Homework 2: Baseline

How to submit. Sumbissions are through the Courses page. For this homework, submit a PDF file. Include the path to your experiment directory on Rocket (execute `pwd` in the directory, copy the output). Don't forget to change permissions of your home directory, because other users cannot access it by default (execute `chmod -R 755 .` in your home directory, the directory where you end up when you have just logged in to Rocket). If you think any additional explanations are necessary to understand what you did, include them in the PDF as well.

Training an Estonian → English translation model

In this homework you will finally train an actual translation model.

START IN ADVANCE. Training may take several days. There might be queues for the GPUs as well. You will not lose points if your model has not finished training (i.e. validation perplexity has not yet stagnated for 3 consecutive checkpoints) by the deadline, but the translation quality has to be reasonably good by then. You might want to use email notifications (`#SBATCH --mail-type=ALL` and `#SBATCH --mail-user=your@email` in your SLURM script). Then you will have more chance to notice if your job fails, and if huge queues suddenly appear and you cannot finish your homework on time, you will have something to show how long you have been waiting for your jobs to start.

Note that you will need your trained model to do the next homework as well.

Task 1: Preprocessing (3 points)

The data is on Owncloud: https://owncloud.ut.ee/owncloud/index.php/s/DnXp9QXsCFN6LrH. It is a `.tar.gz` file, which contains 5 pairs of parallel files (e.g. `data_1.et` is parallel to `data_1.en`). The translation direction is from Estonian into English.

First, preprocess the data:

- concatenate and shuffle the data (do not forget that parallel lines have to remain parallel after you shuffle them),
- clean,
- separate a development set of **1,000** sentence pairs and a test set of **2,000** sentence pairs, all remaining sentence pairs go into the training set,
- truecase,
- SentencePiece (**32,000** subword units, **shared** between source and target sides).

Note that you should only use the **training set** to train your truecasing and SentencePiece models.

Task 2: Training (5 points)

Train a Sockeye translation model with the following specifications:

- each batch should consist of **7000 words**,
- initial learning rate is **0.001**,
- word embeddings should be of size **300**,
- vocabulary is **shared** between sourse and target sides,
- use **2** layers of **LSTMs** in both the encoder and the decoder,
- the hidden dimension of the LSTMs should be **512**,
- **dot** attention,

- a checkpoint should be created every **2000** batches,
- decode and evaluate the **whole** development set at each checkpoint,
- the minimum number of epochs is **3**,
- the model should stop training when the validation perplexity has not improved for **3** consecutive checkpoints.

Do not change any other hyperparameters, and do not add anything extra to the model. Use 1 GPU on Rocket.

Some practical hints:

- It will be more convenient if you do Sockeye's data preparation step before starting training.
- Use the parameter `--disable-device-locking`.
- `#SBATCH --mem=40GB` should be enough.
- If you want to view progress plots, install `mxboard` before training.

Task 3: Applying the model (2 points)

Translate the following sentence with the best checkpoint of your model:

Õppija põhiline eesmärk on teha üldistusi eelneva kogemustehulga põhjal.

Report the translated sentence in your submission.