Training and Adapting Multilingual NMT for Less-resourced and Morphologically Rich Languages

Matīss Rikters, Mārcis Pinnis, Rihards Krišlauks
Tilde
{matiss.rikters, marcis.pinnis, rihards.krislauks}@tilde.lv

Problem

Estonian ↔ Russian

- Both are morphologically rich languages
- A rather small parallel corpus
- Larger corpora available for language pairs including Estonian ↔ English and Russian ↔ English

Model Configuration

	RNN			FConv	Transformer
	MLSTM	GRU			
		Shallow	Deep		
Subword unit vocabulary	25 000 tokens	50 000 tokens			
Layers	1 encoder, 1 decoder	1 encoder, 1 decoder	4 encoder, 4 decoder	15 encoder, 15 decoder	6 encoder, 6 decoder
Maximum sentence length	50			128	

Data

Language pair	Before filtering (Total/Unique)	After filtering (Unique)
En↔Et	62.5M / 24.3M	18.9M
En↔Ru	60.7M / 39.2M	29.4M
Ru↔Et	6.5M / 4.4M	3.5M

Results

Poster

We analyse:

- Translation quality between different NMT architectures (MLSTM, GRU, FConv, and Transformer)
- Translation quality between one-way (U) and multi-way (M) NMT models, deep (D) and shallow (S) NMT models
- Performance in terms of training time, translation speed and resource usage

Resource Usage

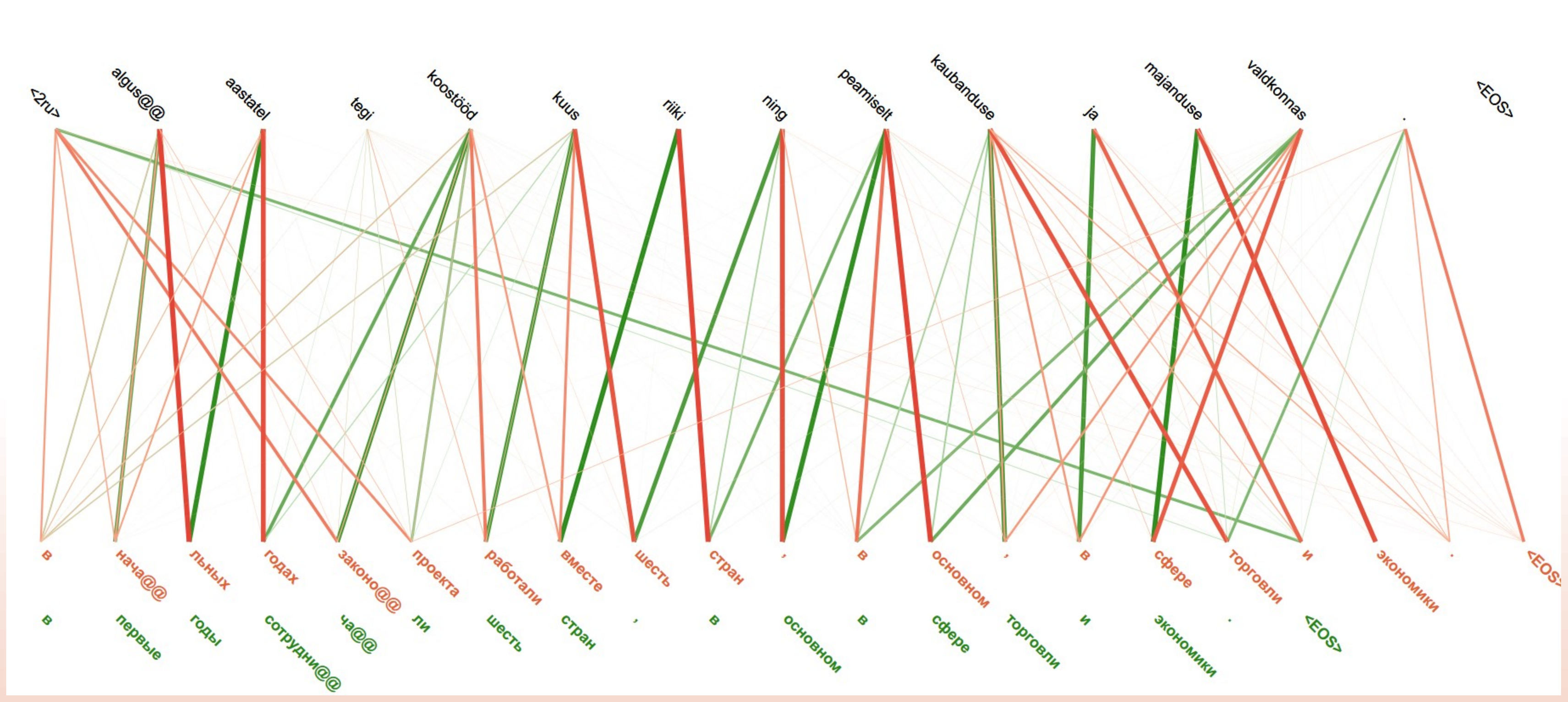
	Seconds		Sentences per second	GPU RAM, MB	Train time, days
	Translation	Per sentence			
Theano-based Nematus					
MLSTM-SM	274.57	0.54	1.86	651	16.4
GRU-SM	211.51	0.41	2.42	**611**	8.5
GRU-DM	460.07	0.9	1.11	979	36.6
MXNet-based Sockeye					
FConv-M	177.19	**0.35**	**2.89**	971	4.5
Transformer-M	191.05	0.37	2.68	1391	**3.8**

Translation Quality

	Development				Test			
	Ru→Et	Et→Ru	En→Et	Et→En	Ru→Et	Et→Ru	En→Et	Et→En
SMT	*27.74*	*25.48*	17.99	25.89	9.88	7.27	21.44	29.69
MLSTM-SU	17.51	18.46	23.79	34.45	11.11	12.32	26.14	36.78
GRU-SM	13.7	13.71	17.95	27.84	10.66	11.17	19.22	27.85
GRU-DU	17.03	17.42	23.53	33.63	10.33	12.36	25.25	36.86
GRU-DM	17.07	17.93	23.37	33.52	13.75	14.57	25.76	36.93
FConv-U	15.24	16.17	21.63	33.84	7.56	8.83	24.87	36.96
FConv-M	14.92	15.8	18.99	30.25	10.65	10.99	21.65	31.79
Transformer-U	17.44	18.9	**25.27**	**37.12**	9.10	11.17	**28.43**	**40.08**
Transformer-M	**18.03**	**19.18**	23.99	35.15	**14.38**	**15.48**	25.56	37.97

Example

Source	Algusaastatel tegi koostööd kuus riiki ning peamiselt kaubanduse ja majanduse valdkonnas.
GRU-DU	в начальных годах законопроекта работали вместе шесть стран, в основном, в сфере торговли и экономики.
(translated into English)	In the initial years of the bill project, six countries worked together, mainly in the sphere of trade and economy.
GRU-DM	в первые годы сотрудничали шесть стран, в основном в сфере торговли и экономики.
(translated into English)	In the first years, six countries cooperated, mainly in the sphere of trade and economy.
Reference	в первый год сотрудничество вели шесть стран , в основном в сфере торговли и экономики.
English Reference	In the early years, the cooperation was between six countries and mainly about trade and the economy.



Conclusions

- Low-resource language pairs benefit in translation quality from adding other language data.
- Multi-way NMT systems improved translation quality for all architectures - deep GRU, FConv, and Transformer.
- The largest improvements and highest overall BLEU scores were achieved using the Transformer model.
- The multi-way approach degraded performance for high-resource language pairs by several BLEU points.
- The most stable NMT architecture for multi-way model training was the deep GRU model, showing improvements for both low-resource and high-resource language pairs on both development and evaluation data sets.
- When training one-way systems for the low-resource language pairs, Fconv and Transformer models under-performed and the best results were achieved by the MLSTM-based models.

Acknowledgements

In accordance with the contract No. 1.2.1.1/16/A/009 between the "Forest Sector Competence Centre" Ltd. and the Central Finance and Contracting Agency, concluded on 13th of October, 2016, the study is conducted by Tilde Ltd. with support from the European Regional Development Fund (ERDF) within the framework of the project "Forest Sector Competence Centre".

NACIONĀLAIS ATTĪSTĪBAS PLĀNS 2020

EIROPAS SAVIENĪBA
Eiropas Reģionālās attīstības fonds

IEGULDĪJUMS TAVĀ NĀKOTNĒ