UNIVERSITY OF TARTU

Machine Translation Shared Task

Team D

Mikhail Papkov
Elizaveta Korotkova

Institute of Computer Science

Repository:
git.io/vbobt

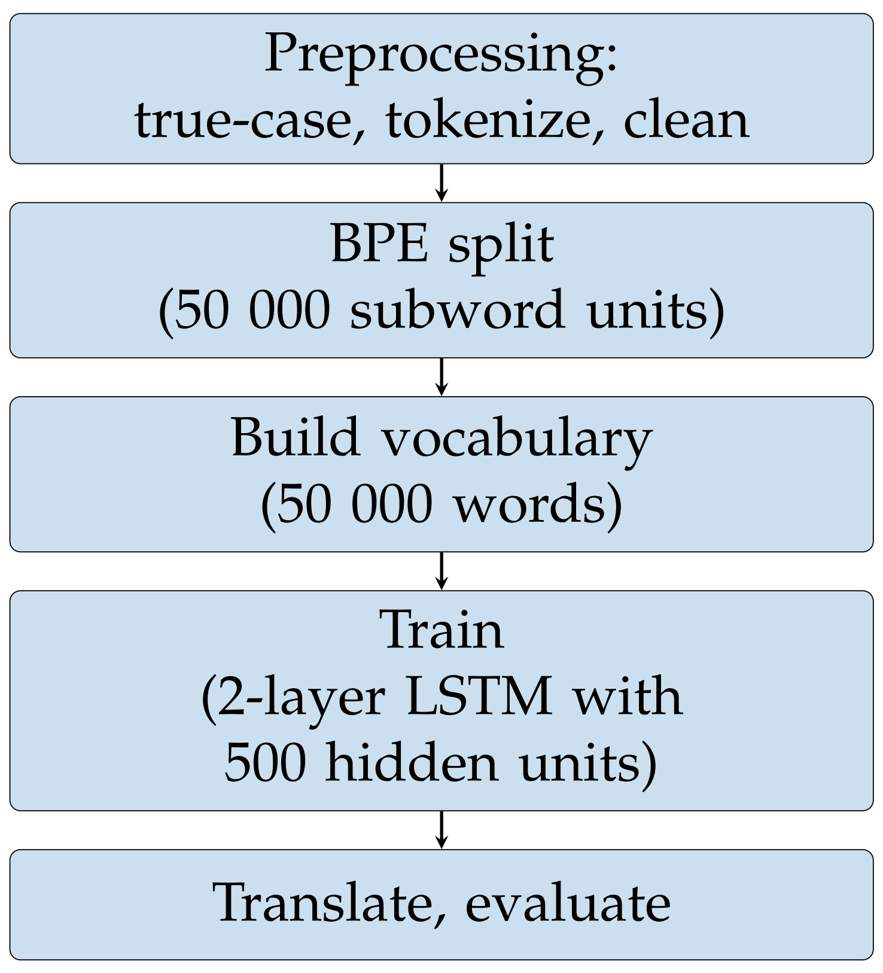
Contacts:
mikhail.papkov@gmail.com
elizaveta.korotkova@gmail.com

Abstract

This poster presents the results of our team's participation in the shared task, which was a part of the Machine Translation course at the Institute of Computer Science, University of Tartu, during the fall semester of 2017. We have built, analyzed and improved a neural machine translation system of Estonian into English.

Baseline Training

- OpenNMT framework [1]
- Standard preprocessing
- $\sim 19M$ sentence pairs total:
  - $\sim 16M$ training set
  - $\sim 2M$ test set
  - $\sim 1M$ development set
- Best model trained for 5 days, 13 epochs (development set perplexity 4.15)
- BLEU (on shared development set): **22.61**


Preprocessing:
true-case, tokenize, clean

BPE split
(50 000 subword units)

Build vocabulary
(50 000 words)

Train
(2-layer LSTM with
500 hidden units)

Translate, evaluate

Figure 1: NMT system pipeline

Solved Baseline Problems

Bad sentence splitting

Two sentences can appear on the same line in the parallel corpus. The system sees a dot (end of sentence) and does not translate the rest.

S: Meie veebisaidi sellest osast leiate teavet, kuidas korraldab parlament erinevate komisjonide süsteemi abil oma tööd. **Euroopa Parlamendi töö on oluline seepärast, et otsuseid uute Euroopa seaduste kohta teevad paljudes poliitikavaldkondades ühiselt parlament ja ministrite nõukogu, mis esindab liikmesriike.**

T: This part of our website will find information on how Parliament will organise its work through a system of different committees .

Fix: We used an additional regular expression to check for sentences which have not been split properly.

FT: This part of our website will find information on how Parliament will organise its work through different committees . **The work of the European Parliament is therefore important because decisions on new European laws do jointly with Parliament and the Council of Ministers , which represents the Member States .**

Missing parts of compound words

In compound words some roots were missing, supposedly because the system tends to translate words one-to-one. Compound words are quite frequent in Estonian, and this issue reappeared many times.

S: Surve käsitleda tõhusalt ja korrektselt nii **tööprotsesse** kui personaliküsimusi on tungiv.

R: The pressure to handle the **working processes** as well as staff questions in an effective and correct way is urgent .

T: Pressure to address effectively and correctly both **processes** and personnel issues is pressing .

Fix: We used `estnltk` [2] to split compound words longer than 10 characters into parts. Whitespaces were added between roots to force the system to see these roots as separate words and direct more attention to each of them.

FT: Pressure to deal effectively and correctly with both **work processes** and personnel issues .

Yet Unsolved Baseline Problems

Unnecessary future tense

This probably has to do with the lack of grammatical future tense in Estonian: the system cannot effectively differentiate between present and future tenses.

S: Euroopa Parlamendi mõju EL asjaajamises suureneb ning 1979. aastal saavad kodanikud esmakordselt parlamendi liikmeid otse valida.

R: The European Parliament increases its influence in EU affairs and in 1979 all citizens **can**, for the first time, **elect** their members directly .

T: The European Parliament's influence in the EU's business is increasing , and citizens **will be able to choose** directly from Parliament in 1979 .

Fix(?):

- Build "tense classifier" for Estonian sentences based on the tense of corresponding English translations in the train set.
- Add a tag for future to Estonian sentences (train set — true tag, test set — predicted) to emulate grammatical tense.

We achieved **83%** classification accuracy but did not retrain the system with the obtained tags.

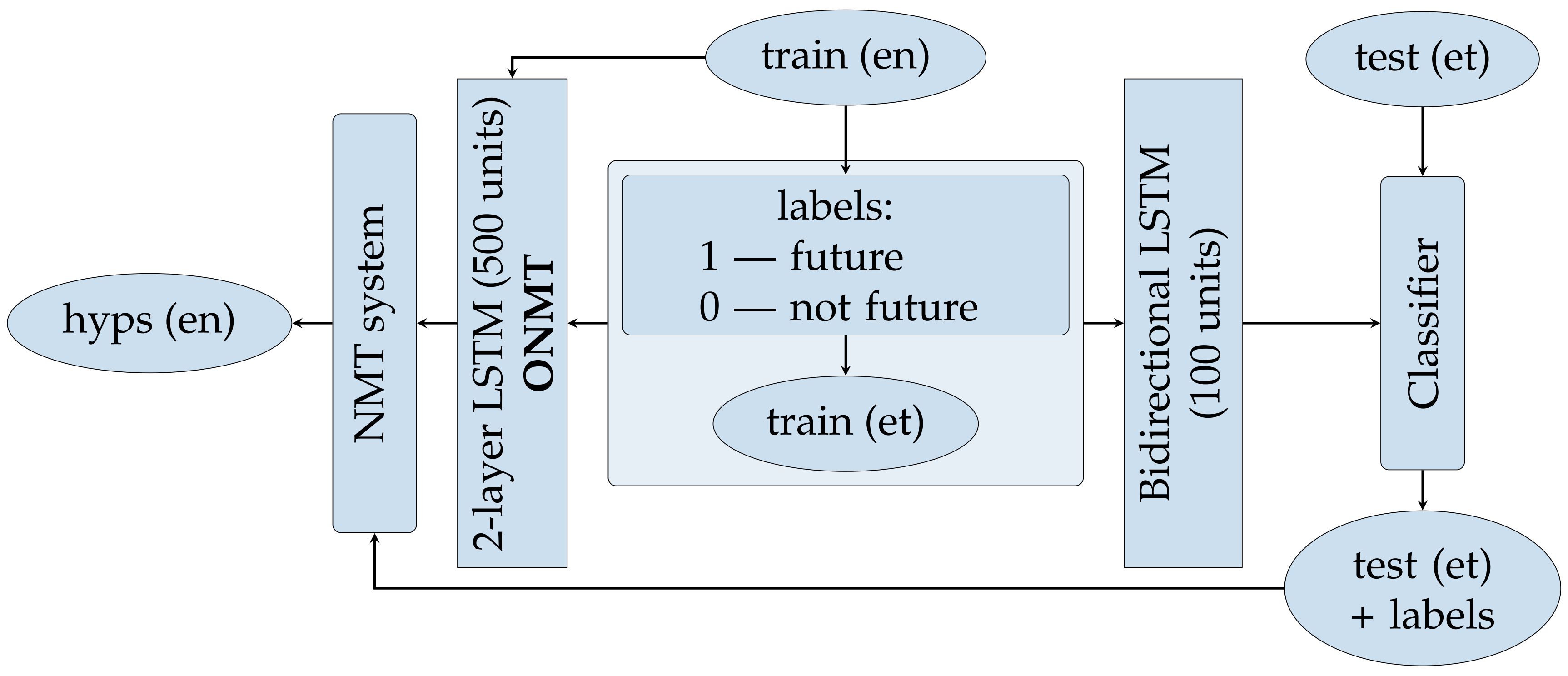


Figure 2: Pipeline for annotation with future tense labels

Lack of special knowledge and terms

The system does not know technical terms, for example, in physics.

S: **Kvarkide** kohta üteldakse, et nad on **vangis**.

R: **Quarks** are said to be **confined** .

T: They say **they're in prison** .

Fix(?): Add more specialized texts to the training set.

Results

We have built, evaluated and improved an Estonian-English neural machine translation system. BLEU score (**22.54**) did not improve over the baseline system (**22.61**), but manual evaluation shows that the mistakes we targeted were fixed in many cases.

References

[1] G. Klein et al. "OpenNMT: Open-Source Toolkit for Neural Machine Translation". In: *ArXiv e-prints* (). eprint: 1701.02810.

[2] Siim Orasmaa et al. "EstNLTK - NLP Toolkit for Estonian". In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. May 2016.