大規模言語モデルによるテキスト平易化のための意味的類似性と表層的非類似性に基づくパラレルコーパスフィルタリング

前川 大輔 [1]　梶原 智之 [2]　二宮 崇 [2]
[1] 愛媛大学工学部　[2] 愛媛大学大学院理工学研究科
{maekawa@ai.cs., kajiwara@cs., ninomiya.takashi.mk@}ehime-u.ac.jp

概要

大規模言語モデルは小規模なファインチューニングで高い性能を発揮できると他のタスクで報告されているが，テキスト平易化タスクにおいて必要なデータ量は未知である．本研究では，テキスト平易化のためのパラレルコーパスフィルタリングの手法を提案し，大規模言語モデルのファインチューニングに必要なデータ量を削減する．日本語における実験の結果，テキスト平易化のタスク遂行能力は 16〜64 文対という非常に少量の訓練データから獲得できることがわかった．充分なドメイン知識を得るにはより多くの訓練データが必要となるが，それでも提案手法によって，約 7 割の訓練データを削減しつつ全件で訓練するよりも高い性能を達成できた．

1　はじめに

テキスト平易化 [1] は，所与のテキストを理解しやすく変換するタスクであり，非母語話者 [2] や子ども [3] の文章読解を支援する技術である．本研究では，大規模言語モデル (LLM: Large Language Model) [4–6] による文の平易化に取り組む．

大規模言語モデルは，その知識と能力の大部分を事前訓練時に獲得している [7] と言われている．そのため，小規模な訓練データのみでファインチューニングできることが示されており，対話では 1,000 件 [7]，機械翻訳では 32 件 [8] のみのファインチューニングで高い性能が得られている．しかし，テキスト平易化タスクにおけるファインチューニングのために必要な訓練データ量は明らかになっていない．

先行研究 [7, 9] では，大規模言語モデルのファインチューニング用データは量よりも質が重要であると述べられており，単純に小規模な訓練用データを与えれば良いわけではない．テキスト平易化においては，後述するように，無作為に訓練データを削減すると性能は悪化してしまう．また，テキスト平易化のためのパラレルコーパスフィルタリングの手法も研究されているが，既存手法 [10] は訓練データに含まれる少量の低品質なノイズを除外するものであり，本研究で目指すような，少量の高品質なデータを選択するための技術ではない．

本研究では，テキスト平易化コーパスから高品質な文対を抽出するパラレルコーパスフィルタリングの手法を提案し，大規模言語モデルの訓練コストを削減する．まず，文対の意味的類似度に基づく既存のパラレルコーパスフィルタリング手法が本タスクには適さないことを説明する．そして，積極的な平易化を評価する表層的非類似度を考慮するパラレルコーパスフィルタリング手法を提案する．

3 つのドメインにおける日本語のテキスト平易化コーパスを用いた実験の結果，提案手法がパラレルコーパスフィルタリングの既存手法よりも高品質な文対を抽出する性能が高いことを確認できた．そして，提案手法で抽出した 16 文対から 64 文対という非常に少量の訓練データによるファインチューニングによって，大規模言語モデルがテキスト平易化のタスク遂行能力を充分に獲得できることが明らかになった．充分なドメイン知識を獲得するためにはより多くの訓練データが必要となるが，約 7 割のデータを削減した 4,000 文対のみの訓練によって，全件で訓練するよりも高い性能を達成できた．

2　関連研究

パラレルコーパスフィルタリング [11–13] は，大規模なパラレルコーパスを利用可能な機械翻訳タスクを中心に研究されてきた．教師なし手法としては，多言語文符号化器の LASER [14] や LaBSE [15] の余弦類似度に基づく手法 [16, 17] が提案されている．教師あり手法としては，翻訳確率などの特徴量を用いた機械学習に基づく手法 [18] や多言語マスク

言語モデルの XLM-R [19] を用いた深層学習に基づく手法 [20] が提案されている．本稿では，これらの機械翻訳における先行研究と比較しつつ，同一言語内の系列変換タスクであるテキスト平易化に適したパラレルコーパスフィルタリングについて考える．

3 提案手法

本研究では，大規模言語モデルの訓練コストを削減するために，テキスト平易化のためのパラレルコーパスフィルタリングの手法を提案する．小規模なパラレルコーパスで高性能なテキスト平易化モデルを訓練できることが明らかになれば，それはひいては本タスクの少資源問題への対策にもなり得る．

3.1 意味的類似度によるフィルタリング

まず，機械翻訳におけるパラレルコーパスフィルタリングの先行研究 [16, 17] を援用し，文埋め込みの余弦類似度に基づくテキスト平易化のパラレルコーパスフィルタリングの予備実験を実施する．

方法 日本語テキスト平易化コーパス SNOW [21, 22] の訓練データから LaBSE [15] の余弦類似度が低い文対を除外し，日本語 LLM の Swallow[1] を LoRA[2] [23] を用いてファインチューニングした．

結果 余弦類似度の高い文対でファインチューニングした LLM は，入力文をそのまま出力する過度に保守的なテキスト平易化モデルとなってしまった．これは，図 1 に示すように，意味的類似度（LaBSE の余弦類似度）が高い文対の多くは表層的類似度（BLEU [24]）も高いため，難解文と平易文に差分のない文対で訓練することになってしまったのが原因であると考えられる．なお，同じく図 1 からわかるように，表層的類似度の非常に高い文対が極めて多いため，無作為に訓練データを削減した場合にも同様の傾向が見られた．

3.2 意味的類似度と表層的非類似度

前節の課題に対処するために，本研究では，表層的類似度の低い文対，つまり難解文と平易文の間で多くの編集が行われている文対を積極的に残すようなパラレルコーパスフィルタリングの手法を提案する．しかし，単純に表層的類似度の低い文対を残すと，“カエルはヘビが恐い → 水の近くに住む緑のよく飛ぶ動物は長くて細い手と足がない動物が怖い”

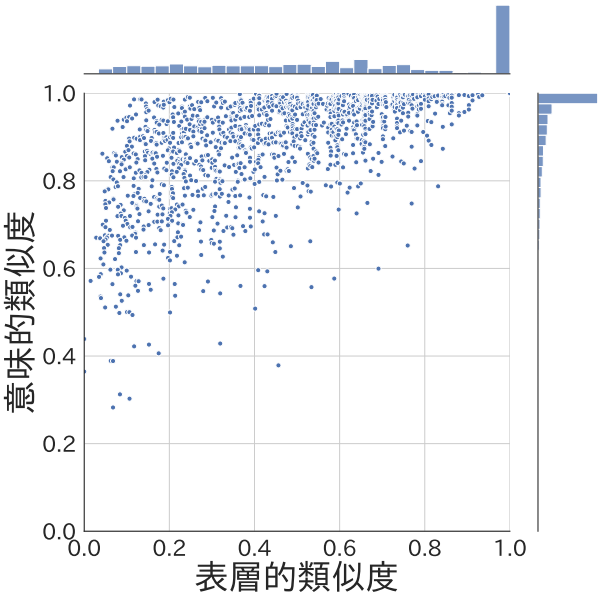

図 1 日本語テキスト平易化コーパスの SNOW における難解文と平易文の間の意味的および表層的類似度の分布

のようなノイズ文対を選んでしまう．そのため，意味的類似度が高く，かつ，表層的類似度が低い文対を選択するように提案手法を設計する．

具体的には，難解文 x および平易文 y からなる文対の品質 $Q(\cdot)$ を以下のように定義する．

$$Q(x, y) = \sqrt{(1 - \theta(x, y))^2 + (0 - \phi(x, y))^2} \quad (1)$$

ここで，$\theta(\cdot)$ は文対の意味的類似度であり，本研究では LaBSE[3] [15] の余弦類似度を用いる．また，$\phi(\cdot)$ は文対の表層的類似度であり，本研究では BLEU[4] [24] を用いる．これは，図 1 における左上端からのユークリッド距離を表しており，意味的類似度の高さと表層的類似度の低さのバランスをとるものである．パラレルコーパスフィルタリングにおいては，この値の高い順に訓練データから除外する．

4 評価実験

日本語のテキスト平易化タスクにおいて，提案手法を既存のパラレルコーパスフィルタリングの手法と比較しつつ，大規模言語モデルのファインチューニングに必要な訓練データ量を明らかにする．また，同じ設定で他の事前訓練済み系列変換モデルもファインチューニングすることにより，必要な訓練データ量の差分を明らかにする．

1) https://huggingface.co/tokyotech-llm/Llama-3.1-Swallow-8B-Instruct-v0.2
2) https://github.com/unslothai/unsloth
3) https://huggingface.co/sentence-transformers/LaBSE
4) https://github.com/mjpost/sacrebleu

4.1 実験設定

データ 大規模な日本語テキスト平易化コーパスである MATCHA[5] [25] および SNOW[6][7] [21, 22] を訓練と評価の両方に使用した．また，小規模な日本語テキスト平易化コーパスである JADES[8] [26] を評価用に使用した．各データの文対数を表 1 に示す．

モデル テキスト平易化モデルは，日本語 Wikipedia 上で事前訓練された系列変換モデルの BART[9] [27] および大規模言語モデルの Swallow[1] [5, 6] をファインチューニングして構築した．なお，出力文のデコーディングには貪欲法を使用した．

ハイパーパラメータ 訓練データを 2 件，4 件，8 件と 2 倍ずつ増やし，MATCHA は 8,000 件まで，SNOW は 64,000 件までの訓練データを用いて，それぞれテキスト平易化の性能を評価した．ファインチューニングの際には，バッチサイズを 16 文対，学習率を 5×10^{-5} に設定し、最適化手法には AdamW [28] を用いた。そして、検証用データにおける交差エントロピー損失が 3 エポック連続で改善しない場合に訓練を終了する early stopping を適用した。大規模言語モデルのファインチューニングには LoRA[2] [23] を使用し，ランクを $r = 16$，スケーリング係数を $\alpha = 16$，dropout 率を 0.05 に設定した．

比較手法 提案手法を以下の 6 つのパラレルコーパスフィルタリング手法と比較した．評価には，EASSE[10] に実装されている SARI [29] を用いた．

w/o Filtering 訓練データの全体を用いる．

0-shot ファインチューニングせず，「入力文を平易に言い換えてください」というプロンプトを与えてテキスト平易化を実行する（LLM のみ）

Random 無作為に文対を選択する．

Bicleaner-AI 機械翻訳タスクの教師ありパラレルコーパスフィルタリング手法[11] [20] を，テキスト平易化の訓練データで訓練して用いる．

COS 意味的類似度の降順に文対を選択する．提案手法から表層的類似度の考慮を外したもの．

BLEU 表層的類似度の昇順に文対を選択する．提案手法から意味的類似度の考慮を外したもの．

5) https://github.com/EhimeNLP/matcha
6) https://www.jnlp.org/GengoHouse/snow/t15
7) https://www.jnlp.org/GengoHouse/snow/t23
8) https://github.com/naist-nlp/jades
9) https://huggingface.co/ku-nlp/bart-large-japanese
10) https://github.com/feralvam/easse
11) https://github.com/bitextor/bicleaner-ai

表 1 日本語テキスト平易化パラレルコーパスの文対数

	訓練用	検証用	評価用
MATCHA	14,000	1,000	1,000
SNOW	82,300	2,000	700
JADES	-	-	3,907

4.2 実験結果

実験結果を図 2 に示す．この図は 6 つのグラフからなり，それぞれ訓練および評価に使用したテキスト平易化コーパスが異なる．上段は MATCHA で訓練した結果，下段は SNOW で訓練した結果，1 列目は MATCHA で評価した結果，2 列目は SNOW で評価した結果，3 列目は JADES で評価した結果である．それぞれのグラフは，横軸が訓練に使用した文対数，縦軸が SARI による評価値，実線が LLM の結果，破線が BART の結果を表している．

LLM によるドメイン内の評価 基本的には訓練データが多いほど高い性能を発揮した．しかし，提案手法において少量のノイズのみを削減する設定では，全件を用いて訓練するよりも高い性能を達成した．具体的には，MATCHA における提案手法の 4,000 件（約 3 割）および SNOW における提案手法の 64,000 件（約 8 割）がこれに該当し，訓練コストを減らしつつ性能を改善できた．また，訓練データ量の多少に関わらず，提案手法は比較手法よりも高い性能を達成する場合が多かった．これらの実験結果から，テキスト平易化のためのパラレルコーパスフィルタリングとしての提案手法の有効性が確認できた．なお，MATCHA は専門家によって構築されたコーパスであり，SNOW は非専門家によって構築されたコーパスであることから，先行研究 [7–9] と同様，高品質な訓練データであれば LLM は小規模なファインチューニングのみで高い性能を発揮できることが示唆された．その他，注目すべき点としては，COS ベースラインが他の手法とは異なる傾向を示した．COS ベースラインにおいては，MATCHA では下位 1,000 件，SNOW では下位 16,000 件までは，訓練データを増やしても平易化性能が改善されなかった．これは，意味的類似度の高すぎる文対が訓練を妨げるノイズであることを示唆している．

BART によるドメイン内の評価 BART は，大規模な訓練データが得られる状況（SNOW の 16,000 件以上）では LLM と同等の性能を示したものの，訓練データの減少に伴って著しく性能が低下した．

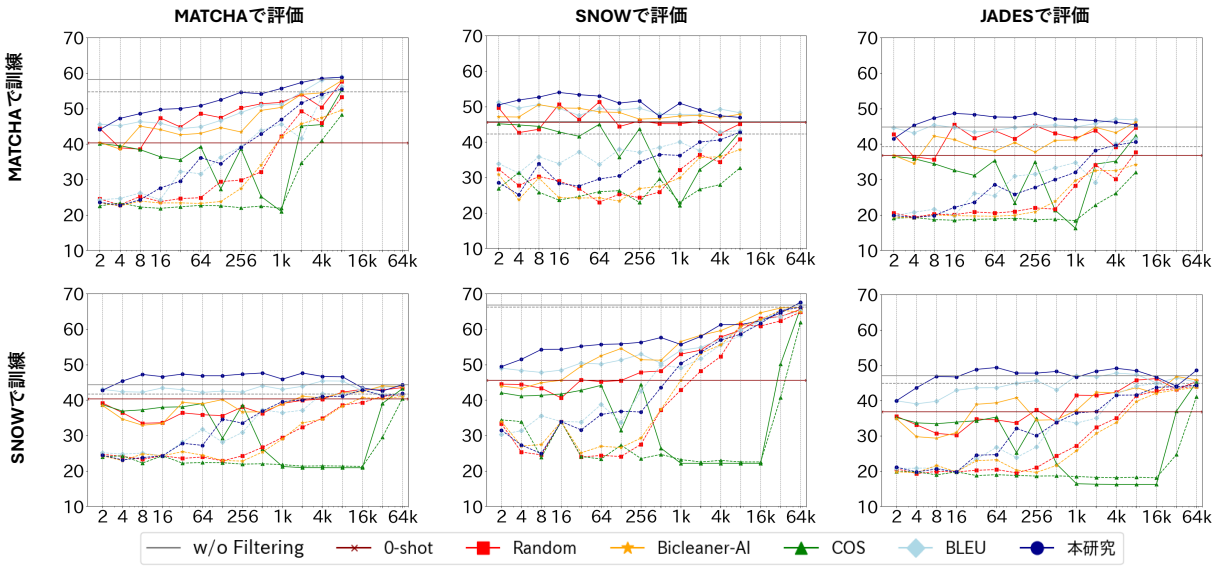

図 2　実験結果．横軸は訓練データの文対数で縦軸は SARI によるテキスト平易化性能．実線は LLM で破線は BART.

表 2　大規模言語モデルによる MATCHA コーパスのテキスト平易化の例．16 文対の訓練で文分割などの平易化能力を獲得できる．より多くの訓練によって，難しい語に英語訳を補足するという MATCHA コーパスの特性に適応できる．

入力文	花粉症とは、植物の花粉によって引き起こされるアレルギー反応のことで、日本の場合、スギやヒノキが原因となることが多いです。
出力文（16 文対で訓練）	花粉症は、植物の花粉が原因で起こるアレルギーです。日本では、スギやヒノキの花粉が原因になることが多いです。
出力文（全文対で訓練）	花粉症は、植物の花粉 (pollen) で起こるアレルギーです。日本では、スギやヒノキが原因になることが多いです。

BART においても，MATCHA では 4,000 件の訓練によって全件で訓練するのと同等の性能が得られていることから，高品質な訓練データであれば提案手法によって訓練コストを大きく削減できることが示唆された．なお，COS ベースラインについては，LLM による実験結果と同様の傾向が見られた．

LLM によるドメイン外の評価　ドメイン内の評価とは異なり，訓練データが多いほど高性能とはならなかった．提案手法では，MATCHA で訓練する場合には 16 件，SNOW で訓練する場合でも 64 件の訓練データで最高性能に到達し，より多くの訓練データを用いても平易化性能は改善されなかった．これらの実験結果から，テキスト平易化のタスク遂行能力は 16〜64 件程度の非常に少量のデータから獲得できることがわかる．一方で，ドメイン内の評価において訓練データ量を増やすほど性能が改善されたのは，ドメインに関する知識を獲得するために大規模な訓練データが必要であると言える．表 2 に実例を示す．なお，提案手法が他の手法よりも全体的に高性能を達成するのは，ドメイン内の評価と一貫しており，提案手法の有効性を確認できた．

BART によるドメイン外の評価　ドメイン内の評価と同様に，BART は訓練データの減少に伴って性能が低下した．そのため，LLM とは異なり，数十件という非常に少量のデータで高性能なテキスト平易化モデルを訓練することは困難である．

5　おわりに

本研究では，意味的類似度と表層的非類似度に基づくテキスト平易化のためのパラレルコーパスフィルタリングの手法を提案した．日本語における実験の結果，提案手法は既存手法よりも高品質な文対を抽出する性能が高いことを確認できた．そして，数十文対という少数の訓練データのみを用いたファインチューニングによって，大規模言語モデルがテキスト平易化のタスク遂行能力を獲得できることを明らかにした．充分なドメイン知識を獲得するためにはより多くの訓練データが必要だが，提案手法では約 7 割のデータを削減した 4,000 文対の訓練によって，全件で訓練するよりも高い性能を達成できた．

今後は，英語のテキスト平易化や他のスタイル変換タスクにおいて提案手法の有効性を検証したい．

謝辞

本研究は，戦略的イノベーション創造プログラム（SIP）「統合型ヘルスケアシステムの構築」JPJ012425の助成を受けたものです．

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] Sarah E. Petersen and Mari Ostendorf. Text Simplification for Language Learners: A Corpus Analysis. In **SLaTE**, pp. 69–72, 2007.

[3] Jan De Belder and Marie-Francine Moens. Text Simplification for Children. In **SIGIR**, pp. 19–26, 2010.

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

[5] Kazuki Fujii, Taishi Nakamura, Mengsay Loem, Hiroki Iida, Masanari Ohi, Kakeru Hattori, Hirai Shota, Sakae Mizuki, Rio Yokota, and Naoaki Okazaki. Continual Pre-Training for Cross-Lingual LLM Adaptation: Enhancing Japanese Language Capabilities. In **CoLM**, 2024.

[6] Naoaki Okazaki, Kakeru Hattori, Hirai Shota, Hiroki Iida, Masanari Ohi, Kazuki Fujii, Taishi Nakamura, Mengsay Loem, Rio Yokota, and Sakae Mizuki. Building a Large Japanese Web Corpus for Large Language Models. In **CoLM**, 2024.

[7] Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. LIMA: Less Is More for Alignment. In **NeurIPS**, pp. 55006–55021, 2023.

[8] Dawei Zhu, Pinzhen Chen, Miaoran Zhang, Barry Haddow, Xiaoyu Shen, and Dietrich Klakow. Fine-Tuning Large Language Models to Translate: Will a Touch of Noisy Data in Misaligned Languages Suffice? In **EMNLP**, pp. 388–409, 2024.

[9] Lichang Chen, Shiyang Li, Jun Yan, Hai Wang, Kalpa Gunaratna, Vikas Yadav, Zheng Tang, Vijay Srinivasan, Tianyi Zhou, Heng Huang, and Hongxia Jin. AlpaGasus: Training A Better Alpaca with Fewer Data. In **ICLR**, 2024.

[10] Koki Hatagaki, Tomoyuki Kajiwara, and Takashi Ninomiya. Parallel Corpus Filtering for Japanese Text Simplification. In **TSAR**, pp. 12–18, 2022.

[11] Philipp Koehn, Huda Khayrallah, Kenneth Heafield, and Mikel L. Forcada. Findings of the WMT 2018 Shared Task on Parallel Corpus Filtering. In **WMT**, pp. 726–739, 2018.

[12] Philipp Koehn, Francisco Guzmán, Vishrav Chaudhary, and Juan Pino. Findings of the WMT 2019 Shared Task on Parallel Corpus Filtering for Low-Resource Conditions. In **WMT**, pp. 54–72, 2019.

[13] Philipp Koehn, Vishrav Chaudhary, Ahmed El-Kishky, Naman Goyal, Peng-Jen Chen, and Francisco Guzmán. Findings of the WMT 2020 Shared Task on Parallel Corpus Filtering and Alignment. In **WMT**, pp. 726–742, 2020.

[14] Mikel Artetxe and Holger Schwenk. Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond. **TACL**, Vol. 7, pp. 597–610, 2019.

[15] Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. Language-agnostic BERT Sentence Embedding. In **ACL**, pp. 878–891, 2022.

[16] Vishrav Chaudhary, Yuqing Tang, Francisco Guzmán, Holger Schwenk, and Philipp Koehn. Low-Resource Corpus Filtering Using Multilingual Sentence Embeddings. In **WMT**, pp. 261–266, 2019.

[17] Akshay Batheja and Pushpak Bhattacharyya. Improving Machine Translation with Phrase Pair Injection and Corpus Filtering. In **EMNLP**, pp. 5395–5400, 2022.

[18] Víctor M. Sánchez-Cartagena, Marta Bañón, Sergio Ortiz-Rojas, and Gema Ramírez. Prompsit's submission to WMT 2018 Parallel Corpus Filtering shared task. In **WMT**, pp. 955–962, 2018.

[19] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised Cross-lingual Representation Learning at Scale. In **ACL**, pp. 8440–8451, 2020.

[20] Jaume Zaragoza-Bernabeu, Gema Ramírez-Sánchez, Marta Bañón, and Sergio Ortiz Rojas. Bicleaner AI: Bicleaner Goes Neural. In **LREC**, pp. 824–831, 2022.

[21] Takumi Maruyama and Kazuhide Yamamoto. Simplified Corpus with Core Vocabulary. In **LREC**, pp. 1153–1160, 2018.

[22] Akihiro Katsuta and Kazuhide Yamamoto. Crowdsourced Corpus of Sentence Simplification with Core Vocabulary. In **LREC**, pp. 461–466, 2018.

[23] 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 **ICLR**, 2022.

[24] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a Method for Automatic Evaluation of Machine Translation. In **ACL**, pp. 311–318, 2002.

[25] 宮田莉奈, 惟高日向, 山内洋輝, 柳本大輝, 梶原智之, 二宮崇, 西脇靖紘. MATCHA：専門家が平易化した記事を用いたやさしい日本語パラレルコーパス. 自然言語処理, Vol. 31, No. 2, pp. 590–609, 2024.

[26] Akio Hayakawa, Tomoyuki Kajiwara, Hiroki Ouchi, and Taro Watanabe. JADES: New Text Simplification Dataset in Japanese Targeted at Non-Native Speakers. In **TSAR**, pp. 179–187, 2022.

[27] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In **ACL**, pp. 7871–7880, 2020.

[28] Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. In **ICLR**, 2019.

[29] 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.