2025年03月15日 言語処理学会第31回年次大会 併設ワークショップ
日本語言語資源の構築と利用性の向上 (JLR2025)

複数の作業工程を経た 翻訳コンテンツの学習における課題

NHK 放送技術研究所
美野 秀弥，河合 吉彦，山田 一郎

背景と目的

■ 背景

- ● 外国人に対する情報発信の強化
  - ✓ やさしいことばニュース (2024年9月～ NHKラジオ第1（R1）)
    - 「日本に住んでいる外国人の皆さんや，子どもたちに，できるだけやさしい日本語でニュースを伝えます．」 (https://www3.nhk.or.jp/news/easy/)
    - NHKがラジオニュースで始めた"お堅く"ない "やさしいことば"に込めた思いとは
      - https://www.nhk.or.jp/info-blog/674856.html
  - ✓ NEWS WEB EASY (～2024年9月 NHK NEWS WEB)

■ 目的

- ● 情報発信の効率化の検討
  - ✓ やさしい日本語の機械翻訳器の構築

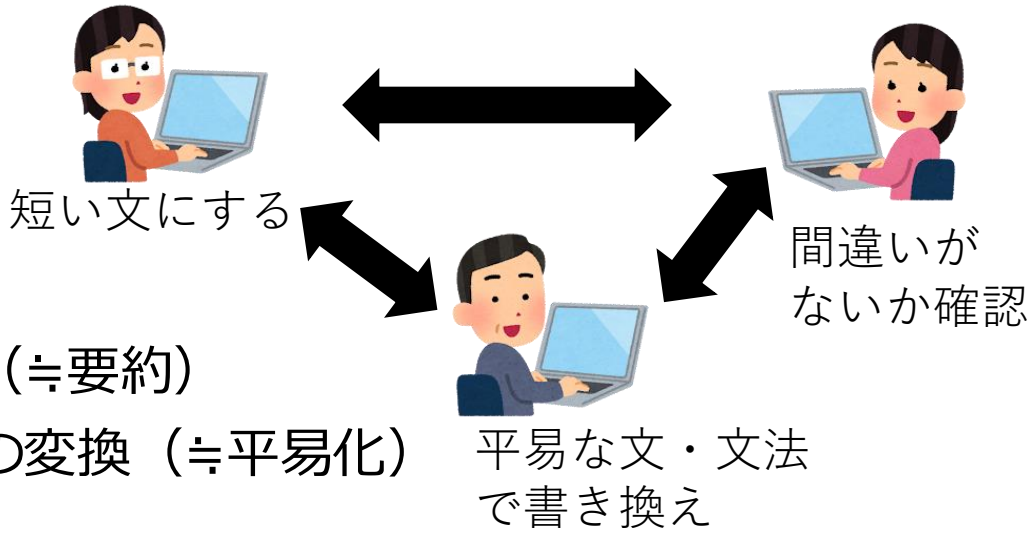
やさしい日本語機械翻訳器構築の課題

課題１：低リソース

- NEWS WEB EASYが年間で制作していたニュースは約950本
  - ✓ 平日のみ1日4本程度
- 高品質な機械翻訳器の構築には100万文程度の対訳データが必要

課題２：作業工程の複雑さ

- 複数の観点での作業が必要
  - ✓ 内容の正確性と伝えるべき内容の取捨選択（≒要約）
  - ✓ 外国人や小学生が理解できる語彙・文法への変換（≒平易化）
  - ✓ 内容の適切さなどを総合的に判断（≒校閲）
- タスクの複雑さは精度に直結する



課題解決へのアプローチ

課題１：低リソース → 大規模言語モデル (LLM)の追加学習

- LLMの追加学習は少ない学習データでも精度向上が報告
  - ✓ ただし，得意なタスク，不得意なタスクがある
    - 得意なタスク例：質問応答，翻訳，要約，対話，ソースコード生成等

課題２：作業工程の複雑さ → 作業工程を複数のタスクで再定義

- やさしい日本語翻訳作業を「1.要約作業→2.平易化作業→3.校閲作業」と定義

提案手法：再定義したタスクのデータによるLLMの追加学習

提案手法：工程ごとのデータを用いたLLMの追加学習

■ 学習手順

① 定義した工程ごとの対訳データを抽出

（一般ニュース文，やさしい日本語文）	← 翻訳データ対
＋（一般ニュース文，記者OBの作業結果）	← 要約データ対
＋（記者OBの作業結果，日本語教師の作業結果）	← 平易化データ対
＋（途中結果，報道局の作業結果＝やさしい日本語文）	← 校閲データ対

＊ システムのログからデータ対は抽出

② ①のデータに工程を表すメタ情報をタグとして付与してLLMを追加学習

■ 推論手順

① 要約 → 平易化 → 校閲の順に3回推論してやさしい日本語文を生成

翻訳実験

- 実験データ
  - 学習データ：NEWS WEB EASY (2017/04 - 2021/08)　48万文対
    - ✓ やさしい日本語翻訳：10万文対，平易化：13万文対，要約：16万文対，校閲：９万文対
  - テストデータ：NEWS WEB EASY (2021/09)　779文対

- 大規模言語モデル（LLM）
  - Llama-3.1-8B，3-2-3B，Llama-3.1-Swallow-8B (decoder-only transformer)
    - ✓ Llama：Meta社が商用利用可で公開しているLLM（学習データの大半は英語）
    - ✓ Swallow：東科大がLlamaをベースに日本語の能力を強化したLLM

- 追加学習手法
  - PEFT LoRA
    - ✓ パラメータの少ない別のモデルを用意して追加学習する手法

- 比較手法
  - やさしい日本語翻訳の対訳データ（10万文対）のみで学習したモデル

実験結果

- 提案手法の精度が高い
- パラメータ数が小さくても後発LLMのほうが精度が高い
- 日本語特化型LLMのほうが精度が高い

	翻訳精度: BLEU		
	3.1-8B	3.2-3B	Swallow-8B
比較手法	10.0	12.9	14.0
提案手法	11.4	13.3	15.9

まとめ

■ やさしい日本語機械翻訳器の試作

- LLMの追加学習手法を用いたやさしい日本語翻訳器を試作
  - ✓ 少ない学習データでも翻訳精度が向上
  - ✓ 翻訳過程のデータを活用することで翻訳精度が向上
- 今後，定義した工程の適切な推論手順を検討

■ 作業過程のデータの重要性

- ログデータも重要な学習データとなりうる