BERT and Beyond

2019.5.17 Liu Chang

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

- Backbone: Transformer

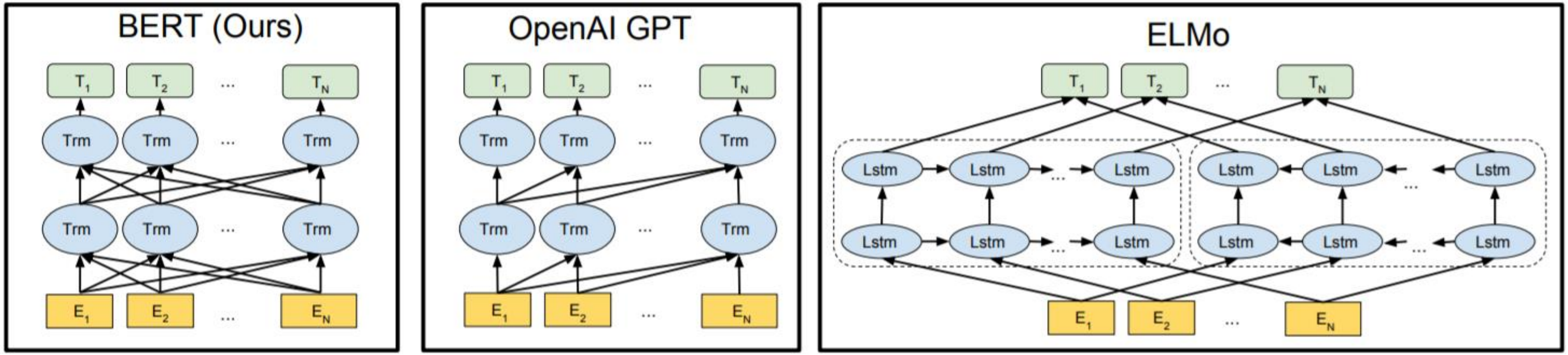
- Training objectives:
  - Masked language modeling(cloze task)
  - Next sentence prediction(Quick thoughts)

- Ability:
  - Text / text pair classification
  - General purpose sentence embeddings / contextualized word embeddings for other tasks, such as sequence labelling…

Language model types:

- ELMo: unidirectional, both left-to-right and right-to-left
- GPT: unidirectional, left-to-right
- BERT: bidirectional

Unidirectional language model is essentially a decoder, while bidirectional language model is an encoder.



Followers:

- What can BERT do, and how?

  Empirical studies about BERT(and other pretrained language models):

  - *Understanding the Behaviors of BERT in Ranking(arxiv)*
  - *Linguistic Knowledge and Transferability of Contextual Representations (NAACL19 long)*
  - *What do you learn from context? Probing for sentence structure in contextualized word representations (ICLR19)*
  - …

Followers:

- What can BERT do for us?

  Using BERT as the backbone to do a wide range of downstream tasks:

  - *BERTSCORE: Evaluating Text Generation with BERT(arxiv)*
  - *BERT for Joint Intent Classification and Slot Filling(arxiv)*
  - *Utilizing BERT for Aspect-Based Sentiment Analysis via Constructing Auxiliary Sentence(NAACL19 short)*
  - …

Followers:

- What can we do for BERT?

  Methods which can be directly incorporated to BERT:

  - *Reducing BERT Pre-Training Time from 3 Days to 76 Minutes(arxiv)*
  - …

Followers:

- What can’t BERT do and how can BERT be improved?

  Modifying model architecture, data type, training objective and … :

  - *Cross-lingual Language Model (arxiv)*

  - *MASS Masked Sequence to Sequence Pre-training for Language Generation (arxiv)*

  - *Unified Language Model Pre-training for natural language understanding and generation (NAACL19 short)*

  - …

Cross-lingual Language Model Pretraining

——Arxiv2019, Facebook

Towards Multilingual BERT

Training objectives

- Causal Language Modeling(CLM): unidirectional, left-to-right
- Masked Language Modeling(MLM): bidirectional
- Translation Language Modeling(TLM): bidirectional and cross lingual

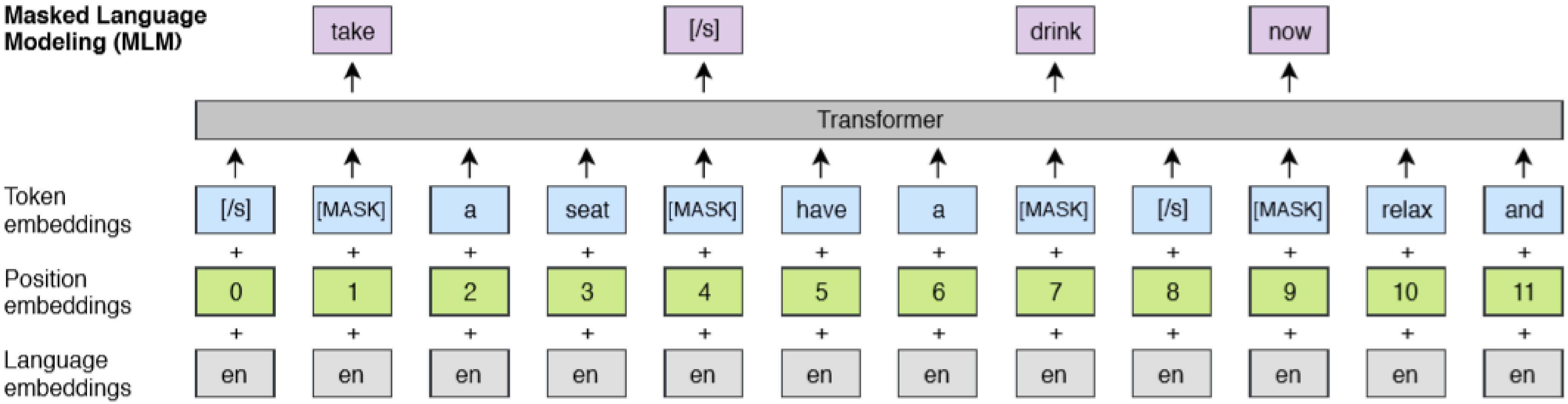
Training objectives

- Causal Language Modeling(CLM):

A Transformer language model trained to model the probability of a word given the previous words in a sentence $P(w_t|w_1,...,w_{t-1};\theta)$.

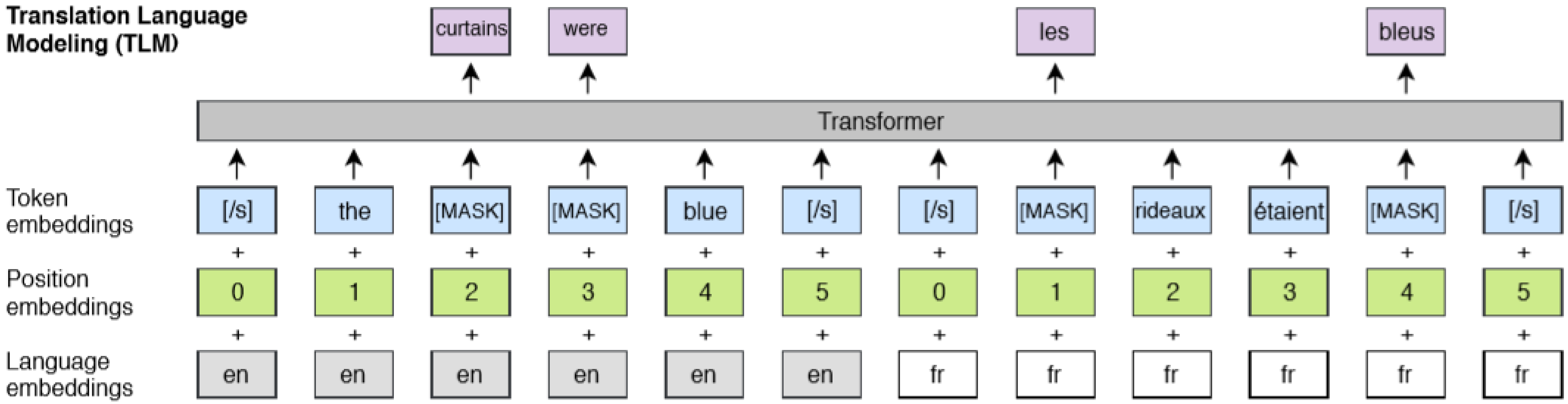
Training objectives

- Masked Language Modeling(MLM)



Training objectives

- Translation Language Modeling(TLM)



Experiments

- Cross-lingual text pair classification(XNLI)

	en	fr	es	de	el	bg	ru	tr	ar	vi	th	zh	hi	sw	ur	Δ
Machine translation baselines (TRANSLATE-TRAIN)																
Devlin et al. (2018)	81.9	-	77.8	75.9	-	-	-	-	70.7	-	-	76.6	-	-	61.6	-
XLM (MLM+TLM)	85.0	80.2	80.8	80.3	78.1	79.3	78.1	74.7	76.5	76.6	75.5	78.6	72.3	70.9	63.2	76.7
Machine translation baselines (TRANSLATE-TEST)																
Devlin et al. (2018)	81.4	-	74.9	74.4	-	-	-	-	70.4	-	-	70.1	-	-	62.1	-
XLM (MLM+TLM)	85.0	79.0	79.5	78.1	77.8	77.6	75.5	73.7	73.7	70.8	70.4	73.6	69.0	64.7	65.1	74.2
Evaluation of cross-lingual sentence encoders																
Conneau et al. (2018b)	73.7	67.7	68.7	67.7	68.9	67.9	65.4	64.2	64.8	66.4	64.1	65.8	64.1	55.7	58.4	65.6
Devlin et al. (2018)	81.4	-	74.3	70.5	-	-	-	-	62.1	-	-	63.8	-	-	58.3	-
Artetxe and Schwenk (2018)	73.9	71.9	72.9	72.6	73.1	74.2	71.5	69.7	71.4	72.0	69.2	71.4	65.5	62.2	61.0	70.2
XLM (MLM)	83.2	76.5	76.3	74.2	73.1	74.0	73.1	67.8	68.5	71.2	69.2	71.9	65.7	64.6	63.4	71.5
XLM (MLM+TLM)	**85.0**	**78.7**	**78.9**	**77.8**	**76.6**	**77.4**	**75.3**	**72.5**	**73.1**	**76.1**	**73.2**	**76.5**	**69.6**	**68.4**	**67.3**	**75.1**

Table 1: **Results on cross-lingual classification accuracy.** Test accuracy on the 15 XNLI languages. We report results for machine translation baselines and zero-shot classification approaches based on cross-lingual sentence encoders. XLM (MLM) corresponds to our unsupervised approach trained only on monolingual corpora, and XLM (MLM+TLM) corresponds to our supervised method that leverages both monolingual and parallel data through the TLM objective. Δ corresponds to the average accuracy.

Experiments

- Unsupervised machine translation
- Supervised machine translation
- Low-resource language model
- Unsupervised cross-lingual word embeddings

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

——ICML2019, NJU, MSRA

Towards Seq2seq BERT

Motivation:

- BERT: bidirectional LM, only encoder
- GPT: unidirectional LM (standard LM), only decoder
- If unified together → Pretrained model for Seq2seq tasks.

Model:

- Architecture:
  - Mask: bidirectional LM

    for encoder’s self-attention and decoder-encoder attention
  - Sequence prediction: unidirectional LM (standard LM)

    for decoder’s generation ability

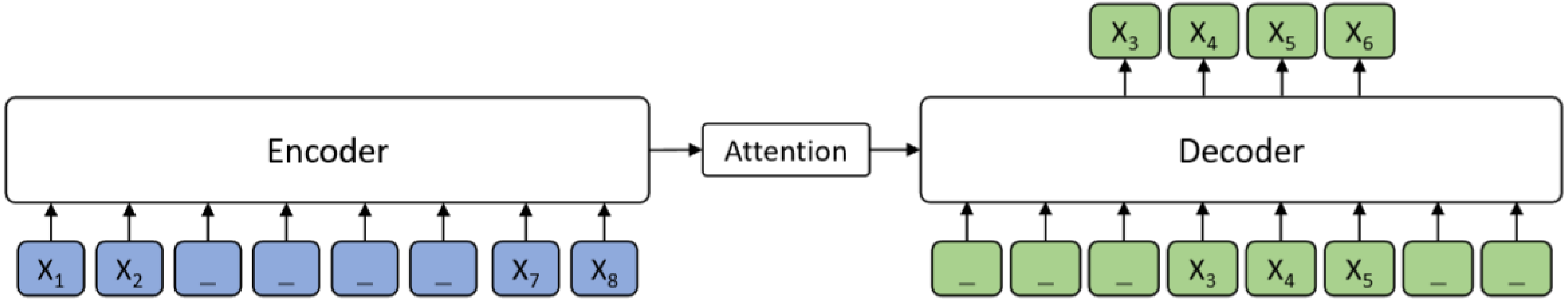


Figure 1. The encoder-decoder framework for our proposed MASS. The token “_” represents the mask symbol [M].

Model:

- Architecture:
  - Mask: bidirectional LM
    for encoder's self-attention and decoder-encoder attention
  - Sequence prediction: unidirectional LM (standard LM)
    for decoder's generation ability

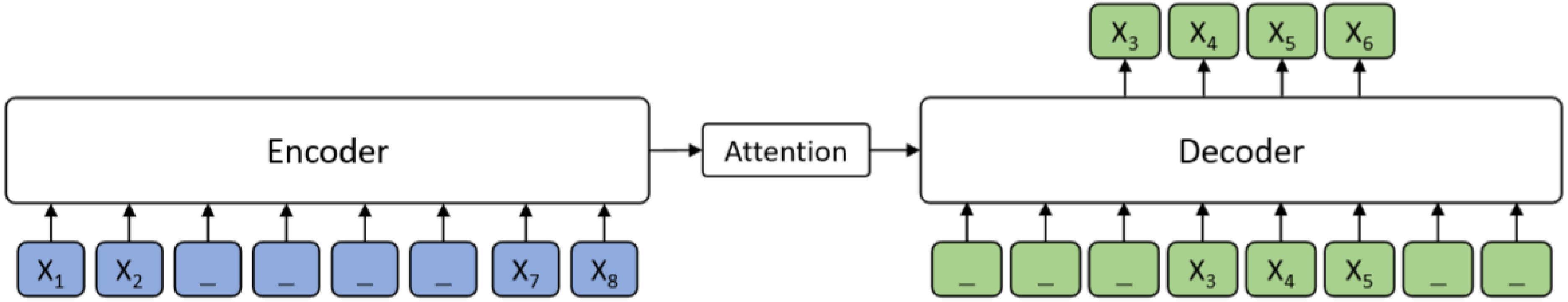


Figure 1. The encoder-decoder framework for our proposed MASS. The token "_" represents the mask symbol [M].

Model:

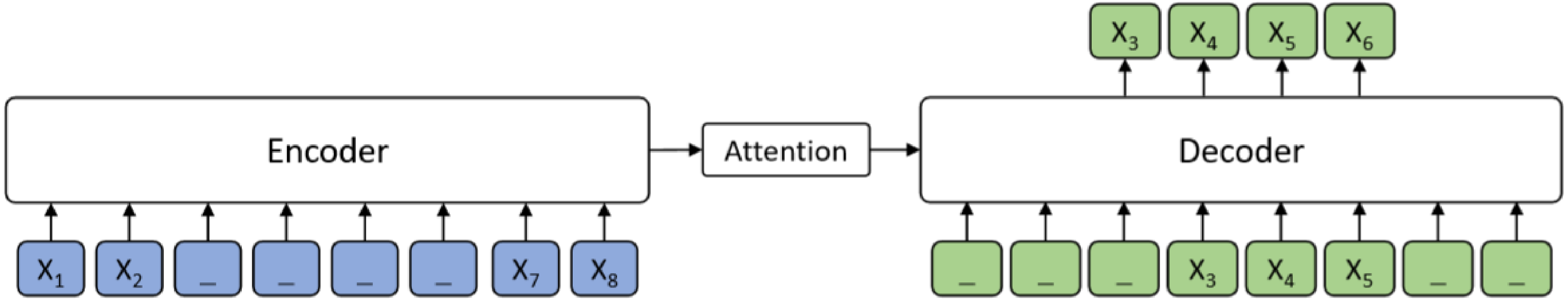


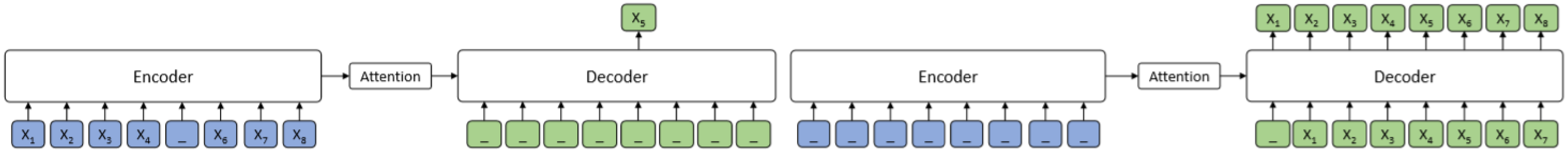
Figure 1. The encoder-decoder framework for our proposed MASS. The token "_" represents the mask symbol $[\mathbb{M}]$.

- Objective function:

$$L(\theta; \mathcal{X}) = \frac{1}{|\mathcal{X}|}\Sigma_{x\in\mathcal{X}} \log P(x^{u:v}|x^{\setminus u:v};\theta)$$

$$= \frac{1}{|\mathcal{X}|}\Sigma_{x\in\mathcal{X}} \log \prod_{t=u}^{v} P(x_t^{u:v}|x_{<t}^{u:v}, x^{\setminus u:v};\theta).$$

Unifying BERT and GPT:



(a) Masked language modeling in BERT ($k = 1$) (b) Standard language modeling ($k = m$)

Figure 2. The model structure of MASS when $k = 1$ and $k = m$. Masked language modeling in BERT can be viewed as the case $k = 1$ and standard language modeling can be viewed as the case $k = m$.

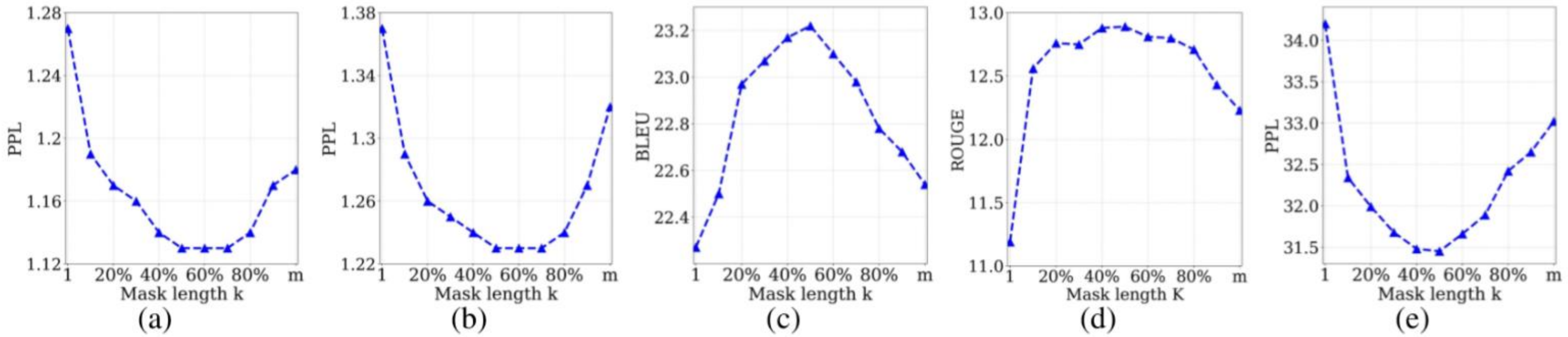
- k=1: becomes BERT

  The decoder can be considered as a non-linear classifier, analogous to the softmax matrix used in BERT.

- k=m: becomes GPT
  - The decoder-encoder attention can’t bring useful information, only noise.
- 1<k<m: methods in between

Unifying BERT and GPT:

- The choice of k: m/2 results in the best results in a wide range of tasks



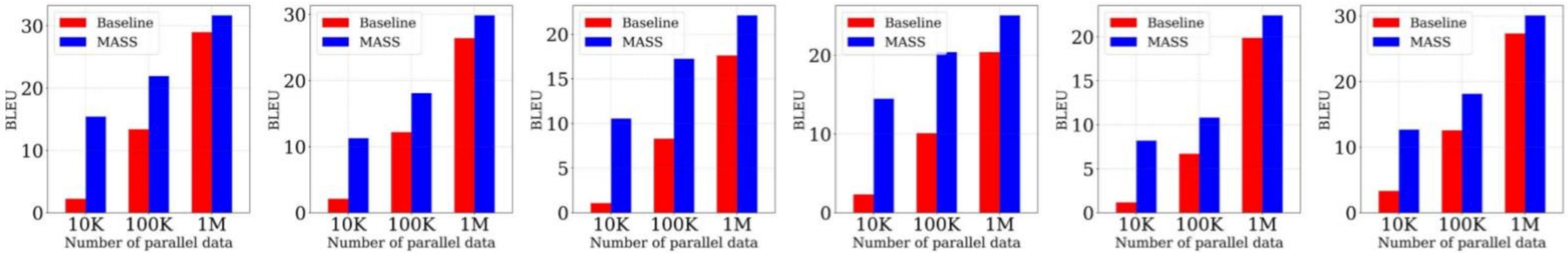
Pretraining details

- Dataset:
  - 50M for English, German and French respectively, 2.9M for Romanian.
  - BPE encoding and vocabulary sharing. (as XLM)

- Other details:
  - Removing the padding in the decoder (the masked tokens) but keeping the positional embedding of the unmasked tokens unchanged.
  - The fragment length k is set as roughly 50% of the total number of tokens in the sentence. (m/2)

Experiments

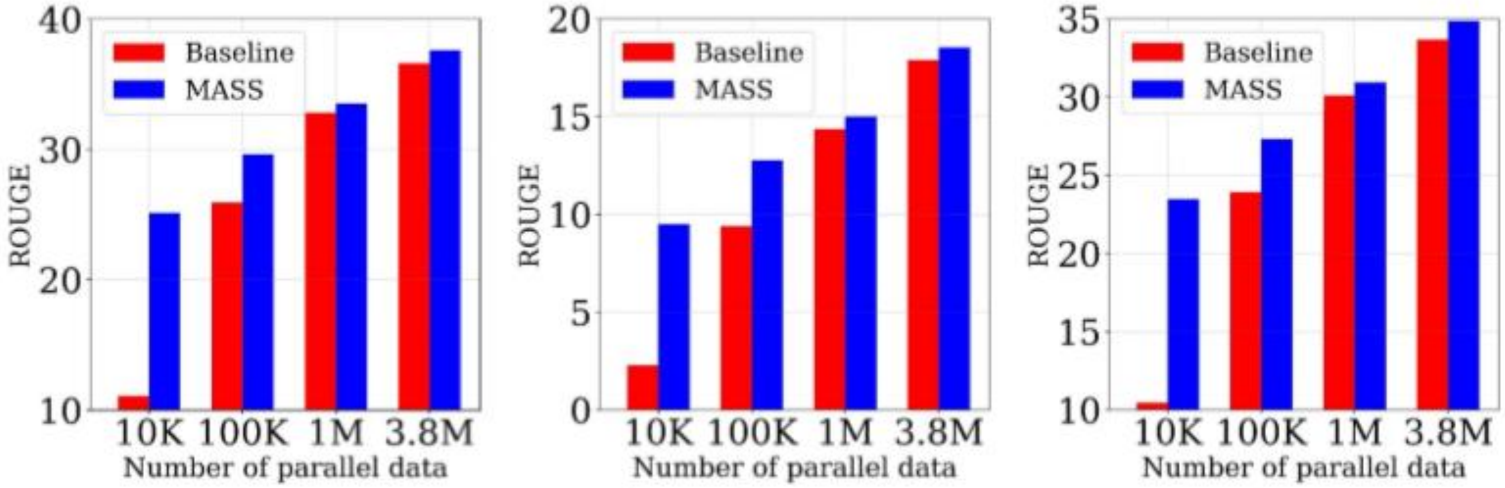
- Unsupervised NMT
  - Monolingual data only, with back translation
- Low-Resource NMT
  - Respectively sampling 10K, 100K, 1M paired sentence from the bilingual training data



(a) en-fr (b) fr-en (c) en-de (d) de-en (e) en-ro (f) ro-en

Experiments

- Text Summarization(ROUGE)



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

Experiments

- Conversational Response Generation(PPL)

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

Unified Language Model Pre-training for Natural Language Understanding and Generation

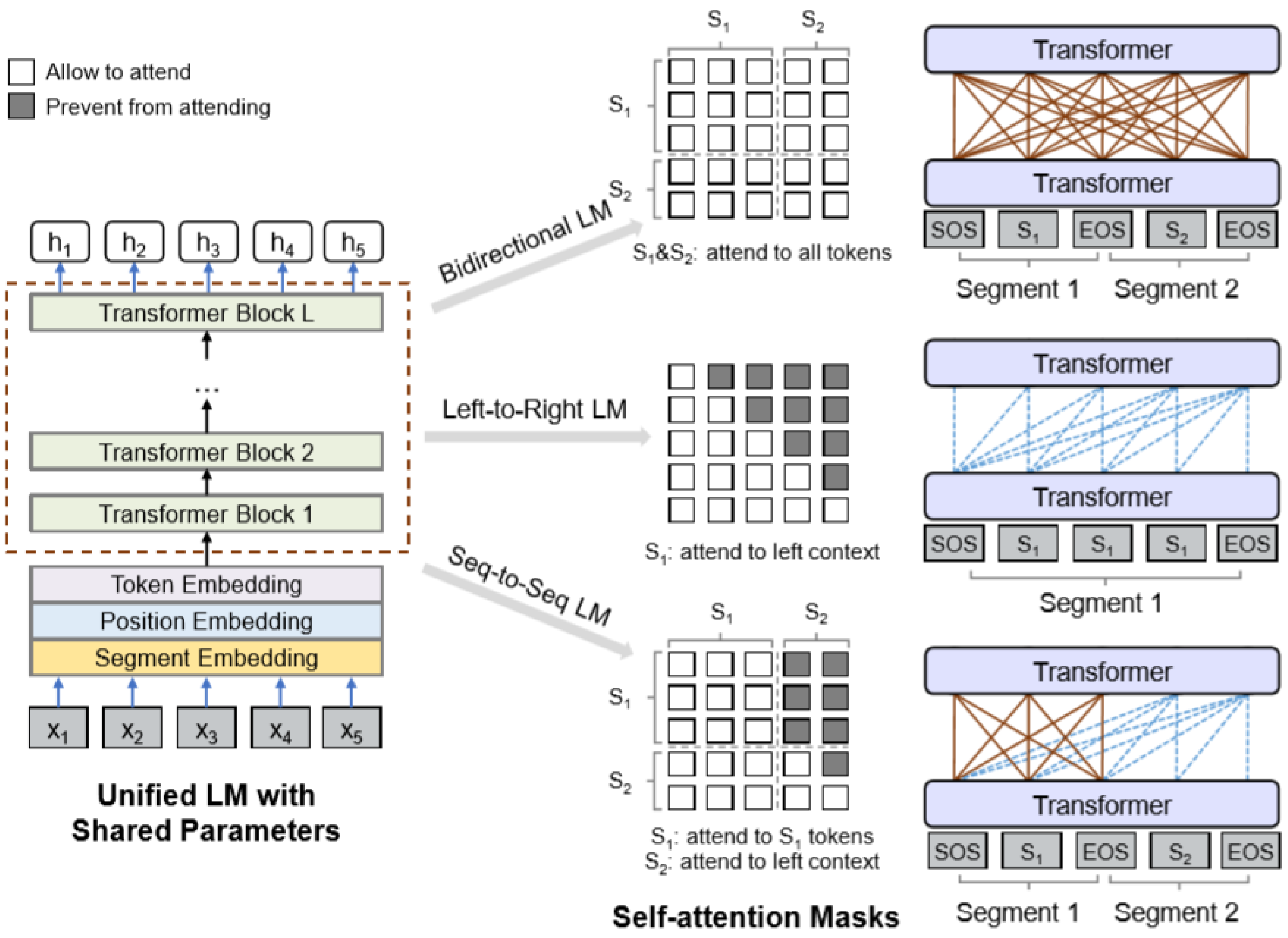
——arxiv, MSRA

Different types of LMs:

	ELMo	GPT	BERT	UniLM
Left-to-Right LM	✓	✓		✓
Right-to-Left LM	✓			✓
Bidirectional LM			✓	✓
Seq-to-Seq LM				✓

Backbone Network	**LM Objectives of Unified Pre-training**	**What Unified LM Learns**	**Example Downstream Tasks**
Transformer with shared parameters for all LM objectives	Bidirectional LM	Bidirectional encoding	GLUE benchmark Extractive question answering
	Unidirectional LM	Unidirectional decoding	Long text generation
	Sequence-to-Sequence LM	Unidirectional decoding conditioned on bidirectional encoding	Abstractive summarization Question generation Generative question answering

Training objectives: 4 LM and quick-thoughts



Experiments

- For natural language understanding:

  compares favorably with BERT on the GLUE benchmark.

- For natural language generation:
  - improving the CNN/DailyMail abstractive summarization ROUGE-L to 40.63 (2.16 absolute improvement),
  - pushing the CoQA generative question answering F1 score to82.5(37.1 absolute improvement),
  - and the SQuAD question generation BLEU-4 to22.88(6.50 absolute improvement).

Thank you for listening!