Applied Machine Learning Days
Lausanne, Jan 28, 2019

LILT

Interactive and Adaptive Translation for Professionals

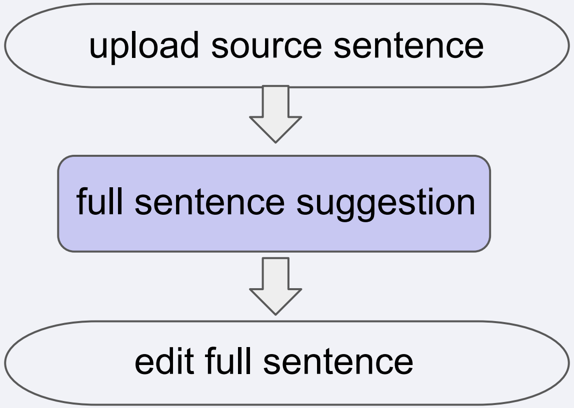
Joern Wuebker
joern@lilt.com

Live Demo

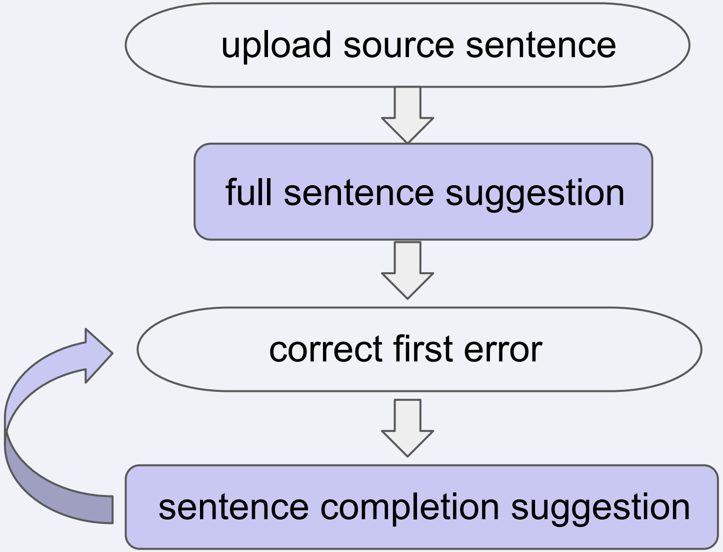
Interactive Machine Translation

Post-editing vs. Interactive MT

Post-editing



Interactive MT



Notation: *user action* *machine action*

Interactive Machine Translation (MT)

- Generally feels more **natural** to translators than post-editing
- Interactive machine translation leads to more edits and **higher end translation quality** [Green et al., 2014] [*Client Evaluation*, 2017]
  - Error frequency, detected by review, was 1.1% for post-editing & 0.3% for interactive MT.
  - Throughput with interactive MT was 700+ words/hour, double a typical unassisted speed.

Green, Spence, et al. "Predictive Translation Memory: A mixed-initiative system for human language translation." Proceedings of the 27th annual ACM symposium on User interface software and technology. ACM, 2014.

Adaptive Machine Translation

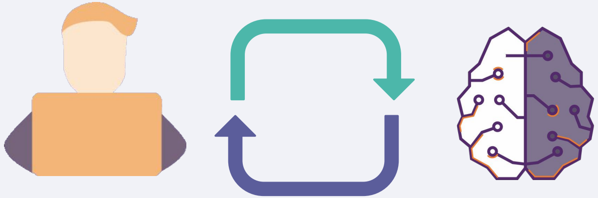
Adaptive Machine Translation

- Personalized adaptation means that translation suggestions **improve as the translators work**:
  - Model **immediately** learns from every translated sentence.
  - **Fewer errors** for translators to correct.
  - Adaptation on a **per-user** basis.
  - Teams become more **consistent** when the system learns from all of them.

Personalized Neural Adaptation

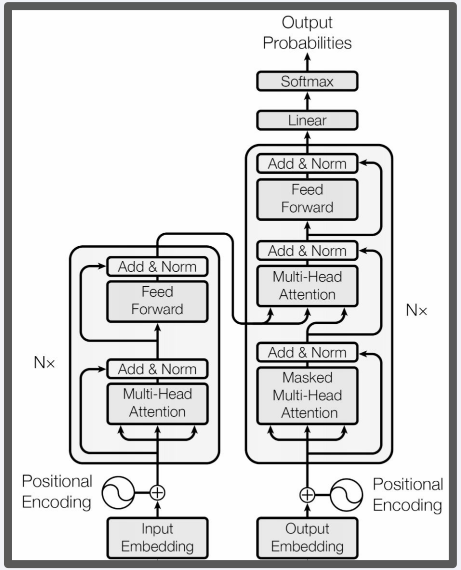
Real-Time
Feedback Loop

Interactive
Translator Interface



Adaptive **Neural**
MT System

- **Gradient descent** on single training instance
- Relearning **all parameters** in a high-quality neural translation model (Transformer architecture) is **too slow** to keep up with a proficient translator.
  ➪ Translations need to be generated at typing speed
- *Lilt's solution*: learn a **dynamic subset** of parameters to be adapted **for each user**.



Adaptive 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

Group Lasso

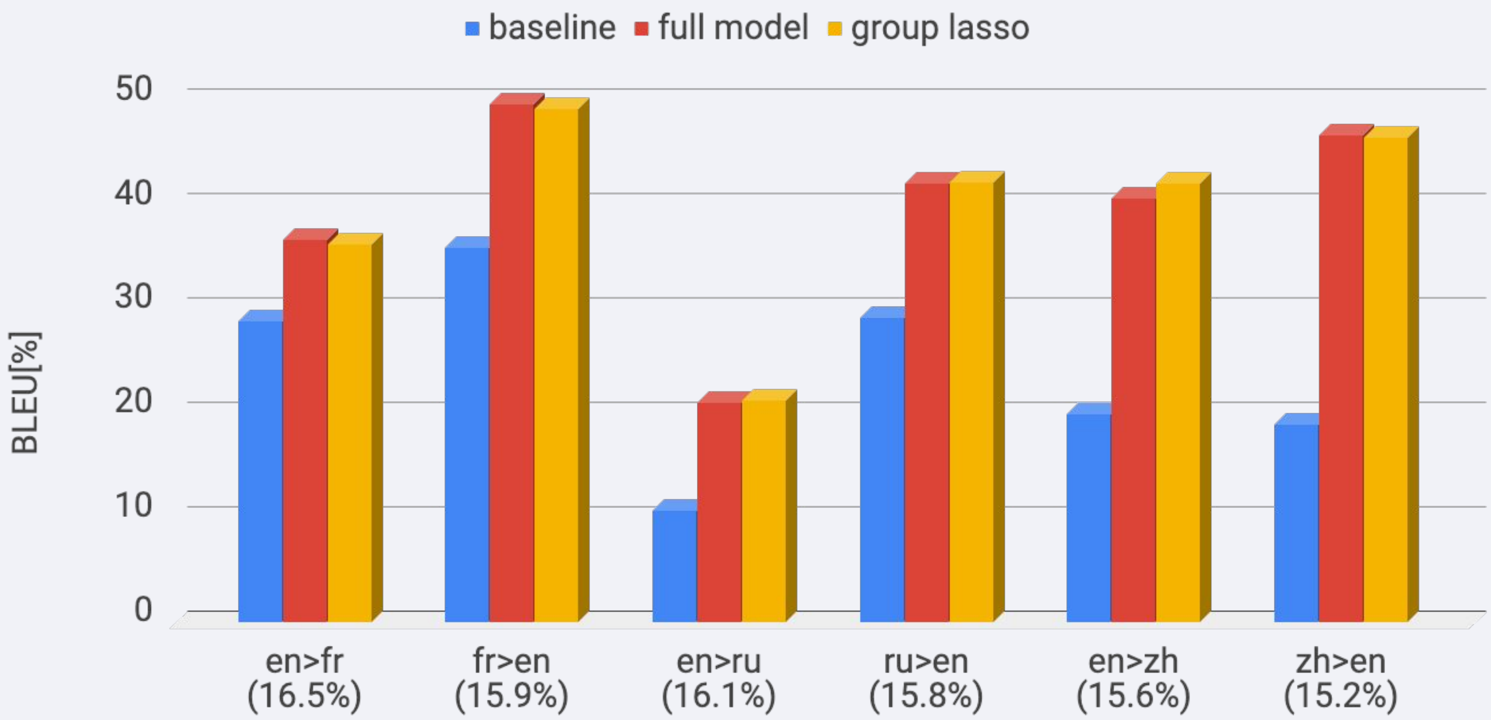
- Simultaneous regularization and tensor selection
- Store personalized models as offsets from baseline model $W = W_b + W_u$
- Regularize offsets W_u, define each tensor as one group *g* for L1/L2 regularization

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

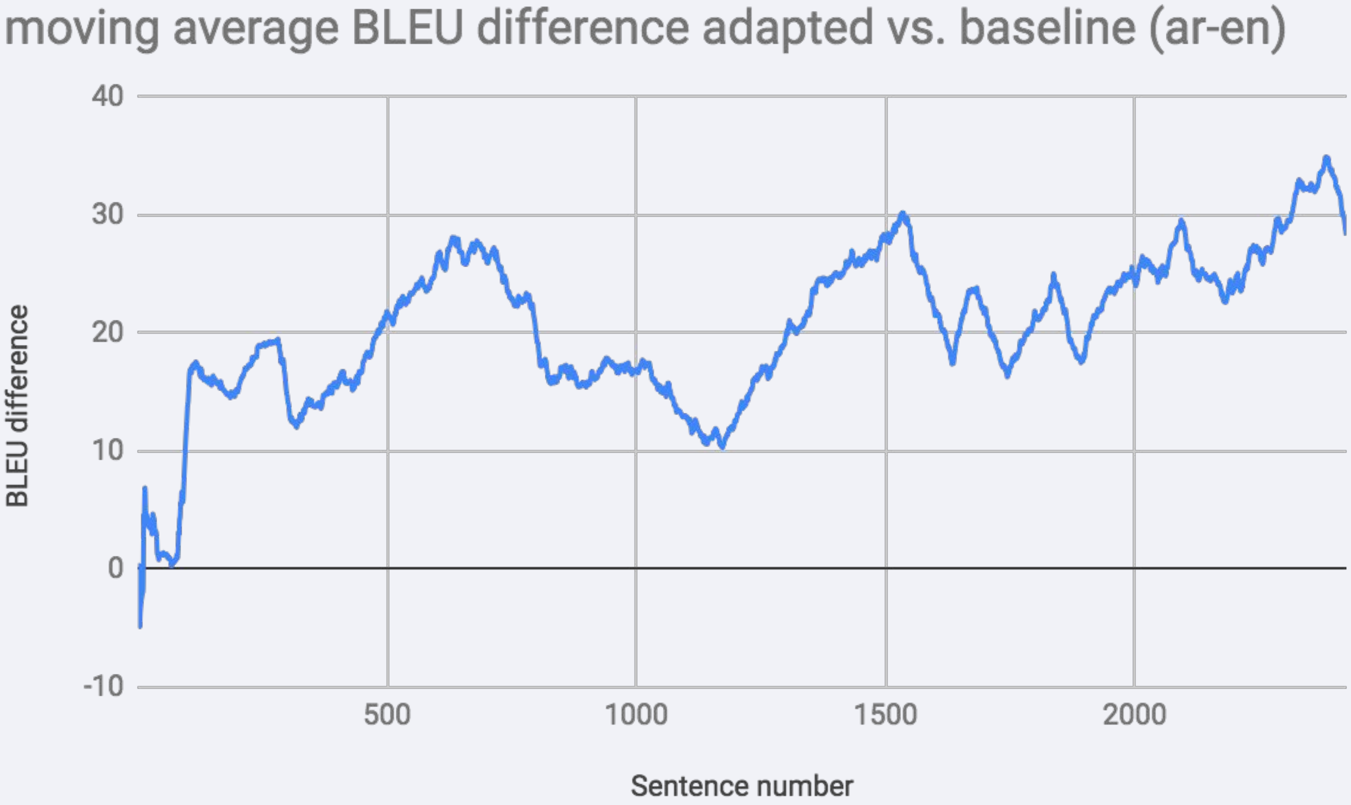
- Total loss: $\mathcal{L} = \mathcal{L}_{seq}(W_b + W_u) + \lambda R_{\ell_{1,2}}(W_u)$
- Cut off all tensors *g* with $\frac{1}{|g|} \sum_{w \in g} |w| < \theta$

Wuebker et al. "Compact Personalized Models for Neural Machine Translation", Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP 2018), Brussels, Belgium, October 2018

Adaptation results



Incremental Adaptation Results



Summary

- Interactive machine translation yields **higher quality** than from-scratch translation or post-editing
- Model is adapted **immediately** after every translated sentence
- **Group lasso**
  - Select a different subset of parameters **for each user**
  - Translation quality similar to full model adaptation
  - Reduces number of adapted parameters by ~85%

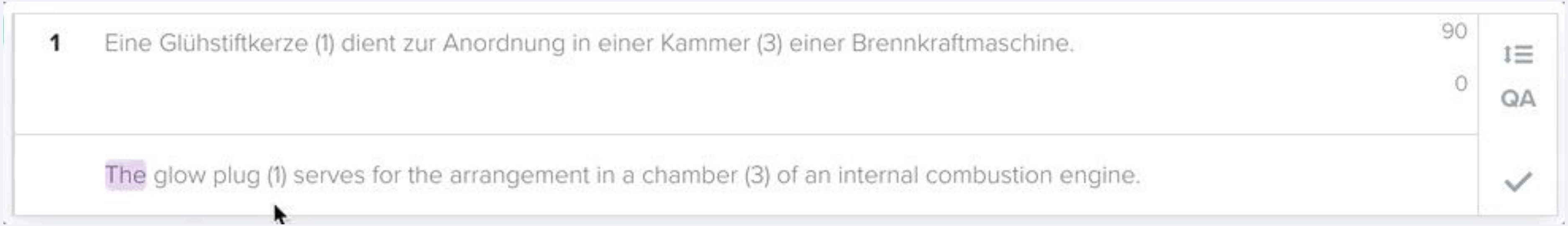
Applied Machine Learning Days
Lausanne, Jan 28, 2019

LILT

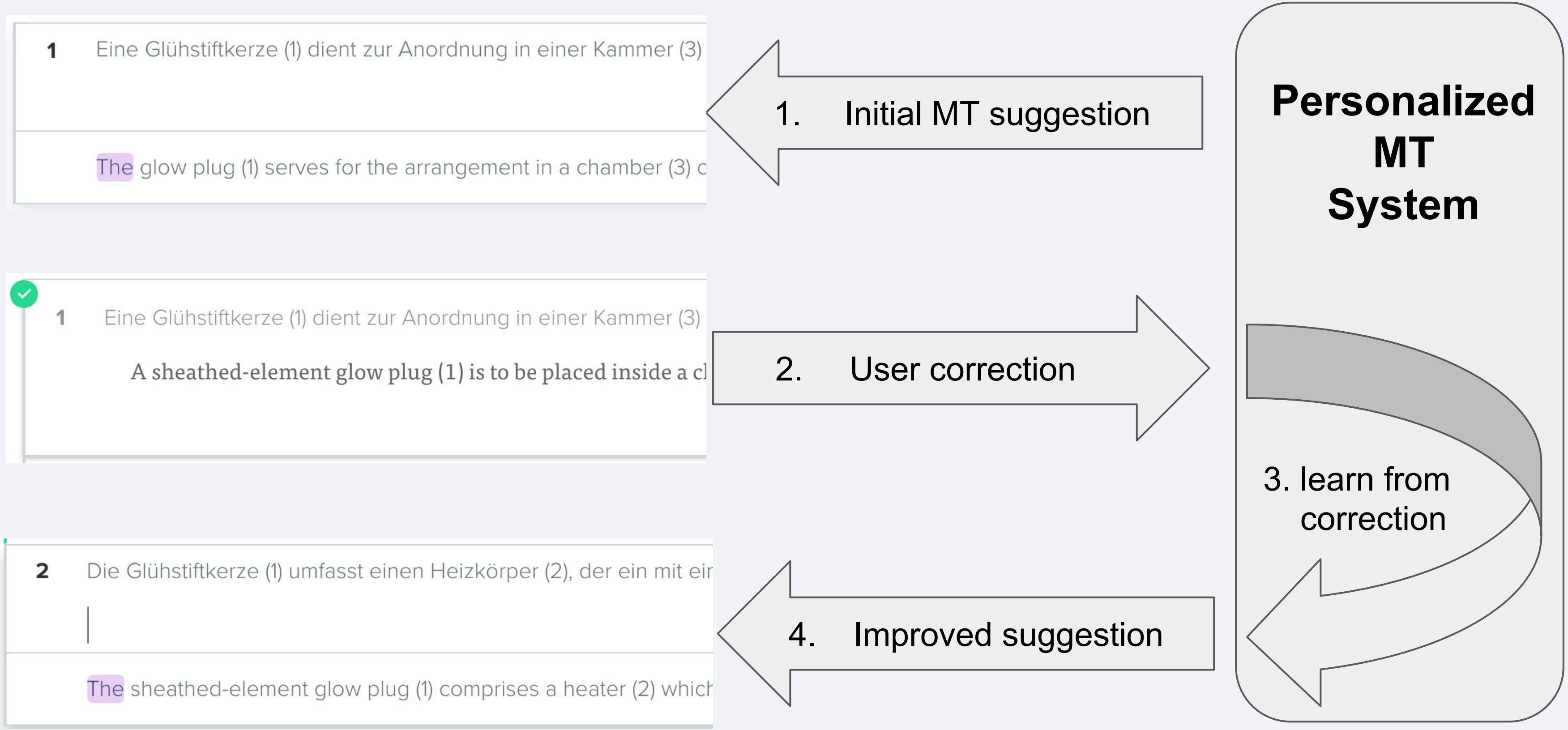
Thank you!

Joern Wuebker
joern@lilt.com

Interactive Machine Translation



Adaptive Machine Translation: Example



Adaptive Machine Translation

- **Personalized** machine translation: Models are **adapted** towards each user
  - **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