Reading Assignment 3

COMP 550, Fall 2024

Due date: November 21, 2024 9:00pm

Read the following paper and write a critical summary of it:

> Liu et al., 2020. Multilingual Denoising Pre-training for Neural Machine Translation. *TACL.* https://aclanthology.org/2020.tacl-1.47/

Your summary should include a description of the paper's contents, its strengths and limitations, and any points that you did not understand. Some aspects of this paper may not be covered in the lectures, I expect that there will be concepts that you do not understand! Do not worry, and document this in your write-up.

Also, discuss the following issues in your writeup:

1. What are the major differences in the pre-training of mBART vs. mBERT (or XLMR) in terms of architecture, pre-training task/noise function, and pre-training data?
2. Why did the authors learn a tokenizer on more languages than the languages covered in mBART25? Do you think this is a good idea? Discuss.
3. Do high-resource language pairs benefit from mBART pre-training at the fine-tuning stage? Contrast with low-resource language pairs and discuss.

Your summary should be **at most 1 page**. It will be graded on the basis of its coverage, linguistic quality, and clarity of argumentation.

Resources

- Jason Eisner's guide to reading papers: https://www.cs.jhu.edu/~jason/advice/how-to-read-a-paper.html
- It may help if you read the original BART paper before reading this https://aclanthology.org/2020.acl-main.703/