Machine Translation: Project Exam

Mock shared-task competition

Sharid Loáiciga, PhD

The exam for this course has the form of a final research project. The final project should implement a MT system using the OpenNMT library and test it. The project is documented in a report that adheres to the standards of practice in computational linguistics.

1 WMT

WMT (`http://www.statmt.org/wmt20/`) organizes a yearly competition on machine translation of news stories. WMT provides training and test data for a variety of language pairs. A subset of these language pairs has been chosen based on your interests. The goal of this project is to train a neural machine translation model from scratch in a realistic setup and to practice to participate in the WMT competition.

The selected language pairs are:
English $\rightarrow$ German,
German $\rightarrow$ English,
German $\rightarrow$ French,
French $\rightarrow$ German,
Russian $\rightarrow$ English,
English $\rightarrow$ Russian.

Please select one of these language pairs for your experiments.

2 Data

You are allowed to use all parallel data available here for training:

`http://www.statmt.org/wmt20/translation-task.html#download`

`http://www.statmt.org/wmt19/translation-task.html#download`

For development, the datasets will be provided directly on Moodle. These datasets correspond to plain text versions of the news dev sets available at the same place. You will also be provided with the final test sets at a later date, one week before the final deadline.

No other datasets are allowed.

3 Modelling options

Your task is to train a model that performs as well as possible. This involves:

- careful preprocessing with sensible options (e.g., number of BPE units),
- choosing a good model architecture with good parameters,
- training the model for long enough,
- choosing useful training data.

I suggest that you to start with a simple model (e.g., one using the settings proposed for one of the assignments) as quickly as possible to have a baseline, and then gradually try out new options and check whether they improve BLEU scores on the development set.

One option that we did not cover in the assignments is the attention option. You can add this parameter to the usual command:

`-global_attention` the options are `general, dotprod, mlp`

4 Benchmarks

You can see the results obtained by leading research groups on `http://matrix.statmt.org/`. Make sure to select the appropriate test set.

Please do not hesitate to ask for help if you have any questions either by email or on Moodle.

5 Submission

Give your system a name. You will submit your translation of the test set on Moodle by the deadline, **August 20th, 2021**, and Sharid will report your BLEU scores.

You must also submit a report (PDF format, 6-10 pages) that describes your experiments. It is strongly recommended that you start taking notes from the beginning, detailing your modelling choices. Also, do not hesitate to discuss your failed experiments.

6 Evaluation

The evaluation will be based on the following criteria:

- Soundness and originality of the chosen methods as described in the report,
- BLEU scores obtained on the final test set,
- Quality of the analysis described in the report.