Pretrained Models for Text Generation

How to ultilize BERT for language generation

- **Distilling Knowledge Learned in BERT for Text Generation ACL(2020)**
- **BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension (ACL2020)**

Distilling Knowledge Learned in BERT for Text Generation

- BERT is learned with a generative objective via Masked Language Modeling (MLM).
- This training objective should have learned essential, bidirectional, contextual knowledge that can help enhance text generation.
- However, **MLM objective is not auto-regressive.**

Finetune BERT with Conditional MLM

The input is a sequence pair (X,Y), 15% of the tokens are randomly masked. The trained BERT model aims to estimate the joint probability:

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

Conditional-MLM allows further finetuning of pre-trained BERT on target dataset. We randomly masks 15% of the tokens only in Y , then train the network to model the joint probability:

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

Knowledge Distillation for Generation

- The distribution for a given word $P(y_t^m | X, Y^u)$ contains information from both **backward and forward** contexts, which is a desirable benefit for providing **sequence-level global guidance**.

Knowledge Distillation for Generation

- BERT as Teacher & Seq2Seq as Student:

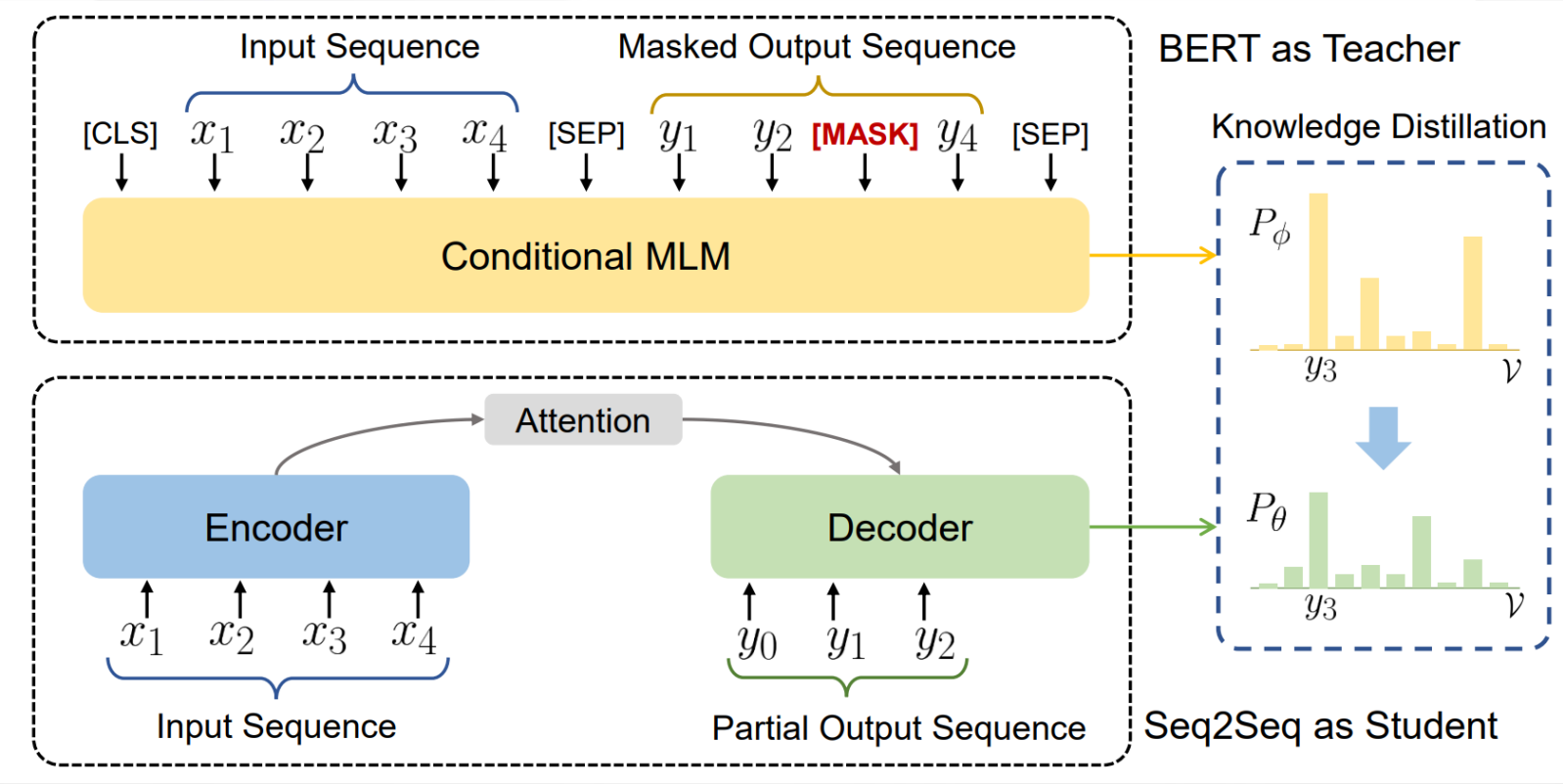
$$L_{bidi}(\theta) = -\sum_{w \in \nu}[P_\phi(y_t = w|Y^u, X) \cdot \log P_\theta(y_t = w|y_{1:t-1}, X)]$$

$P_\phi(y_t)$ is the soft target estimated by the finetuned BERT with learned parameters ϕ (fixed).

- To further improve the Seq2Seq student model, hard-assigned labels are also utilized. The final objective:

$$L(\theta) = \alpha L_{bidi}(\theta) + (1 - \alpha)L_{xe}(\theta)$$

Knowledge Distillation for Generation



Experiments

- Machine Translation:
  - IWSLT15 English-Vietnamese
  - IWSLT14 German-English
  - WMT14 English-German
- Abstractive Summarization
  - Gigaword summarization dataset

Results on Machine Translation

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)◇	-	34.4†
Lightweight Conv◇	-	34.8†
Dyn. Convolution◇	-	35.2†

Table 1: BLEU scores for IWSLT14 German-English translation. (†) tuned with checkpoint averaging. (‡) from Edunov et al. (2018). (◇) from Wu et al. (2019).

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◇	-	29.3
CVT◇	-	29.6

Table 2: BLEU scores for IWSLT15 English-Vietnamese translation. (†) from Luong et al. (2017). (⋆) from Chen et al. (2019). (◇) from Clark et al. (2018).

En-De Models	NT2013	NT2014
Our Implementations		
Transformer (base)	25.95	26.94
+ BERT teacher	**26.22**	**27.53**
Other Reported Results		
Transformer (base)◇	25.8	27.3†
Transformer (big)⋆‡	26.5	29.3†
Dyn. Convolution•‡	**26.9**±0.2	**29.7**†

Table 3: BLEU scores for WMT14 English-German translation. (†) tuned with checkpoint averaging. (‡) trained on WMT16, a slightly different version of training data. (◇) from Vaswani et al. (2017). (⋆) from Ott et al. (2018). (•) from Wu et al. (2019).

Results on Abstractive 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**

Table 4: ROUGE F_1 scores for Gigaword abstractive summarization on our internal test-dev split.

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

Table 5: ROUGE F_1 scores for Gigaword abstractive summarization on the official test set (Trm: Transformer). (†) from Nallapati et al. (2016). (‡) from Lin et al. (2018). (⋆) from Cao et al. (2018b). (◇) from Amplayo et al. (2018). (•) from Cao et al. (2018a).

Ablation Study

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**

Table 6: Ablation study. (Trm: Transformer)

- BERT_sm: use a smaller BERT (6 layers) for C-MLM finetuning
- BERT_l2r: use the full BERT model but finetune it using left-to-right LM

Qualitative Examples

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)

Table 7: Qualitative examples from IWSLT German-English translation. Numbers inside the parenthesis are sentence-level BLEU scores. Red word is where the baseline Transformer makes a mistake without considering the possible future phrase and fails to recover. On the other hand, our model makes the right decision at the blue word, hence generates more coherent sentence. Please refer to Section 4.7 for detailed explanation.

Conclusion

- utilize a pretrained model to improve text generation without explicit parameter sharing, feature extraction, or augmenting with auxiliary tasks.
- distillation approach indirectly influences the text generation model by providing soft-label distributions only, hence is model-agnostic.
- How to extend the Conditional MLM to multimodal input such as image captioning?

BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension

- BART is a denoising autoencoder built with a sequence-to-sequence model which combines the power of BERT (due to the bidirectional encoder), GPT (with the left-to-right decoder).
- BART is particularly effective when fine tuned for text generation but also works well for comprehension tasks.

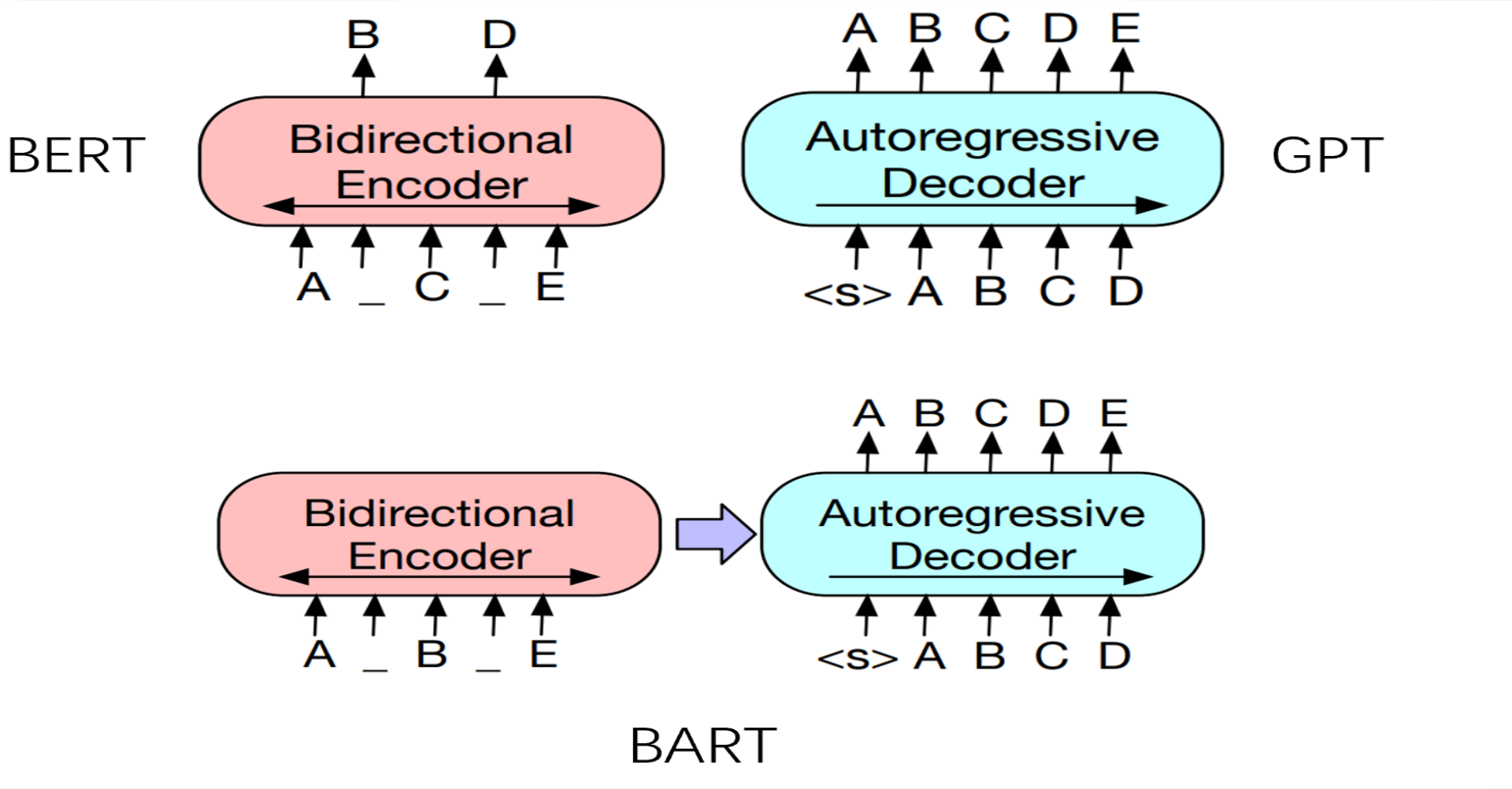
How does it work

> decoupling language models and the functions with which the text are corrupted is useful to compare different pre-training techniques

pre-training is a sequence of repeated steps:

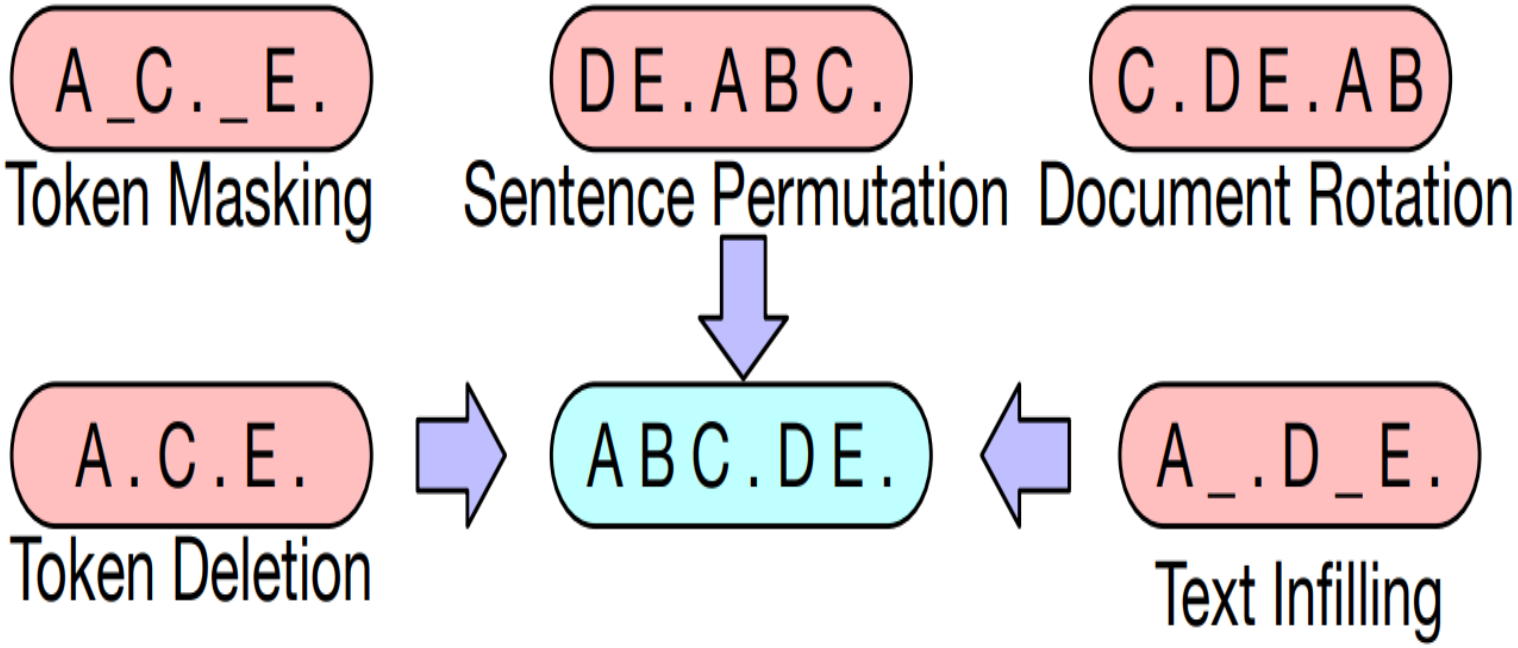
- Apply a noising function to the text
- The language model attempts to reconstruct the text
- Then calculate the loss function (typically cross entropy over the original text) and then back-propagate the gradients and update the model's weights.

BERT, GPT and BART



Pre-training BART

- BART is trained by corrupting documents and then optimizing a reconstruction loss.
- several previously proposed and novel transformations:



Pre-training BART

- **Token Masking**: random tokens are sampled and replaced with [MASK]
- **Token Deletion**: similar to masking but the sampled tokens are deleted and the model has to add a new token in their place.
- **Token Infilling**: a number of text spans, i.e. contiguous group tokens, are sampled, and then they are replaced by the [MASK] token.
- **Sentence Permutation**: random shuffling of the document's sentences.

Pre-training BART

- **Document Rotation**: a token is chosen randomly to be the start of the document, the section before the starting token is appended at the end.

Fine-tuning BART

- Sequence Classification Tasks
  - Final hidden state of the final decoder token is fed into new multi-class linear classifier
- Token Classification Tasks
  - Top hidden state of the decoder is used as a representation for each word.
- Sequence Generation Tasks
  - Because BART has an autoregressive decoder, it can be directly fine tuned for sequence generation tasks.

Fine-tuning BART

- Machine Translation
  - Add a new set of encoder parameters that are learned from bitext

Comparing Pre-training Objectives

- **Autoregressive, left to right, LM** (GPT-2)
- **Masked LM** (BERT) replace 15% of the token with the [MASK] token and predict the corresponding words.
- **Permuted LM** (XLNet) left to right, autoregressive LM training but with the order of the words to predict chosen at random.
- **Multitask Masked LM** (UniLM) combination of right-to-left, left-to-right and bidirectionality. ⅓ of the time using each with shared parameters.
- **Masked Seq2Seq** (MASS) masking a span containing 50% of the tokens and train to predict the masked tokens.

Results

Model	SQuAD 1.1 F1	MNLI Acc	ELI5 PPL	XSum PPL	ConvAI2 PPL	CNN/DM PPL
BERT Base (Devlin et al., 2019)	88.5	**84.3**	-	-	-	-
Masked Language Model	90.0	83.5	24.77	7.87	12.59	7.06
Masked Seq2seq	87.0	82.1	23.40	6.80	11.43	6.19
Language Model	76.7	80.1	**21.40**	7.00	11.51	6.56
Permuted Language Model	89.1	83.7	24.03	7.69	12.23	6.96
Multitask Masked Language Model	89.2	82.4	23.73	7.50	12.39	6.74
BART Base						
w/ Token Masking	90.4	84.1	25.05	7.08	11.73	6.10
w/ Token Deletion	90.4	84.1	24.61	6.90	11.46	5.87
w/ Text Infilling	**90.8**	84.0	24.26	**6.61**	**11.05**	5.83
w/ Document Rotation	77.2	75.3	53.69	17.14	19.87	10.59
w/ Sentence Shuffling	85.4	81.5	41.87	10.93	16.67	7.89
w/ Text Infilling + Sentence Shuffling	**90.8**	83.8	24.17	6.62	11.12	**5.41**

Results

Several trends are clear:

- Token masking is crucial
- Left-to-right pre-training improves generation
- Bidirectional encoders are crucial for QA

Large-scale Pre-training Experiments

- downstream performance can dramatically improve when pre-training is scaled to large batch sizes and corpora.
- pretrain a large model with 12 layers in each of the encoder and decoder, and a hidden size of 1024.
- use a batch size of 8000, and train the model for 500000 steps
- use the same pre-training data as BERT consisting of 160Gb of news, books, stories, and web text.

Discriminative Tasks

	SQuAD 1.1 EM/F1	**SQuAD 2.0** EM/F1	**MNLI** m/mm	**SST** Acc	**QQP** Acc	**QNLI** Acc	**STS-B** Acc	**RTE** Acc	**MRPC** Acc	**CoLA** Mcc
BERT	84.1/90.9	79.0/81.8	86.6/-	93.2	91.3	92.3	90.0	70.4	88.0	60.6
UniLM	-/-	80.5/83.4	87.0/85.9	94.5	-	92.7	-	70.9	-	61.1
XLNet	**89.0**/94.5	86.1/88.8	89.8/-	95.6	91.8	93.9	91.8	83.8	89.2	63.6
RoBERTa	88.9/**94.6**	**86.5/89.4**	**90.2/90.2**	96.4	92.2	94.7	**92.4**	86.6	**90.9**	**68.0**
BART	88.8/**94.6**	86.1/89.2	89.9/90.1	**96.6**	**92.5**	**94.9**	91.2	**87.0**	90.4	62.8

Generation Tasks

- Summarization

	CNN/DailyMail			XSum		
	R1	R2	RL	R1	R2	RL
Lead-3	40.42	17.62	36.67	16.30	1.60	11.95
PTGEN (See et al., 2017)	36.44	15.66	33.42	29.70	9.21	23.24
PTGEN+COV (See et al., 2017)	39.53	17.28	36.38	28.10	8.02	21.72
UniLM	43.33	20.21	40.51	-	-	-
BERTSUMABS (Liu & Lapata, 2019)	41.72	19.39	38.76	38.76	16.33	31.15
BERTSUMEXTABS (Liu & Lapata, 2019)	42.13	19.60	39.18	38.81	16.50	31.27
BART	**44.16**	**21.28**	**40.90**	**45.14**	**22.27**	**37.25**

Table 3: Results on two standard summarization datasets. BART outperforms previous work on summarization on two tasks and all metrics, with gains of roughly 6 points on the more abstractive dataset.

- Dialogue

	ConvAI2	
	Valid F1	Valid PPL
Seq2Seq + Attention	16.02	35.07
Best System	19.09	17.51
BART	**20.72**	**11.85**

Table 4: BART outperforms previous work on conversational response generation. Perplexities are renormalized based on official tokenizer for ConvAI2.

Generation Tasks

- Abstractive QA

	ELI5		
	R1	R2	RL
Best Extractive	23.5	3.1	17.5
Language Model	27.8	4.7	23.1
Seq2Seq	28.3	5.1	22.8
Seq2Seq Multitask	28.9	5.4	23.1
BART	**30.6**	**6.2**	**24.3**

Table 5: BART achieves state-of-the-art results on the challenging ELI5 abstractive question answering dataset. Comparison models are from Fan et al. (2019).

Translation

	RO-EN
Baseline	36.80
Fixed BART	36.29
Tuned BART	**37.96**

Table 6: The performance (BLEU) of baseline and BART on WMT'16 RO-EN augmented with back-translation data. BART improves over a strong back-translation (BT) baseline by using monolingual English pre-training.

Qualitative Examples

Source Document (abbreviated)	BART Summary
The researchers examined three types of coral in reefs off the coast of Fiji ... The researchers found when fish were plentiful, they would eat algae and seaweed off the corals, which appeared to leave them more resistant to the bacterium Vibrio coralliilyticus, a bacterium associated with bleaching. The researchers suggested the algae, like warming temperatures, might render the corals' chemical defenses less effective, and the fish were protecting the coral by removing the algae.	Fisheries off the coast of Fiji are protecting coral reefs from the effects of global warming, according to a study in the journal Science.
Sacoolas, who has immunity as a diplomat's wife, was involved in a traffic collision ... Prime Minister Johnson was questioned about the case while speaking to the press at a hospital in Watford. He said, "I hope that Anne Sacoolas will come back ... if we can't resolve it then of course I will be raising it myself personally with the White House."	Boris Johnson has said he will raise the issue of US diplomat Anne Sacoolas' diplomatic immunity with the White House.
According to Syrian state media, government forces began deploying into previously SDF controlled territory yesterday. ... On October 6, US President Donald Trump and Turkish President Recep Tayyip Erdoan spoke on the phone. Then both nations issued statements speaking of an imminent incursion into northeast Syria ... . On Wednesday, Turkey began a military offensive with airstrikes followed by a ground invasion.	Syrian government forces have entered territory held by the US-backed Syrian Democratic Forces (SDF) in response to Turkey's incursion into the region.
This is the first time anyone has been recorded to run a full marathon of 42.195 kilometers (approximately 26 miles) under this pursued landmark time. It was not, however, an officially sanctioned world record, as it was not an "open race" of the IAAF. His time was 1 hour 59 minutes 40.2 seconds. Kipchoge ran in Vienna, Austria. It was an event specifically designed to help Kipchoge break the two hour barrier.	Kenyan runner Eliud Kipchoge has run a marathon in less than two hours.
PG&E stated it scheduled the blackouts in response to forecasts for high winds amid dry conditions. The aim is to reduce the risk of wildfires. Nearly 800 thousand customers were scheduled to be affected by the shutoffs which were expected to last through at least midday tomorrow.	Power has been turned off to millions of customers in California as part of a power shutoff plan.

Conclusion

- Pre-training techniques can be viewed as corrupting text with an arbitrary noising function while the Language Model is tasked with denoising it.
- Performance of pre-training methods varies significantly across tasks.
- Explore new methods for corrupting documents for pre-training, perhaps tailoring them to specific end tasks.