Targeted Source Text Editing for Machine Translation: Exploiting Quality Estimators and Large Language Models

Hyuga Koretaka (Ehime U.) Atsushi Fujita (NICT) Tomoyuki Kajiwara (Ehime U.)

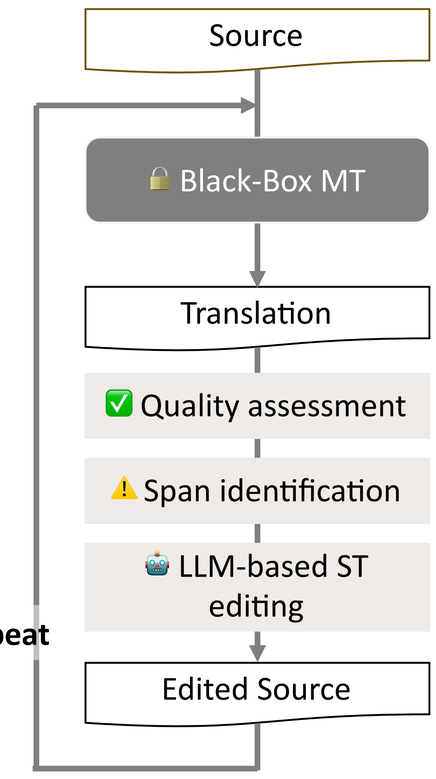
1. Source Text Editing for MT

- Editing source texts so that they can be better translated by MT
  - An effective way for exploiting black-box MT systems
  - Orthogonal to other strategies
    - e.g., system combination, automatic post-editing
- Findings in manual investigations of "targeted source text editing" (Miyata & Fujita, 2017, 2021)
  - Potential improvement: 88%-100% segments were eventually translated w/o any error.
  - Diverse linguistic operations: 50+ types ranging from syntactic/phrasal alternations to symbol substitutions/normalizations.
- Existing automatic "pre-editing" methods are
  - performing limited types of edit operations
  - with no reference to actual translation errors

} We addressed these issues !

2. Proposed Method

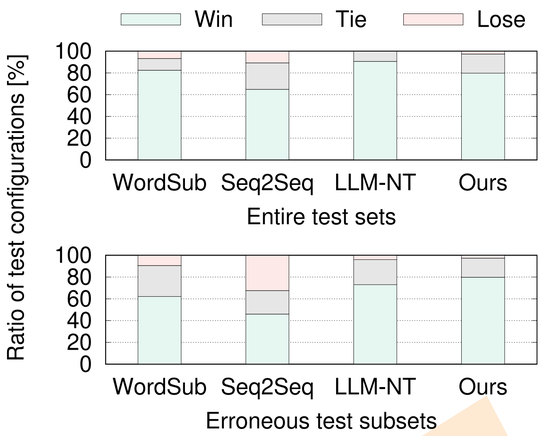
- Our method attempts to reduce translation errors caused by a given translation system (*T*), relying on
  - Segment-level quality estimator (`XCOMET-XL`, *Q*) to ✅ search for the best translation (as in Ki & Carpuat, 2025)
  - Span-level quality estimator (`XCOMET-XL`, *E*) and word aligner (`OTAlign`, *A*) to ⚠️ determine the source span to be edited
  - Large language model (LLM, *P*) to 🤖 flexibly and accurately edit the identified source text span



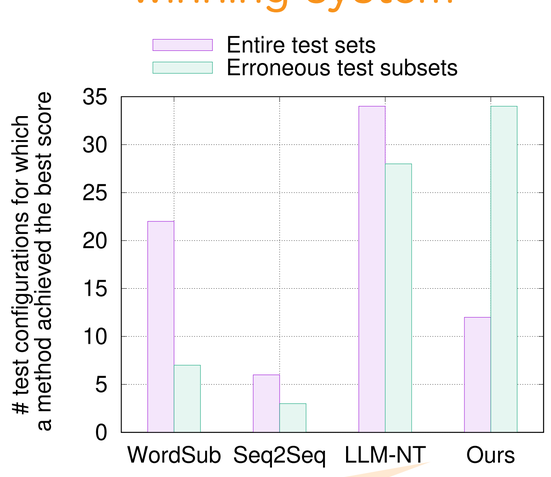
3. Evaluation & Results

- 74 test configurations
  - 8 MT systems
    - 4 NMT: `NLLB-3.3B` (NLLB Team, 2022), 3 variants pre-trained on `JParaCrawl` (Morishita et al., 2022)
    - 4 LLM-based: `Llama-3.1-70B-Instruct` (Grattafiori et al., 2024), `Llama-3.1-Swallow-70B-Instruct-v0.1` (Fujii et al., 2024)
      - with RAG based on either BM25 or LaBSE+Faiss
  - 10 datasets (see [a] to [j] on the right panel →)
    - Covering Japanese (Ja), English (En), and Chinese (Zh)
    - We focus on erroneous test subsets: sets of source segments for which each MT system leads to translation errors
- 5 methods compared
  - Baseline: no edits
  - Word-Sub: Word substation (Koretaka et al., 2023)
  - Seq2Seq-B: Sequence-to-sequence (Koretaka et al., 2023)
  - LLM-NT: Iterative version of Ki & Carpuat (2025)
  - Ours: Our proposed method
- Evaluation metric: COMET (`wmt22-comet-da`)

Significant COMET changes over the Baseline



Distribution of the winning system

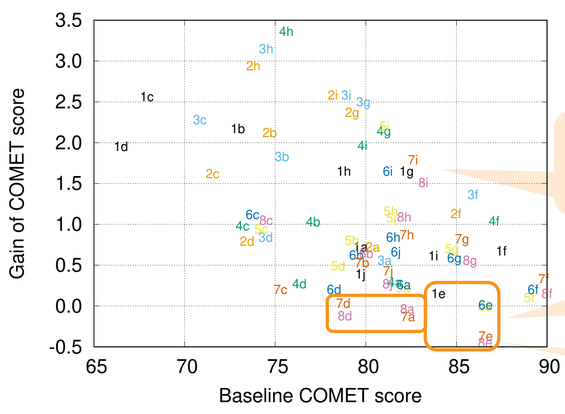


When translating erroneous test subsets, our targeted method performed better than non-targeted counterpart (LLM-NT),

4. Analyses

- Our method and erroneous test subsets

Correlation between Baseline COMET score and its gain

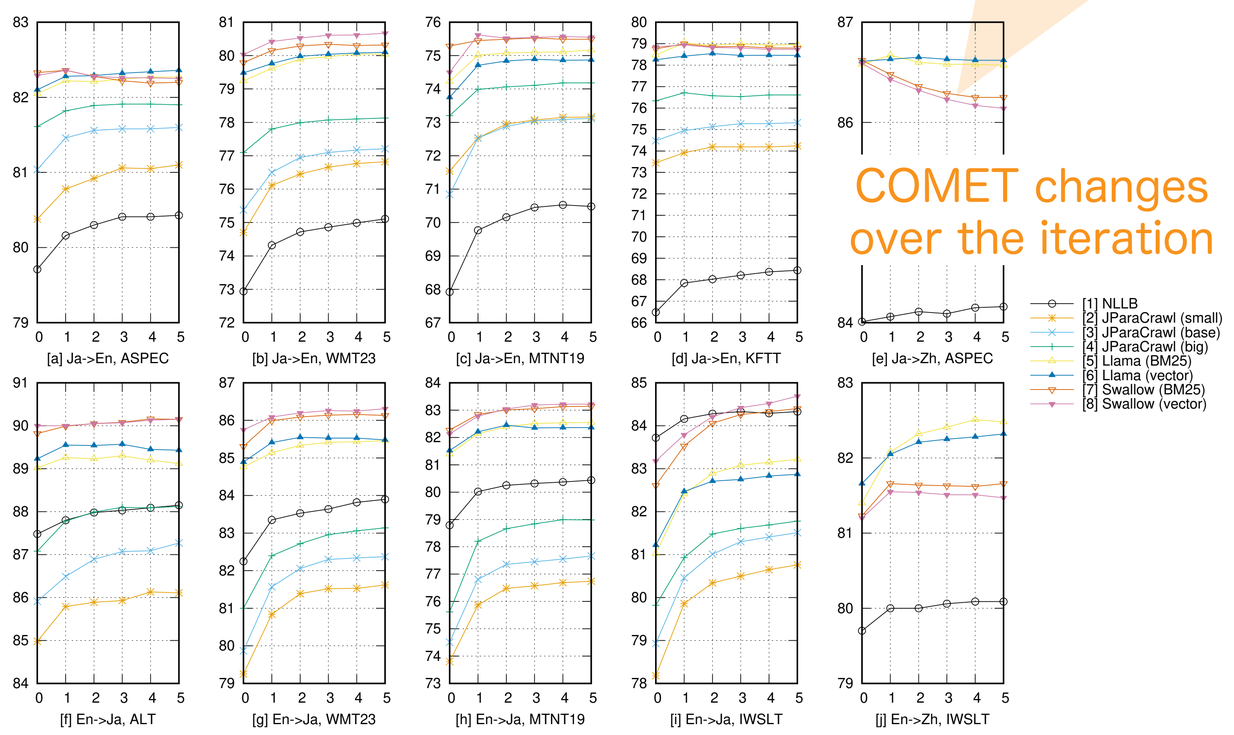


Naively translated English templates worked well for En→* tasks ([f] to [j])

Gain depends on datasets: [a][d][e] were difficult to improve

The two sig. worse configurations

COMET changes over the iteration



First iteration focused on the severest error and led to the biggest jump

Better segment-level QE will minimize the risk of quality deterioration

Acks & Contact

- This work was done during an internship of Hyuga Koretaka at NICT.
- A part of the results was obtained from the commissioned research contract between Ehime U. and NICT.

- Corresponding author: atsushi.fujita@nict.go.jp