テキスト平易化のための難易度に基づく検索拡張生成

宮田 莉奈 [1]
[1] 愛媛大学大学院理工学研究科
miyata@ai.cs.ehime-u.ac.jp

梶原 智之 [1,2]
[2] 大阪大学 D3 センター
kajiwara@cs.ehime-u.ac.jp

概要

本研究では、大規模言語モデルによるテキスト平易化において検索拡張生成の事例選択を改良し、難易度制御の性能向上に取り組む。近年、検索拡張生成の技術が広く利用されているが、従来の意味的類似度に基づく検索ではテキスト平易化において重要な難易度が考慮されていないため、事例選択に改良の余地がある。本研究では、類似度・難易度・その組み合わせの 3 種類の方法で、テキスト平易化の検索拡張生成における事例選択の有効性を検証する。英語の文難易度制御に関する評価実験の結果、難易度を考慮する検索拡張生成の有効性が確認できた。

1 はじめに

テキスト平易化 [1] とは、文の主要な意味を保持したまま難解な表現を平易に言い換えることによって、テキストの読みやすさを改善する技術である。この技術は、子供や言語障害を持つ人々など多様な読者の文章読解支援 [2–4] や、各人の言語能力に適した教材提供による効率的な言語学習支援 [5] に有用である。このような背景から、対象読者に応じて文難易度を制御するテキスト平易化の手法 [6–8] が盛んに研究されている。本タスクは、言い換えや要約などの関連タスクとは異なり、入力文との同義性と、目標難易度に適した平易性の両立が求められる点に特徴がある。

近年、他の自然言語処理タスクと同様に、テキスト平易化においても大規模言語モデルに基づく手法 [9–14] が注目を集めている。大規模言語モデルの活用においては、入力文に関連する事例をプロンプトに含める検索拡張生成 (Retrieval-Augmented Generation; RAG) [15] によって、ドメイン知識などを補完し、性能を改善できることが知られている。テキスト平易化における先行研究 [16] でも、入力文と意味的に類似する事例をプロンプトに含める検索拡張生成によって、性能改善が報告されている。

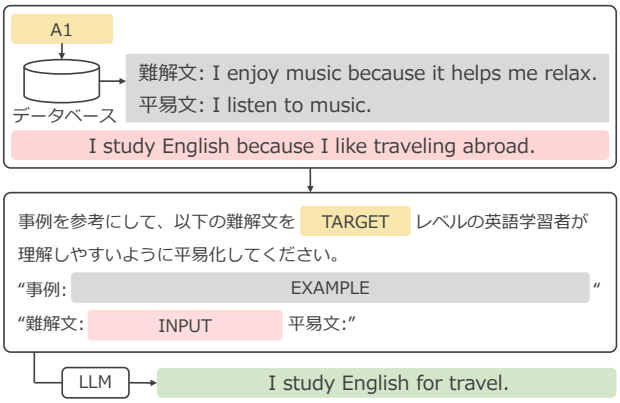

図 1 提案手法の概要図

しかし、意味的類似度のみに基づく一般的な検索拡張生成は、テキスト平易化タスクの特性を必ずしも充分に捉えていない。入力文と意味的に類似した事例であっても、その難易度が目標難易度から大きく乖離している場合がある。そのような事例をプロンプトに含めると、同義性の改善は期待できる一方で、平易性の向上には寄与しない可能性がある。特に、多様な目標難易度を扱いたい場合や、特定の対象読者に向けて難易度を制御したい場合には、意味的類似度のみに基づく既存の検索拡張生成では充分な平易化性能を得ることが難しい。

本研究では、テキスト平易化における検索拡張生成のために、入力文との意味的類似度ではなく、目標難易度の一致（図 1）に着目してプロンプトに含める事例を選ぶ。これにより、大規模言語モデルは目標難易度に適した表現の具体例を文脈内学習 [17] でき、難易度制御の性能改善が期待できる。さらに、類似度と難易度の両方を考慮する検索拡張生成を提案し、同義性と平易性の両立を目指す。

英文の難易度制御に関する実験の結果、難易度に基づく検索拡張生成は、類似度に基づく検索拡張生成と同等の同義性を維持しつつ、平易性を顕著に改善できた。さらに、類似度と難易度の両方を考慮する検索拡張生成は、全ての評価指標において一貫して最高性能を達成した。

表 1　データセットの統計

Newsela-Auto	訓練	検証	評価
to A1（平易）	7,174	1,153	1,196
to A2	43,239	6,081	6,063
to B1	42,850	4,696	4,748
to B2（難解）	2,809	272	379
合計	96,072	12,202	12,386

2　提案手法

本研究では、検索拡張生成をテキスト平易化に特化させるために、目標難易度の一致に着目して事例を選択する。具体的には、平易化したい入力文および目標難易度が与えられた際に、外部知識源から目標難易度が一致する事例を K 件選択する。それらの事例は、プロンプトの一部として大規模言語モデルに入力され、文脈内学習に使用される。検索の手法として、以下の 2 つのアプローチを提案する。

2.1　難易度に基づく検索

本手法では、入力文の意味内容は考慮せず、目標難易度が一致する事例のみを検索拡張生成に利用する。外部知識源として、本研究では難解文と平易文の文対からなるテキスト平易化パラレルコーパスを使用する。ここで、各平易文には事前に難易度ラベルを付与しておき、検索時には目標難易度と同一のラベルを持つ文対を候補とする。それらの候補の中から K 件の文対を無作為に選択し、プロンプトとして利用する。無作為選択を採用することで、意味的類似度の影響を排除し、難易度情報のみが平易化性能に与える影響を分析できる。本手法によって、大規模言語モデルは目標難易度に適した表現の具体例を参照しつつテキスト平易化に取り組める。

2.2　難易度と意味の両方に基づく検索

本手法では、前節と同様に目標難易度によって事例をフィルタリングし、さらに入力文との意味的類似度によって事例をリランキングする。つまり、目標難易度と一致する文対集合の中から、入力文との意味的類似度が高い上位 K 件を抽出し、プロンプトとして利用する。本手法によって、大規模言語モデルは目標難易度に適した表現とドメインやトピックの近い表現の両方の具体例を参照できるため、平易化性能の更なる改善が期待できる。

3　評価実験

英語の文難易度制御の実験によって、提案手法の有効性を評価した。

3.1　実験設定

データセット　本実験には英語のテキスト平易化パラレルコーパスである Newsela-Auto [18] を使用した。本コーパスは、学年レベルの異なる複数の読者に向けてニュースを人手で平易化して作られたパラレルコーパスであり、多様な難易度の文が含まれるため難易度制御の実験に適している。ただし、学年レベルはニュース記事に対して付与されており、記事中の各文には難易度が定義されていないことには注意が必要である。先行研究 [6–8] では、記事中の各文の難易度として記事の学年レベルを使用してきたが、この実験設定では文難易度に多くのノイズが含まれてしまう。そこで本研究では、文難易度推定器[1] [19] を使用し、Newsela-Auto に含まれる各文に CEFR 基準[2] の文難易度を自動的に付与した。ここで、C レベルの難解な文は Newsela-Auto にはほとんど見られなかったため、本実験では A1 から B2 までの 4 段階の CEFR レベルを用いることにした。

表 1 にデータセットの統計情報として、目標難易度ごとの文対数を示す。これは、Newsela-Auto の公式の訓練用／検証用／評価用の分割に基づくが、目標難易度が C レベルの文対を除外したものである。また、難解文が対応する平易文以上に平易な難易度ラベルを持つ場合も、それはノイズ文対であると考えて除外した。これらのうち、訓練用データを検索拡張生成のための外部知識源として使用し、評価用データを難易度制御の実験に使用した。

モデル　大規模言語モデルには、Llama-3.1（以降 Llama と表記）[20] の `8B-Instruct` モデル[3] および `70B-Instruct` モデル[4] と、Qwen-2.5（以降 Qwen と表記）[21] の `7B-Instruct` モデル[5] および `72B-Instruct` モデル[6] を使用した。ここで、大規模言語モデルによる効率的な推論のために vLLM[7] [22] を用いた。

1）https://github.com/yukiar/CEFR-SP
2）CEFR は言語運用能力の国際標準であり、A1（平易）, A2, B1, B2, C1, C2（難解）の 6 段階の難易度を持つ。
3）https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct
4）https://huggingface.co/meta-llama/Llama-3.1-70B-Instruct
5）https://huggingface.co/Qwen/Qwen2.5-7B-Instruct
6）https://huggingface.co/Qwen/Qwen2.5-72B-Instruct
7）https://github.com/vllm-project/vllm

表 2　Newsela-Auto における英文の難易度制御の実験結果。SARI は参照文との表層的な類似性を評価し、BERTScore は文脈化埋め込みに基づく参照文との類似性を評価し、Pearson は出力文の推定難易度と目標難易度の相関を評価する。

Model	Method	SARI	BERTScore	Pearson	Model	Method	SARI	BERTScore	Pearson
Llama 8B	Zero-shot	0.389	0.898	0.441	Llama 70B	Zero-shot	0.382	0.892	0.457
	Few-shot	0.407	0.904	0.410		Few-shot	0.404	0.904	0.457
	類似度 RAG	0.411	0.904	0.418		類似度 RAG	0.409	0.904	0.470
	難易度 RAG	0.413	0.905	0.586		難易度 RAG	0.408	0.904	0.612
	両方で RAG	**0.419**	**0.906**	**0.626**		両方で RAG	**0.416**	**0.906**	**0.646**
Qwen 7B	Zero-shot	0.364	0.903	0.486	Qwen 72B	Zero-shot	0.382	0.904	0.499
	Few-shot	0.377	0.904	0.460		Few-shot	0.388	0.905	0.489
	類似度 RAG	0.379	0.905	0.452		類似度 RAG	0.387	0.906	0.474
	難易度 RAG	0.384	0.905	0.563		難易度 RAG	0.394	0.906	0.564
	両方で RAG	**0.386**	**0.906**	**0.591**		両方で RAG	**0.398**	**0.907**	**0.607**

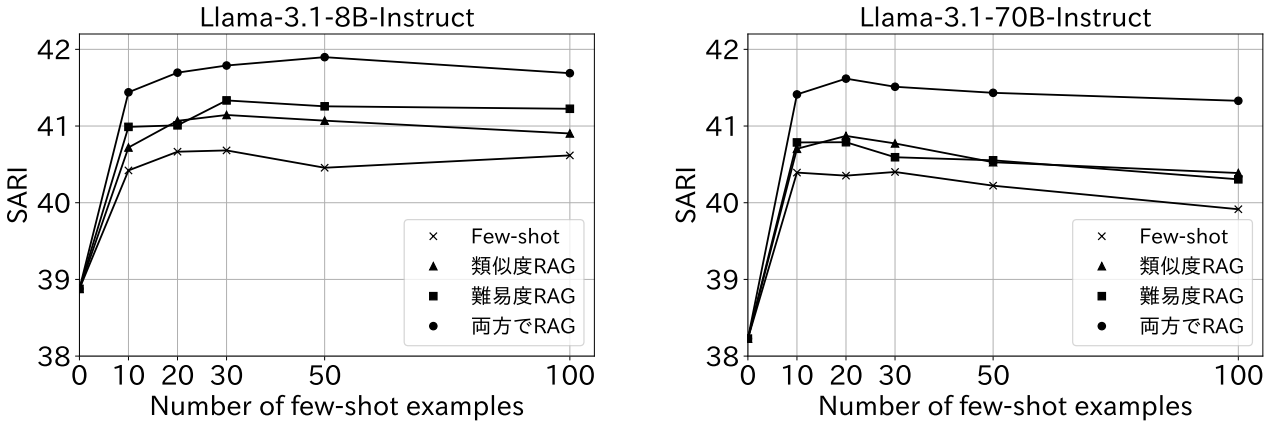

図 2　Few-shot 事例数に対する SARI の変化

検索拡張生成における意味的類似度の推定には、E5 [23] の base-v2 モデル[8]を用いた。ベクトル間の余弦類似度によって文間の意味的類似度を求めた。

大規模言語モデルに与えるプロンプトは、CEFR 難易度の制御に関する先行研究 [14] で使用された Barayan et al. のプロンプト P4 に従った。これは、入力文と目標難易度に加えて、特定の CEFR レベルの定義やテキスト平易化タスクの詳細についても指示を与えるものである。なお、Few-shot 設定においては、プロンプトに含める事例数を $\{10, 20, 30, 50, 100\}$ の範囲で変化させた際の最高性能を報告する。

評価　難易度制御の品質評価には SARI[9] [25] を採用した。参照文との表層マッチングに基づく評価指標である SARI に加えて、同義性を評価するために BERTScore[10] [26] を採用した。また、平易性については、出力文の推定難易度[11]と目標難易度の間の Pearson 相関を評価した。

比較手法　難易度に基づく検索拡張生成の有効性を検証するために、以下の 5 つの手法を比較した。

Zero-shot　事例を与えずに難易度制御を指示する。

Few-shot　各難易度の事例を均一にサンプリングしてプロンプトに含める。つまり、類似度と難易度の両方の影響を排除した事例を与える。

類似度 RAG　先行研究 [16] と同じく、入力文との類似度が高い順に事例を選択してプロンプトに含める。提案手法とは異なり、この比較手法は事例選択の際に目標難易度を考慮しない。

難易度 RAG　2.1 節で説明した提案手法。難易度のみを考慮して選択した事例を与える。

両方で RAG　2.2 節で説明した提案手法。難易度と類似度の両方を考慮して事例を選択する。

8）https://huggingface.co/intfloat/e5-base-v2
9）実装は EASSE [24] https://github.com/feralvam/easse
10）https://huggingface.co/spaces/evaluate-metric/bertscore
11）データセットの前処理と同様、CEFR 基準の文難易度推定器 [19] を使用した。https://github.com/yukiar/CEFR-SP

表 3 Llama-3.1-8B-Instruct による難易度ランキングの評価。C レベルの文を B2 から A1 まで 4 段階に平易化し、5 文が期待通りの難易度順になるかどうかを評価した。

Method	nDCG	ρ	τ
Zero-shot	0.947	0.639	0.581
Few-shot	0.957	0.703	0.639
類似度 RAG	0.957	0.692	0.627
難易度 RAG	0.970	0.826	0.769
両方で RAG	**0.973**	**0.842**	**0.785**

3.2 実験結果

表 2 に難易度制御の実験結果を示す。全てのモデルで一貫して、Zero-shot ベースラインよりも他の手法が SARI において高性能を示し、テキスト平易化における大規模言語モデルの文脈内学習の有効性を確認できた。次に、既存手法の類似度 RAG と提案手法の難易度 RAG を比較すると、SARI や BERTScore は Llama においては同等であり、Qwen においても BERTScore は同等の性能であった。一方で、目標難易度との一致に関する Pearson 相関の評価を見ると、全てのモデルで一貫して、難易度 RAG が類似度 RAG を大きく上回った。これらの結果から、提案手法の難易度 RAG は、既存の類似度 RAG と同等の同義性を維持しつつ、平易性を顕著に改善するという有効性を持つことが確認できた。さらに、類似度と難易度の両方を考慮する提案手法（両方で RAG）は、全てのモデルにおいて一貫して、全ての評価指標で最高性能を達成した。

文脈内学習に用いる事例数と性能の関係を図 2 に示す。Llama-8B の小規模モデルでは 50 件程度まで、Llama-70B の大規模モデルでは 20 件程度まで、事例数の増加とともに性能も向上し、それ以上に事例数を増やしても性能は改善しないことがわかった。同様の傾向が Qwen に対しても観察され、英文の難易度制御において、事例の有無や選択方法は性能に大きく影響するが、大規模モデルにおいては事例数は少なくても良いことが明らかになった。

3.3 分析

難易度の制御性　各手法の難易度の制御性能について、より詳細に分析した。Newsela-Auto の評価用データから C レベルの難解文を無作為に 500 件抽出し、それを B2 から A1 までの 4 レベルにそれぞれ平易化した。このように用意した 5 文の組に対し、文

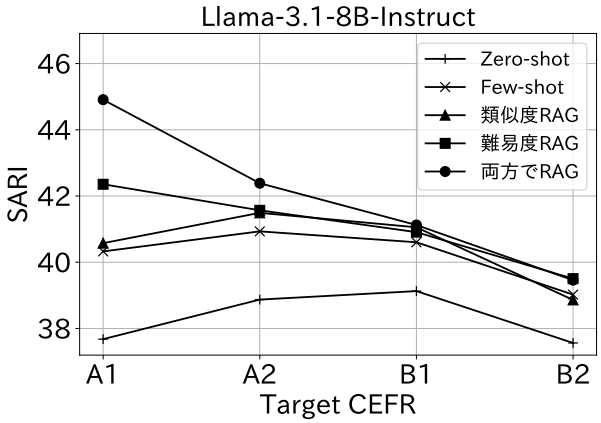

図 3　目標難易度ごとの SARI の変化

難易度推定器[11]で各文の CEFR レベルを評価して難易度順にランキングした。そして、ランキング性能を nDCG・スピアマンの順位相関係数 (ρ)・ケンドールの順位相関係数 (τ) の 3 つの方法で自動評価した。

Llama-8B モデルによる分析結果を表 3 に示す。nDCG による評価は表 2 の SARI とも一貫しており、Zero-shot よりも Few-shot や類似度 RAG の方が優れた性能を示し、難易度を考慮する提案手法はそれらを更に上回った。順位相関による評価では、難易度を考慮しない類似度 RAG が低評価となり、一方で提案手法は比較手法たちよりも顕著に高い性能を達成した。つまり、難易度を考慮する検索拡張生成によって、難易度の制御性を大きく改善できた。

目標難易度ごとの性能の変化　各手法がどのレベルへの平易化を得意とするのか分析した。評価用データを目標難易度ごとに分け、SARI を評価した。

Llama-8B による分析結果を図 3 に示す。3 つの比較手法は、A2 または B1 を目標とする際に比較的高い性能を発揮し、B2 では性能が最も悪化した。これは表 1 のデータ量と一貫しており、大規模言語モデルが一般的によく観測される目標難易度のテキスト生成を得意としている可能性が示唆される。対照的に 2 つの提案手法は、目標難易度が平易であるほど性能が向上した。

4 おわりに

本研究では、大規模言語モデルによる難易度制御の性能向上のために、難易度を考慮する検索拡張生成に取り組んだ。英語の文難易度制御に関する評価実験の結果、提案手法の有効性を確認し、類似度も併せて考慮することで更なる性能向上を確認した。

謝辞

本研究は、JSPS 科研費（基盤研究 B，課題番号：JP25K03233）の助成を受けて実施した。

参考文献

[1] Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. Data-Driven Sentence Simplification: Survey and Benchmark. **CL**, Vol. 46, No. 1, pp. 135–187, 2020.

[2] John Carroll, Guido Minnen, Yvonne Canning, Siobhan Devlin, and John Tait. Practical Simplification of English Newspaper Text to Assist Aphasic Readers. In **Proceedings of the AAAI-98 Workshop on Integrating Artificial Intelligence and Assistive Technology**, pp. 7–10, 1998.

[3] Sarah E Petersen and Mari Ostendorf. Text Simplification for Language Learners: A Corpus Analysis. In **Proceedings of the Workshop on Speech and Language Technology in Education**, pp. 69–72, 2007.

[4] Jan De Belder and Marie-Francine Moens. Text Simplification for Children. In **Proceedings of the SIGIR 2010 Workshop on Accessible Search Systems**, pp. 19–26, 2010.

[5] Batia Laufer. How Much Lexis is Necessary for Reading Comprehension? **Vocabulary and Applied Linguistics**, pp. 126–132, 1992.

[6] Carolina Scarton and Lucia Specia. Learning Simplifications for Specific Target Audiences. In **Proc. of ACL**, pp. 712–718, 2018.

[7] Daiki Nishihara, Tomoyuki Kajiwara, and Yuki Arase. Controllable Text Simplification with Lexical Constraint Loss. In **Proc. of ACL-SRW**, pp. 260–266, 2019.

[8] Daiki Yanamoto, Tomoki Ikawa, Tomoyuki Kajiwara, Takashi Ninomiya, Satoru Uchida, and Yuki Arase. Controllable Text Simplification with Deep Reinforcement Learning. In **Proc. of AACL**, pp. 398–404, 2022.

[9] Tannon Kew, Alison Chi, Laura Vásquez-Rodríguez, Sweta Agrawal, Dennis Aumiller, Fernando Alva-Manchego, and Matthew Shardlow. BLESS: Benchmarking Large Language Models on Sentence Simplification. In **Proc. of EMNLP**, pp. 13291–13309, 2023.

[10] Joseph Marvin Imperial and Harish Tayyar Madabushi. Flesch or Fumble? Evaluating Readability Standard Alignment of Instruction-Tuned Language Models. In **Proc. of GEM**, pp. 205–223, 2023.

[11] Joseph Marvin Imperial, Gail Forey, and Harish Tayyar Madabushi. Standardize: Aligning Language Models with Expert-Defined Standards for Content Generation. In **Proc. of EMNLP**, pp. 1573–1594, 2024.

[12] Asma Farajidizaji, Vatsal Raina, and Mark Gales. Is It Possible to Modify Text to a Target Readability Level? An Initial Investigation Using Zero-Shot Large Language Models. In **Proc. of COLING**, pp. 9325–9339, 2024.

[13] Ali Malik, Stephen Mayhew, Christopher Piech, and Klinton Bicknell. From Tarzan to Tolkien: Controlling the Language Proficiency Level of LLMs for Content Generation. In **Findings of ACL**, pp. 15670–15693, 2024.

[14] Abdullah Barayan, Jose Camacho-Collados, and Fernando Alva-Manchego. Analysing Zero-Shot Readability-Controlled Sentence Simplification. In **Proc. of COLING**, pp. 6762–6781, 2025.

[15] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In **Proc. of NeurIPS**, pp. 9459–9474, 2020.

[16] François Ledoyen, Gaël Dias, Jeremie Pantin, Alexis Lechervy, Fabrice Maurel, and Youssef Chahir. Facilitating Cognitive Accessibility with LLMs: A Multi-Task Approach to Easy-to-Read Text Generation. In **Proc. of EMNLP**, pp. 11782–11808, 2025.

[17] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language Models are Few-Shot Learners. In **Proc. of NeurIPS**, pp. 1877–1901, 2020.

[18] 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.

[19] Yuki Arase, Satoru Uchida, and Tomoyuki Kajiwara. CEFR-Based Sentence Difficulty Annotation and Assessment. In **Proc. of EMNLP**, pp. 6206–6219, 2022.

[20] Llama Team. The Llama 3 Herd of Models. **arXiv:2407.21783**, 2024.

[21] Qwen Team. Qwen2.5 Technical Report. **arXiv:2412.15115**, 2025.

[22] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient Memory Management for Large Language Model Serving with PagedAttention. In **Proceedings of the 29th Symposium on Operating Systems Principles**, p. 611–626, 2023.

[23] Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. Text Embeddings by Weakly-Supervised Contrastive Pre-training. **arXiv:2212.03533**, 2022.

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

[25] 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.

[26] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. BERTScore: Evaluating Text Generation with BERT. In **Proc. of ICLR**, 2020.