

Improving Neural Machine Translation

Mateusz Krubiński

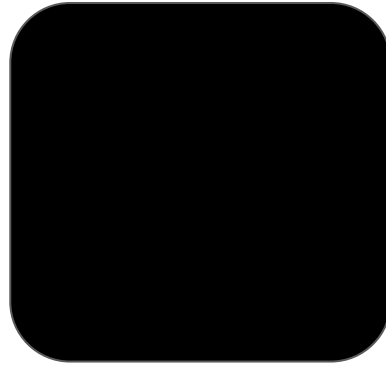


Charles University
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics



unless otherwise stated

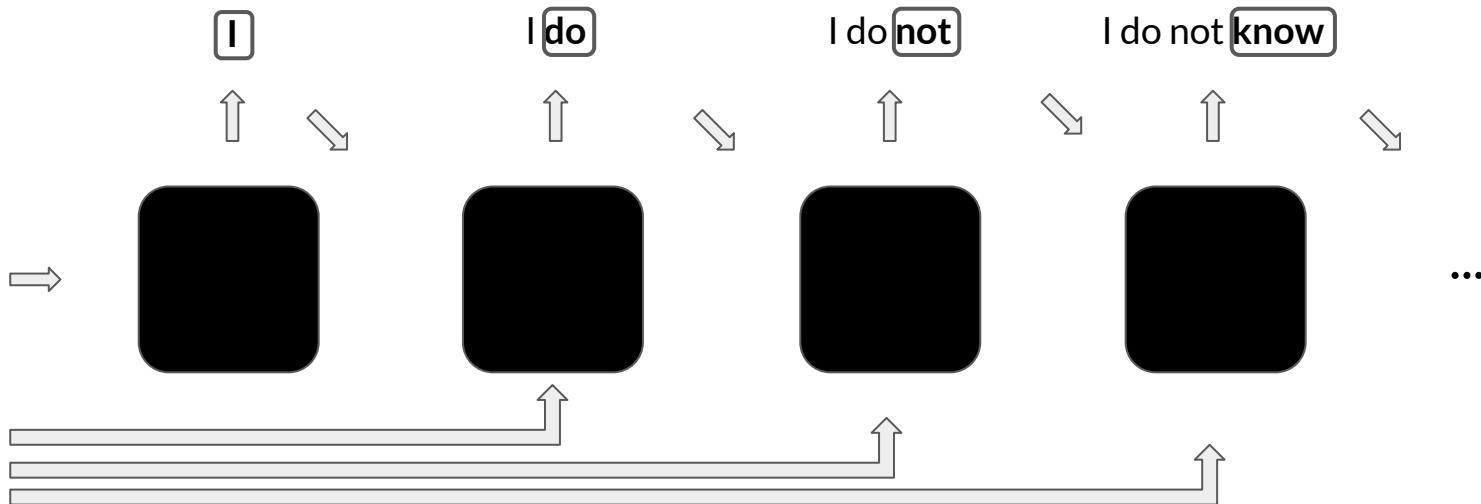
I do not know what I am talking about .



Nie wiem o czym mówię .



Nie wiem o czym mówię.



Training data - Parallel corpus

Zjadłem dziś pyszny obiad.	I had a great dinner today.
Gdy jadę samochodem zapinam pasy.	When I drive a car I close my seat belt.
Mamy dziś piękną pogodę!	We have a great weather today!
Masz czas dziś wieczorem?	I don't like pasta with garlic.
My internet browser isn't working.	My internet browser isn't working.
03.05.1999	03.05.1999

- Depending on language pair, roughly 30k - 100M sentence pairs used for training



Evaluation

- Human evaluation
 - Linguist, according to some guideline
 - Does not require a reference translation
- Automatic evaluation
 - BLEU metric - modified precision of n-grams
 - Requires a reference translation
 - 0-100, the higher the better



	SOURCE	REFERENCE	MODEL OUTPUT	BLEU
#1	Uber's big changes come as the company struggles to turn a profit months after it went public in May .	Duże zmiany w Uberze mają miejsce po wejściu na giełdę w maju i braku zysków od miesięcy .	Wielkie zmiany Ubera się jak firma walczy o obrót zysk miesięcy po tym , jak wszedł na giełdę w maju .	13.52
#2			Duże zmiany w Uberze następują , gdy firma walczy o zysk kilka miesięcy po wejściu na giełdę w maju .	37.52
#3			Wielkie zmiany w Uber nastąpiły , gdy firma walczy o zwrot z zysku kilka miesięcy po tym , jak stała się publiczna w maju .	4.9



How do we improve the baseline?

- Add more parallel data.
- Train bigger model.



How do we improve the baseline?

- Add more parallel data.
- Train bigger model.
- But what if we are constrained to use only the data we are given?

EMNLP 2020 FIFTH CONFERENCE ON MACHINE TRANSLATION (WMT20)

**November 19-20, 2020
Online**

Home

- Given certain data, train THE BEST machine translation system.
- What is THE BEST? Performing best on test set, according to human judgment



Samsung R&D Institute Poland submission to WMT20 News Translation Task

**Mateusz Krubiński, Marcin Chochowski, Bartłomiej Boczek,
Mikołaj Koszowski, Adam Dobrowolski,
Marcin Szymański, Paweł Przybysz**
Samsung R&D Institute Poland



Corpus filtering

Zjadłem dziś pyszny obiad.	I had a great dinner today.	0.95
Gdy jadę samochodem zapinam pasy.	When I drive a car I close my seat belt.	0.75
Mamy dziś piękną pogodę!	We have a great weather today!	0.96
Masz czas dziś wieczorem?	I don't like pasta with garlic.	0.2
My internet browser isn't working.	My internet browser isn't working.	0.6
03.05.1999	03.05.1999	0.65

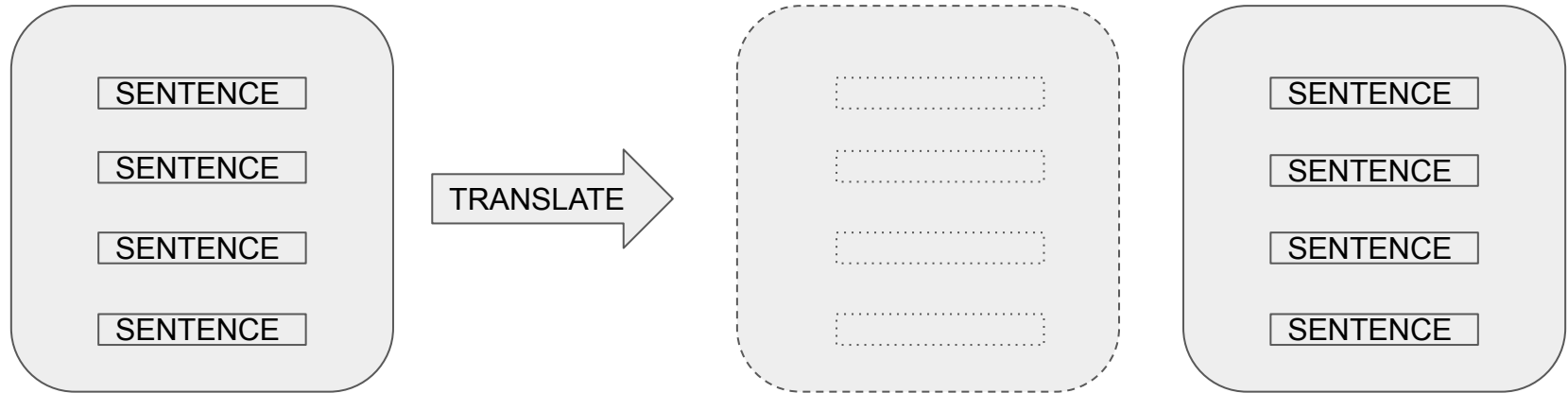


	Orig.	+ clean-up	+ fast-align
En \leftrightarrow Cs	62.5M	61.8M	43.4M
En \leftrightarrow Iu	2.6M	1.2M	1.1M
En \leftrightarrow Pl	11.2M	10.7M	8.6M



Back-translation

- Much more monolingual than parallel data - how to use it?
- Translate it with the baseline system, use the artificial parallel corpus



System	newsdev2020	
	En → Cs	Cs → En
Quadro-huge	26.0	32.7
+ finetune	26.5	33.5
+ ensemble	27.3	33.8
+BT	27.4	37.4



Ensembling with in-domain language model



MT System

In-domain LM

Ciało Galois to pierwsze o czym
pomyślałem.



Galois ...



System	newsdev2020	
	En → Pl	Pl → En
Quadro-huge	27.3	32.3
+ finetune	27.4	32.8
+ ensemble	28.7	32.9

Google Translate: “Galois's body was the first thing I thought about.”



Polish→English

Ave.	Ave. z	System
77.2	0.131	SRPOL
76.7	0.097	Online-G
77.7	0.096	NICT-Rui
77.9	0.094	Online-B
78.1	0.085	SJTU-NICT
76.6	0.083	Online-A
75.2	0.050	OPPO
77.3	0.006	Online-Z
78.1	-0.003	CUNI-Transformer
76.1	-0.038	NICT-Kyoto
73.3	-0.041	VolcTrans
73.2	-0.048	PROMT-NMT
74.3	-0.072	Tilde
74.0	-0.130	zlabs-nlp

English→Polish

Ave.	Ave. z	System
88.6	0.672	HUMAN
76.4	0.493	SRPOL
75.6	0.435	eTranslation
76.3	0.383	VolcTrans
74.0	0.348	Tilde
70.6	0.316	Online-G
72.0	0.310	OPPO
72.4	0.299	NICT-Kyoto
69.7	0.272	Tilde
71.8	0.255	CUNI-Transformer
70.1	0.236	Online-B
69.0	0.219	SJTU-NICT
64.5	0.097	Online-A
63.9	-0.060	Online-Z
47.7	-0.538	zlabs-nlp

Inuktitut→English

Ave.	Ave. z	System
73.1	0.168	NiuTrans
72.9	0.167	Facebook-AI
71.2	0.100	CUNI-Transfer
70.7	0.096	Groningen
70.3	0.072	SRPOL
71.1	0.066	Helsinki
70.2	0.055	NRC
70.2	0.054	UEDIN
70.1	0.047	UQAM-TanLe
68.8	0.006	NICT-Kyoto
68.4	-0.035	OPPO



Dziękuję za uwagę!



Charles University
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics



unless otherwise stated