An Empirical Study of Multilingual Representations from Language Modeling and Translation

- a principled standpoint and train comparable MT and LM systems to contrast their cross-lingual and monolingual downstream performances;
- an empirical study on publicly available pretrained LM and MT systems and study whether continued training on MT helps or hinders the emergence of cross-lingual capabilities.
- Data: UNPC (Ziemski et al., 2016) and OpenSubtitles (Tiedemann, 2012)
- Languages: Arabic, Chinese, English, French, Russian, and Spanish
- Models
  1. Masked Language Modeling (MLM) with the BERT architecture (Devlin et al., 2019);
  2. Causal Language Modeling (CLM) with the GPT-2 architecture (Radford et al., 2019);
  3. Translation Language Modeling (TLM) with the GPT-2 architecture, where the input is the concatenation of a language pair following a setup similar to Conneau and Lample (2019);
  4. Denoising Sequenece-to-Sequence Langauge Modeling with BART architecture (Lewis et al., 2020);
  5. Machine Translation (MT) with the classic encoder-decoder transformer architecture (Vaswani et al., 2017) and the BART architecture (Lewis et al., 2020).