令和 5 年度 修士論文要旨

機械学習を用いた文体の自動変換についての検討

T22M065　安野 俊也
指導教員　三好 力　教授

1　はじめに

日本語は品詞や文法のみならず敬語など様々な要素によって構成されている。これらの要素は話者の状況や立場に合わせて使い分けなければならない。しかし、機械翻訳などの二言語間の変換を行う研究はされているが、常体から敬体など文体の変換はあまり研究されていない。

そこで、本研究では機械学習によって画像変換を行う手法である CycleGAN を応用し、文体変換を実現する手法について述べる。また、提案手法の精度を確認するために、実際に文体変換を行い、その結果を評価する。

2　提案手法

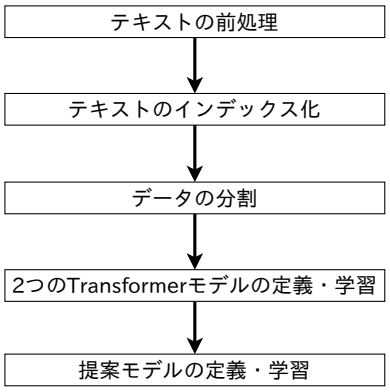

図 1　学習フェーズの概要

Transformer では対応する表現のペアが大量に必要であるが、これを作成することは大きなコストがかかる。そこで、画像変換を行う手法を参考に入力データのみでも学習を行う手法を提案する。

提案手法の学習方法を図 1 に示す。まず、学習に用いるデータセットを用意する。次に、データセットのテキストに対してクリーニングなどの前処理を行い、テキストをインデックス化する。その後、事前学習用・追加学習用・評価用・テスト用の 4 つにデータセットを分割する。まず、既存技術である Transformer のモデルの定義をし、変換および逆変換の事前学習を行う。その後、変換および逆変換を連続的に行う提案手法のモデルを定義し、事前学習済みのモデルを用いて追加学習を行う。

3　評価実験

精度を評価するために、事前学習を行った変換用の Transformer を用いた文体変換と提案手法を用いた文体変換を行い、その結果を比較する。データセットには丸山らのやさしい日本語コーパスを用いた。

事前学習データ数が 1000 個の時の実験の結果を表 1 に示す。その結果、提案手法の精度は既存技術である Transformer を用いた文体変換よりも精度が低いことがわかった。また、追加学習データの数と精度には相関関係がないことがわかった。

表 1　事前学習データが 1000 個の場合の実験結果

	BLEU スコア	増減率
Transformer	0.49046	-
追加データ (100)	0.44408	-9.45643%
追加データ (500)	0.40978	-16.44986%
追加データ (1000)	0.45380	-7.47462%
追加データ (5000)	0.44967	-8.31668%
追加データ (10000)	0.41657	-15.06545%

4　おわりに

本稿では、機械学習を用いた文体変換についての検討について述べた。今回の評価実験において、提案手法では既存技術の Transformer の精度を上げるには至らないことがわかった。今後は今回の結果を踏まえ、課題に対して変更を行うことで、精度を向上させていく予定である。