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

Github Link:

Li Dong* Nan Yang* Wenhui Wang* Furu Wei*† Xiaodong Liu Yu Wang
Jianfeng Gao Ming Zhou Hsiao-Wuen Hon
Microsoft Research
{lidong1,nanya,wenwan,fuwei}@microsoft.com
{xiaodl,yuwan,jfgao,mingzhou,hon}@microsoft.com

Abstract

Unified Modeling: shared Transformer network and utilizing specific **self-attention masks** to control what context the prediction conditions on

Unified Pre-training: cloze-style tasks for language model (LM) pre-training
- Left-to-right unidirectional LM
- Right-to-left unidirectional LM
- Bidirectional LM
- Sequence-to-sequence LM

Unified Fine-tuning: UniLM (the **same** model) can be fine-tuned as a **unidirectional decoder**, a **bidirectional encoder**, or a **sequence-to-sequence model** to support various downstream natural language **understanding** and **generation** tasks

Overview

- Comparison between language model (LM) pre-training objectives:

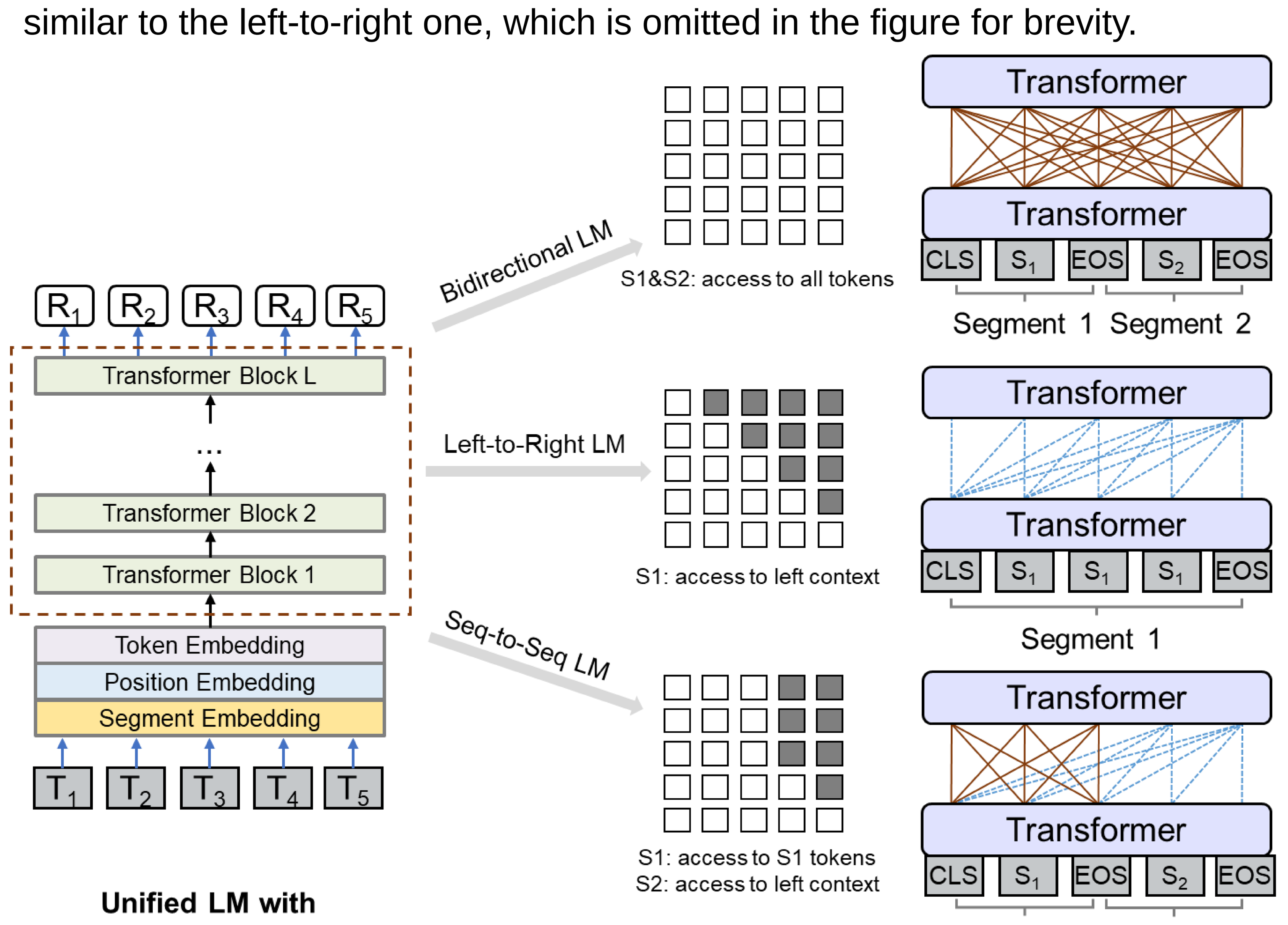
	ELMo	GPT	BERT	UniLM
Left-to-Right LM	✓	✓		✓
Right-to-Left LM	✓			✓
Bidirectional LM			✓	✓
Sequence-to-Sequence LM				✓

- The unified LM is jointly pre-trained by multiple language modeling objectives, sharing the same parameters. We fine-tune and evaluate the pre-trained unified LM on various datasets, including both language understanding and generation tasks.

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

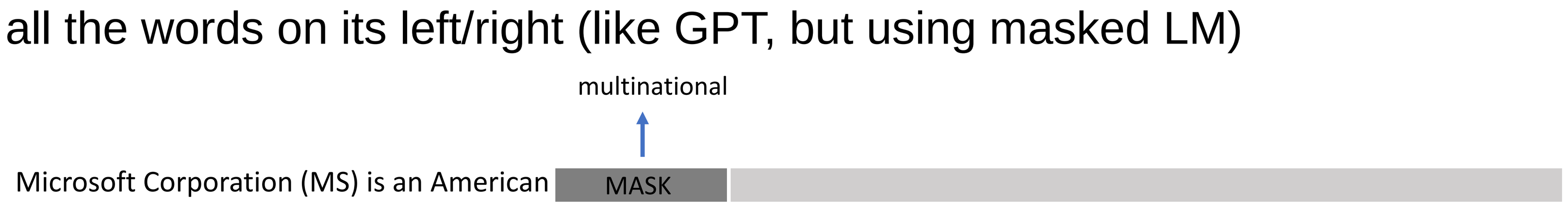
UniLM Pre-training

- The model parameters are shared across the LM objectives (i.e., bidirectional LM, unidirectional LM, and sequence-to-sequence LM). We use different self-attention masks to control the access to context for each word token. The right-to-left LM is similar to the left-to-right one, which is omitted in the figure for brevity.

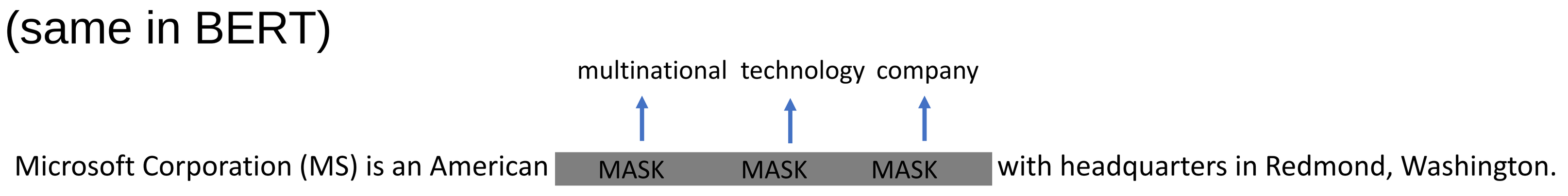


Unified LM with Shared Parameters **Self-attention Masks**

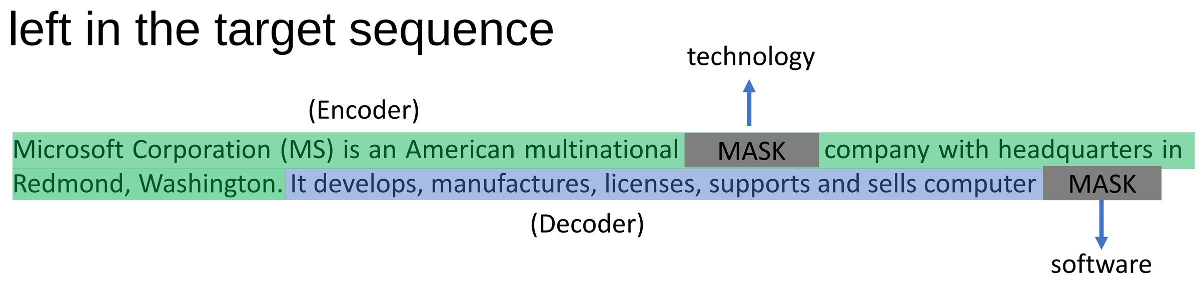
- Pre-training tasks
  - Unidirectional LM: the context of the masked word to be predicted consists of all the words on its left/right (like GPT, but using masked LM)



  - Bidirectional LM: the context consists of the words on both the right and the left (same in BERT)



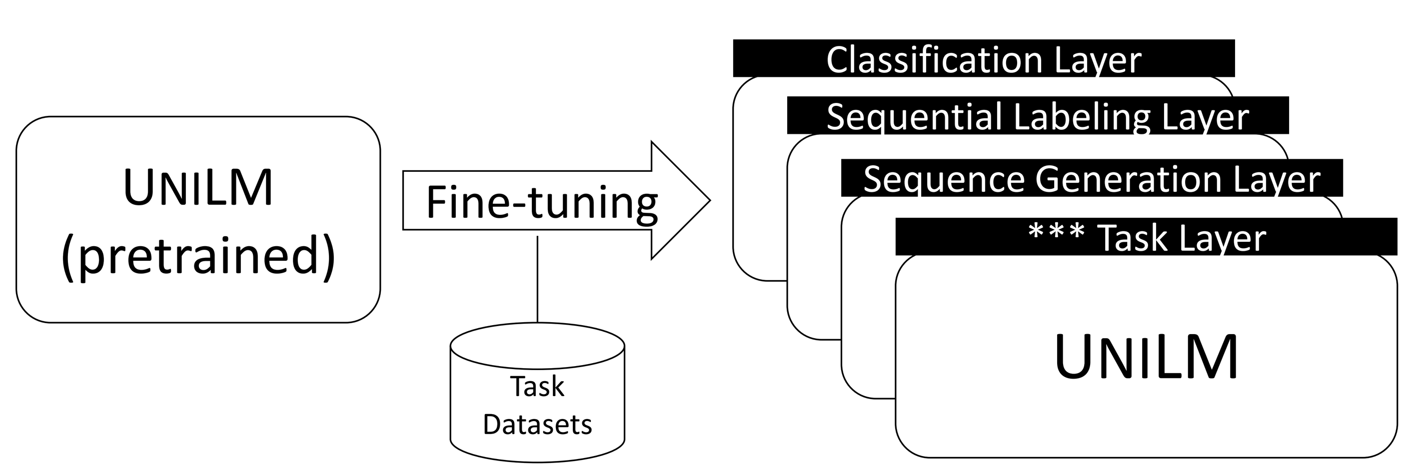
  - Sequence-to-sequence LM: the context of the to-be-predicted word in the target sequence consists of all the words in the source sequence and the words on the its left in the target sequence



- Data: Wikipedia (11G) + BookCorpus (4G)

UniLM Fine-tuning

- Adding simple task-specific layers upon UNILM
- Fine-tuning several epochs on the downstream task
- Natural language understanding tasks (e.g., classification, and sequential labeling)
  - Fine-tune UniLM as a bidirectional Transformer encoder
- Natural language generation tasks
  - Fine-tune UniLM as a sequence-to-sequence model



Experiments

- Abstractive Summarization
  - Document -> summary: sequence-to-sequence fine-tuning

CNN / Dailymail

	RG-1	RG-2	RG-L
Extractive Summarization			
LEAD-3	40.42	17.62	36.67
Best Extractive [27]	43.25	**20.24**	39.63
Abstractive Summarization			
PGNet [37]	39.53	17.28	37.98
Bottom-Up [16]	41.22	18.68	38.34
S2S-ELMo [13]	41.56	18.94	38.47
UniLM	**43.33**	20.21	**40.51**

Gigaword

	RG-1	RG-2	RG-L
10K Training Examples			
Transformer [43]	10.97	2.23	10.42
MASS [39]	25.03	9.48	23.48
UniLM	**32.96**	**14.68**	**30.56**
Full Training Set			
OpenNMT [23]	36.73	17.86	33.68
Re3Sum [4]	37.04	19.03	34.46
MASS [39]	37.66	18.53	34.89
UniLM	**38.45**	**19.45**	**35.75**

- Question Answering (QA)
  - Extractive QA: classify answer spans

Stanford Question Answering Dataset (SQuAD)

	EM	F1
RMR+ELMo [20]	71.4	73.7
$\text{BERT}_{\text{LARGE}}$	78.9	81.8
UniLM	**80.5**	**83.4**

Conversational Question Answering (CoQA)

	F1
DrQA+ELMo [35]	67.2
$\text{BERT}_{\text{LARGE}}$	82.7
UniLM	**84.9**

  - Generative QA: generate answers as a sequence-to-sequence model

Conversational Question Answering (CoQA)

	F1
Seq2Seq [35]	27.5
PGNet [35]	45.4
UniLM	**82.5**

- Question Generation
  - Generate a question that asks for the given passage and answer

Stanford Question Answering Dataset (SQuAD)

	BLEU-4	MTR	RG-L
CorefNQG [11]	15.16	19.12	-
SemQG [50]	18.37	22.65	46.68
UniLM	**22.12**	**25.06**	**51.07**
MP-GSN [51]	16.38	20.25	44.48
SemQG [50]	20.76	24.20	48.91
UniLM	**23.75**	**25.61**	**52.04**

  - Question generation based on UniLM improves question answering results

Stanford Question Answering Dataset (SQuAD)

	EM	F1
UniLM QA Model (Section 3.2)	80.5	83.4
+ UniLM Generated Questions	**84.7**	**87.6**

- Dialog Response Generation
  - multi-turn conversation history + document -> response

DSTC7 Shared Task

	NIST-4	BLEU-4	METEOR	Entropy-4	Div-1	Div-2	Avg len
Best System in DSTC7 Shared Task	2.523	1.83	8.07	9.030	0.109	0.325	15.133
UniLM	**2.669**	**4.39**	**8.27**	**9.195**	**0.120**	**0.391**	14.807
Human Performance	2.650	3.13	8.31	10.445	0.167	0.670	18.76

- GLUE Benchmark
  - a collection of nine language understanding tasks, including question answering, linguistic acceptability, sentiment analysis, text similarity, paraphrase detection, and natural language inference

General Language Understanding Evaluation (GLUE)

Model	CoLA MCC	SST-2 Acc	MRPC F1	STS-B S Corr	QQP F1	MNLI-m/mm Acc	QNLI Acc	RTE Acc	WNLI Acc	AX Acc	**Score**
GPT	45.4	91.3	82.3	80.0	70.3	82.1/81.4	87.4	56.0	53.4	29.8	72.8
$\text{BERT}_{\text{LARGE}}$	60.5	**94.9**	89.3	86.5	**72.1**	86.7/**85.9**	**92.7**	70.1	65.1	39.6	80.5
UniLM	**61.1**	94.5	**90.0**	**87.7**	71.7	**87.0/85.9**	**92.7**	**70.9**	65.1	38.4	**80.8**