The RWTH Aachen University English-German and German-English Unsupervised Neural Machine Translation Systems for WMT 2018

Miguel Graça, Yunsu Kim, Julian Schamper, Jiahui Geng, Hermann Ney
Human Language Technology and Pattern Recognition, RWTH Aachen University, Aachen, Germany

Lehrstuhl Informatik 6
Human Language Technology and Pattern Recognition

RWTH AACHEN UNIVERSITY

Core System

Model:
- 4-layer Transformer
- 300 hidden nodes, 2048 feed-forward hidden units, 6 attention heads, 0.1 dropout
- Shared encoder, decoder and output layer

Optimization:
- AdaM with learning rate $3 \cdot 10^{-4}$ and $\beta_1 = 0.5$
- Batch size: 32 sentences
- Noise model [Lample et. al. 2017] applied to all inputs

Online Back-translation [Artxexe et. al 2017]

- Back-translation during training for the next 10 mini-batches
- Trained for 500k updates $\longrightarrow \approx 3$ epochs

Batch (Iterative) Back-translation [Lample et. al 2017]

- Initialize with an unsupervised word-by-word translation [Conneau et. al. 2017]
- Back-translation after 1 epoch (160k updates)
- Each epoch the model sees **one** back-translation of a given sentence
- Trained for 800k updates $\longrightarrow$ 5 epochs

Data Processing

Training data:
- 100M sentences from NewsCrawl 2014 to 2017: Word embedding training
- 5M subset of above 100M sentences: Translation model training

Vocabularies:
- Shared joint BPE with 50k merge operations
- Shared and unshared word-based vocabularies of top 50k frequent words

Pre-processing:
- Tokenization, numbers / URLs mapped to categories, lower-casing

Post-processing:
- Unknown and category carry-over, frequent-casing, de-tokenization

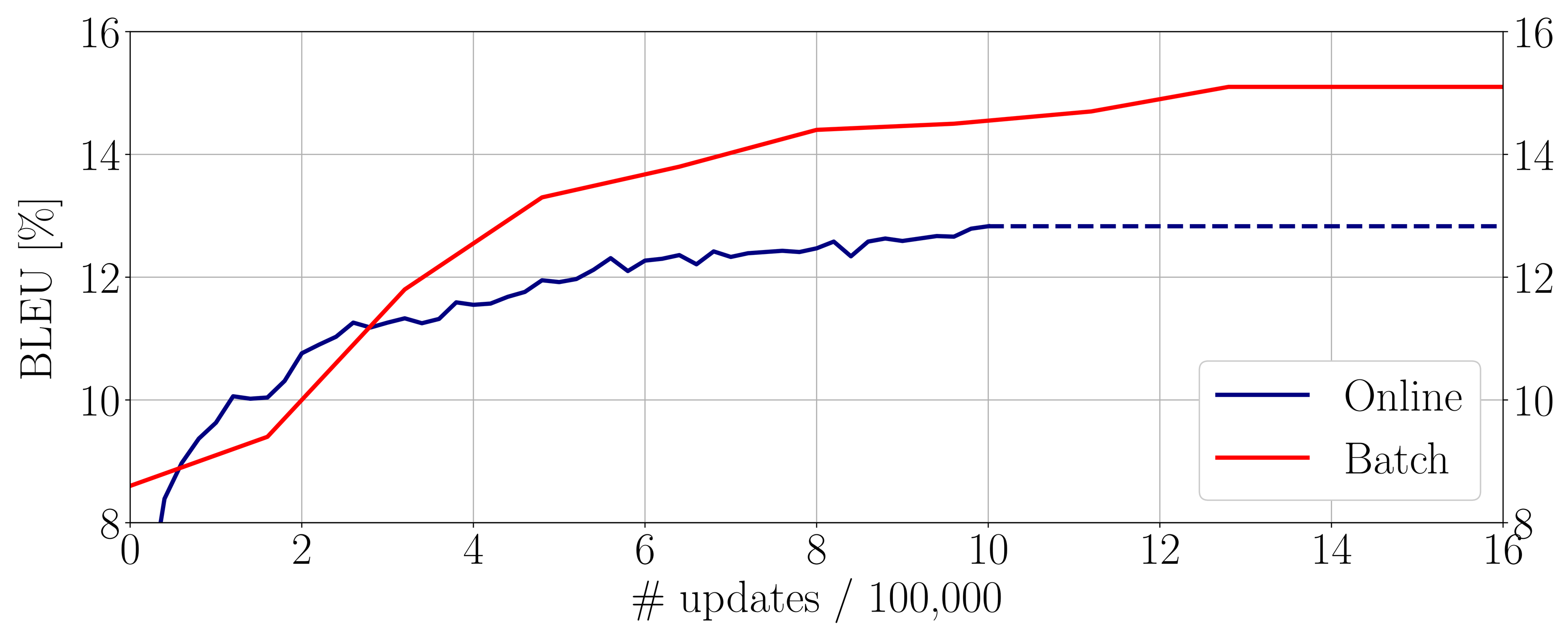
Model selection: BLEU on **newstest2015** German $\longrightarrow$ English (see Submission)

Embedding Initialization Experiments

		German → English				English → German	
		newstest2017		newstest2018		newstest2018	
		BLEU [%]	TER [%]	BLEU [%]	TER [%]	BLEU [%]	TER [%]
random	online	4.9	92.7	4.9	91.7	4.9	96.5
monolingual		7.5	88.2	8.2	85.7	5.9	93.0
cross-lingual		**13.1**	**75.5**	**15.4**	**70.8**	**12.0**	**79.5**
+ frozen		12.7	76.3	15.1	71.6	10.9	78.9
random	batch	14.5	73.6	17.6	68.2	13.7	78.0
monolingual		14.3	73.3	17.2	68.0	13.9	76.9
cross-lingual		**14.9**	**72.7**	**18.1**	**67.1**	**14.0**	**77.0**
+ frozen		14.0	75.8	16.9	71.5	12.6	83.6

Remarks:
- Online system: Weak / implicit cross-lingual signal by embedding initialization
- Batch system: Strong / explicit word-by-word translation

Training Progress on newstest2017 German $\longrightarrow$ English



Vocabulary Experiments

		German → English				English → German	
		newstest2017		newstest2018		newstest2018	
		BLEU [%]	TER [%]	BLEU [%]	TER [%]	BLEU [%]	TER [%]
words	batch	**14.9**	**72.7**	**18.1**	**67.1**	**14.0**	**77.0**
unshared		14.5	73.3	17.2	67.8	13.6	77.2
words	online	11.9	75.7	14.2	71.0	10.6	81.5
unshared		10.6	77.7	13.2	73.1	9.7	81.9
BPE 20k		11.8	77.9	13.6	73.9	10.8	81.1
BPE 50k		**13.1**	**75.5**	**15.4**	**70.8**	**12.0**	**79.5**

Word Embedding Gating Mechanism

$$\bar{w} = \Big(g(w) \odot E_{pre-train}(w) + (1 - g(w)) \odot E_{random}(w)\Big)$$

Gate weights:

$$g(w) = \sigma\Big(b + W \cdot [E_{pre-train}(w), E_{random}(w)]\Big)$$

Motivation:
- Pre-trained embeddings are rich in information and have the cross-lingual property
- However: adaptation to the task at hand might cancel out its benefits
- Embedding pre-training is not normalized $\longrightarrow$ apply weight normalization

Additional Feature Experiments

	German → English				English → German	
	newstest2017		newstest2018		newstest2018	
	BLEU [%]	TER [%]	BLEU [%]	TER [%]	BLEU [%]	TER [%]
baseline	**14.9**	**72.7**	**18.1**	**67.1**	**14.0**	**77.0**
+ frozen emb.	14.0*	75.8*	16.9*	71.5*	12.6*	83.6*
+ gating	14.4*	72.5	17.6*	67.3	14.2	77.2
+ emb. WN	14.5*	73.4*	17.5*	68.4*	13.6	77.7*
+ emb. WN	**14.7**	**72.8**	**18.2**	**67.1**	**14.4***	**76.9**
+ adversarial loss	13.9*	74.2*	16.9*	69.0*	12.8*	79.6*
+ unshared decoder	14.3*	73.3*	17.3*	68.0*	13.9	77.4
+ drop AE & noise	**15.2**	**72.6**	**18.3**	**66.9**	**14.4***	**76.5***

* denotes a p-value of $< 1\%$ w.r.t. the baseline

Remarks:
- Don't freeze your embeddings!
- Adversarial loss needs to be adjusted for the Transformer architecture
- Noise and auto-encoding can be dropped during training after the 4th epoch

Final Results

	newstest2018			
	German → English		English → German	
	BLEU [%]	TER [%]	BLEU [%]	TER [%]
RWTH submission	18.6	66.3	14.8	75.3
LMU submission	17.9	68.4	15.5	76.2
RWTH internal	**24.4**	**60.6**	**17.4**	**73.7**
best supervised systems	48.4	38.1	48.3	40.7

- Model selection: round-trip BLEU on **newstest2017**
- Internal setup: BPE, same data, larger models, larger batch size

Acknowledgements

This work has received funding from the European Research Council (ERC) (under the European Union's Horizon 2020 research and innovation programme, grant agreement No 694537, project "SEQCLAS") and the Deutsche Forschungsgemeinschaft (DFG; grant agreement NE 572/8-1, project "CoreTec"). The GPU computing cluster was supported by DFG (Deutsche Forschungsgemeinschaft) under grant INST 222/1168-1 FUGG. The work reflects only the authors' views and none of the funding agencies is responsible for any use that may be made of the information it contains.