

梶原 智之

大阪大学データバリエティフロンティア機構
特任助教

自然言語処理の真価を引き出す言い換え生成

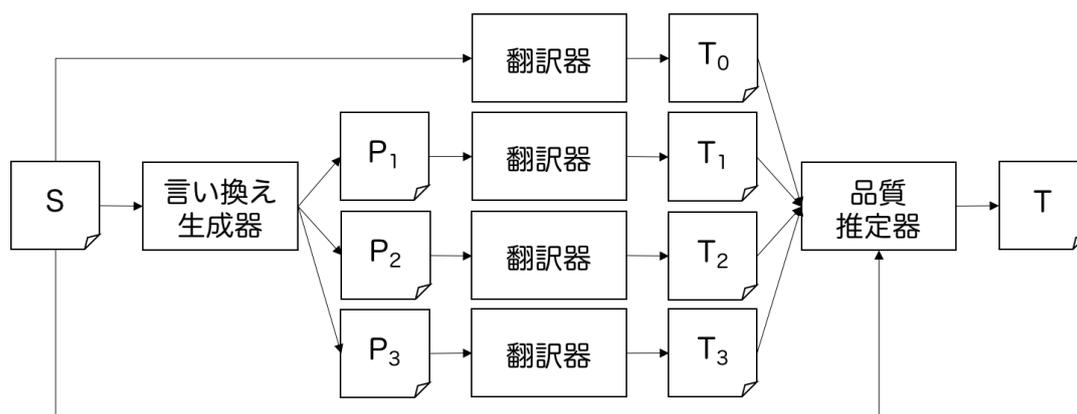
§ 1. 研究成果の概要

東京五輪や大阪万博に向けて、機械翻訳の精度向上が急務である。本研究では、機械翻訳を中心に、自然言語処理における入力感度の問題に取り組む。入力文の意味が同じでも、例①のようなモデルにとって得意な表現の翻訳は成功し、例②のような苦手な表現の翻訳は失敗する。

① 東大を受験します。 → I will take an exam of the University of Tokyo.

② 東大を受けます。 → I receive the University of Tokyo.

②の入力文を予め①へ言い換えることによって、所与の自然言語処理モデルを用いてより高品質な出力を得ることができる。本研究では、このような前編集技術を実現する。



2019 年度は、上図のような流れで機械翻訳の前編集に取り組んだ。入力文 S について数種類の言い換え P_i を生成し、それぞれを翻訳した。これらの翻訳文 T_i と入力文 S を用いて翻訳品質を推定し、最も高品質な翻訳を利用する。

機械翻訳の自動評価では、予め用意した正解文と翻訳文を比較する BLEU という指標が一般的

に利用される。そこで事前調査として、入力文と翻訳文から翻訳品質を推定する代わりに、翻訳文と正解文を比較する BLEU によって最良の翻訳を選択した。なお、言い換えは 100 種類生成した。

WMT-2017 のコンペで利用された英語からドイツ語への翻訳タスクにおける実験の結果、コンペ優勝モデルの BLEU=27.05 を前編集によって 35.08 と大幅に改善できた。また、標準的な設定の機械翻訳モデルの性能は BLEU=18.60 であるが、提案手法によってこれは 28.74 に改善でき、コンペ優勝モデルの性能を超えることを確認した。今後は正解文を用いずに翻訳品質を推定する。