T0: Multitask Prompt Training

Sasha Rush /w

BigScience
A one-year long research workshop on large multilingual models and datasets

https://bigscience.huggingface.co/
Language Models are Few-Shot Learners

TriviaQA zero-shot performance

Accuracy vs. Parameters (Billions)
PaLM: Scaling Language Modeling with Pathways

TriviaQA zero-shot performance

Chowdhery et al. 2022
Q: ‘Nude Descending A Staircase’ is perhaps the most famous painting by which 20th century artist?

A: ....
Prompt Template

Q: {Question}
A: {Answer}
Q: Which President of the Philippines was deposed in 1986?
A: Marcos

Q: Who was president of the USA at the outbreak of World War I?
A: Wilson

Q: ‘Nude Descending A Staircase’ is perhaps the most famous painting by which 20th century artist?
A: ...
"Multitask prompted training enables zero-shot task generalization"

Punchline ->
Training on many NLP tasks improves generalization to new unseen tasks.

Artifact ->
T0 - A smaller model with strong zero-shot prompting abilities

Sanh et al. (2022)
Outline

● Preliminary Work
  ○ Datasets
    ○ How many data points is a prompt worth

● T0

● Context: BigScience
Preliminary Work: Datasets

(Lhoest et al, 2021)
Datasets: Tour of the library

```python
from datasets import load_dataset

dataset = load_dataset("boolq")

# Each dataset has a features schema and metadata.
print(dataset.features, dataset.info)

# Any slice of data points can be accessed directly without loading the full dataset into memory.
dataset["train"][start:end]

# Processing can be applied to every data point in a batched and parallel fashion using standard libraries such as NumPy or Torch.
tokenized = dataset.map(tokenize, num_proc=32)
```
Datasets: Internals

Apache Arrow:

- language-independent columnar memory format
- memory-mapping to load terabytes of data without using RAM
- zero-copy reads for fast data access without serialization overhead
  - <1ms latency even on billion-scale datasets
  - end-to-end zero-copy to deep-learning frameworks

*jax not fully end-to-end
Dataset cards

- document the datasets
- community-driven
- dynamic
- search by task/lang/etc.

- standardized types
- get feature names
- types across dataset

Dataset Summary

The EL15 dataset is an English-language dataset of questions and answers gathered from three subreddits were users ask factual questions requiring paragraph-length or longer answers. The dataset was created to support the task of open-domain long form abstractive question answering, and covers questions about general topics in its r/explainlikeimfive subset, science in its r/askscience subset, and History in its r/AskHistorians subset.

Supported Tasks and Leaderboards

- abstractive-qa, open-domain-qa: The dataset can be used to train a model for Open Domain Long Form Question Answering. An LFQA model is presented with a non-factoid and asked to retrieve relevant information from a knowledge source (such as Wikipedia), then use it to generate a multi-sentence answer. The model performance is measured by how high its ROUGE score to the reference is. A BART-based model with a dense retriever trained to draw information from Wikipedia passages achieves a ROUGE-1 of 0.149.
Datasets: Meta-Datasets

- Benchmarks: LM Evaluation Harness
- Workshops / Shared tasks: GEM
- Robustness evaluation: Robustness Gym
Preliminary Work:
How many data points is a prompt worth?

(Le Scao et al, 2021)
Finetuning with Prompting

1. Start from pre-trained language model

2. Modify labeled training data to prompted form

Goals

- Sanity check the use of prompts in training.
- Does training with prompts improve over standard labels?
- How can we measure that difference?
Experimental setup

RoBERTa-Large
Testing on SuperGLUE + MNLI
Best of 4 runs on every data size

- Linear classification head
- Fine-tuned via backpropagation on the predicted class
- Task-adaptation with a prompt (3-4 different prompts per task)
- Fine-tuned via backpropagation on the predicted output token
Choice of prompts

Prompts from *It’s Not Just Size That Matters* (Schick and Schütze 2020) For BoolQ, for example:

- `{passage}. Question: {question}? Answer: ....
- `{passage}. Based on the previous passage, {question}?....
Data Advantage

Performance vs. dataset size on BoolQ for the classifier model.
Data Advantage

Performance vs. dataset size on BoolQ for the classifier and prompting models.
Data Advantage

The **prompted** model reaches 0.75 accuracy with **1132 data points** less than the **classifier**.
Data Advantage

Over the whole region, the prompted model is 752 data points ahead of the classifier on average.
Data advantage (all tasks)

### BoolQ
752±46

### CB
90±2

### COPA
288±242

### MNLI
3506±536
(x log scale)

### MultiRC
384±378

### RTE
282±34

### WSC
281±137

### WiC
-424±74
What we know

● Does the model understand the prompt?
  ○ Probably not. (Webson & Pavlick, 2022)

● Does the prompt need to be human understandable?
  ○ Not clear, particularly in few-shot versions.

● What can we say?
  ○ Language is a convenient modality for task encoding.
T0

BigScience

(Sanh et al, 2022)
Research Question

- Can we induce zero-shot task transfer through pretraining on prompts?
- Practical benefit → Smaller models with zero-shot ability
- Research → Generic pretraining versus targeted induction.
Review: T5

Text-to-Text Transfer Transformer
The cabs __ the same rates as those ___ by horse-drawn cabs and were ___ quite popular, ___ the Prince of Wales (the ___ King Edward VII) travelled in ___. The cabs quickly ___ known as "hummingbirds" for ___ noise made by their motors and their distinctive black and ___ livery. Passengers ___ ___ the interior fittings were ___ when compared to ___ cabs but there ___ some complaints ___ the ___ lighting made them too ___ to those outside ___.

charged, used, initially, even, future, became, the, yellow, reported, that, luxurious, horse-drawn, were that, internal, conspicuous, cab
T0 Recipe

- Produce templates for turning a large set of datasets to prompts.
- Pretrain T5 LM on those prompts for a significant amount of time.
- Evaluate model on tasks it has not seen before.
<table>
<thead>
<tr>
<th>Generalization</th>
<th>Task</th>
<th>Dataset</th>
<th>Prompt</th>
<th>Instances</th>
</tr>
</thead>
<tbody>
<tr>
<td>(i.i.d.) new examples</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>new instructions</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>new domain</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>new “skill”</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Increasing generalization
Summarization

The picture appeared on the wall of a Poundland store on Whymark Avenue [...] How would you rephrase that in a few words?

Paraphrase identification

“How is air traffic controlled?” “How do you become an air traffic controller?” Pick one: these questions are duplicates or not duplicates.

Question answering

I know that the answer to “What team did the Panthers defeat?” is in “The Panthers finished the regular season [...]”. Can you tell me what it is?

Multi-task training

Zero-shot generalization

Natural language inference

Suppose “The banker contacted the professors and the athlete”. Can we infer that "The banker contacted the professors"?
PromptSource: Prompts for Training
Closed-book question answering
http://www.autosweblog.com/cat/trivia-questions-from-the-50s
  who was frank sinatra? a: an american singer, actor, and producer.

Paraphrase identification
https://www.usingenglish.com/forum/threads/60200-Do-these-sentences-mean-the-same
  Do these sentences mean the same? No other boy in this class is as smart as the boy. No other boy is as smart as the boy in this class.

Natural Language Inference
https://ell.stackexchange.com/questions/121446/what-does-this-sentence-imply
  If I say: He has worked there for 3 years. does this imply that he is still working at the moment of speaking?

Summarization
https://blog.nytsoi.net/tag/reddit
  ... Lately I've been seeing a pattern regarding videos stolen from other YouTube channels, reuploaded and monetized with ads. These videos are then mass posted on Reddit by bots masquerading as real users. tl;dr: Spambots are posting links to stolen videos on Reddit, copying comments from others to masquerade as legitimate users.

Pronoun resolution
  Jennifer is a vegetarian, so she will order a nonmeat entrée. In this example, the pronoun she is used to refer to Jennifer.
QQP (Paraphrase)

<table>
<thead>
<tr>
<th>Question1</th>
<th>How is air traffic controlled?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Question2</td>
<td>How do you become an air traffic controller?</td>
</tr>
<tr>
<td>Label</td>
<td>0</td>
</tr>
</tbody>
</table>

{Question1} {Question2}
Pick one: These questions are duplicates or not duplicates.

I received the questions "{Question1}" and "{Question2}". Are they duplicates?

{Choices[label]}
The picture appeared on the wall of a Poundland store on Whymark Avenue...

Graffiti artist Banksy is believed to be behind...

How would you rephrase that in a few words?

First, please read the article: How would you rephrase that in a few words?

Now, can you write me an extremely short abstract for it?
Prompt Template Language

**Jinja template**

*Input template*

```python
{{ premise }}
Question: {{ hypothesis }}  True, False, or Neither?
```

*Target template*

```python
{{ answer_choices[label] }}
```
The SNLI corpus (version 1.0) is a collection of 570k human-written English sentence pairs manually labeled for the task of NLI.

```json
{  premise: "A person...",  hypothesis: "A person...",  label: 1 }

{  premise: "The kids...",  hypothesis: "All kids...",  label: 2 }
```
S2 + S3 + S4: Creation

Based on the previous passage, adapted from the BoolQ prompts in Schick & Schütze 2021.

Original Task

Yes ||| No ||| Maybe

Choices in Prompt

Based on the previous passage, is it true that "{{hypothesis}}"?
Yes, no, or maybe? |||

{{ answer_choices[label] }}
S5: Review

The SNLI corpus (version 1.0) is a collection of 570k human-written English sentence pairs manually labeled for the task of NLI...

“A person…” Based on the previous passage, is it true that “A person…”? Yes, no, or maybe? ||| Maybe

“The kids…” Based on the previous passage, is it true that “All kids…”? Yes, no, or maybe? ||| No
Number of prompted datasets: 180
Number of prompts: 2085
My body cast a shadow over the grass.

I am hesitating between two options. Help me choose the more likely cause:

- The sun was rising.
- The grass was cut.

Y

The sun was rising.
<table>
<thead>
<tr>
<th>Task Type</th>
<th>Input</th>
<th>Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>RE</td>
<td>Taken together, these results make it clear that @chemical@-bound forms of ORC and @protein@ are likely to be required for productive interactions and pre-RC formation.</td>
<td>bind</td>
</tr>
<tr>
<td>COREF</td>
<td>We investigated the potential of the @aryl hydrocarbon receptor@ (@AHR@) to suppress NF-kappaB regulated-gene expression, especially acute-phase genes, such as serum amyloid A (Saa).</td>
<td>coref</td>
</tr>
<tr>
<td>EAE</td>
<td>v-erbA @Gene_expression@ is required to @Negative_regulation@ c-erbA function in erythroid cell differentiation and regulation of the erbA target gene CAII.</td>
<td>cause</td>
</tr>
</tbody>
</table>
Comparison: Natural Instructions v2

- PromptSource was post-hoc instruction generation
- PromptSource has less tasks, but multiple instructions per task
- PromptSource tasks are single language.

Wang et al. 2022
T0 - Experiments
<table>
<thead>
<tr>
<th>Multiple-Choice QA</th>
<th>Closed-Book QA</th>
<th>Structure-To-Text</th>
<th>Sentence Completion</th>
<th>BIG-Bench</th>
</tr>
</thead>
<tbody>
<tr>
<td>CommonsenseQA</td>
<td>Hotpot QA</td>
<td>Common Gen</td>
<td>COPA</td>
<td>Code Description</td>
</tr>
<tr>
<td>DREAM</td>
<td>Wiki QA</td>
<td>Wiki Bio</td>
<td>HellaSwag</td>
<td>Conceptual</td>
</tr>
<tr>
<td>QuAIL</td>
<td></td>
<td></td>
<td>Story Cloze</td>
<td>Hindu Knowledge</td>
</tr>
<tr>
<td>QuaRTz</td>
<td></td>
<td></td>
<td></td>
<td>Known Unknowns</td>
</tr>
<tr>
<td>Social IQA</td>
<td></td>
<td></td>
<td></td>
<td>Language ID</td>
</tr>
<tr>
<td>WiQA</td>
<td></td>
<td></td>
<td></td>
<td>Logic Grid</td>
</tr>
<tr>
<td>Cosmos QA</td>
<td></td>
<td></td>
<td></td>
<td>Logical Deduction</td>
</tr>
<tr>
<td>QASC</td>
<td></td>
<td></td>
<td></td>
<td>Misconceptions</td>
</tr>
<tr>
<td>QuaRel</td>
<td></td>
<td></td>
<td></td>
<td>Movie Dialog</td>
</tr>
<tr>
<td>SciQ</td>
<td></td>
<td></td>
<td></td>
<td>Novel Concepts</td>
</tr>
<tr>
<td>Wiki Hop</td>
<td></td>
<td></td>
<td></td>
<td>Strategy QA</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Syllogisms</td>
</tr>
<tr>
<td>Extractive QA</td>
<td></td>
<td></td>
<td></td>
<td>Vitamin C</td>
</tr>
<tr>
<td>Adversarial QA</td>
<td></td>
<td></td>
<td></td>
<td>Winowhy</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Experimental Details

- Based on T5-LM model, 11B parameters
- Comparison to GPT-3 (6.7, 13, 175 B parameters)
  - GPT-3 (6.7B)
  - GPT-3 (13B)
  - GPT-3 (175B)
  - T5+LM (11B)
  - T0 (11B)
- Uniformly sampled from datasets and prompts
- Evaluated on held out task types, across prompts
Performance on held-out tasks
BIG-Bench

- Evaluation data set meant to test very different tasks
- Comparison with 3 Google LMs (8.5B, 28B, 68B)
- Three versions of T0 11B trained with different tasks.
BIG-Bench

Based only on the information contained in a brief quote from Wikipedia, answer whether the related claim is True, False or Neither. Use Neither when the Wikipedia quote does not provide the necessary information to resolve the question. Input: {claim} ....
Performance on BIG-Bench subset

[Bar charts showing performance across different tasks and models]
More prompts are better than one
More prompts are better than one
Adding datasets (usually) helps
Adding datasets (usually) helps
Finetuned Language Models Are Zero-Shot Learners

Jason Wei* Maarten Bosma* Vincent Y. Zhao* Kelvin Guu* Adams Wei Yu Brian Lester Nan Du Andrew M. Dai Quoc V. Le
Google Research

Finetune on many tasks ("instruction-tuning")

Input (Commonsense Reasoning)
Here is a goal: Get a cool sleep on summer days. How would you accomplish this goal?
OPTIONS:
- Keep stack of pillow cases in fridge.
- Keep stack of pillow cases in oven.
Target keep stack of pillow cases in fridge

Input (Translation)
Translate this sentence to Spanish:
The new office building was built in less than three months.

Target
El nuevo edificio de oficinas se construyó en tres meses.

Inference on unseen task type

Input (Natural Language Inference)
Premise: At my age you will probably have learnt one lesson.
Hypothesis: It's not certain how many lessons you'll learn by your thirties.
Does the premise entail the hypothesis?
OPTIONS:
- yes
- it is not possible to tell
- no

FLAN Response
It is not possible to tell

Performance on unseen task types

<table>
<thead>
<tr>
<th>Task Type</th>
<th>GPT-3 175B zero-shot</th>
<th>GPT-3 175B few-shot</th>
<th>FLAN 137B zero-shot</th>
</tr>
</thead>
<tbody>
<tr>
<td>Natural language inference</td>
<td>42.9</td>
<td>56.2</td>
<td>56.6</td>
</tr>
<tr>
<td>Reading Comprehension</td>
<td>63.7</td>
<td>72.6</td>
<td>77.4</td>
</tr>
<tr>
<td>Closed-Book QA</td>
<td>49.8</td>
<td>55.7</td>
<td>56.6</td>
</tr>
</tbody>
</table>
B

Performance on **held-out** tasks

- Instruction tuning
- Untuned model

Average zero-shot accuracy on 13 held-out tasks (%) against Model Size (number of parameters).
Performance on **held-out** tasks

Average zero-shot accuracy on 13 held-out tasks (%)

- Instruction tuning
- Untuned model

$T0 = 11B$ parameters
Caveats

- Task accuracy is dependent on the prompt format / wording
- For each of these tasks numbers are low in an absolute sense (zero-shot)
- Approach does not extend automatically to in-context learning (Natural instructions Wang et al. 2022)
- No evidence (in this work) of prompt understanding in a complex sense
Usage
Premise: World leaders expressed concern that North Korea will quit six-party nuclear disarmament talks and will bolster its nuclear weapons arsenal.

Hypothesis: North Korea says it has a stockpile of nuclear weapons and is building more.

from “Prompt Consistency for Zero-Shot Task Generalization”, Zhou et al. 2022
Susie loves her grandma’s banana bread. Susie called her grandma and asked her to send some. Grandma lived very far away. A week passed and grandma surprised Susie by coming to visit. What is a possible continuation for the story?

from "Few-Shot Parameter-Efficient Fine-Tuning is Better and Cheaper than In-Context Learning", Liu et al. 2022
http://prompt.vizhub.ai/

From PromptIDE, Strobelt et al. 2022
Epilogue: BLOOM
Large-scale Public Compute

Jean Zay supercomputer at Orsay, France.

Accelerated partition (or GPU partition)

- 261 four-GPU accelerated compute nodes with:
  - 2 Intel Cascade Lake 6248 processors (20 cores at 2.5 GHz), namely 40 cores per node
  - 192 GB of memory per node
  - 4 Nvidia Tesla V100 SXM2 GPUs (32 GB)

- 31 eight-GPU accelerated compute nodes, currently dedicated to the AI community with:
  - 2 Intel Cascade Lake 6226 processors (12 cores at 2.7 GHz), namely 24 cores per node
  - 20 nodes with 384 GB of memory and 11 nodes with 768 GB of memory
  - 8 Nvidia Tesla V100 SXM2 GPUs (32 GB)

- Extension in the summer of 2020, 351 four-GPU accelerated compute nodes with:
  - 2 Intel Cascade Lake 6248 processors (20 cores at 2.5 GHz), namely 40 cores per node
  - 192 GB of memory per node
  - 4 Nvidia Tesla V100 SXM2 GPUs (16 GB)

- Cumulated peak performance of 28 Pflop/s with a total of **2696 Nvidia V100 GPUs**
- JZ 3 expands to 3,152 GPUs (V100s and A100s) - use time: 3 months
- **Omni-PAth interconnection** network 100 Gb/s : 4 links per converged node
- Parallel storage w/capacity of **2.2 PB SSD disks** (GridScaler GS18K SSD)
How to Train a Language Model

Carbon footprint
Biomedical/Mathematics/Historical
Interpretability & Visualization
Tokenization
Evaluation
Legal
Data
Model
Privacy

Extrinsic
Bias & Fairness
Few-shot
Intrinsic
Multilinguality

Governance
Sourcing
Tooling
Approach
Architecture
Metadata
Multilinguality
Prompt engineering
Retrieval

Accessibility
Engineering/scaling
Organization/social impact

LLM

Approach
Architecture
Metadata
Multilinguality
Prompt engineering
Retrieval
Here are the last 3 months of 104B GPT2 trial-and-errors at @BigscienceW in pictures and lessons learned:

github.com/bigscience-wor...

It's hard!

Wishing great breakthroughs to all in the New Year!

Scaling loss curves for the @BigScienceLLM training are nice and smooth - training the beast still feels a bit terrifying but at least the loss curve for the big model is on trend.
Ya puedes usar BLOOM, una IA de código abierto más potente que GPT-3 que es capaz de generar texto en 59 lenguajes
https://github.com/bigscience-workshop/t-zero
https://huggingface.co/bigscience/T0{p,pp,-3B}