Neural Machine Translation

Authors: Dmytro Tkachuk, Raman Shapaval and Yevhen Tyshchenko

UNIVERSITY OF TARTU
Institute of Computer Science

Project Goal

The project aim was to improve the BLEU score of the baseline neural machine translation system using any of the approaches, and available tools and packages.

Data

The shared dataset consists of 19.051.439 Estonian-English parallel sentence pairs.

Corpus preparation

- **Corpora concatenating**
- **Data shuffling**
- **Data splitting:** 18976439 training examples, 50000 test examples, and 25000 development examples
- **Data tokenization**
- **Data true-casing**
- **Byte-pair encoding**

Model Training

- We used 1 Tesla P100 GPU Machine provided by HPC center of the University of Tartu to train our model with vocabulary of size 50000. The model we trained is the default OpenNMT-py model, which consists of a 2-layer LSTM with 500 hidden units on both the encoder/decoder.
- In addition to this we spent some time for dropout tuning and tried next values: 0.3, 0.5, 0.7.
- We trained our models for approximately 5 days each.
- Lastly, we have also trained Sockeye model (1-layer bi-LSTM encoder, 1-layer LSTM decoder with dot attention).

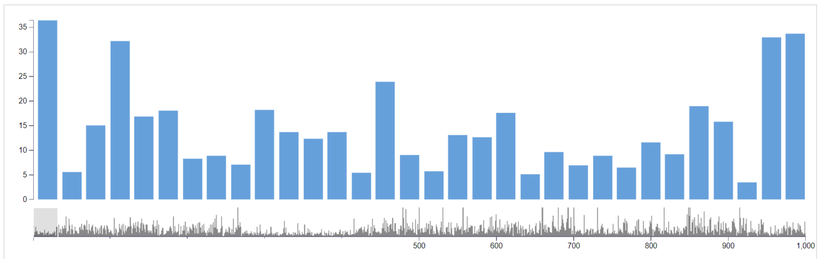
Results

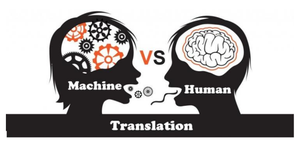
Best model BLEU score

BLEU:	24.56
Precision x brevity:	26.22 x 93.65

Type	1-gram	2-gram	3-gram	4-gram
Individual	61.57	32.82	19.54	11.98
Cumulative	57.66	42.10	31.89	24.56

Export data: CSV





Challenges

The main difficulties are connected to the hardware we have to work with (rocket.hpc.ut.ee) as we always suffered from being queued for a long time, which then resulted in time losses for model training and evaluating of generated translations.

We are not claiming that model tuning process is insufficient but it is always has high time cost which often is not payed with appropriate improvements.

Conclusion

- Increased vocabulary size along with dropout tuning improved our previous BLEU score from 21.14 to 24.56 points.
- The more data you feed to the model - the better results you achieve.
- Unfortunately, sockeye model did not show expected improvements.

Manual Evaluation

BEFORE

Human	100.00	1.00	As a result, the EU schemes for educational exchanges and trans-border partnerships like Erasmus and Leonardo are bywords among students and other learners.
Machine	4.04	0.74	Thanks to this, every student and other student are aware of Erasmus, Leonardo and other cross-border partnerships.

Human	100.00	1.00	Even the traditional website is dying, killed by the relentless force of constant streaming information from social networks.
Machine	16.02	0.83	Even traditional websites die, which has killed a continuous flow of information from social networks.

Human	100.00	1.00	The two sides also explored positions ahead of the international talks on climate change in Copenhagen in December.
Machine	10.71	0.72	Two sides discussed the views of the international climate change debate in Copenhagen.

Human	100.00	1.00	Latvia is represented in the Arab Republic of Egypt by Ambassador Maris Selga.
Machine	39.65	1.23	Mr Maris Selga, who is represented in the Arab Republic of Egypt, is represented by Latvia.

Human	100.00	1.00	To understand the magnitude of this number, 1 TeV is the approximate amount of energy used for motion… by a mosquito in flight.
Machine	5.55	0.91	In order to understand the size of this figure, 1 TEN is about this energy, which is used by flying mosquito.

AFTER

Sentence 198	BLEU	Length ratio	Text
Human	100.00	1.00	As a result , the EU schemes for educational exchanges and trans-border partnerships like Erasmus and Leonardo are bywords among students and other learners .
Machine	4.34	0.80	Thanks to this , every student and other learner know the Erasmus , Leonardo and the other cross-border partnerships .

Sentence 488	BLEU	Length ratio	Text
Human	100.00	1.00	Even the traditional website is dying , killed by the relentless force of constant streaming information from social networks .
Machine	21.80	1.15	Even the traditional website dies , which has been killed by the constant force of the continuous information voiced by social networks .

Sentence 602	BLEU	Length ratio	Text
Human	100.00	1.00	The two sides also explored positions ahead of the international talks on climate change in Copenhagen in December .
Machine	30.80	0.84	Two sides discussed the views of the international climate change debate in Copenhagen in December .

Sentence 643	BLEU	Length ratio	Text
Human	100.00	1.00	Latvia is represented in the Arab Republic of Egypt by Ambassador Maris Selga .
Machine	41.19	0.93	Latvia is represented by Maris Selga since the Arab Republic of Egypt .

Sentence 707	BLEU	Length ratio	Text
Human	100.00	1.00	To understand the magnitude of this number , 1 TeV is the approximate amount of energy used for motion… by a mosquito in flight .
Machine	18.27	0.96	In order to understand the size of this size , 1 TEV is about the energy that is used to move a mosquito .