別添４

厚生労働科学研究費補助金
（政策科学総合研究事業（臨床研究等 ICT 基盤構築・人工知能実装研究事業））
分担研究報告書

課題名　　：新薬創出を加速する症例データベースの構築･拡充/創薬ターゲット推定アルゴリズムの開発

研究分担者名：荒瀬 由紀
国立大学法人大阪大学大学院 情報科学研究科 マルチメディア工学専攻 准教授

研究要旨
本研究ではブログ等の患者自身が記述するテキストを対象とし、患者のみしか知り得ない精神・神経症状も含めた反応の記述を特定し、医薬品の奏功及び副作用知識を抽出することで、創薬や新たな薬効の発見への貢献を目指す。患者テキストは記述者によって使用する語彙や記述の粒度、スタイルも大きく異なる。そこでこれら表現の多様性に頑健な手法として、患者テキストを医療用語およびフォーマルかつ簡潔な文体を用いたものに自動的に書き換えるテキスト正規化手法を開発する。深層学習による言語生成モデルに対し語彙制約を付与することで、精度の高いテキスト正規化を実現する。同種のタスクである英語テキスト平易化において既存研究との性能を網羅的に比較評価したところ、提案手法は現時点での State-of-the-art を保持していることを確認した。さらに患者テキストを人手で正規化したコーパスを構築し、提案手法の性能を評価した。その結果、提案手法は流暢性を多少犠牲にする傾向はあるが、高い水準で制約を満たした文を生成できることを確認した。

A．研究目的

患者自身が記述する闘病ブログ等のテキストには、患者のみしか知りえないつぶさな症状の記録がなされる。このような患者テキストからの情報抽出は、これまでの医師主導の評価に対し、Patient Reported Outcome（PRO）として注目を浴びつつある。しかし、PRO の対象としては、患者の幸福度など余命に関する受け止め方や QOL に関する指標が多く、新薬開発やドラッグリポジショニングに必要な医学的な症状への適応は少ない。本研究では自然言語処理技術により、文脈を補いながら、患者のみしか知り得ない精神・神経症状も含めた反応の抽出を行う。

患者テキストは記述者によって使用する語彙や記述の粒度、スタイルも大きく異なるため、これら表現の多様性に頑健な手法の開発が必要である。そこで自然言語処理分野で活発に研究が進められている深層学習による言語生成モデルを応用した表現の正規化手法を開発する。多様性の高い患者テキストに対し、所与の標準的語彙と、フォーマルかつ簡潔な文体を用いるよう自動で書き換える正規化を行うことで、後段の様々なテキスト処理の品質を改善できると期待される。本手法を開発・評価するには、患者テキストと正規化後の文からなるパラレルコーパスが必要となるが、そのようなコーパスは存在しない。そこで本研究では人手によるコーパス構築にも取り組む。

B．研究方法

本研究ではテキストに現れる多様な表現を正規化するよう自動的に書き換えるモデルを開発する。具体的には事前学習済みの言語生成モデルに対し、所与の語彙を用いるよう入力文を書き替える語彙制約を課すことで正規化を実現する。以下では出力文に出現すべき単語の制約を正の制約、出力文に出現すべきでない単語の制約を負の制約と呼ぶ。提案手法では生成した正・負の制約を用い、Neurologic Decoding (Lu et al. 2021) により言語生成モデルに語彙制約を付加し、生成確率が高く、かつ制約をできるだけ満たしたテキストの生成を行う。

Neurologic Decoding は言語生成モデルの再学習を必要とせず、学習済みモデルの推論（デコーディング）において語彙制約を付与する手法である。ビームサーチにおいて正・負の制約が満たされたかどうかの状態を追跡しつつ出力候補を生成することで、生成確率が高く、かつ制約を満たした候補を探索する。既存の語彙制約手法では制約数が増えるに従って計算量が大幅に増大する問題があったが、Neurologic Decoding は

語彙制約を目的関数のペナルティとし、状態追跡における計算を再利用することで効率的な語彙制約を実現している。本研究では Neurologic Decoding により、正・負双方の語彙制約を考慮することで患者テキストの正規化を行う。

患者テキスト正規化コーパスの構築

作業者によって作文の質が変化することを防ぎ質の高いコーパスを構築するため、正規化におけるルールを定めたアノテーションガイドラインを作成する。患者テキストには文境界があいまいなものも多く存在するため、まずは自動で行った分割を人手で修正する文分割を行う。その後、テキスト中の症状の記述に対し、対応する MedDRA 標準名を使用するよう書き換えを行った。さらに様式を新聞文体に統一し、周辺の記述から分かる範囲で 5W1H を補完した。

（倫理面への配慮）

本研究は、大阪大学の研究倫理審査委員会で承認を受け、実施している。

C．研究結果

正規化コーパス

本研究でこれまで構築してきた闘病ブログコーパス中のテキストについて、人手で作文した正規化文を加えることで、既存リソースの拡張を行った。727 記事から抽出した 2,009 文について正規化を行い、正規化前後の文対からなるパラレルコーパスを作成した。作文の例を下表に示す。

患者文	正規化後
今　抗がん剤終了から 7 時間半経ちましたが、体がぽっぽと熱いです	現在、抗がん剤終了後 7 時間半経過、ほてりがある。
顔がまた暑くてホワホワするので 今日はこんなところで	ほてりが出ているので、今日はここまで。
ホットフラッシュは気温が涼しくなってきているので特に辛くはないです	気温が下がってきているので、ほてりは辛くない。

テキスト平易化での網羅的評価実験

本研究で目指す患者テキスト正規化と同種のタスクであり、また自然言語処理分野において活発に研究されている英語テキストの平易化タスクに提案手法を適用することで、既存研究との網羅的比較実験を実施した。標準的データセットである Newsela-Auto を対象とし、言語生成モデルには事前学習済み系列変換モデルである BART を用いた。提案手法では BART を Newsela-Auto によりファインチューニングしたモデルをベースラインとし、それに対して語彙制約を付与する。

単語の「追加」「維持」「削除」の性能に基づき、テキスト平易化の総合的な評価値を算出する SARI の結果を下表に示す。提案手法は既存手法に比べ顕著に高い SARI スコアを達成しており、現時点で state-of-the-art の性能を保持している。特に既存手法では単語の追加や削除といった書き換えに消極的という課題があったが、提案手法は「追加」「削除」においても顕著に高い性能を達成している。

手法	SARI	追加	維持	削除
(Dong+2019)	37.4	1.0	34.8	76.5
(Kajiwara 2019)	38.3	**4.4**	40.5	70.0
(Agrawal+2021)	40.5	1.2	**44.7**	75.5
BART Fine-tuning	39.7	4.1	39.2	75.7
提案手法	**43.1**	**4.4**	42.7	**82.1**

さらに Amazon Mechanical Turk を用い、英語母語話者に平易化前後の文を比較し評価してもらう人手評価を実施した。提案手法では、文の流暢性と書き換え前後の意味の保持性能が既存手法に比べ改善されていることを確認できた。

患者テキスト正規化コーパスでの評価実験

患者テキスト正規化コーパス（2,009 文対）により実験を行った。患者テキスト正規化コーパスからランダムに 500 文をサンプルしてテストセットとし、残りを訓練セットとした。日本語言語生成モデルとして京都大学より公開されている BART 日本語 Pretrained モデル（田中 et al. 2021）を利用した。言語生成モデルは一般ドメインのコーパスで言語モデルとしての事前訓練をされているため、そのままでは言い換え生成を行うことは難しい。そこで医療ドメインへの適応と言い換えの訓練を行うため、（正規化した部分を除いた）患者ブログテキストおよび Wikipedia 日本語テキストを折り返し翻訳することで疑似的な訓練コーパス（約 70 万文対）を作成、追加訓練を行った。さらに一般的な言い換えではなく、テキスト正規化を行うよう調整するため、患者テキスト正規化コーパスの訓練セットにより fine-tuning を行った。NeuroLogic Decoding で必要な制約の作成には MedDRA 辞書を用いた。

語彙制約を用いない BART モデルと比較した際の SARI スコアおよび制約の充足率を下表に示す。提案手法の SARI スコアは BART に比べやや低い値となった。一方で、制約の充足率を検証したところ、BART では正負の制約の充足率がそれぞれ 35.3%、48.1%という低い値となったのに対し、提案手法ではそれぞれ 70.0%, 60.2%となった。すなわち、提案手法は制約を満たすためにやや流暢性を犠牲にする傾向があるものの、高い水準で指定した語彙を用いた正規化を行えることが示された。

手法	SARI	制約の充足率（%）	
		正	負
BART Fine-tuning	**52.0**	35.3	48.1
提案手法	50.8	**70.0**	**60.2**

BART・提案手法それぞれによる正規化文の出力例を下表に示す。表が示す通り、提案手法では「目が痛い→眼痛」、「ムーンフェイス→満月様顔貌」、「頬が痛い→顔面痛」のように、MedDRA 用語を用いるよう正規化できている。一方で、制約を満たす文の候補が見つからず、文を出力できない場合もあり、これが SARI スコアを下げる要因となっていた。BART は流ちょうな文を生成するものの、制約を完全に満たすような文は少数であった。また 1 行目の例（「目が痛い→胃痛」）のように、元の文とは異なる症状に書き換えてしまう場合もあった。

患者テキスト	BART Fine-tuning	提案手法
いつもの如く、エンドキサンをし始めると目が痛くなりました	いつものように胃痛が出た。	いつものようにエンドキサンをし始めると、眼痛が悪化した。
ムーンフェイスはほんの少しマシになった気はしている	ムーンフェイスが少しマシになった気はある。	満月様顔貌が少し良くなった気がする。
火照って、頬が痛い	ほてりがある。	ほてりがあり、顔面痛がある。
足の発疹	足に発疹がある。	下肢 皮疹がある。
爪、真っ黒	爪の障害がある。	爪の色素沈着がある。

D．結論

本研究では患者テキストに現れる多様な表現を正規化することを目的とし、患者テキスト正規化コーパスの構築および自動正規化手法の開発を行った。提案手法の性能を網羅的に評価するため、同種のタスクである英語テキスト平易化における既存手法との比較を実施、顕著な性能を達成することを確認した。提案手法を患者テキスト正規化コーパスに適用したところ、流暢性の改善に課題はあるものの、高い水準で制約を満たした正規化文を生成できることを確認できた。提案手法が流暢性を犠牲にする要因として、追加訓練データの不足が考えられる。今後患者テキストを追加でマイニングし、折り返し翻訳することでコーパスを拡張し、流暢性の改善を行いたい。

E．研究発表

1．論文発表

1) 舌 達也, 梶原 智之, 荒瀬 由紀. 編集操作予測に基づく語彙制約付きデコーディングによるテキスト平易化の難易度制御. 自然言語処理（2023 年 4 月 3 日時点で条件付き採録）.

2．学会発表

1) Tatsuya Zetsu, Tomoyuki Kajiwara, and Yuki Arase. 2022. Lexically Constrained Decoding with Edit Operation Prediction for Controllable Text Simplification. In Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022), pages 147-153.

F．知的財産権の出願・登録状況（予定を含む。）

1．特許取得

なし

2．実用新案登録

なし

3．その他

なし