11/16 課題

調べた論文の中で最も興味が惹かれたものは、小説等の執筆を支援するシステムに関する論文であった [1]. この論文では，文学作品の執筆者が良い作品を作り上げるために豊かな文章表現を推敲することは重要な過程の一つであるとし，このような過程における支援を目的としている．文章表現を推敲する際，コロケーションや比喩表現，類語の辞典を調べることは執筆者の表現の幅を広げる上で役立つものであるが，これらは多くの場合網羅的であり，その中から執筆者の書いた文章の文脈に沿った文章表現を探し出す必要がある。そこで，この論文では Encoder-Decoder モデルを応用することによって, 文脈を考慮した文章表現の提案を行う手法について研究されている。

データセットとしては新字体および現代仮名遣いで書かれた青空文庫の作品約 469 万文字のテキストデータが使用されており，これに対して形容語の除去，直喩法の除去, 反復法の除去, 技巧的な言葉の平易化を行うことによって，経験の少ない執筆者をを想定した文書集合を疑似的に作成している。このデータから訓練データ，検証データおよびテストデータを作成し，Transformer を用いた生成モデルが構築されている。語彙数は 50,000 語で，それ以外のトークンは未知語トークンとして処理される。このモデルを用いて，ユーザの入力から生成された出力をもとに候補文を提案するシステムが構築されている。出力に含まれる未知語トークンは，Transformer モデルから抽出した Attention スコアをもとに最も関連度の高い入力のトークンへ置換することによって，未知語処理される。またこのシステムでは複数の候補を提案することが目的であるため，推論に 10 窓のビームサーチを用い，スコア上位の 6 つの出力を候補文としてユーザに提案される仕組みになっている。ただし，入力と同じ出力文である場合や，出力文に含まれる未知語トークンの数が入力文より多い場合は候補に含まれない。

このシステムを用いた例としては,「草木が揺れていた」という文章を入力した場合，形容語や擬態法が付加された「草木がゆらゆらと揺れていた」や「草木が静かに揺れていた」というような文が出力される。このほかに直喩法や反復法を付加したり，言い換えを行ったりした文の出力も存在する。実験の評価としてはモデルの自動評価とシステムの人手評価が行われており，前者は未知言語処理を行っていないモデルと行ったモデルに対して自動評価を行い，後者は豊かさ（付加された文章表現が豊かであるか，乏しいか），流暢さ（文として自然か，不自然か），意味の保持（意味内容が保たれているか，ずれているか），総合評価（小説や随筆の執筆支援システムとして，入力文からその文へ変更したいと思うか）の 4 つの評価尺度で 5 段階評価をしてもらっており，総合的に良い評価であることが確かめられている。

私は創作活動をする人の支援ができるような研究をしたいと考えており，この論文で研究されている内容に興味を持った。生成される提案モデルの傾向をユーザ側でもある程度いじれるようにできれば，さらに実用的なものになるのではないかと感じた。また趣味として「小説家になろう」といったサイトに投稿されているネット小説をよく読むため，経験が少ないと考えられる執筆者の文章を多く読んできたことも理由の一つである。

参考文献

[1] 鈴木勘太, 杉本徹. Encoder-decoder モデルを用いた文章表現を豊かにする執筆支援システム．言語処理学会 第 27 回年次大会 発表論文集, pp. 1862–1866, 2021.