日本語テキストの多段階平易化に向けたBERTによる難易度推定

BERT-based Readability Estimation towards Multi-level Simplification of Japanese Text

静岡大学情報学部講師
綱川　隆司

2008 年東京大学大学院情報理工学系研究科コンピュータ科学専攻博士課程単位取得退学。博士（情報理工学）。静岡大学情報学部学術研究員、同助教を経て、2019 年より静岡大学情報学部講師。自然言語処理の研究に従事。

tuna@inf.shizuoka.ac.jp　TEL 053-478-1487

静岡大学情報学部情報科学科
郷原　聖士

静岡大学情報学部情報科学科

1 はじめに

文章を書く際には読み手を想定して適切な難易度と専門性をもったテキストにすることが必要である。例えば、日本語を学習中の小学生や留学生に対しては易しい日本語を用いないと伝えるべき情報が正しく伝わらないおそれがある。一方で、様々な読み手が想定される文章について難易度に応じた複数のテキストを作成することは、ニュースなどの一部の例を除けば現実的ではない。この課題に対して文章を自動的に易しいテキストに変換する平易化タスクの研究が進められている [1]。

文章の平易化タスクにおいて出力テキストが満たすべき条件の一つとして、入力文に対して出力文がより平易である [1] ことが挙げられる。ここで、入力文および出力文の難易度がどの程度であるか明らかになっていることで、平易化された文章を利用する際の利便性が向上するとともに、出力文の難易度を調整できる多段階平易化モデルの構築にも寄与するものと考えられる。そこで本稿では、現代日本語書き言葉コーパス（BCCWJ）中の図書館サブコーパスおよび日本語教科書コーパスを規準コーパスとして用い日本語の事前学習 BERT モデルをファインチューニングし、日本語テキストの難易度推定を行うモデルを提案する。また、従来の難易度推定指標を用いた場合の結果との比較を行う。

1　入力文がすでに十分平易である場合は、そのまま出力文とする。

2 関連研究

テキストの難易度を測る手法として、語彙や構文の難しさを示す様々な特徴量を用いて計算される公式に基づく方法のほか、ある規準に基づく難易度が各文書に付与されている規準コーパスを用いて構成した分類器を用いる方法が提案されている。日本語の難易度推定について述べると、前者の方法としては、李らによる日本語教育のリーダビリティ公式 [2] などが知られている。一方、後者の手法による例として小島ら [3] は BCCWJ の日本語教科書コーパスにおける教科書の学年を難易度とみなし、文字 bigram を用いた言語モデルの尤度による難易度推定を行う手法「帯」を提案している。

また、英語における平易化のための代表的なデータセットの一つである Newsela [4] は、同じ文章に対して異なる複数の難易度の平易化を行っている。このデータにより出力する平易文の難易度制御を行う研究 [5,6]

もなされている。また、本稿で扱う BERT による難易度推定を行った研究として Martinc らによるものがある [7]。日本語では同様のデータセットは存在しないが、任意のテキストに対する難易度推定を高精度で行うことができれば、同じ文に対して複数の平易文の生成と難易度推定を行うことで擬似的に同様のデータセットを構築できることが期待される。

3 BERT による難易度推定

本稿では BCCWJ 中の図書館サブコーパスおよび日本語教科書コーパスを規準コーパスとして BERT による難易度クラス分類器の学習を行う。各文書を BERT で扱うため、文書をトークン化した後に最初の 512 トークンのみを残した。難易度は図書館サブコーパスの文体情報に含まれる 5 段階の「専門度」をそのまま用いて設定した。

- 難易度 1：専門家向き
- 難易度 2：やや専門的な一般向き
- 難易度 3：一般向き
- 難易度 4：中高生向き
- 難易度 5：小学生・幼児向き

また、日本語教科書コーパスの文書についてはその学習対象学年に応じて難易度 4 または 5 の文書として扱った。各コーパスにおける文書数を表 1 に示す。性能評価を行うため各コーパスを 8:1:1 の割合で訓練・開発・評価用データに無作為に分割し、実験評価には図書館サブコーパスの評価用データのみを用いた。さらに、ラベルの偏りを均衡化するため、オーバーサンプリングおよび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。

各データセットに対して、東北大学乾研究室による日本語 BERT 訓練済みモデル[2] を用いて分類 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もつモデルを選出して評価用データによる評価を行った。

2 https://github.com/cl-tohoku/bert-japanese

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
計	141	929	7065	702	396	9233

4 難易度を示す特徴量による難易度推定

テキストの難易度に関連する特徴量として、先行研究 [8,9] に従い以下の 6 種 16 個を用いた。

平均文長 文章中の一文あたりの文字数。

漢字比率 文章中の漢字数（重複含む）を総文字数で割った値。

Type Character Ratio（TCR） 文章中の異なり文字数を総文字数で割った値。

平均出現頻度 文章中に出現する各文字（重複除く）の Wikipedia における出現頻度の対数値を平均したもの。

レベル別漢字頻度 以下の基準で定めた各レベルの漢字が文章中に出現した頻度。小学校の 1 ～ 6 年生で習う漢字をそれぞれレベル 1 ～ 6、それ以外の常用漢字をレベル 7、常用漢字でない漢字をレベル 8 とする。

係り受け関係の相対頻度 文章中の各文を SciPy と GINZA を用いて係り受け解析して得られたすべての係り受け関係を、係り元と係り先の間の距離に応じて近い方から順に 4 段階にカテゴリー分けし、それぞれのカテゴリーの係り受け関係の相対頻度を求める。

これらの特徴量を用いて、SVM、SVR、線形回帰、ロジスティック回帰、LightGBM を用いた分類器を学習し入力テキストに対する難易度推定を行った。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 モデルから取得して max-pooling した特徴量を用いた LightGBM による分類器（LightGBM（emb））も用いた。

5 評価実験

5.1 実験結果

難易度推定の評価指標として、中町ら [9] と同様に、平均絶対値誤差（MAE）、ピアソンの相関係数（Pearson）、スペアマンの順位相関係数（Spearman）、正解率（Acc）、F1 スコアを求めた。ここで、線形回帰、

表 2　実験結果

推定器	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
SVR	0.218	0.625	0.527	0.804	0.406
線形回帰	0.241	0.588	0.544	0.780	0.266
ロジスティック回帰	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	0.183	0.684	0.614	0.841	0.495
BERT(+samp)	0.181	0.719	0.657	0.837	0.527
BERT(+text)	0.178	0.707	0.634	0.841	0.483
BERT(+text+samp)	0.283	0.053	0.071	0.788	0.181

表 3　評価データに対する混合行列

(a) BERT

正解＼推定	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 (+samp)

正解＼推定	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

SVR、ロジスティック回帰については難易度推定結果の値を小数点以下四捨五入して評価した。

評価用データに対する実験結果を表 2 に示す。各推定器に対する値は図書館サブコーパスのみを学習に用いた場合を示し、(+text) は日本語教科書コーパスを学習・開発データに加えた場合、(+samp) はアンダーサンプリング・オーバーサンプリングを適用した場合の評価結果を示す。BERT による難易度推定手法による F1 スコアは 0.495 で、他の特徴量による推定手法の性能を上回った。また、サンプリングを行うことで F1 スコアは 0.527 まで向上した。

表 3 は BERT および BERT（+samp）に対する混同行列であり、サンプリングの効果で少数データの分類精度が向上していることを示している。

5.2　特徴量に関する分析

4 節で示した特徴量 6 種 16 個について主成分分析を行った結果、図 1 に示すように第 3 主成分までで 5 割程度の寄与率を示した。そこで、第 1 主成分（PC1）～第 3 主成分（PC3）と各特徴量との因子負荷量（相関係数）を求め（表 4）、その絶対値が 0.3 を超えるものに着目すると、全特徴量の中では漢字比率、平均出現頻度、レベル別漢字頻度（主に小学校 3 ～ 6 年生）が難易度推定に与える影響が大きいことが示唆された。

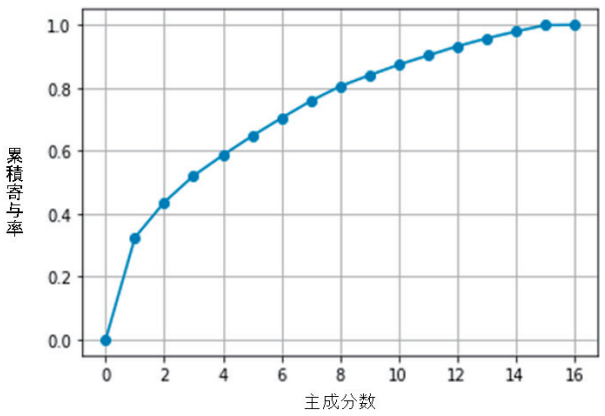

図 1　各主成分の累積寄与率

表 5 および図 2、3 は実際の各難易度のテキストに

表４　第１～第３主成分と各特徴量との因子負荷量（相関係数）（太字は絶対値が 0.3 を超えるもの）

主成分	平均文長	漢字比率	TCR	平均出現頻度	漢字頻度Lv1	漢字頻度Lv2	漢字頻度Lv3	漢字頻度Lv4	漢字頻度Lv5	漢字頻度Lv6	漢字頻度Lv7	漢字頻度Lv8	係り受けC1	係り受けC2	係り受けC3	係り受けC4
PC1	0.072	0.367	0.322	-0.115	0.181	0.189	0.352	0.413	0.400	0.315	0.160	0.030	0.089	-0.002	-0.064	-0.289
PC2	-0.016	0.149	0.470	-0.176	0.331	0.029	-0.103	-0.192	-0.307	-0.036	0.370	0.322	0.142	0.180	-0.025	0.393
PC3	0.010	-0.060	-0.086	0.078	-0.155	0.002	0.044	0.044	0.079	0.026	-0.069	-0.124	0.317	0.901	-0.115	0.024

表５　各特徴量の難易度別平均値

難易度	平均文長	漢字比率	TCR	平均出現頻度	漢字頻度Lv1	漢字頻度Lv2	漢字頻度Lv3	漢字頻度Lv4	漢字頻度Lv5	漢字頻度Lv6	漢字頻度Lv7	漢字頻度Lv8	係り受けC1	係り受けC2	係り受けC3	係り受けC4
1	187.43	0.371	0.283	5.567	41.53	57.60	55.28	44.57	40.57	18.04	30.93	2.07	0.652	0.229	0.134	0.047
2	108.51	0.356	0.285	5.675	44.26	57.05	53.67	40.83	36.31	17.02	27.18	2.24	0.653	0.230	0.134	0.049
3	49.57	0.293	0.280	5.754	44.44	49.96	39.45	26.44	19.81	12.56	26.49	2.88	0.651	0.230	0.135	0.056
4	39.83	0.214	0.250	5.810	34.32	38.02	27.41	15.40	9.66	8.52	23.08	2.49	0.647	0.230	0.136	0.063
5	34.11	0.136	0.194	6.021	29.95	27.69	16.60	8.15	4.55	3.88	8.84	0.85	0.644	0.229	0.135	0.062

対して求めた各特徴量の値の平均を示している。漢字比率、平均出現頻度、レベル別漢字頻度については概ね難易度が上がるのに伴い特徴量の数値も増加する傾向がみられる一方で、平均文長や係り受け関係の相対頻度については難易度による顕著な差はみられなかった。また、

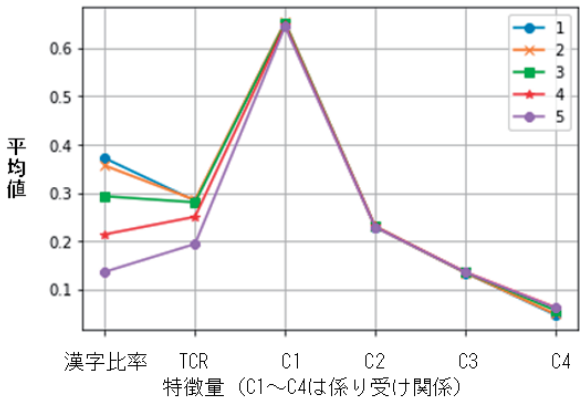

図２　各特徴量の難易度別平均値（難易度別）
（漢字比率、TCR、係り受け関係の相対頻度）

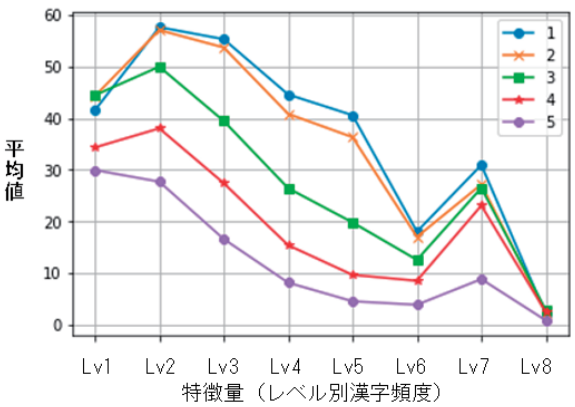

図３　各特徴量の難易度別平均値（レベル別漢字頻度）

難易度１と２については大きな差はみられず分類が難しいことが考えられる。

以上のことから、漢字や出現頻度といった語彙に関する特徴、および長距離の係り受け関係を含む複雑な構文を検出する特徴については難易度推定に対して一定の手がかりが得られているものと考えられる。

5.3　誤判別例についての分析

表６に BERT モデルによる誤判別の例を示す。入力は冒頭部分のみを示している。１番目の例は最も専門性が高い難易度１の文章であるが、カタカナ語が多く含まれており、専門性は高いものの BERT モデルによるトークン化の過程で難易度が実際より低く見積もられた可能性がある。また難易度１のデータが少ないため表３に示す通り BERT モデルではすべての文章に対して難易度１と判定することはできなかったが、この例ではサンプリングにより推定結果が難易度３から２に改善している。同様に、難易度４および５についてもデータが少なく、２番目の例のように難易度３と判定される例が多かった。

３番目の例については実際の難易度は３であるが、冒頭部分に平易な文が多く存在しているため難易度５と推定されたものである。より長い文章に対してどの部分を抽出して判定すべきかについて、特徴量を用いた簡便な抽出器で予め判定部分を抽出した上で判別する手法等も検討する余地があると思われる。

6 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、BERT を用いた分類器による手法は難易度を示す特徴量に基づく機械学習による手法に比べて高い性能が得られた。また、難易度推定のために有効な特徴量を調査し、漢字比率、平均出現頻度、小学校各年次に習得する漢字の頻度、および長距離の係り受け関係の有無が比較的有効であることが示された。

表 6　BERT の誤判別例
（※括弧内は BERT (+samp) による推定値）

No	正解	推定※	入力文章の冒頭部分
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
2	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
3	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 …。

参考文献

[1] 梶原智之 , 山本和英（2015）.“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,pp.983-992.

[2] 李在鎬（2016）.“日本語教育のための文章難易度研究 .” 早稲田日本語教育学 ,Vol.21,pp1-16.

[3] 小島健輔 , 佐藤理史 , 藤田篤（2009）.“文字 bigram モデルを用いた日本語テキストの難易度推定 .” 言語処理学会第 15 回年次大会発表論文集 ,pp.897-900.

[4] Xu,W.,Callison-Burch, C., and Napoles, C.（2015）.“Problems in Current Text Simplification Research : New Data Can Help.” *Transactions of the Association for Computational Lingustics,*Vol.3,pp.283-297.

[5] Scarton, C. and Spacia,L.（2018）.“Learning Simplifications for Specific Target Audiences.” In *Proc. of the 56th Annual Meeting of the Association for Computational Linguistics,* pp.712-718.

[6] 西原大貴 , 梶原智之 , 荒瀬由紀（2020）.“テキスト平易化における語彙制約に基づく難易度制御 .” 自然言語処理 ,Vol.27,No.2,pp.189-210.

[7] Martinc, M., Pollak, S., and Robnik-Šikonja, M.（2021）.“Supervised and Unsupervised Neural Approaches to Text Readability.” *Computational Linguistics,* Vol. 47, No.1, pp. 141-179.

[8] 劉志宇 , 内田理 .（2012）.“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 ,Vol.2012-NL-205,No.11,pp.1-5.

[9] 中町礼文 , 佐藤敏紀 , 西内紗恵 , 浅原正幸 , 奥村学 .（2022）.“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 ,pp.658-663.