知能情報学科
（設置構想中）

LLMを活用したタスク特化型ニューラル機械翻訳の研究

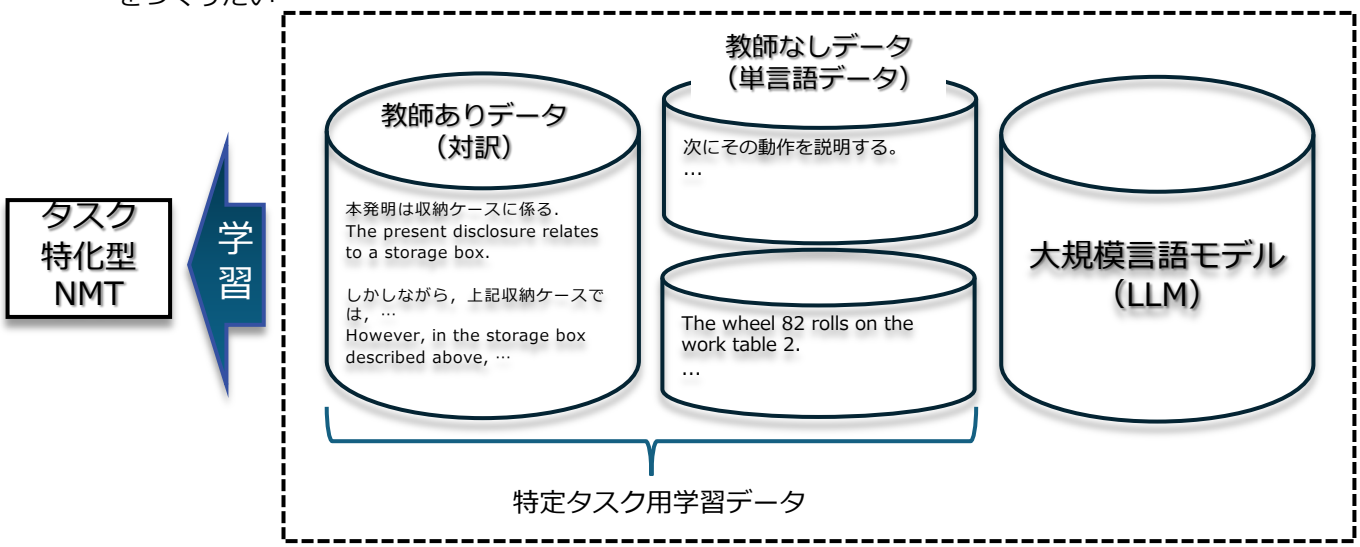
愛知産業大学・教授　塚田 元

研究の背景と目的

- 教師ありデータだけでなく教師なしデータを活用する半教師あり学習に基づくニューラル機械翻訳(NMT)[1][2][3]を研究してきた
- 学習データ以外に、大規模言語モデル(LLM)が利用可能になりつつある

▼

- LLMに少量の特定タスク用学習データを併用して、タスク特化型の翻訳機（マイナー言語用など）をつくりたい



大規模言語モデル（LLM）とは

- 過去に出てきた単語列から次の単語を予測する統計モデル
- 大規模なテキストデータ（教師なしデータ）で事前学習されている
  （新聞記事数十万年分相当量のデータを使い，GPUだけで約1千億円もかかる計算機環境で
  数か月の期間をかけて学習）
- タスクに応じた学習データを用意しなくても，汎用モデル一つで様々な自然言語処理タスクにおいて最先端の性能を達成
  - 質問応答（ChatGPT，Geminiなど），機械翻訳，文書分類，情報抽出など

▼

- パラメータ数を抑えた小規模のオープンウェイトモデル（ローカルLMM）が公開されてきており，<u>現実的な計算機環境でもLLMの推論や追加学習が可能になりつつある</u>

今後の研究計画

マイナー言語の機械翻訳

・言語資源の乏しい言語の機械翻訳をLLMの追加学習アプローチで検討

特徴量の言語化

・健康診断結果の自動所見作成のように，特徴量を言語化するタスクへの適用を検討

[1] T. Tanigawa, T. Akiba, H. Tsukada, Analysis on Unsupervised Acquisition Process of Bilingual Vocabulary through Iterative Back-Translation, The 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), 2024.
[2] 紺谷, 秋葉, 塚田，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年3月12日
[3] 森田, 秋葉, 塚田，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年4月1日

問い合わせ先：tsukada@asu.ac.jp