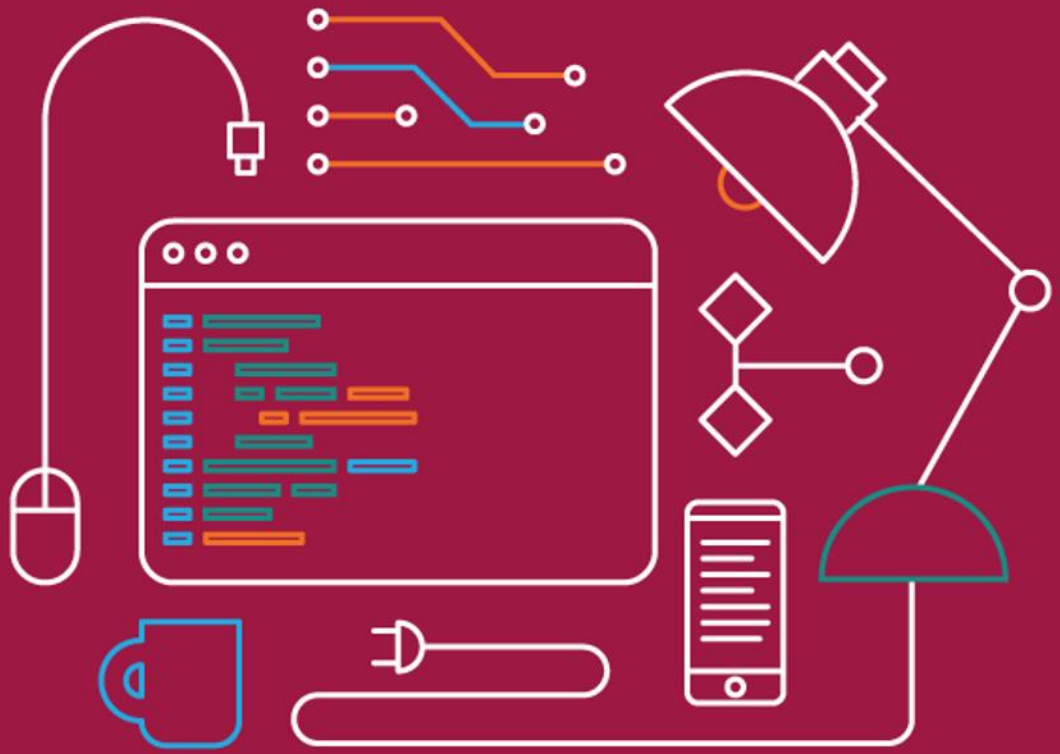


MULTILINGUAL AND MORPHOLOGY-AWARE NMT

Dr. Raivis SKADIŅŠ

Tilde, Director of Research and Development

Second Finnish Workshop on Machine Translation, Helsinki, November 1, 2017

In my talk

- How the WMT 2017 Shared Task was Won (EN-LV-EN)
- What we do with Estonian (Multilingual NMT, CNN etc.)

How the WMT 2017 Shared Task was Won

Data

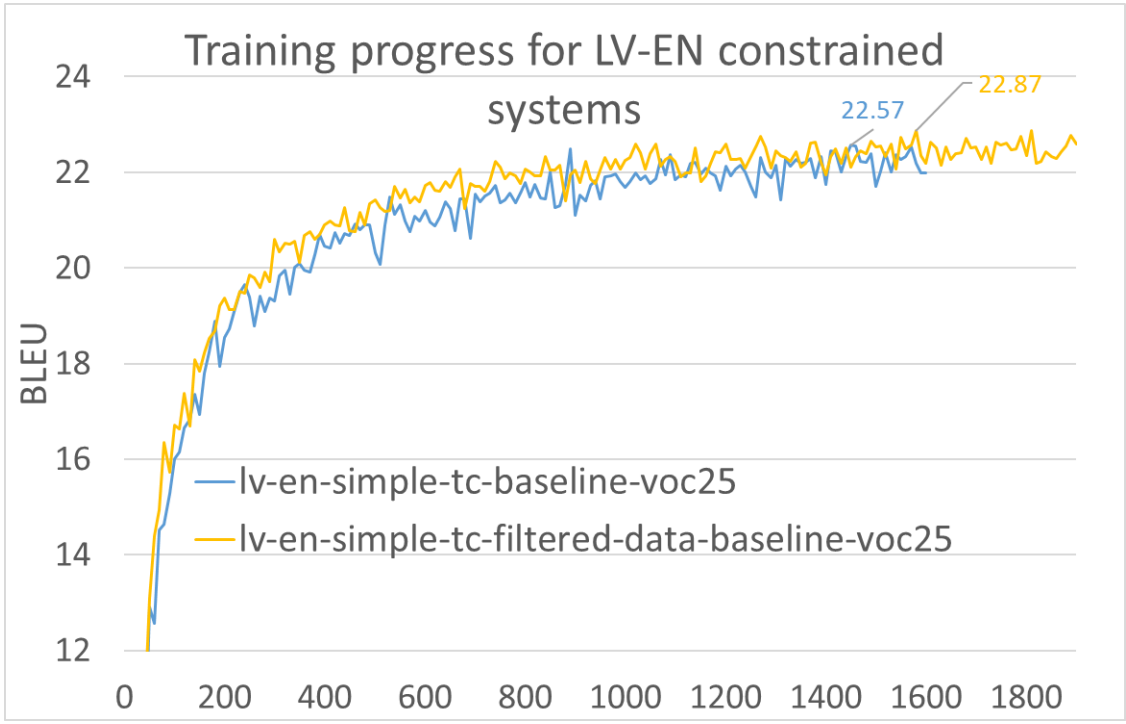
- WMT official training data for constrained systems
  - Parallel 4.51M
  - Mono: 27.75M (LV), 330.23M (EN)
- Data was noisy
  - we filtered out a lot
  - only 1.61M left
- For unconstrained systems
  - Parallel: 12.69M
  - Mono: 81.68M (LV), 351.99M (EN).
- For back-translation
  - random 1.61M sentences from the mono corpus

Preprocessing

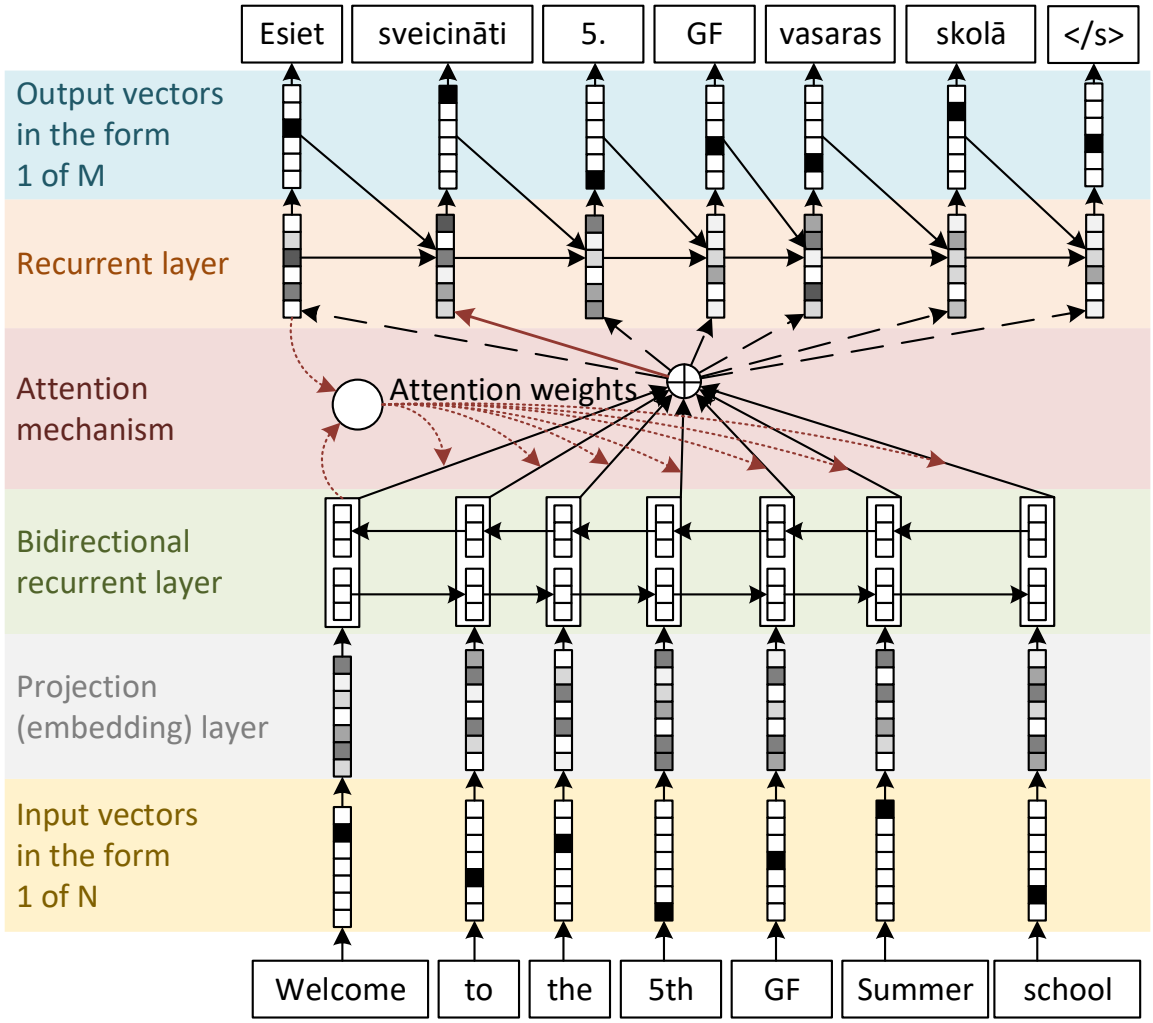
- Normalization (-, “, «, ‘, etc.)
- Non-translatable tokens
  - emails → $EMAIL$
  - urls, filenames → URL
  - XML tags → XML
- Tokenization
- Truecasing

Data filtering

- Original 4.5M
- Baseline: 1.74M
- Filtered: 1.61M



Baseline NMT architecture

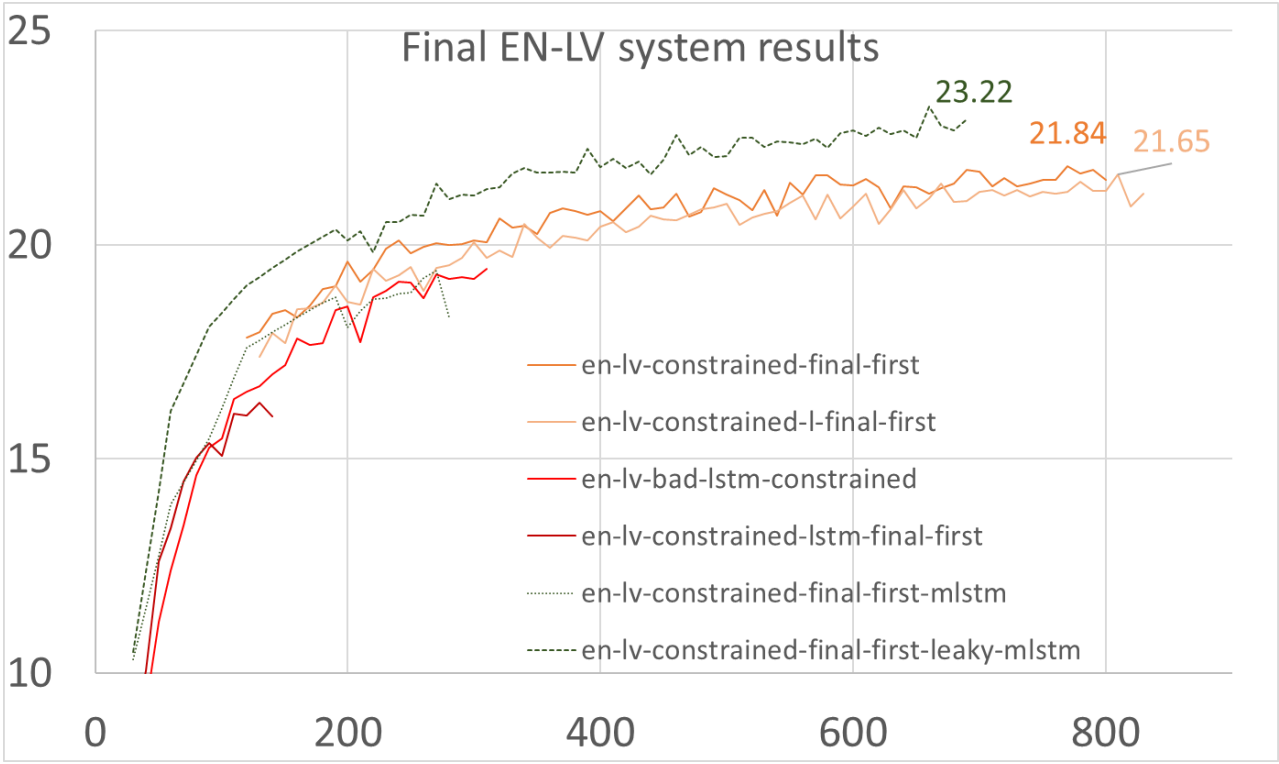


- Nematus and AmuNMT toolkits
- end-to-end NMT
- sub-word tokens (BPE)

Recurrent Units

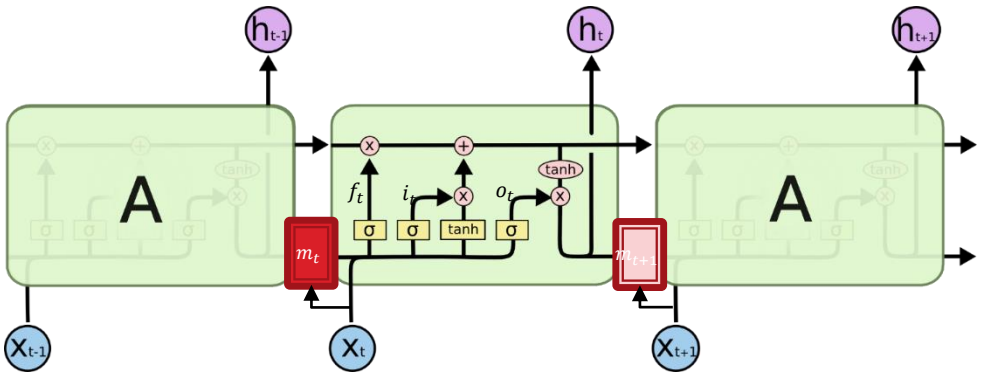
- GRU
- LSTM
- Multiplicative LSTM

- **MLSTM won**



Multiplicative LSTM

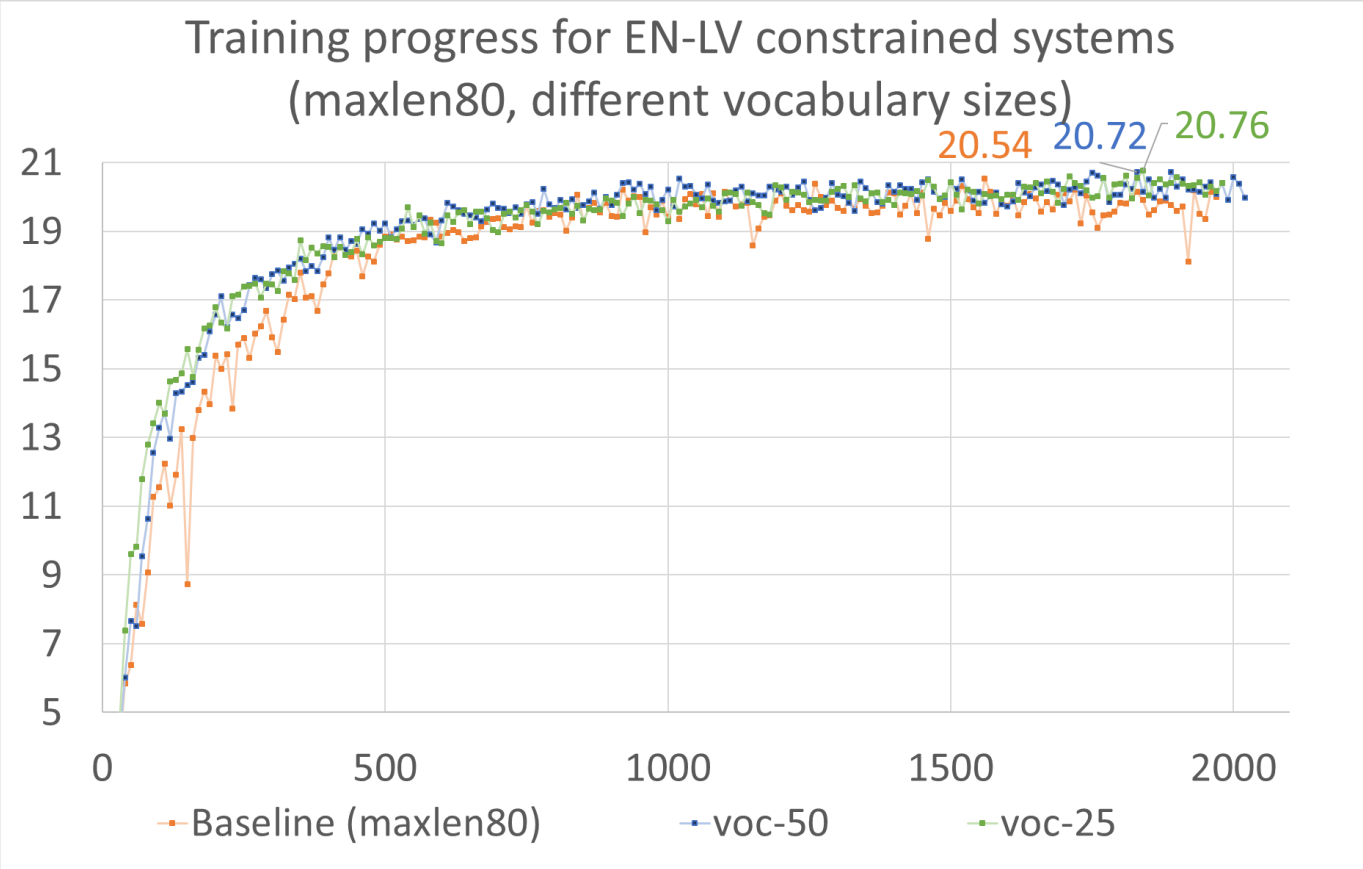
- Slightly modified LSTM
  - intermediate result m_t is used to calculate parameters for f_t, i_t, o_t
    LSTM uses the previous result h_{t-1}



$$m_t = W_m \cdot h_{t-1} \cdot x$$

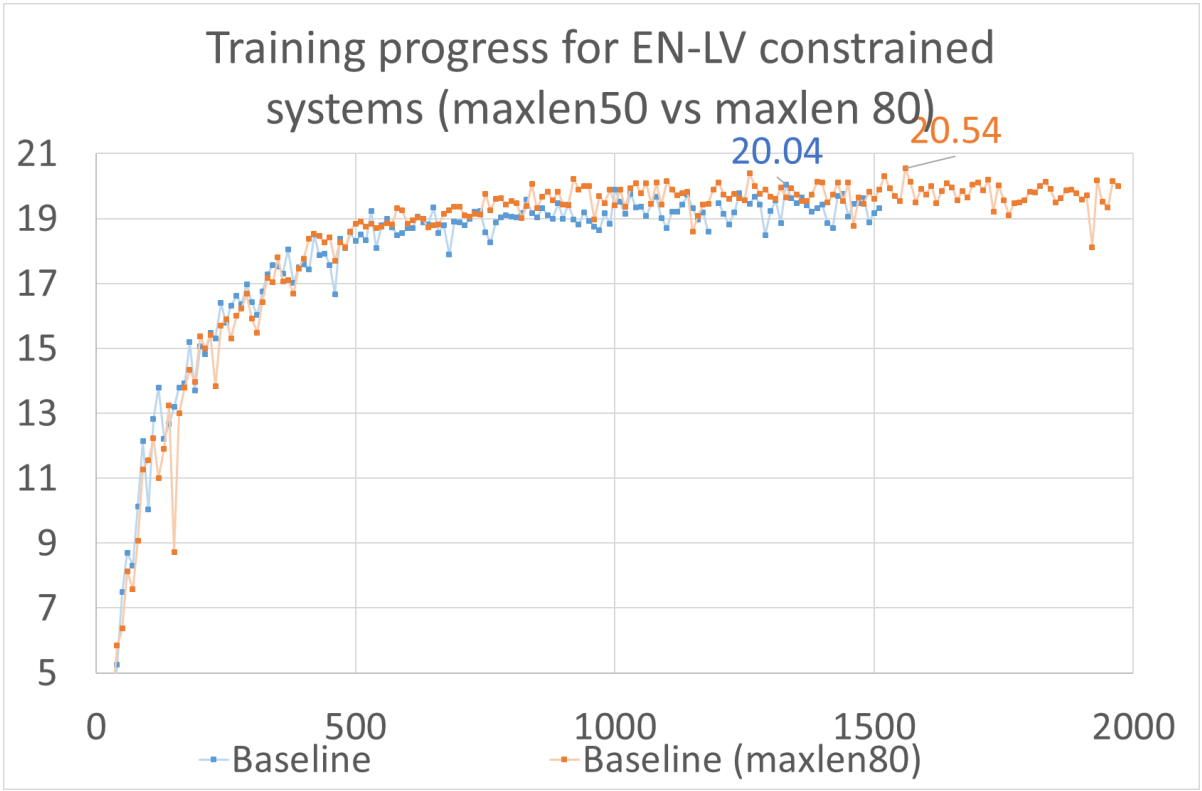
Vocabulary size

- We tried 25K, 50K and 100K
- but we should not



Context length (maxlen)

- We tried 50 and 80
- Not that important



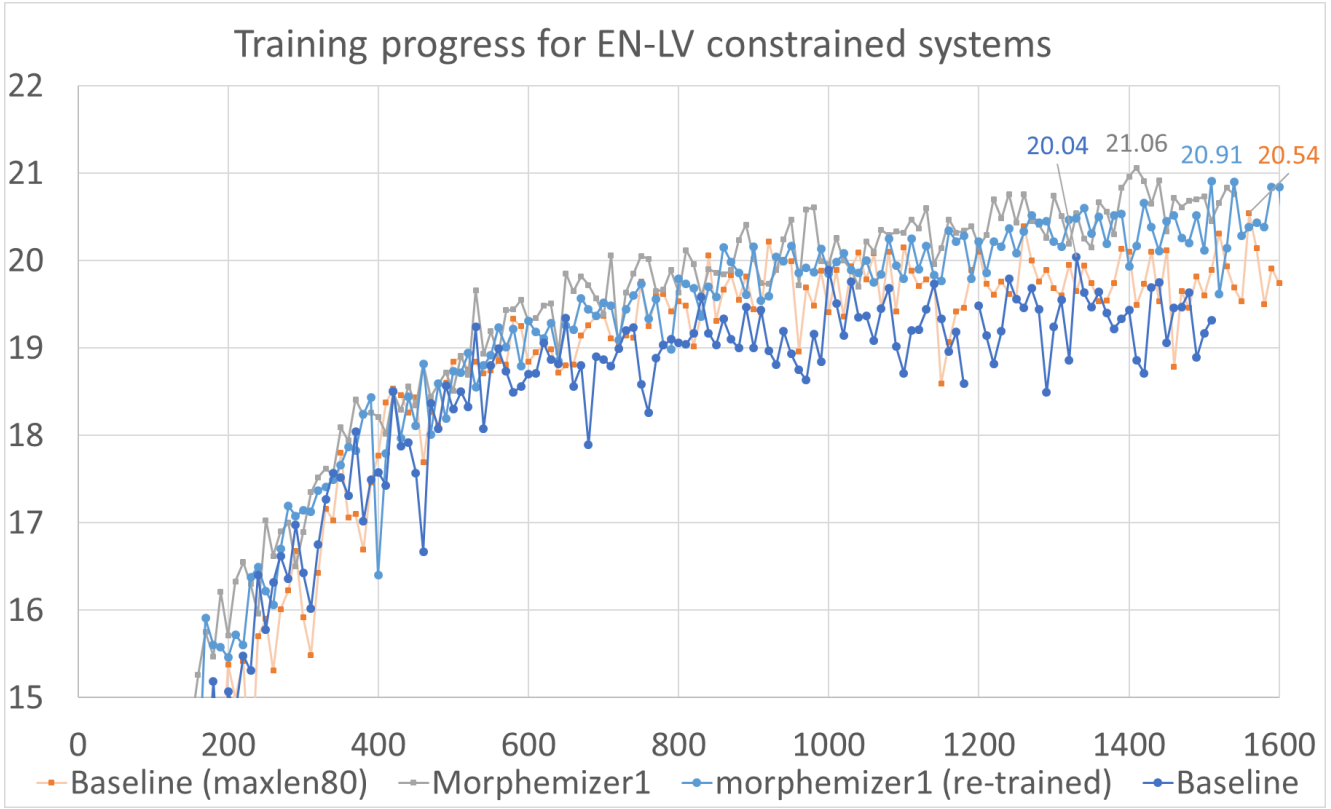
Sub-word units

- We used morphological analyzer to split off
  - prefixes
  - endings
  - compounds

Word	BPE	MWS
English		
legalization	leg@@ alization	legal@@ ization
legalize	leg@@ alize	legal@@ ize
legalized	leg@@ alized	legal@@ iz@@ ed
legalizes	legaliz@@ es	legal@@ izes
legalizing	legaliz@@ ing	legal@@ iz@@ ing
Latvian (”atbalss” is translated as ”echo”)		
atbalss	at@@ balss	atbals@@ s
atbalsis	atbal@@ sis	atbals@@ is
atbalsi	atbal@@ si	atbals@@ i
atbalsīs	at@@ bals@@ īs	atbals@@ īs
atbalsīm	at@@ balsīm	atbals@@ īm

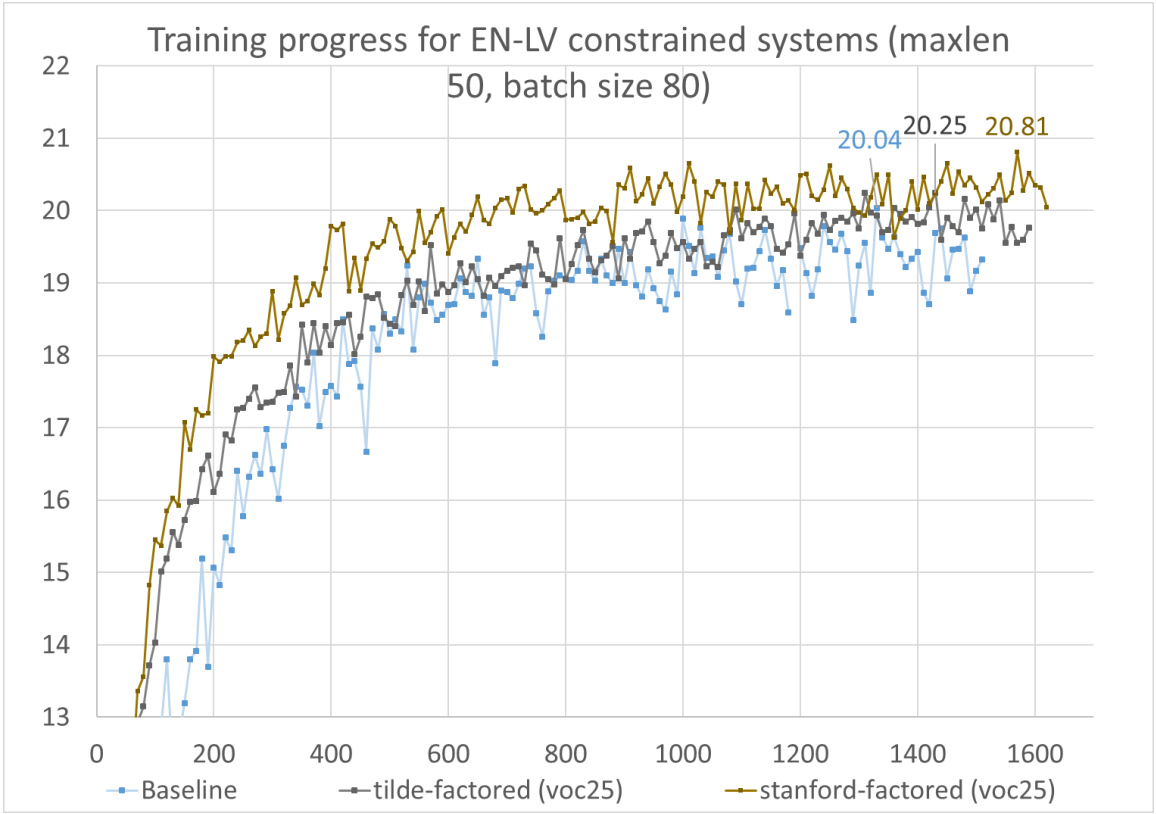
Sub-word units

- Morphology helps



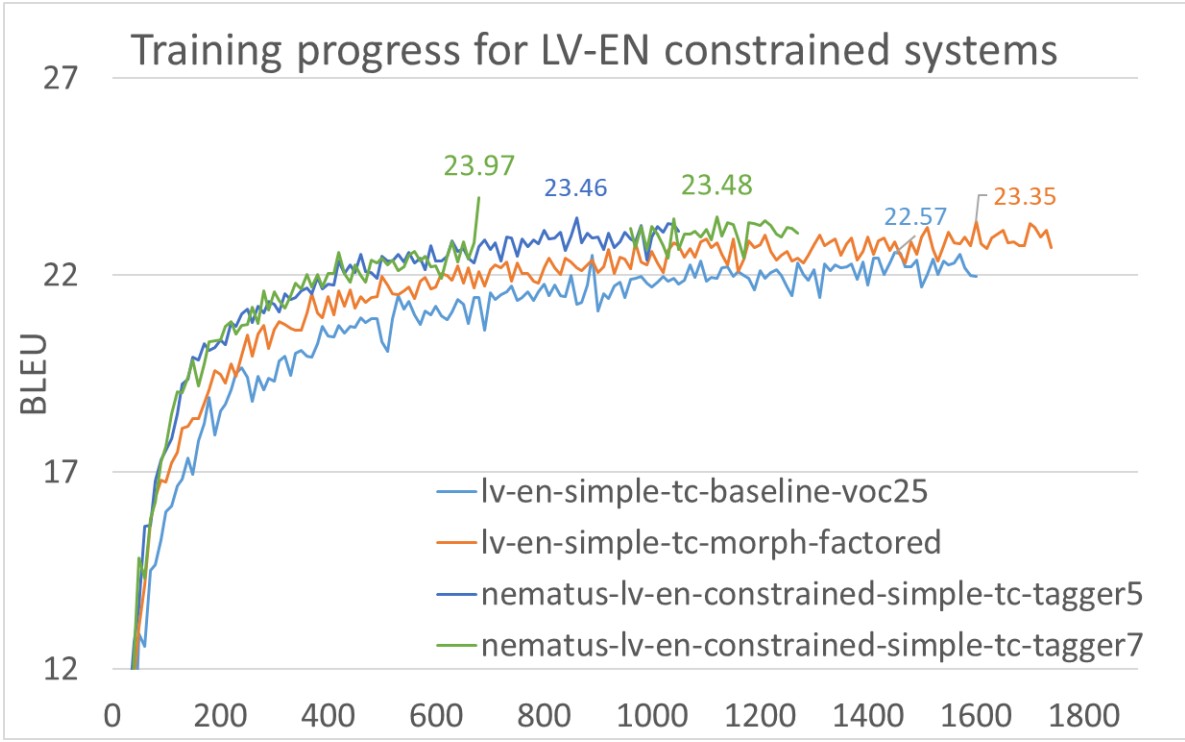
Factored models

- EN-LV
  Stanford parser



Factored models

- LV-EN
  Latvian POS tagger



NMT does not like rare words

- Rare named entities cause NMT to output unpredictable output
- We replaced all rare words with ID in training data
- NMT learned to leave them untranslated
- We translated ID with SMT or left untranslated

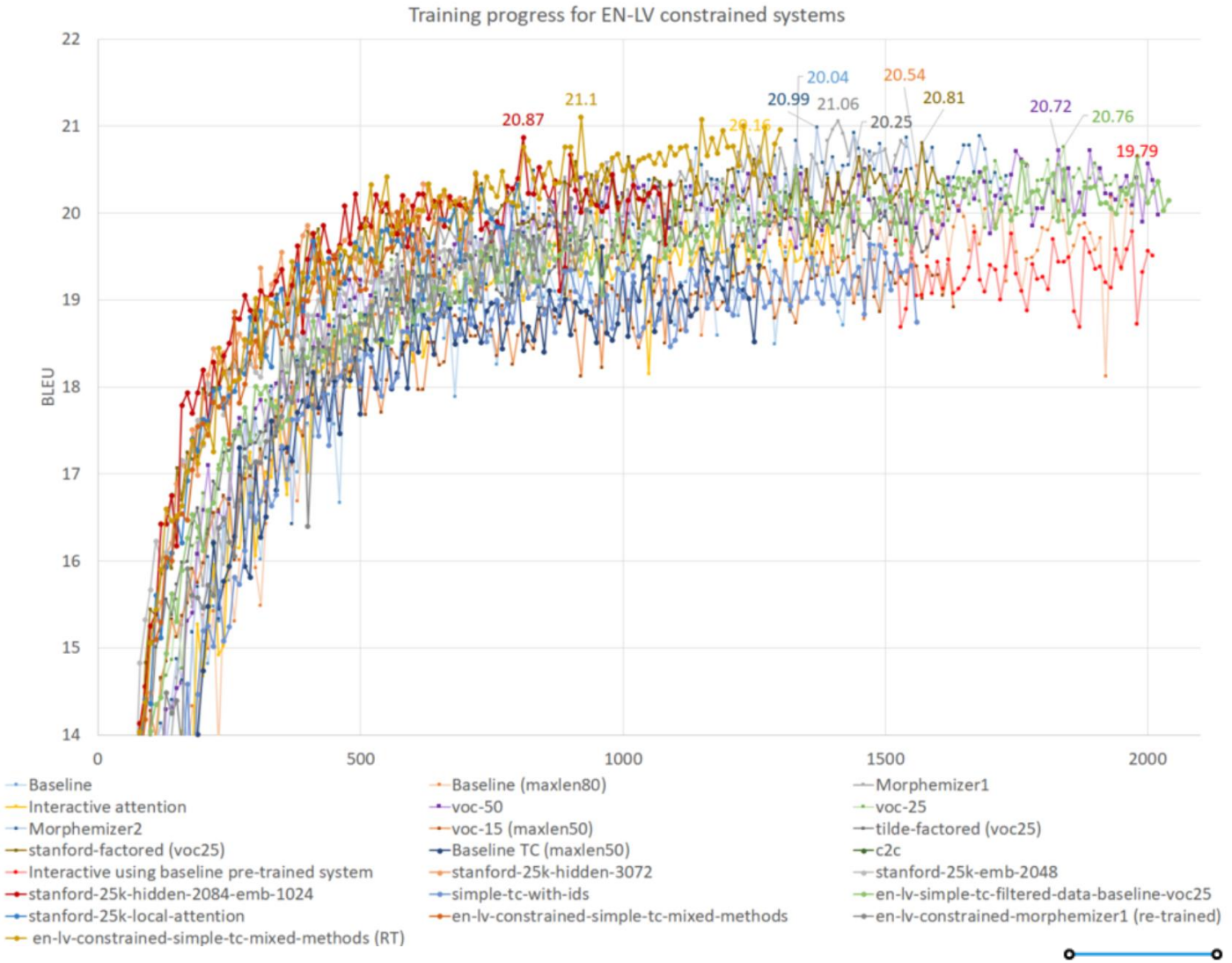
NMT-SMT Hybrid

- Impact on BLEU was minimal

Translation step	Example sentence
Source text	Šodien skatieties **Ikaunieces-Admidiņas** startu Rio spēlēs.
Pre-processed text	šodien skat@@ ieties **I@@ kaun@@ iec@@ es - Ad@@ mi@@ di@@ ņas** start@@ u Rio spēlē@@ s .
Text with identified rare words	šodien skat@@ ieties **βIDβ - βIDβ** start@@ u Rio spēlē@@ s .
NMT translation	watch the **βIDβ - βIDβ** start at the Rio Games today .
Moses XML with untranslated rare words	<nmt translation="watch the">šodien skatieties </nmt>**Ikaunieces** <nmt translation="-">-</nmt>**Admidiņas** <nmt translation="start at the Rio Games today">šodien startu Rio spēlēs</nmt><nmt translation=".">.</nmt>
Moses XML with identified untranslated person names	<nmt translation="watch the">šodien skatieties </nmt><**ne translation="Ikauniece" prob="1.0">Ikaunieces</ne**> <nmt translation="-">-</nmt><**ne translation="Admidina||Admidins" prob="0.95||0.05">Admidiņas</ne**> <nmt translation="start at the Rio Games today">šodien startu Rio spēlēs</nmt><nmt translation=".">.</nmt>
SMT translation	watch the **Ikauniece - Admidina** start at the Rio Games today .
Post-processed translation	Watch the **Ikauniece-Admidina** start at the Rio Games today.
NMT only transl. (for comparison)	Today, look at the start of the **Isolence-Admidias** in the Rio Games.

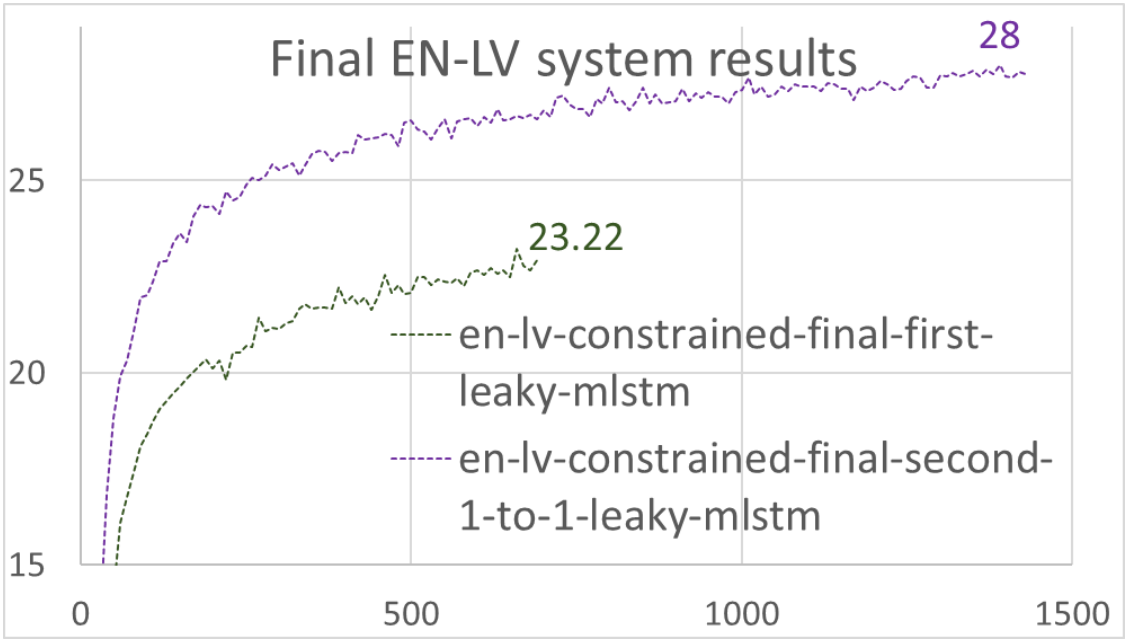
The final system

- We merged everything together

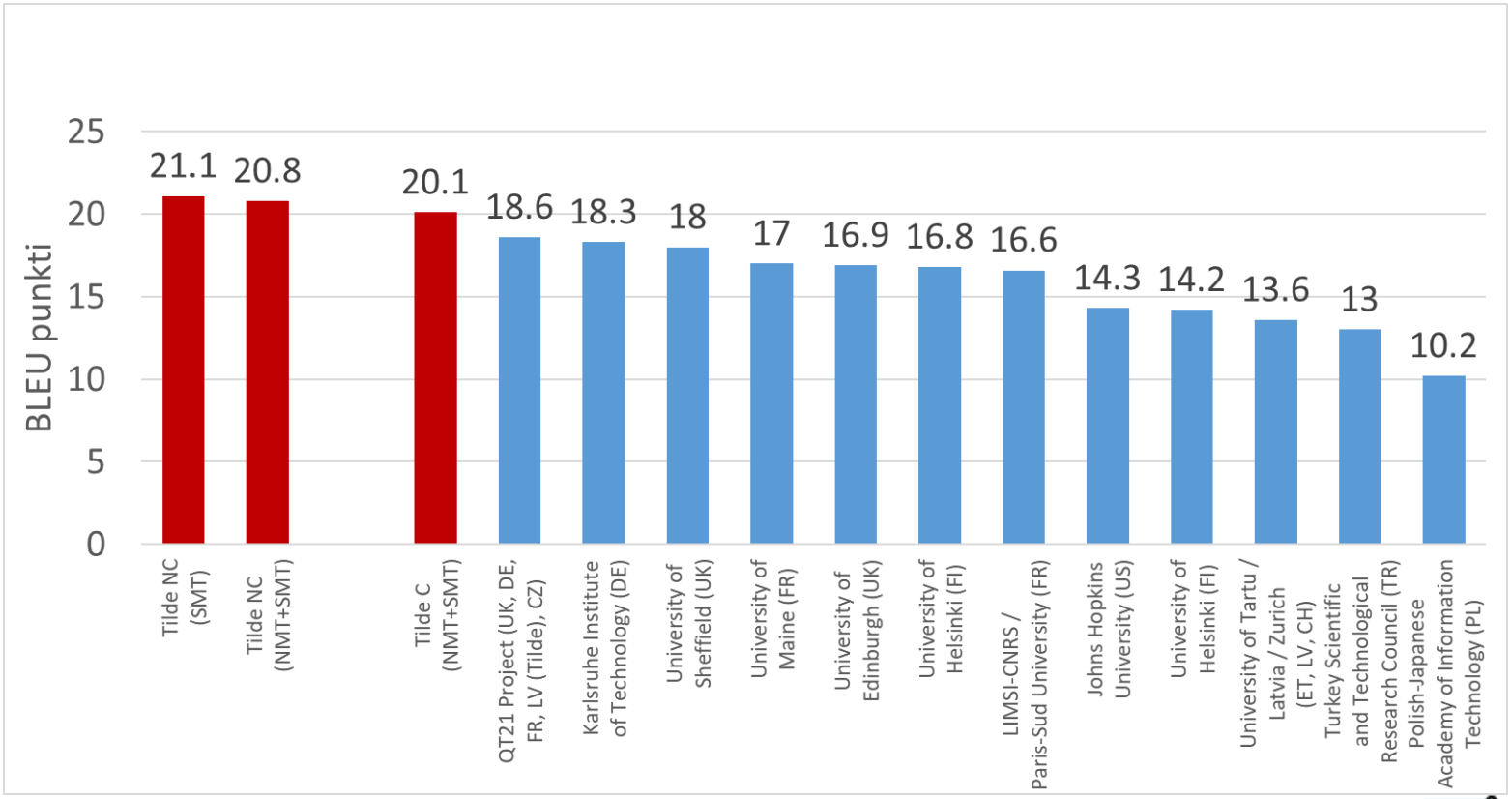


Back-translation

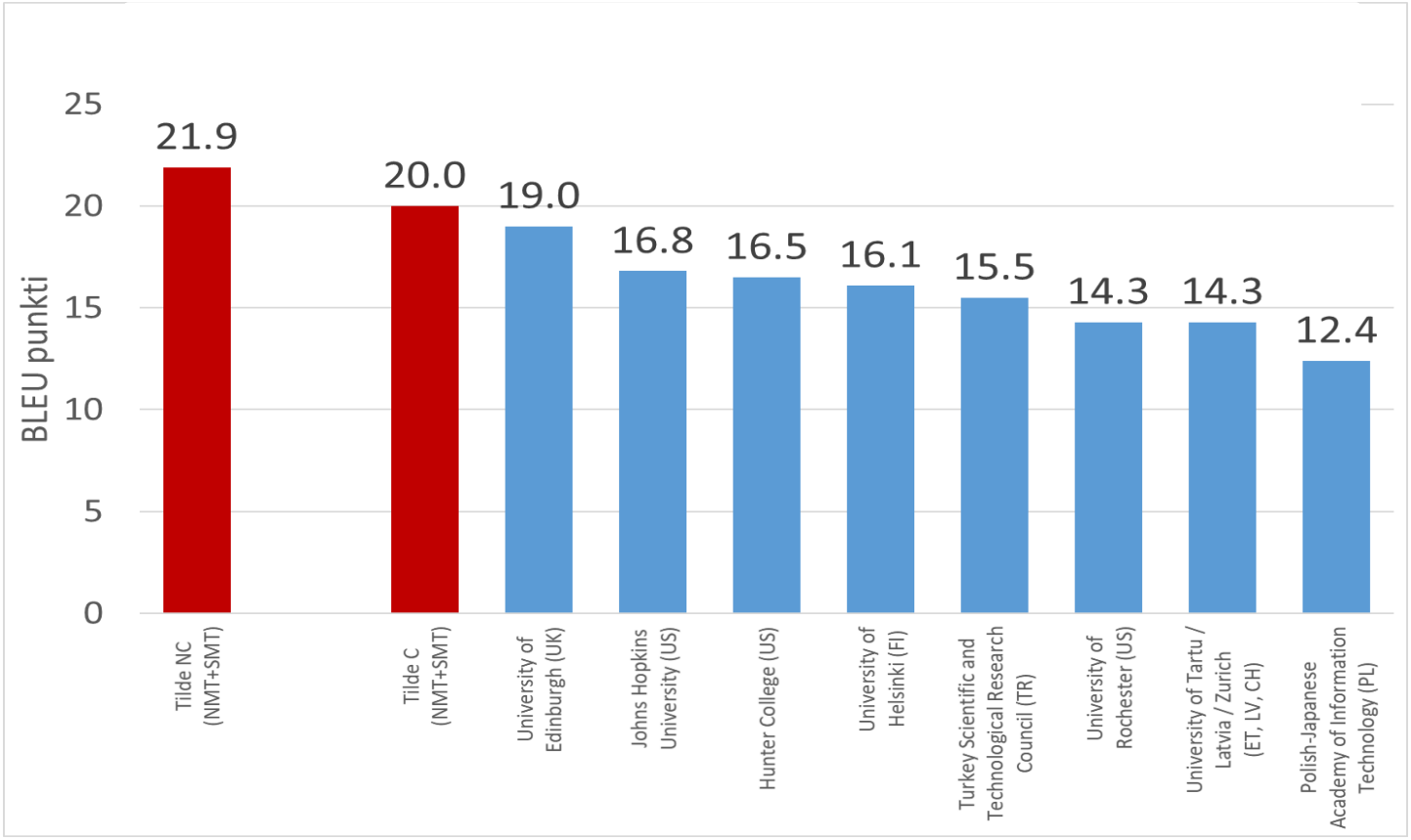
- It works
- Probably for everybody ☺



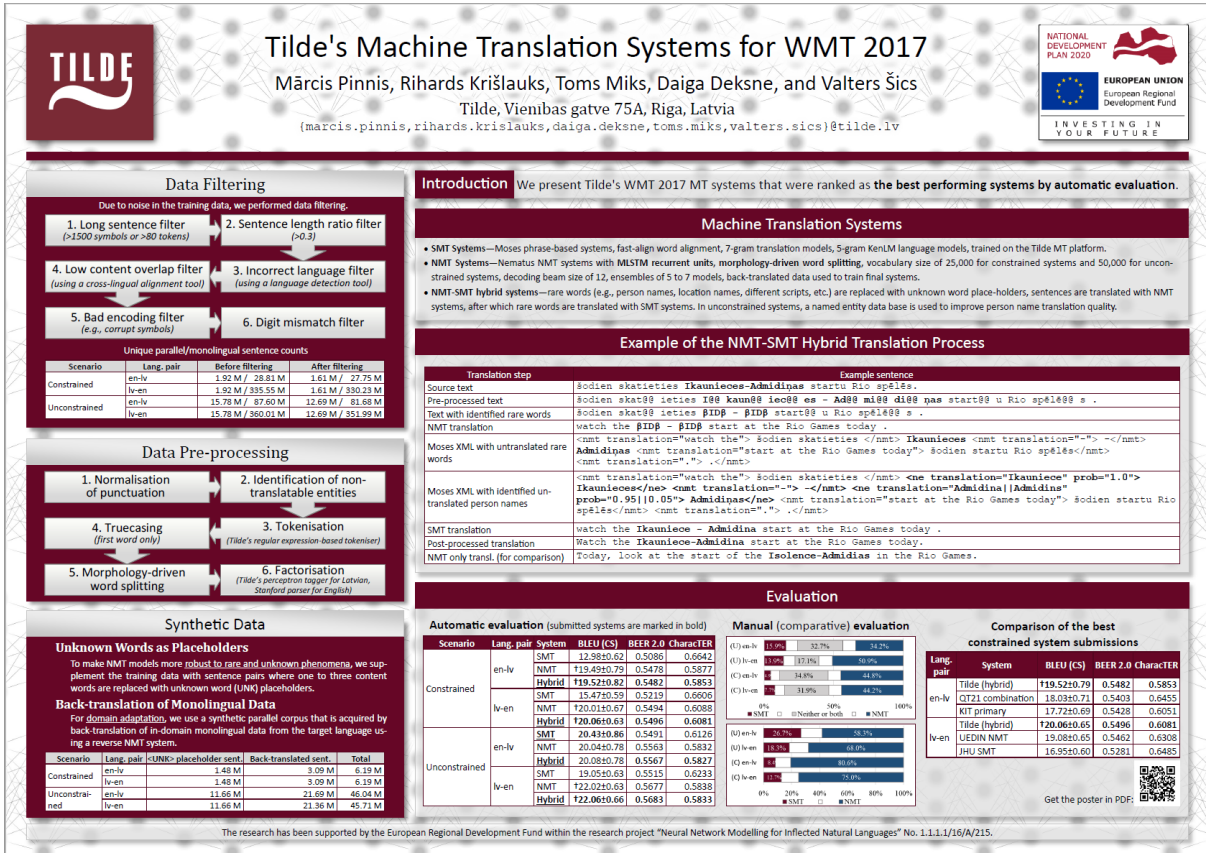
WMT 2017 EN-LV



WMT 2017 LV-EN



- ~55 NMT systems trained
- ~25 SMT systems trained



Tilde's Machine Translation Systems for WMT 2017

Mārcis Pinnis, Rihards Krišlauks, Toms Miks, Daiga Deksne, and Valters Šics
Tilde, Vienības gatve 75A, Riga, Latvia
{marcis.pinnis,rihards.krislauks,daiga.deksne,toms.miks,valters.sics}@tilde.lv

NATIONAL DEVELOPMENT PLAN 2020 · EUROPEAN UNION European Regional Development Fund · INVESTING IN YOUR FUTURE

Data Filtering

Due to noise in the training data, we performed data filtering.

1. Long sentence filter (>1500 symbols or >80 tokens)
2. Sentence length ratio filter (>0.3)
3. Incorrect language filter (using a language detection tool)
4. Low content overlap filter (using a cross-lingual alignment tool)
5. Bad encoding filter (e.g., corrupt symbols)
6. Digit mismatch filter

Unique parallel/monolingual sentence counts

Scenario	Lang. pair	Before filtering	After filtering
Constrained	en-lv	1.92 M / 28.81 M	1.61 M / 27.75 M
	lv-en	1.92 M / 335.55 M	1.61 M / 330.23 M
Unconstrained	en-lv	15.78 M / 87.60 M	12.69 M / 81.68 M
	lv-en	15.78 M / 360.01 M	12.69 M / 351.99 M

Data Pre-processing

1. Normalisation of punctuation
2. Identification of non-translatable entities
3. Tokenisation (Tilde's regular expression-based tokeniser)
4. Truecasing (first word only)
5. Morphology-driven word splitting
6. Factorisation (Tilde's perceptron tagger for Latvian, Stanford parser for English)

Synthetic Data

Unknown Words as Placeholders

To make NMT models more robust to rare and unknown phenomena, we supplement the training data with sentence pairs where one to three content words are replaced with unknown word (UNK) placeholders.

Back-translation of Monolingual Data

For domain adaptation, we use a synthetic parallel corpus that is acquired by back-translation of in-domain monolingual data from the target language using a reverse NMT system.

Scenario	Lang. pair	<UNK> placeholder sent.	Back-translated sent.	Total
Constrained	en-lv	1.48 M	3.09 M	6.19 M
	lv-en	1.48 M	3.09 M	6.19 M
Unconstrained	en-lv	11.66 M	21.69 M	46.04 M
	lv-en	11.66 M	21.36 M	45.71 M

Introduction

We present Tilde's WMT 2017 MT systems that were ranked as **the best performing systems by automatic evaluation**.

Machine Translation Systems

- **SMT Systems**—Moses phrase-based systems, fast-align word alignment, 7-gram translation models, 5-gram KenLM language models, trained on the Tilde MT platform.
- **NMT Systems**—Nematus NMT systems with **MLSTM recurrent units**, **morphology-driven word splitting**, vocabulary size of 25,000 for constrained systems and 50,000 for unconstrained systems, decoding beam size of 12, ensembles of 5 to 7 models, back-translated data used to train final systems.
- **NMT-SMT hybrid systems**—rare words (e.g., person names, location names, different scripts, etc.) are replaced with unknown word place-holders, sentences are translated with NMT systems, after which rare words are translated with SMT systems. In unconstrained systems, a named entity data base is used to improve person name translation quality.

Example of the NMT-SMT Hybrid Translation Process

Translation step	Example sentence
Source text	Šodien skatieties **Ikaunieces-Admidiņas** startu Rio spēlēs.
Pre-processed text	šodien skat@@ ieties **Ī@@ kaun@@ iec@@ es - Ad@@ mi@@ di@@ ņas** start@@ u Rio spēl@@ s .
Text with identified rare words	šodien skat@@ ieties **βIDβ - βIDβ** start@@ u Rio spēl@@ s .
NMT translation	watch the **βIDβ - βIDβ** start at the Rio Games today .
Moses XML with untranslated rare words	<nmt translation="watch the"> šodien skatieties </nmt> **Ikaunieces** <nmt translation="-"> -</nmt> **Admidiņas** <nmt translation="start at the Rio Games today"> šodien startu Rio spēlēs</nmt> <nmt translation="."> .</nmt>
Moses XML with identified untranslated person names	<nmt translation="watch the"> šodien skatieties </nmt> **<ne translation="Ikauniece" prob="1.0"> Ikauniece</ne> <nmt translation="-"> -</nmt> <ne translation="Admidina||Admidins" prob="0.95||0.05"> Admidiņas</ne>** <nmt translation="start at the Rio Games today"> šodien startu Rio spēlēs</nmt> <nmt translation="."> .</nmt>
SMT translation	watch the **Ikauniece - Admidina** start at the Rio Games today .
Post-processed translation	Watch the **Ikauniece-Admidina** start at the Rio Games today.
NMT only transl. (for comparison)	Today, look at the start of the **Isolence-Admidias** in the Rio Games.

Evaluation

Automatic evaluation (submitted systems are marked in bold)

Scenario	Lang. pair	System	BLEU (CS)	BEER 2.0	CharacTER
Constrained	en-lv	SMT	12.98±0.62	0.5086	0.6642
		NMT	↑19.49±0.79	0.5478	0.5877
		Hybrid	**↑19.52±0.82**	**0.5482**	**0.5853**
	lv-en	SMT	15.47±0.59	0.5219	0.6606
		NMT	↑20.01±0.67	0.5494	0.6088
		Hybrid	**↑20.06±0.63**	**0.5496**	**0.6081**
Unconstrained	en-lv	**SMT**	**20.43±0.86**	0.5491	0.6126
		NMT	20.04±0.78	0.5563	0.5832
		Hybrid	20.08±0.78	**0.5567**	**0.5827**
	lv-en	SMT	19.05±0.63	0.5515	0.6233
		NMT	↑22.02±0.63	0.5677	0.5838
		Hybrid	**↑22.06±0.66**	**0.5683**	**0.5833**

Manual (comparative) **evaluation**

Comparison of the best constrained system submissions

Lang. pair	System	BLEU (CS)	BEER 2.0	CharacTER
en-lv	Tilde (hybrid)	**↑19.52±0.79**	**0.5482**	**0.5853**
	QT21 combination	18.03±0.71	0.5403	0.6455
	KIT primary	17.72±0.69	0.5428	0.6051
lv-en	Tilde (hybrid)	**↑20.06±0.65**	**0.5496**	**0.6081**
	UEDIN NMT	19.08±0.65	0.5462	0.6308
	JHU SMT	16.95±0.60	0.5281	0.6485

Get the poster in PDF:

The research has been supported by the European Regional Development Fund within the research project "Neural Network Modelling for Inflected Natural Languages" No. 1.1.1.1/16/A/215.

- (Pinnis et al., 2017)

Multilingual NMT

- Some language pairs are really under resourced (ET-RU)
- Inspired from Google Zero-Shot NMT

Multilingual NMT

	Development				Test			
	RU-ET	ET-RU	EN-ET	ET-EN	RU-ET	ET-RU	EN-ET	ET-EN
SMT					12.52	14.74	22.53	32.52
MLSTM Shallow	17.51	18.46	23.79	34.45	11.11	12.32	26.14	36.78

Multilingual NMT

	Development				Test			
	RU-ET	ET-RU	EN-ET	ET-EN	RU-ET	ET-RU	EN-ET	ET-EN
SMT					12.52	14.74	22.53	32.52
MLSTM Shallow	17.51	18.46	23.79	34.45	11.11	12.32	26.14	36.78
MLSTM Shallow Multiling.	<10	<10	<10	<10				

Multilingual NMT

	Development				Test			
	RU-ET	ET-RU	EN-ET	ET-EN	RU-ET	ET-RU	EN-ET	ET-EN
SMT					12.52	14.74	22.53	32.52
MLSTM Shallow	17.51	18.46	23.79	34.45	11.11	12.32	26.14	36.78
MLSTM Shallow Multiling.	<10	<10	<10	<10				
GRU Shallow	13.70	13.71	17.95	27.84	10.66	11.17		
GRU Deep	17.03	17.42			10.33	12.36		

Multilingual NMT

	Development				Test			
	RU-ET	ET-RU	EN-ET	ET-EN	RU-ET	ET-RU	EN-ET	ET-EN
SMT					12.52	14.74	22.53	32.52
MLSTM Shallow	17.51	18.46	23.79	34.45	11.11	12.32	26.14	36.78
MLSTM Shallow Multiling.	<10	<10	<10	<10				
GRU Shallow	13.70	13.71	17.95	27.84	10.66	11.17		
GRU Deep	17.03	17.42			10.33	12.36		
GRU Deep Multiling.	17.07	**17.93**	23.37	33.52	**13.75**	**14.57**	25.76	**36.93**

Fully Convolutional NMT

- Latest experiments with Estonian-English

Method	BLEU
MLSTM-SU	36.78
GRU-DM	36.93
SMT	32.52
Fully convolutional	**40.54**

Acknowledgements & related papers

- The research has been supported by the European Regional Development Fund within the research project "Neural Network Modelling for Inflected Natural Languages" No. 1.1.1.1/16/A/215.
- The research has been supported by The Ministry of Education and Research (*https://www.hm.ee/en*) and by The National Programme for Estonian Language Technology (*https://www.keeletehnoloogia.ee/en*).

Pinnis, M., Krišlauks, R., Deksne, D., & Miks, T. (2017). Neural Machine Translation for Morphologically Rich Languages with Improved Sub-word Units and Synthetic Data. In Proceedings of the 20th International Conference of Text, Speech and Dialogue (TSD2017). Prague, Czechia.

Pinnis, M., Krišlauks, R., Miks, T., Deksne, D., & Šics, V. (2017). Tilde's Machine Translation Systems for WMT 2017. In Proceedings of the Second Conference on Machine Translation (WMT 2017), Volume 2: Shared Task Papers (pp. 374–381). Copenhagen, Denmark: Association for Computational Linguistics. Retrieved from http://www.aclweb.org/anthology/W17-4737

Pinnis, M., Krišlauks, R., Deksne, D., & Miks, T. (2017). Evaluation of Neural Machine Translation for Highly Inflected and Small Languages. In Proceedings of the 18th International Conference on Intelligent Text Processing and Computational Linguistics (CICLING 2017). Budapest, Hungary.