ライトノベルからの情報抽出に向けた初期検討

井上　大輝

要旨

近年，ライトノベルは若者を中心に広く普及し，アニメや漫画化される作品も多く，ポップカルチャーの中で重要な位置を占めるようになった．しかし，ライトノベルの特有の言語スタイルや表現は，情報抽出において固有の課題を生じさせている．この分野の研究は，文学研究だけでなく，マーケティングや文化研究にも重要な意義を持つ．

本研究では，ライトノベルから情報を効果的に抽出するためのアプローチとして，テキスト中で重要な役割を果たす固有表現の抽出に着目する．ライトノベルにおける固有表現は，作品のキャラクターや設定の理解に不可欠であるが，文体の多様性や比喩的な表現，造語，方言などによりその抽出は容易ではない．この問題に対処するための第一歩として，本研究では，基本データセットの構築および固有表現抽出器の構築に取り組む．

基本データセットの構築に関しては，11 種の固有表現を設定すると共に，ライトノベル 9 冊を対象にアノテーションを行った．また固有表現抽出器の構築に関しては，前処理としてデータ拡張を適用すると共に，BERT モデルをファインチューニングする形で学習を行った．

評価実験では，必ずしも十分な精度での固有表現抽出は実現できなかったが，その過程において，適用したデータ拡張手法 5 種が必ずしも性能の向上に繋がらなかったことや固有表現の種類によって大きく精度が異なることなど，更なる精度向上に向けて有益な知見を得ることができた．