相互学習に基づく機械翻訳の半教師あり学習の研究

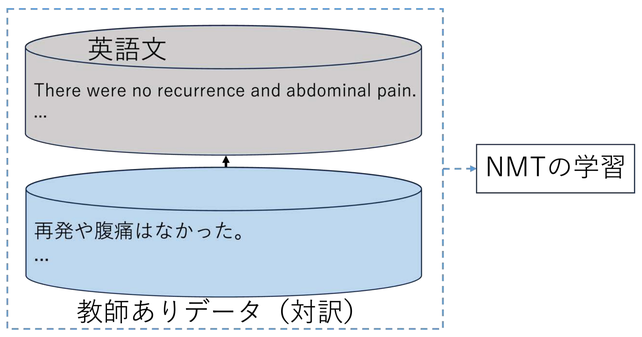
造形学部 スマートデザイン学科　塚田　元

概要

- 近年の機械翻訳技術は、大量の翻訳の対（対訳データ）から翻訳の仕方（翻訳モデル）をニューラルネットで自動的に学習
- 翻訳モデルの学習の際、対訳データ（教師ありデータ）だけでなく単言語データ（教師なしデータ）を活用する半教師あり学習法を提案
- 提案法は入出力両言語の単言語データを活用できることが特徴で、双方向の翻訳機が反復的に教え合うことで実現
- 種々の言語対で提案法の効果を確認するとともに、対訳データには存在しない単語レベルの対応（対訳語彙）も自動獲得されることを確認

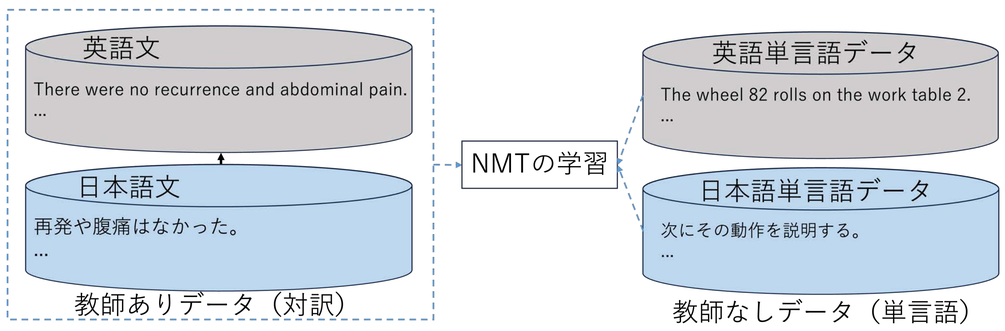
ニューラル機械翻訳（NMT）の問題点

NMTの学習には大量の対訳データが必要だが、低資源言語ではその収集・構築は困難



NMTの半教師あり学習

- 対訳になっていない単言語コーパスは比較的収集・構築が容易
- 対訳コーパスに、原言語および目的言語の単言語コーパスを併用して、NMTの精度を改善

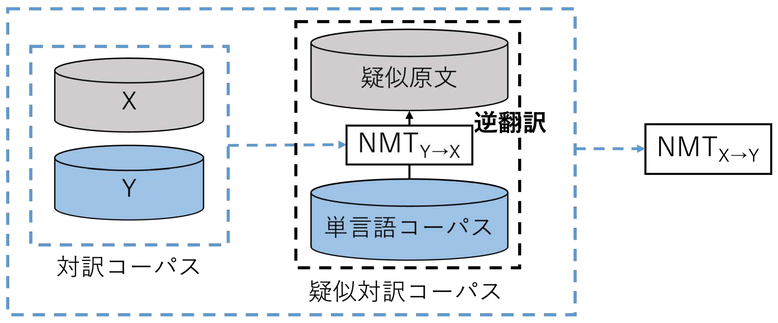


相互学習のイメージ

英語を学びたい日本語ネイティブと日本語を学びたい英語ネイティブが相互に教え合うことで、両者の言語能力の向上が可能

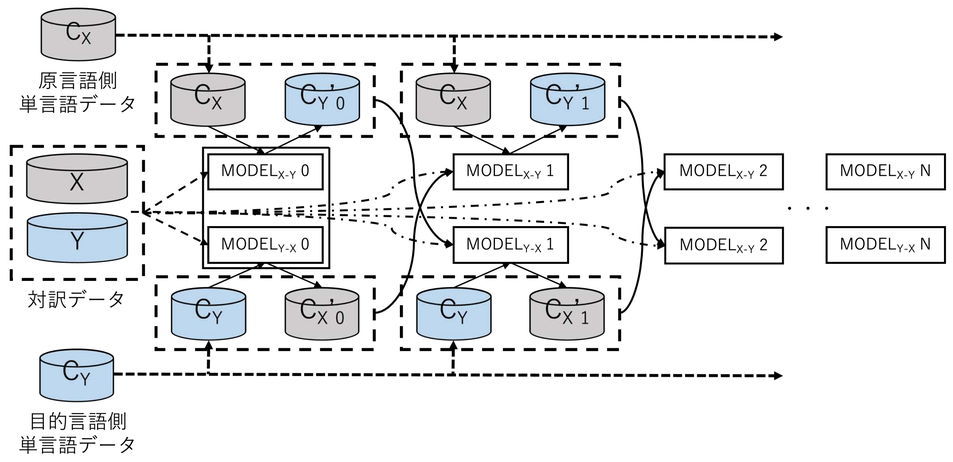
先行研究（逆翻訳）

- 目的言語側の単言語コーパスを用いた半教師あり学習（Sennrich+ 2015）
- 目的言語側の単言語コーパスを逆翻訳して疑似対訳コーパスを生成し，対訳コーパスと組み合わせて学習

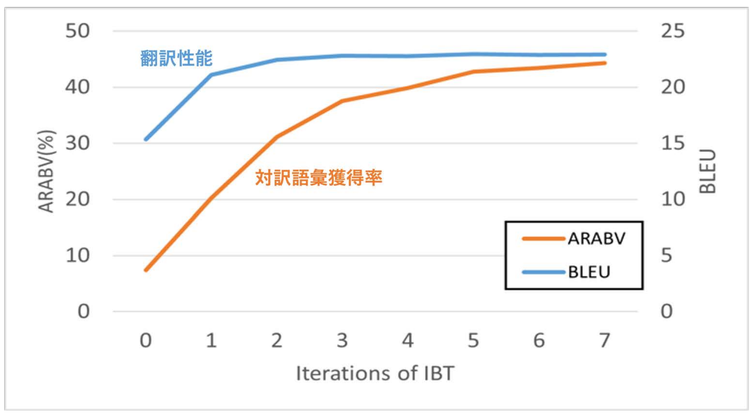


IBT: Iterative Back-Translation【提案法】

Sennrichら（2015）の手法を双方向に反復適用



英独翻訳の対訳語彙獲得率と翻訳性能



対訳語彙獲得の過程

入力	a sacrificial **anode** was successfully developed afterwards, and the safety was raised .
正解訳	その後犠牲**陽極**の開発に成功し，安全性を高めた。
翻訳結果（反復 0）	後から犠牲が出て安全になった。
翻訳結果（反復 1）	犠牲的分子はその後成長し，安全性を高めた。
翻訳結果（反復 2）	その後犠牲**陽極**が開発され，安全性が向上した。
翻訳結果（反復 6）	その後犠牲**陽極**の開発に成功し安全性が高められた

応用分野：データからのテキスト生成（D2T: Data-to-Text Generation）

- 現代の機械翻訳技術は汎用的な系列変換モデルに基づくものであり、入出力は単語列に縛られない
- センサーデータ、生理データ、金融データ、人間の行動データなど様々なデータを、「言語化」によって可視化する技術（D2T）にも応用可

主な研究成果（豊橋技科大との共同研究）

- Taisei Sone, Tomoyoshi Akiba and Hajime Tsukada, “Incorporating Curriculum Learning into Iterative Back-Translation for Neural Machine Translation,” ICAICTA 2025, 2025.
- Takuma Tanigawa, Tomoyosi Akiba, Hajime Tsukada, “Analysis on Unsupervised Acquisition Process of Bilingual Vocabulary through Iterative Back-Translation,” LREC-COLING 2024, 2024.
- 紺谷 優志, 秋葉 友良, 塚田 元, 双方向翻訳モデルの相互学習におけるデータ多様化の適用, 言語処理学会第30回年次大会, 2024.
- 森田 知熙, 秋葉 友良, 塚田 元, ニューラル機械翻訳の反復的逆翻訳に基づくデータ拡張のための混成サンプリング手法, 電子情報通信学会論文誌D 情報・システム J106-D(4) 298-306, 2023.
- 森田 知熙, 秋葉 友良, 塚田 元, 双方向の逆翻訳を利用したニューラル機械翻訳の教師なし適応の検討, 第5回自然言語処理シンポジウム, 2018.

ASU AICHI SANGYO UNIVERSITY　問い合わせ先：tsukada@asu.ac.jp

カラーコード　青ベースカラー：＃235BA8　研究題目・副題目・所属・氏名文字：＃FFFFFF　本文・見出し文字：＃3C3D3C