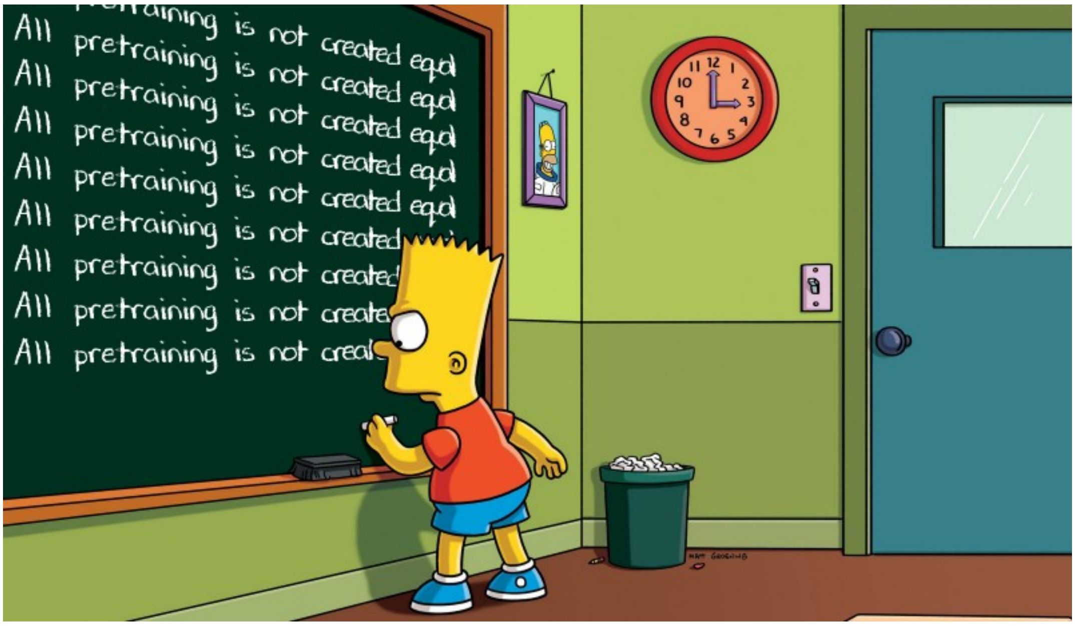
Upgrade Open in app

Published in DAIR.AI · Follow

Antonio Lopardo · Follow

May 15, 2020 · 5 min read · Listen

BART: Are all pretraining techniques created equal?


raining is not created equal
All pretraining is not created equal
All pretraining is not created equal
All pretraining is not created equal
All pretraining is not created equal
All pretraining is not created
All pretraining is not created
All pretraining is not create
All pretraining is not creat

Created with http://www.ranzey.com/generators/bart/index.html

Why is this important?

In this paper, Lewis et al. present valuable comparative work on different pre-training techniques and show how this kind of work can be used to guide large pre-training experiments reaching state-of-the-art (SOTA) results.

What does it propose?

The authors propose a framework to compare pre-training techniques and language model (LM) objectives. This framework focuses on how these techniques can be viewed as ***corrupting text with an arbitrary noising function while the language model is tasked with denoising it.*** After some comparative experiments using this framework, BART is introduced as a transformer-based LM that reaches SOTA performance.

How does it work?

The Framework

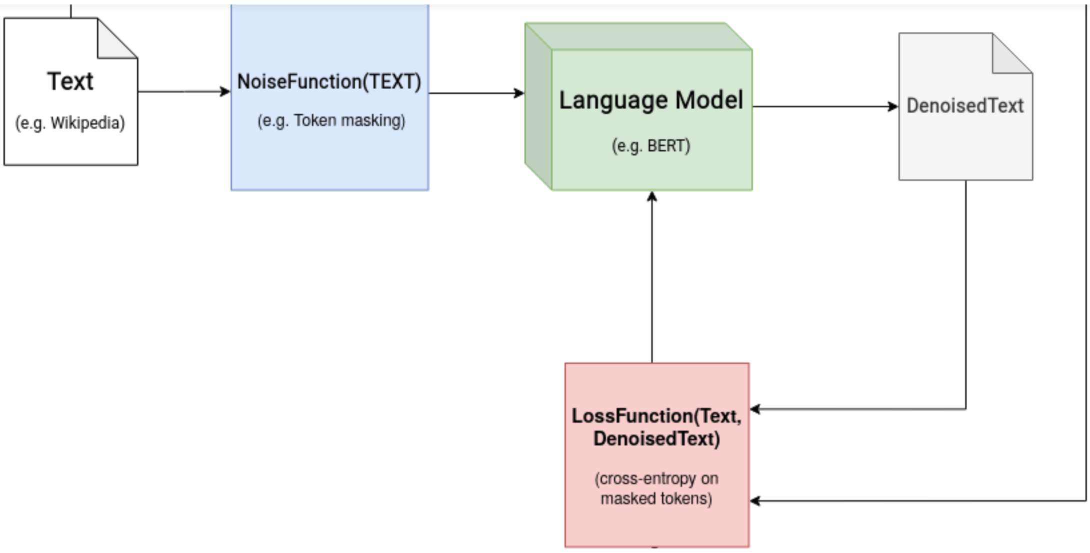


Figure 1: Diagram of the framework introduced in the paper

The idea behind the proposed framework is simple, they suggest that decoupling language models and the functions with which the texts are corrupted are useful to compare different pre-training techniques and see how they perform on similar models and diverse benchmarks. Viewed this way, 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.

Comparing different text-noising techniques and LM Objectives

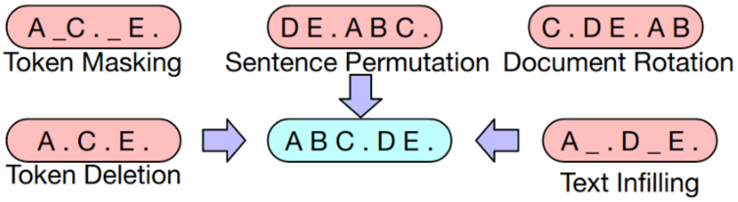


Figure 2: How different text-noising techniques corrupt the text

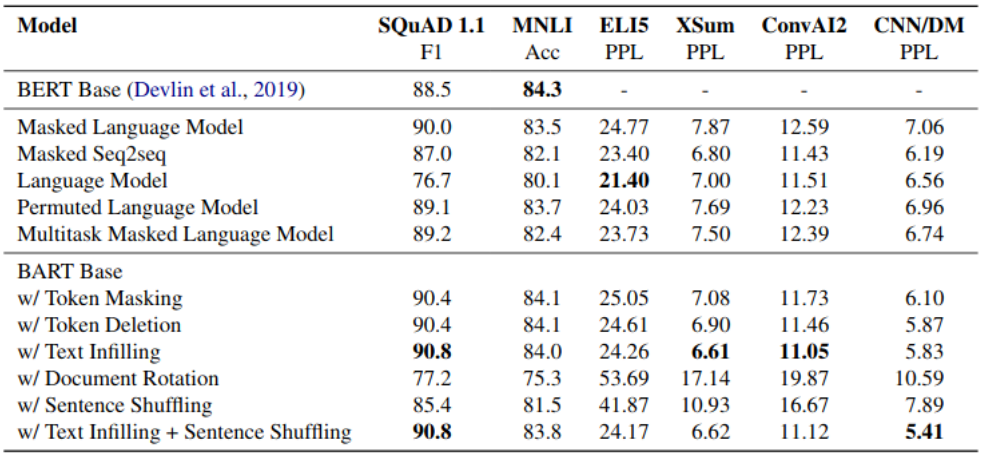
In the first experiment, using the framework introduced at the beginning of the article, the authors compared different pre-training techniques and LM objectives on a smaller than usual model, BART-base. The model uses a 6 layered, transformer-based, seq2seq architecture for autoencoding as introduced by Vaswani et al. The pre-training techniques compared in the experiments can be divided between those that work at the token level and those that work at the sentence level:

- **Token Masking** random tokens are sampled and replaced with [MASK]
- **Token Deletion** similar to masking but the sampled token is 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.
- **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.

Besides the pre-training techniques, the authors also compare different LM objectives focusing on the ones used by BERT and GPT as well as techniques that tried to incorporate the best of both worlds:

- **Autoregressive, left to right, LM** (GPT-2)
- **Masked LM**(BERT) replace 15% of the token with [MASK] 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, using bi-direction. 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 of the first experiment



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**

Figure 3: Table of results for the first experiment

From the results of these first experiments, the authors draw some important conclusions.

Token masking is crucial

Only the configurations with token masking or its variations achieve consistently great performance on different tasks.

Left-to-right pre-training improves NLG

The Classical Language Model objective despite not doing well in inference or question answering tasks achieves SOTA on ELI5(Explain Like I'm 5).

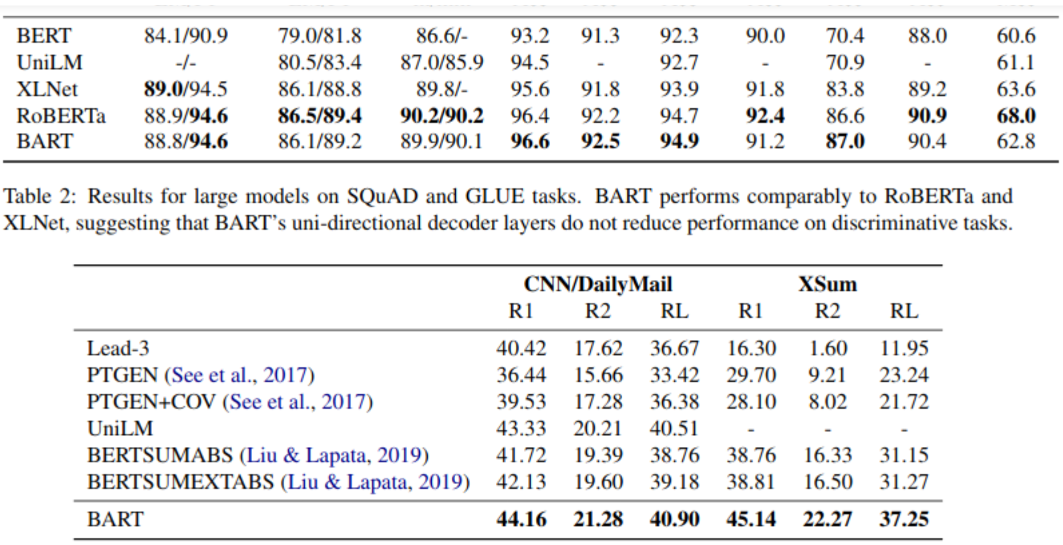
Bidirectional encoders are crucial for QA

Ignoring future context hinders the performance of left-to-right models.

While pre-training techniques and LM objectives are important, the authors make note of the fact that they do not provide the full picture. They report that their permuted language model performs much worse than XLNet because BART lacks some of the valuable architectural innovations introduced in XLNet.

Results of the large-scale pre-training experiment

After the comparative experiment, the authors trained a 12 layered, transformer-based architecture for autoencoding, and using similar hyperparameters to RoBERTa. They used both a form of token masking at 30% and sentence permutation as pre-training text-noising techniques and run the model on 160GB of news, books, stories, and web text, similar to what's done in RoBERTa.



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

Table 2: Results for large models on SQuAD and GLUE tasks. BART performs comparably to RoBERTa and XLNet, suggesting that BART's uni-directional decoder layers do not reduce performance on discriminative tasks.

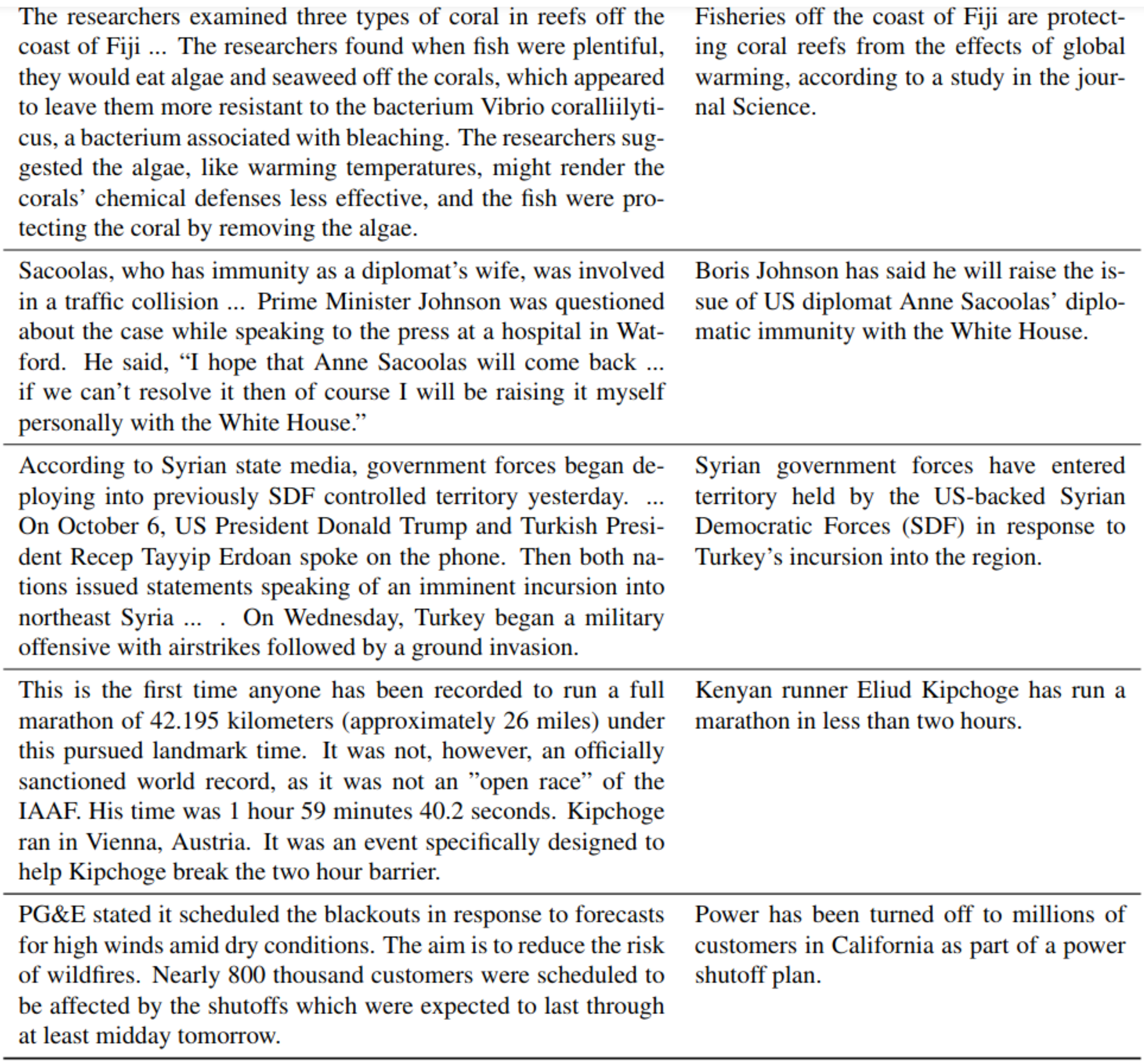
	CNN/DailyMail			XSum		
	R1	R2	RL	R1	R2	RL
Lead-3	40.42	17.62	36.67	16.30	1.60	11.95
PTGEN (See et al., 2017)	36.44	15.66	33.42	29.70	9.21	23.24
PTGEN+COV (See et al., 2017)	39.53	17.28	36.38	28.10	8.02	21.72
UniLM	43.33	20.21	40.51	-	-	-
BERTSUMABS (Liu & Lapata, 2019)	41.72	19.39	38.76	38.76	16.33	31.15
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**

Figure 4: Table of results for the Large Scale pre-training

BART performs best in abstractive summarization tasks especially in the **XSum** benchmark that contains very few examples of summaries where phrases are present both in the summary and the original text. Besides surpassing the previous best systems in summarization by a considerable margin, BART does well also in natural language inference (NLI) tasks and QA, where it is on par with SOTA results.

Qualitative Analysis

The paper also features examples of summaries of WikiNews articles produced by a version of BART fine-tuned on the XSum dataset:



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.

Figure 5 Table of summaries produced by BART from WikiNews articles

From these examples, BART appears capable of producing coherent grammatical sentences that capture the sense of the text it should summarize. It highlights names and places why ignoring other details like dates and figures.

If you want to summarize some text of your own we have set up a Google Colab notebook using the Hugging Face library.