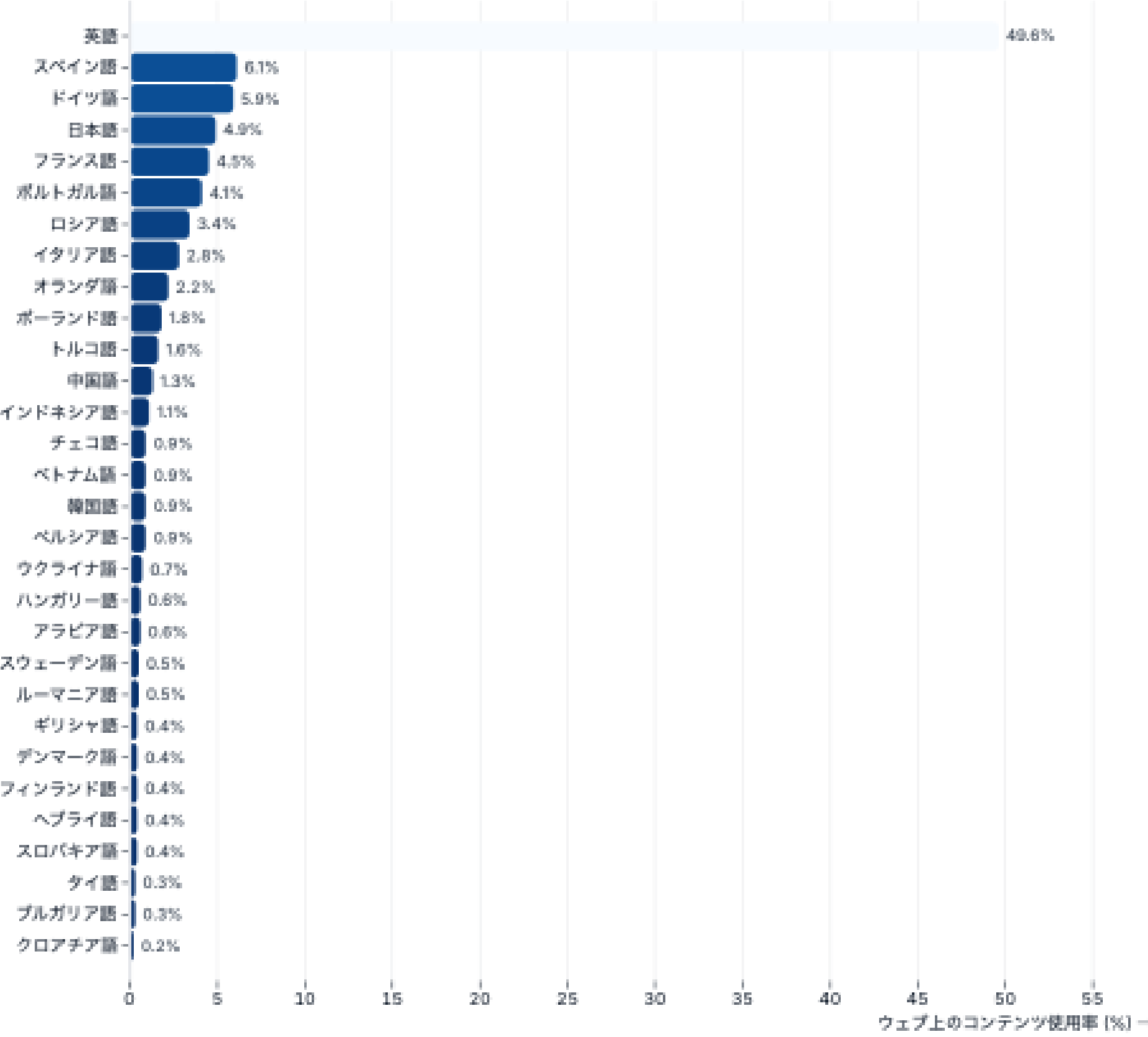


低資源言語の機械翻訳に向けて

愛知産業大学 情報学部 知能情報学科（設置認可申請中） 塚田 元

世界のテキストデータの大部分は英語

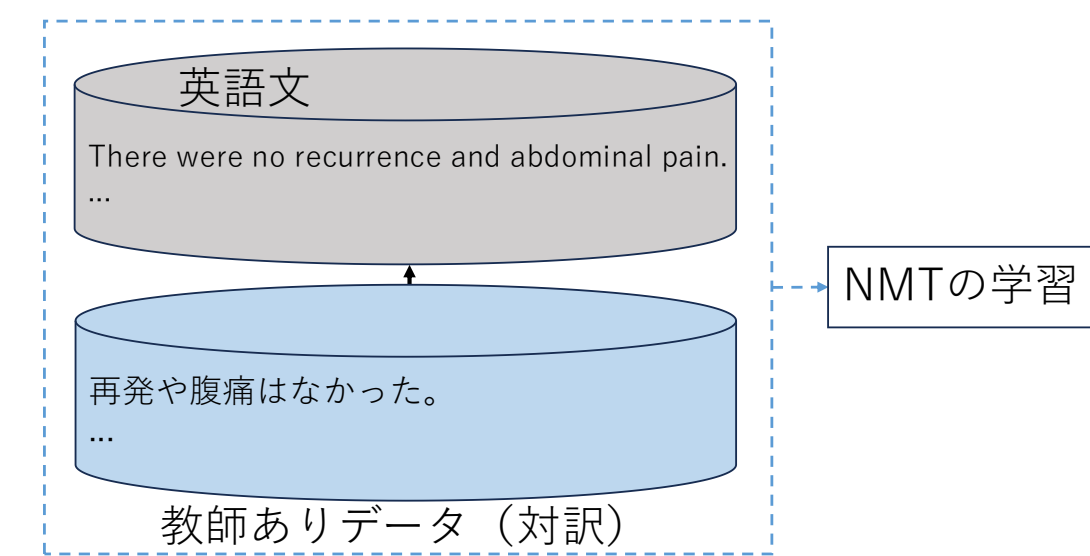
世界には数千もの言語がある。しかし、大量のテキストデータが存在する言語は限られている。



ウェブ上のテキストデータ量ランキング W3Techs (Web Technology Surveys)より

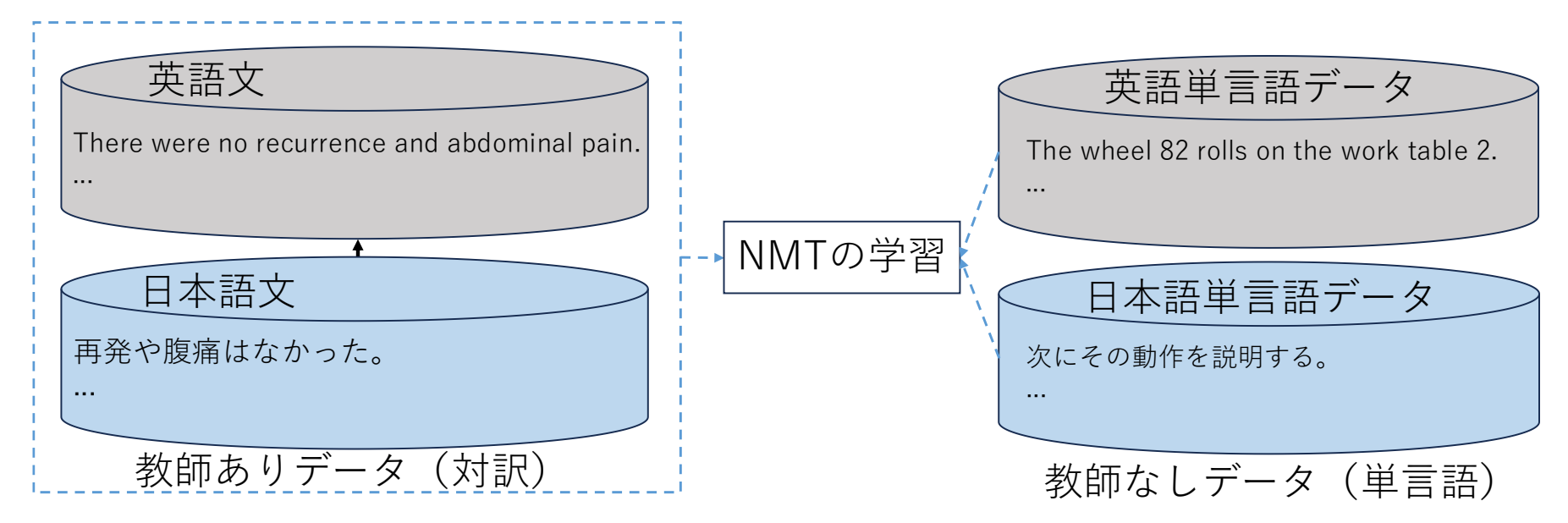
ニューラル機械翻訳（NMT）の問題点

- 近年の機械翻訳は、機械学習型の技術であるニューラル機械翻訳(NMT)によって実現
- NMTの学習には大量の対訳データが必要。しかし、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データであれば、比較的収集・構築は容易
- 半教師あり学習とは、教師ありデータ(対訳)に教師なしデータ(原言語および目的言語の単言語データ)を併用する学習法
- 低資源言語のNMT実現に向けて、半教師あり学習法を研究中



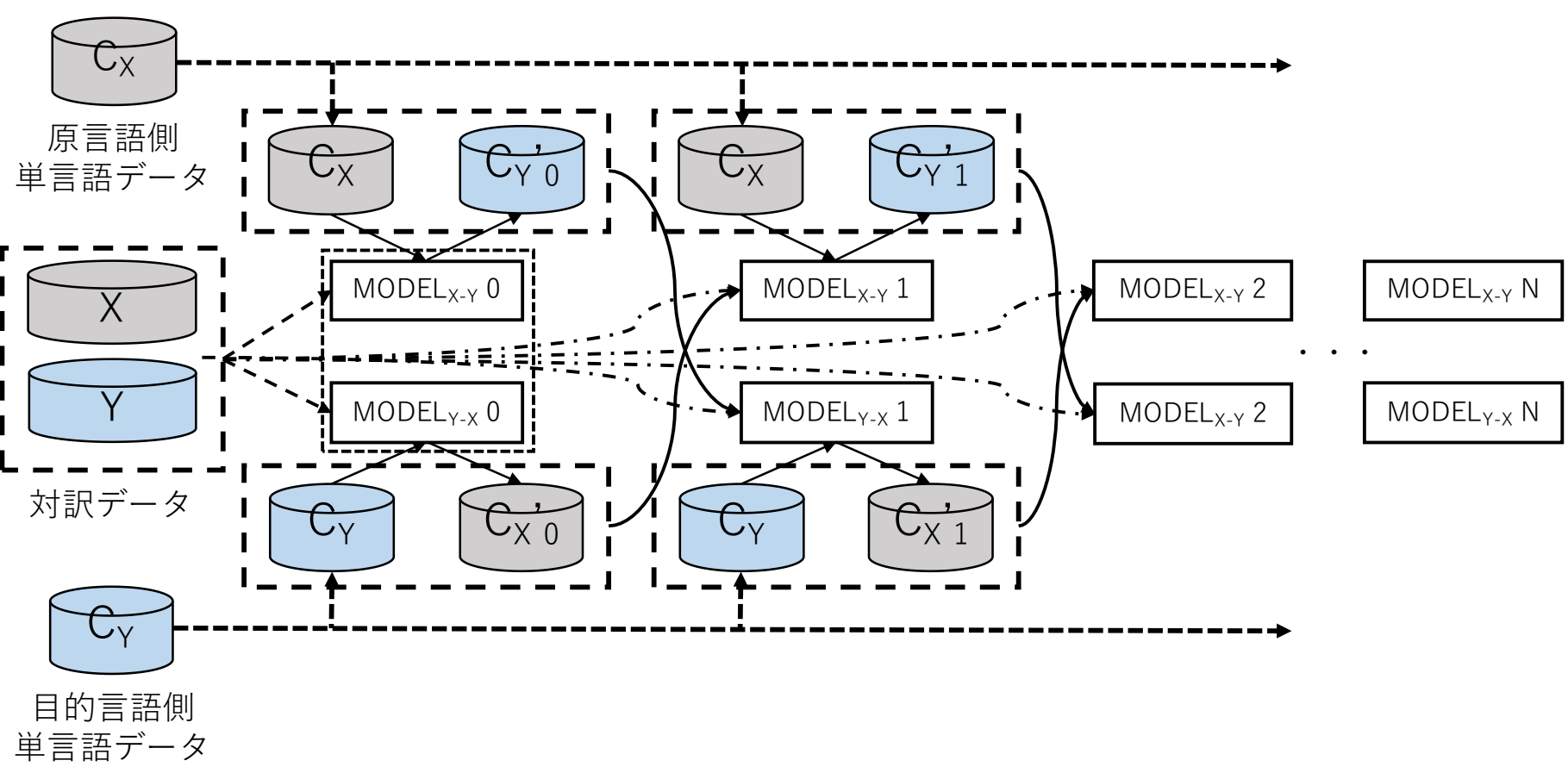
半教師あり学習を実現する「相互学習」

- 「相互学習」によって、半教師あり学習を実現
- 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、両者の言語能力を向上

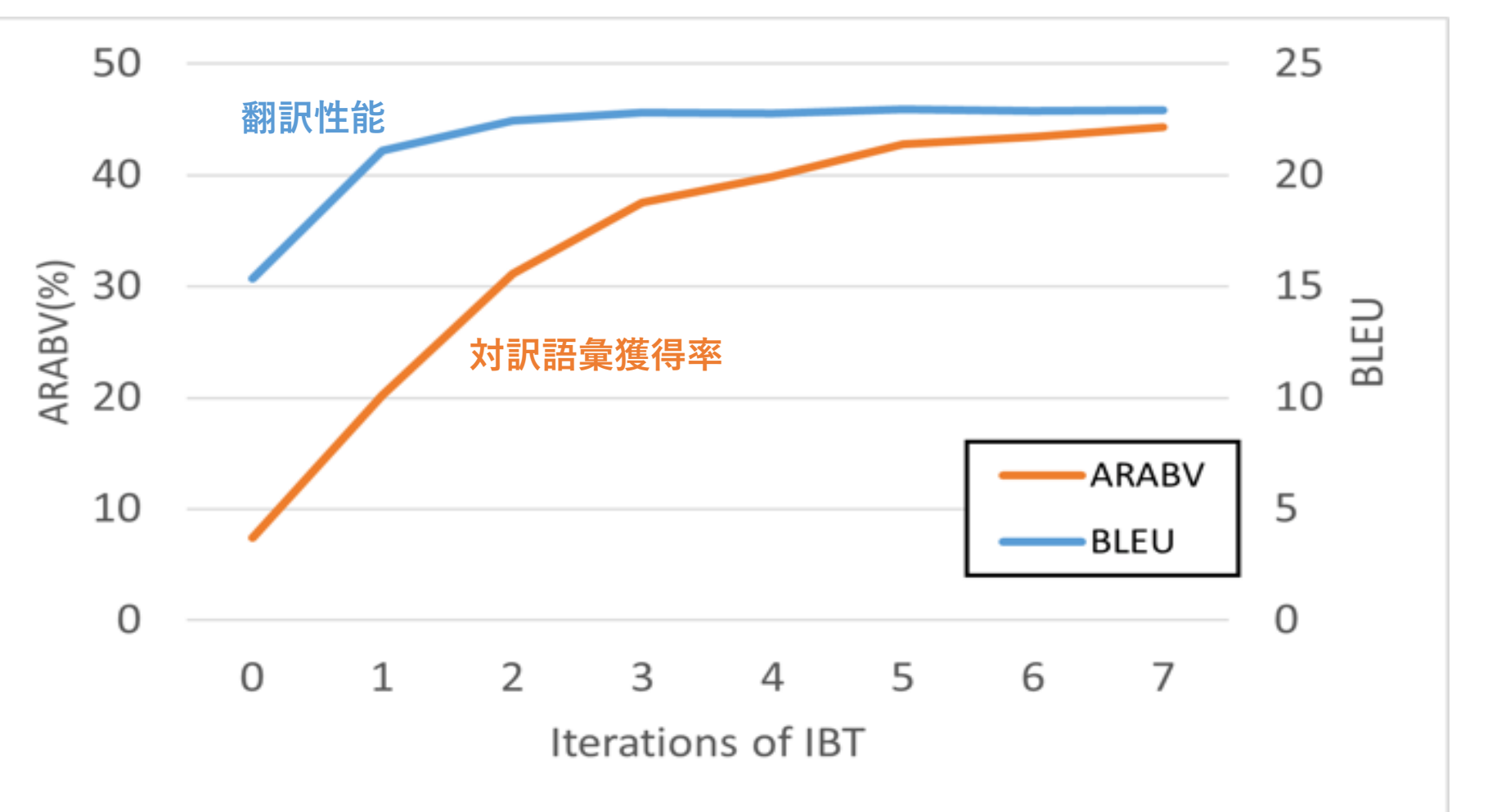


IBT: Iterative Back-Translation【提案法】

- 逆翻訳で疑似対訳データを作成し学習に活用
- それを双方向に繰り返す



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

参考文献

- Takuma Tanigawa, Tomoyosi Akiba, Hajime Tsukada, “Analysis on Unsupervised Acquisition Process of Bilingual Vocabulary through Iterative Back-Translation,” LREC-COLING 2024, 2024.