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[1], 井川 朋樹[2], 梶原 智之[2], 二宮 崇[2]

[1]愛媛大学工学部情報工学科, [2]愛媛大学大学院理工学研究科

https://github.com/yu-hiro/qents

背景：テキスト平易化

- テキストの意味を保持したまま、難しい文法や単語を簡単にする

  例）原文：The cat perched on the mat.

  平易文：The cat sat on the mat.

- テキスト平易化システムは、子供や語学学習者の学習支援や読解支援につながる

背景：テキスト平易化の自動評価

- BLEU やSARIといった参照文に基づく指標で求められることが多い
- 人手評価と比較し低コストで求められる
- 参照文が必要

- 先行研究[1, 2]では、BLEUと人手評価の相関が低いと報告されている

→人手評価と比較し、信頼度が低い

[1] Xu et al., **Optimizing Statistical Machine Translation for Text Simplification**, TACL, 2016
[2] Sulem et al., **BLEU is Not Suitable for the Evaluation of Text Simplification**, EMNLP, 2018

背景：テキスト平易化の人手評価

- 文法性 (G), 同義性 (M), 平易性 (S), 総合評価 (Overall)の4つの観点で評価
- 出力した平易文を3~5人の英語母語話者が4~5段階で評価
- 自動評価と比較し信頼度が高い
- 自動評価と比較しコストがかかる

→人手評価を自動化する品質推定の研究が行われている

関連研究：品質推定の既存データセット

QATSデータセット[3]

- 631の英文の文対
- フレーズベース統計的機械翻訳に基づく手法が評価対象
- 文法性 (G), 同義性 (M), 平易性 (S), 総合評価 (Overall)の4つの観点について Good, OK, Bad の3 段階で人手評価

課題:
・631文対と小規模である
・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

[3] Stajner et al., **Shared Task on Quality Assessment for Text Simplification**, LREC, 2016

関連研究：品質推定の既存手法

- Kajiwara-2017[4]
  - 単語分散表現に基づく特徴量を使用
  - SVMを使用し，Good, OK, Bad の3 クラス分類のモデルを作成
- Martin-2018[5]
  - BLEUなどの機械翻訳の評価指標やFKGLなどのリーダビリティ指標に基づく特徴量を使用
  - SVMなどを使用し，回帰モデルと分類モデルを作成

課題:
・既存データセットが小規模のため、深層学習に基づく品質推定モデルの学習が難しい

[4] Kajiwara and Atsushi Fujita, **Semantic Features Based on Word Alignments for Estimating Quality of Text Simplification**, IJCNLP, 2017
[5] Martin et al., **Reference-less Quality Estimation of Text Simplification Systems**, ATA, 2018

本研究の目的

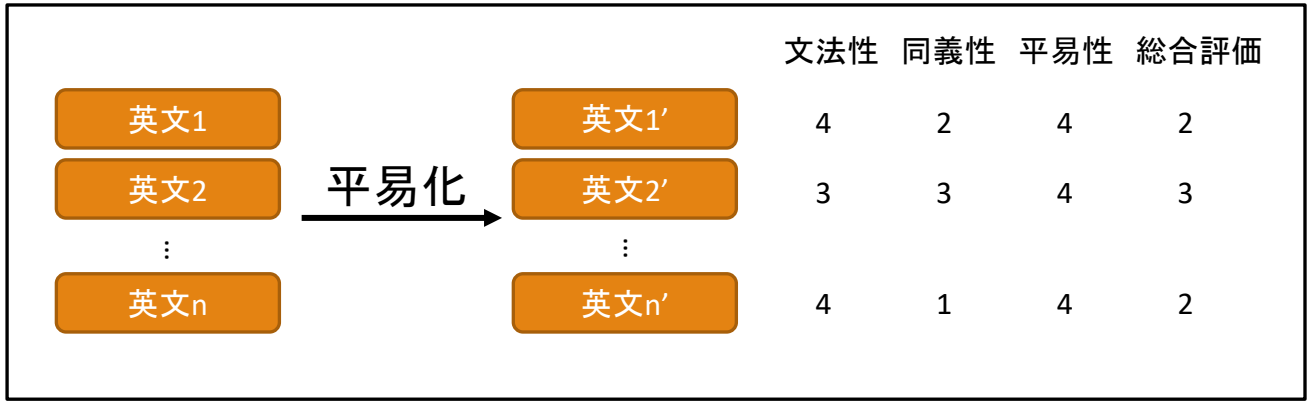
品質推定の性能向上

- 新たな大規模データセットの構築
  - 現在主流の深層学習に基づくテキスト平易化モデルを対象
  - データセットの規模を**約17倍**に拡張 (631文対→10,770文対)

- 最新の深層学習モデルを用いた品質推定器の作成
  - 構築したデータセットを用いて，BERTを学習する
  - 既存手法との比較実験を行う

本研究の流れ

①データセット構築



9種類のテキスト平易化モデルの
出力文と参照文に対して、
人手評価を実施

②評価実験

- 品質推定器（文法性）
- 品質推定器（同義性）
- 品質推定器（平易性）
- 品質推定器（総合評価）

構築したデータセットを用いて品質推定器を作成

品質推定器の出力と人手評価の相関を計算

[3] Stajner et al., **Shared Task on Quality Assessment for Text Simplification**, LREC, 2016

データセット構築

QENTSデータセットの構築

- Amazon Mechanical Turk を使用
- US在住, Master資格保有, 過去のタスク承認率が95%以上の評価者1名を雇用
  （事前に小規模のアノテーションを実施し，10名の候補者の中から評価者を選定）
- 深層学習に基づく代表的なテキスト平易化モデルの出力文が対象
- Newselaコーパス[6] の評価用データ1,077文×10モデルの出力文を人手評価
- 文法性 (G), 同義性 (M), 平易性 (S), 総合評価 (Overall)の4つの観点で4段階で評価

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

[6] Xu et al., **Problems in Current Text Simplification Research: New Data Can Help** , TACL, 2015

対象とするテキスト平易化モデル

- フレーズベース統計的機械翻訳 (PBMT) に基づく手法
  - PBMT-R
  - Hybrid
- RNNベースのニューラル機械翻訳に基づく手法
  - EncDecA
  - DRESS
  - S2S-Attl-FA
  - EditNTS
- Transformerベースのニューラル機械翻訳に基づく手法
  - Transformer
  - DMASS
  - BERT

対象とするテキスト平易化モデル

フレーズベース統計的機械翻訳 (PBMT) に基づく手法

- PBMT-R：PBMTの出力を，入力との非類似度で
  リランキングする手法
- Hybrid：文分割などの前処理を行った後，PBMTモデルで
  平易化する手法

対象とするテキスト平易化モデル

RNNベースのニューラル機械翻訳に基づく手法

- EncDecA：注意機構によるニューラル機械翻訳に基づく手法
- DRESS：EncDecAを強化学習によって再訓練する手法
- S2S-All-FA：EncDecAの出力を単語難易度によって
  リランキングする手法
- EditNTS：単語の編集操作をRNNによって推定する手法

対象とするテキスト平易化モデル

Transformerベースのニューラル機械翻訳に基づく手法

- Transformer：自己注意機構によるニューラル機械翻訳に基づく手法
- DMASS：Transformerに言い換え知識を統合した手法
- BERT：Transformerの符号化器として事前訓練された
  BERTを用いる手法

データセットの信頼性評価

- US在住, Master資格保有, 過去のタスク承認率が95%以上の評価者3名を雇用し、構築したデータセットの一部（100文）を人手評価
- Quadratic Weighted Kappa[7]を用いて、人手評価の一致率を計算

	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

→全体的に高い一致率が確認できた

[7] Jacob Cohen, **Weighted Kappa: Nominal Scale Agreement Provision for Scaled Disagreement or Partial Credit**, Psychological Bulletin, 1968

データセットの分析１

テキスト平易化モデルの人手評価の平均と自動評価を計算

	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

- SARI↑：
  入力文，出力文，参照文のn-gramから正しく編集された割合
- selfBLEU↓：
  入力文と出力文から計算するBLEU
- FKGL↓：
  文の難易度を数値化した指標

データセットの分析１

テキスト平易化モデルの人手評価の平均と自動評価を計算

	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.**
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

・同義性と selfBLEU が高い
→入力文の大部分を出力文にコピーしている

・平易性が低く、FKGLが高い
→平易な出力とは言えない

データセットの分析１

テキスト平易化モデルの人手評価の平均と自動評価を計算

	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

・同義性が低い

・平易性が高く、FKGLが低い
→平易な出力と言える

データセットの分析１

テキスト平易化モデルの人手評価の平均と自動評価を計算

	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

・文法性が高い
・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

→高品質なモデル

データセットの分析２

自動評価と人手評価のピアソン相関係数を計算

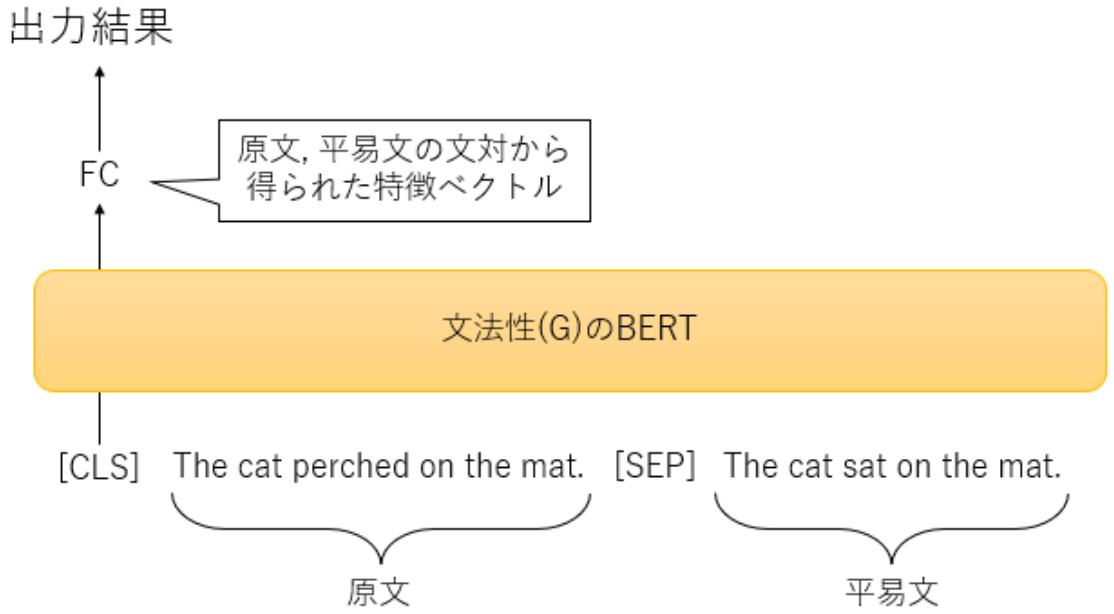
- SARIは，平易性と正の相関を持つ一方，同義性および総合評価とは負の相関を持つ
- BLEUは，すべての人手評価の項目において正の相関が見られた

	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

評価実験

品質推定器の作成

- 事前学習済みのBERT[8]を用いて，平易化した英文の品質推定器を作成
  - 文法性 (G), 同義性 (M), 平易性 (S), 総合評価 (Overall)に関して，それぞれの回帰モデルを作成



[8] Devlin et al., **BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding**, NAACL, 2019

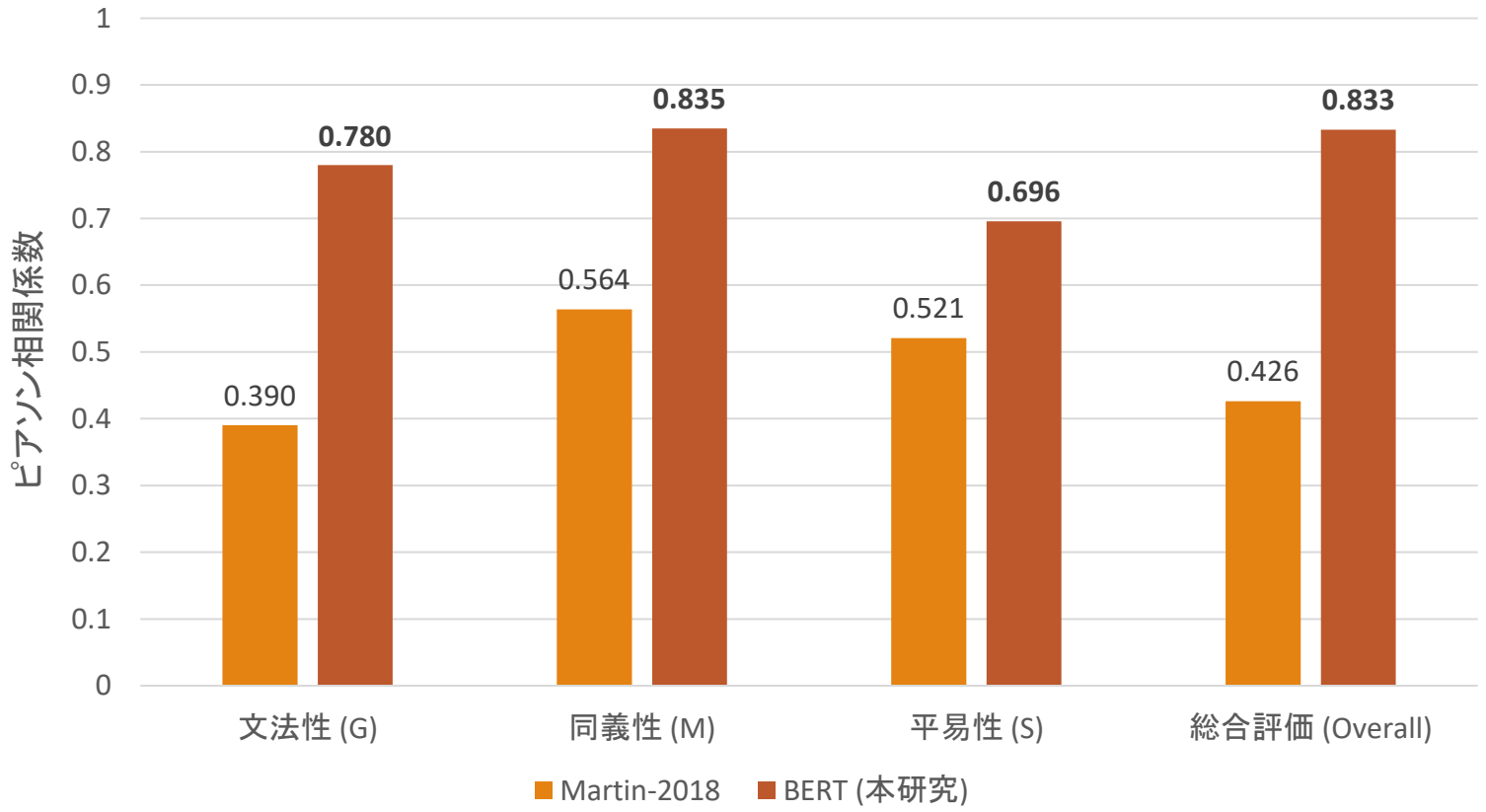
実験設定

- データセット：QENTSデータセット
- 提案手法：BERTによる回帰モデル
- 比較手法：Martin-2018 で提案された回帰モデル
- 評価指標：人手評価との比較（ピアソン相関係数）

データセット

Train	Dev	Test
8770文	1000文	1000文

実験結果


1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
ピアソン相関係数
0.390
0.780
0.564
0.835
0.521
0.696
0.426
0.833
文法性 (G)
同義性 (M)
平易性 (S)
総合評価 (Overall)
Martin-2018
BERT (本研究)

まとめ

https://github.com/yu-hiro/qents

- 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を構築
  - 深層学習に基づくテキスト平易化モデルを対象
  - データセットの規模を**約17倍**に拡張（631文対→10,770文対）

- BERTによる品質推定器を作成し，既存手法の性能を上回る結果となった