2V-05

# 機械学習による英文の難易度推定手法の検討

髙橋 里紗 †　　来住 伸子 ‡
津田塾大学 †　　津田塾大学 ‡

## 1 はじめに

英語が母国語ではない英語学習者にとって, 適した難易度の英語学習教材を利用することは重要である. 英語テキストの難易度を判断する上で, 英語学習者にとっての難易度と, ネイティブスピーカーにとっての難易度は異なる. 例えば, 語彙や文法が, 第二言語の読みやすさにおいて重要な役割を果たしていることが示されている. しかし, 英文難易度に関する研究の多くは, ネイティブスピーカーが判断したテキストの難易度を評価している. 本研究では, 英語学習者向けに, 英文の難易度 (英語の運用能力の国際基準である CEFR レベル) を推定することを目指している. 今回は, 機械学習による自然言語処理モデル BERT を使い, CLC（Cambridge Learner Corpus）コーパスから抽出した CLC FCE(First Certificate in English) データセット [1] を使って, 文書分類することで CEFR レベルを推定した. BERT による分類結果を, CEFR レベルの語彙数を特徴量としたクラスタリングによる分類と比較し, 結果を考察した.

## 2 関連研究

これまでの筆者らの研究 [2] では, レベル分けされた英語学習教材で使用されている, TED Talks のスクリプト 40 件の難易度を推定する手法を検討した. 有名なリーダビリティ指標の 1 つである Flesch-Kincaid Readability Ease は, 基本的な文章の特徴から計算した値であるため, 指標のスコアと教材のレベル分けは必ずしも一致しなかった. その原因として, 語彙の難易度や文法構造を考慮できていないことが考えられる. Schmalz ら [3] は, 言語学習者の試験に CEFR レベルを自動的に割り当てるために, BERT ベースモデルを使用することを提案した. モデルの入力や種類を変えながら, いくつかの実験を行ったところ, BERT ベースのアーキテクチャは, 数値的に有意なデータを用いて, 原文テキストから CEFR の習熟度を分類できることを証明した. しかし, 学習素材が少なく, 入力テキストの評価が一貫していないコーパスでは, 分類精度は高くなかった.

## 3 方法

### 3.1 使用データ

CLC FCE データセットは, 2000 年と 2001 年に Cambridge ESOL(English for Speakers of Other Languages) の FCE 試験を受けた受験者によって書かれた試験スクリプトのセットである. 受験者が作成したテキストの点数は 0.0 から 5.3 の範囲にあり, 4 段階の CEFR レベル（A2,B1,B2,C1）に対応する.

### 3.2 語彙によるクラスタリング

CEFR-J Wordlist[4] の語彙数を特徴量に用いた, 学習語彙によるクラスタリングを行った. 分類パターンは, 4 分類で予測するパターン 1(A2,B1,B2,C1) と, 2 分類で予測するパターン 2(A2 と B1,B2 と C1 をグループにする) の 2 通りである.

### 3.3 BERT を使った分類

先行研究 [3] と同様に, 事前学習済みの BERT base-uncased アーキテクチャを用いて, 4 レベ

A Study on Estimating Difficulty Level of English Learner's Text Using Machine Learning
† Risa Takahashi, Tsuda University
‡ Nobuko Kishi, Tsuda University

ルの CLC FCE データセットを分類する実験を行った．3.2 と同様の 2 種類のパターンで分類する．

## 4　結果

### 4.1　語彙によるクラスタリングの結果

クラスタリングによる分類結果を図 1,2 に示す．パターン 1,2 ともに，ある程度のクラスタに分けることができたが，各クラスタには，A2 から C1 レベルのテキストが混在しており，CEFR レベルごとの分類はできなかった．

### 4.2　BERT を使った分類の結果

BERT による分類結果を表 1 に示す．精度は，複数回の実行の平均値である．比較的高い精度で，CEFR レベルを予測することができた．また，4 分類の実行結果から 1 つを抜粋し，予測レベル，正解レベルの各データ数を，混同行列にまとめたものを表 2 に示す．

## 5　まとめ・今後の課題

英語学習者向けに，英文の難易度を CEFR レベルで推定することを目指し，自然言語処理モデル BERT を使って，文書分類を行った．クラスタリングによる分類と比較して，2 分類と 4 分類ともに，高い精度で CEFR レベルを推定することができると分かった．ただし，実用的に十分な精度にはまだ達していない．今後は，モデル変更や学習データの改善を行い，より高い精度の分類を目指す．

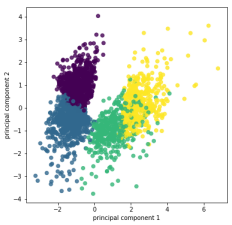


図 1　語彙による 4 分類のクラスタリング結果

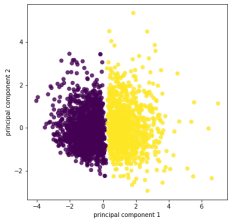


図 2　語彙による 2 分類のクラスタリング結果

表 1　BERT による分類結果

| 分類パターン | 精度 |
|---|---|
| パターン 1 (4 分類) | 70% |
| パターン 2 (2 分類) | 84% |

表 2　BERT による 4 分類結果の混同行列

| | | 予測した CEFR レベル | | | |
|---|---|---|---|---|---|
| | | A2 | B1 | B2 | C1 |
| 実際の CEFR レベル | A2 | 0 | 0 | 1 | 0 |
| | B1 | 0 | 2 | 35 | 0 |
| | B2 | 0 | 0 | 161 | 2 |
| | C1 | 0 | 0 | 32 | 1 |

## 参考文献

[1] CLC FCE データセット https://ilexir.co.uk/datasets/index.html

[2] 髙橋里紗, 来住伸子.”リーディング用英語学習教材の難易度推定手法の検討”,FIT 2022,D-005,2022.8

[3] Schmalz, V. J., & Brutti, A. (2021). Automatic Assessment of English CEFR Levels Using BERT Embeddings. In http://ceur-ws.org/Vol-3033/ (Vol. 3033). CEUR Workshop Proceedings.

[4]『CEFR-J Wordlist Version 1.6』東京外国語大学投野由紀夫研究室．（http://www.cefr-j.org/download.html より 2022 年 5 月ダウンロード）