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ため、これら表現の多様性に頑健な手法として、患者テキストを所与の標準的な語彙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テキスト平易化データセットによる評価実験の結果、既存の state-of-the-art を大幅に上回る性能を達成することを確認した。

Ａ．研究目的

患者が記述する闘病ブログ等の Social Networking Service 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。本研究目的と同種のタスクであり、かつ自然言語処理分野において標準的に用いられるデータセットが利用可能なテキスト平易化により手法の開発と評価を行う。

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。語彙制約は入力文から正規化すべき箇所を予測する編集操作予測モデルにより自動的に生成する。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。得られた正・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。

編集操作予測モデル

テキストを正規化する上で、入力文中のどの単語を書き換えるべきかを人手で指定するのは利用者の負荷が高く、実用的でない。そこで本研究では、入力文中で書き換えるべき単語を自動で推定する編集操作予測モデルを構築する。具体的には、入力文中の各単語に対してどのような編集操作を行うべきか予測する。編集操作は、「削除」「保持」「置換」の３種類とする。「削除」は出力すべきでない単語、「保持」は出力にそのまま出現してもよい単語である。そして「置換」は別の単語に言い換えるべき単語である。予測モデルには事前学習済み言語モデルを用い、入力文の各単語に対してそれぞれ編集操作を予測する。編集操作の正解ラベルを人手でアノテーションするのは非常にコストが高い。そこで平易化前後の文について単語アラインメ

ントを行い、疑似的な正解ラベルを作成する。
予測ラベルが「削除」であった単語は負の制約、「保持」であった単語は正の制約とする。また予測ラベルが「置換」であった単語については、入力文中の該当単語を負の制約、その言い換えとして尤もらしい単語を正の制約とする。この言い換え単語予測には語彙言い換えデータセットを用いて訓練した事前学習済み言語モデルを用いる。

語彙制約付き言語生成
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。本研究では編集操作予測に基づき作成する正・負の語彙制約を用い、言語生成を行う。

（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
実験設定
テキスト平易化において標準的に用いられるデータセットである Newsela-Auto を用いて評価実験を実施した。編集操作予測モデルには事前学習済み言語モデルである BERT を用いた。Newsela-Auto の訓練データに対し単語アラインメントにより疑似正解ラベルを付与し、BERT のファインチューニングを実施した。また「置換」ラベルが付与された単語の言い換え候補を推定する語彙言い換えモデルには RoBERTa を用い、語彙言い換えデータセットによりファインチューニングを実施した。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。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、それに語彙制約を付与する提案手法と比較した。テキスト平易化で標準的に用いられる SARI を評価指標として評価を行った。

結果と考察
評価実験の結果、BART をファインチューニングしたベースラインの SARI スコアは 37.1、提案手法では 41.6 となり、顕著な性能改善を達成した。またテキスト平易化における既存研究における state-of-the-art 手法の SARI スコアは 36.6 であり、提案手法は現時点の世界最高性能である。
また理想的な語彙制約が生成できた場合のオラクルの性能についても調査した。正解文を参照し、編集操作予測モデルの疑似正解ラベルと同様の方法で理想的な（編集操作予測モデルの予測が全て正解であった場合の）語彙制約を生成した。この場合の SARI スコアは 62.2 であった。編集操作予測モデルの改善により、今後さらなるテキスト平易化の性能改善が見込めることが明らかとなった。

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手法の開発を行い、同種の問題設定であるテキスト平易化において顕著な性能を達成することを確認した。今後は患者テキストでの実験を実施する予定である。我々の研究グループではこれまで、IPF 及びがん患者の闘病ブログ 1,019 記事から抽出したテキストに対し、投与を受けている薬剤名、それによって起こった反応を特定し、ICD-10 及び MedDRA 分類を付与したデータセットを構築してきた。データセット中の文に対し、MedDRA で定義されたラベルを症状名として記述するよう言い換えるアノテーションを実施することで、患者テキスト正規化のパラレルコーパスを構築する。本データを用い、提案手法の評価とさらなる改善を実施する予定である。

E．研究発表
1．論文発表
なし

2．学会発表
なし

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