MLM の継続事前学習と口調マルチラベル分類による ライトノベルにおける話者推定

興津 舞花

要旨

小説において，会話文は登場人物の特徴や関係性を示す比重の大きい要素である．小説内の登場人物の発話とその話者を紐づける行為を話者推定という．計算機による話者推定手法は，会話文の直前・直後に記載される話者名や，連続する会話文では話者が交代するという一般的な規則等に着目するルールに基づく手法と，発話内容や口調に対して機械学習手法を適用する方法の２つに大別できる．小説にはライトノベルと呼ばれるジャンルがあり，2010 年代からウェブ小説の書籍化により市場の多くを占めつつある．ライトノベルはキャラクターを重視する特徴から口調が多彩であり，会話を中心に物語が進行するため話者が明示されない場合も多い．これらのことから，ライトノベルを対象とした話者推定の先行研究において，会話文から話者ごとの口調の特徴を反映したベクトルを出力する口調エンコーダを用いることが提案されている．口調エンコーダとは，会話文を口調の違いによって分類できるよう追加学習したベクトル変換器である．先行研究では，エンティティに特化した事前学習済みモデルである LUKE に対し，一般的な小説データを利用した継続事前学習および口調タイプ分類をタスクとする追加学習を行うことで口調エンコーダを獲得している．しかし，得られたエンコーダを用いた推定モデルは，BERT を基礎とするモデルと比較して高い話者推定精度を示しているが，実用において必ずしも十分な精度が達成されているわけではない．また先行研究では，口調タイプの分類をマルチクラス分類問題として定式化しているが，小説においては登場人物の差異や個性を演出するために様式を組み合わせることが少なくなく，口調を排他的に分類することは大きな情報の損失につながると考えられる．これらのことを背景に，本研究では，ライトノベルにおける話者推定の精度向上を目的に，ライトノベルデータを用いた継続事前学習と，口調タイプのマルチラベル分類をタスクとする追加学習を行うことで，口調エンコーダを構築した．また実験を通じ，提案する同種データを用いた継続事前学習と口調のマルチラベル分類が，ライトノベルにおける話者推定の精度向上に貢献することを確認した．