テキスト平易化の品質推定のための擬似訓練

廣中 勇希　　梶原 智之　　二宮 崇
愛媛大学大学院理工学研究科
{hironaka@ai., kajiwara@, ninomiya@}cs.ehime-u.ac.jp

概要

本研究では，テキスト平易化の品質推定の性能改善のために，事前訓練モデルのファインチューニングの前に擬似的な品質推定の訓練の実施を提案する．品質推定のラベル付きデータは構築コストが高く，テキスト平易化の品質推定のために利用可能な英語の既存データは 600 件と小規模である．この少資源問題を緩和するために，品質ラベルを持たない既存の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判定する擬似的な品質推定の訓練を大規模に実施する．3 種類の事前訓練モデルを対象とする英語のテキスト平易化の品質推定における実験の結果，提案手法は平易性だけでなく同義性に関しても性能を改善できた．

1　はじめに

テキスト平易化 [1] は，意味を保持しつつ難解な表現を平易に変換するタスクである．自動的な文の平易化は，子ども [2] や語学学習者 [3] の学習支援や読解支援に貢献し，関係抽出 [4] や機械翻訳 [5] などの他の自然言語処理タスクの性能改善にも役立つ．

テキスト平易化モデルの品質は，文法性・同義性・平易性の観点からの人手評価や，参照文に基づく SARI [6] や BLEU [7] およびリーダビリティ指標である FKGL [8] などの自動評価によって評価されている．しかし，人手評価にはコストや再現性の課題があり，自動評価には人手評価との相関が低いという課題がある．また，実世界においてテキスト平易化モデルが使用される際には参照文が存在しない場合が多く，参照文に基づく自動評価は活用できない．このような背景から，参照文を用いないテキスト平易化の品質推定 [9–12] が研究されている．

テキスト平易化における教師あり品質推定の先行研究 [10, 11] は，単語分散表現や単語一致率に基づく評価指標を用いた素性抽出を行い，機械学習モデルを訓練している．文脈を考慮可能な深層学習モデルを用いることで品質推定の性能を改善できると期待されるが，テキスト平易化の品質推定のための既存のデータセットは，統計的機械翻訳に基づくモデルを対象とする QATS[1)] [9] もニューラル機械翻訳に基づくモデルを対象とする Simplicity-DA[2)] [12] も，約 600 文対で構成されており，深層学習モデルを充分に訓練するためには小規模である．

このようなテキスト平易化の品質推定における少資源問題に対処するために，本研究では，事前訓練モデルのファインチューニングの前に，品質推定によく似た擬似タスクの訓練を行う．これによって，品質推定の小規模なラベル付きデータ上での効率的な訓練を促進する．本研究では，この擬似タスクとして，既存の大規模なテキスト平易化パラレルコーパス [13] を用いて，難解文と平易文の識別に取り組むことを提案する．Simplicity-DA [12] を用いた英語のテキスト平易化の品質推定における評価実験の結果，期待通り平易性に関する品質推定の性能が改善するだけでなく，扱う深層学習モデルによっては同義性にも改善が見られた．

2　関連研究

2.1　テキスト平易化の品質推定

難解な英文と平易な英文から構成されるパラレルコーパス [13] を用いて，系列変換タスクとしてのテキスト平易化の研究が行われている．2010 年代の前半には，フレーズベース統計的機械翻訳 [14] に基づくテキスト平易化 [15–17] が研究された．ニューラル機械翻訳 [18] の成功を受け，2010 年代の後半からはリカレントニューラルネットワークなどの深層学習に基づくテキスト平易化 [19–21] が研究されている．近年は，機械翻訳などの他の系列変換タスクと同じく，Transformer [22] に基づくテキスト平易化 [23–25] が主流となっている．

1） https://qats2016.github.io/
2） https://github.com/feralvam/metaeval-simplification

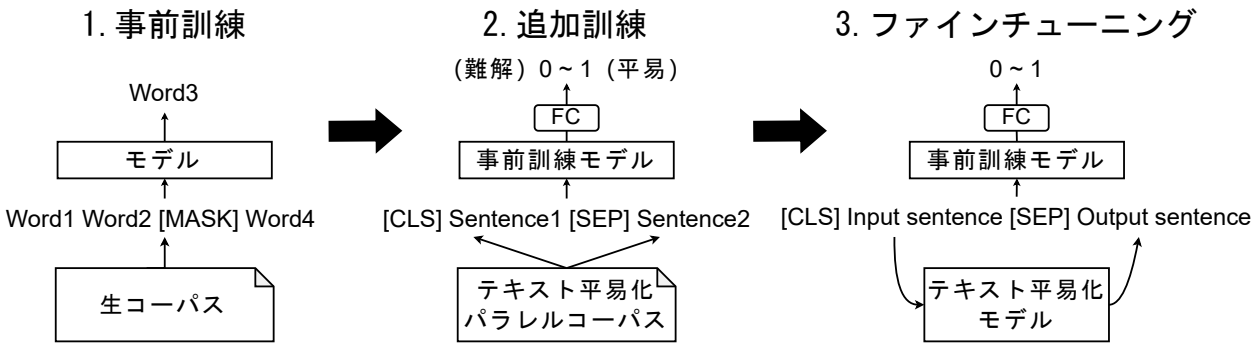

図 1　提案手法の概要（追加訓練とファインチューニングの両方において [SEP] の後の文を評価対象とする）

品質推定は，入力文と出力文の文対から出力文の品質を推定するタスクである．テキスト平易化における先行研究としては，QATS データセット [9] を用いて，サポートベクトルマシンやリッジ回帰などの機械学習に基づく品質推定モデルが訓練されている．Kajiwara and Fujita [10] は，単語分散表現 [26] に基づく素性抽出を行い，Good・OK・Bad の 3 クラス分類としての品質推定を行った．Martin et al. [11] は，BLEU [7] などの機械翻訳の評価指標や FKGL [8] などのリーダビリティ指標に基づく素性抽出を行い，回帰および分類の品質推定を行った．Alva-Manchego et al. [12] は，深層学習に基づくテキスト平易化モデルを評価対象とする Simplicity-DA データセットを構築した．本データセットは，近年のテキスト平易化モデルを対象としているものの，QATS データセットと同様にデータ数が 600 件と少ない．本研究では，このような小規模なラベル付きデータからの品質推定の性能改善に取り組む．

2.2　目的タスクに特化した訓練

近年の自然言語処理では，マスク言語モデル [27–29] などの事前訓練モデルを目的タスクでファインチューニングする転移学習のアプローチが，様々な応用タスクにおいて高い性能を達成している．最終的なタスクの性能は，事前訓練や事前訓練に続く追加訓練として，ファインチューニングの前に目的タスクに近い特性を持つ訓練を行うことで，さらに改善できることが知られている．

例えば，要約タスクにおいては入力文章中の一部の文をマスクし復元する事前訓練 [30] が有効であり，言い換え生成タスクにおいては折り返し翻訳を復元する事前訓練 [31] が有効であることが報告されている．分類や回帰のタスクにおいても，事前訓練とファインチューニングの間で句の言い換えを分類する追加訓練を実施することで，文の類似度推定の性能を改善できる [32] ことが報告されている．しかし，テキスト平易化の品質推定のための効果的な追加訓練の方法は明らかになっていない．

3　提案手法

本研究では，図 1 のように事前訓練モデルを 2 ステップで訓練することで，テキスト平易化の品質推定モデルを構築する．テキスト平易化の品質推定のためのラベル付きデータは小規模にしか存在しないが，我々が提案する追加訓練は品質推定のラベルを必要とせず，既存のテキスト平易化パラレルコーパスのみを用いるため，大規模に訓練できる．先行研究 [11] に従い，文法性・同義性・平易性の各観点で回帰問題としての品質推定モデルを訓練する．

3.1　事前訓練：マスク言語モデリング

品質推定モデルには Transfomer [22] エンコーダを用いる．小規模なラベル付きデータから効率的に訓練するために，まず大規模な生コーパス上で事前訓練する．本研究では，事前訓練のタスクとしてマスク言語モデリングを採用し，BERT などの事前訓練モデル [27–29] を用いる．

3.2　追加訓練：擬似的な品質推定

テキスト平易化の品質推定における少資源問題に対処するために，品質推定のラベル付きデータでファインチューニングする前に，擬似的な品質推定タスクで事前訓練モデルを追加訓練する．追加訓練では，図 1 の中央に示すように，難解な文と平易な文の文対を連結して入力し，連結用の特殊トークン [SEP] に続く 2 文目が 1 文目と比べて相対的に難解であるか平易であるかを判定する．これは，テキスト平易化の品質推定における平易性の評価に近い問

表 1　本実験で使用したデータセットの統計

タスク	名称	分割	文対数
擬似的な品質推定	Wiki-Auto [13]	訓練用	488,332
	Turk Corpus [6]	検証用	2,000
	Newsela-Auto [13]	訓練用	394,300
		検証用	43,317
品質推定	Simplicity-DA [11]	訓練用	400
		検証用	100
		評価用	100

題設定であるため，この擬似的な品質推定タスクの追加訓練を挟むことによって，平易性に関する品質推定の性能改善が期待できる．なお，この訓練には人手で付与された品質推定のラベルは不要であり，既存のテキスト平易化パラレルコーパス [13] のみがあれば良いため，大規模な訓練が可能である．

3.3　ファインチューニング：品質推定

3.2 節の擬似的な品質推定の追加訓練を経たモデルを用いて，入力文とテキスト平易化システムの出力文の文対に対して，図 1 の右に示すように実際の品質推定タスクでファインチューニングを行う．擬似的な品質推定の訓練で粗いレベルで評価できるようになったテキスト平易化の品質を，実際の品質推定でのファインチューニングによって，より細かいレベルで評価できるようになることが期待される．

4　評価実験

提案手法の有効性を検証するために，テキスト平易化モデルの文単位の品質推定を行う．本実験では，文法性・同義性・平易性の各項目で回帰モデルとしての品質推定モデルを訓練した．図 1 中央の追加訓練は，分類タスクとしても回帰タスクとしても設計できるが，予備実験で高性能を達成した回帰タスクを採用した．品質推定の性能は，推定値と人手評価値の間のピアソン相関によって自動評価した．

4.1　実験設定

データ　表 1 に，本実験で使用するデータセットを示す．擬似的な品質推定の追加訓練のために，英語のテキスト平易化モデルの訓練によく使われる Wikipedia および Newsela の 2 種類のパラレルコーパスを使用した．Wikipedia は，訓練用に Wiki-Auto[3] [13]，検証用に Turk Corpus[4] [6] を用いた．Newsela は，訓練用および検証用に Newsela-Auto[3] [13] を用いた．品質推定のファインチューニングのために，Simplicity-DA[2] [12] を使用した．600 件のデータセットを，訓練用に 400 件，検証用および評価用に 100 件ずつ，無作為分割して用いた．

表 2　品質推定の実験結果（ピアソン相関）

	文法性	同義性	平易性
Kajiwara-17	0.405	0.670	0.373
Martin-18	0.462	0.680	0.320
BERT	0.766	0.638	0.482
+ pseudoQE (Wikipedia)	0.662	**0.690**	**0.535**
+ pseudoQE (Newsela)	0.717	**0.741**	**0.556**
RoBERTa	0.790	0.779	0.543
+ pseudoQE (Wikipedia)	0.667	0.739	**0.545**
+ pseudoQE (Newsela)	0.732	0.751	**0.578**
DeBERTa	0.716	0.734	0.473
+ pseudoQE (Wikipedia)	0.670	**0.774**	**0.598**
+ pseudoQE (Newsela)	**0.750**	**0.770**	**0.622**

モデル　BERT[5] [27]，RoBERTa[6] [28]，DeBERTa[7] [29] の 3 種類の事前訓練モデルを用いて品質推定モデルを構築した．事前訓練モデルごとに，品質推定のみを訓練するベースライン，Wikipedia での擬似的な品質推定の訓練後に品質推定のファインチューニングを行う pseudoQE (Wikipedia)，同様に Newsela を用いる pseudoQE (Newsela) の 3 種類の品質推定モデルを訓練した．

擬似的な品質推定の追加訓練においては，バッチサイズは 1,024，学習率は 5e-5，損失は平均二乗誤差，最適化手法は AdamW [33] を用いて，回帰タスクを訓練した．3 エポックの訓練を行い，検証データにおける Accuracy（ラベルの閾値は 0.5 に設定）が最も高いエポックのモデルを品質推定タスクのファインチューニングに使用した．

品質推定においては，バッチサイズは 32，損失は平均二乗誤差，最適化手法は AdamW を用いて，検証データにおけるピアソン相関が 10 エポック連続で改善しなくなるまで訓練し，最も相関の高いエポックのモデルを選択した．学習率は，5e-5, 4e-5,

3) https://github.com/chaojiang06/wiki-auto
4) https://github.com/cocoxu/simplification
5) https://huggingface.co/bert-base-uncased
6) https://huggingface.co/roberta-base
7) https://huggingface.co/microsoft/deberta-base

表 3　追加訓練のタスク設定ごとの平易性の品質推定性能

	分類	回帰
BERT	0.482	
+ pseudoQE (Wikipedia)	0.503	**0.535**
+ pseudoQE (Newsela)	0.470	**0.556**
RoBERTa	0.543	
+ pseudoQE (Wikipedia)	0.517	**0.545**
+ pseudoQE (Newsela)	0.568	**0.578**
DeBERTa	0.473	
+ pseudoQE (Wikipedia)	0.522	**0.598**
+ pseudoQE (Newsela)	0.519	**0.622**

3e-5, 2e-5 の中から，検証データにおけるピアソン相関が最高となる値を選択した．各モデルは，シード値を変更して 5 回ずつ実験を行い，最大値と最小値を除いた 3 回の評価の平均値を報告する．

比較手法　機械学習に基づく 2 つの既存手法と，深層学習に基づく提案手法の性能を比較する．Kajiwara-17 [10] は，scikit-learn[8] を用いて実装した．Martin-18 [11] は，著者らの実装[9] を用いた．

4.2　実験結果

実験結果を表 2 に示す．全ての事前訓練モデルおよびドメインにおいて，提案手法が平易性に関する品質推定の性能を一貫して改善できた．また，提案手法は全ての観点において，既存の品質推定手法の性能を一貫して上回った．本研究では，平易性に関連する擬似的な品質推定の訓練を追加したため，平易性に関する品質推定の性能が改善することを期待していたが，BERT および DeBERTa においては，同義性にも改善が見られた．

4.3　分析

表 3 に，擬似的な品質推定の追加訓練を，分類タスクとして実施する場合と回帰タスクとして実施する場合の性能の比較を示す．平易性に関する品質推定の性能を比較したところ，追加訓練を回帰タスクとして実施した方が一貫して高い性能を達成した．また，分類タスクとしての追加訓練では，追加訓練なしのベースラインよりも性能が悪化することもあった．これらの結果から，目的タスクに近い設定で追加訓練を行うと良いことが示唆される．

8）https://scikit-learn.org
9）https://github.com/facebookresearch/text-simplification-evaluation

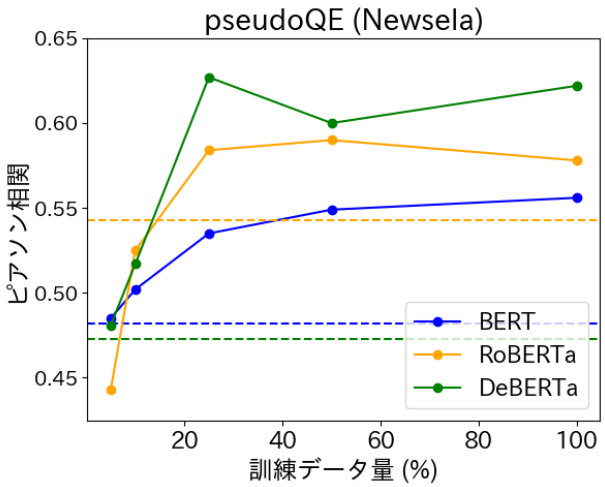

図 2　追加訓練のデータ量ごとの平易性の品質推定性能

図 2 に，追加訓練のデータサイズを 50%，25%，10%，5%と削減した際の平易性に関する品質推定の性能の変化を示す．全ての事前訓練モデルにおいて，10 万文対（25%）までは性能が改善するものの，それ以降は性能向上が鈍化することがわかる．また，点線で示した追加訓練なしのベースラインと比較すると，BERT および DeBERTa においては 10%の 4 万文対，RoBERTa においても 10 万文対のパラレルコーパスがあれば，提案手法によって品質推定の性能を改善できることがわかる．

5　おわりに

本研究では，小規模なラベル付きデータを用いてテキスト平易化の品質推定を効率的に訓練するために，事前訓練モデルのファインチューニングの前に行う擬似的な品質推定の追加訓練を提案した．提案手法では，品質推定のラベルを持たない一般的なテキスト平易化パラレルコーパスを用いて，文対のどちらがより平易かを評価する訓練を行う．英語のテキスト平易化の品質推定における評価実験の結果から，提案手法によって平易性に関する品質推定の性能が改善するだけでなく，事前訓練モデルによっては同義性も改善できることが明らかになった．

今後の課題として，文法性や同義性に焦点を当てた追加訓練の手法を検討することや，英語以外の言語における品質推定へ取り組むことが考えられる．

謝辞

本研究は JSPS 科研費（基盤研究 B，課題番号：JP22H03651）の助成を受けたものです．

参考文献

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

[2] Jan De Belder and Marie-Francine Moens. Text Simplification for Children. In **Proc. of the SIGIR 2010 Workshop on Accessible Search Systems**, pp. 19–26, 2010.

[3] Sarah E Petersen and Mari Ostendorf. Text Simplification for Language Learners: A Corpus Analysis. In **Proc. of SLaTE**, pp. 69–72, 2007.

[4] Makoto Miwa, Rune Sætre, Yusuke Miyao, and Jun'ichi Tsujii. Entity-Focused Sentence Simplification for Relation Extraction. In **Proc. of COLING**, pp. 788–796, 2010.

[5] Sanja Štajner and Maja Popovic. Can Text Simplification Help Machine Translation? In **Proc. of EAMT**, pp. 230–242, 2016.

[6] Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. Optimizing Statistical Machine Translation for Text Simplification. **TACL**, Vol. 4, pp. 401–415, 2016.

[7] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a Method for Automatic Evaluation of Machine Translation. In **Proc. of ACL**, pp. 311–318, 2002.

[8] 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.

[9] Sanja Štajner, Maja Popovic, Horacio Saggion, Lucia Specia, and Mark Fishel. Shared Task on Quality Assessment for Text Simplification. In **Proc. of QATS**, pp. 22–31, 2016.

[10] Tomoyuki Kajiwara and Atsushi Fujita. Semantic Features Based on Word Alignments for Estimating Quality of Text Simplification. In **Proc. of IJCNLP**, pp. 109–115, 2017.

[11] Louis Martin, Samuel Humeau, Pierre-Emmanuel Mazaré, Éric de La Clergerie, Antoine Bordes, and Benoît Sagot. Reference-less Quality Estimation of Text Simplification Systems. In **Proc. of ATA**, pp. 29–38, 2018.

[12] Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. The (Un)Suitability of Automatic Evaluation Metrics for Text Simplification. **CL**, Vol. 47, No. 4, pp. 861–889, 2021.

[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] Philipp Koehn, Franz J. Och, and Daniel Marcu. Statistical Phrase-Based Translation. In **Proc. of NAACL**, pp. 127–133, 2003.

[15] Lucia Specia. Translating from Complex to Simplified Sentences. In **Proc. of PROPOR**, pp. 30–39, 2010.

[16] Sander Wubben, Antal van den Bosch, and Emiel Krahmer. Sentence Simplification by Monolingual Machine Translation. In **Proc. of ACL**, pp. 1015–1024, 2012.

[17] Tomoyuki Kajiwara and Mamoru Komachi. Building a Monolingual Parallel Corpus for Text Simplification Using Sentence Similarity Based on Alignment between Word Embeddings. In **Proc. of COLING**, pp. 1147–1158, 2016.

[18] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to Sequence Learning with Neural Networks. In **Proc. of NIPS**, pp. 3104–3112, 2014.

[19] Sergiu Nisioi, Sanja Štajner, Simone Paolo Ponzetto, and Liviu P. Dinu. Exploring Neural Text Simplification Models. In **Proc. of ACL**, pp. 85–91, 2017.

[20] Xingxing Zhang and Mirella Lapata. Sentence Simplification with Deep Reinforcement Learning. In **Proc. of EMNLP**, pp. 584–594, 2017.

[21] Daiki Nishihara, Tomoyuki Kajiwara, and Yuki Arase. Controllable Text Simplification with Lexical Constraint Loss. In **Proc. of ACL-SRW**, pp. 260–266, 2019.

[22] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, L ukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. In **Proc. of NIPS**, pp. 5998–6008, 2017.

[23] Sanqiang Zhao, Rui Meng, Daqing He, Andi Saptono, and Bambang Parmanto. Integrating Transformer and Paraphrase Rules for Sentence Simplification. In **Proc. of EMNLP**, pp. 3164–3173, 2018.

[24] Tomoyuki Kajiwara. Negative Lexically Constrained Decoding for Paraphrase Generation. In **Proc. of ACL**, pp. 6047–6052, 2019.

[25] Daiki Yanamoto, Tomoki Ikawa, Tomoyuki Kajiwara, Takashi Ninomiya, Satoru Uchida, and Yuki Arase. Controllable Text Simplification with Deep Reinforcement Learning. In **Proc. of AACL-IJCNLP**, pp. 398–404, 2022.

[26] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space. In **Proc. of ICLR**, 2013.

[27] 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.

[28] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach. **arXiv:1907.11692**, 2019.

[29] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. DeBERTa: Decoding-enhanced BERT with Disentangled Attention. In **Proc. of ICLR**, 2021.

[30] Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization. In **Proc. of ICML**, 2020.

[31] Tomoyuki Kajiwara, Biwa Miura, and Yuki Arase. Monolingual Transfer Learning via Bilingual Translators for Style-Sensitive Paraphrase Generation. In **Proc. of AAAI**, pp. 8042–8049, 2020.

[32] Yuki Arase and Jun'ichi Tsujii. Transfer Fine-Tuning: A BERT Case Study. In **Proc. of EMNLP-IJCNLP**, pp. 5393–5404, 2019.

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