5 以上 0.6 未満の範囲に属する文の特徴は，表 6 から分かる通り，0.6〜に属する文以上により関連度が低いトークンも見られるようになった．特に，”about”のような前置詞ではなく，”came”のような一般動詞も抽出されている点が特徴である．その他にも，関係代名詞”which”も抽出されている．

表 6: ”that”との類似度が 0.5〜0.6 の範囲に属する文の例．比較元の文中の”that”に最も類似している単語をダブルクォーテーションで示す．

本文	類似度
An ambulance ”came” and took it to an ...(略)	0.56
This is a film ”which” was made by George Lucas.	0.53
This is a book ”that” shows houses in Asia.	0.57

6.2.2 調査 2 の結果と考察

BERT の fine-tuning において，BERT の後に用いる層とテストの正答率の関係に関する調査の結果は表 7 に示す通りとなった．ただし，batch はバッチのサイズ，dropout は dropout 層における確率を表す．

層	正答率 [%]	学習率	batch	dropout
線形層	45.7	3.8810e-5	32	-
線形層（拡張後）	62.7	5.8546e-6	8	-
畳み込み層	53.4	3.2957e-6	8	0.763
畳み込み層（拡張後）	63.2	1.1176e-5	8	0.733

表 7: 層と正答率の関係を表す表

6.2.3 調査を通しての考察

調査 1 の結果から最初に得られることは，関係代名詞と類似した単語が含まれる文は中学三年生の文データに特に多かったという結果であり，この教師データでは関係代名詞が現れる学年は中学三年生のみであることを踏まえると，この結果は想定通りの結果と言える．しかしながら，”that”との類似度が 0.6〜0.7 となるトークンが含まれている文はすべて中学三年生の英文であるが，関係代名詞ではない用法も含まれていることから，この表の結果から，直ちに単一の埋め込みベクトルのみを使って関係代名詞の抽出を高精度で行えると結論付けることは難しいと考えられる．更に，基準となる that の埋め込みベクトルとの Cosine 類似度が 0.5〜0.6 の間となる文については，関係代名詞を含む文がより多く抽出されているが，55%を境目に，その値を下回ると著しく関係代名詞が含まれていない文が増加した．一方，”that”との類似度が 0.7 を超えている文は 3 件しかないが全て関係代名詞の用法であり，更にこれらの文は”that”が用いられているだけではなく，後に場所を指す修飾語や出自を表す修飾語が現れている点で，比較元の文と文法的に共通点の多い構造をとっていることがわかる．これらの結果から，文法抽出の精度と件数はトレードオフの関係になっていることが確かめられた．続いて，調査 2 の結果について論じる．まず，データ拡張なしにおける結果を比較すると，average-pooling の後に線形層を用いるよりも，畳み込み層を用いた処理を行った場合の方がより高い Accuracy を見せた．データ拡張の有無で比較すると，線形層と畳み込み層各々の場合で，データ拡張を行っていない場合に対してデータ拡張を行った場合の Accuracy は 10%，17%程度の上昇がみられた．このように，同一のエポック数の条件下では，データ拡張を用いた方がより精度が高められることを確かめられた．

7. 今後の課題

今後の課題としては，より多様な種類の教科書データを用いることである．本研究で用いた教科書は 1 社のみで

あったため，この検証方法では他の英語の教科書の文に対する学年への分類性能を評価できない．したがって，今後教科書を増やし，こういった比較を行うことが必要と考えている．

参考文献

[1] 古川昭夫：SSS 英語学習法のご案内，https://www.seg.co.jp/sss/learning/ (2024/1/28 に閲覧).
[2] Kincaid, J.P., F. R. R. R. . C. B.: Derivation of new readability formulas (automated readability index, fog count, and flesch reading ease formula) for Navy enlisted personnel., (1975).
[3] 福井正康，小篠敏明：英文リーダビリティ測定システムの開発，https://www.heisei-u.ac.jp/ba/fukui/pdf/kiyou2007-1.pdf (2024/1/28 に閲覧).
[4] Ozasa, T., Weir, G. and Fukui, M.: DEVELOPMENT OF A READABILITY INDEX ATTUNED TO THE NEW ENGLISH COURSE OF STUDY IN JAPAN, *ICERI2014 Proceedings*, 7th International Conference of Education, Research and Innovation, IATED, pp. 2446–2453 (2014).
[5] 古川昭夫：ＳＳＳ推薦・多読用基本洋書のご紹介, https://www.seg.co.jp/sss/review/osusume.html (2024/1/28 に閲覧).
[6] Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding, *arXiv preprint arXiv:1810.04805* (2018).
[7] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. and Polosukhin, I.: Attention is all you need, *Advances in neural information processing systems*, Vol. 30 (2017).
[8] HuggingFace: bert-base-uncased, https://huggingface.co/bert-base-uncased (2024/1/28 に閲覧).
[9] 株式会社 Preferred Networks PreferredNetworks：Optuna ハイパーパラメータ自動最適化フレームワーク，https://www.preferred.jp/ja/projects/optuna/ (2024/1/28 に閲覧).