

逆翻訳における元の対訳コーパスおよび外部の単言語データの有効性の比較 ～アイヌ語と日本語の機械翻訳に着目して～

菅原 葵

本研究は、低資源言語であるアイヌ語と日本語の機械翻訳の精度を高めることを目的とした。研究手法として、はじめに、NLLB-200というマルチリンガル事前学習済みモデルを利用し、アイヌ語と日本語の機械翻訳モデルを構築した。次に、元の対訳コーパスのみを用いて、反復的逆翻訳とランダムサンプリングによる疑似対訳文生成の手法を利用し、翻訳精度の向上を目指した。さらに、外部の日本語の単言語データを利用して、逆翻訳を行い、疑似対訳文を生成することで、さらなる精度向上に繋がるのではないかと考え、実験を行った。

結果として、逆翻訳の有効性は確認されたが、反復的逆翻訳に関しては、はっきりした精度の上昇傾向が見られなかった。また、ランダムサンプリングによる疑似対訳文生成に関しては、サンプリングによって疑似対訳データの多様性をある程度増大させる方法の有効性が確認されたが、生成する疑似対訳文の数が多すぎる場合には、逆に精度の向上度合いが低下する傾向が見られた。さらに、外部の日本語の単言語データを利用した実験では、逆翻訳の回数が増えるほど精度が低下する傾向が見られ、外部の日本語データを利用するよりも、元の対訳コーパスのみを用いた場合の方が精度向上に効果的であることが確認された。