University of Leeds, School of Languages, Cultures and Societies
Centre for Translation Studies

Automatic Translation Quality and its Estimation

Serge Sharoff

Introduction to Computer-Assisted Translation

1 Outline of exercises

On completion of these exercises, you should be able to:

- Understand differences between the Large Language Models and MT tools;
- Gain experience on the issues of hallucination, cohesion and automated translation evaluation;
- Design LLM prompts for producing translations and discussing issues in translated sentences;
- Design LLM prompts for translation quality assessment;
- Design methodologically correct experiments in evaluation of translation quality

2 Large Language Models as Translation engines

- Choose two of the following LLMs:
  - OpenAI's ChatGPT: `https://chat.openai.com/`
  - Google's Gemini: `https://gemini.google.com/`
  - Microsoft's CoPilot: `https://chat.bing.com/`
  - Anthropic's Claude: `https://claude.ai/`
  - Inflection's Pi: `https://pi.ai/`
- Take an ST and its TT with **known major** flaws
  Feel free to use your Translation test marked by your tutor or any other translated text with **known major** flaws
- Translate its problematic sentences with GAI tools
  Generative AI (AI) is a broad term, which covers both Machine Translation and LLMs
- Produce another translation for the same sentences with more traditional MT tools, such as Google Translate, DeepL, MateCat, Phrase,...
- Question the known flaws with the LLM tools
- Beware of hallucinations in MT/LLM output

3 Terminology extraction with Large Language Models

Feel free to experiment with different prompts aiming at extracting **useful** terms on the sentence level or on the document level. Add them to your term base (MultiTerm, MemoQ, Phrase or simply Excel).

4 Assessment of translation quality

- Experiment with translations on the sentence level vs the document level in LLMs
- Experiment with different quality assessment prompts:
  - Simple: *Quality on the scale of 0 to 100*
  - More specific: separately for ST Comprehension, Terminology, etc
  - Very specific: for example, provide the Marking criteria for your Specialised Translation modules
- Enrich quality prompts with Chain of thought
  either with explicit examples or with self-reflection

Please check the following examples of templates:

> Score the following translation with respect to the human reference with one to five stars.
> Where one star means "Nonsense/No meaning preserved",
> two stars mean "Some meaning preserved, but not understandable",
> three stars mean "Some meaning preserved and understandable",
> four stars mean "Most meaning preserved with few grammar mistakes",
> and five stars mean "Perfect meaning and grammar".
> Source: <SRC>
> *Reference: <REF>*
> Translation: <TGT>

Also try this template **without** the human reference.

- A prompt can include more specific instructions:

> Source: <SRC>
> *Reference: <REF>*
> Translation: <TGT>
> Identify the major and minor errors in this translation. Note that Major errors refer to actual translation or grammatical errors, and Minor errors refer to smaller imperfections, and purely subjective opinions about the translation.
> Count the number of major and minor errors and compute the final score for this translation. Deduct 5 points for each major error. Deduct 1 point for each minor error. If the translation has no errors, its score will be 0.

Again, the reference translation in italics can be omitted.

- It is also possible to add examples (Chain of thought prompts):

  Source: They were addressed to her son, who has autism and lives in a private care facility, she said. But instead of her son's name inside…
  Translation: Sie wurden an ihren Sohn gerichtet, der Autismus hat und in einer privaten Pflegeeinrichtung lebt, sagte sie. Aber anstelle des Namens…
  Deduct 5 points for each major error. Deduct 1 point for each minor error. If the translation has no errors, its score will be 0.
  A: Major errors:
  (1) Span "Dear Maine 's Department of Health and Human Services" – Untranslated text
  (2) Span "im Inneren" – Mistranslation
  Minor errors:
  (1) "Briefe „" – Punctuation
  (2) "wurden" – Grammar
  (3) Span "im Inneren, als Sie sie öffneten, sagten die Briefe" – Awkward Style
  Based on the above evaluation, The final score for this translation is -5-5-1-1-1=-13.

  Now evaluate the following Source-Translation pair:
  Source: <SRC>
  Translation: <TGT>

- Try using reflective chain-of-thought prompts:

  Source: <SRC>
  Translation: <TGT>
  Given the source text and its translation, please list the parameters to define whether the translation contains major errors or not.

  After that please consider applying these criteria to list the major errors in this translation.

You can also use more explicit translation quality criteria.

5 Homework

Take a test text and evaluate your most preferred prompt template for each sentence:

- Record the number of cases LLMs predicted the errors, and
- Report the accuracy of your prompt

Check the terminology for:

- False positives (Type I errors)
- False negatives (Type II errors)
- Precision and recall

Compare this terminology against the spam filtering example:

- A spam filtering tool wrongly classifies a legitimate email as spam;
- It fails to detect a realistic spam email as spam.