1 Description of the reading material

The paper "Multilingual Denoising Pre-training for Neural Machine Translation" [3] introduces mBART, a **multilingual Seq2Seq [6] autoencoder for Neural Machine Translation**. Using a denoising pre-training objective, mBART trains on monolingual data from 25 languages by applying noise and reconstructing the original text. It achieves state-of-the-art BLEU [4] improvements in low- and medium-resource settings and effectively transfers knowledge to language pairs without bi-text or pre-training data.

► ***Strength.*** This study addresses key limitations in previous works on pre-training for machine translation, such as focusing only on encoders or English corpora. **mBART provides a flexible parameter set that can be fine-tuned for any language pair in both supervised and unsupervised scenarios, without requiring task-specific or language-specific adjustments**, making it highly practical and adaptable. Another strength lies in its **thorough empirical approach**, featuring detailed ablation studies that investigate factors such as bi-text types, data volume, and back-translation (BT) [5], which underscore mBART's versatility and effectiveness. Experimentally, **mBART achieves state-of-the-art performance in both sentence-level and document-level translation tasks**, significantly enhancing machine translation outcomes.

► ***Weakness.*** In high-resource settings, **pre-training with mBART negatively impact performance**, a phenomenon the paper attributes to gradients being "washed out" during fine-tuning. However, **this explanation requires deeper investigation**, such as analyzing the volume of bi-text at which pre-training performance begins to decline and assessing the role of pre-training steps in high-resource scenarios. Moreover, the "washing out gradients" hypothesis is not empirically validated; **more analysis are needed to quantify the importance of gradients**. Another limitation lies in the evaluation metrics, as the paper **only focuses on BLEU scores**, ignoring other metricss like GLEU [7], which could provide a more nuanced evaluation. These additional metrics might also help clarify why performance declines in high-resource settings, offering more insights beyond BLEU.

► ***Questions.*** I am confused by the finding: "Pre-trained Transformer layers learn universal properties of language that generalize well even with minimal lexical overlap." Is this learning due to the **inherent capacity of the Transformer architecture**, or is it driven by the **structural similarities between different languages**? Additionally, the observation that **mBART25 offers marginal improvements** over mBART06 and mBART02 in tasks such as Ni-En MT. **Is pre-training on major languages sufficient to achieve robust generalization?** It seems like that **more data leads to better performance.** It is a common question in machine learning field, with no much insights presented.

2 Discussion

► ***Comparison between mBART and mBERT.*** For **architecture**, mBART uses a Seq2Seq framework, pre-training both the encoder and decoder for translation task, while mBERT and XLM-R [1] focus on encoder-only masked language models. For **pre-training tasks**, mBART employs a denoising objective, reconstructing corrupted inputs using noise functions such as sentence permutation and span masking, whereas mBERT and XLM-R predict randomly masked tokens, optimizing for representation learning. For **data**, mBART is trained on monolingual corpora from 25 languages, while BART [2] is restricted to a single-language dataset, and XLM-R leverages the broader CC100 dataset covering 100 languages to enhance cross-lingual representation learning.

► ***Reasons to learn a tokenizer on more languages.*** The authors trained the tokenizer on full CC data to **enhance generalization**, enabling fine-tuning on additional languages and supporting smoother transfer learning to languages not included in the pre-training phase. **The tokenizer is a good tool** because it allows the model to better capture **linguistic patterns specific to each language** while also learning **shared features among languages**, thereby improving its generalization capability. Additionally, a well-trained tokenizer supports the creation of more effective **embeddings**, as it ensures that semantically similar subwords across languages are represented in a unified manner.

► ***Comparison of effects of using high- and low-resource language pairs.*** High-resource language pairs often experience **diminished or slightly negative effects** from mBART pre-training. In contrast, **low-resource pairs benefit significantly**, as pre-training compensates for the lack of parallel data by leveraging **universal linguistic features** learned from monolingual corpora. This disparity may arise because **high-resource pairs are mainly task-specific parallel data**, reducing the impact of **generalizations**, whereas low-resource pairs, being **scarce**, rely on pre-training to **bridge data gaps**. Furthermore, in high-resource settings, negative effects in high-resource setting could also stem from being **unable to handle long dependencies** or **complex linguistic patterns**.

References

[1] Alexis Conneau and Guillaume Lample. Cross-lingual language model pretraining. *Advances in neural information processing systems*, 32, 2019.

[2] M Lewis. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, 2019.

[3] Y Liu. Multilingual denoising pre-training for neural machine translation. *arXiv preprint arXiv:2001.08210*, 2020.

[4] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

[5] Rico Sennrich. Improving neural machine translation models with monolingual data. *arXiv preprint arXiv:1511.06709*, 2015.

[6] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

[7] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016.