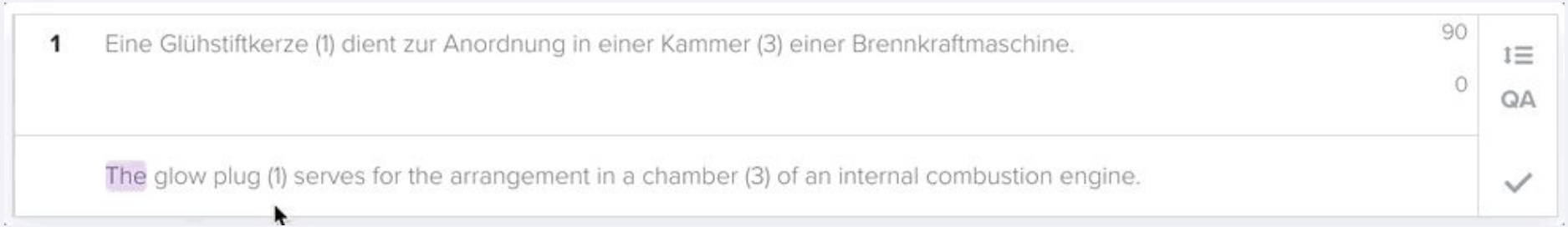
EMNLP 2018, Nov 2

Compact Personalized Models for Neural Machine Translation

Joern Wuebker, Patrick Simianer, John DeNero
{joern,patrick,john}@lilt.com

Personalized interactive MT

- **interactive** MT:



Personalized interactive MT

- **Personalized** MT: Models are **adapted** towards each user
  - **Batch adaptation**: User uploads domain-relevant bilingual data
  - **Online adaptation**: Model immediately learns from every translated sentence

- Strict **latency constraints**
  - Translations need to be generated at typing speed

- **Large number** of adapted models
  - One model per user
  - New user model after every translated sentence

Personalized MT: Inference process

1. **Load** User X's model from cache or persistent storage
2. **Apply** model parameters to computation graph
3. Perform **inference**

(1.) + (2.) ⇒ max. ~10M parameters for personalized model (**latency** constraints)

Full model: ~36M parameters

Solution: - Store personalized models as offsets from baseline model $W = W_b + W_u$

- Select sparse parameter subset W_u

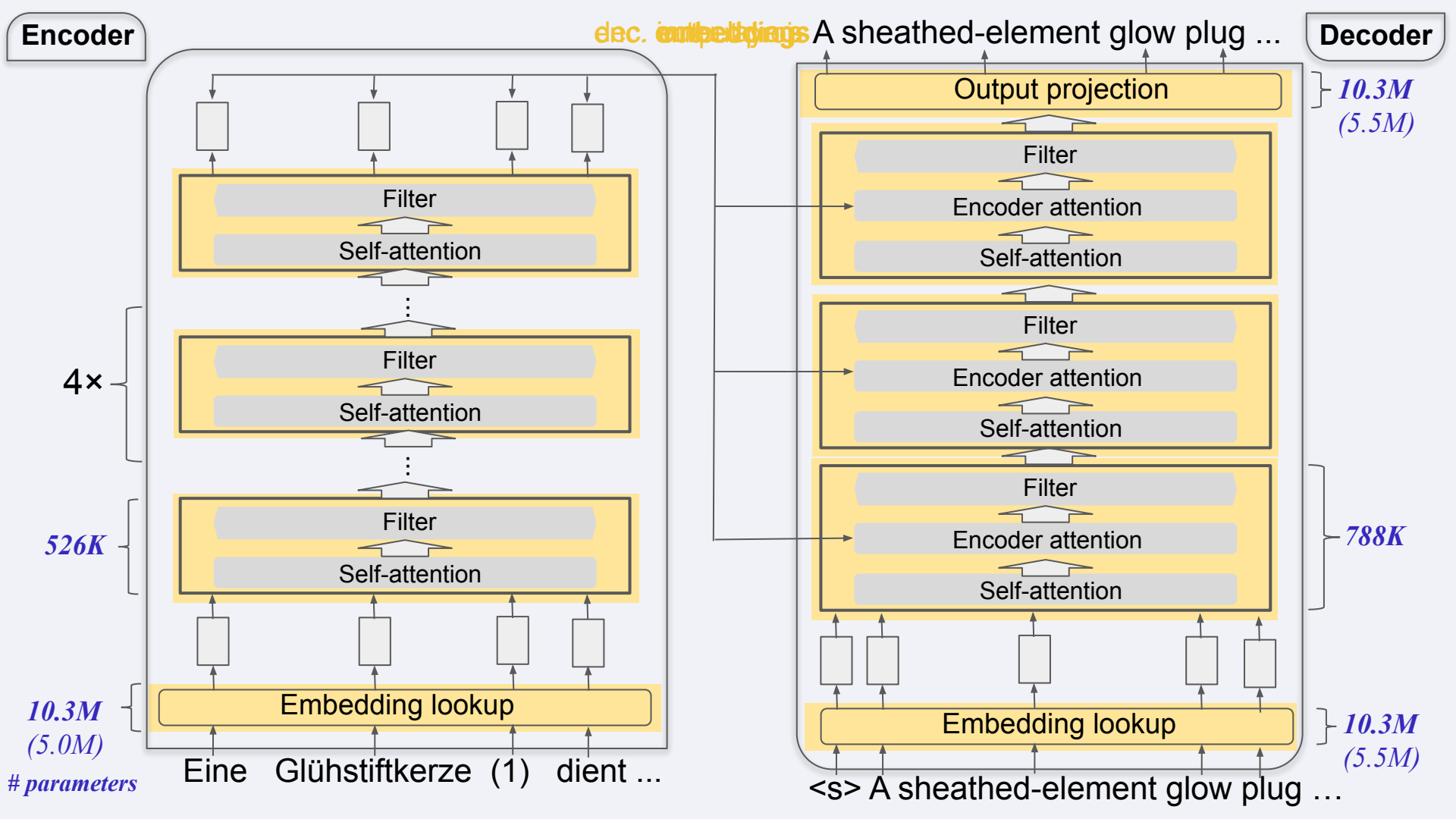
Experimental setup

- Down-sized self-attentive transformer network (Vaswani et al., 2017)
- 40k BPE tokens
- Adaptation: Fine tuning with SGD

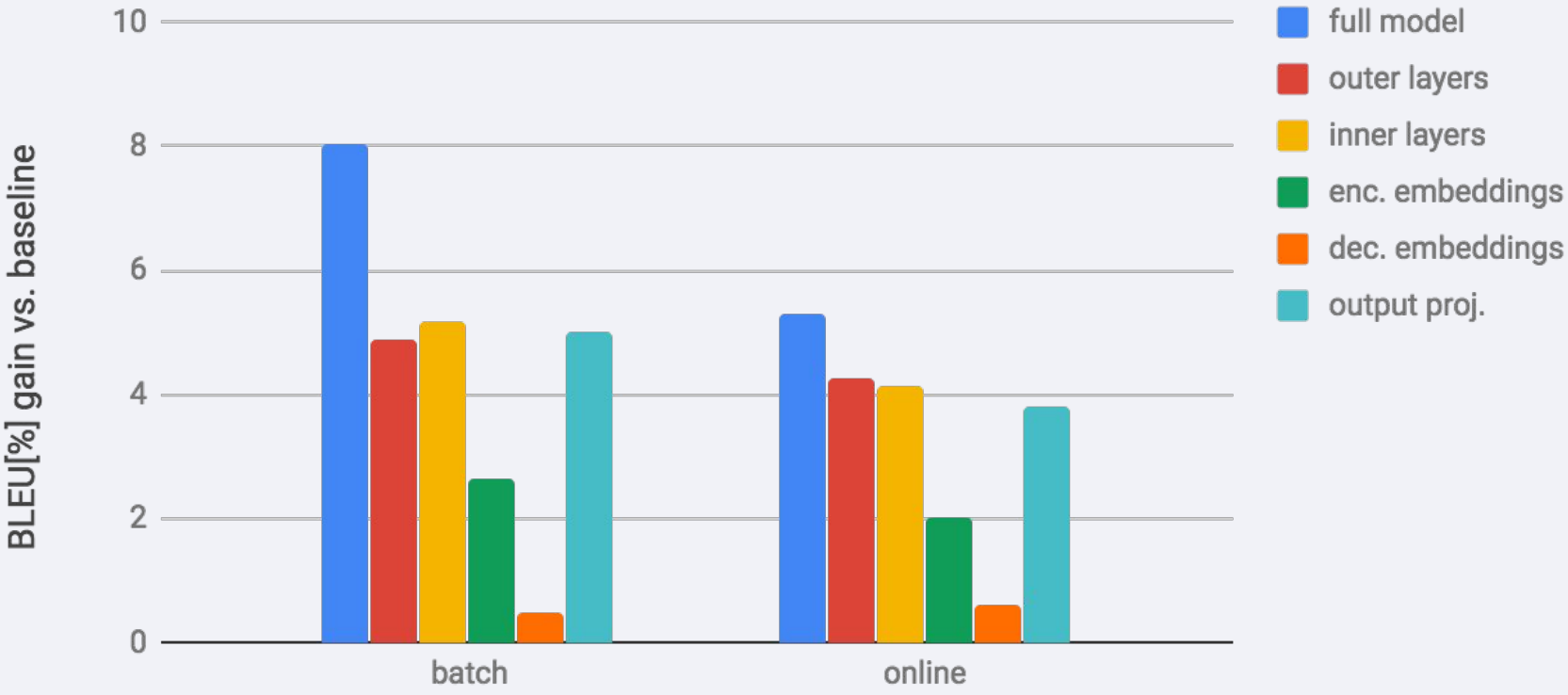
- **Main experiments:**
  - German→English production system
  - Here: Results are averages over four test sets (for individual scores see paper)
  - Separate experiments for batch and online adaptation
- **Final experiments:**
  - Six different production systems: English↔French, English↔Russian, English↔Chinese
  - Joint batch and online adaptation

Idea 1:
Select specific network regions

Freezing Subnetworks to Analyze Domain Adaptation in Neural Machine Translation,
Thompson et al., WMT 2018

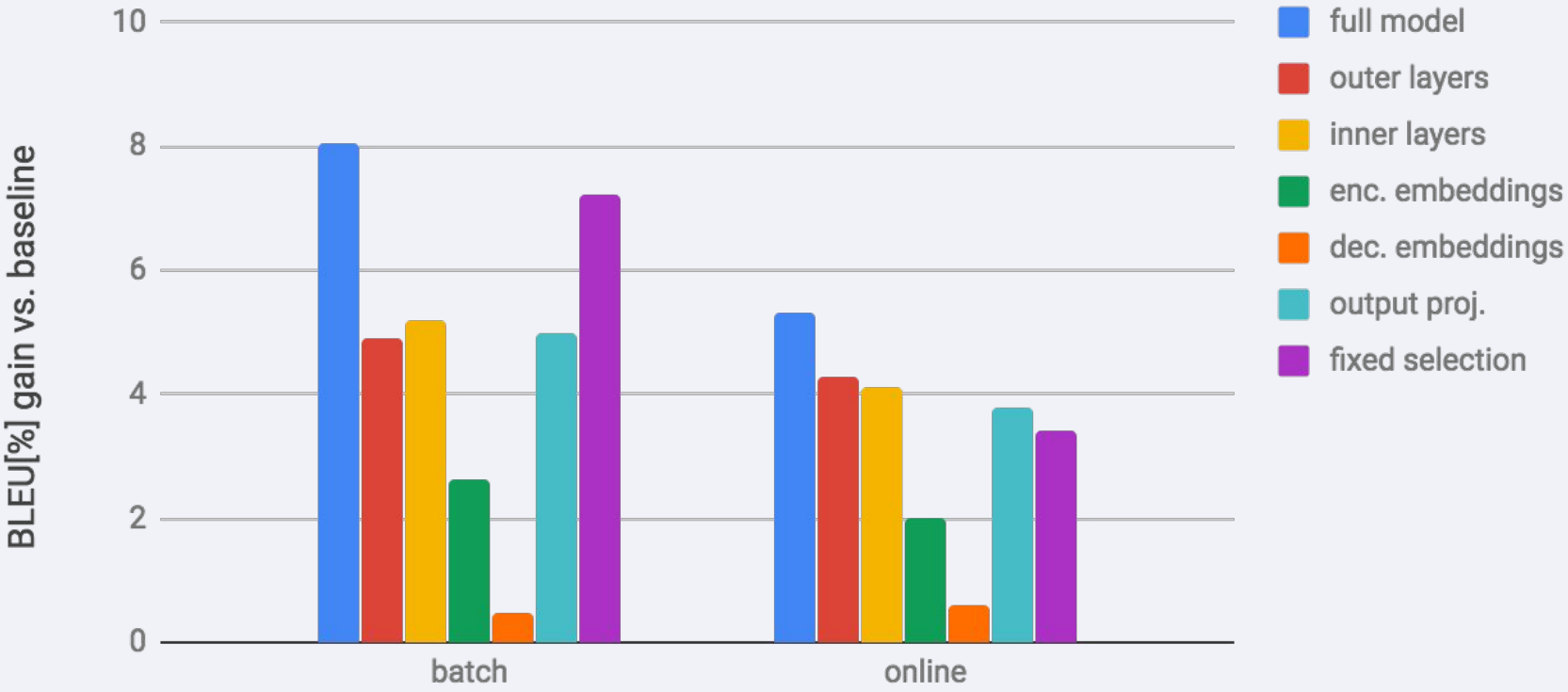

Encoder
dec. embeddings
A sheathed-element glow plug ...
Decoder
Output projection
10.3M
(5.5M)
Filter
Encoder attention
Self-attention
Filter
Encoder attention
Self-attention
Filter
Encoder attention
Self-attention
788K
4×
Filter
Self-attention
Filter
Self-attention
526K
Filter
Self-attention
Embedding lookup
10.3M
(5.0M)
parameters
Eine Glühstiftkerze (1) dient ...
Embedding lookup
10.3M
(5.5M)
<s> A sheathed-element glow plug …

Idea 1: Select specific network regions



Idea 2:
Select most relevant tensors on development set

Idea 2: Select most relevant tensors on dev



Idea 3:
Group Lasso

Idea 3: Group Lasso

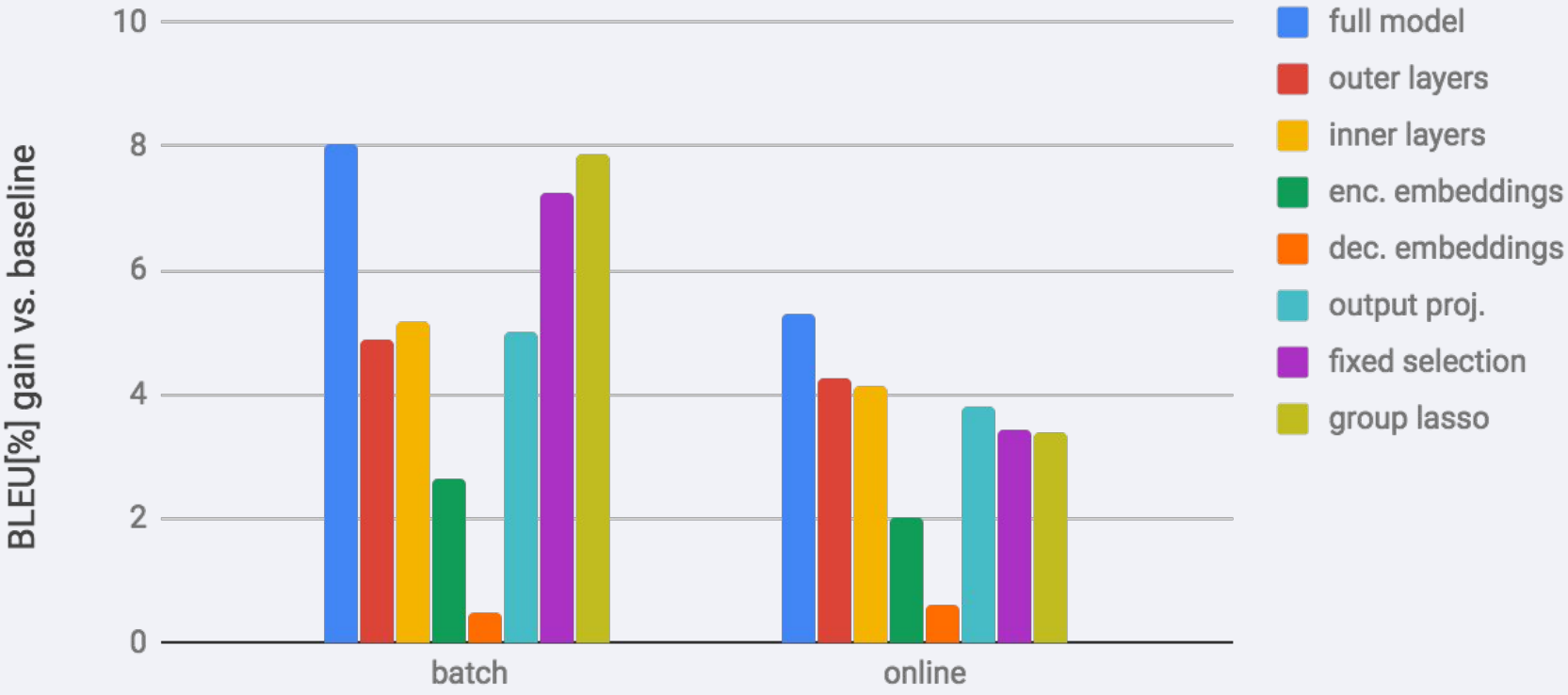
- Simultaneous regularization and tensor selection
- Regularize offsets W_u, define each tensor as one group g for L1/L2 regularization

$$R_{\ell_{1,2}}(W_u) = \sum_{g \in W_u} \sqrt{|g|}\ \|g\|_2$$

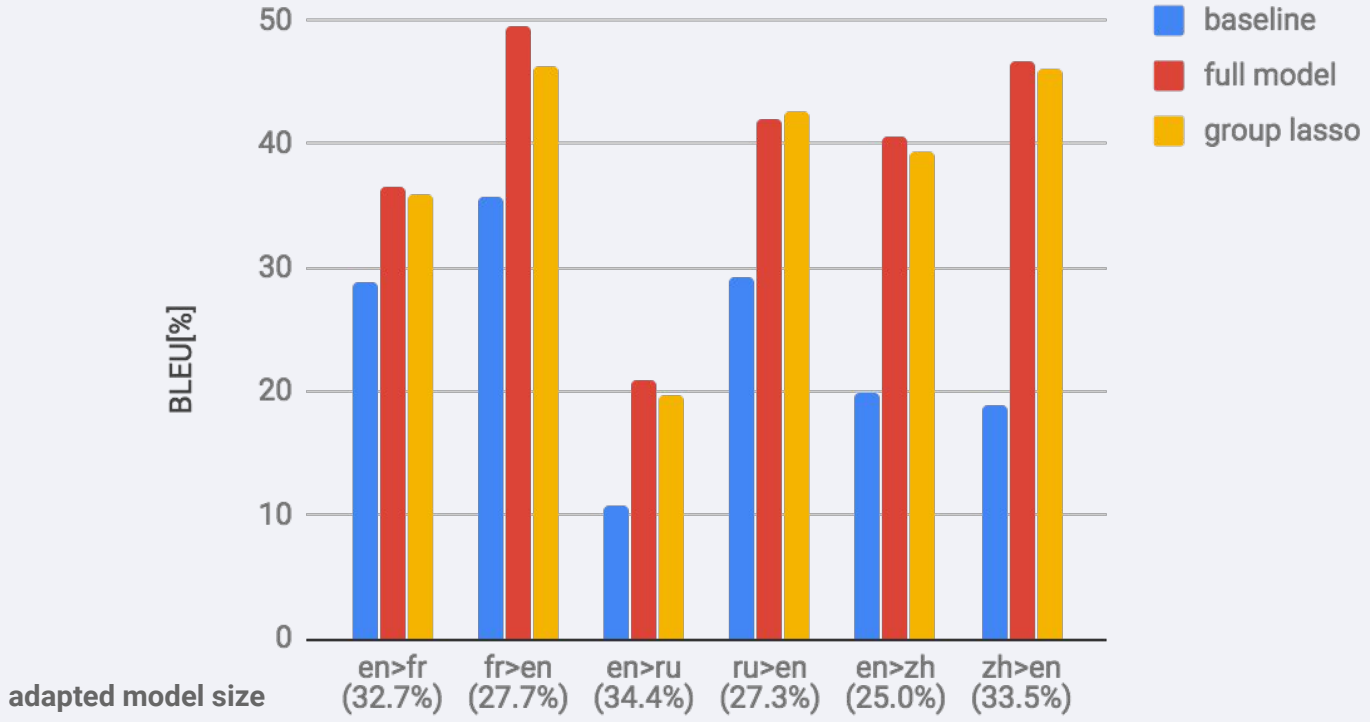
- Total loss: $\mathcal{L} = \mathcal{L}_{seq}(W_b + W_u) + \lambda R_{\ell_{1,2}}(W_u)$

- Cut off all tensors g with $\frac{1}{|g|} \sum_{w \in g} |w| < \theta$

Idea 3: Group Lasso



Final results (batch + online)



Conclusion

- Personalized interactive machine translation requires sparse adaptation
- Define adapted models by their parameter **offsets** to the baseline model
- **Group lasso**:
  - Regularize and select the parameter offsets
  - Quality similar to full model adaptation
  - Reduces number of adapted parameters by ~70%

EMNLP 2018, Nov 2

We’re hiring! (San Francisco & Berlin)

Joern Wuebker
joern@lilt.com