**Assignment 3: Abstractive Summarization**

**Due: November 17 11:59pm (AOE)**

1. Task: In this assignment you will be implementing an abstractive summarization system.

* Dataset: CNN/Daily Mail
	+ You may download the preprocessed version with

wget <https://cdn-datasets.huggingface.co/summarization/cnn_dm_v2.tgz>

* Input: a news article

*(CNN)A Duke student has admitted to hanging a noose made of rope from a tree near a student union, university officials said Thursday. The prestigious private school didn't identify the student, citing federal privacy laws. In a news release, it said the student was no longer on campus and will face student conduct review.*

*...*

* Output: a summary

*Student is no longer on Duke University campus and will face disciplinary review . School officials identified student during investigation and the person admitted to hanging the noose, Duke says . The noose, made of rope, was discovered on campus about 2 a.m.*

* Additional resources you may found useful:
* Download CNN\_STORIES\_TOKENIZED : <https://github.com/abisee/cnn-dailymail>
* Preprocess into non-tokenized cased sample format expected by BPE preprocessing: <https://github.com/artmatsak/cnn-dailymail>
* Fairseq (preprocessing for BART): <https://github.com/pytorch/fairseq/blob/main/examples/bart/README.summarization.md>
* Huggingface: <https://huggingface.co/datasets/cnn_dailymail>
* Tensorflow: <https://www.tensorflow.org/datasets/catalog/cnn_dailymail>
* Paperswithcode: <https://paperswithcode.com/dataset/cnn-daily-mail-1>
* Evaluation: Compute ROUGE scores on test set and report them in the final submission. You can run a ROUGE scorer at:

<https://github.com/pltrdy/files2rouge>

2. [5pts] You should implement at least one baseline system. Baselines to consider:

* (Sequence-to-sequence model with regular attention) Neural Machine Translation by Jointly Learning to Align and Translate

<https://arxiv.org/abs/1409.0473>

* Point-gen: pointer-generator model with copying mechanism added on seq2seq

<https://arxiv.org/abs/1704.04368>

* BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension

<https://arxiv.org/abs/1910.13461>

<https://github.com/pytorch/fairseq/tree/main/examples/bart>

Note: BART is expected to run on 1 node with 8 32gb-V100. Expected training time is about 5 hours.

3. [2pts] Possible Enhancements to consider:

* Add entity constraints as introduced in lecture 18 to deal with the hallucination problem, you can use the entity tagger you developed for assignment 1, or any open-source tools.
* Remove duplicates by adding constraints in decoding or adding semantic parsing constraints as introduced in lecture 18.
* Invent new evaluation metrics, by introducing more measures based on semantic similarity and informative content checking.

4. [3pts] Qualitative Analysis and Report Writing

* If you have done any successful enhancements to the baseline methods, report gains and concrete examples to explain why they worked.
* Look into your remaining errors, categorize remaining challenges and suggest possible solutions with concrete examples.

5. Submission:

* Please send your **{netid}.zip** to both TAs through email (qizeng2@illinois.edu and pengfei4@illinois.edu) and use “**CS546 ASSIGNMENT 3**” as the subject.
* Your zip file should include your **source code folder** and your **report pdf**. Any supplementary materials can be uploaded as external downloadable links.