Enhancing Localization Workflows

A Deep Dive into Automated Post-Editing with GenAI

Speaker: Maciej Modrzejewski

BIG LANGUAGE SOLUTIONS

NMT Systems:
The Industry Standard in MT

- NMT Systems:
  - Leveraging MTPE (Machine Translation Post-Editing) for optimized workflows
  - rawMT
- Addressing Key Challenges in NMT
- Maximizing the Impact of Large Language Models (LLMs)
- Hybrid MT workflows
  - NMT-based workflows augmented by LLMs-based components
  - Use of Quality Estimation (QE) models

Limitations of Neural Machine Translation Systems

Challenges with Out-of-Domain Scenarios

Handling unfamiliar content & Maintaining quality

Model Robustness

Incorporating Specialized Terminology

Adapting to client-specific terms

Using specialized glossaries

Handling Language Nuances

Idiomatic expressions

Formal vs. informal language

Contextual understanding

LLMs enable paragraph-by-paragraph translations

Addressing Ambiguity and Bias

Managing bias

Cultural nuances

LLMs address limitations of NMT systems

Automated Post-Editing (APE)

"Automated Post-Editing (APE) is the process of refining MT content by sending the source text and the initial NMT output (hypothesis) to a Generative AI engine for linguistic review."

BIG LANGUAGE SOLUTIONS

APE Prompt

You will act as an Engine for Automated Post-Edition, specializing in the [domain_name] domain. You will receive {len(x)} source segments in {source_language} and {len(y)} machine-translated outputs in {target_language} from a custom, domain-adapted NMT engine.

Your task is to:
- Improve the fluency and translation quality of the output.
- Ensure 100% accuracy without introducing any new facts.
- Retain all relevant information.
- Match the capitalization of the source text.

Your final output must be in the target language: {target_language}.

Source: {x}
Custom Model Output: {y}   MT Hypothesis

Case 1: APE in out-of-domain scenario

- Use APE to post-edit an out-of-domain test set
- Test Data Set:
  - Khresmoi Summary Translation Test Data 2.0
  - Medical domain
  - Language pair: ENG-DEU
  - 500 segments

Case 1: APE in out-of-domain scenario

- BLEU scores:

	Big Language generic model	Google Translate	DeepL
Regular Translation	32.3	30.8	32.6
After APE with GPT-4o	34.3 (Δ +2.0)	34.2 (Δ +3.4)	33.7 (Δ +1.1)

- Average BLEU improvement: +2.15 BLEU

Case 1: APE in out-of-domain scenario

- COMET-20 and COMET-22 scores:

	Big Language generic model	Google Translate	DeepL
Regular Translation	COMET-20: 0.6340 COMET-22: 0.8626	COMET-20: 0.6977 COMET-22: 0.8808	COMET-20: 0.6958 COMET-22: 0.8797
After APE (gpt-4o)	COMET-20: 0.6968 COMET-22: 0.8810	COMET-20: 0.7037 COMET-22: 0.8824	COMET-20: 0.7023 COMET-22: 0.8819

Case 2: APE with a fine-tuned NMT system

- Perform APE on the output from a fine-tuned NMT system
- Training data size: 60k segments
- Test Data Set:
  - True Hold-Out Test set (not used in training)
  - Domain: Healthcare
  - Language pair: ENG-SPA-US
  - 1000 segments

Case 2: APE with a fine-tuned NMT system

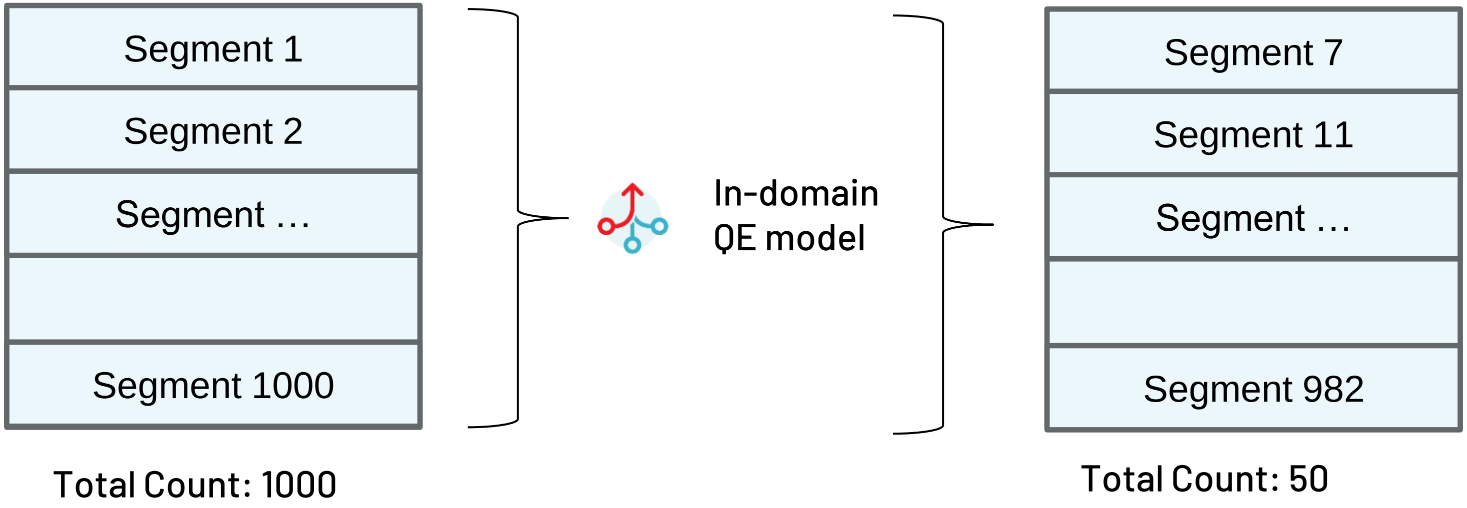
- BLEU and COMET scores:

	Big Language fine-tuned model	Google Translate
Regular Translation	BLEU: 72.8 COMET-20: 0.9675 COMET-22: 0.9119	BLEU: 49.9 COMET-20: 0.8238 COMET-22: 0.8835
After APE (gpt-4o)	BLEU: 63.5 COMET-20: 0.9218 COMET-22: 0.9066	BLEU: 51.5 COMET-20: 0.7961 COMET-22: 0.8730

➢ APE doesn't bring any improvement for fine-tuned NMT systems!

Case 3: APE with a in-domain ref-free QE model

- Idea: Identify **worst translations** from fine-tuned NMT system with an in-domain, reference-free QE model
- Perform APE only on those segments



Case 3: APE with a in-domain ref-free QE model

- Identical test data set as in Case #2
- Evaluation metrics for 50 worst translations (out of 1000 segments):

	Big Language fine-tuned model	Δ
Regular Translation	BLEU: 52.2 COMET-20: 0.5307 COMET-22: 0.8372	BLEU: -1.8 COMET-20: +0.053 COMET-22: +0.024
After APE (gpt-4o)	BLEU: 50.4 COMET-20: 0.5834 COMET-22: 0.8612	

Case 3: APE with a in-domain ref-free QE model

- Human Evaluation for the 50 worst performing segments
- APE-enhanced translation is preferred:
    - Translation 1: Regular Translation
    - Translation 2: APE-output
    - Linguist's review: *"In my opinion, [Translation 2] was a better translation, because, even though, glossary terms were not translated as per glossary like in Translation 1, there were no missing words, or issues with Spanish style, or inaccurate translations. Additionally, Translation 2 was more natural sounding and clear."*

Case 3: APE with a in-domain ref-free QE model

- Example for translation improvement:

Source	Division of Neighborhood Health Research	
Reference	División de Investigación Sanitaria del Vecindario	
Fine-tuned NMT model	División de Investigación Médica orientada a los vecindarios	Ref-free QE score: 0.5140
After APE	División de Investigación de Salud en los Vecindarios	Ref-free QE score: 0.5690

Case 3: APE with a in-domain ref-free QE model

- Edit Distance Report

<table>
<tr><th></th><th>Regular Translation</th><th>APE-output</th><th>Δ</th></tr>
<tr><td>Edited Segments</td><td>26/50</td><td>20/50</td><td>-23%</td></tr>
<tr><td>Absolute Edit Distance</td><td>783</td><td>655</td><td>-16%</td></tr>
<tr><td>Normalized Edit Distance</td><td>0.196</td><td>0.125</td><td>-36%</td></tr>
<tr><td>Total PE Time [in mins]</td><td>35</td><td>20</td><td rowspan="2">-43%</td></tr>
<tr><td>TTE [words/s]</td><td>6,84</td><td>3,91</td></tr>
</table>

Conclusions

1. APE Effectiveness
   i. Enhances baseline translation quality in out-of-domain NMT systems.
   ii. Identifies and corrects issues that commonly arise in these systems.
2. In-Domain Scenarios:
   i. APE may not always improve results; potential for performance decline.
   ii. Despite this, APE benefits approximately 5% of the worst translations.
3. Optimization Strategy:
   i. Use in-domain QE model to identify problematic segments.
   ii. Targeted APE application reduces post-editing time by 40% for these segments.
4. Use of GenAI engine interchangeable
   i. Use a fine-tuned LLM for APE for better results