LLMs for translation
Introduction
The task is to translate two types of texts:
- Phone conversations
- Interviews with refugees

in two Arabic dialects:
- Tunisian
- North Levantine

Into English
Traditional neural machine translation: trained on sentence pairs (source sentence, target sentence)

LLMs: trained mostly just on unrelated text in multiple languages (in larger chunks than sentences)

Challenges:
- The dialects itself (not very common in the parallel data) -> advantage LLMs
- Code-switching (e.g. French words in Tunisian sentence) -> advantage LLMs
- Context is important, one sentence does not contain enough info -> advantage LLMs
Problems
Problems

- Not that many LLMs are trained on Arabic (finetuning is possible, but complicated bc. of different script)
- Not that many LLMs are trained for translation -> finetuning
- Finetuning requires large GPU memory
- We tested:
  - Baseline: Google Translate, NLLB 1.3B
  - Mistral
  - Gemma
  - Aya-101
  - Cohere4AI/Command-R
  - Jais
  - Falcon
  - Bloom-z
  - ALMA
Some LLMs fail completely, some fail in the scenario with context.

Few-shot for some reason harms the output quality in our scenario?

Context-less prompt:

- Translate this line from Tunisian Arabic into English. Arabic: \{source_example\}
  English: \{target_example\}
- Translate this line from Tunisian Arabic into English. Arabic: \{src_line\}
  English:

- **Source_example, target_example** is a sentence pair from some data similar to our test set (1-shot learning)
- **Src_line** is our input sentence
### Forcing LLMs to translate

<table>
<thead>
<tr>
<th>Model</th>
<th>MADLAD 10B</th>
<th>eTranslation</th>
<th>Google translate</th>
<th>NLLB 1.3B</th>
<th>Bloomz</th>
<th>aya-101</th>
<th>Command-R</th>
<th>jais-13B-chat</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLEU</td>
<td>18.6</td>
<td>15.1</td>
<td>28.7</td>
<td>20.7</td>
<td>7.2</td>
<td>17.6</td>
<td>22.1</td>
<td>20.7</td>
</tr>
</tbody>
</table>

- Sentence-level results are quite promising, we should try to use the context
Forcing LLMs to translate

- We want to use context, but the evaluation is still sentence-level -> same number of sentences in the output

- We need to split the translation of a larger chunk back into sentences:
  - Sentence splitter - unreliable, introducing another tool
  - Asking the model to keep the same number of lines: only works for few lines and it still moves the content between the sentences
  - Using separator token, e.g. [s]: a bit better, especially after finetuning, but still not 100% reliable
  - Only translating one sentence at a time, but providing the whole context (source document and previously translated prefix) in the prompt
Forcing LLMs to translate

- Context prompt:

"We need to translate a single line from conversation in Tunisian Arabic into English. This is the conversation {src_context}
The start of the conversation is already translated into English: {prev_context}
Translate the following line from Arabic to English. Be very literal, and only translate the content of the line, do not add any explanations: {src_line}"

- The only “document-level” prompt that work reliably at least for some models (Command-R)

- BLEU 22.1 -> 26.1

- We finetuned with this prompt (maybe the other approaches would start to work after finetuning as well)
We need to translate the following conversation in Levantine Arabic into English. It is very important to keep the same number of lines in the translation:

مَاشَىَ * بَلََّلََْشَدَ * الرَّبِيعَ العَرَبِيَّة مِنْها مَا يَقُولُوا بِيَقُولُوا بِالْأَلْفِيَدِ أوُرَبِطَهُشَّ إِذَا مَا نَلَّدُ خَطَّمَتْ بَلوُسَبَدَ * دَوْلَتَانِ يَنْتَهِيَانِ

فَوُقَّتَ الْلُوَصْلَةَ * الرَّبِيعَ العَرَبِيَّةَ مُسَيَّدَتْ الْعَالَمِ بِبَنْسَتْ النَّهَدِ الشَّيْعَةِ سَيَاسَةً

لَاذَ الْحُكُومَةَ أوُرَّدَتْ بِالرَّبِيعَ الْمُهْدَرَةَ قَلْبَهُ رُفَهَّتْ بِفَضْلَةِ الْفَسُنَّاءَ

وَكَانَ فَرْصَارَ بَيْنَشَدَأَ أَمَّنَمْدُعَوا التَّجَمَّعَةَ أوُجَعْتُهُ يَنْتَهِيَانِ

فَبَلََّلََْشَدَ الْعَالَمِ بِتَطَّرَدُتْ بِيَطَهَّرَاتٍ

إِلَّا بِبَسْبَدََّ حَرَةً وَعَدَالَةً اجتماعِيَّةً وَمَعْدَادٍ الْحُكُومَةََ أَ لْمَا بَدَّلَ بِالْبَلَّادِ يَتَغِيِّرُ لَا نَظَامَ وَلَا شَيْءٍ * أَوْكَمَ

Yes.
At first, the Arab Spring started, as they say, I think in 2014, if I'm not mistaken, in Tunisia.
Then it spread to other countries.
When the Arab Spring reached Egypt the people in Syria started demanding the same thing for political reasons.
Because the government or the president that is in power now has been there since the sixties.
And there was a start of security tightening, gatherings were banned and things like that.
So the people started demanding, protesting.
They just wanted freedom, social justice and all that.
They didn't want to change the regime or anything at first, okay.
Translate the following line from Arabic to English. Arabic: بينما و قاعلا موضوع إنو يكون عندهم خصوصيتهم كمان شغة مهمة

English: While and actually the fact that they have their own privacy is also a big thing.
Forcing LLMs to translate

- **Finetuning:**
  - Command-R has 34B parameters
  - With QLoRA and 4-bit quantization, it fits on a 40GB GPU and takes 2-3 hours for our data, so <10$ even if you need to rent a GPU

- **Final results:**

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Google translate</td>
<td>28.7</td>
</tr>
<tr>
<td>NLLB</td>
<td>20.7</td>
</tr>
<tr>
<td>NLLB+finetuning</td>
<td>27.1</td>
</tr>
<tr>
<td>Command-R</td>
<td>22.1</td>
</tr>
<tr>
<td>Command-R+context</td>
<td>26.1</td>
</tr>
<tr>
<td><strong>Command-R+context+finetuning</strong></td>
<td><strong>32</strong></td>
</tr>
</tbody>
</table>