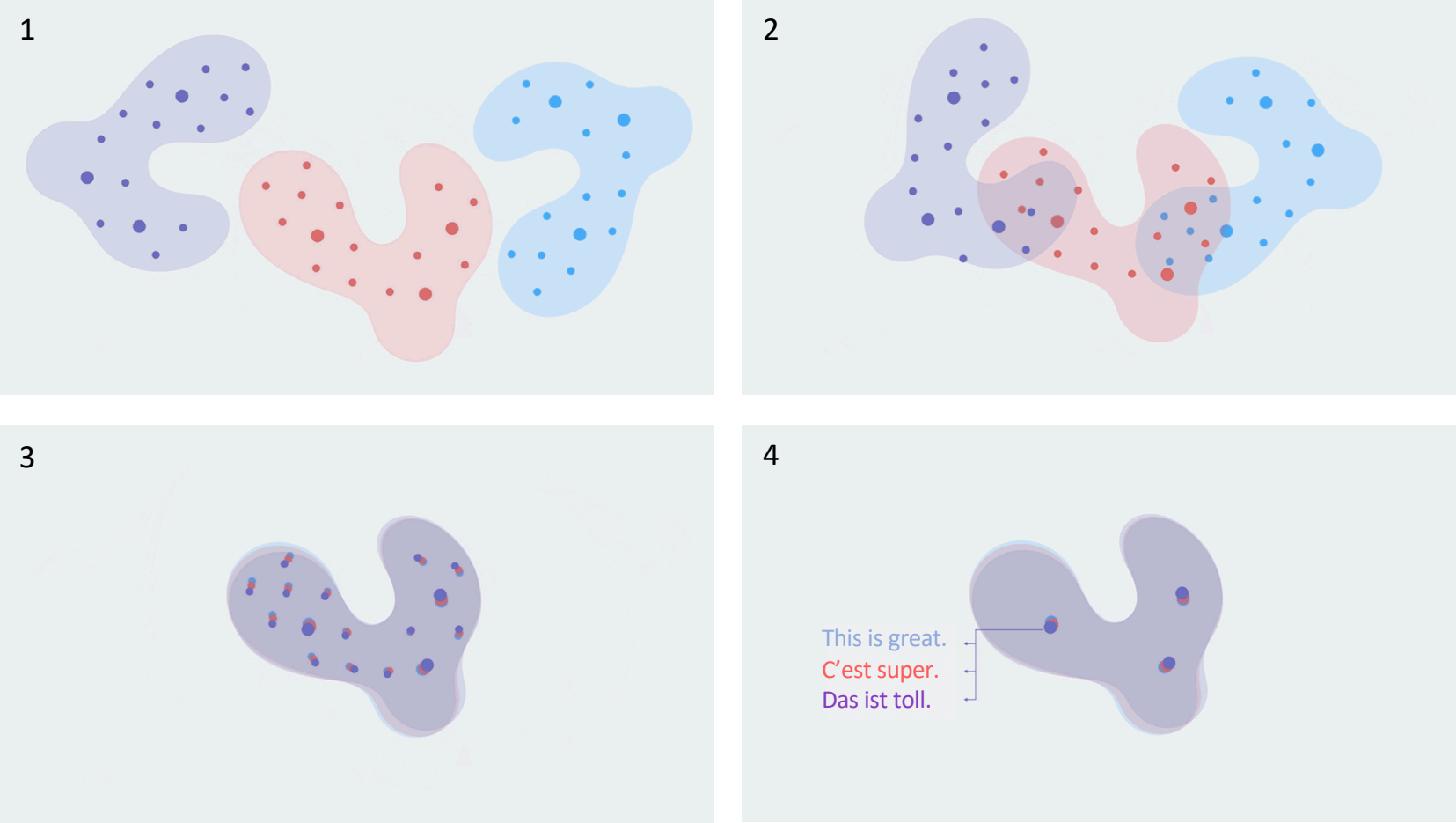
Cross-lingual language model pretraining

Alexis Conneau and Guillaume Lample

Facebook AI Research

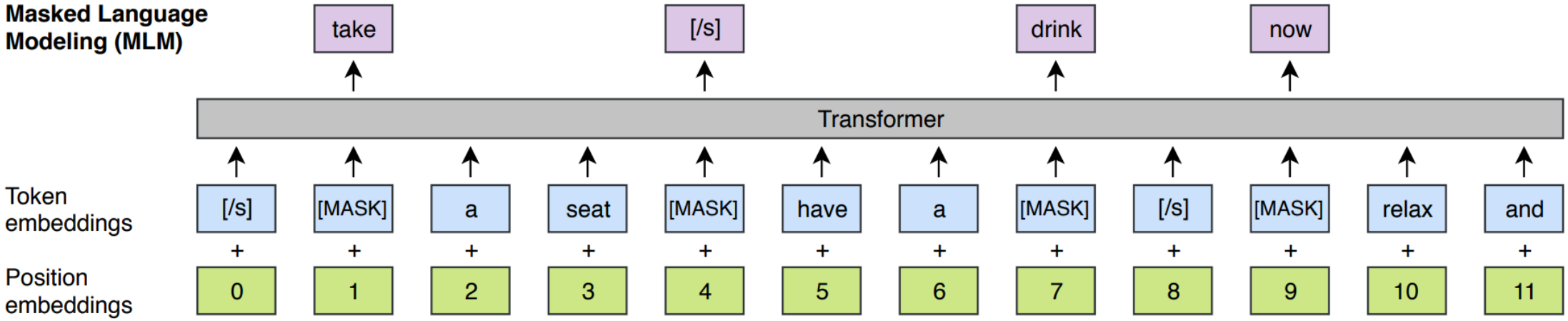
Why learning cross-lingual representations?



Cross-lingual language models

Mult. Masked Language Modeling (MLM)

Similar to BERT, we pretrain a Transformer model with MLM but in many languages:



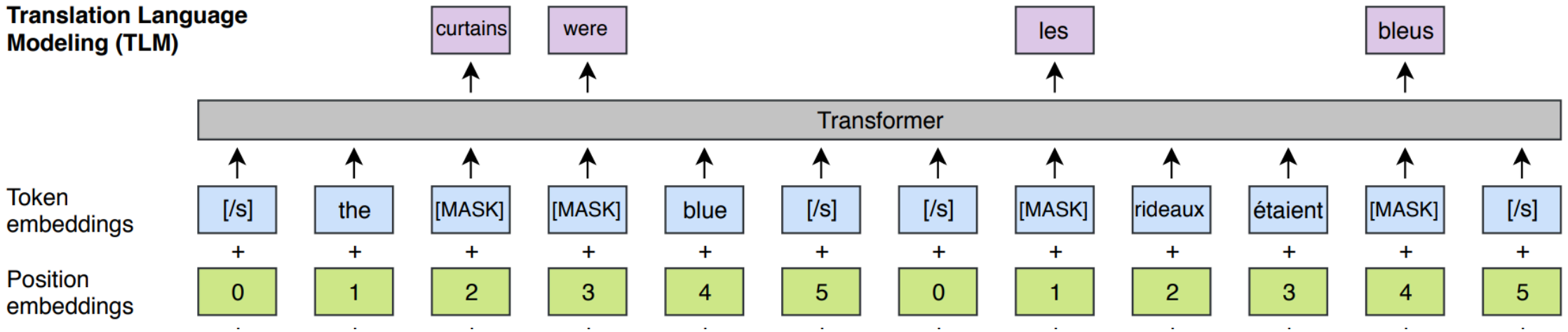
Multilingual Masked language modeling pretraining

.. multilingual representations emerge from a single MLM trained on many languages.

Devlin et al. – BERT: Pretraining of Deep Bidirectional Transformers for Language Understanding (+ mBERT)

Translation Language Modeling (TLM)

Multilingual MLM is unsupervised, but we leverage parallel data with TLM:



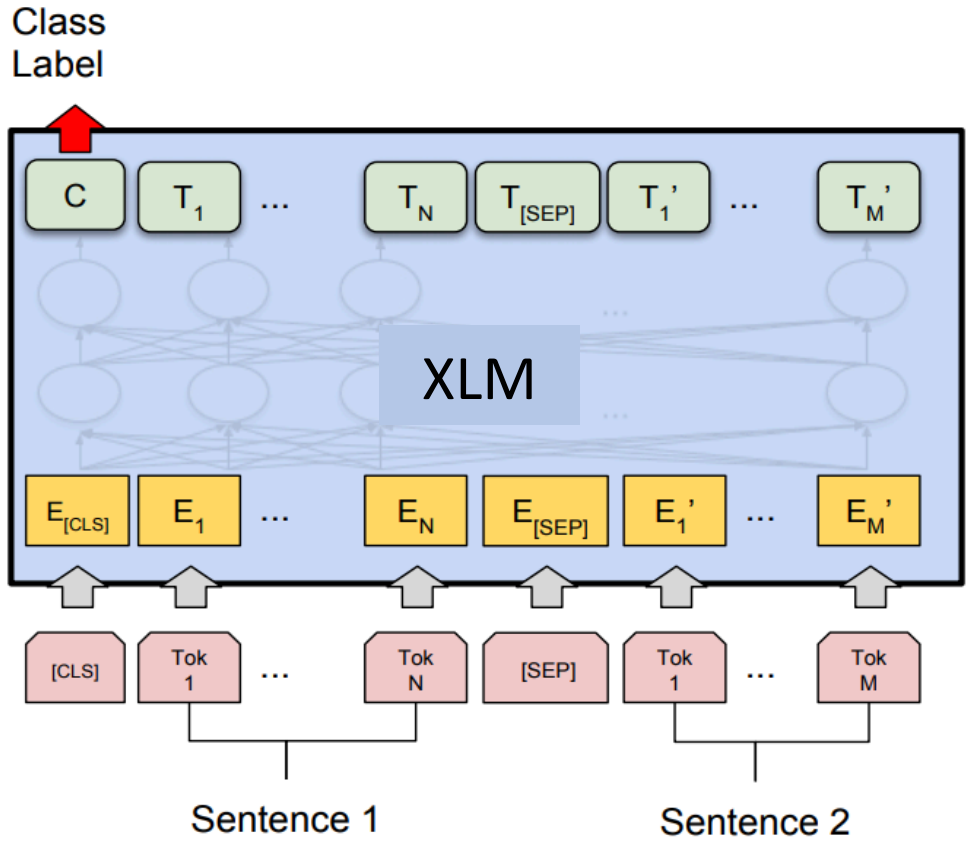
Translation language modeling (TLM) pretraining

.. to encourage the model to leverage cross-lingual context when making predictions.

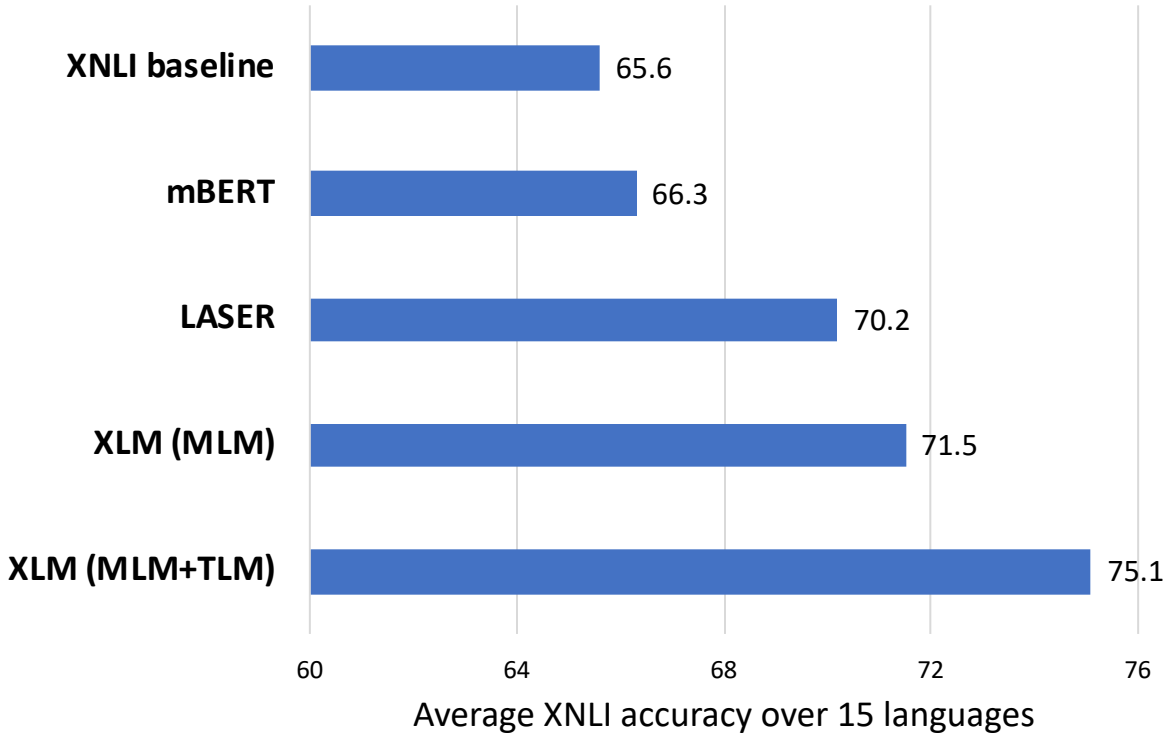
Results on XLU benchmarks

Results on Cross-lingual Classification (XNLI)

The pretrained encoder is fine-tuned on the English XNLI(*) training data and then tested on 15 languages



Average XNLI accuracy on the 15 languages of XNLI for zero-shot cross-lingual classification

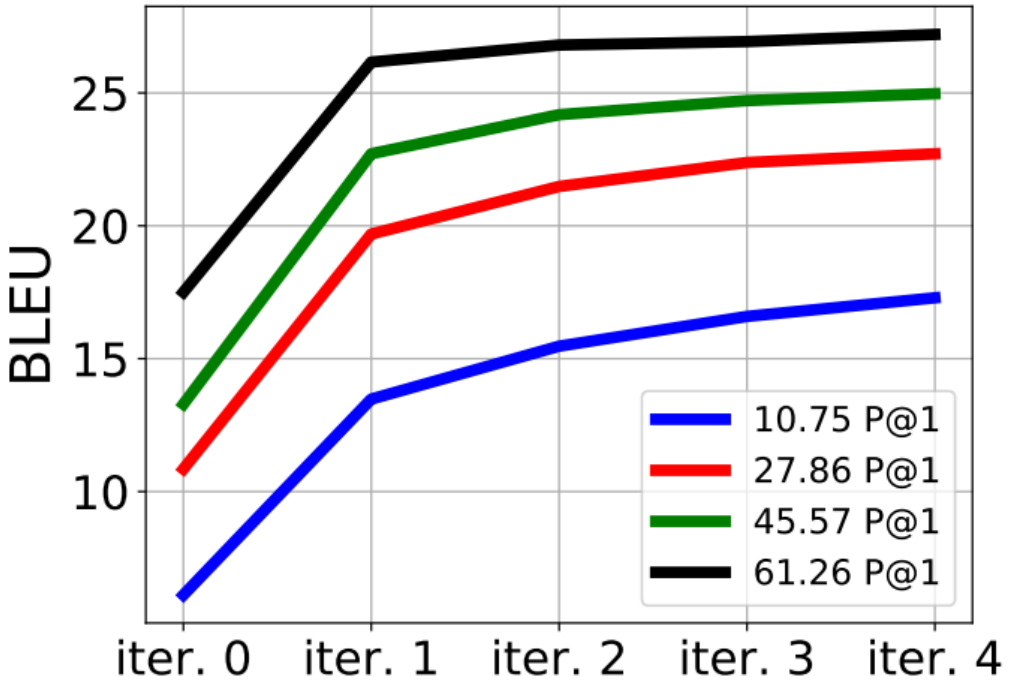


(*) Conneau et al. – XNLI: Evaluating Cross-lingual Sentence Representations (EMNLP 2018)

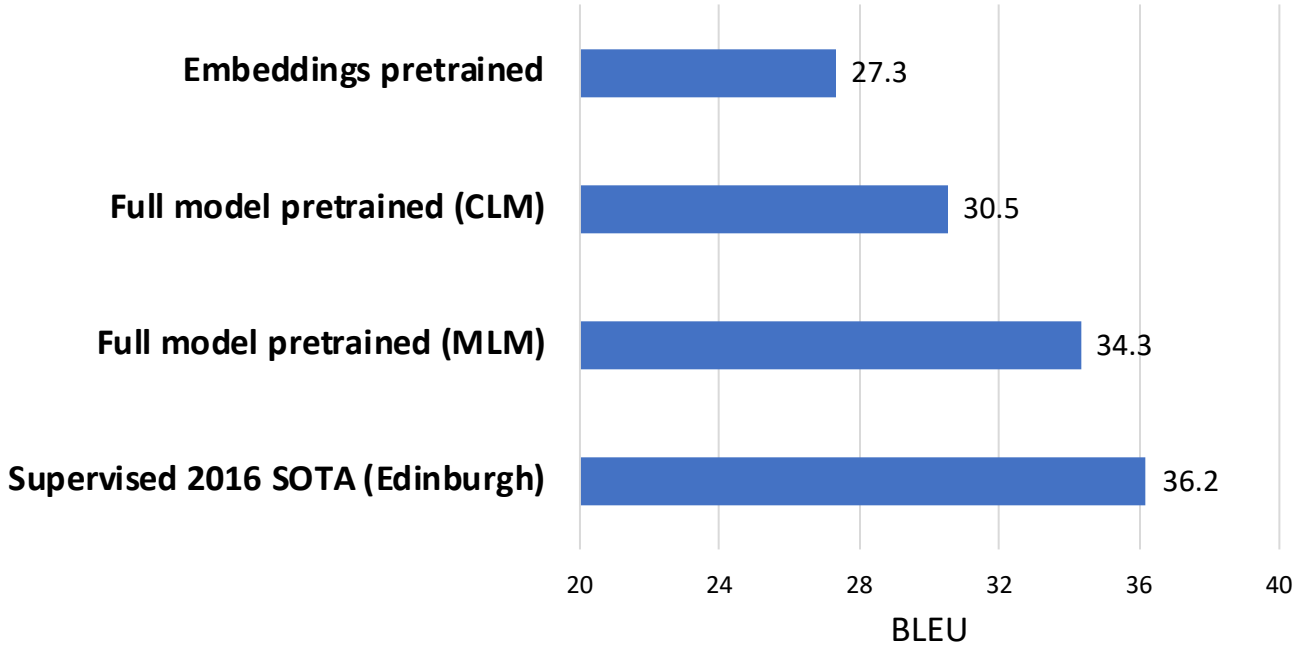
Results on Unsupervised Machine Translation

Initialization is key in unsupervised MT to bootstrap the iterative BT process

Embedding layer initialization
is essential for neural unsupervised MT (*)



Full Transformer model initialization
significantly improves performance (+7 BLEU)

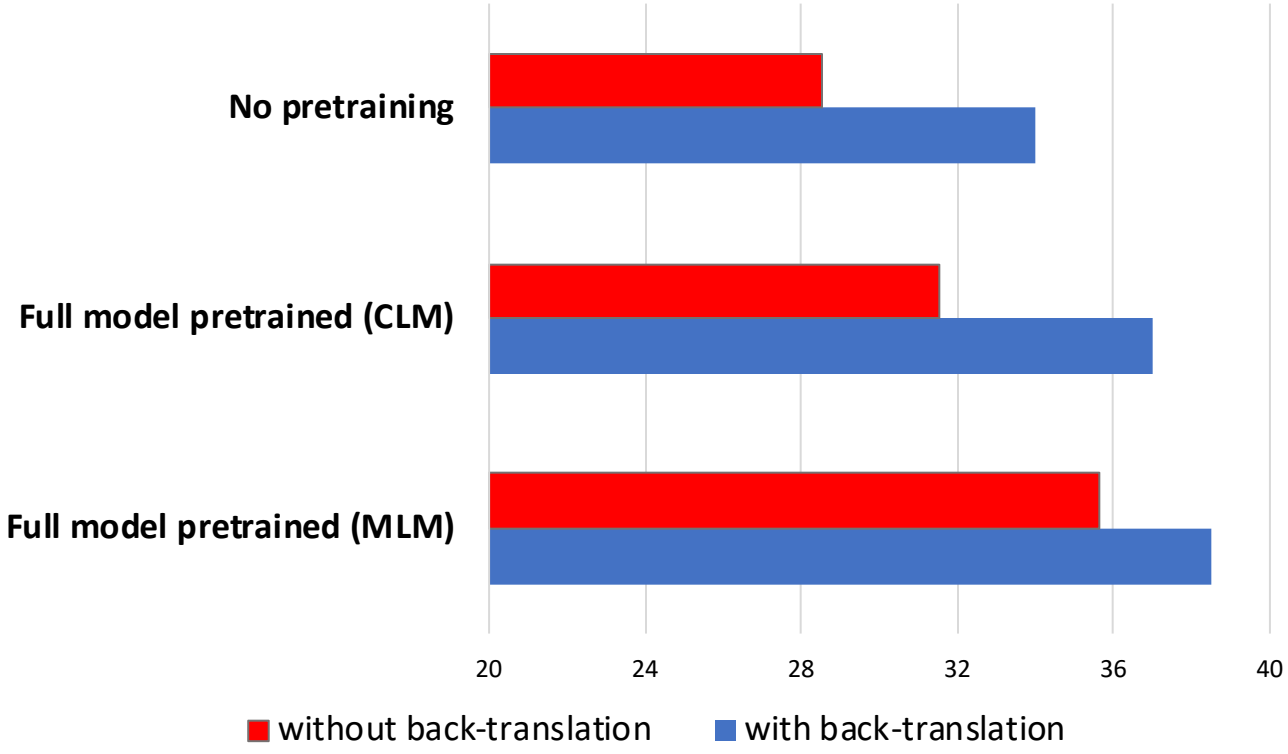


(*) Lample et al. – Phrase-based and neural unsupervised machine translation (EMNLP 2018)

Results on Supervised Machine Translation

We also show the importance of pretraining for generation

- Pretraining both the encoder and decoder improves BLEU score
- MLM better than LM pretraining
- Back-translation + pretraining leads to the best BLEU score
- Pretraining is more important when supervised data is small



Conclusion

- Cross-lingual language model pretraining is very effective for XLU
- New state of the art for cross-lingual classification on XNLI
- Reduces the gap between unsupervised and supervised MT
- Recent developments have improved XLM/mBERT models

Thank you!

Code and models available at github.com/facebookresearch/XLM

Lample & Conneau – Cross-lingual Language Model Pretraining (NeurIPS 2019)