S5-P39

機械翻訳タスクにおける日本語文書単位容認性判断データセット構築と途中補完モデル

松田 陵佑[1] 工藤 慧音[1,2] 藤井 諒[1,3] 伊藤 拓海[1,4] 森下 睦[1,3] 鈴木 潤[1,2,5]

([1]東北大学, [2]理化学研究所, [3]フューチャー株式会社, [4]マシンラーニング・ソリューションズ株式会社, [5]NII LLMC)

概要

- WMT2026 General MT Task の英日・中日方向に参加
- 文書文脈での日本語の自然さを人手評価した容認性データセット **CAJALA** を構築
- CAJALAと対訳コーパスで**途中補完翻訳**を学習
  推論は **MBRデコーディング → 途中補完翻訳**

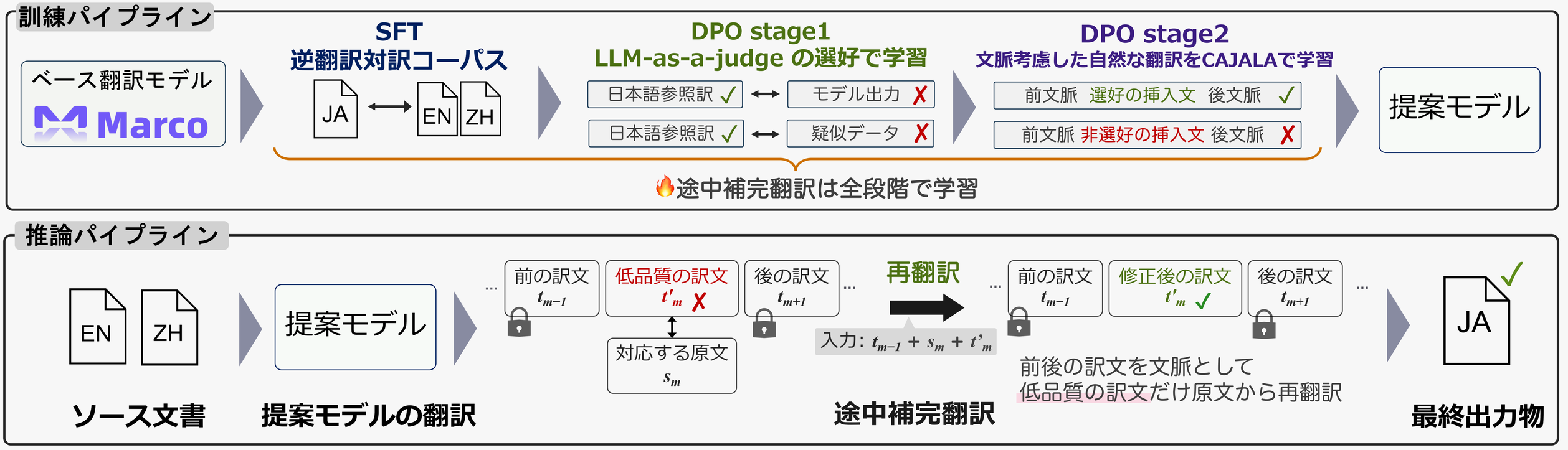
性能向上:ベース比 英日 **+8.3** / 中日 **+22.6 points**

背景・課題

- 文単位で自然でも文書中では**文体・用語**が不整合に
- 自然さは自動評価では捉えにくく**人手選好**が必要
  → 文書文脈つきの**容認性**を考慮した学習をしたい
- 翻訳後に残る局所的な誤りを**翻訳後に修正**したい

CAJALA の構築 と **途中補完翻訳** を組み合わせる

提案システム

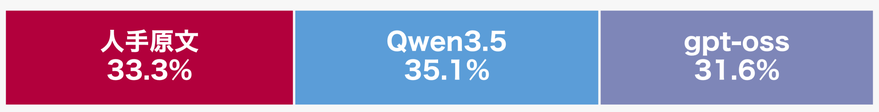


CAJALA データセット構築

7,349 サンプル / 計**2,000** ワーカー

前文脈 …
候補A：人間が書いた原文
候補B：Qwen3.5-397B の言い換え文 ✓
候補C：gpt-oss-120b の言い換え文
… 後文脈

候補別の選択率 (3者一致率 85.9%)



人手原文が選ばれたのは **33.3% のみ** → **LLM の言い換えも好まれる**

実験設定

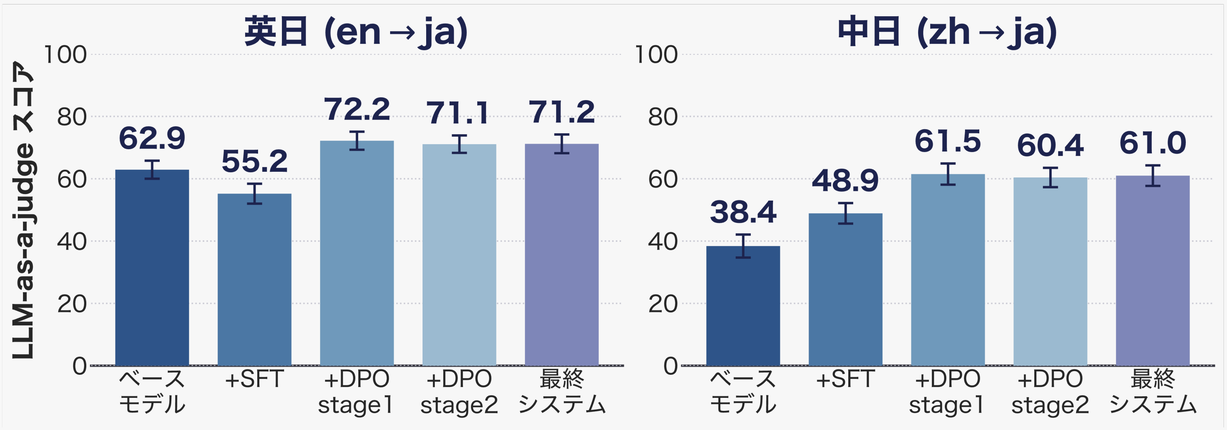
学習/推論設定

- ベースモデル: **Marco-MT-Algharb** [Wang+'2025]
- **SFT**: 日本語文書を逆翻訳した対訳コーパスで学習
- **DPO**: 2段階に分けた選好学習
- **推論**: MBR デコーディング(XCOMET-lite) + 途中補完翻訳

評価

評価セット: WMT2026 blind set
評価方法: LLM-as-a-judge（gemini-3.1-pro）0~100 点

実験結果



LLM-as-a-judge (gemini-3.1-pro) の平均スコア（±95% 信頼区間）

DPO stage1 で大きな改善 (英日 **+17.0** / 中日**+12.6**)
→ 未翻訳のソース言語が残る問題を解消
途中補完翻訳が発生した英日 **4 件全てでスコア上昇↑**

定性分析

DPO stage2(CAJALAによる学習)→文体の不整合が改善
(英日, 45 → 85 点)

原文: Okay. And you're going uphill. I mean, this is really good stuff here. I know that you're probably feeling it a bit.
stage1: はい。そして上りですね。いや、これ、めっちゃいいもんやで。ちょっと気持ち入ってるんやろうな、わかるけど。
stage2: はい。そして上りですね。いや、これホントいいもんなんですよ。ちょっと感じてるんだろうな、ってのは分かってるんですけど。

途中補完による改善例 →未翻訳の箇所を修正
(英日, 33 → 50 点)

原文: *they send an email saying "thanks for sending through the design"*
編集前: *they send an email saying "thanks for sending through the design"*
編集後: 「デザインを送って下さってありがとうございます」というメールを送る