Distilling Knowledge Learned in BERT for Text Generation

Yen-Chun Chen[1], Zhe Gan[1], Yu Cheng[1], Jingzhou Liu[2], Jingjing Liu[1]

[1]Microsoft D365 AI Research

[2]Carnegie Mellon University

BERT is Dominating NLU

GLUE

Rank	Name	Model	Score
1	PING-AN Omni-Sinitic	ALBERT + DAAF + NAS	90.6
2	ERNIE Team - Baidu	ERNIE	90.4
3	Alibaba DAMO NLP	StructBERT	90.3
4	T5 Team - Google	T5	90.3
5	Microsoft D365 AI & MSR AI & GATECH	MT-DNN-SMART	89.9
6	Zihang Dai	Funnel-Transformer (Ensemble B10-10-10H1024)	89.7
7	ELECTRA Team	ELECTRA-Large + Standard Tricks	89.4
8	Huawei Noah's Ark Lab	NEZHA-Large	88.7
9	Microsoft D365 AI & UMD	FreeLB-RoBERTa (ensemble)	88.4

SQuAD2.0
The Stanford Question Answering Dataset

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Apr 06, 2020	SA-Net on Albert (ensemble) *QIANXIN*	**90.724**	**93.011**
2 May 05, 2020	SA-Net-V2 (ensemble) *QIANXIN*	90.679	92.948
2 Apr 05, 2020	Retro-Reader (ensemble) *Shanghai Jiao Tong University* http://arxiv.org/abs/2001.09694v2	90.578	92.978
3 May 04, 2020	ELECTRA+ALBERT+EntitySpanFocus (ensemble) *SRCB_DML*	90.442	92.839
4 Mar 12, 2020	ALBERT + DAAF + Verifier (ensemble) *PINGAN Omni-Sinitic*	90.386	92.777
5 Jan 10, 2020	Retro-Reader on ALBERT (ensemble) *Shanghai Jiao Tong University* http://arxiv.org/abs/2001.09694v2	90.115	92.580

HellaSwag

Rank	Model	Overall accuracy	In-domain accuracy	Zero-shot accuracy	ActivityNet accuracy	WikiHow accuracy
	Human Performance *University of Washington* (Zellers et al. '19)	95.6	95.6	95.7	94.0	96.5
March 23, 2020	ALUM *MSR* https://github.com/namisan/mt-dnn	**85.6**	86.5	**84.6**	**77.1**	90.1
2 July 25, 2019	RoBERTa *Facebook AI*	85.2	**87.3**	83.1	74.6	**90.9**
3 February 7, 2020	G-DAug-inf *Anonymous*	83.7	85.6	81.8	73.0	89.6
4 January 19, 2020	HighOrderGN + RoBERTa *USC MOWGLI/INK Lab*	82.2	84.3	80.2	71.5	88.1
5 July 25, 2019	Grover-Mega *Univerity of Washington* https://rowanzellers.com/grover	75.4	79.1	71.7	64.8	81.2
6 July 25, 2019	Grover-Large *Univerity of Washington* https://rowanzellers.com/grover	57.2	60.7	53.6	53.3	59.2

Wang et al., "GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding", EMNLP 2018
Rajpurkar et al., "Know What You Don't Know: Unanswerable Questions for SQuAD", ACL 2018
Zellers et al., "HellaSwag: Can a Machine Really Finish Your Sentence?", ACL 2019

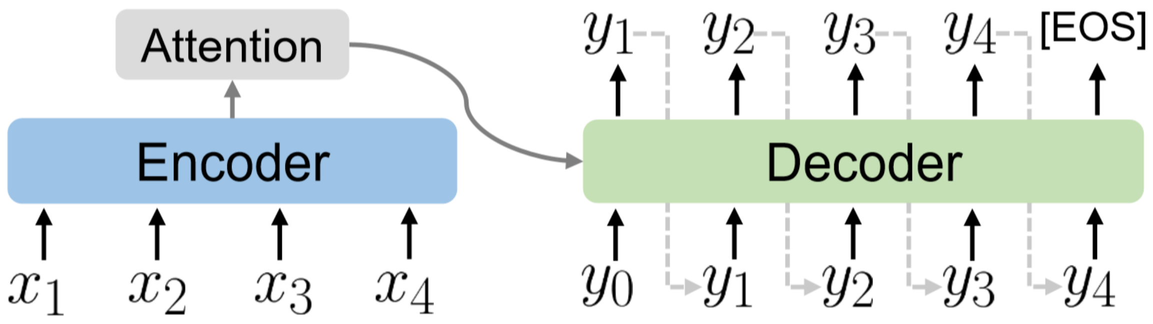
What about Text Generation?

- Machine Translation
  - Bing Microsoft Translator, Google Translate
- Automatic Text Summarization
- Image Captioning / Alt Text

SOTA: Sequence-to-Sequence (Seq2Seq)

- Predict **one word** at a time, from **left to right**
- Conditional Language Model

$$\mathcal{L}_{xe}(\theta) = -\log P_\theta(Y|X) \tag{1}$$
$$= -\sum_{t=1}^{N} \log P_\theta(y_t|y_{1:t-1}, X)\,,$$



Sutskever et al., “Sequence to Sequence Learning with Neural Networks”, NeurIPS 2014

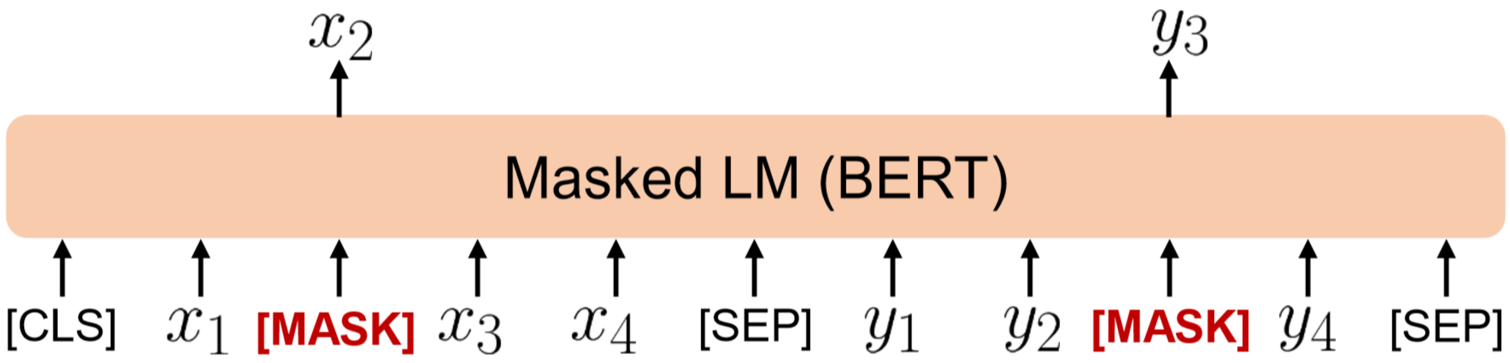
Why is Seq2Seq Insufficient?

- Left-to-right only: context on the right is not utilized
- BERT is bidirectional and encodes rich contextual information from large-scale corpus
- Can we apply BERT to Text Generation?

BERT: Masked Language Modeling (MLM)

- Predict 15% of masked tokens at the same time

$$P(x_1^m, \ldots, x_i^m, y_1^m, \ldots, y_j^m | X^u, Y^u), \qquad (2)$$



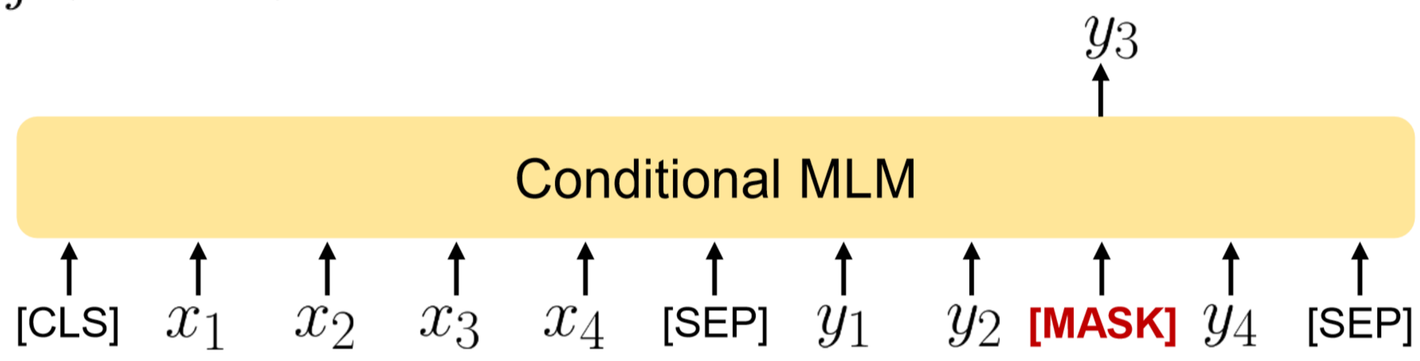
- Why not apply BERT to text generation directly?
  - Can't go conditional or sequential during inference

Devlin et al., "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", NAACL 2019

Our Proposal: Conditional MLM (C-MLM)

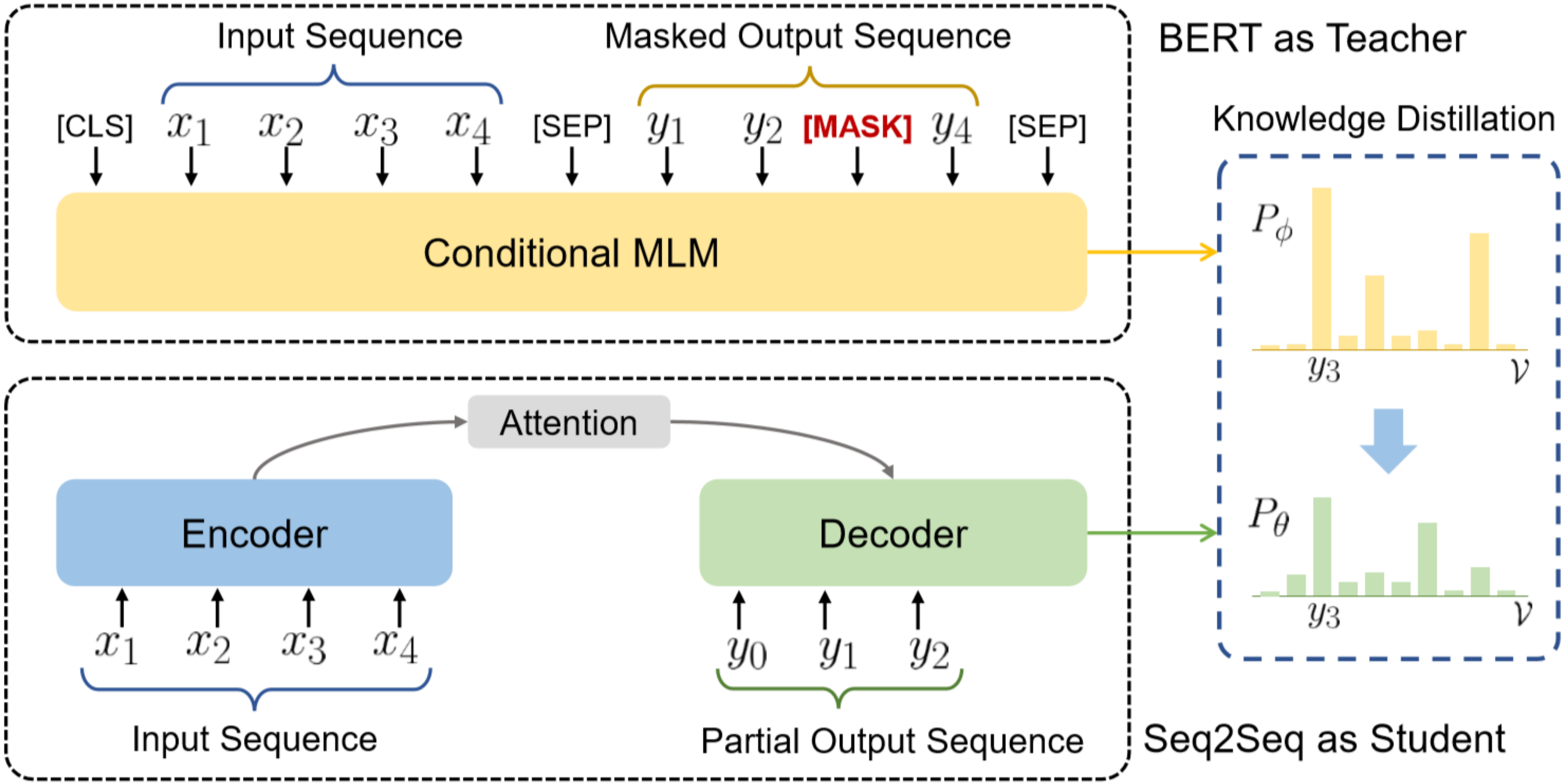
- We propose an MLM variant for Seq2Seq generation

$$P(y_1^m, \ldots, y_j^m | X, Y^u) . \qquad (3)$$

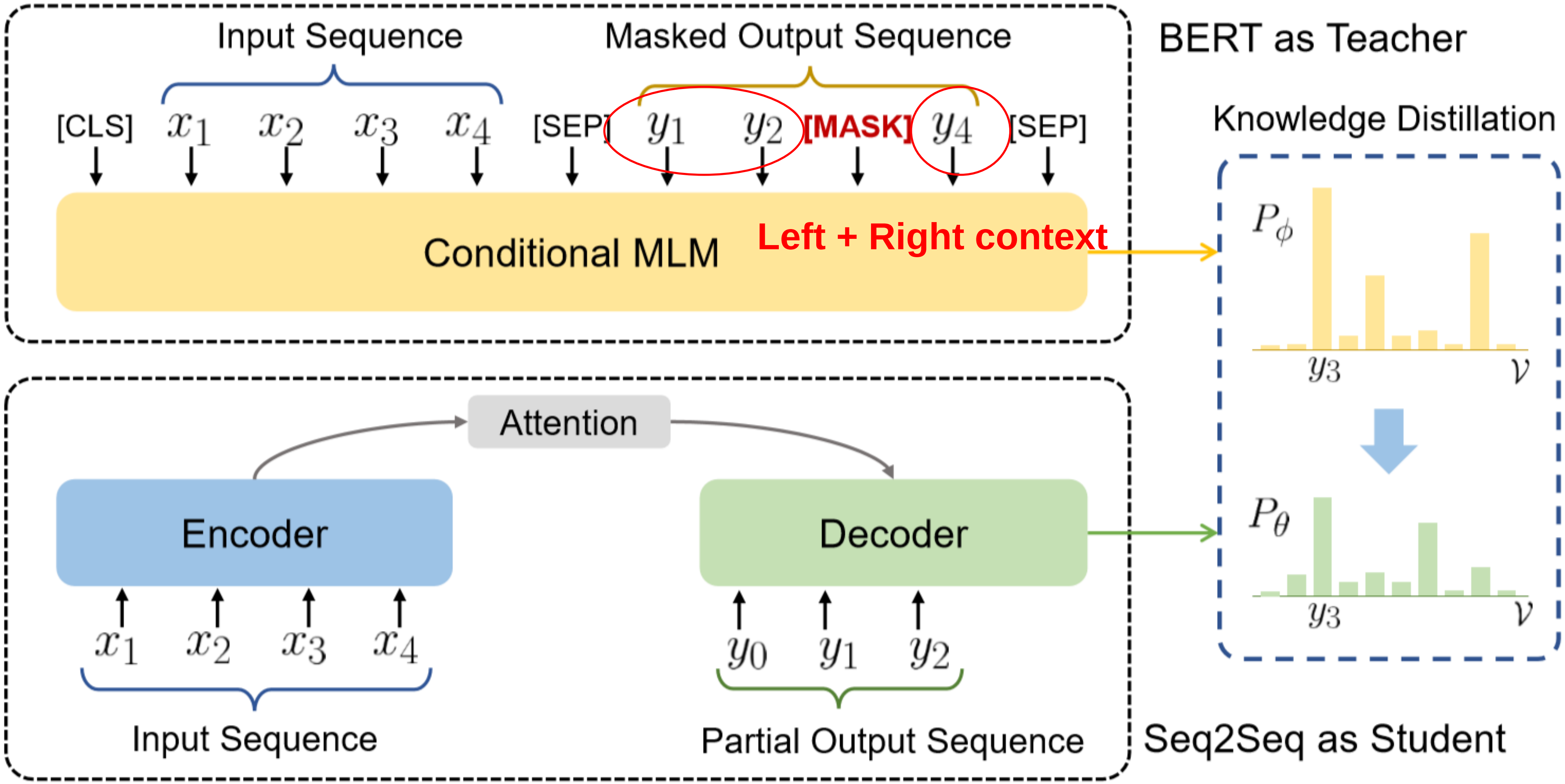


- Now it's conditional, but how to make it sequential for text generation?

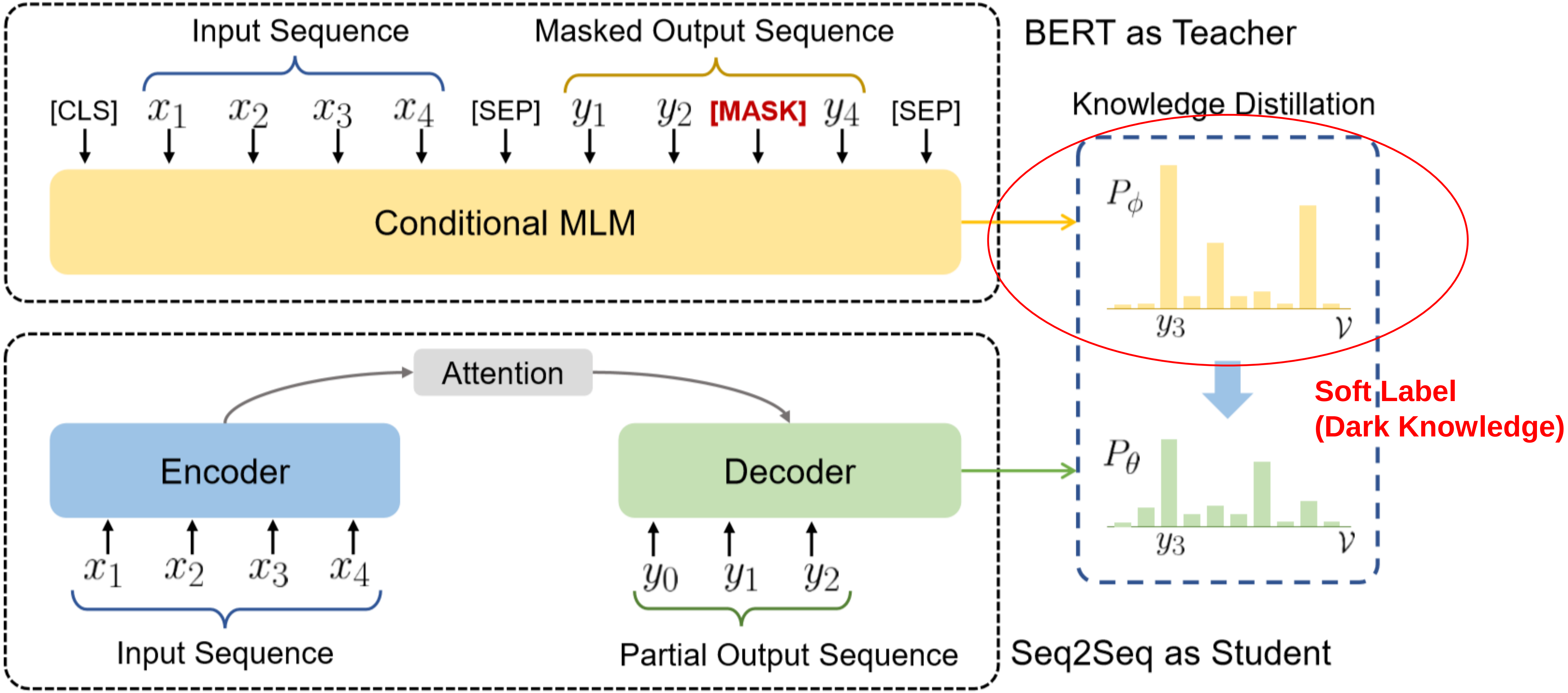
Solution: Seq2Seq Enhanced by C-MLM



Looking into the Future (Right Context)



Dark Knowledge (Soft Label)



Hinton et al., “Distilling the Knowledge in a Neural Network”, 2015

Experiments

- Text Generation Tasks
  - Machine Translation (MT)
    - Small (IWSLT, <200k): English to Vietnamese (En-Vi), German to English (De-En)
    - Medium (WMT, ~4.5M): English to German (En-De)
  - Abstractive Summarization
    - Gigaword summarization (3.8M): Generate headlines for news articles

- Evaluation Metrics
  - BLEU: geometric mean of *N*-gram precision (*N*=1, 2, 3, 4)
  - ROUGE: F-1 scores of *N*-gram matching (with longest common subsequence)

Rush et al., “A Neural Attention Model for Abstractive Sentence Summarization”, EMNLP 2015
Papineni et al., “BLEU: A Method for Automatic Evaluation of Machine Translation”, ACL 2002
Lin, “ROUGE: A Package for Automatic Evaluation of Summaries”, 2004

Results on MT Task

- Our approach is *model-agnostic*
  - Generalizable to different Seq2Seq models (e.g., RNN, Transformer)

- Our approach can *scale* to larger datasets

En-Vi Models	tst2012	tst2013
Our Implementations		
RNN	23.37	26.80
+ BERT teacher	25.14	27.59
Transformer (base)	27.03	30.76
+ BERT teacher	**27.85**	**31.51**

En-De Models	NT2013	NT2014
Our Implementations		
Transformer (base)	25.95	26.94
+ BERT teacher	**26.22**	**27.53**

Vaswani et al., “Attention Is All You Need”, NeurIPS 2017

Results on Summarization Task

- Our approach is *generalizable* to different text generation tasks

- Our generic approach is *comparable* to task-specific SOTA customized for summarization

GW Models	R-1	R-2	R-L
Dev			
Transformer (base)	46.64	24.37	43.17
+ BERT teacher	**47.35**	**25.11**	**44.04**
Test-Dev			
Transformer (base)	46.84	24.80	43.58
+ BERT teacher	**47.90**	**25.75**	**44.53**

GW Models	R-1	R-2	R-L
Seq2Seq†	36.40	17.77	33.71
CGU‡	36.3	18.0	33.8
$\text{FTSum}_g{}^{\star}$	37.27	17.65	34.24
$\text{E2T}_{cnn}{}^{\diamond}$	37.04	16.66	**34.93**
$\text{Re}^3\text{Sum}^{\bullet}$	37.04	**19.03**	34.46
Trm + BERT teacher	**37.57**	18.59	34.82

State-of-the-art on Two MT Tasks

En-Vi Models	tst2012	tst2013
Our Implementations		
RNN	23.37	26.80
+ BERT teacher	25.14	27.59
Transformer (base)	27.03	30.76
+ BERT teacher	**27.85**	**31.51**
Other Reported Results		
RNN†	-	26.1
Seq2Seq-OT*	24.5	26.9
ELMo$^{\diamond}$	-	29.3
CVT$^{\diamond}$	-	29.6

De-En Models	dev	test
Our Implementations		
Transformer (base)	35.27	34.09
+ BERT teacher	**36.93**	**35.63**
Other Reported Results		
ConvS2S + MRT‡	33.91	32.85
Transformer (big)$^{\diamond}$	-	34.4†
Lightweight Conv$^{\diamond}$	-	34.8†
Dyn. Convolution$^{\diamond}$	-	35.2†

Clark et al., “Semi-Supervised Sequence Modeling with Cross-View Training”, EMNLP 2018
Wu et al., “Pay Less Attention with Lightweight and Dynamic Convolutions”, ICLR 2019

*SOTA at the time of submission

Ablation Study

- Bidirectional nature?
  - $BERT_{l2r}$: train teacher with left-to-right LM objective
- Extra parameter?
  - $BERT_{sm}$: train smaller teacher (6-layer BERT)

Methods	De-En (dev)	En-Vi (tst2012)
Transformer (base)	35.27	27.03
Trm + BERT_{l2r}	35.20	26.99
Trm + BERT_{sm}	36.32	27.68
Trm + BERT	**36.93**	**27.85**

- BERT pre-training is still essential

Why Distillation?

- C-MLM takes both **left and right context**
  - Look-ahead generation / plan for future (implicit)

Why Distillation?

- C-MLM takes both left and right context
- Soft distribution has more information
    - *“Dark knowledge”* contains useful learning signal

Why Distillation?

- C-MLM takes both left and right context
- Soft distribution has more information
- No-explicit parameter sharing / feature extraction
    - Model agnostic
    - Same inference speed / model size

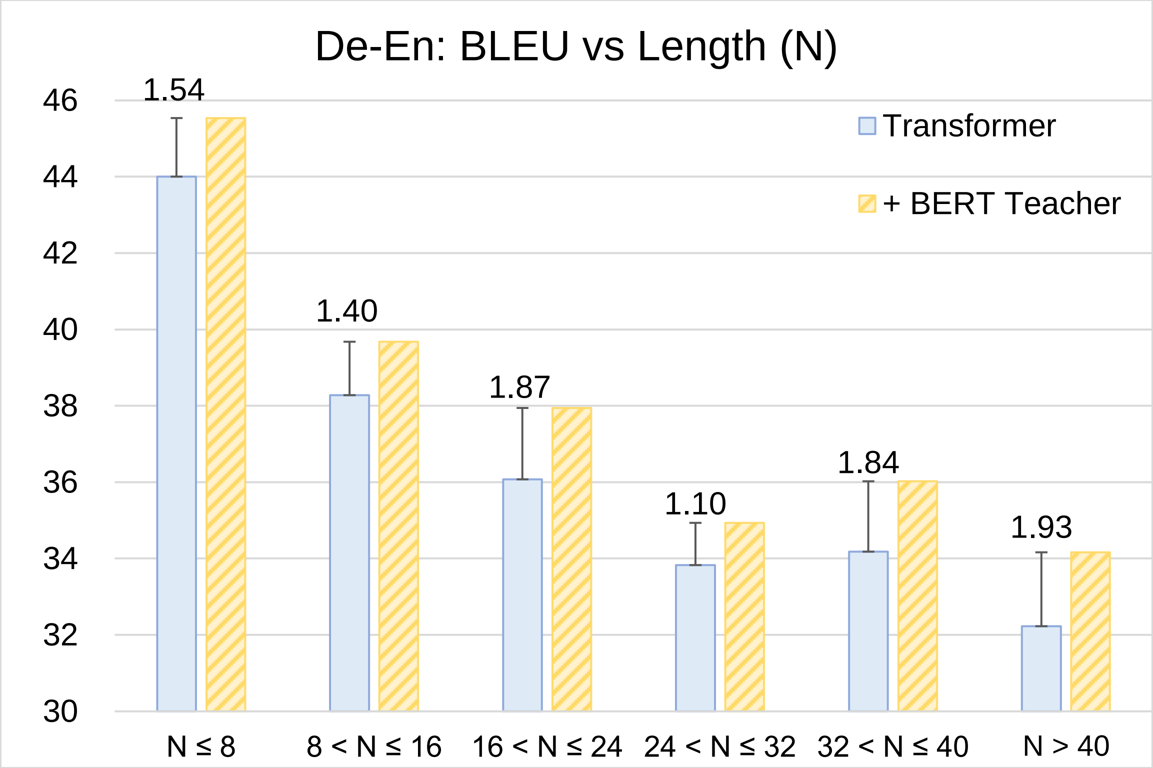
Why Distillation?

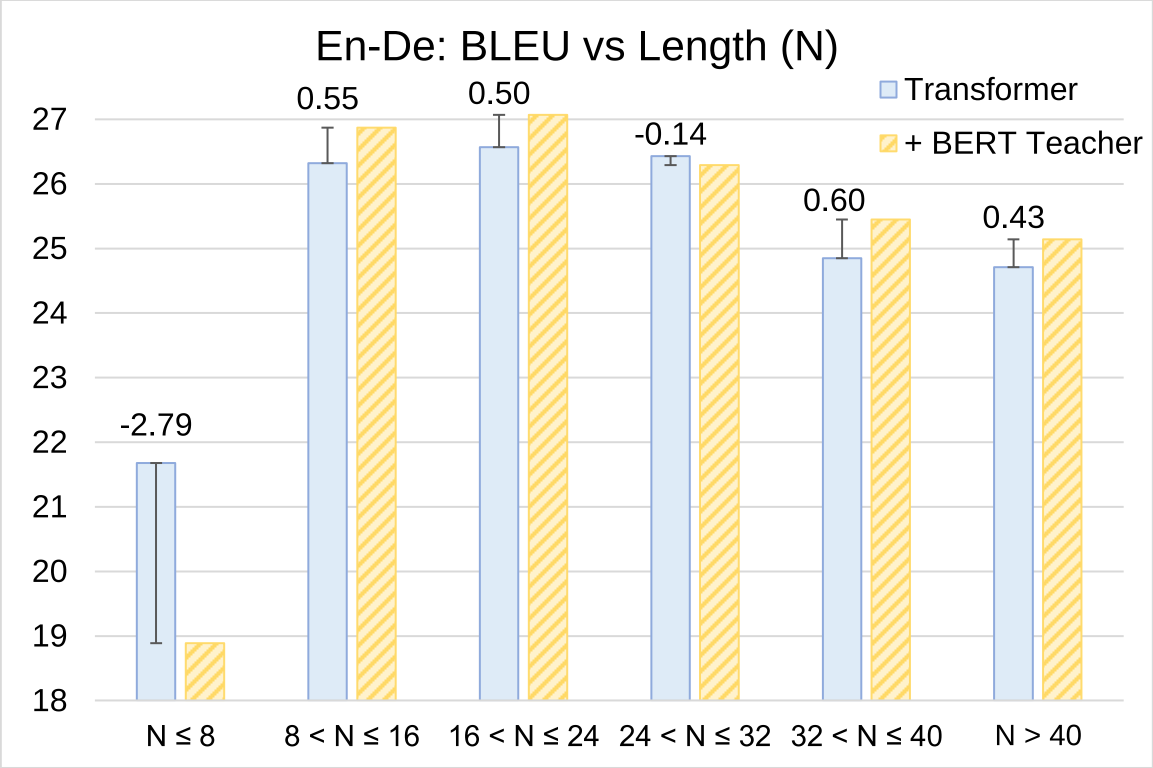
- C-MLM takes both left and right context
- Soft distribution has more information
- No-explicit parameter sharing / feature extraction

- NOT model compression
  - C-MLM / BERT cannot generate, but can provide better training target

Analysis

- Our approach achieves higher performance gain on longer sequences





Examples (MT)

Reference	my mother says that i started reading at the age of two , although i think four is probably close to the truth .
Transformer	my mother says that i started reading with two years , but i think that four of them probably correspond to the truth . (39.6)
Ours	my mother says that i started reading at the age of two , but i think four is more likely to be the truth . (65.2)
Reference	we already have the data showing that it reduces the duration of your flu by a few hours .
Transformer	we 've already got the data showing that it 's going to crash the duration of your flu by a few hours . (56.6)
Ours	we already have the data showing that it reduces the duration of your flu by a few hours . (100.0)
Reference	we now know that at gombe alone , there are nine different ways in which chimpanzees use different objects for different purposes .
Transformer	we know today that alone in gombe , there are nine different ways that chimpanzees use different objects in different ways . (35.8)
Ours	we now know that in gombe alone , there are nine different ways that chimpanzees use different objects for different purposes . (71.5)

Summary

- Use Knowledge Distillation to pass on bidirectional contextual information from pre-trained C-MLM (Teacher) to Seq2Seq model (Student) for text generation
  - **Bidirectional**: C-MLM takes both left and right context
  - **Model-agnostic**: No-explicit sharing / feature extraction
  - **Generalizable** to different text generation tasks

- State-of-the-art on two machine translation tasks

Code: https://github.com/ChenRocks/Distill-BERT-Textgen

Thank You!