Prof. Dr. Goran Glavaš,
M.Sc. Fabian David Schmidt
M.Sc. Benedikt Ebing
Lecture Chair XII for Natural Language Processing, Universität Würzburg

7. Exercise for "Multilingual Natural Language Processing"

07.07.2023

1 Paper Readings

We segment the literature on neural machine translation (NMT) as follows:

1. **Bitext Mining**
   - CCMatrix: Mining Billions of High-Quality Parallel Sentences on the Web
2. **NMT with Large Language Models**
   - Multilingual Machine Translation with Large Language Models: Empirical Results and Analysis
3. **Translate-Test for Sequence-Level Cross-Lingual Transfer**
   - Revisiting Machine Translation for Cross-lingual Classification

2 Bitext Mining

Modern supervised NMT systems like NLLB achieve remarkable translation performance even on low-resource languages. In this part of the exercise, we will explore one key building block of such NMTs called bitext mining.

1. Briefly explain bitext mining.

> Automatically retrieving parallel sentences (i.e., sentences that are translations of each other) from large corpora (e.g., the web).

2. Why is bitext mining important for NMT? Explain the motivation.

> Parallel sentences are crucial for the training of multilingual sentence encoders and neural machine translation. For higher resource languages, comparably large parallel corpora exist (e.g., United Nations Parallel Corpus or Europarl). However, specifically for lower resource languages, there are less efforts to manually construct parallel corpora, creating the need for automatically mined parallel data.

3. Summarize the bitext mining method presented in *CCMatrix: Mining Billions of High-Quality Parallel Sentences on the Web (Schwenk et al., 2021)*.

> a) Text extraction: Extracting text form the CCNet JSON, deduplicate the sentences, perform language identification
>
> b) Get all sentence embeddings using a sentence encoder (storing them in a compressed from using FAISS library)
>
> c) For all pairs of sentences of two languages, compute the margin-based criterion for both directions (sentence embeddings that are "close" in their respective neighborhood). Built the union of the forward and backward direction, sort the candidates and omit source or target sentences that are already used. Then apply a threshold (hyperparameter) to decide whether two sentences are mutual translations.

3 NMT with Large Language Models

Large language models like ChatGPT, Bloomz, or mT0 gain increasing attention as "generalists", i.e., they are able to solve tasks with no or few examples seen. In this section, we examine this hypothesis and investigate whether large language models outperform supervised NMT systems in automatic translation.

1. Do large language models outperform supervised NMT models? Briefly summarize the main results from the paper.

Compare four large language models, two that are "only" pretrained (XGLM, OPT) and two instruction-tuned (BLOOMZ, ChatGPT) against two open-source translation models (M2M-100 and NLLB) ChatGPT underperforms NLLB in 83.33% of the tested translation directions (ChatGPT performs best among the large language models), but shows competitive performance compared to M2M-100. BLOOMZ outperforms supervised NMT models on some language families. Language models without instruction tuning perform generally worse than supervised NMT models. All in all, supervised NMT models still outperform large language models, particularly in lower resource languages and translations from English to the target language.
Remarks:

- Instruction-tuned models are typically fine-tuned on translation tasks
- NLLB-1.3B is not the largest NMT model from that family
- There is no comparison against a commercial NMT model like Google Translate

2. Large language models are prone to produce certain translation errors. Name and describe the three typical translation errors presented in the paper.

- *Off-target translations*: Translating into the wrong target language (e.g., into Azerbaijani instead of Turkish).
- *Hallucination*: Generating highly pathological translations that are unrelated with the source sentence.
- *Monotonic translations*: Translating sentences word-by-word lacking effective word-reordering of the target language.

3. Describe the issue of "data leakage" when evaluating large language models on publicly available datasets.

Large language (and instruction-tuned) models might be trained on publicly available evaluation data. This effect might inflate the performance on certain evaluation datasets. It is particularly an issue for commercial models that do not open-source their training corpora which hinders fair comparison.

4. How important is the choice of the template for prompting? Briefly describe the corresponding results from the paper.

> The performance of large language models relies on the template choice. There is a gap in the average performance of up to 16 BLEU between the best and the worst performing template. However, even unreasonable templates may produce decent translations (e.g., instruct the model to summarize). The role of the template is still an open research area.

4 Translate-Test for Sequence-Level Cross-Lingual Transfer

Thanks to ever improving machine translation, translation-based approaches (re-)gain a lot of popularity. Another paradigm is `translate-test`, in which test instances in the target language are translated to a high-resource language, typically English, in which you perform inference on models trained on high(er) quality annoations than in the target language.

As part of this exercise, we focus on a very recent paper that showcases important developments relevant for cross-lingual transfer.

Reading: Revisiting Machine Translation for Cross-lingual Classification

Elaborate on the key ingredients which the authors lever to materially improve `translate-test` over prior work!

> - **Better translation models**: the original translations by today's standards are subpar and state-of-the-art models produce much better performance to translate to better inference
> - **Account for distribution shifts**: Using the original human-generated data does align well with what models see at (translate-)test time. Machine translation models introduce what is referred to as "translationese" (e.g., copying certain words from the input, etc.). Hence, models need to be trained on both original and translated data.
> - **Monolingual models**: rather than evaluating `translate-test` on multilingual models, evaluate the approach on stronger, focused monolingual models (RoBerta & DeBerta). The model's tokenizers are tailored to English which typically improves performance vis-a-vis multilingual models by slightly less than 2 points.