テキスト平易化における自動評価指標のメタ評価の検討

早川 明男 [1]　大内 啓樹 [1,3]　梶原 智之 [2]　渡辺 太郎 [1]
[1] 奈良先端科学技術大学院大学　[2] 愛媛大学　[3] 理化学研究所
hayakawa.akio.gv6@is.naist.jp
hiroki.ouchi@is.naist.jp　kajiwara@cs.ehime-u.ac.jp　taro@is.naist.jp

概要

テキスト平易化は，語彙の置換や非重要部分の削除などの操作によって文を平易に書き換えることを目的とする．こうした平易化の操作の多様さは，出力文の自動評価を難しくしてきた．平易化されたテキストを評価する観点として，平易性・同義性・文法性が挙げられる．本研究では，自動評価指標のメタ評価の一環として，各観点について妥当な指標が満たす条件を検討する．また既存の指標の振る舞いを確認し，それらがどのように利用されるべきかを提言する．

1　はじめに

テキスト平易化は文の意味を保ちつつ平易に書き換えるタスクであり，語彙の平易化，構文の平易化，文の分割，非重要部分の削除などの操作によって実現される [1]．平易なテキストは，子ども [2] や非母語話者 [3]，また自閉症 [4] や読字障害 [5] など，読解能力が低い人を補助すると考えられ，テキスト平易化への期待は高まっている．一方，こうした人の属性ごとに読解をより促進するテキストの種類は異なりうる．例えば，子どもに対しては平易な構造の文が，非母語話者に対しては平易な語彙による文がより読解を促進する，といった差が生じる可能性がある．

かねてのテキスト平易化研究は多様な操作の一部を実現できるにすぎなかったが，近年の文生成の発展により，平易文の多様さについて出力を制御できるモデルが提案されるようになった [6, 7]．一方で，その多様さゆえに平易文の評価は難しい．平易 (出力) 文の人手評価においては，複雑 (入力) 文と比較した際の平易さ (**平易性**)，重要部分の意味を保持できている程度 (**同義性**) に加えて，出力文そのものの文法的正しさ (**文法性**) という 3 つの観点を利用することが確立している [8] が，自動評価に利用する指標については明確な合意がないのが現状である．

テキスト平易化の自動評価指標としては，文生成タスクで一般に利用されるもののほか，テキスト平易化のために設計されたものもいくつか提案されている [9, 10]．こうした提案の多くは，その妥当性の根拠として，人手評価と自動評価の相関の高さを挙げる．このとき，ある入力文に対する出力文を平易化モデルなどを利用して生成し，それらに人手でスコアをつけたデータセットが利用される．しかしながら，この人手スコアデータセットそのものが問題をはらむ場合がある．とりわけ平易文の多様化という主題に対して，出力文が特定の操作のみを含んでいる，という指摘 [11] はデータセットの根底を揺るがすものである．こうした現状は評価指標，そして提案モデルの妥当性に疑問を投げかけ，テキスト平易化研究の発展を妨げているといえる．

ところで，自動評価指標の妥当性を示す手段は人手評価との相関に限らない．本研究では，出力文の人手評価に用いられる平易性・同義性・文法性という 3 つの観点それぞれについて，自動評価指標の妥当性を示す他の要素を検討する．また自動評価指標のメタ評価として，既存の指標がそれらを満たす程度を確認する．その際，とくに多様な操作について意識的に要素を検討する．しかしながら，複雑文と平易文のペアから操作を自動的に分類するのは難しい [12]．そこで，それぞれの操作を直接的に分類するのではなく，入力文から出力文への書き換えにおける各単語の編集 (置換，挿入，削除) をもとに編集タグを設定し，検討の材料とする．

2　関連研究

機械翻訳など他の文生成タスクと同様，テキスト平易化研究の多くで BLEU [13] が評価に利用されている．しかしながら，BLEU はテキスト平易化の評価には不適であるとしばしば指摘されている．例えば [14] は，とくに複数の参照文があるとき，入力文をそのまま出力する低質なシステムでも BLEU による評価が高くなってしまうとしている．

こうした BLEU の欠陥を補うため，[9] では SARI という新たな評価指標を提案している．SARI は入力

文との n-gram の重複が大きい出力文にペナルティを与えるように設計され，語彙の平易化の妥当性を評価できるという著者らの主張から，以降テキスト平易化タスクのほとんどで利用されている．一方 [15] は，とくに複数の操作を含む書き換えについて，同義性や文法性についての人手評価と SARI との相関の低さを指摘している．

FKGL [16] は英文のリーダビリティを示すために設計された指標で，テキスト平易化の評価にもしばしば利用されている．この FKGL についても，参照文を必要としないその特徴から，高い評価を得られる出力文を容易に生成できるとして，テキスト平易化の評価指標としては不適であるとの指摘がある [17]．

その他にも BERTScore [18, 11]，SAMSA [10]，QUESTEVAL [19] などさまざまな自動評価指標の導入が提案され，今後の増加も予想される．これらの指標も出力文の質に応じた評価を与えることが期待され，その程度や応用範囲についてメタ評価を受けるべきである．本研究ではメタ評価の要素を検討し，また実施することを目標とする．

3 手法

本研究では，入力文に対して多様なタイプの出力文の集合を作成し，各指標で評価する．同時に，入力文から出力文への編集を計算し，出力文に編集タグを付与した上で，タグごとの出力文の評価も行う．

これらをもとに，平易性・同義性・文法性という 3 つの観点について，各指標が評価できるべき要素を仮定する．また，それらを満たす程度をもとにメタ評価を行う．

3.1 編集タグ

出力文に付与される編集タグは，編集距離と同時に得られる編集列をもとに決定される．置換・挿入・削除の割合 r_{ADD}，r_{DEL}，r_{SUB} を以下のように定義する．

$$
\begin{aligned}
r_{\mathrm{ADD}} &= \text{編集列中の挿入数} \;/\; \text{入力文の単語数} \\
r_{\mathrm{DEL}} &= \text{編集列中の削除数} \;/\; \text{入力文の単語数} \\
r_{\mathrm{SUB}} &= \text{編集列中の置換数} \;/\; \text{入力文の単語数}
\end{aligned}
$$

このとき，以下のような条件で編集タグを設定し，出力文ごとに条件を満たすタグすべてを付与する．

$$
\begin{aligned}
&\mathrm{ADD} : r_{\mathrm{ADD}} > 0.1 \\
&\mathrm{DEL} : r_{\mathrm{DEL}} > 0.2 \\
&\mathrm{SUB} : r_{\mathrm{SUB}} > 0.2 \\
&\mathrm{NONE} : r_{\mathrm{ADD}} \leq 0.1, r_{\mathrm{DEL}} \leq 0.2, r_{\mathrm{SUB}} \leq 0.2
\end{aligned}
$$

それぞれの文は，NONE を除いた複数のタグが付与されうる．図 1 は編集タグのイメージを示したものである．

3.2 利用データ

以下に記述するように，入力文・参照文・出力文を準備する．

3.2.1 入力文・参照文

文を人手で平易化したデータセットに ASSET [15] がある．ASSET は，Wikipedia の英語版に記述された 359 文に対し，10 人のアノテーターそれぞれによる平易な参照文を含む．参照文は多様な操作を含みうるように設計されており，本研究ではこれらを十分妥当な平易文とみなす．

3.2.2 出力文

出力文については，(1) 人手による平易文，(2) 低質なシステムによる出力文の 2 つのタイプを準備する．

1. 人手による平易文
   ASSET のそれぞれの入力文に対する 10 の参照文について，1 文を出力文，残りの 9 文を参照文とみなして評価した結果を平均する．

2. 低質なシステムによる出力文
   以下のように，入力文を利用して簡易に出力文を生成する 5 つのシステムを設定する．Random は，前項で定義した特定の編集タグが付与されるように設計されている．

Identical	: 入力文をそのまま出力する
Split	: 入力文のランダムな位置にピリオドを挿入する
$\mathrm{Random}_{\mathrm{ADD}}$	: 入力文の 1 割にあたる単語列をランダムな位置に挿入する
$\mathrm{Random}_{\mathrm{DEL}}$	: 入力文の 2 割にあたる単語列を削除する
$\mathrm{Random}_{\mathrm{SUB}}$	: 入力文の 2 割にあたる単語列をランダムな単語列に置換する

3.3 利用する自動評価指標

テキスト平易化に用いられる評価指標の代表として，BLEU，SARI，BERTScore，FKGL を選択する．なお，評価に参照文を必要とする BLEU，SARI，

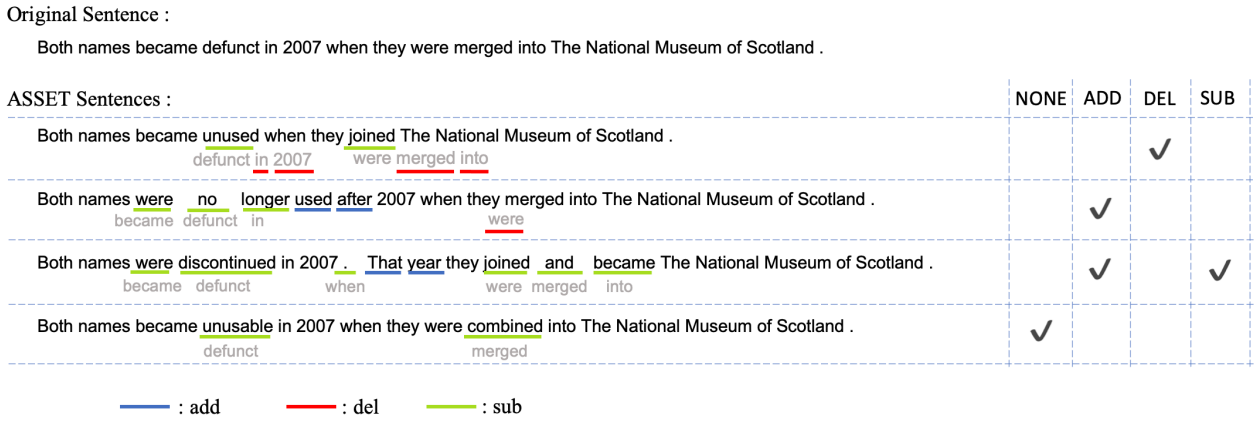

図 1　編集タグのイメージ

BERTScore については，複数の参照文を用いた評価 (Multi) と，単一の参照文を用いた評価の平均値 (Avg) をそれぞれ算出する．各指標の評価値の算出には EASSE [20] の実装を利用する．

3.4　メタ評価の観点

人手評価で用いられる平易性・同義性・文法性の 3 つの観点を中心に，自動評価指標の妥当性を示す要素を以下のように仮定する．

3.4.1　平易性

平易性について，各指標は入力文と比較してより平易な出力文に高い評価を，そうでない出力文に低い評価を与えることが期待される．

本研究では，前者には ASSET が，後者には Identical が該当し，これらを正しく評価できる程度をもってメタ評価を行う．

3.4.2　同義性

同義性について，各指標は入力文における重要な意味が保持されている出力文に高い評価を，そうでない出力文に低い評価を与えることが期待される．すなわち，入力文の重要な意味を同等に保持する出力文は同等に評価されるべきである．

本研究では，ASSET の平易文それぞれがもつ同義性が同等であるとみなすため，構文などによらず評価も同等であることが望ましい．すなわち，編集タグ間の評価の差の小ささをもってメタ評価を行う．

3.4.3　文法性

文法性について，各指標は文法的に正しい出力文を高い評価を，文法的に誤った出力文に低い評価を与えることが期待される．

表 1　各出力文の編集タグ付与割合

	NONE	ADD	DEL	SUB
ASSET	30.36%	21.95%	**41.64%**	33.37%
Identical	**100.00%**	0.00%	0.00%	0.00%
Split	**96.10%**	3.90%	0.00%	0.00%
$\text{Random}_{\text{ADD}}$	0.00%	**100.00%**	0.00%	0.00%
$\text{Random}_{\text{DEL}}$	0.00%	0.00%	**100.00%**	0.00%
$\text{Random}_{\text{SUB}}$	14.48%	0.00%	0.00%	**85.52%**

本研究では，ASSET と Identical が前者に，Split と Random が後者に該当し，これらを正しく評価できる程度をもってメタ評価を行う．

4　結果

4.1　編集タグの分布

表 1 は，それぞれのシステムの出力文に対してどのような割合で編集タグが付与されるかを示したものである．この結果からはまず，人手による平易文 (ASSET) に多様さがあることがわかる．また，低質なシステムの大部分に特定のタグが付与されていることが確認できる．

4.2　自動評価指標のメタ評価

表 2 は各指標で出力文を評価した結果で，太字の数字は指標ごとにもっとも良い評価を意味する．なお，表 2 上部は各システムに対する評価を，下部は ASSET のうち各編集タグが付与された出力文のみに対する評価を表す．以下，3.4 項における要素をもとに，各指標がそれぞれの観点を満たすかを確認する．

表 2　各指標による出力文の評価

	BLEU↑		SARI↑		BERTScore↑		FKGL↓
	Multi	Avg	Multi	Avg	Multi	Avg	
ASSET	65.11	28.37	**42.69**	**41.09**	77.47	61.51	6.95
Identical	**91.60**	**41.36**	20.56	19.48	**91.07**	**69.81**	10.34
Split	80.28	36.50	24.72	23.51	85.02	65.65	**5.45**
Random							
ADD	73.37	33.63	24.78	23.47	75.67	58.42	10.42
DEL	80.37	31.27	32.83	30.94	71.36	54.71	8.33
SUB	61.54	27.88	32.46	30.49	56.88	43.10	9.27
ASSET	65.11	28.37	42.69	41.09	77.47	61.51	6.95
N0NE	**80.97**	**39.37**	41.70	40.67	**86.40**	61.66	7.92
ADD	58.63	26.80	**44.61**	**42.08**	75.55	61.99	6.11
DEL	58.80	22.18	41.82	40.46	72.18	**62.08**	6.70
SUB	48.25	19.39	43.21	41.01	69.84	61.61	**6.06**

4.2.1　BLEU

まず平易性という観点では，Identical を高く評価してしまっている．前提として，ASSET の約 3 割は NONE，すなわち編集の少ない平易文である．こうした参照文は n-gram の重複が大きい Identical を高く評価してしまう．同様の理由で，NONE への評価の突出を説明でき，同義性についての要素を満たさないといえる．また，ASSET より Random に対する評価のほうが高く，文法性についての要素も満たさない．編集タグごとについても同様の関係が見られるが，これは Random における編集部分以外の n-gram の重複によるものといえる．

このように，結果は 3 つの観点すべてについて BLEU の不適さを示している．これは，テキスト平易化の性質である参照文の多様さに由来するといえる．

4.2.2　SARI

まず平易性については，Identical を正しく低く評価できている．これは，入力文に含まれず参照文に含まれる n-gram を含む出力文を高く評価し，入力文に含まれ参照文に含まれない n-gram を含む出力文を低く評価する設計に原因すると考えられる．この設計によって，結果的に編集タグ間の評価の差が小さくなっており，同義性についての要素も満たすといえる．一方で，入力文に含まれる n-gram にペナルティを与えるあまり，Random の評価が Identical より高くなっており，文法性については正しく評価できていないといえる．

4.2.3　BERTScore

平易性については BLEU と同様，BERTScore は Identical を高く評価してしまっている．BERTScore は文全体をとらえた評価をするが，編集の少ない参照文と Identical について，各単語が文全体のなかでもつ意味の差が小さいことが推測される．これは，ASSET に対する Random への評価の低さも説明できる．すなわち，編集を受けていない単語の意味もある程度変化すると考えられる．

同義性については，BLEU と同様に BERTScore-Multi では NONE の評価が高いが，BERTScore-Avg では編集タグ間の評価の差が小さくなっている点には注目したい．これは，BERTScore が構造の異なる文に対しても一定程度その類似性を評価できることを示唆している．

4.3　FKGL

FKGL は Split のようにただ平均単語数が小さい出力文を高く評価してしまい，文法性の説明には不適である．また参照文を必要としないため，同義性を評価するのは難しい．

一方で平易性については，Identical に対して ASSET に良い評価を与えられている．しかし，編集タグ別では ADD や SUB への評価が高くなっており，これはピリオドの挿入・置換によると考えられる．このことは，編集された単語の情報をふまえられないこの手法の弱点を示しているといえる．

5　おわりに

本研究の結果は，メタ評価に利用したどの自動評価指標もいずれかの観点では短所があることを示した．現状では平易化の評価には複数の指標を用いるのが望ましいといえ，例えば文法性については BERTScore で，平易性については SARI で評価する，といった方法が考えられる．

また，これらの考察は本研究で利用したような編集タグでも十分に導くことが可能であり，今後のメタ評価における一つの要素となるといえる．

今後の課題として，やはり多様な操作によって書き換えられた文についての人手スコアデータセットが必要である．多様さの確認においては，本研究で定義したような編集タグを利用することもできるといえよう．

謝辞

本研究は JSPS 科研費 JP19K20351 の助成を受けたものである．

参考文献

[1] Sandra M. Aluísio, Lucia Specia, Thiago A. S. Pardo, Erick G. Maziero, Helena M. Caseli, and Renata P. M. Fortes. A corpus analysis of simple account texts and the proposal of simplification strategies: First steps towards text simplification systems. Association for Computing Machinery, 2008.

[2] Jan De Belder and Marie-Francine Moens. Text simplification for children. Proceedings of the SIGIR workshop on accessible search systems, 2010.

[3] Menglin Xia, Ekaterina Kochmar, and Ted Briscoe. Text readability assessment for second language learners. Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications, 2016.

[4] John Carroll, Guido Minnen, Yvonne Canning, Siobhan Devlin, and John Tait. Practical simplification of english newspaper text to assist aphasic readers. Proc. of AAAI-98 Workshop on Integrating Artificial Intelligence and Assistive Technology, 1998.

[5] Luz Rello, Ricardo Baeza-Yates, Stefan Bott, and Horacio Saggion. Simplify or help? text simplification strategies for people with dyslexia. W4A 2013 - International Cross-Disciplinary Conference on Web Accessibility, 2013.

[6] Louis Martin, Éric de la Clergerie, Benoît Sagot, and Antoine Bordes. Controllable sentence simplification. Proceedings of the 12th Language Resources and Evaluation Conference, 2020.

[7] Louis Martin, Angela Fan, Éric de la Clergerie, Antoine Bordes, and Benoît Sagot. Muss: Multilingual unsupervised sentence simplification by mining paraphrases. arXiv preprint arXiv:2005.00352, 2021.

[8] Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. Data-driven sentence simplification: Survey and benchmark. Computational Linguistics, 2020.

[9] Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. Optimizing Statistical Machine Translation for Text Simplification. Transactions of the Association for Computational Linguistics, 2016.

[10] Elior Sulem, Omri Abend, and Ari Rappoport. Semantic structural evaluation for text simplification. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2018.

[11] Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. The (Un)Suitability of Automatic Evaluation Metrics for Text Simplification. Computational Linguistics, 2021.

[12] Fernando Alva-Manchego, Joachim Bingel, Gustavo Paetzold, Carolina Scarton, and Lucia Specia. Learning how to simplify from explicit labeling of complex-simplified text pairs. Proceedings of the Eighth International Joint Conference on Natural Language Processing, 2017.

[13] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, 2002.

[14] Elior Sulem, Omri Abend, and Ari Rappoport. BLEU is not suitable for the evaluation of text simplification. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018.

[15] Fernando Alva-Manchego, Louis Martin, Antoine Bordes, Carolina Scarton, Benoît Sagot, and Lucia Specia. ASSET: A dataset for tuning and evaluation of sentence simplification models with multiple rewriting transformations. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020.

[16] Kincaid J. Peter, P. Fishburne Robert, L. Rogers Richard, and S. Chissom Brad. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Institute for Simulation and Training, 1975.

[17] Teerapaun Tanprasert and David Kauchak. Flesch-kincaid is not a text simplification evaluation metric. Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021), 2021.

[18] Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. International Conference on Learning Representations, 2020.

[19] Thomas Scialom, Louis Martin, Jacopo Staiano, Éric Villemonte de la Clergerie, and Benoît Sagot. Rethinking automatic evaluation in sentence simplification. arXiv preprint arXiv:2104.07560, 2021.

[20] Fernando Alva-Manchego, Louis Martin, Carolina Scarton, and Lucia Specia. EASSE: Easier automatic sentence simplification evaluation. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations, 2019.