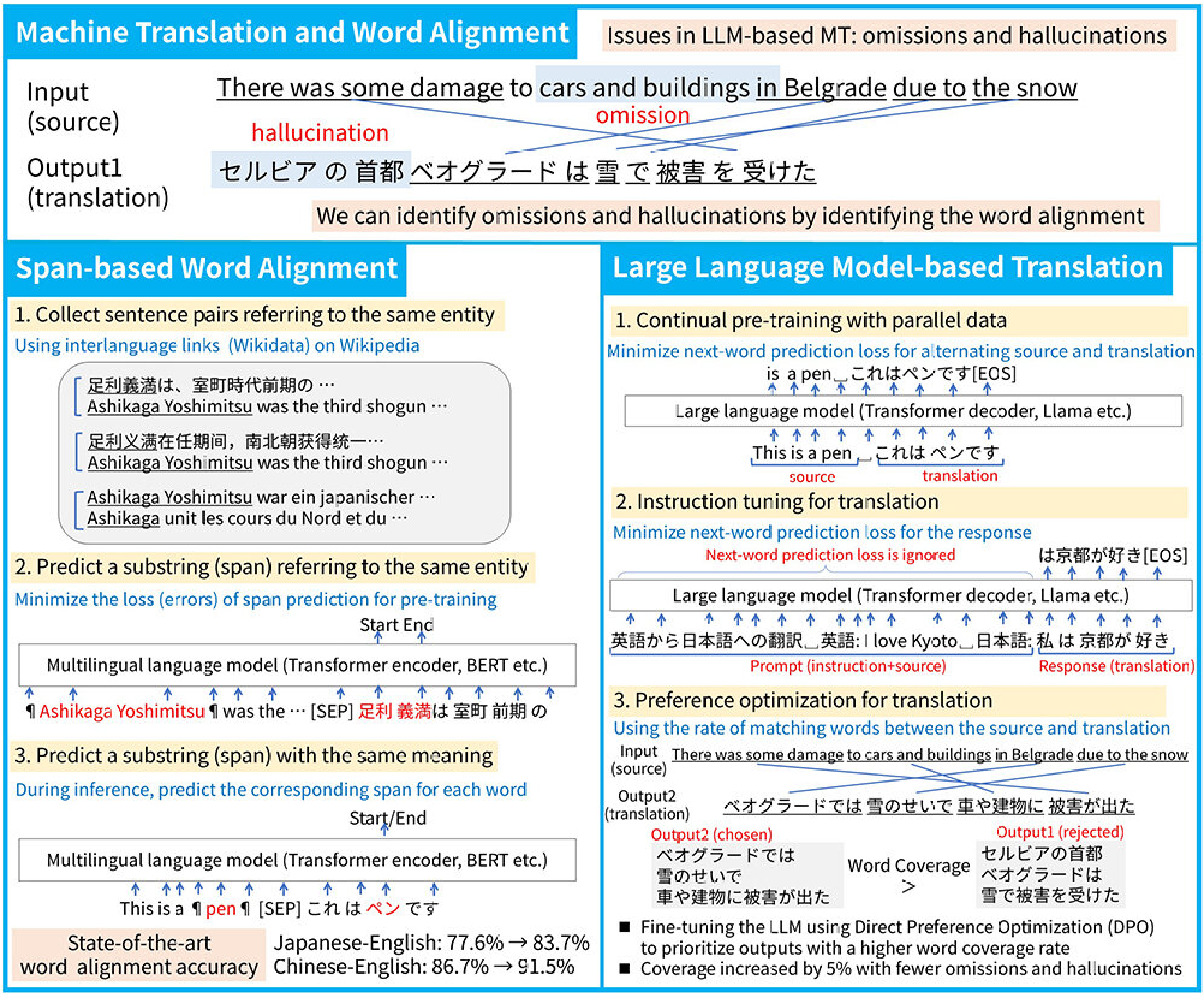
13

Preference optimization for LLM-based translation

Faithful translation without excess or deficiency

Abstract

Machine translation using Large Language Models (LLMs) can lead to errors such as “missing translation,” where parts of the source text are omitted, and “hallucination,” where the translation includes content not present in the source text. In this study, we first developed a method to train a highly accurate word alignment model using pairs of sentences in different languages that refer to the same entity in Wikipedia. We then developed a method for training a translation model with fewer omissions and hallucinations by maximizing pairs of words with equivalent meanings in the source text and its translation. In the future, we aim to improve machine translation technology in areas that require precise translations, including patents, law, and medicine. We will improve the fidelity of LLM-based translation, which excel at generating fluent and lengthy translations while maintaining consistency with the source text.



Machine Translation and Word Alignment

Issues in LLM-based MT: omissions and hallucinations

Input (source) There was some damage to cars and buildings in Belgrade due to the snow

omission

hallucination

Output1 (translation) セルビアの首都ベオグラードは雪で被害を受けた

We can identify omissions and hallucinations by identifying the word alignment

Span-based Word Alignment

1. Collect sentence pairs referring to the same entity

Using interlanguage links (Wikidata) on Wikipedia

足利義満は、室町時代前期の …
Ashikaga Yoshimitsu was the third shogun …

足利义满在任期间，南北朝获得统一…
Ashikaga Yoshimitsu was the third shogun …

Ashikaga Yoshimitsu war ein japanischer …
Ashikaga unit les cours du Nord et du …

2. Predict a substring (span) referring to the same entity

Minimize the loss (errors) of span prediction for pre-training

Start End

Multilingual language model (Transformer encoder, BERT etc.)

¶ Ashikaga Yoshimitsu ¶ was the … [SEP] 足利 義満は 室町 前期 の

3. Predict a substring (span) with the same meaning

During inference, predict the corresponding span for each word

Start/End

Multilingual language model (Transformer encoder, BERT etc.)

This is a ¶ pen ¶ [SEP] これ は ペン です

State-of-the-art word alignment accuracy	Japanese-English: 77.6% → 83.7% Chinese-English: 86.7% → 91.5%

Large Language Model-based Translation

1. Continual pre-training with parallel data

Minimize next-word prediction loss for alternating source and translation

is a pen␣これはペンです[EOS]

Large language model (Transformer decoder, Llama etc.)

This is a pen␣これは ペンです

source translation

2. Instruction tuning for translation

Minimize next-word prediction loss for the response

Next-word prediction loss is ignored

は京都が好き[EOS]

Large language model (Transformer decoder, Llama etc.)

英語から日本語への翻訳␣英語: I love Kyoto␣日本語: 私 は 京都が 好き

Prompt (instruction+source) Response (translation)

3. Preference optimization for translation

Using the rate of matching words between the source and translation

Input (source) There was some damage to cars and buildings in Belgrade due to the snow

Output2 (translation) ベオグラードでは 雪のせいで 車や建物に 被害が出た

Output2 (chosen)
ベオグラードでは
雪のせいで
車や建物に被害が出た

Word Coverage
>

Output1 (rejected)
セルビアの首都
ベオグラードは
雪で被害を受けた

■ Fine-tuning the LLM using Direct Preference Optimization (DPO) to prioritize outputs with a higher word coverage rate

■ Coverage increased by 5% with fewer omissions and hallucinations

References

[1] Q. Wu, M. Nagata, Y. Tsuruoka, “WSPAlign: Word alignment pre-training via large-scale weakly supervised span prediction,” in *Proc. The 61st Annual Meeting of the Association for Computational Linguistics (ACL 2023)*, 2023.

[2] Q. Wu, M. Nagata, Z. Mao, Y. Tsuruoka, “Word alignment as preference for machine translation,” in *Proc. The 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP 2024)*, 2024.

Contact

Masaaki Nagata, Linguistic Intelligence Research Group, Innovative Communication Laboratory