QENTS：テキスト平易化の品質推定のためのデータセット

廣中 勇希 † 井川 朋樹 ‡ 梶原 智之 ‡ 二宮 崇 ‡

† 愛媛大学工学部情報工学科 ‡ 愛媛大学大学院理工学研究科
{hironaka, ikawa}@ai.cs.ehime-u.ac.jp {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ための英語データセットを構築し，公開した．本タスクにおける既存のデータセットには，数百文と小規模であり，また現在主流である深層学習に基づくテキスト平易化モデルを対象としていないという課題があった．我々は，深層学習に基づく手法を含む9種類の代表的なテキスト平易化モデルを対象に，10,770文に対して文法性・同義性・平易性・総合評価の人手評価を行い，これらの課題を解決した．

1 はじめに

テキスト平易化 [1] は，意味を保持したまま難解な文法的表現や語句を平易に変換するタスクである．自動的な文の平易化システムは，子どもや語学学習者などの学習支援や読解支援に貢献する．

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 [2] などの自動評価によって評価されている．しかし，前者にはコストや再現性の課題があり，後者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使用者は参照文を持っていない場合が多く，SARIなどの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 [3–5] が研究されている．

2016年に開催されたShared Task on Quality Assessment for Text Simplification (QATS) [3] では，テキスト平易化の品質推定のためのデータセットが構築された．QATSデータセットでは，631文の英語文とそれに対応する平易化モデル [6–10] の出力文に対して，文法性・同義性・平易性・総合評価の4観点からGood・OK・Badの3段階の人手評価が付与されている．テキスト平易化の品質推定に関する以降の研究 [4, 5] ではQATSデータセットが用いられているものの，小規模である点および現在主流となっている深層学習に基づくテキスト平易化モデルが対象となっていない点の2点から，QATSデータセットは近年のテキスト平易化モデルのための品質推定には適していないと考えられる．

本研究では，これらの課題を解決するテキスト平易化の品質推定のための新たなデータセットQENTS[1)] (Quality Estimation for Neural Text Simplification) を構築し，BERT [11] による品質推定のベンチマークを行う．本データセットは，深層学習に基づく手法を含む9種類のテキスト平易化モデルを対象に，10,770文対の英語文とシステム出力文について，文法性・同義性・平易性・総合評価の4観点から4段階で人手評価値を付与したものである．

2 関連研究

難解な英文と平易な英文の文対 [12] 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年から2016年ごろまで，フレーズベース統計的機械翻訳 [13] に基づくテキスト平易化 [7, 14–16] が研究された．ニューラル機械翻訳 [17] の成功を受け，2017年から深層学習に基づくテキスト平易化 [18–20] が研究されている．近年は，Transformer [21] に基づくテキスト平易化 [22, 23] が主流となっている．QATSデータセット [3] では，フレーズベース統計的機械翻訳に基づく初期のテキスト平易化モデル [7] を対象に人手評価を実施しているため，深層学習に基づく近年のテキスト平易化モデルの品質推定には適していないと考えられる．本研究では，深層学習に基づく手法を含む代表的なテキスト平易化モデルを対象に人手評価を行い，大規模な品質推定のデータセット（表1）を構築する．

品質推定の先行研究としては，QATSデータセット [3] 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テキスト平易化の品質

1） https://github.com/yu-hiro/qents

表 1　本研究で構築した QENTS データセットの例

入力文：Weariness tears up their voices , but they ’re still on the freedom highway .					
モデル	出力文	文法性	同義性	平易性	総合評価
Hybrid	weariness tears up voices they ’re .	2	1	3	1
EditNTS	weariness tears up their voices , but they ’re still on the street .	4	2	3	2
BERT	they ’re still on the freedom highway .	4	3	4	3

推定モデルが訓練されている．Kajiwara and Fujita [4] は，単語分散表現 [24] 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. [5] は，BLEU [25] などの機械翻訳の評価指標や FKGL [26] などのリーダビリティ指標に基づく素性抽出を行い，回帰モデルとしての品質推定を行った．本研究では，QENTS データセットのベンチマークとして，先行研究 [5] の適用に加えて，深層学習に基づく強力なベースラインである BERT [11] による回帰モデルの性能を調査する．

3　データセットの構築

3.1　対象とするテキスト平易化モデル

本研究では，2 節で紹介した代表的なテキスト平易化モデルのうち，以下の 9 種類を用いて品質推定のためのデータセットを構築する．

フレーズベース統計的機械翻訳（PBMT）に基づく手法のうち，PBMT-R [15] および Hybrid [16] の 2 モデルを用いる．PBMT-R は，PBMT モデルの出力を入力文との非類似度によってリランキングする手法である．Hybrid は，文分割などの前処理を行った後に PBMT モデルによる平易化を行う手法である．

RNN ベースのニューラル機械翻訳に基づく手法のうち，EncDecA [18]・DRESS [18]・S2S-All-FA [19]・EditNTS [20] の 4 モデルを用いる．EncDecA は，注意機構によるニューラル機械翻訳モデル [17] に基づく手法である．DRESS は，EncDecA モデルを強化学習によって再訓練する手法である．S2S-All-FA は，EncDecA モデルの出力を単語難易度によってリランキングする手法である．EditNTS は，単語の追加・削除・保持の明示的な編集操作を RNN によって推定する編集ベースの手法である．

Transformer ベースのニューラル機械翻訳に基づく手法のうち，Transformer [23]・DMASS [22]・BERT [23] の 3 モデルを用いる．Transformer は，自己注意機構によるニューラル機械翻訳モデル [21] に基づく手法である．DMASS は，Transformer モデルに言い換え知識 [27] を統合した手法である．BERT は，Transformer モデルの符号化器として事前訓練された BERT [11] を用いる手法である．

全てのモデルは Newsela コーパス [12] を用いて訓練されている．PBMT-R・Hybrid・EncDecA・DRESS の 4 モデルは，Zhang and Lapata [18] によって公開[2] されている出力文を用いた．S2S-All-FA および DMASS は，Kriz et al. [19] によって公開[3] されている出力文を用いた．EditNTS は，Dong et al. [20] によって公開[4] されている出力文を用いた．Transformer および BERT は，Jiang et al. [23] によって公開[5] されている出力文を用いた．

3.2　人手評価のアノテーション

Newsela コーパス [12, 18] の評価用データ 1,077 文に対応する，3.1 節で述べた 9 種類のテキスト平易化モデルによる出力文および参照文の合計 10,770 文について，文法性・同義性・平易性・総合評価の 4 観点の人手評価を行う．以降では，Newsela コーパスに含まれる入力文を難解文，テキスト平易化モデルによる出力文および Newsela コーパスに含まれる参照文を平易文と表記する．以下に評価基準を示す．

文法性　平易文の文法的な正しさを評価する．Xu et al. [2] の人手評価の基準（4. 文法的である，3. 1-2 件の文法誤りを含む，2. 複数件の文法誤りを含む，1. 文法的ではない）に従って 4 段階評価した．

同義性　難解文と平易文の間の意味的な等価性を評価する．Xu et al. [2] の人手評価の基準（4. 同一，3. わずかに異なる，2. 異なる，1. 大幅に異なる）に従って 4 段階評価した．

平易性　難解文と比較した際の平易文の理解しやすさを評価する．文法性や同義性と同様に，4 段階（4. 理解しやすい，3. わずかに理解しやすい，2. 変化なし，1. 理解しにくい）で評価した．

2）https://github.com/XingxingZhang/dress
3）https://github.com/rekriz11/sockeye-recipes
4）https://github.com/yuedongP/EditNTS
5）https://github.com/chaojiang06/wiki-auto

表 2　Quadratic Weighted Kappa によるアノテーションの一致率

	文法性	同義性	平易性	総合評価	全体の一致率
評価者 A vs. 評価者 B	0.710	0.677	0.603	0.579	0.688
評価者 A vs. 評価者 C	0.666	0.680	0.749	0.514	0.679
評価者 A vs. 評価者 D	0.724	0.805	0.711	0.626	0.736

総合評価　文法性・同義性・平易性の 3 つの評価を踏まえて，4 段階で総合評価を行った．

クラウドソーシングの Amazon Mechanical Turk[6] を用いて，評価者を雇用した．データセットの品質を担保するために，US 在住者のうち，質の高い回答を行う Master 資格を保有し，過去のタスク承認率が 95%以上の評価者を採用した．また，10 文× 3 モデル（PBMT-R・DRESS・参照文）の小規模な予備アノテーションを実施し，10 人の評価者候補のうち外れ値となる 2 人を除外した．残りの 8 人の評価者候補の中から 1 人の評価者を選び，10,770 文の全体のアノテーションを依頼した．なお，評価者には時給 7.5USD と見積もり合計 700USD の報酬を支払った．

3.3　アノテーションの信頼性評価

3.2 節の人手評価は 1 人の評価者（評価者 A）が行ったため，本節ではアノテーションの信頼性を評価する．同じく Amazon Mechanical Turk を用いて，3.2 節の条件を満たす評価者を新たに 3 名（評価者 BCD）雇用し，QENTS データセットの一部（10 文 × 10 モデル）へのアノテーションを追加で行った．そして，Quadratic Weighted Kappa を用いて，評価者間のアノテーションの一致率を計算した．

表 2 に評価者間の一致率を示す．文法性・同犠性・平易性においては，$K > 0.6$ の substantial agreement を確認できた．総合評価については $0.4 < K < 0.6$ の moderate agreement も含まれるが，4 つの観点をまとめた全体の一致率としては全評価者間で substantial agreement が得られた．この結果から，テキスト平易化の人手評価は評価者間の揺れが少なく，3.2 節のアノテーションも充分に一般的な評価だと言える．

4　アノテーション結果の分析

4.1　テキスト平易化モデルの品質

表 3 に，各テキスト平易化モデルの人手評価および自動評価の結果を示す．自動評価には，テキスト平易化の研究においてよく使用される SARI [2]・BLEU [25]・FKGL [26] を用いる．また，変換の非積極性を表す selfBLEU（入出力間の BLEU）も用いる．

初期のモデル（PBMT-R および EncDecA）は selfBLEU が高く，入力文の大部分を出力文にコピーしている．これらは同義性が高いものの，平易性が低く FKGL が高いため，平易な出力とは言えない．

対照的に，近年の RNN ベースのモデル（S2S-All-FA および EditNTS）は，同義性は低いものの，平易性が高く FKGL が低いため，平易な出力を行っている．これらは参照文と似た傾向であると言える．

高度な訓練を行ったモデル（強化学習の DRESS および転移学習の BERT）は，最も高品質なモデルであると言える．これらは文法性が高く，参照文と同程度の同義性を持ちつつ，平易性も比較的高い．

4.2　自動評価と人手評価の相関

表 4 に，自動評価と人手評価の間のピアソン相関係数を示す．テキスト平易化の自動評価に最もよく用いられる SARI は，平易性との正の相関を持つものの，文法性とは無相関であり，同義性および総合評価とは負の相関を持つことが明らかになった．SARI は消極的な変換にペナルティを与えるため，同義性の人手評価とは対照的な挙動が見られた．

テキスト平易化の自動評価の文脈では否定的に語られることの多い BLEU は，本研究においては全ての人手評価の項目において正の相関が見られた．これは，BLEU と人手評価の相関が低いと報告している先行研究 [2, 28] が Simple Wikipedia に基づくデータセットを用いている一方で，本研究は Newsela を用いていることが要因と考えられる．Simple Wikipedia が平易に記述されていないことは先行研究 [12] でも指摘されており，Zhang and Lapata [18] の人手評価においても，Simple Wikipedia の参照文は低い平易性と高い同義性を持ち，Newsela の参照文は高い平易性と低い同義性を持つことが報告されている．Newsela のような平易な参照文に対しては，BLEU による自動評価が人手評価との良い相関を持つと言える．

6）https://www.mturk.com/

表 3　テキスト平易化モデルの人手評価と自動評価の結果

	人手評価				自動評価			
	文法性	同義性	平易性	総合評価	SARI	BLEU	selfBLEU	FKGL
PBMT-R	3.14	**2.82**	1.96	**2.87**	26.24	18.19	75.60	8.13
Hybrid	2.52	1.86	2.60	1.87	34.73	14.46	25.64	4.52
EncDecA	3.43	2.34	2.55	2.41	35.61	21.70	52.91	5.83
DRESS	3.47	2.22	2.97	2.27	38.37	**23.26**	39.96	4.65
S2S-All-FA	3.43	1.73	**3.27**	1.77	**39.80**	19.51	21.96	**3.51**
EditNTS	3.21	1.88	3.08	1.90	39.28	19.96	23.82	3.80
Transformer	2.90	1.71	2.57	1.74	39.21	15.58	26.52	4.47
DMASS	1.76	1.10	1.67	1.11	38.72	11.99	**20.67**	4.44
BERT	**3.51**	2.21	3.04	2.26	39.06	20.74	32.97	4.50
参照文	3.95	2.31	3.44	2.48	-	-	18.30	3.82

表 4　自動評価と人手評価のピアソン相関係数

	文法性	同義性	平易性	総合評価
SARI	0.005	-0.686	0.528	-0.675
BLEU	0.928	0.642	0.659	0.655
selfBLEU	0.351	0.873	-0.343	0.873
FKGL	0.080	0.734	-0.577	0.730

5　品質推定の実験

本研究で構築した QENTS データセットを用いて，テキスト平易化モデルの文単位の品質推定を行う．

5.1　実験設定

データセットは，8,770 件（877 文× 10 モデル）の訓練用データ，1,000 件（100 文× 10 モデル）ずつの検証用データおよび評価用データに分割して実験した．本実験では，文法性・同義性・平易性・総合評価の項目ごとに回帰モデルを訓練した．品質推定モデルの性能は，モデルが推定した評価値と人手評価値の間のピアソン相関係数を用いて自動評価した．

ベンチマークとして，既存の品質推定モデル（Martin-2018） [5] と BERT [11] による品質推定モデルの性能を比較した．Martin-2018 モデルは，EASSE[7] [29] を用いて実装した．BERT モデルは，HuggingFace Transformers[8] [30] を用いて実装した．バッチサイズは 32，学習率は 5e-5，最適化手法は Adam を用いて，30 エポックの訓練を行った．

7）https://github.com/feralvam/easse
8）https://huggingface.co/bert-base-uncased

表 5　品質推定の実験結果（ピアソン相関係数）

	文法性	同義性	平易性	総合評価
Martin-2018	0.356	0.564	0.519	0.331
BERT	**0.780**	**0.835**	**0.696**	**0.833**

5.2　実験結果

実験結果を表 5 に示す．文法性・同義性・平易性・総合評価の全項目において，BERT による品質推定モデルが先行研究 [5] の性能を大幅に上回った．

6　おわりに

本研究では，現在主流の深層学習に基づくテキスト平易化モデルの品質推定を実現するために，品質推定モデルを訓練および評価するための大規模なデータセットを構築した．深層学習モデルを含む 9 種類のテキスト平易化モデルの出力文および人手で平易化された参照文を対象に，10,770 文に対して，文法性・同義性・平易性・総合評価の 4 つの観点から 4 段階の人手評価を行った．

既存のデータセットは 631 文と小規模なため，テキスト平易化の品質推定には深層学習に基づくモデルを適用することが難しかった．本研究では，約 17 倍の規模にデータセットを拡大できたため，深層学習に基づく品質推定モデルを訓練可能になった．

本研究の副産物として，自動評価と人手評価の相関分析から，Newsela のような平易な参照文を使用できる場合には，BLEU による自動評価が人手評価との良い相関を持つという知見が得られた．

謝辞

本研究は JSPS 科研費（若手研究，課題番号：JP20K19861）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. **Computational Linguistics**, Vol. 46, No. 1, pp. 135–187, 2020.

[2] 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.

[3] Sanja Stajner, 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.

[4] 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.

[5] 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.

[6] Goran Glavaš and Sanja Štajner. Event-Centered Simplification of News Stories. In **Proc. of RANLP Student Research Workshop**, pp. 71–78, 2013.

[7] Sanja Štajner, Hannah Béchara, and Horacio Saggion. A Deeper Exploration of the Standard PB-SMT Approach to Text Simplification and its Evaluation. In **Proc. of ACL-IJCNLP**, pp. 823–828, 2015.

[8] Or Biran, Samuel Brody, and Noémie Elhadad. Putting it Simply: a Context-Aware Approach to Lexical Simplification. In **Proc. of ACL**, pp. 496–501, 2011.

[9] Colby Horn, Cathryn Manduca, and David Kauchak. Learning a Lexical Simplifier Using Wikipedia. In **Proc. of ACL**, pp. 458–463, 2014.

[10] Goran Glavaš and Sanja Štajner. Simplifying Lexical Simplification: Do We Need Simplified Corpora? In **Proc. of ACL-IJCNLP**, pp. 63–68, 2015.

[11] 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.

[12] 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.

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

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

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

[16] Shashi Narayan and Claire Gardent. Hybrid Simplification using Deep Semantics and Machine Translation. In **Proc. of ACL**, pp. 435–445, 2014.

[17] Thang Luong, Hieu Pham, and Christopher D. Manning. Effective Approaches to Attention-based Neural Machine Translation. In **Proc. of EMNLP**, pp. 1412–1421, 2015.

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

[19] Reno Kriz, João Sedoc, Marianna Apidianaki, Carolina Zheng, Gaurav Kumar, Eleni Miltsakaki, and Chris Callison-Burch. Complexity-Weighted Loss and Diverse Reranking for Sentence Simplification. In **Proc. of NAACL**, pp. 3137–3147, 2019.

[20] Yue Dong, Zichao Li, Mehdi Rezagholizadeh, and Jackie Chi Kit Cheung. EditNTS: An Neural Programmer-Interpreter Model for Sentence Simplification through Explicit Editing. In **Proc. of ACL**, pp. 3393–3402, 2019.

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

[22] 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.

[23] 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.

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

[25] 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.

[26] 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, DTIC Document**, 1975.

[27] Ellie Pavlick and Chris Callison-Burch. Simple PPDB: A Paraphrase Database for Simplification. In **Proc. of ACL**, pp. 143–148, 2016.

[28] Elior Sulem, Omri Abend, and Ari Rappoport. BLEU is Not Suitable for the Evaluation of Text Simplification. In **Proc. of EMNLP**, pp. 738–744, 2018.

[29] Fernando Alva-Manchego, Louis Martin, Carolina Scarton, and Lucia Specia. EASSE: Easier Automatic Sentence Simplification Evaluation. In **Proc. of EMNLP**, pp. 49–54, 2019.

[30] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-Art Natural Language Processing. In **Proc. of EMNLP**, pp. 38–45, 2020.