編集操作によるデータ拡張を用いたテキスト平易化の自動評価

山中 光　徳永 健伸
東京工業大学 情報理工学院
{yamanaka.h.ac@m,take@c}.titech.ac.jp

概要

テキスト平易化とは文の意味を保ちつつ平易に書き換えるタスクである．既存のテキスト平易化の研究では様々な自動評価指標が使用されているが，平易性を測定する適切な自動評価指標に関する合意が存在しない．また，既存の自動評価指標は評価するために参照文を必要とするものが多いので，参照文を集めるコストも問題となる．本研究では評価時に参照文を必要としない，テキスト平易化のための新しい自動評価指標を提案する．提案手法では，難易の順序関係を判別するためのランキングモデルを学習し，さらに学習データを拡張するために編集操作を用いたデータ拡張手法を提案する．また，平易性を評価するために適切な評価データについても議論する．評価実験の結果，提案手法が既存の評価指標と同等以上の評価性能を示した．

1　はじめに

テキスト平易化とは文の意味を保ちつつ，語彙平易化や構文平易化などの複数の編集操作を通じて平易に書き換えるタスクであり [1]，子供や非母語話者など平易な文を必要とする人々にとって重要な支援技術の 1 つとしても期待されている [2]．先行研究ではテキスト平易化を難解な文から平易な文への単一言語内の翻訳問題として捉え，近年では深層学習を用いた手法によって流暢な平易文を生成することが可能になっている [3]．

テキスト平易化の評価に関しては，流暢性（出力文の文法的正しさ），同義性（出力文が元の文（原文）の意味を保持している程度），平易性（出力文が原文よりも平易になった程度）という 3 つの観点から人手評価が行われる [4, 1]．既存研究では，これら 3 つの観点において人手評価と相関する自動評価指標を用いて平易化の評価を行う．中でも平易性に関する自動評価指標としては，文中の単語数と平均音節数から計算される FKGL [5]，入出力文と参照文間における N-gram 一致度を測る SARI [6] が主に使用されており，近年では出力文と参照文における BERT の埋め込みで類似性を測る BERTScore [7] を用いることが提案されている [1]．しかし，こういった自動評価指標それぞれに関して，平易性の評価のために用いることが不適切であることが報告されており [8, 9, 10]，平易性のための適切な自動評価指標に関する合意は取れていない．さらに，SARI と BERTScore に関しては計算時に参照文を必要とするため，人手で作られた適切な参照文を収集するコストが存在する．

参照文を用いずに評価値を算出できる自動評価指標は，テキスト平易化のみならず他分野でも重要であり，処理の前後のパラレルデータのみを用いて学習できる枠組みが要約 [11] や文法誤り訂正 [12] の分野で提案されている．これらの研究では，何らかの方法で測った文の良さを順序付けするランク学習と，順序付けのために必要なデータ拡張を行うことで評価モデルを作成している [11, 12]．本研究はこれらの研究を参考にし，原文と参照文から成るパラレルデータのみを用いて学習し，計算時に参照文を用いない，平易化のための新しい自動評価指標 **SIERA**（**SI**mplification metric based on **E**dit operation through learning to **RA**nk）を提案する．

SIERA では，平易性の順序のついた文対を用いてランク学習を行い，さらに評価性能の向上を目的に編集操作に基づいた拡張データも学習に活用する．またテキスト平易化の自動評価指標の評価データから，平易性を測るのに適切な評価データ抽出を，Inter-class correlation（IC）と Inter-annotator agreement（IAA）の観点から行った．実験の結果，拡張データを用いた提案モデルが既存の自動評価指標と同等以上の評価性能を持つことを示した．

2 ランキングモデル

提案モデルは，平易性の順序関係を判別するランキングモデルであり，リーダビリティ指標NPRM[13] に着想を得たモデルである．

学習 n を総文対数，x_{i1} を原文，x_{i2} を平易文(参照文) としたとき，原文と平易文の対集合 $X = \{(x_{11}, x_{12}), \ldots, (x_{n1}, x_{n2})\}$ から入力 I_{ijk} とラベル y_{ijk} を以下のように作成する．

$$I_{ijk} = concat(x_{ij}; \mathrm{SEP}; x_{ik}) \tag{1}$$

$$y_{ijk} = \begin{cases} [0,0,1] & \text{if } j < k \\ [1,0,0] & \text{if } j > k \\ [0,1,0] & \text{if } j = k \end{cases} \tag{2}$$

ただし，$concat(a; \mathrm{SEP}; b)$ は文 a に右から文 b を SEP トークンで結合する操作であり，$(j,k) \in \Lambda$, $\Lambda = \{(1,2),(2,1),(1,1),(2,2)\}$ とする．作成した入力 I_{ijk} とラベル y_{ijk} を用いて，以下のように損失関数 L で評価モデルを学習する．

$$S_{ijk} = softmax(\mathrm{FFNN}(\mathrm{BERT}(I_{ijk}))) \tag{3}$$

$$L = -\sum_{i=1}^{n} \sum_{(j,k)\in\Lambda} y_{ijk} \log(S_{ijk}) \tag{4}$$

ただし，BERT(·) は BERT の出力における CLS トークンを表し，FFNN は 1 層の全結合層とする．

評価時 評価スコア $score$ を推論するには，原文 $orig$ と平易文 $simp$ から入力 I_{eval} を以下のように作成し，$score$ を得る．

$$I_{eval} = concat(orig; \mathrm{SEP}; simp) \tag{5}$$

$$score = softmax(\mathrm{FFNN}(\mathrm{BERT}(I_{eval})))_{[3]} \tag{6}$$

ただし，[3] はベクトルの 3 次元目の要素を表す．

3 編集操作を用いたデータ拡張

2 節のランキングモデルの訓練データを拡張する手法を本節で提案する．具体的には，原文と平易文の対から成るパラレルデータにおいて，原文と平易文の中間の平易性を持つ文（本研究では**中間文**と呼ぶ）を，編集操作を用いて作成し，中間文を含む文対を訓練データとして用いる．編集操作で中間文が作成できる根拠として，平易化のための編集操作を適用した文がそうでない文よりも平易性が高いと人間によって評価される [9] ことが挙げられる．つまり，原文から参照文となる平易文に変換するために必要な編集操作を，原文に部分的に適応した文を考えると，これは原文と参照文の中間の平易性を持つ文（中間文）になる．中間文を学習に用いることで，より細かい粒度で平易性のランキングモデルが学習できることを期待する．

中間文作成の手順 本研究では編集操作を，**トークン単位編集操作（TE）**と**スパン単位編集操作（SE）**に分類する．中間文は，原文と参照文から自動的に抽出した TE から SE を構成し，SE を原文に適応することで作成する．

TE は，原文を参照文に変換するために，原文におけるそれぞれのトークンに対して適用する変換操作である．本研究では Dong ら [14] に倣って TE の種類を ADD（トークンの挿入），DEL（トークンの削除），KEEP（トークンの保持）とし，同研究の方法[1)]を用いて自動的かつ一意に抽出する．SE は KEEP 以外の連続する 1 つ以上の，平易化操作として意味を持つような TE のことであり，以下の 3 種類が存在する．

- ADD-DEL スパン：連続する 1 つ以上の ADD と DEL がこの順序で結合している部分．語彙平易化や文分割に該当する．
- DEL スパン：ADD-DEL スパン以外の連続する 1 つ以上の DEL 部分．不必要な情報削除に該当する．
- ADD スパン：ADD-DEL スパン以外の連続する 1 つ以上の ADD 部分．必要な情報追加に該当する．

表 1 に TE と SE の抽出例を示す．抽出できた SE が N 個だった場合，何も適用しない場合と全て適用する場合を除いた $2^N - 2$ 通りの組み合わせで原文に SE を適用し，$2^N - 2$ 件の中間文の集合を作成する．そこから m 件ランダムサンプリングしたものをデータ拡張に用いる．中間文を用いて提案モデルを学習する際は，中間文と原文，参照文それぞれをペアとした入力を拡張データとして学習に使用する．

4 平易性のための評価データ

テキスト平易化の自動評価指標の評価データ Simplicity-DA [1] では，システムが出力した文に対して人手評価が付与されている．しかし，Simplicity-DA のようにシステムが出力した文を使用する評価データは，流暢性，同義性，平易性間の Inter-class

1) https://github.com/YueDongCS/EditNTS/blob/master/label_edits.py

表 1 トークン単位編集操作とスパン単位編集操作を抽出した例．それぞれのハイライトの色は，ADD-DEL スパン，DEL スパン，ADD スパンを表す．この例ではスパン単位編集操作が合計 4 個抽出されている．

原文	According to Ledford , Northrop executives said they would build substantial parts of the bomber in Palmdale , creating about 1,500 jobs .
トークン単位編集操作	KEEP KEEP KEEP KEEP KEEP DEL KEEP KEEP KEEP KEEP ADD(most) DEL DEL KEEP KEEP KEEP ADD(parts) KEEP KEEP ADD(.) ADD(It) ADD(would) ADD(create) DEL DEL DEL KEEP KEEP KEEP
スパン単位編集操作	KEEP KEEP KEEP KEEP KEEP DEL KEEP KEEP KEEP KEEP ADD(most) DEL DEL KEEP KEEP KEEP ADD(parts) KEEP KEEP ADD(.) ADD(It) ADD(would) ADD(create) DEL DEL DEL KEEP KEEP KEEP
参照文	According to Ledford , Northrop said they would build most of the bomber parts in Palmdale . It would create 1,500 jobs .

correlation（IC）が高いので，平易性の評価データとして使用することは不適切であることが示唆されている [10]．システムが出力した文はそもそも流暢性や同義性が低い文を含んでおり，そのような質の低い文に対して人間が平易性の評価を行う際には，流暢性や同義性に強く依存した評価値が付与されてしまうことを意味する．したがって，流暢性や同義性に問題があるデータを含む Simplicity-DA をすべてそのまま平易性の評価に用いるのは好ましくない．そこで本研究では，以下の分析に基づき，Simplicity-DA から流暢性，同義性と平易性の依存性 (IC) が低い部分集合を抽出する．まず，以下のように Simplicity-DA の部分集合を作成する．

- all：Simplicity-DA 全体
- f_high：流暢性が全体の中央値よりも高い部分集合
- m_high：同義性が全体の中央値よりも高い部分集合
- fm_high：流暢性と同義性がそれぞれの全体の中央値よりも同時に高い部分集合

それぞれの部分集合において，平易性に対する流暢性と同義性の IC をピアソンの相関係数 ρ で計算する．また，人手評価値そのものの信頼性を分析するために，ICC$(1,k)$ [15] を用いて Inter-annotator agreement（IAA）を計算する．ICC$(1,k)$ は k 人の評価値を集計した値に対する信頼性を表す指標であり，0.5 未満なら poor，0.5〜0.75 なら fair，0.75〜0.90 なら good，0.90 以上なら exellent と解釈される [16]．結果を表 2 に示す．まず ρ に関しては，all が最も高い傾向を示すことが分かる．この結果は，Simplicity-DA の平易性に関する人手評価は IC が高く，平易性の評価データとしてそのまま用いるのには適さないことを示唆する．また，fm_high が最も低い傾向を示すことから，Simplicity-DA においては

表 2 Simplicity-DA における ρ と ICC$(1,k)$.

部分集合	$\rho\downarrow$		ICC$(1,k)\uparrow$
	流暢性	同義性	
all	.7705	.7575	.9042
f_high	.3660	.5086	.8961
m_high	.4755	.3829	.8969
fm_high	.2884	.2895	.8992

流暢性と同義性の両方の質の高いデータが平易性の評価データとして信頼性が高いことを確認できる．一方，ICC$(1,k)$ の結果から，評価値そのものの信頼性に関しては all の信頼性が最も高い（exellent）が，fm_high に関しても高い（good）信頼性を持つことが分かる．以上より Simplicity-DA において，IC と IAA の観点から fm_high が平易性の人手評価として適切であると判断した．

5 実験

SIERA に関する実験を行い，人手評価との相関を調べた．

5.1 実験設定

訓練データ 訓練データとして Newsela [17] を用いて 2 節の提案モデルをファインチューニングする．Newsela は原文を人手により平易文に書き直したデータセットである．Newsela の特徴として最終的な平易文に至る途中段階の文も 4 段階にわたり付与されていることが挙げられる．本研究では，難易度の差が 4 である文対のみをベースラインとして用いるデータ（Base）とした．この Base に，3 節の手法を適用し中間文を拡張したデータ（Silver）を作成した．また Base から，Newsela の人手による途中段階の平易文を中間文とみなして文対を追加し，データ拡張をおこなった (Gold)．これらの訓練データを

表 3 既存の評価指標（上）と提案手法（下）の人手評価との相関．SIERA に関しては 10 回実験を行った平均（左）と標準偏差（右）を示す．

	Simplicity-DA (fm_high)	Human-Likert
BERTScore_p	.2629	.4176
BERTScore_r	.0484	.3742
BERTScore_f1	.1355	.3938
SARI	.2270	.3907
BLEU	.1655	.3499
FKGL	-.1542	-.3531
SIERA_Base	.2204 (.0300)	.5441 (.0527)
SIERA_Silver	.2548 (.0339)	.5571 (.0326)
SIERA_Gold	.2986 (.0306)	.5644 (.0347)

用いて提案モデルを学習する際は，FFNN のパラメータと BERT の最終層のパラメータを学習した．

評価データ 4 節の結果から，評価データとして Simplicity-DA の fm_high 分割を用いた．さらに，人手で書き直された質の高い参照文で構成された評価データであり，IC が低いことが確認されている Human-Likert [10] も用いた．また提案手法との比較を行うために，参照文が必要な指標として BERTScore（精度，再現率，F1），SARI，BLEU，参照文が必要ない指標として FKGL を用いて人手評価との相関を調べた．人手評価との相関は，評価データに付与されている平易性の人手評価スコアと自動評価指標が算出したスコアとのピアソンの相関係数を用いて計算する．

5.2 実験結果

表 3 に実験結果を示す．既存の評価指標の中では，Simplicity-DA と Human-Likert のどちらにおいても BERTScore_p が最も人手評価との相関が高い．BERTScore_p と提案手法を比較すると，Simplicity-DA に関しては Base では SIERA が劣るが Silver や Gold で同等以上の評価性能を示している．また，Human-Likert では Base，Silver，Gold のいずれにおいても SIERA が高い性能を持つ．この結果から，参照文なしで評価スコアを計算できる SIERA は，参照文が計算時に必要な既存評価指標と比較しても同等以上の性能であるといえる．

また，表 3 下から，中間文を活用している Silver と Gold が中間文を使用しない Base よりも高い相関を示す傾向があることが分かる．これは，提案したランキングモデルは中間文を用いると，より細かい粒度の平易性の違いを考慮することが可能になり，平易性に関する評価性能が上がることを示唆する．

表 4 同義性が低い中間文の 1 例．

原文	Wars of the future will be fought with American women on the front lines and the public has no problem with it .
参照文	American women will soon be able to fight in wars .
中間文	Wars of the future will be able to fight in wars .

一方で，Silver で学習したモデルは Gold で学習したモデルよりも Base からの上がり幅が小さい傾向にある．Gold の中間文は人手で作成されたものなので流暢性や同義性に関しての質が担保されている一方で，Silver の中間文は自動的に抽出した SE を原文に適用しただけの文をランダムサンプリングしているので，質が低い文も含まれる．実際に作成できた同義性が低い Silver の中間文の例を表 4 に示す．この例では，原文の “fought with American women on the front lines and the public has no problem with it” が参照文の “able to fight in wars” に置き換わって作成された中間文を示している．この部分の置き換えだけでは文全体として同義性を損ねたものとなってしまうので，同義性を保つためには原文の主語である “Wars of the future” を参照文の “American women” に同時に置換する必要がある．このような Gold と Silver の中間文の質に関する違いが，Silver で学習した SIERA の評価性能の上がり幅が低くなる原因となっている可能性がある．考えられる改善方法としては，Silver で拡張した $2^N - 2$ 件の中間文の中から流暢性や同義性の高いものだけをサンプリングして学習に使用することが挙げられるので，これを今後の課題とする．

6 おわりに

本研究では編集操作を用いたデータ拡張手法で拡張した中間文を用いてランキングモデルを学習することで，参照文なしで計算できるテキスト平易化のための新しい自動評価指標 SIERA を提案した．また，平易性を評価するのに適切な評価データの抽出方法を IC と IAA に観点から行った．評価実験を通じて，中間文を学習に用いることによってランキングモデルの評価性能が向上することを示し，SIERA が既存の評価指標と同等以上の人手評価との相関を示すことを確認した．今後の課題としては，流暢性や同義性の高い中間文を自動的に抽出する方法を模索することが挙げられる．

参考文献

[1] Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. The (un)suitability of automatic evaluation metrics for text simplification. **Computational Linguistics**, Vol. 47, No. 4, pp. 861–889, December 2021.

[2] Sian Gooding. On the ethical considerations of text simplification. In **Ninth Workshop on Speech and Language Processing for Assistive Technologies (SLPAT-2022)**, pp. 50–57, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[3] Sergiu Nisioi, Sanja Štajner, Simone Paolo Ponzetto, and Liviu P. Dinu. Exploring neural text simplification models. In **Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)**, pp. 85–91, Vancouver, Canada, July 2017. Association for Computational Linguistics.

[4] 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 **Proceedings of the 1st Workshop on Automatic Text Adaptation (ATA)**, pp. 29–38, Tilburg, the Netherlands, November 2018. Association for Computational Linguistics.

[5] J. Peter Kincaid, Robert P. Fishburne, R 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. 1975.

[6] 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**, Vol. 4, pp. 401–415, 2016.

[7] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with BERT. In **8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020**. OpenReview.net, 2020.

[8] Teerapaun Tanprasert and David Kauchak. Flesch-kincaid is not a text simplification evaluation metric. In **Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)**, pp. 1–14, Online, August 2021. Association for Computational Linguistics.

[9] 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. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 4668–4679, Online, July 2020. Association for Computational Linguistics.

[10] Thomas Scialom, Louis Martin, Jacopo Staiano, Éric Villemonte de la Clergerie, and Benoît Sagot. Rethinking automatic evaluation in sentence simplification. **CoRR**, Vol. abs/2104.07560, , 2021.

[11] Hanlu Wu, Tengfei Ma, Lingfei Wu, Tariro Manyumwa, and Shouling Ji. Unsupervised reference-free summary quality evaluation via contrastive learning. In **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 3612–3621, Online, November 2020. Association for Computational Linguistics.

[12] Koki Maeda, Masahiro Kaneko, and Naoaki Okazaki. IMPARA: Impact-based metric for GEC using parallel data. In **Proceedings of the 29th International Conference on Computational Linguistics**, pp. 3578–3588, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics.

[13] Justin Lee and Sowmya Vajjala. A neural pairwise ranking model for readability assessment. In **Findings of the Association for Computational Linguistics: ACL 2022**, pp. 3802–3813, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[14] Yue Dong, Zichao Li, Mehdi Rezagholizadeh, and Jackie Chi Kit Cheung. EditNTS: An neural programmer-interpreter model for sentence simplification through explicit editing. In **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics**, pp. 3393–3402, Florence, Italy, July 2019. Association for Computational Linguistics.

[15] Patrick E. Shrout and Joseph L. Fleiss. Intraclass correlations: uses in assessing rater reliability. **Psychological bulletin**, Vol. 86 2, pp. 420–8, 1979.

[16] Terry K. Koo and Mae Y. Li. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. **Journal of chiropractic medicine**, Vol. 15, No. 2, pp. 155–163, 2016.

[17] Wei Xu, Chris Callison-Burch, and Courtney Napoles. Problems in current text simplification research: New data can help. **Transactions of the Association for Computational Linguistics**, Vol. 3, pp. 283–297, 2015.