CSE573 - NLP Applications
Assignment 2: Neural Machine Translation

Deadline
11:55pm, 29 February, 2020

Data:

The data for this assignment is available on the drive link **here**. It is a parallel corpus for the following languages: English Bangla Gujurati Hindi Konkani Malayalam Marathi Punjabi Tamil Telugu.
Anyone with the link can access the data.

Choose a pair (En-IL; where IL is one of [Bangla Gujurati Hindi Konkani Malayalam Marathi Punjabi Tamil Telugu]).

Keep in mind while choosing a pair that you have to also evaluate the performance by going through test data. Each pair has 3 sets of data- train, dev and test.

Task Description:

In this assignment, you are expected to implement the following papers, and check their performance on the test data:

1. Sequence to Sequence Learning with Neural Networks [**paper link**], which is the baseline.
2. Neural Machine Translation By Jointly Learning To Align And Translate [**paper link**], which is a basic attention model.
3. Effective Approaches to Attention-based Neural Machine Translation [**paper link**]. You are expected to implement the three global attention options, and you can ignore local attention.
4. Modeling Coverage for Neural Machine Translation [**paper link**].

You are free to choose the library to implement the above, and for those who do not have an ADA account, check out Google colabs [**link**] to use GPUs if needed.

There are a lot of tutorials to start with those libraries, and the official ones like **this** and **this** are a good resource. You are expected to learn and code them up on your own.

Evaluation metrics can be any of the popular metrics for machine translation [Eg: Bleu Score]. You are expected to have at least one.

A write-up should be submitted along with the code, where the implementation choices have to be explained, comparison of the models, and analysis of 5 sentences or more from the test data. There will be a viva at the time of evaluation.

The results that are reported should be on the test data provided.

Language:

Python 3.x

Evaluation:

20 - Paper 1
25 - Paper 2
25 - Paper 3
30 - Paper 4
20 - Complete report and analysis
(+20) Bonus: Based on any extra experiments
Total - 120(+20)

Submission:

On moodle, you are expected to submit the code [in the form of .py files] and the complete report. The code submitted should be in a single directory named 'src'.

The report should have the link to your google drive which should contain the the models that you would have trained for the results mentioned in your report, as the model's size is expected to be big.

Note:

- There will strictly be no extensions for this assignment as you will have to start working on the major project after this.
- Please make sure you start the work early, and submit it in time, keeping in mind the mid sem exams.
- You are expected to write your own code. Feel free to discuss the assignment with other students and collaborate on developing algorithms at a high level. Your report and the code that you submit must be yours. Plagiarism either from the internet or from other students will result in 0 on this assignment, and possibly more.
- Use the moodle thread for any queries that you might have.