E-030

機械学習を用いた日本語文体の自動変換について

安野俊也† 三好 力††
†††龍谷大学大学院理工学研究科

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，本稿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，提案手法の学習に用いる損失の評価のための実験を行う．

2 関連研究

本節では，先行研究として機械翻訳，機械学習による画像変換および同一言語における変換について述べる．

2.1 機械翻訳

機械翻訳はコンピュータを用いてある言語から異なる言語に機械的に翻訳する手法である．機械翻訳は基本となる仕組みにより 3 つに分類される．一つ目はルールベース機械翻訳である．これは定められたルールに基づいて，蓄積された翻訳文や単語などのデータを用いて翻訳文を出力するシステムである．文法に則った構造ではない文章では正確な翻訳が期待できないという問題が存在する．二つ目は統計ベース機械翻訳である．蓄積された対訳を元に統計的なモデルを学習することにより翻訳を行うシステムである．問題点として原文の文法と訳する先の言語の文法構造が異なる場合にクオリティが下がってしまうということがある．三つ目はニューラル機械翻訳である．ニューラルネットワークを用いて学習を行うことにより，翻訳を行うシステムである．これまでの二つの機械翻訳と比べて高精度であるだけではなく，表現の違いに強いという利点が存在する．本研究では，このうちの一つである Seq2Seq[1] を用いる．

2.2 機械学習による画像変換

CycleGAN[2] は主に画像のスタイル変換で使用される GAN(敵対的生成ネットワーク) の一種である．これは 2 枚の画像の関係を学習するのではなく，2 つの画像データセットのドメイン関係を学習し画像変換を実現する手法である．

2.3 同一言語における変換

同一言語における変換に関する従来研究として，下地ら [3] は LSTM-RNN を用いた文体の変換手法を提案した．また，林ら [4] は書き言葉から話し言葉に変換する手法を提案した．

3 提案手法

Seq2Seq では対応する表現のペアを学習データで大量に必要とするが，これを作成することは大きなコストがかかる．そこで画像変換の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．

提案手法による学習を図 1 に示す．また，本章では提案手法について解説する．

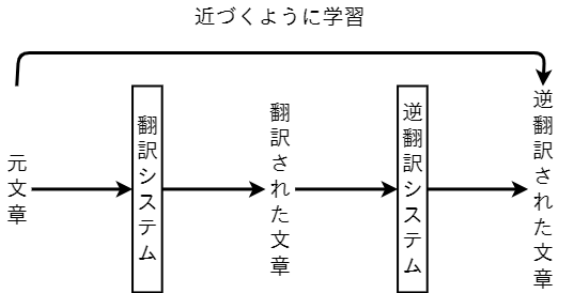

図 1 学習手法

3.1 学習フェーズ

本節では，文章を学習し，モデルを生成するフェーズについて解説する．

まず，学習データには対応させた文体の異なる 2 種類のテキストデータを収集し，それぞれのテキストデータの文章を文ごとに分割する．そして，学習に用いる文章に対して Mecab を用いて形態素解析をし，分かち書きを行う．その後，出現単語をまとめた単語辞書を

Automatic Style Conversion of Japanese Sentence Using Machine Learning
Toshiya Yasuno† and Tsutomu Miyoshi††
†††Graduate School of Science and Technology, Ryukoku University

作成する．これにより，文章を単語ベクトルとして表現する．

さらに，機械学習を用いて元の文体の文章がもう一方の目的とする文体の文章になるように変換システムの学習を行う．それと並行して，先ほどとは反対に目的とする文体の文章が元の文体の文章になるように逆変換システムの学習を行う．その後，変換および逆変換を連続して行い，学習モデルを生成する．

本稿では，その際の損失は 2 つの Seq2Seq の損失を統合した損失を開発して用いた．

連続学習時の学習データは対応の必要がないため，多数の新規データが利用可能である．機械翻訳において翻訳前後の意味の同一性が担保されていないことに着目し，本研究では機械学習の学習時に意味の同一性を保証するために，変換および逆変換を連続で行うモデルとした．

3.2　変換フェーズ

本節では，3.1 節で生成したモデルにより入力文の文体変換を行うフェーズについて解説する．

3.1 節と同様にして，入力文に対して形態素解析を行い，分かち書きを行う．そして，ベクトル化を行い，3.1 節で生成した学習モデルに入力する．その後，出力されたベクトルを単語辞書を用いて文章の形に逆変換することで，目的とする文体に変換された文章を得る．

4　評価実験

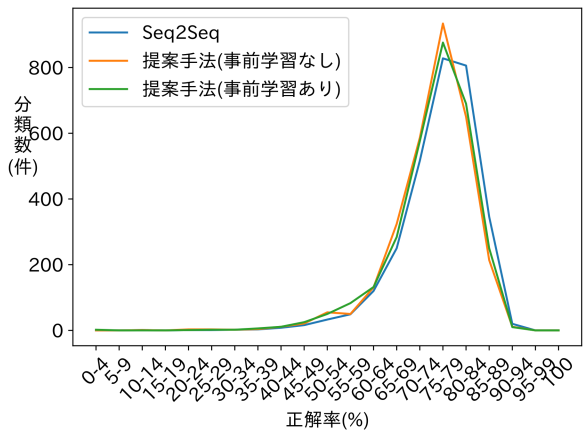

図 2　正答率の分布

提案手法の損失を確認するために評価実験を行った．データには, 勝田らのやさしい日本語コーパス [5] を学習データにおいて 6000 文，テストデータにおいて 3,000 文使用した．本実験では，損失関数には交差エントロピー関数を用いた．評価は出力文と正解文の特定の位置の単語が同じである割合である単語適合率で行った．

評価実験の結果である正答率の分布を図 2 に示す．正解率が 70%後半の結果が既存技術の Seq2Seq が提案手法の損失と比較して分類数が多い結果となった．しかしながら，それ以外の正解率については大きな差異が見られず，近しい結果となった．よって，これは学習データの不足が原因であると考えられるため，十分な量のデータが用意されれば解決されると考える．

したがって，提案手法の損失は学習用の損失として扱っても問題がないと考えられる．

5　おわりに

本稿では，機械学習を用いた文体変換する手法の提案を行った．3000 文に対して，提案手法の損失に関する精度評価を行った．その結果，既存技術である Seq2Seq と比較しても遜色ない結果となった，

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．

参考文献

[1] Sutskever, I., Vinyals, O. and Le, Q. V.: Sequence to sequence learning with neural networks, *Advances in neural information processing systems*, Vol. 27 (2014).

[2] Zhu, J.-Y., Park, T., Isola, P. and Efros, A. A.: Unpaired image-to-image translation using cycle-consistent adversarial networks, *Proceedings of the IEEE international conference on computer vision*, pp. 2223–2232 (2017).

[3] 下地健太，森田和宏，泓田正雄：LSTM-RNN を用いた文体変換手法，人工知能学会全国大会論文集第 33 回 (2019)，一般社団法人 人工知能学会，pp. 2L4J901–2L4J901 (2019).

[4] 林由紀子，松原茂樹：ニュース記事の自然な音声出力のためのテキスト変換，言語処理学会第 14 回年次大会発表論文集，pp. 790–793 (2008).

[5] Katsuta, A. and Yamamoto, K.: Crowdsourced corpus of sentence simplification with core vocabulary, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)* (2018).