Translation User Group

Translation using AI Techniques - *Proof of Concept*

Rory Davidson
2024-12-10

The Task

- Localising SNOMED CT to a non-English speaking locale is a costly and time-consuming exercise, with the translation of concepts accounting for a significant proportion of the costs.

- To accelerate this process Neural Machine Translation (NMT) services like DeepL, Amazon Translate, Google Translate and Microsoft Translator have been used.

- Although these are mature and well-developed products, translating clinical terms accurately is difficult. Furthermore, translations into “lower resource” languages are often of lower quality than translations into “higher resource” languages.

The Task

- Recent developments in Large Language Models (LLMs) offer an alternative technology for translation. Unlike NMT models, LLMs can be provided with "context" to help them solve natural language tasks. In this proof-of-concept exercise, we study whether an LLM can exploit the structure within SNOMED CT to provide better translations than the leading NMT model, DeepL.

- For this exercise, we selected the **Aya model developed by Cohere.AI**. Aya is an open-source, 11B parameter LLM that has been fine-tuned on a large corpus of multi-lingual text (also open sourced). Aya's fine-tuning dataset contains 513M training examples evenly balanced across 114 languages.

What We Did

- We evaluated Aya and DeepL on 15,000 concepts sampled from the Refsets of the following member countries: Sweden, The Netherlands, Korea and Estonia.
- Our samples were balanced across the following dimensions – which we believed would have an impact on translation accuracy: hierarchy, depth of concept, length of description and whether parent concepts had been translated.
- Our samples were taken from the following hierarchies: Substance, Body Structure, Finding, Disorder, Procedure, Morphological Abnormality
- We evaluated the translations using two metrics: (1) exact matches to an existing reference translation, (2) the Levenshtein Ratio: a character-level measure of similarity between a translation and a reference translation.

Results

1. “Out of the box”, translations by Aya are slightly less accurate than those by DeepL.
2. To exploit the “context awareness” of Aya, we exploited the structured nature of the terminology to collate a set of relevant translation examples that could be presented to the model before it was asked to perform a translation. (This is called Retrieval Augmented Generation, or “RAG”). These examples included:
    - Reference translations for parent concepts.
    - Reference translations for the attributes of inferred relationships.
    - Reference translations of other concepts where the corresponding English preferred term is syntactically similar to the term being translated.
3. Where this additional information is available (i.e. concepts with “rich context”), Aya’s translations are significantly more accurate than DeepL’s. This effect is consistent across both the high resource languages (Swedish, Dutch) and low resource languages (Estonian, Korean).

Example of a rich prompt used for Aya translations

Translate the following clinical concept into Estonian: "Cardiac pacemaker, device". Südamestimulaator.
Translate the following clinical concept into Estonian: "Implantation of biventricular cardiac pacemaker system". Biventrikulaarse südamestimulaatori implanteerimine.
Translate the following clinical concept into Estonian: "Cardiac pacemaker procedure". Südamestimulaatoriga seotud protseduur.
Translate the following clinical concept into Estonian: "Maintenance procedure for cardiac pacemaker system".

Translation accuracy for concepts with “rich context”

	Exact Matches Aya	Exact Matches DeepL	Levenshtein Ratio Aya	Levenshtein Ratio DeepL
Dutch	0.43	0.11	0.86	0.71
Estonian	0.28	0.07	0.83	0.66
Korean	0.38	0.07	0.85	0.74
Swedish	0.52	0.12	0.88	0.72

Further Progress

- We then sent 100 random translations of each model (Deepl & Aya) to the following NRCs, to manually verify & score the quality of translations of each model:
  - Korean NRC
  - Netherlands NRC
  - Estonian NRC
  - Swedish NRC
- The results were as follows:

Count of sctid	Column Labels					
Row Labels	Aya	Both are good	DeepL	Neither is good	<<NO RESPONSE>>	Grand Total
Dutch	31	1	12	56		100
Estonian	20	21	31	28		100
Korean	10	26	33	30	1	100
Swedish	41	12	16	31		100
Grand Total	**102**	**60**	**92**	**145**	**1**	**400**

What Do We Learn From This?

- The evaluation was done using the “retrieval-augmented” Aya translations – i.e. where we gave Aya the contextual hints from related translations. As expected, the two languages for which we have plenty of context (Dutch and Swedish) are the ones where Aya is preferred.
- The preference ratios (Aya : DeepL) are roughly of the order 3:1 for Dutch, Korean and Swedish – which is a wider margin than we might have expected given the differences in the evaluation metric scores that we saw during the translation exercise.
- It underlines the point that there really is no substitute for human evaluations if one is looking to answer questions of the form “how much better is A than B?”.
- The feedback from the Dutch team was interesting: 34% of the translations were marked as “Neither is good” with the explanation that it “does not adhere to our internal guidelines”. This is where fine-tuning would help, getting Aya to learn the “house style” for each team and produce translations that conform to it.
- Note: we only had time to run a very brief fine-tuning run and didn’t use the fine-tuned model in the human evaluation but we believe it could work well for Dutch and for Swedish.

Next Steps

- **Translation Maintainers Option:**
  - We perform a more extensive fine-tuning run on Dutch and Swedish to see if we can teach Aya the “house style” and further improve the preference score over DeepL. This needs buy-in from both NRCs that (A) it could be useful to them and (B) they’d be willing to take on this task.
- **New Translators Option PoC:**
  - We re-run using the latest Aya, Claude, GPT or Gemini model (but keep the retrieval augmentation step) to see if they improve the translations.
  - We run a further experiment, this time using a “basket” of example translations into other languages to see if that offers any improvements. The focus here could be on an as-yet untranslated language like Italian, or it could be an existing Romance language.
- Regardless, all of this is available in a GitHub repository open to any NRCs (and anyone) who are interested in continuing the experimentation - https://github.com/IHTSDO/snomed_translation_poc

Questions?

SNOMED International

Leading healthcare terminology, worldwide