Microsoft

南京理工大学
NANJING UNIVERSITY OF SCIENCE & TECHNOLOGY

MASS: Masked Sequence to Sequence Pre-training for Language Generation

Tao Qin
Joint work with Kaitao Song, Xu Tan, Jianfeng Lu and Tie-Yan Liu
Microsoft Research Asia
Nanjing University of Science and Technology

Motivation

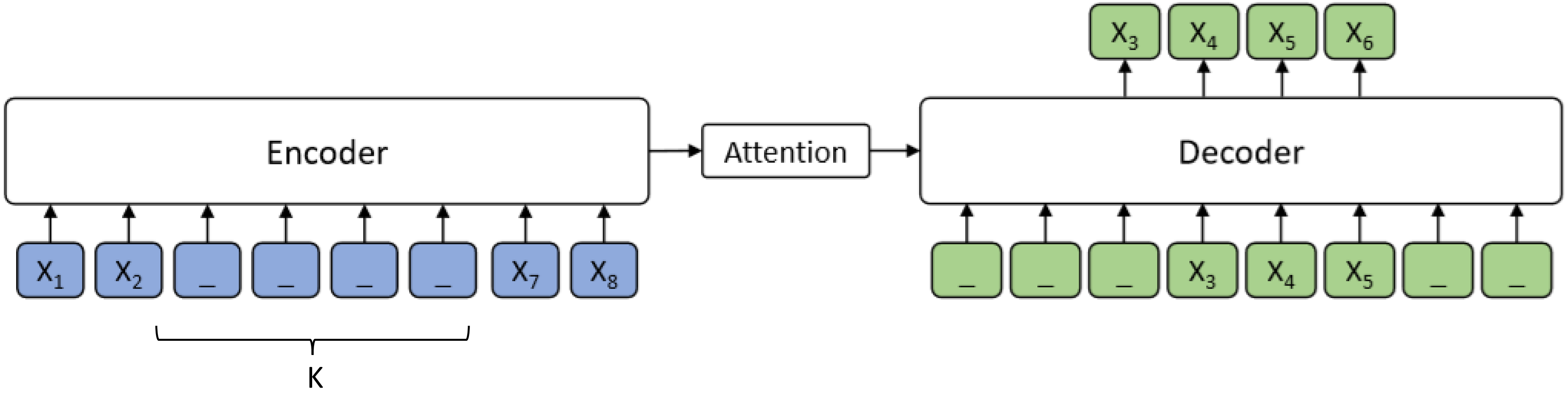
- BERT and GPT are very successful
  - BERT pre-trains an encoder for language understanding tasks
  - GPT pre-trains a decoder for language modeling.
- However, BERT and GPT are suboptimal on sequence to sequence based language generation tasks
  - BERT can only be used to pre-train encoder and decoder separately.
  - Encoder-to-decoder attention is very important, which BERT does not pre-train.

Method	**BLEU**
Without attention	**26.71**
With attention	**36.15**

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. "Neural machine translation by jointly learning to align and translate." ICLR 2015.

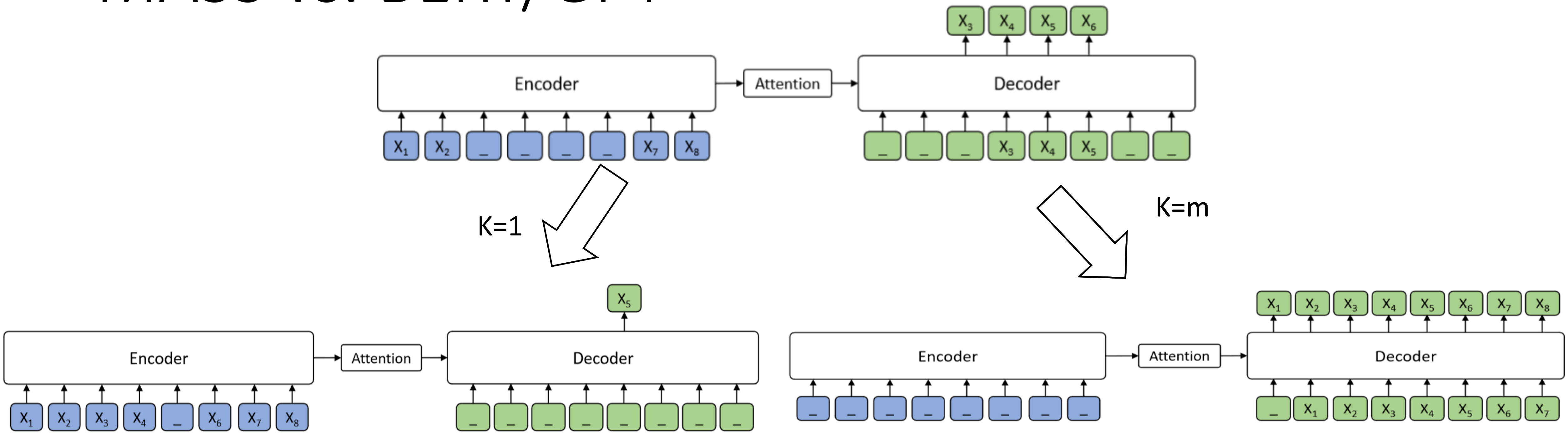
MASS: Pre-train for Sequence to Sequence Generation

- MASS is carefully designed to jointly pre-train the encoder and decoder



- Mask k consecutive tokens (segment)
  - Force the decoder to attend on the source representations, i.e., encoder-decoder attention
  - Force the encoder to extract meaningful information from the sentence
  - Develop the decoder with the ability of language modeling

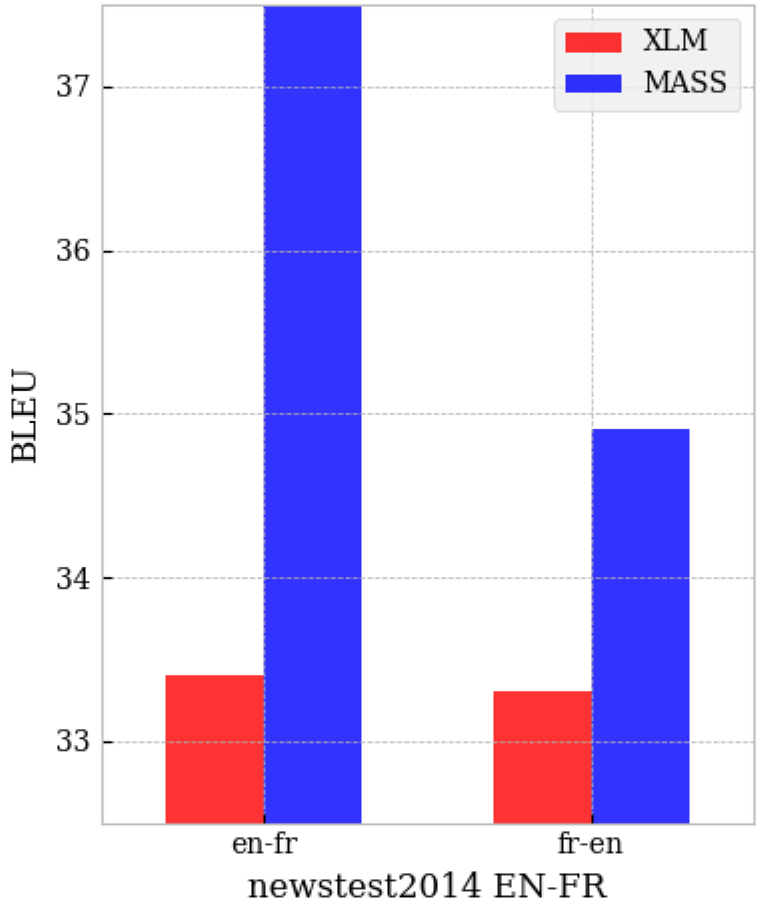
MASS vs. BERT/GPT

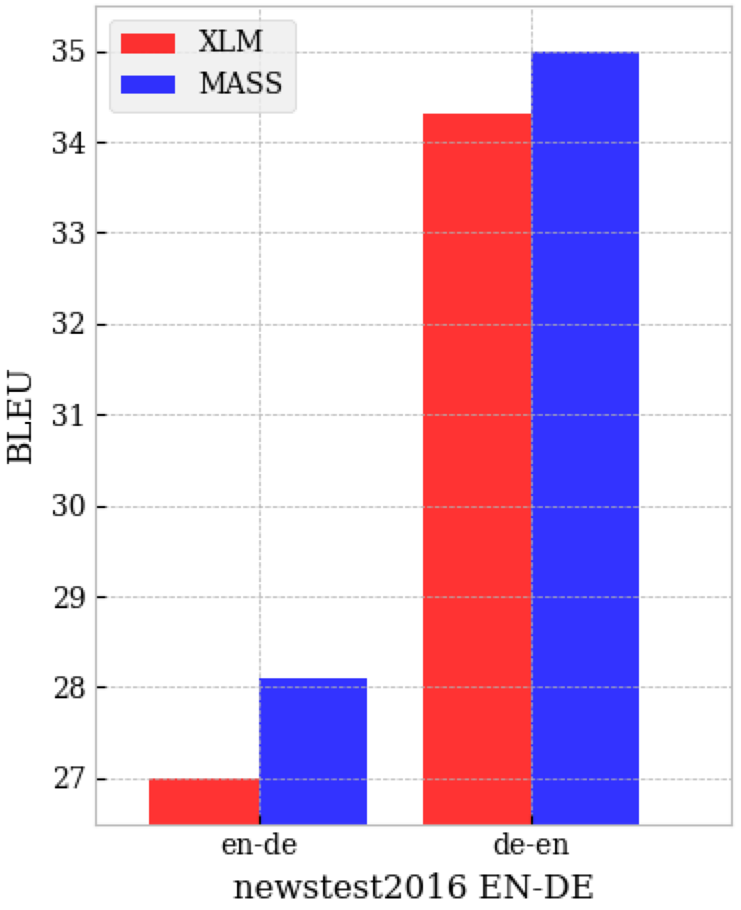


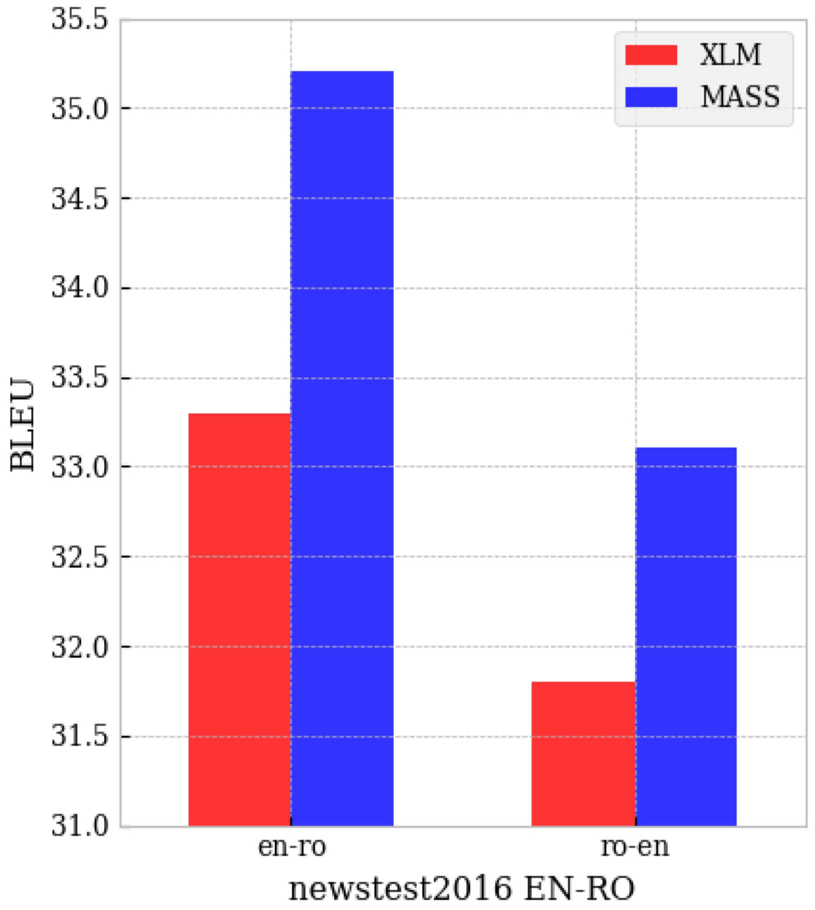
Length	Probability	Model
$k = 1$	$P(x^u \| x^{\setminus u}; \theta)$	masked LM in BERT
$k \in [1, m]$	$P(x^{u:v} \| x^{\setminus u:v}; \theta)$	MASS

Length	Probability	Model
$k = m$	$P(x^{1:m} \| x^{\setminus 1:m}; \theta)$	standard LM in GPT
$k \in [1, m]$	$P(x^{u:v} \| x^{\setminus u:v}; \theta)$	MASS

Unsupervised NMT

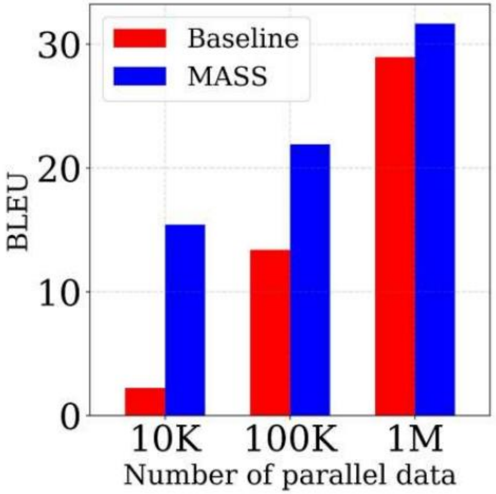




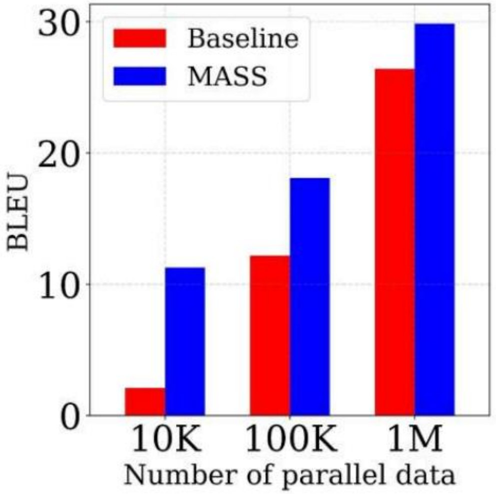


XLM: Cross-lingual language model pretraining, CoRR 2019

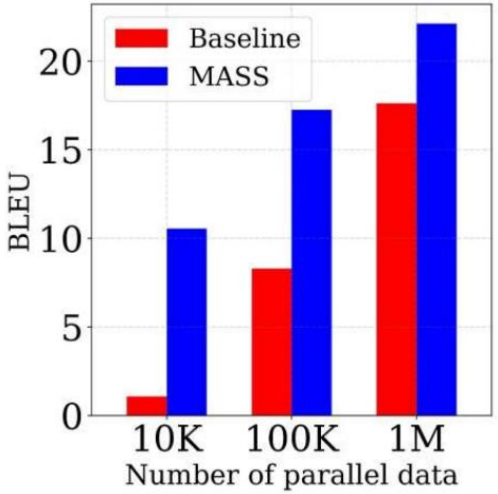
Low-resource NMT



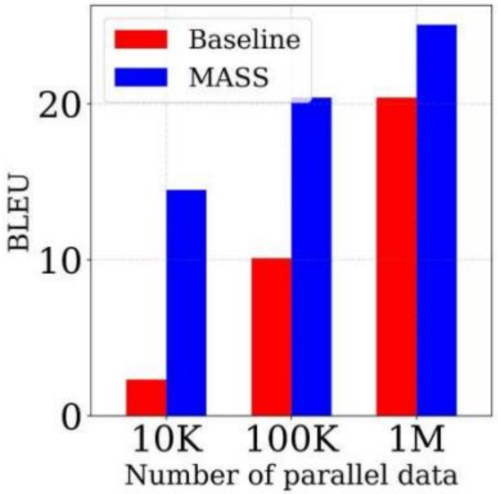
(a) en-fr



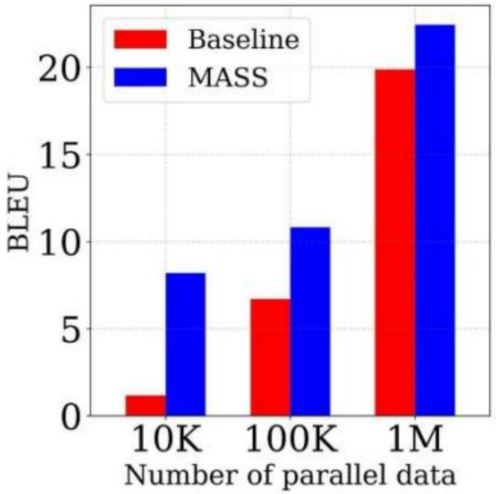
(b) fr-en



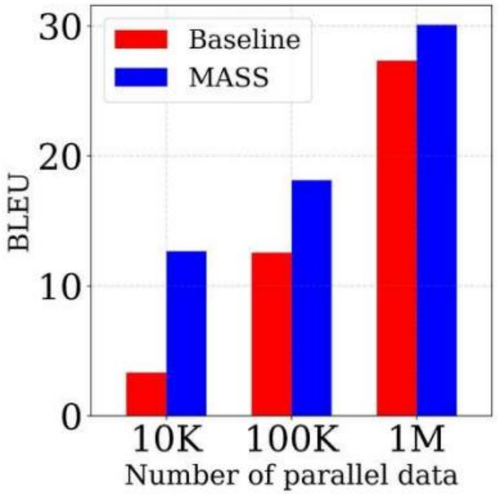
(c) en-de



(d) de-en

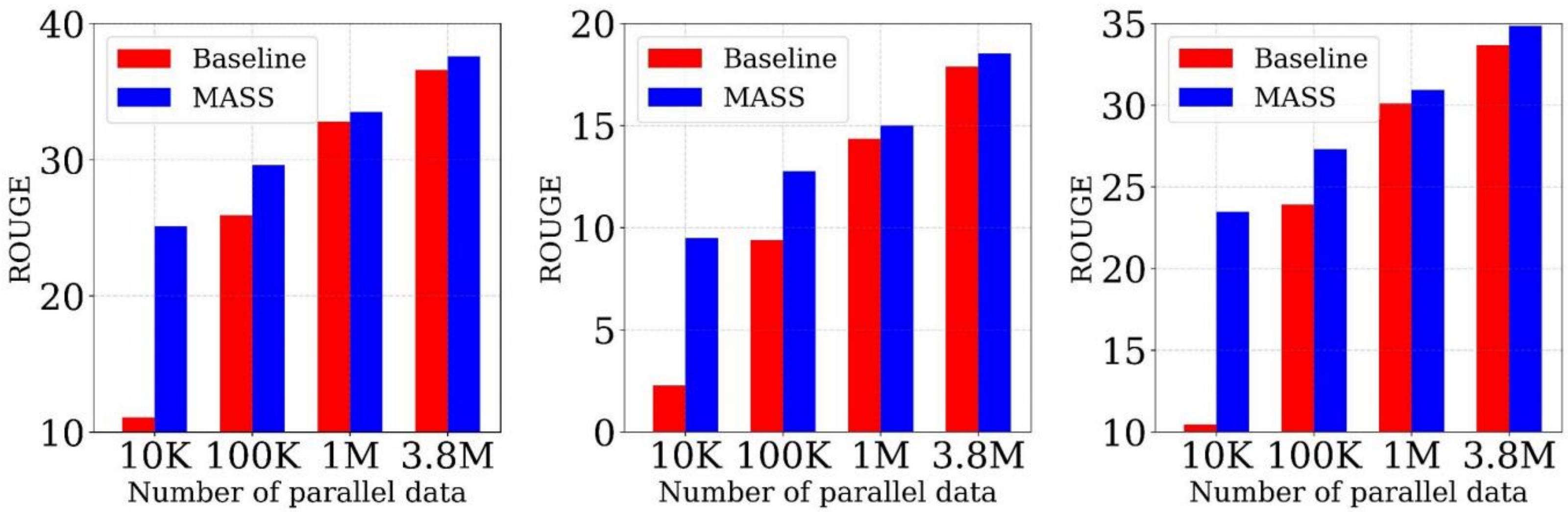


(e) en-ro



(f) ro-en

Text summarization

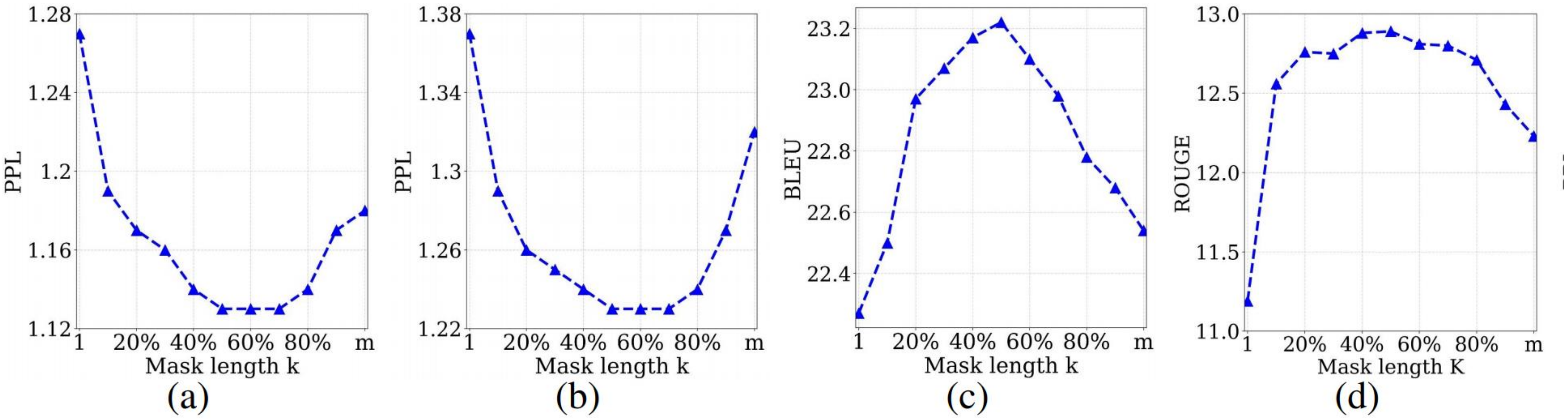


(a) RG-1 (F) (b) RG-2 (F) (c) RG-L (F)

Gigaword Corpus

Analysis of MASS: length of masked segment

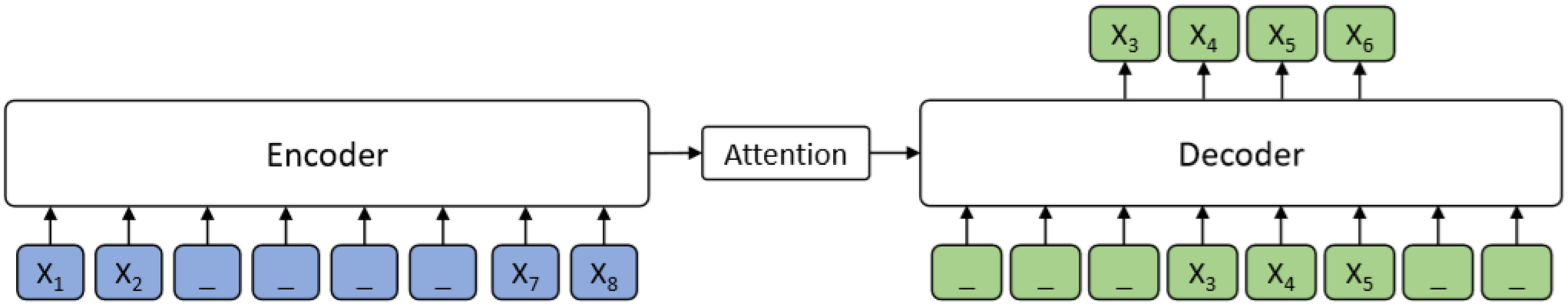
(a), (b): PPL of the pre-trained model on En and Fr
(c): BLEU score of unsupervised En-Fr
(d): ROUGE of text summarization



- K=50%m is a good balance between encoder and decoder
- K=1 (BERT) and K=m (GPT) cannot achieve good performance in language generation tasks.

Summary

- MASS jointly pre-trains the encoder-attention-decoder framework for sequence to sequence based language generation tasks
- MASS achieves significant improvements over the baselines without pre-training or with other pre-training methods on zero/low-resource NMT, text summarization and conversational response generation.



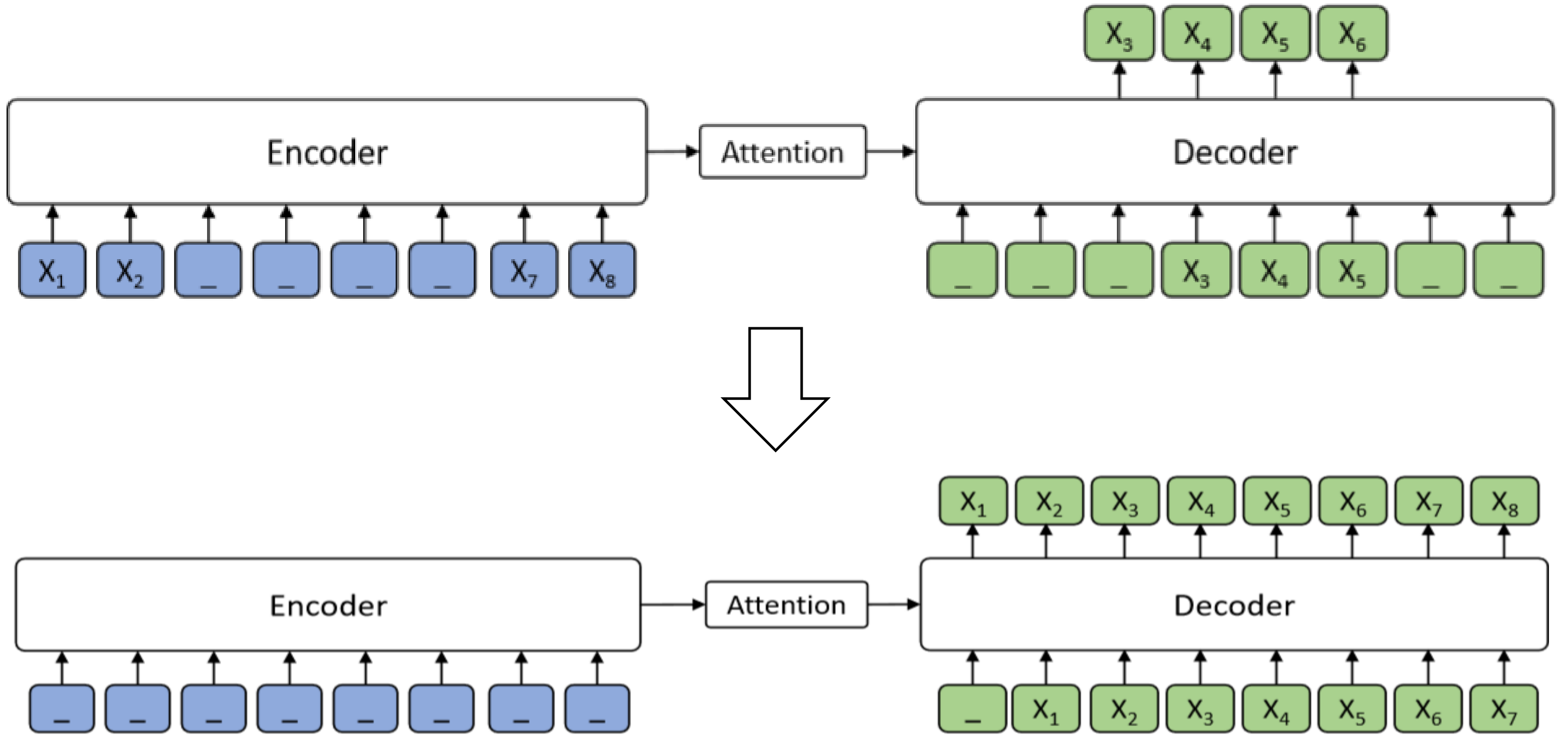
Thanks !

Backup

MASS pre-training

- Model configuration
  - Transformer, 6-6 layer, 1024 embedding.
  - Support cross-lingual tasks such as NMT, as well as monolingual tasks such as text summarization, conversational response generation.
  - English, German, French, Romanian, each language with a tag.
- Datasets
  - We use monolingual corpus from WMT News Crawl. Wikipedia data is also feasible.
  - 190M, 65M, 270M, 2.9M for English, French, German, Romanian.
- Pre-training details
  - K=50%m, 8 V100 GPUs, batch size 3000 tokens/gpu.

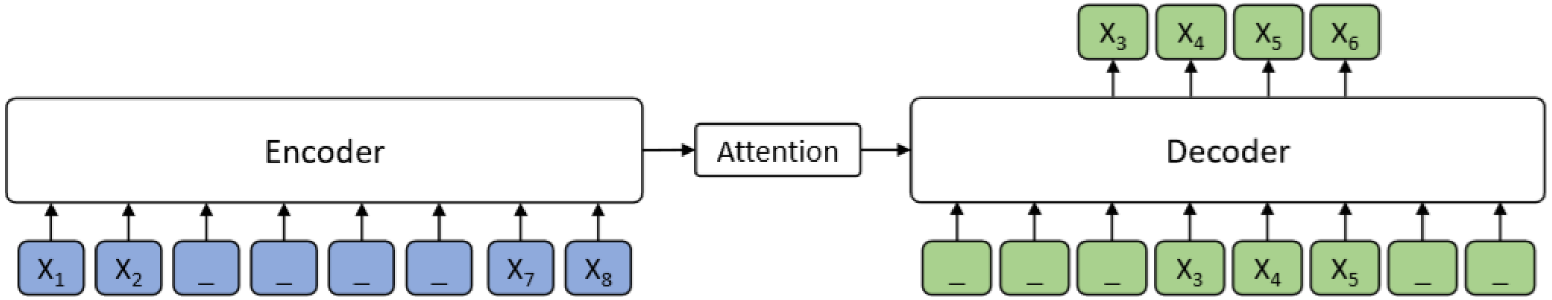
MASS (k=m) ➔ GPT



Length	Probability	Model
$k = m$	$P(x^{1:m} \mid x^{\setminus 1:m}; \theta)$	standard LM in GPT
$k \in [1, m]$	$P(x^{u:v} \mid x^{\setminus u:v}; \theta)$	MASS

Analysis of MASS

- Ablation study of MASS



Method	BLEU	Method	BLEU	Method	BLEU
Discrete	26.76	*Feed*	25.56	MASS	27.41

- Discrete: instead of masking continuous segment, masking discrete tokens
- Feed: Feed the tokens to the decoder that appear in the encoder

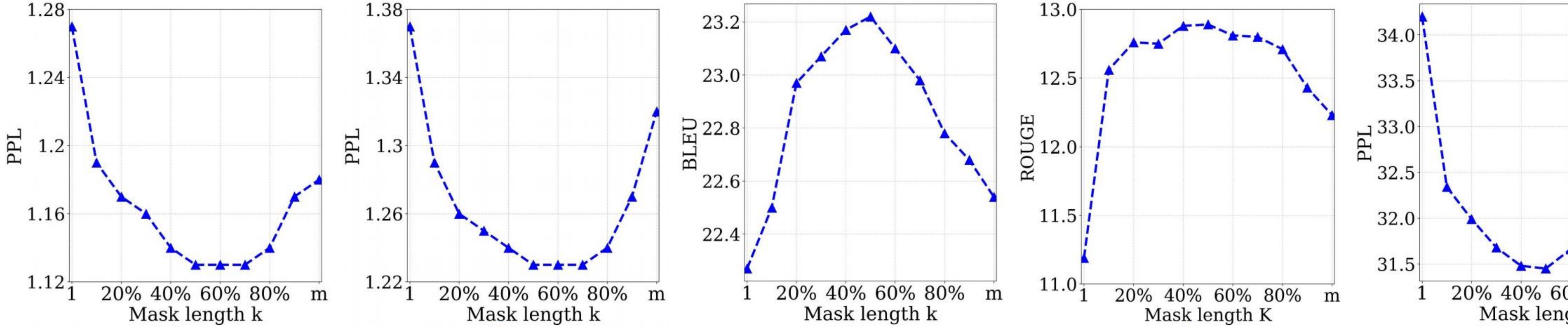
Fine-tuning on conversation response generation

- We fine-tune the model on the Cornell movie dialog corpus, and simply use PPL to measure the performance of response generation.

Method	Data = 10K	Data = 110K
Baseline	82.39	26.38
BERT+LM	80.11	24.84
MASS	**74.32**	**23.52**

Analysis of MASS: length of masked segment

(a), (b): PPL of the pre-trained model on En and Fr
(c): BLEU score of unsupervised En-Fr
(d), (e): ROUGE and PPL on text summarization and response generation



(a) (b) (c) (d) (e)

- K=50%m is a good balance between encoder and decoder
- K=1 (BERT) and K=m (GPT) cannot achieve good performance in language generation tasks.