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.
2. Why is bitext mining important for NMT? Explain the motivation.
3. Summarize the bitext mining method presented in *CCMatrix: Mining Billions of High-Quality Parallel Sentences on the Web (Schwenk et al., 2021)*.

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.
2. Large language models are prone to produce certain translation errors. Name and describe the three typical translation errors presented in the paper.
3. Describe the issue of "data leakage" when evaluating large language models on publicly available datasets.
4. How important is the choice of the template for prompting? Briefly describe the corresponding results from the paper.

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!