E-031

BERT による日本語文章の難易度推定
Readability Estimation of Japanese Text Using BERT

郷原 聖士† 綱川 隆司† 西田昌史† 西村雅史†
Seiji Gobara Takashi Tsunakawa Masafumi Nishida Masafumi Nishimura

1. はじめに

情報伝達にあたって，留学生のような非母語話者や小学生のような幼い子供に対して通常用いられるレベルの文章で情報を伝えても，使用語彙や文構造の複雑さから，情報が正しく伝わらないという問題がある．この問題の解決策として，文章の平易化に関する取り組みがある[1]．平易化は，文章を元の文章と同じ意味の分かりやすい文章に変換する，自然言語処理における文章生成タスクの一つである．平易化によって情報を分かりやすい形に変換することで，情報伝達を効率化することが可能である．

一方で，平易化された文章は，元の文章よりもやさしい表現が得られるが，利用者にとって適当な難易度とは限らない．そこで，文書利用の参考にあたり，予め文章がどの程度の難易度であるかを分類する，難易度推定を行う必要がある．日本語での難易度推定に関する取り組みとしては，文章に含まれる語彙の難易度や文章の構造の複雑さを特徴量として，重回帰分析や LightGBM，英語では BERT を用いた手法などがある[2][3][4]．先行研究[3][4] の取り組みは，文の難易度推定において一定の成果を見せた一方で，依然として推定精度が低いという問題がある．

本研究では，文章の難易度推定を行うため，日本語書き言葉コーパス（BCCWJ）[5]に含まれる図書館サブコーパスとその文体情報[6]に日本語教科書コーパス[5]を加えた，9233 個の文書を用いて，日本語の事前学習 BERT モデルをファインチューニングし，文章の難易度がより正確に捉えられることを，比較実験を通して示す．

2. 文章の難易度推定器の作成

2.1. データセットの作成

文章の難易度推定実験に用いるデータセットとして，BCCWJ に含まれる文章の中から，図書館サブコーパスと日本語教科書コーパスの文書を抽出し，各文書は冒頭 512 トークンのみ残した．ここでトークンは，BERT モデル [1] に対応するトークナイザを用いて取得した．サブコーパスについては，文体情報[6] に含まれる「専門性」を 5 段階の難易度（難易度 1（専門家）～難易度 5（小学生））として設定した（以下，本設定を BERT Class と呼ぶ．以降も同様とする）．日本語教科書コーパスについては，コーパスに含まれる学習対象学年を小学生と中高生に分割し，図書館サブコーパスの難易度 4（中高生）及び 5（小学生）に対応付けてデータセットに加えた（BERT Class +text）．ここで，日本語教科書コーパスは，小学生と中学生で学年毎に分類されているが，図書館サブコーパスに含まれる文章には難易度 3（一般レベル）が多くあり，少数データの難易度 4 と難易度 5 の文章を補う目的で利用している．

また，図書館サブコーパスに対応する文体情報が存在する文書及び日本語教科書コーパスについて，訓練，開発，評価用データを無作為に 8:1:1 に分割した．くわえて，データセット内のラベルは表 1 に示すように偏りが存在している．そこでこのバランスを均衡化するため，imbalanced-learn [2]を用いて，オーバーサンプリングとアンダーサンプリングを行ったデータセットも構築した（BERT Class +samp，BERT Class +samp+text）．

2.2.BERT モデルのファインチューニング

構築した各データセットに対して，東北大学の乾研究室が公開している BERT モデル [1]を用いた分類 BERT のファインチューニングを行った．エポック毎にモデルを保存し，開発データに対する最良の正解率を持つモデルを選出した．

3. ベースライン手法の実装

ベースライン手法の分類器として，先行研究[3][4]に従い，重回帰分析，SVR，LightGBM，SVM，ロジスティック回帰を用いた．また，中町ら[4]と同様に，文章に対する埋め込みベクトルを BERT モデルから取得し，max-pooling した特徴量を LightGBM に与えた場合についても学習を行った． さらに，SVM 及び LightGBM については，グリッドサーチと 5 分割のクロスバリデーションにより，パラメータのチューニングを行った．ここで，LightGBM のハイパーパラメータのうち，Optuna[3]のサンプルコードに含まれる 7 つのパラメータをチューニングの対象とした．

まず，構築したデータセットに含まれる文章に対して，先行研究[3][4]をベースに，各特徴量を求めた．また，係り受け解析には SciPy [4]と GINZA [5]を用いた．その後，Min-Max 法によって各特徴量に対する正規化を行った．

特徴量には，平均文長，漢字比率，平均出現頻度，常用漢字初出年，係り受け距離の相対頻度，文字の種類率の 6 種類計 16 個を用いた．ここで，平均出現頻度は Wikipedia の日本語記事に出現する文字の回数を平均して，常用対数を取った数である．また，係り受け距離の相対頻度は，劉ら[3]と同様に，係り受け距離に基づくカテゴリ分類を行い，係り受け自体の総数でカテゴリ別に割った数である．

表 1　各コーパスにおける難易度別文書数

難易度	1	2	3	4	5	合計
図書館	141	929	7065	384	302	8821
教科書	0	0	0	318	94	412
図書館 +教科書	141	929	7065	702	396	9233

†静岡大学 Shizuoka University

1) https://github.com/cl-tohoku/bert-japanese
2) https://imbalanced-learn.org/
3) https://github.com/optuna/optuna
4) https://scipy.org/
5) https://megagonlabs.github.io/ginza/

4. 評価実験

4.1.実験結果と考察

難易度推定器の評価指標として，中町ら[4]と同様に，平均絶対値誤差（MAE），ピアソンの相関係数（Pearson），スピアマンの順位相関係数（Spearman），正解率（Acc），F1 スコアを求めた．評価データに対する実験結果を表 2 に示す．ここで，表 2 における回帰モデル（Linear reg，SVR，Logistic reg）の推定結果は，小数点第 1 位を四捨五入した時の整数値を推測値として計算した．表 2 より，全ての評価指標において提案手法（BERT Class）がベースライン手法を上回っていることが分かる．なお，BERT Class（+samp+text）の推定結果は全て難易度 3 であった．これにより，BERT モデルによる分類手法が，日本語文章の難易度推定においても有効であることが示された．

また，表 3 に BERT Class 及び BERT Class （+samp）における評価データに対する混同行列を示す．表 3 において，BERT Class（+samp）は，BERT Class に比べて，少数データの分類精度が向上したことから，サンプリングによるデータの増加が上手く機能したと考えられる．

4.2.分析

入力文のうち，BERT モデル（BERT Class，BERT Class +samp）が誤判別した入力文の冒頭を表 4 に示す．表 4 の例 1 について，入力文には一定レベルの漢字の使用や文の長さがある一方で，カタカナも多く含むことから，文章の「専門性」は高いが，「難易度」は高くないため，「3」と推定したと考えられる．また，BERT Class +samp では，少数データを補ったため，実際の値にやや近い「2」と推定したと予測される．例 2 について，[UNK]トークンのように，辞書に未登録の語彙やその他専門性の高い語彙が多く含まれている一方で，難易度「1」に関するサンプル数が少なかったことから，やや難易度が高い「2」と推定したと考えられる．例 3 について，語彙は中高生レベルだが，例 2 と同様に，難易度「4」「5」のサンプル数が少ないため，難易度「3」に割り当てられたと考えられる．例 4 について，冒頭に平易な文が多く存在しており，文書から抽出した冒頭 512 トークンの難易度は，「文章全体の難易度と同等の難易度になる」とする暗黙の仮定に反して，文章の難易度にばらつきがある文が含まれていたことが誤判別の原因である可能性が高い．上記より，本研究では，「専門性の高さ」を文章の難易度と見立ててモデルを作成し，低難易度から中難易度にかけての難易度推定は一定の成功が見られた．一方で，中難易度から高難易度にかけては，「専門性」が高くても文章の難易度は低い場合が存在するため，難易度推定が不正確になることが分かった．したがって，本データセットをベースにより精度良く難易度推定を行うためには，専門性と難易度について，それぞれラベル付けをする必要があると考えられる．

表 2　図書館サブコーパスによる評価結果

	MAE	Pearson	Spearman	Acc	F1
SVM (linear)	0.228	0.526	0.415	0.813	0.347
SVM (rbf)	0.290	0.553	0.445	0.739	0.410
SVM (sig)	0.305	0.519	0.439	0.745	0.415
SVM (poly)	0.221	0.575	0.503	0.816	0.380
Linear reg	0.241	0.588	0.544	0.780	0.266
SVR	0.218	0.625	0.527	0.804	0.406
Logistic reg	0.211	0.595	0.511	0.822	0.386
LightGBM	0.204	0.640	0.532	0.821	0.432
LightGBM (emb)	0.195	0.648	0.578	0.832	0.455
BERT Class	0.183	0.684	0.614	**0.841**	0.495
BERT Class (+samp)	0.181	**0.719**	**0.657**	0.837	**0.527**
BERT Class (+text)	**0.178**	0.707	0.634	**0.841**	0.483
BERT Class (+samp+text)	0.283	0.053	0.071	0.788	0.181

5. おわりに

本研究では BCCWJ に含まれる文章とその文体情報を用いて，日本語の BERT モデルをファインチューニングし，5 段階の難易度に分類する難易度推定を行った．評価実験から，提案手法は既存手法と比較して，複数の評価指標において高い評価が得られ，提案手法の有効性が示された．今後は作成した難易度推定器を用いて難解文章と平易文章の分類を行い，文章の平易化モデルを作成する予定である．

参考文献

[1] 梶原智之, 山本和英, “語釈文を用いた小学生のための語彙平易化”, 情報処理学会論文誌, Vol.56, No.3 (2015).

[2] Matej Martinc, Senja Pollak, and Marko RobnikŠikonja. “*Supervised and Unsupervised Neural Approaches to Text Readability*”. Computational Linguistics, Vol.47, No.1 (2021).

[3] 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”, 情報処理学会研究報告, No.11(2012).

[4] 中町礼文, 佐藤敏紀, 西内紗恵, 浅原正幸, 奥村学, “日本語能力試験に基づく日本語文の難易度推定”, 言語処理学会第 28 回年次大会発表論文集 (2022).

[5] 国立国語研究所コーパス開発センター, “『現代日本語書き言葉均衡コーパス』利用の手引 第 1.0 版”, 国立国語研究所コーパス開発センター (2011).

[6] 柏野和佳子,”BCCWJ 図書館サブコーパスの文体情報（第 1 版）”, 国立国語研究所 (2015).

表 3　評価データに対する混同行列

(a) BERT Class

	1	2	3	4	5
1	0	11	7	0	0
2	0	30	48	0	0
3	0	18	669	6	1
4	0	2	32	11	4
5	0	0	13	0	32

(b) BERT Class (+samp)

	1	2	3	4	5
1	0	16	2	0	0
2	0	54	24	0	0
3	0	48	641	3	1
4	0	0	36	11	0
5	0	0	12	1	32

表 4　BERT Class の誤判別例

No	True	Pred (+samp)	Input
1	1	3 (2)	現状 打開 クーデタ 後 に 成立 し た 少数 派 アラブ ・ スンニ 派 の 新し
2	1	2 (2)	この 他 に も パラコート 剤， 有機 [UNK] 剤, 臭 化 メチル 剤,りん 化 亜
3	4	3 (3)	イタリア の 政治 家 ・ 文化 人 ・ 金融 業者 (1449 ～ 1492) フィレンツ
4	3	5 (5)	目 を しっかり と 閉じ、 口もと に は うっすら と 笑み を うかべ て...。