テキスト平易化パラレルコーパスに基づく教師なし文難易度推定

宮田 莉奈 [1,2]　浦川 通 [2]　田森 秀明 [2]　梶原 智之 [1]
[1] 愛媛大学大学院理工学研究科　[2] 株式会社朝日新聞社
{miyata@ai.,kajiwara@}cs.ehime-u.ac.jp　{urakawa-t, tamori-h}@asahi.com

概要

本研究では、絶対的な文難易度が付与されていないコーパスから文難易度推定器を訓練し、所与の文集合を難易度でランキングする課題に取り組む。文難易度は読み手の知識に依存するため、客観的かつ絶対的な難易度の判定には専門家による高コストな評価が必要となる。そのため、絶対的な文難易度が付与されたコーパスは少ないが、2 文間の相対的な文難易度が付与されたテキスト平易化パラレルコーパスは比較的多くの言語で利用できる。本研究では、多言語展開を念頭に置き、テキスト平易化パラレルコーパスに基づく文難易度推定の手法を提案する。英語における評価実験の結果、提案手法は既存の教師なし文難易度推定の性能を上回るとともに、絶対的な文難易度のラベル付きコーパスで訓練した教師あり手法にも匹敵する性能を達成した。

1　はじめに

テキストの難易度推定は、単語・文・文章などのテキストに対して読みやすさの値を付与するタスクである。この技術は、子ども [1] や言語学習者 [2]、認知障害を持つ人々 [3] など、幅広い読み手の言語能力に合わせたテキストの提供や作成の支援に利用される。本研究では、難解な文を平易に言い換えるテキスト平易化 [4] において主に使用される文の単位で難易度推定に取り組む。

文難易度は読み手の知識に依存するため、客観的かつ絶対的な難易度の判定には専門家による高コストな評価が必要となる。そのため、絶対的な文難易度が付与されたコーパス [5, 6] は英語においても 1 千文から 1 万文の規模でしか存在せず、その他の言語においてはほとんど利用できない。この少資源問題は、高品質な教師あり文難易度推定の研究開発を困難にしている。

本研究では、文難易度推定の多言語展開を念頭に置き、絶対的な文難易度のラベル付きデータを用いない教師なし文難易度推定に取り組む。絶対的な文難易度のラベル付きコーパスが少ない一方で、後述するように、相対的な文難易度の付与されたテキスト平易化パラレルコーパスは比較的多くの言語で利用できる。そこで我々は、テキスト平易化パラレルコーパスに含まれる難解文と平易文の対を用いて、所与の 2 文のどちらがより平易かを推定する相対的な文難易度推定器を訓練する。そして、これを用いて文集合を一対比較し、難易度ランキングを得る。

英語における相対的な文難易度推定に関する評価実験の結果、提案手法は既存の教師なし手法を上回る性能を達成した。また、絶対的な文難易度を考慮した教師あり手法にも匹敵する性能を示した。

2　関連研究

2.1　テキストの難易度推定

テキストの難易度推定には、可読性指標・品詞や構文解析に基づく言語学的特徴・言語モデル尤度などを素性とする教師あり学習手法 [2, 7, 8] が提案されてきた。これらの教師あり手法は、絶対的な難易度が付与されたコーパスに基づいており、ラベル付きコーパスを利用できない言語には適用できない。

ラベル付きコーパスを必要としない教師なし手法には、単語数や音節数から対象学年を推定する Flesch-Kincaid Grade Level (FKGL) [9] などの可読性指標や、相対的な難易度推定に基づくランキング手法 [10] がある。ただし、これらは文書単位の難易度推定に取り組んでいる。文単位でも適用可能な手法には、言語モデル尤度に基づく Ranked Sentence Readability Score (RSRS) [11] や大規模言語モデルの文脈内学習 [12] がある。本研究も、絶対的な文難易度を用いない教師なし文難易度推定に取り組む。

2.2 テキスト平易化パラレルコーパス

テキスト平易化パラレルコーパスは、難解な文と平易な同義文の対からなる。英語では数十万文対 [1, 13]、日本語では数千から数万文対 [14–18]、その他、デンマーク語 [19] スペイン語 [1] イタリア語 [20] フランス語 [21] ロシア語 [22] などでも数万から数十万文対のテキスト平易化パラレルコーパスを利用できる。通常、テキスト平易化パラレルコーパスに含まれる各文には詳細な難易度は付与されておらず、各文対において難解文よりも平易文の方が平易だということだけがわかる（他の文対とは比較できない）ため、テキスト平易化パラレルコーパスは相対的な文難易度が付与されたコーパスだと考えることができる。本研究では、テキスト平易化パラレルコーパスから得られる相対的な文難易度の情報のみを用いて教師なし文難易度推定に取り組む。

3 提案手法

テキスト平易化パラレルコーパスに含まれる難解文と平易文の対を用いて、所与の 2 文のどちらがより平易かを推定する相対的な文難易度推定器を訓練する（3.1 節）。そして、所与の文集合に対する一対比較を実施し、難易度ランキングを得る（3.2 節）。

3.1 相対的な文難易度推定器

文難易度推定器は BERT [23] などのマスク言語モデルをファインチューニングして構築する。テキスト平易化パラレルコーパスに含まれる難解文と平易文の文対から、先頭と文境界を示す特殊トークンを用いて、“[CLS] 難解文 [SEP] 平易文” という形式の入力系列を作成する。ただし、作成時に 50%の確率で入力系列の難解文と平易文の位置を入れ替える。このデータセットを用いて、所与の 2 文のどちらがより平易かを推定する 2 値分類器を訓練する。

3.2 難易度ランキング

3.1 節の相対的な文難易度推定器を用いて、任意の文集合に含まれる各文の難易度を順位付けする。本研究では、一対比較法によってこの難易度ランキングを得る。つまり、所与の文集合に含まれる全ての 2 文の組み合わせに対して、相対的な文難易度を推定し、どちらがより平易かを評価する。最終的に、各文が「より平易である」と評価される確率に従って難易度ランキングを得る。

4 評価実験

提案手法の有効性を評価するために、英語の文集合を難易度でランキングする実験を行った。文章単位の難易度ランキングの先行研究 [24] に従い、Normalized Discounted Cumulative Gain (NDCG)、スピアマンの順位相関係数 (ρ)、ケンドールの順位相関係数 (τ)、ランキングの完全一致 (Ranking Accuracy; RA) の 4 指標により難易度ランキングを評価した。

4.1 実験設定

データ Newsela[1] [1, 13] のテキスト平易化パラレルコーパスのうち、38.5 万文対の訓練用および 4.2 万文対の検証用データを用いて、相対的な文難易度推定器を訓練した。そして、4.3 万文対の評価用データを用いて、難易度ランキングのための文集合を構築した。Newsela は英語のニュース記事を人手で 4 段階に平易化して構成されたパラレルコーパスであり、本実験では同一の原文に対する平易化である同義文集合を用いて難易度ランキングを行った。また、絶対的な文難易度が付与された英語コーパスである CEFR-SP[2] [6] からも、難易度ランキングのための文集合を構築した。CEFR-SP はパラレルコーパスではないため、Newsela とは異なり、非同義文の集合であることに注意されたい。CEFR-SP は 1.7 万文のそれぞれに対して 6 段階の難易度が付与されており、各難易度の文を無作為に 1 文ずつ選択して文集合を得た。Newsela の評価用文集合は 5 文 1 組の合計 4,478 組、CEFR-SP の評価用文集合は 6 文 1 組の合計 165 組である。

モデル 文難易度推定器は BERT[3] [23] をファインチューニングして構築した。訓練のハイパーパラメータとして、バッチサイズを 128 文対、学習率を 5×10^{-5} に設定し、最適化手法には AdamW [25] を用いた。そして、検証用データにおける交差エントロピー損失が 3 エポック連続で改善しない場合に訓練を終了する early stopping を適用した。

4.2 比較手法

教師なし手法 教師なし文難易度推定のベースラインとして、言語モデル尤度に基づく RSRS [11] および大規模言語モデル（Large Language Model; LLM）

1) https://github.com/chaojiang06/wiki-auto
2) https://github.com/yukiar/CEFR-SP
3) https://huggingface.co/google-bert/bert-base-uncased

表 1　文難易度ランキングの評価結果

		Newsela の同義文集合				CEFR-SP の非同義文集合			
	教師あり	NDCG	ρ	τ	RA	NDCG	ρ	τ	RA
RSRS	×	0.913	0.402	0.341	0.081	0.851	0.082	0.060	0.000
LLM (0-shot)	×	0.888	0.207	0.178	0.041	0.861	0.034	0.027	0.000
本研究	×	**0.985**	**0.865**	**0.799**	**0.421**	**0.958**	**0.749**	**0.619**	**0.048**
Pointwise	○	0.980	0.841	0.769	0.369	0.949	0.661	0.529	0.012
Pairwise	○	**0.986**	**0.874**	**0.811**	**0.438**	0.961	0.755	0.621	0.048
LLM (10-shot)	△	0.953	0.644	0.550	0.130	**0.967**	**0.764**	**0.636**	**0.073**

に基づく手法の 2 種類を用いた。これらの教師なし手法との比較によって、提案手法の有効性を確認する。LLM ベースの文難易度推定においては、文章単位の難易度推定に取り組む先行研究 [26] で使用されたプロンプトを本タスク用に調整した以下のプロンプトを用いた。

System Prompt:
Evaluate the readability of the text using the following eleven levels (reading difficulty):
[score: 2]: Most Easy
[score: 12]: Most Difficult
Based on the provided text examples, assign a readability score to new text and display it in the following format: "[score: X]"
User Input:
Text: {example 1}
[score: 2]
...
Text: {example n}
[score: 12]
New text:
Text: "{}"

LLM ベースの文難易度推定は、LLaMA[4] [27] を使用し、事例（上記プロンプトの “example” の部分）を提示しない 0-shot の設定と各難易度 10 件ずつの事例を提示する 10-shot の設定の両方を評価した。これらの事例は、Newsela の検証用データから無作為に抽出した。なお、10-shot の設定は少量の教師データを用いるため、教師なし文難易度推定ではないことに注意されたい。

教師あり手法　教師あり文難易度推定のベースラインとして、1 文入力の Pointwise 法および 2 文入力の Pairwise 法の 2 種類を用いた。これらはいずれも、提案手法と同じくマスク言語モデルに基づく文難易度推定であり、提案手法とは異なり Newsela コーパス [1, 13] から得られる絶対的な難易度ラベル[5]を用いて訓練した。これらのマスク言語モデルに基づく教師あり手法との比較によって、相対的な難易度情報と絶対的な難易度情報の有用性の差異について明らかにする。Pointwise 法は、マスク言語モデルを用いて入力文の難易度を推定する回帰モデルであり、最終的には各文の推定難易度を用いて難易度ランキングを得る。Pairwise 法は、提案手法と同様に 2 文を入力するが、2 文の難易度の差を推定する回帰モデルである。Pairwise 法の出力が正の値を取るか負の値を取るかによって、どちらの文がより平易かという判定ができるため、提案手法と同様に難易度ランキングが得られる。

4.3　実験結果

同義文集合の難易度ランキング　Newsela の同義文集合に対する難易度ランキングの実験結果を表 1 の左側に示す。上段の教師なし手法の中で、提案手法が最高性能を達成した。教師あり手法の中では、Pointwise 法よりも Pairwise 法の方が高い性能を示し、相対的な難易度推定のために文間の関係を考慮することが重要であることを示唆している。全体としては教師ありの Pairwise 法が最高性能を示したものの、教師なしの提案手法もそれに匹敵する性能を達成した。また、提案手法は教師ありの Pointwise 法や few-shot の LLM よりも優れた性能を示し、その有効性が明らかになった。

4）　https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct

5）　テキスト平易化の先行研究 [28–30] に従い、ある文が含まれる文書の難易度をその文の難易度とした。

表 2　文対の難易度差が難易度推定に与える影響の分析

文対の難易度差	難易度推定の正解率
1	0.759
2	0.886
3	0.954
4	0.990

非同義文集合の難易度ランキング　CEFR-SP の非同義文集合に対する難易度ランキングの実験結果を表 1 の右側に示す。同義文集合に対する難易度ランキングの実験結果と同様、上段の教師なし手法の中で、提案手法が最高性能を達成した。教師あり手法と比較すると、やはり同義文集合に対する難易度ランキングの実験結果と同様、提案手法は Pointwise 法の性能を上回り、Pairwise 法に匹敵する性能を達成した。また、Few-shot の LLM は他の教師あり手法を上回り、非同義文集合に対する難易度ランキングの最高性能を達成した。

4.4　分析

Newsela の同義文集合に対する難易度ランキングの実験について、詳細な分析を行う。

文対の難易度の差が大きいほど相対的な難易度推定は容易であるか？ → Yes.　これを明らかにするために、実験を追加する。平易化の難易度差ごとに文対を分け、難易度推定の正解率を評価した結果を表 2 に示す。この分析の結果から、難易度の差が大きくなる（平易化の段階が増える）ほど、相対的な難易度推定の正解率が向上することがわかる。期待通り、難易度の差が大きい文対ほど容易に難易度推定ができると言える。

具体例を見ると、1 段階の平易化である “Sub-Saharan Africa has benefited from **high** oil and other commodities **prices**, which have started to decline sharply. → Sub-Saharan Africa has benefited from **high prices for** oil and other commodities, which have started to decline sharply.” のような差分の少ない文対は難易度の差も小さく、後者の文が平易であることを判定するのも難しい。一方で、4 段階の平易化である “Any artifacts linked to an emperor would bring tremendous pride to Mexico. → Finding remains of those leaders would make Mexico proud.” のような差分の多い文対は難易度の差も大きく、後者の文が平易であることを判定するのは難しくない。実際に提案手法は、上の例では難易度推定に失敗し、下の例では成功した。

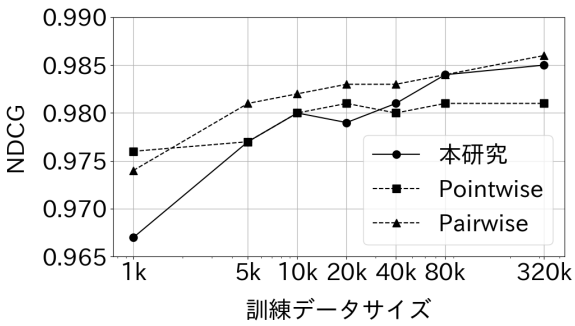

図 1　訓練用文対数が難易度推定に与える影響の分析

何件の訓練データがあれば提案手法は有効に機能するのか？ → 5 千文対。　これを明らかにするために、実験を追加する。38.5 万文対の訓練データを、32 万文対から 1 千文対まで減らしつつ、難易度ランキングの品質（NDCG）を評価した結果を図 1 に示す。この分析の結果から、提案手法の訓練用のテキスト平易化パラレルコーパスは、1 千文対だけでも NDCG=0.967 をとり、RSRS や LLM の教師なし文難易度推定よりも高性能である。この規模のテキスト平易化パラレルコーパスは、2.2 節で述べたように、日本語をはじめとして複数の言語において利用できるため、本手法は文難易度推定の多言語展開において有望である。また、5 千文対のテキスト平易化パラレルコーパスを用意できれば、教師あり文難易度推定と同等の性能に到達できる。

5　おわりに

本研究では、絶対的な文難易度の情報を用いない教師なし文難易度推定に取り組んだ。提案手法では、テキスト平易化パラレルコーパスを用いて、所与の 2 文のどちらがより平易かを推定する相対的な文難易度推定器を訓練する。そして、一対比較によって文集合に対する難易度ランキングを得た。

英語における実験の結果、同義文集合と非同義文集合の両方において、提案手法は既存の教師なし文難易度推定の性能を上回るとともに、絶対的な文難易度を考慮して訓練した教師あり手法にも匹敵する性能を達成した。詳細な分析の結果、難易度の差が大きい文対ほど相対的な難易度推定が容易であること、提案手法は 1 千文対の訓練のみでも既存の教師なし手法よりも高性能であること、5 千文対のテキスト平易化パラレルコーパスがあれば教師あり手法と同等の性能が得られることが明らかになった。

今後は、本手法を応用して、強化学習 [30, 31] の枠組みなどでテキスト平易化の性能を改善したい。

参考文献

[1] Wei Xu, Chris Callison-Burch, and Courtney Napoles. Problems in current text simplification research: New data can help. **TACL**, Vol. 3, pp. 283–297, 2015.

[2] Menglin Xia, Ekaterina Kochmar, and Ted Briscoe. Text Readability Assessment for Second Language Learners. In **Proc. of BEA**, pp. 12–22, 2016.

[3] Victoria Yaneva, Constantin Orăsan, Richard Evans, and Omid Rohanian. Combining Multiple Corpora for Readability Assessment for People with Cognitive Disabilities. In **Proc. of BEA**, pp. 121–132, 2017.

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

[5] Sanja Stajner, Simone Paolo Ponzetto, and Heiner Stuckenschmidt. Automatic Assessment of Absolute Sentence Complexity. In **Proc. of IJCAI**, pp. 4096–4102, 2017.

[6] Yuki Arase, Satoru Uchida, and Tomoyuki Kajiwara. CEFR-Based Sentence Difficulty Annotation and Assessment. In **Proc. of EMNLP**, pp. 6206–6219, 2022.

[7] Sowmya Vajjala and Ivana Lučić. OneStopEnglish corpus: A new corpus for automatic readability assessment and text simplification. In **Proc. of BEA**, pp. 297–304, 2018.

[8] Tovly Deutsch, Masoud Jasbi, and Stuart Shieber. Linguistic features for readability assessment. In **Proc. of BEA**, pp. 1–17, 2020.

[9] J. Peter Kincaid, Robert P. Fishburne Jr., Richard L. Rogers, and Brad S. Chissom. Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel. **Technical report, Defence Technical Information Center (DTIC) Document**, 1975.

[10] Kumiko Tanaka-Ishii, Satoshi Tezuka, and Hiroshi Terada. Sorting texts by readability. **CL**, Vol. 36, No. 2, pp. 203–227, 2010.

[11] Matej Martinc, Senja Pollak, and Marko Robnik-Šikonja. Supervised and Unsupervised Neural Approaches to Text Readability. **CL**, Vol. 47, No. 1, pp. 141–179, 2021.

[12] Sean Trott and Pamela Rivière. Measuring and Modifying the Readability of English Texts with GPT-4. In **Proc. of TSAR**, pp. 126–134, 2024.

[13] Chao Jiang, Mounica Maddela, Wuwei Lan, Yang Zhong, and Wei Xu. Neural CRF Model for Sentence Alignment in Text Simplification. In **Proc. of ACL**, pp. 7943–7960, 2020.

[14] Takumi Maruyama and Kazuhide Yamamoto. Simplified Corpus with Core Vocabulary. In **Proc. of LREC**, pp. 1153–1160, 2018.

[15] Akihiro Katsuta and Kazuhide Yamamoto. Crowdsourced Corpus of Sentence Simplification with Core Vocabulary. In **Proc. of LREC**, pp. 461–466, 2018.

[16] Akio Hayakawa, Tomoyuki Kajiwara, Hiroki Ouchi, and Taro Watanabe. JADES: New Text Simplification Dataset in Japanese Targeted at Non-Native Speakers. In **Proc. of TSAR**, pp. 179–187, 2022.

[17] Toru Urakawa, Yuya Taguchi, Takuro Niitsuma, and Hideaki Tamori. A Japanese News Simplification Corpus with Faithfulness. In **Proc. of COLING-LREC**, pp. 659–665, 2024.

[18] 宮田莉奈, 惟高日向, 山内洋輝, 柳本大輝, 梶原智之, 二宮崇, 西脇靖紘. MATCHA：専門家が平易化した記事を用いたやさしい日本語パラレルコーパス. 自然言語処理, Vol. 31, No. 2, pp. 590–609, 2024.

[19] Sigrid Klerke and Anders Søgaard. DSim, a Danish parallel corpus for text simplification. In **Proc. of LREC**, pp. 4015–4018, 2012.

[20] Dominique Brunato, Andrea Cimino, Felice Dell'Orletta, and Giulia Venturi. PaCCSS-IT: A parallel corpus of complex-simple sentences for automatic text simplification. In **Proc. of EMNLP**, pp. 351–361, 2016.

[21] Rémi Cardon and Natalia Grabar. French biomedical text simplification: When small and precise helps. In **Proc. of COLING**, pp. 710–716, 2020.

[22] Andrey Sakhovskiy, Alexandra Izhevskaya, Alena Pestova, Elena Tutubalina, Valentin Malykh, Ivan Smurov, and Ekaterina Artemova. RuSimpleSentEval-2021 Shared Task: Evaluating Sentence Simplification for Russian. In **Proceedings of the International Conference on Computational Linguistics and Intellectual Technologies "DIALOGUE"**, pp. 607–617, 2021.

[23] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In **Proc. of NAACL**, pp. 4171–4186, 2019.

[24] Justin Lee and Sowmya Vajjala. A Neural Pairwise Ranking Model for Readability Assessment. In **Findings of ACL**, pp. 3802–3813, 2022.

[25] Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. In **Proc. of ICLR**, 2019.

[26] Ziyang Wang, Sanwoo Lee, Hsiu-Yuan Huang, and Yunfang Wu. FPT: Feature Prompt Tuning for Few-shot Readability Assessment. In **Proc. of NAACL**, pp. 280–295, 2024.

[27] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. LLaMA: Open and Efficient Foundation Language Models. **arXiv:2302.13971**, 2023.

[28] Carolina Scarton and Lucia Specia. Learning simplifications for specific target audiences. In **Proc. of ACL**, pp. 712–718, 2018.

[29] Daiki Nishihara, Tomoyuki Kajiwara, and Yuki Arase. Controllable text simplification with lexical constraint loss. In **Proc. of ACL-SRW**, pp. 260–266, 2019.

[30] Daiki Yanamoto, Tomoki Ikawa, Tomoyuki Kajiwara, Takashi Ninomiya, Satoru Uchida, and Yuki Arase. Controllable text simplification with deep reinforcement learning. In **Proc. of AACL**, pp. 398–404, 2022.

[31] Akifumi Nakamachi, Tomoyuki Kajiwara, and Yuki Arase. Text simplification with reinforcement learning using supervised rewards on grammaticality, meaning preservation, and simplicity. In **Proc. of AACL-SRW**, pp. 153–159, 2020.