MTAT.06.055 Machine Translation: Projects

April 15, 2020
Project topics

Improving ET→EN model

Projects will be individual

If you hate this idea, you can choose your topic yourself (for example, you can try to replicate some paper that you find interesting). Then you have to discuss your topic with us (Lisa or Mark) separately.
First steps

- If your baseline model is well-trained:
  - BLEU on external test set around 15 or higher,
  - you used around 10M training sentence pairs,
  - output looks reasonable, etc.

  let your model **continue training** (for example, until it does not improve for 10 consecutive checkpoints)

- If it is not well-trained (very low BLEU, you had problems with preprocessing/training), **retrain** it properly
Next steps

Do several re-trainings with modifications. Do not just try random things, have some reasoning.

Expected process for each step would be:

1. Find a specific problem via manual analysis
2. Think: what kind of modifications could help?
3. Retrain your model with modifications
4. Look at sentences which had the errors you are trying to correct: are they better now?
5. Calculate new BLEU: did it improve?
Important tips

- It is better to introduce modifications one at a time
- BLEU is not the ultimate goal
- **Do manual analysis**
- Save all scripts and logs
- Using a Transformer model is probably a good idea
- Parallel corpora: [http://opus.nlpl.eu/](http://opus.nlpl.eu/) (baseline training data included parts of DGT, OpenSubtitles, JRC-Acquis, Europarl and EMEA corpora)
- If you are using new data, check it manually
Submissions

April 19: Project topic and brief outline

If you work on ET→EN system, you don’t have to think about the topic. Just describe your initial ideas (what you want to try) with some justification.
Questions?