Reading Assignment 3
Hamza Rashid, 260971031
COMP 550, Fall 2024

The paper, *"Multilingual Denoising Pre-training for Neural Machine Translation"* (Liu et al., TACL 2020) presents mBART—a sequence-to-sequence denoising auto-encoder pre-trained on large-scale monolingual corpora in many languages using the BART objective (Lewis et al., 2019). The paper demonstrates that multilingual denoising pre-training produces significant performance gains across a wide variety of Machine Translation (MT) tasks, including supervised sentence-level and document-level MT with and without Back-Translation (BT), and unsupervised MT. The pre-training is done on 25 languages extracted from the Common Crawl (CC) corpora (CC25). Due to the skewed distribution of corpus sizes, they used up and down sampling on each language's corpus based on the percentage each one takes up in CC25. As for preprocessing, they used the SentencePiece (language-independent) subword-tokenizer learned on the full CC data. While this model is learned on more languages than present in CC25, the authors deemed it useful for fine-tuning on additional languages. This can be particularly effective on OOV items when fine-tuning on unseen languages sharing a family in CC25—the pre-trained embeddings might be able to encode meaningful information due to the subword overlap.

Previous MT methods, such as XLM-R (XLM-RoBERTa), pre-trained BERT on the multilingual Masked Language Modelling (MLM) objective (replacing individual tokens with designated a mask token, and having the model predict the correct value) using only monolingual data. While bidirectional tranformers produce contextual representations that consider the entire text, BERT's encoder-only architecture is difficult to fine-tune for text generation tasks, where each generated token depends only on the previosly generated tokens. On the other hand, BART's encoder-decoder architecture gives it the representational capacity of a bidirectional transformer (its encoder) and the ability to generate sequences (through its autoregressive transformer decoder). This sequence to sequence (Seq2Seq) architecture allows greater flexibility in the noising method, and hence can improve results in predictive and generative tasks. In particular, the noising function used in mBART's pre-training removes spans of text and replaces them with a mask token (in contrast to masking individual tokens in MLM), and permutes the order of sentences within each document. While mBART is pretrained on 25 languages from the CC corpora, XLM-R is pretrained on 100 languages over CC.

References

- Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual Denoising Pre-training for Neural Machine Translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. *In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.