Exploring the Importance of Source Text in Automatic Post-Editing for Context-Aware Machine Translation

Chaojun Wang[1], *Christian Hardmeier*[2,3], *Rico Sennrich*[4,1]

[1] *University of Edinburgh* [2] *IT University of Copenhagen* [3] *Uppsala University* [4] *University of Zurich*

Q&A

- Accurate translation requires document-level information.
- A recent work reports improvement on document-level consistency by **only target-language** automatic post-editing.
- **What are the effects of lack of using source-language information on post-editing translation quality?**
  According to our human evaluation, lack of source-knowledge would hurt the adequacy of post-editing translation.
- **Is there anything else we can take away?**
  The benefit of source-information is not reflected by automatic evaluation, highlighting blind spots of automatic evaluation for discourse-level MT.

Model and Data

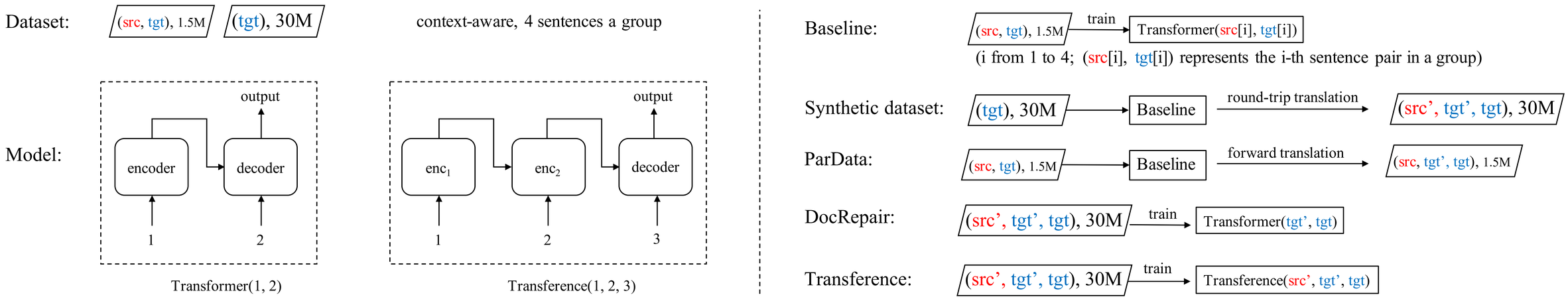

Fig. 1: Diagram of the process of data generation and model training. src: source language (English), tgt:target language (Russian). The prime denotes synthetic data.

Starting with two types of datasets and model architectures, we finally got **six models** to evaluate or generate synthetic data, which are:

- two sentence-level baselines (forward and backward);
- DocRepair and Transference trained with only synthetic dataset(as illustrated in the diagram);
- DocRepair and Transference trained with both synthetic dataset and ParData.

Automatic Evaluation

- **Multi-source APE (Transference)** better:
  - ellipsis test sets, especially on VP ellipsis.
- **Monolingual APE (DocRepair)** better:
  - T/V pronouns ("deixis");
  - Transliteration consistency ("lexical cohesion").
- **Equal**:
  - BLEU score on general test set (when ParData included).

Fig. 2: Automatic evaluation results of accuracy on contrastive test sets (left) and BLEU score on general test set (right).

Human Evaluation

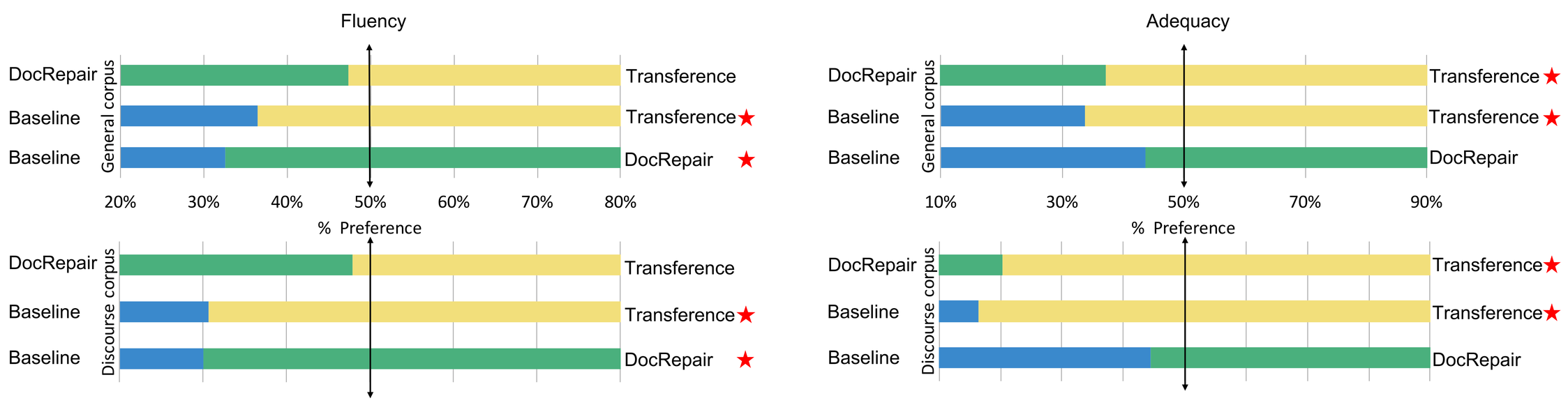

Fig. 3: Proportion of win in pairwise comparisons in terms of fluency (left two) and adequacy (right two) over general and discourse corpus. The red star indicates the statistically significant win.

- Quantitative Analysis: (both general and discourse corpus show same pattern, discourse corpus just more significant.)
  - **for fluency: Multi-source APE (Transference) = Monolingual APE (DocRepair) > Baseline;**
  - **for adequacy: Multi-source APE (Transference) > Monolingual APE (DocRepair) = Baseline.**
- Qualitative Analysis: source information permits Transference to correct certain recurring problems more reliably, such as agreement errors, mistranslations of proper names (e.g., Lena as Sarah), or the incorrect use or omission of subjunctive mood in conditional sentences.