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

The paper, *"Multilingual Denoising Pre-training for Neural Machine Translation"* (Liu et al., 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, and unsupervised MT. However, they do not experiment on language pairs where neither one is English, such as Norwegian-Spanish (Hareide et al., 2013), and thus the results are not necessarily generalizable to all language pairs. They experiment with various baselines, including transformers with random initializations and BART models with varying degrees of multilingual pre-training (monolingual, bilingual, hexalingual), which are compared against the flagship 25-language mBART25 model. Training required 256 GPUs and took 2.5 weeks to complete, which is relatively expensive compared to the gains in MT tasks.

The pre-training was done on 25 languages extracted from the Common Crawl (CC) corpora (CC25). For preprocessing, they used the SentencePiece (Kudo et al., 2018) subword-tokenizer learned on the full CC data. Although the tokenizer is learned on more languages than present in CC25, the authors deemed it useful for fine-tuning on additional languages. Its language-independent design makes it an appropriate tool for capturing morphological details in multilingual data. This can be effective when fine-tuning on unseen languages sharing a family in CC25—the pre-trained embeddings might encode meaningful information due to the subword overlap.

Previous NMT methods, such as XLM-RoBERTa (Conneau et al., 2019), pre-trained BERT on the MLM objective in a multilingual setting (100 languages over the CC corpus). BERT's encoder-only architecture can only predict missing tokens independently, limiting the scope of noising methods. In contrast, BART's encoder-decoder architecture not only resembles the originally proposed transformer (Vaswani et al., 2017), but also gives it the representational capacity of a bidirectional transformer (its encoder, capturing left-and-right contexts), and the ability to generate sequences with its autoregressive transformer (decoder) using the encoder's embeddings and left context. This sequence to sequence (Seq2Seq) architecture allows greater flexibility in the noising method and generative task. The noising function used in mBART's pre-training performs text-infilling (replaces spans of text with a mask token, in contrast to individual tokens in MLM), and permutations of entire sentences. As a result, mBART can learn relationships within and across sentences more effectively. Due to their architectural differences, XLM-R is pre-trained so that its resulting parameters can be later used to initialize a translation model encoder, while mBART simultaneously pre-trains the encoder and the decoder due to the Seq2Seq architecture. Finally, mBART is pre-trained to maximize the log-likelihood of the orginal input sequence given its post-noise transformation.

For supervised sentence-level MT, they fine-tuned mBART on a single pair of bi-text data at a time (covering all languages, but not all pairs in CC25), feeding the source language into the encoder and decoding the target with teacher forcing (the decoder conditions on the true target's left-context). The mBART25 weights acheived gains on all low and medium-resource pairs (against the randomly initialized baselines), acheiving 12+ BLEU point gains on some pairs having more than $100k$ parallel texts. However, fine-tuning was ineffective in low-resource cases such as En-Gu, which had $\sim 10k$ samples. For high-resource pairs, the pre-trained weights did not acheive consistent gains, even hurting performance when more than 25M samples were present. They suspected that fine-tuning diluted the pre-trained weights in this scenario. The pre-training might have been rendered redundant given large fine-tuning data, or introduced initial weights that hurted fine-tuning convergence. The authors did not explore (or report) any strategies to resolve the issue, such as adjusting training steps, or training on multiple language pairs at a time, so that the model could generalize across them. Lastly, mBART25 improved its BLEU scores with Back-Translation (BT; Sennrich et al., 2016) on low and high-resource language pairs (and outperformed XLM-R even without BT on English-Romanian translation).

For document-level MT, they compared the results of sentence and document-level fine-tuning of mBART25 on the En-De and Zh-En language pairs. Both variants of mBART25 outperformed the random baselines in these tasks, with a 26 point BLEU gain on document-level Zh-En translation. Furthermore, they tested mBART25 in three unsupervised MT scenarios: no bi-text exists for either language, both languages occur in seperate bi-texts, there is another bi-text with a different source language. The last scenario saw significant improvements in low-resource pairs compared to the supervised method. Fine-tuning mBART25 on Hi-En led to a BLEU score of 13.8 on Gu-En, compared to 0.3 in the supervised approach.

More broadly, sentence-level experiments showed that mBART25 can improve performance for languages that did not appear in the pre-training corpora. As a result, the authors suggested that the pre-training weights are at least partially language universal; a claim which is at odds with all tested language pairs including english.

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.

- Conneau, Alexis, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. "Unsupervised Cross-lingual Representation Learning at Scale." In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 8440–8451.

- Lewis, Mike, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves 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*, 7871–7880.

- Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding Back-Translation at Scale. *In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500, Brussels, Belgium. Association for Computational Linguistics.

- Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

- Hareide, Lidun, 2013, The Norwegian-Spanish Parallel Corpus, *Common Language Resources and Technology Infrastructure Norway (CLARINO) Bergen Repository*, http://hdl.handle.net/11509/73.

- Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. *Advances in Neural Information Processing Systems* 30: 5998–6008.