日本語医療テキスト平易化の評価用データセットの構築

堀口 航輝 [1]　梶原 智之 [2]　荒瀬 由紀 [3]　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　[3] 大阪大学大学院情報科学研究科
{horiguchi@ai., kajiwara@, ninomiya@}cs.ehime-u.ac.jp　arase@ist.osaka-u.ac.jp

概要

本研究では、医療用語を患者が理解しやすい表現に言い換える医療テキスト平易化のための日本語の言語資源を構築し、公開する。患者の闘病ブログから抽出した 1,425 文に対して、病名や症状に関する表現に対応する医療用語をアノテーションし、医療テキスト平易化のための評価用パラレルコーパスを構築する。また、訓練用パラレルコーパスが存在しない本タスクに取り組むために、他のドメインにおける日本語テキスト平易化モデルを援用しつつ、指定した表現を避ける出力文のリランキング手法を提案する。評価実験の結果、提案手法が医療テキスト平易化の性能改善に貢献することを確認できた。

1　はじめに

医師によって記述される医療文書には、ドメイン固有の専門用語が多く含まれるため、非専門家である患者はその情報をしばしば充分に理解できない [1]。医療従事者と患者の間のコミュニケーションを改善し、患者に自身の病気や症状、治療方針への詳細な理解を促すために、英語では医療テキスト平易化 [2–4] が盛んに研究されている。しかし、医療ドメインのテキスト平易化コーパスが存在しない日本語では、研究のアプローチが制限されている。

英語においては、医療ドメインに特化した事前訓練モデル [5, 6] やテキスト平易化タスクに特化した事前訓練モデル [7, 8] を用いるのが典型的なアプローチである。しかし、テキスト平易化に特化した事前訓練モデルは日本語では利用できない。

この問題に対処するための第一歩として、本研究では日本語の医療テキスト平易化のための評価用パラレルコーパス JASMINE[1)] を構築し、公開する。また、医療ドメインに特化した訓練用コーパスが存在しない日本語における医療テキスト平易化に取り組むために、テキスト平易化タスクに適応した事前訓練モデルである日本語 SimpleBART[2)] を構築する。さらに、専門用語を回避する生成文のリランキング手法を提案し、テキスト平易化モデルに適用する。JASMINE を用いた実験の結果、提案手法が日本語の医療テキスト平易化に有効であることが示された。

2　関連研究

2.1　テキスト平易化のパラレルコーパス

テキスト平易化タスクでは、難解な文と平易な文の対からなる単言語パラレルコーパスを用いて、系列変換モデルを訓練する。英語においては、対象読者の異なるニュースや Wikipedia の記事の対から自動的な文アライメントによって構築された Newsela [9] や Wiki-Auto [10] などのテキスト平易化パラレルコーパスが公開されている。日本語においては、教科書の例文やニュースを文単位で人手で平易に言い換えて構築された SNOW [11, 12] や JADES [13] が公開されている。

医療テキスト平易化は、医師が書いたテキストから専門用語を取り除き、患者にとって理解しやすい表現に言い換えるタスクである。医療テキスト平易化 [2–4] は英語では盛んに研究されているものの、医療ドメインのパラレルコーパスが存在しない日本語では研究のアプローチが制限されている。

2.2　テキスト平易化の事前訓練

近年の自然言語処理では、大規模な生コーパスで事前訓練した Transformer [14] を目的タスクでファインチューニングする転移学習のアプローチが主流である。テキスト平易化などの系列変換タスクでは、Transformer をノイズ除去自己符号化タスクで事前訓練した BART [15] が広く使用されている [16–18]。

目的のタスクによっては、タスクに特化した事前訓練が有効である。例えば、要約タスクにおいては

1）JASMINE: **JA**panese text **S**implification dataset in the **M**edical doma**IN** for **E**valuation https://github.com/EhimeNLP/JASMINE

2）https://github.com/EhimeNLP/JapaneseSimpleBART

表 1　JASMINE コーパスの事例（専門用語は赤字、専門用語を患者向けに平易化した表現は青字、下線はスタイル変換）

難解文	体調不良は、顔面浮腫と酔っ払った感覚が強くなっているためだろう。
平易文	体調悪いのは、顔面のむくみと酔っ払った感覚が強くなっているからだと思われる。

文章中の文をマスクして復元する事前訓練 [19]、言い換えタスクにおいては折り返し翻訳を復元する事前訓練 [20] の有効性が報告されている。テキスト平易化タスクにおいては、事前訓練済みの BART に対して平易な単語を対象とする単語穴埋めを追加訓練した SimpleBART [8] が、平易な単語の生成能力を改善できると報告されている。SimpleBART は英語のテキスト平易化タスクにおいて成功を収めているものの、日本語における取り組みはない。

3　評価用データセット JASMINE

医療ドメインにおける日本語のテキスト平易化モデルの性能評価のために、医療用語を含む文と含まない文からなるパラレルコーパス JASMINE[1)] を構築する。本研究では、非専門家である患者向けの医療テキストとして、患者によって記述された闘病ブログ[3)] に注目する。闘病ブログに含まれる文に対して、病名や症状に対応する表現を MedDRA 標準名辞書[4)] に登録されている医療用語に置き換えることによって、表 1 に例示するような専門用語を含まない平易文と専門用語を含む難解文からなるパラレルコーパスを構築する。なお、ブログ記事に含まれる口語体の表現は、難解文においては「だ・である」調の文体にスタイルを統一した。

3.1　アノテータによる書き換え

パラレルコーパスを構築するために、日本語テキストのアノテーションに精通した 2 名のアノテータを採用した。医療分野のアノテーション経験を持つアノテータ（医療従事者ではない）が専門用語を用いた書き換えを行い、他方のアノテータが確認と修正を行った。本コーパスは、闘病ブログの約 1,000 記事に含まれる約 17,000 文のうち、MedDRA 標準名辞書を用いて病名や症状を医療用語に置換できた 2,009 文対から構築されている。

3.2　著者による修正

前節で得た 2,009 文対には、文脈なしでは同義性を持たない文対や正式な文として完結していない箇条書きなど、ノイズとなる事例が含まれていた。そこで、全ての文対を著者が人手で確認し、これらのノイズとなる事例を除外し、さらに句読点の補完および絵文字や顔文字の削除などの修正を行った。最終的に、1,425 文対の医療テキスト平易化の評価のための日本語パラレルコーパスを構築した。

4　日本語のテキスト平易化モデル

医療テキスト平易化のために、テキスト平易化に特化した日本語の事前訓練モデルを開発し（4.1 節）、専門用語を避ける手法を提案する（4.2 節）。

4.1　日本語 SimpleBART

英語での先行研究 [8] に従い、生コーパス上での単語穴埋めの事前訓練を経た BART に対して、テキスト平易化パラレルコーパス上での追加の単語穴埋めの事前訓練によって平易な単語の生成能力を強化する SimpleBART の日本語版[2)] を開発する。なお、SimpleBART は、さらにテキスト平易化パラレルコーパス上での系列変換のファインチューニングを経て、テキスト平易化モデルになる。以下では、パラレルコーパスの平易文と難解文を用いたそれぞれの追加事前訓練の手法について説明する。

平易文に対するマスク処理　SimpleBART の追加訓練では、平易な単語をマスク対象として使用する。平易な単語の検出のために、英語の先行研究 [8] は単語難易度推定器 [21] を用いたが、その代わりに日本語では大規模な単語難易度辞書[5)] [22] を利用できる。この辞書は、40,605 単語のそれぞれに 3 段階の難易度（初級・中級・上級）が付与されている。本研究では、このうち初級および中級の単語を平易語と定義し、マスク対象とする。

先行研究 [8, 15] に従い、マスク対象の単語を実際にマスクする確率は 0.15 に設定する。しかし、上級単語や辞書登録外の単語を多く含む文に対してはマスクが適用されにくいため、文ごとにマスク対象単語の割合 t を考慮し、各文において 15% の単語がマスクされるよう調整する。また、より平易な単語をより頻繁にマスクするために、マスク確率に $0 \leq \theta \leq 1$ の重みを掛ける。本研究では、初級単語は

3）https://www.tobyo.jp/
4）http://sociocom.jp/~data/2018-manbyo/index.html
5）https://github.com/Nishihara-Daiki/lsj

$\theta = 1$、中級単語は $\theta = 0.75$、その他の単語は $\theta = 0$ である。最終的に、マスク確率 m は以下のとおり。

$$m = \min(\frac{0.15\theta}{t},\ 1.0) \tag{1}$$

難解文に対するマスク処理　SimpleBART の追加訓練においては、平易な単語の生成を促進するとともに、難解な単語の生成を抑制したい。そこで英語の先行研究 [8] では、語彙平易化辞書 SimplePPDB++ [23] を用いて、難解文中の難解語をマスクし対応する平易語を復元するという、テキスト平易化に特化した単語穴埋めを訓練している。本研究でも同様に、日本語の語彙平易化辞書[5] [22] を用いる。この辞書には 42,642 件の平易化単語対が登録されているが、高品質な言い換えのために、単語分散表現[6]の余弦類似度が 0.25 を超える 18,810 単語対のみを使用する。なお、複数の平易化候補が存在する場合、BERT[7] [24] の単語穴埋め確率が最も高い平易語を選択する。

先行研究 [8, 15] に従い、マスク確率は 0.15 に設定する。ただし、マスク対象は語彙平易化辞書に登録されている単語に限定されるため、文によってはマスクが適用されにくい場合がある。そこで、4.1 節と同様に式 (1) を用いて文ごとに動的にマスク確率を重み付けする。ここでは、語彙平易化辞書への登録単語は $\theta = 1$、その他の単語は $\theta = 0$ である。

4.2　語彙制約リランキング

専門用語を含まない患者向けの平易文を生成するために、複数の候補文の中から指定した語を含まない候補文を選択するリランキング手法を提案する。提案手法では、訓練済みのテキスト平易化モデルを用いて、ビームサーチや Top-p サンプリング [25] など任意のデコーディング手法によって、複数の候補文を生成する。これらの候補文を上位から順に調査し、指定した語を一切含まない候補文が見つかればそれを出力する。ただし、指定した語が全ての候補文に含まれる場合は、先頭の候補文を出力する。

本研究では、MedDRA 標準名辞書に登録されている全ての医療用語を避けるよう指定して実験するが、患者が既知の専門用語は出力文に含めるなど、より良い語彙制約の作成は今後の課題である。

6）https://cl.asahi.com/api_data/wordembedding.html
7）https://huggingface.co/cl-tohoku/bert-base-japanese-whole-word-masking

表 2　コーパスの文対数

	訓練用	検証用	評価用
SNOW	82,300	1,000	100
JADES		2,959	948
JASMINE		425	1,000

5　評価実験

医療ドメインにおけるテキスト平易化の性能を評価するために、本研究で構築した JASMINE を用いて実験を行う。さらに、他のドメインにおけるテキスト平易化の性能も評価するために、既存の日本語テキスト平易化パラレルコーパスである SNOW [11, 12] および JADES [13] を用いた実験も行う。専門用語の辞書[4]を使用可能な JASMINE においては、専門用語を避ける語彙制約リランキングの有効性を検証する。表 2 にコーパスの統計を示す。

5.1　実験設定

モデル　日本語 Wikipedia を用いて事前訓練された BART[8] [15] と、この BART に対して追加の事前訓練を適用する以下の 3 モデルを用いた。

- BART-CP：SNOW 上で BART と同様のマスク言語モデリングを追加訓練する比較手法
- SimpleBART：SNOW 上で 4.1 節の平易な単語に注目したマスク言語モデリングを追加訓練
- SimpleBART-CP：SNOW 上で BART と同様のマスク言語モデリングを追加訓練し、さらに平易な単語に注目したマスク言語モデリングを訓練

これらの事前訓練モデルに対して、SNOW 上でさらに系列変換のファインチューニングをして、テキスト平易化モデルを構築した。

追加の事前訓練　上述の追加の事前訓練は SNOW の訓練セットを用いて 10 エポック行った。前処理として、Juman++ [26] による単語分割および SentencePiece [27] によるサブワード分割を行った。バッチサイズは 64、Dropout 率は 0.1 とし、最適化手法には AdamW [28] を用いた。学習率スケジューリングには polynomial decay を用い、最大の学習率は $5e-5$、warmup ステップ数は 5,000 とした。

ファインチューニング　追加の事前訓練と同じ前処理を施した SNOW の訓練セットをファインチューニングにも用いた。Dropout 率を 0.3、最大の

8）https://huggingface.co/ku-nlp/bart-base-japanese

表 3　「腸閉塞」および「腸管拡張」を語彙制約とするテキスト平易化の出力例

入力文	度重なる腸閉塞と腸管拡張があったため、結局腹部には効いていないとの判断になった。
BART	度重なる腸の閉塞と腸管の拡張があったため、結局腹には効いていないとの判断になった。
SimpleBART	何度も腸の閉塞と腹の臓器の拡張があったため、結局腹部には効果がないとの判断になった。
+ 語彙制約	何度も腸の障害があったため、結局腹には効果がないとの判断になった。

表 4　実験結果：SARI による自動評価

	SNOW	JADES	JASMINE		
Decoding	Beam	Beam	Beam	Beam	Top-p
Reranking				✓	✓
BART	64.55	34.51	33.16	36.88	36.13
BART-CP	64.68	37.45	34.03	37.15	36.67
SimpleBART	64.79	35.50	**36.12**	**38.02**	**37.27**
SimpleBART-CP	**65.06**	**37.70**	34.80	37.05	36.59

学習率を $3e-5$、warmup ステップ数を 2,500 に変更し、検証用データのクロスエントロピー損失を用いて 5 エポックの early stopping にて訓練を終了した。

推論　評価のため、ビーム幅 5 のビーム探索を用いて平易文を生成した。JASMINE では、4.2 節の語彙制約リランキングを用いて平易文を生成する実験も行った。語彙制約リランキングにおいては、ビーム幅 100 のビーム探索および $p = 0.95$ の Top-p サンプリングによって、100 件ずつの候補文を生成した。

評価指標　テキスト平易化の性能は、EASSE [29] を用いて SARI [30] を自動評価した。

5.2　実験結果

実験結果を表 4 に示す。SimpleBART が一貫して BART の性能を上回ったことから、日本語のテキスト平易化に特化した事前訓練の有効性を確認できた。同じく SNOW を用いて追加訓練を行う BART-CP も一貫して BART の性能を上回っているものの、JADES 以外の設定では SimpleBART がより高い性能を達成した。また、ドメインによっては、単純な追加訓練とテキスト平易化に特化した追加訓練の両方を実施した SimpleBART-CP が最高性能を達成した。これらの実験結果は、平易な単語に焦点を当てた追加訓練の有効性を示している。

JASMINE における評価に注目すると、語彙制約リランキングによって、平易化性能を大きく改善できることがわかる。デコーディング手法に関しては、ビーム探索が常に Top-p サンプリングを上回った。表 5 は、語彙制約リランキングにおいて、候補文の数 n を 1 から 200 まで変化させた場合の、専門用語を含まない出力文の割合を示したものである。語彙制約リランキングによって、平易化性能の SARI だけでなく、専門用語を含まない出力文数も大きく改善できることがわかる。リランキング候補の数が多いほど専門用語を含まない文の割合は高くなるが、10 文のみのリランキングでも大きな効果がある。表 3 は医療テキスト平易化の出力例である。

表 5　SimpleBART の語彙制約リランキングで候補数 n を変化させた際の専門用語を含まない出力文の割合（ビーム幅も n に合わせたが、$n = 1$ のときはビーム幅 5 とした）

	$n = 1$	$n = 10$	$n = 50$	$n = 100$	$n = 200$
Beam	77.4	91.8	94.9	96.2	97.2
Top-p	80.8	85.0	86.9	87.4	88.2

6　おわりに

本研究では、患者の闘病ブログに基づく医療テキスト平易化の評価用パラレルコーパス JASMINE[1] を構築し、日本語のテキスト平易化に特化した事前訓練モデル日本語 SimpleBART[2] を開発した。3 つのドメインにおける評価実験の結果、日本語 SimpleBART は一貫して BART の性能を上回り、テキスト平易化タスクに特化した事前訓練の有効性を確認できた。また、医療ドメインにおいては、専門用語を避けるための語彙制約リランキングを提案し、さらに平易化性能を改善できた。

今後の課題として、医療ドメインにおける日本語のテキスト平易化のための大規模なパラレルコーパスを構築し、テキスト平易化モデルを訓練したい。また、患者が既知の専門用語は出力文に含めるなど、より適切な語彙制約についても検討したい。

謝辞

本研究は、厚生労働科学研究費補助金 AC 事業 JPMW21AC5001 および内閣府総合科学技術・イノベーション会議の戦略的イノベーション創造プログラム（SIP）第 3 期「統合型ヘルスケアシステムの構築」（研究推進法人：JST）の助成を受け実施された。

参考文献

[1] Christina Cheng and Matthew Dunn. Health Literacy and the Internet: A Study on the Readability of Australian Online Health Information. **Australian and New Zealand journal of public health**, Vol. 39, No. 4, pp. 309–314, 2015.

[2] Yixin Cao, Ruihao Shui, Liangming Pan, Min-Yen Kan, Zhiyuan Liu, and Tat-Seng Chua. Expertise Style Transfer: A New Task Towards Better Communication between Experts and Laymen. In **Proc. of ACL**, pp. 1061–1071, 2020.

[3] Ashwin Devaraj, Iain Marshall, Byron Wallace, and Junyi Jessy Li. Paragraph-level Simplification of Medical Texts. In **Proc. of NAACL**, pp. 4972–4984, 2021.

[4] Junyu Luo, Junxian Lin, Chi Lin, Cao Xiao, Xinning Gui, and Fenglong Ma. Benchmarking Automated Clinical Language Simplification: Dataset, Algorithm, and Evaluation. In **Proc. of COLING**, pp. 3550–3562, 2022.

[5] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. BioBERT: A Pre-trained Biomedical Language Representation Model for Biomedical Text Mining. **Bioinformatics**, Vol. 36, No. 4, pp. 1234–1240, 2019.

[6] Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. Publicly Available Clinical BERT Embeddings. In **Proceedings of the 2nd Clinical Natural Language Processing Workshop**, pp. 72–78, 2019.

[7] Renliang Sun and Xiaojun Wan. SimpleBERT: A Pre-trained Model That Learns to Generate Simple Words. **arXiv:2204.07779**, 2022.

[8] Renliang Sun, Wei Xu, and Xiaojun Wan. Teaching the Pre-trained Model to Generate Simple Texts for Text Simplification. In **Findings of the Association for Computational Linguistics: ACL 2023**, pp. 9345–9355, 2023.

[9] Wei Xu, Chris Callison-Burch, and Courtney Napoles. Problems in Current Text Simplification Research: New Data Can Help. **TACL**, Vol. 3, pp. 283–297, 2015.

[10] 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.

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

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

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

[14] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, L ukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In **Proc. of NIPS**, pp. 5998–6008, 2017.

[15] 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 **Proc. of ACL**, pp. 7871–7880, 2020.

[16] Louis Martin, Angela Fan, Éric de la Clergerie, Antoine Bordes, and Benoît Sagot. MUSS: Multilingual Unsupervised Sentence Simplification by Mining Paraphrases. In **Proc. of LREC**, pp. 1651–1664, 2022.

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

[18] Tatsuya Zetsu, Tomoyuki Kajiwara, and Yuki Arase. Lexically Constrained Decoding with Edit Operation Prediction for Controllable Text Simplification. In **Proc. of TSAR**, pp. 147–153, 2022.

[19] Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. PEGASUS: Pre-Training with Extracted Gap-Sentences for Abstractive Summarization. In **Proc. of ICML**, 2020.

[20] Tomoyuki Kajiwara, Biwa Miura, and Yuki Arase. Monolingual Transfer Learning via Bilingual Translators for Style-Sensitive Paraphrase Generation. In **Proc. of AAAI**, pp. 8042–8049, 2020.

[21] Chunguang Pan, Bingyan Song, Shengguang Wang, and Zhipeng Luo. DeepBlueAI at SemEval-2021 Task 1: Lexical Complexity Prediction with A Deep Ensemble Approach. In **Proc. of SemEval**, pp. 578–584, 2021.

[22] Daiki Nishihara and Tomoyuki Kajiwara. Word Complexity Estimation for Japanese Lexical Simplification. In **Proc. of LREC**, pp. 3114–3120, 2020.

[23] Mounica Maddela and Wei Xu. A Word-Complexity Lexicon and A Neural Readability Ranking Model for Lexical Simplification. In **Proc. of EMNLP**, pp. 3749–3760, 2018.

[24] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In **Proc. of NAACL**, pp. 4171–4186, 2019.

[25] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The Curious Case of Neural Text Degeneration. In **Proc. of ICLR**, 2020.

[26] Arseny Tolmachev, Daisuke Kawahara, and Sadao Kurohashi. Juman++: A Morphological Analysis Toolkit for Scriptio Continua. In **Proc. of EMNLP**, pp. 54–59, 2018.

[27] Taku Kudo and John Richardson. SentencePiece: A Simple and Language Independent Subword Tokenizer and Detokenizer for Neural Text Processing. In **Proc. of EMNLP**, pp. 66–71, 2018.

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

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

[30] 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.