単語難易度を考慮した反復的な翻訳文の平易化

大鹿雅史[1]　森下睦[2]　平尾努[2]　笹野遼平[1]　武田浩一[1]
[1] 名古屋大学大学院情報学研究科　[2]NTT コミュニケーション科学基礎研究所
oshika.masashi.f6@s.mail.nagoya-u.ac.jp
{makoto.morishita,tsutomu.hirao}@ntt.com
{sasano,takedasu}@i.nagoya-u.ac.jp

概要

近年，機械翻訳の精度は大幅に向上しており，機械翻訳の利用が広まっている．しかし，利用者の年齢によっては，機械翻訳によって出力された文章の語彙の難易度が高く出力文の意味を適切に理解するのが困難な場合がある．そこで，本研究では単語の難しさを語彙の獲得年齢 (Age of Acquisition) [1] と定義し，修正すべき単語を指定した翻訳文の平易化手法を提案する．実験を通して，提案手法が文章の意味を保持したまま平易化を行うことができることを示す．また，本手法を反復的に利用することでより平易化が可能であることを示す．

1　はじめに

近年，ニューラル機械翻訳技術の発展により多くの人々が機械翻訳を利用できるようになってきた．しかし，現在用いられている機械翻訳には出力される翻訳文の難易度を制御する仕組みが存在していない．そのため，利用者の年齢層に合わせた翻訳を生成することができず，特に子供が利用した際，翻訳文の意味を適切に理解できない可能性がある．

翻訳文の意味を適切に理解できない要因の一つとして，難易度の高い単語が使用されていることがある．単語の難しさを示す指標として語彙の獲得年齢 (Age of Acquisition, 以下 AoA) がある．AoA に従って単語をより平易なものに置換することで，翻訳文を年齢に適した難易度に平易化することが可能である．しかし，単純に単語を置き換えるだけでは，翻訳文の意味を損なう恐れがある．また，単語以外の修正ができず，文全体の平易化に繋がらない可能性がある．

そこで，本研究では原言語文と機械翻訳の出力に対し大規模言語モデル (LLM) を用いて翻訳文中の AoA の高い語を反復的に指定する平易化を行う手法を提案する．図 1 に本研究の概要図を示す．LLM を利用することにより，指定した単語だけでなく，周囲の単語も文脈に応じて置き換えることができ，文意を保持した平易化が可能となる．また，翻訳文内に修正すべき単語が複数存在した場合や出力された平易文に再び AoA の高い単語が含まれていた場合，反復的に適用することで全ての単語を平易化できる．また，本手法は AoA を用いて単語の難易度を決定するため，子供のための翻訳器や非母語話者の言語学習支援など教育目的への応用が可能であると考えられる．

Simple-English-Wikipedia をもとに作成したデータセットのでの実験を通じて，MUSS[2] を利用した平易化手法や語彙を制限した生成手法に比べ高い翻訳性能を維持したまま平易化を達成した．

2　関連研究

2.1　平易化と後編集

テキスト平易化や後編集の手法は翻訳文をより理解しやすい文へと変換するために広く用いられている．テキスト平易化には深層強化学習を利用した平易化 [3] や統計的機械翻訳を平易化に応用する手法 [4] などが存在する．また，後編集には翻訳文を反復的に修正することで翻訳の品質を高める手法 [5] や，削除や置き換えなどのタグを利用して後編集を行う手法 [6] などが存在している．しかし，翻訳文に対して後編集を利用した平易化は取り組まれておらず，本研究では翻訳文の品質を維持したままの後編集を利用した平易化を目的とする．

2.2　難易度を考慮した機械翻訳

近年，文章の難易度や複雑さを制御する手法が提案されている．Agrawal ら [7] や Tani ら [8] は機械翻訳モデルから出力される文章の複雑さを制御する多

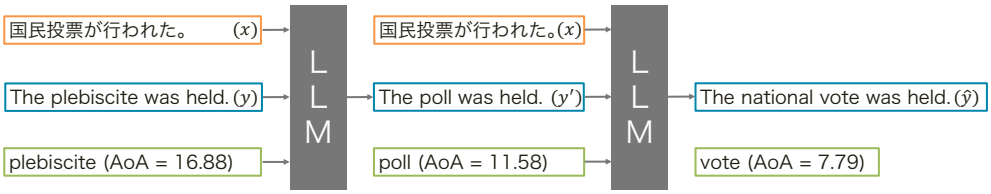

図 1 提案手法の概要図．原言語文 (x)，機械翻訳文 (y)，修正単語の三つ組を LLM に入力し平易文 ($\hat{y}$) を出力する．修正単語は機械翻訳文のうち最も AoA の高い単語と定義し，その AoA を示す．

段階難易度制御翻訳に取り組んでいる．また，谷ら [9] は異なる難易度の参照文を利用した学習を行うことで出力文の難易度を制御した機械翻訳の手法を提案している．これらの手法では文全体の難易度を考慮しているが，単語単位での難易度を考慮しておらず，個々の単語の難易度を制御することができない．

3 反復的な翻訳文の平易化

3.1 提案手法

本研究では，修正すべき単語を指定した上で翻訳文を平易化する手法を提案する．機械翻訳文 (y)，原言語文 (x)，修正単語の三つ組を LLM に入力することで平易文 ($\hat{y}$) を出力するモデルの構築を行う (図 1)．修正単語の同定には AoA を利用し，機械翻訳文の修正すべき単語を < 編集 > タグで囲みモデルへ入力する．本手法は語彙の獲得年齢をもとに修正単語の定義を行うため，機械翻訳の利用者に合わせ対象の年齢を指定することが可能である．すなわち，n 歳向けの機械翻訳を実現する際には AoA が n 以上である単語を修正単語と定義することで AoA が n 以上の単語を出力しない翻訳器を構築する．

本手法は反復的に利用可能なことから，機械翻訳文に複数の修正すべき単語が含まれていた場合，繰り返し単語を指定することで，全ての単語を編集することができるという特徴がある．加えて，一回の平易化で十分に AoA が下がらなかった場合 y' にも，再び修正単語を指定し平易化をすることで，AoA を満たすまで修正することが可能である．

3.2 データセットの構築

提案手法の評価に適したデータセットが存在しないため，本研究では単言語データを元に逆翻訳を適用することでデータセットの作成を行う．図 2 に作成の手順を示す．単言語データに対して逆翻訳を適用することで，原文，中間翻訳文，逆翻訳文の三つ組を得る．この三つ組のうち，原文と逆翻訳文の各

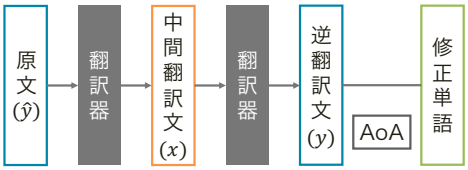

図 2 データセットの作成手順を示した図

文に含まれる語の最大 AoA の差が 0.5 以上である文ペアを抽出する．修正単語は逆翻訳文のうち，最も AoA の高い単語とし，原文，中間翻訳文，逆翻訳文，修正単語の四つ組のデータセットを構築する．

また，n 歳向けの機械翻訳を実現するために生成時には対象となる年齢 n を指定してデータセットの選別を行う．すなわち，逆翻訳文に含まれる最大 AoA が n 以上，かつ，原文に含まれる最大 AoA が n 未満である文のみを利用して生成を行う．

データセットのうち，実際の生成時には中間翻訳文が図 1 内 x，逆翻訳文が図 1 内 y，原文が図 1 内 $\hat{y}$ に相当する．

4 実験

4.1 実験設定

大規模言語モデルとして Hugging Face が公開しているライブラリである Transformers から事前学習済みの GPT-NeoX[1] を利用した．また，モデルのファインチューニングには LoRA[10] を用いる．[2]

本研究では 3.2 節で述べた手法を用いてデータセットの作成を行った．単言語データには Simple-English-Wikipedia[3] を利用した．データセットには見出しなどの非文が含まれていたため，これらを取り除く前処理を行い実験に利用した．逆翻訳には JParaCrawl v3.0 を用いて学習されたニューラル機械翻訳モデル [11] を利用して，英→日→英と機械翻訳を適用することでデータセットを作成した．

1) https://huggingface.co/rinna/bilingual-gpt-neox-4b-instruction-sft
2) 実験設定の詳細は AppendixA に示す．
3) https://huggingface.co/datasets/wikipedia/viewer/20220301.simple

表 1 各手法の生成文を評価した際の実験結果．「生成した文数」は提案手法を反復した際に AoA が 10 未満にならず再度生成した際の文数を示す．なお，評価の際は 6,194 文全体を用いて評価を行っている．

					提案手法				
	機械翻訳文	MUSS	制約付き生成	複数単語指定	反復 1	反復 2	反復 3	反復 4	反復 5
生成した文数	6,194	6,194	6,194	6,194	6,194	1,080	542	349	247
BLEU ↑	38.0	26.7	40.1	43.4	43.4	43.3	**43.5**	**43.5**	**43.5**
COMET ↑	0.870	0.831	0.866	**0.876**	**0.876**	0.875	0.875	0.875	0.875
SARI ↑	52.0	43.7	56.5	60.1	60.2	60.3	**60.4**	**60.4**	**60.4**
FKGL ↓	9.15	**6.42**	8.69	8.60	8.58	8.54	8.52	8.52	8.51
平均 AoA	11.51	8.85	8.48	8.50	8.51	8.24	8.14	8.09	**8.06**
AoA が 10 未満になった割合↑	0.00	0.69	0.90	0.83	0.82	0.91	0.94	0.96	**0.97**

作成したデータセットは 169,672 文からなり，訓練データ：検証データ：テストデータ = 8：1：1 に分割して実験を行った．ファインチューニング時には一つのモデルで全ての年齢を対象として平易化が可能になるように訓練データのすべてのデータを利用した．また，本研究では生成時の対象の年齢を 10 歳として生成を行なった．すなわち，テストデータでは逆翻訳文に含まれる最大 AoA が 10 以上かつ原文に含まれる最大 AoA が 10 未満であるデータのみを利用して生成を行った．生成したデータは全 6,194 文であった．

4.2 評価指標

評価指標には機械翻訳の評価指標，テキスト平易化の評価指標，平均最大 AoA，平易化の成功割合を用いて評価を行う．機械翻訳としての評価には *n*-gram の一致度を元に評価を行う BLEU [12] と類似度ベースに評価を行う COMET [13] を利用する．[4] テキスト平易化の評価指標として単語の言い換えなどを元に評価を行う SARI [14]，一文あたりの単語数や単語あたりの音節数から評価を行う FKGL [15] を利用する．平易化の成功割合についてはテストセットのうち，文章内の最も高い AoA が対象の年齢より低くなった文数の割合を示す．

4.3 比較手法

提案手法の有用性を検証するために三つの比較手法で実験を行った．一つ目はテキスト平易化の手法として事前学習済みの教師なしモデルである MUSS[2] を利用して平易化を行う．テストセットの逆翻訳文を入力とし平易文を出力する手法である．二つ目はデコード時に語彙を制限する翻訳手法である (制約付き生成)．この手法ではデータセットの翻訳文を日英翻訳で英語に翻訳する際に，AoA が 10 以上の単語が出力されないように語彙を制限して翻訳を行う．具体的には，生成時の各時刻で仮説に AoA が 10 以上の単語が含まれている場合，その仮説のスコアを $-\infty$ とする．生成時のビーム幅は 6 で生成を行い，探索に失敗した文については逆翻訳文を出力する設定とする．三つ目は一回の平易化で複数の単語を指定する手法である．この手法は提案手法の派生型として考えることができる．ファインチューニング時には原文の最大 AoA よりも高い AoA を持つ逆翻訳文に含まれる単語を修正単語として指定し，生成時には AoA が 10 以上の単語をすべて修正単語と指定して平易化を行う．この手法により反復的に平易化を行う利点を示す．

4) COMET のモデルには，Unbabel/wmt22-comet-da を使用した．

4.4 実験結果

表 1 に実験結果を示す．MUSS を用いた平易化では逆翻訳文に比べ BLEU や COMET の機械翻訳の評価が著しく低下しているのが確認できる．これによって単純に平易化を行うだけでは翻訳文の意味が損なわれてることがわかる．次に，提案手法の 1 回目の生成では逆翻訳文に比べ全ての評価指標において性能が向上していることがわかる．これにより提案手法では翻訳文の文意を維持したまま平易化が可能であり，提案手法の有効性が確認できる．また，提案手法の 1 回目と制約付き生成を比較すると平均 AoA 及び AoA が 10 未満になった割合は劣るものの，その他の指標で高い値を示していることが確認できる．最後に複数単語指定との比較では翻訳性能は同程度であり，平易化の指標についてはわずかに提案手法が上回っている．一方で，平均 AoA や AoA が 10 未満になった割合については複数単語指定が上回っており，複数の単語を指定することで一回の平易化でより多くの単語を修正することが可能であることがわかる．

表 2 各モデルにおける実際の生成例．一番右の列は文中で最も AoA の高い単語とその AoA を表す．

日本語文	紙幣発祥の地.	
機械翻訳文	The birthplace of banknotes.	banknotes (12.18)
MUSS	The country's banknotes.	banknotes (12.18)
制約付き生成	It is the birthplace of paper money.	birthplace (6.9)
提案手法反復 1	The origin of paper money.	origin (10.25)
提案手法反復 2	The first place of production for paper money.	production (9.21)
原文	The Birthplace of Paper Money.	Birthplace (6.9)

次に，AoA が基準値を超えていた文に対して反復的に提案法を適用した際の比較を行う．提案手法内の比較では反復を行うことで全ての指標が同程度に維持または改善されていることがわかる．これより，反復的に平易化を行うことで翻訳文の意味を維持したままより AoA の低い単語を利用した翻訳文へと平易化することが可能になっていると言える．また，制約付き生成や複数単語指定との比較では 2 回目の反復で AoA が 10 未満になった割合を上回っており，その際，機械翻訳の手法ではわずかに劣るものの平易化の指標において提案手法が高い性能を示している．なお，5 回目の反復時には COMET 以外の全ての指標において比較手法のスコアを上回る結果となった．

最後に，図 3 に横軸に各文における最大 AoA，縦軸に文数を示したグラフを示す．図 3 より提案手法の一回目の生成では AoA が 10 以上の単語が多く使われているが，反復を行うことで低 AoA の単語に平易化されていることが確認できる．制約付き生成は低 AoA の単語を利用しているものの，生成に失敗する例があることによって AoA が 10 以上の文が生成されており，MUSS を用いた平易化の手法についても AoA の高い単語が多く使われている．

4.5 分析

表 1 において MUSS を用いた平易化は FKGL について大幅なスコアの改善が確認できる．MUSS では一文を複数の文に分割する平易化が多く行われていた．FKGL は一文あたりの単語数を考慮する評価指標であり，文の分割によって一文あたりの単語数が減少するため，スコアが改善したと考えられる．

提案手法において平易化に失敗している例の中には反復を行うことで一度修正された単語が再び生成されてしまう例が存在した．また，反復を繰り返すことで平均 AoA が下がっているのに対して BLEU の値は収束している．これは反復によって平易化に

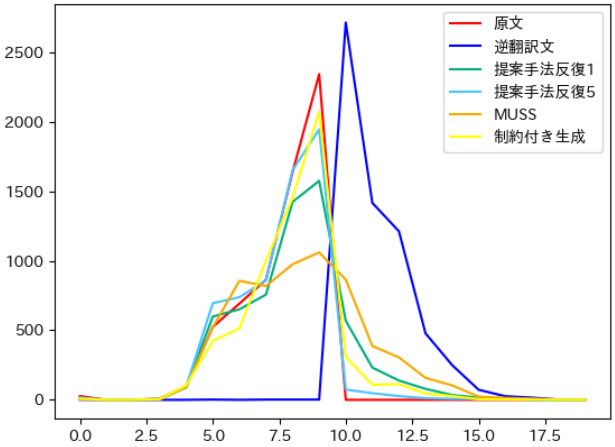

図 3 各手法の生成文の最大 AoA の統計を表したグラフ

は成功しているものの，テストセット全体の文数に対して平易化した文数の割合が少ないため，全体で評価した際に値に変化が見られないと考えられる．

4.6 生成例

表 2 に実際の生成例を示す．MUSS による平易化では最大 AoA が下がっていないのに加え，文意が損なわれていることが確認できる．一方で，提案手法の 1 回目の反復では AoA が下がりきってはいないが，文意を保持した平易化に成功しており，また，2 回目の反復で AoA が 10 未満の単語への平易化が成功していることが確認できる．制約付き生成では原文に最も近い生成を行なっており，生成時に語彙を制限する手法も一定の有効性を示している．

5 おわりに

本研究では修正すべき単語を指定し，反復的な翻訳文の平易化を行なった．実験の結果より提案手法により AoA の高い単語を指定しながら平易化を行うことで翻訳性能を維持したまま平易化を行うことが可能であることを示した．本研究では日英翻訳についてのみ実験を行ったが，英日翻訳や他の言語対に関しても実験を行いたい．

参考文献

[1] Victor Kuperman, Hans Stadthagen-González, and Marc Brysbaert. Age-of-acquisition ratings for 30,000 english words. **Behavior Research Methods**, Vol. 44, pp. 978–990, 2012.

[2] Louis Martin, Angela Fan, Éric de la Clergerie, Antoine Bordes, and Benoît Sagot. MUSS: Multilingual unsupervised sentence simplification by mining paraphrases. In **Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC)**, pp. 1651–1664, 2022.

[3] Xingxing Zhang and Mirella Lapata. Sentence simplification with deep reinforcement learning. In Martha Palmer, Rebecca Hwa, and Sebastian Riedel, editors, **Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 584–594, 2017.

[4] 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 (TACL)**, Vol. 4, pp. 401–415, 2016.

[5] Pinzhen Chen, Zhicheng Guo, Barry Haddow, and Kenneth Heafield. Iterative translation refinement with large language models, 2023.

[6] Prabhakar Gupta, Anil Nelakanti, Grant M. Berry, and Abhishek Sharma. Interactive post-editing for verbosity controlled translation. In **Proceedings of the 29th International Conference on Computational Linguistic (COLING)**, pp. 5119–5128, 2022.

[7] Sweta Agrawal and Marine Carpuat. Controlling text complexity in neural machine translation. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)**, pp. 1549–1564, 2019.

[8] Kazuki Tani, Ryoya Yuasa, Kazuki Takikawa, Akihiro Tamura, Tomoyuki Kajiwara, Takashi Ninomiya, and Tsuneo Kato. A benchmark dataset for multi-level complexity-controllable machine translation. In **Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC)**, pp. 6744–6752, 2022.

[9] 谷和樹, 湯浅亮也, 田村晃裕, 梶原智之, 二宮崇, 加藤恒夫. 異なる難易度の参照文を用いる多段階難易度制御翻訳. 言語処理学会第 29 回年次大会, 2023.

[10] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In **The Tenth International Conference on Learning Representations (ICLR)**, 2022.

[11] Makoto Morishita, Katsuki Chousa, Jun Suzuki, and Masaaki Nagata. JParaCrawl v3.0: A large-scale English-Japanese parallel corpus. In **Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC)**, pp. 6704–6710, 2022.

[12] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In **Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)**, pp. 311–318, 2002.

[13] Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. COMET: A neural framework for MT evaluation. In **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 2685–2702, 2020.

[14] 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 (TACL)**, Vol. 4, pp. 401–415, 2016.

[15] Peter Kincaid, Robert P. Fishburne, 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. 1975.

表 3　実験に利用したプロンプト

指示:< 編集 > で囲まれた単語を使わないで仮説文を元に次の日本語を翻訳してください. ### 日本語: 紙幣発祥の地. ### 仮説文:The birthplace of < 編集 > banknotes < 編集 >. ### 翻訳:

表 4　平易化に失敗した際の生成例

日本語文	彼は 1909 年から 1911 年までその機関の大統領を務めました.	
機械翻訳文	He served as president of the institution from 1909 to 1911.	institution (10.05)
提案手法反復 1	He served as the President of that institution from 1909 to 1911.	institution (10.05)
提案手法反復 2	He served as the President of that agency from 1909 to 1911.	agency (10.58)
提案手法反復 3	He was President of that institution from 1909 to 1911.	institution (10.05)
提案手法反復 4	He was president of the agency from 1909 to 1911.	agency (10.58)
提案手法反復 5	He was president of that institution from 1909 to 1911.	institution (10.05)
原文	He served as president of that body from 1909 to 1911.	served (7.17)

表 5　反復 1 回で平易化に成功した例

日本語文	また，1996 年から 2007 年まで 12 年連続でベスト・イレブンに選出された.	
機械翻訳文	He was also named Best Eleven for 12 consecutive years from 1996 to 2007.	consecutive (10.0)
MUSS	He was also named Best Eleven by the magazine 12 times, from 1996 to 2007.	times (6.7)
制約付き生成	He was also named Best Eleven for 12 consecutive years from 1996 to 2007.	consecutive (10.0)
提案手法反復 1	He was also selected Best Eleven for 12 years in a row from 1996 to 2007.	selected (8.0)
原文	She was also selected Best Eleven for 12 years in a row (1996-2007).	selected (8.0)

A　学習設定

LoRA は全ての線形層に対して適用し学習されたパラメータ数は 25,952,256 で全体の 0.68%のパラメータについてファインチューニングを行なった．また，ハイパーパラメータは r を 16，α を 32 と設定した．ファインチューニングを行う際のパラメータは初期の学習率を 1e-5 とし線形に減衰させ，バッチサイズを 16，エポック数を 5 とし，最適化手法に AdamW を利用した．また，実験に利用したプロンプトの例を表 3 に示す．日本語が中間翻訳文 (図 1 内 x)，仮説文が逆翻訳文 (図 1 内 y)，< 編集 > で区切られた単語が修正単語としてモデルに入力される．

B　平易化に失敗している例

表 4 に平易化に失敗した際の例を示す．機械翻訳文に存在する institution という単語は反復 2 回目で agency という AoA の高い単語へと修正されていることが確認できる．この agency を修正単語と指定して平易化を行うことで，3 回目の反復で再び institution に修正されているが，4 回目の反復で agency が再度生成されていることが確認できる．このように反復を行うことで一度平易化された単語が再び生成されてしまうような例が存在した．これは現在のモデルが変更履歴を考慮できないことが原因の一つとして考えられ，今後の課題である．

C　平易化に成功している例

表 5 に反復をすることなく平易化に成功した例を示す．MUSS を利用した平易化では平易化には成功しているものの，やや文意を損なった平易文が生成されていることが確認できる．また，制約付き生成においては AoA が 10 未満の単語のみでの生成ができず，平易化に失敗している．一方で，提案手法の生成では反復をすることなく一回の平易化で文意を保持したまま AoA の低い単語への修正に成功している．