Toolformer: Language Models Can Teach Themselves to Use Tools

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, Thomas Scialom

Published on
Feb 9, 2023 (arXiv)

Presented by
Giuseppe Attanasio on March 17, 2023

Bocconi

“(LMs) struggle with basic functionality, such as arithmetic or factual lookup, where much simpler and smaller models excel.”

Bocconi

The New England Journal of Medicine is a registered trademark of [QA("Who is the publisher of The New England Journal of Medicine?") → Massachusetts Medical Society] the MMS.

Out of 1400 participants, 400 (or [Calculator(400 / 1400) → 0.29] 29%) passed the test.

The name derives from "la tortuga", the Spanish word for [MT("tortuga") → turtle] turtle.

The Brown Act is California's law [WikiSearch("Brown Act") → The Ralph M. Brown Act is an act of the California State Legislature that guarantees the public's right to attend and participate in meetings of local legislative bodies.] that requires legislative bodies, like city councils, to hold their meetings open to the public.

1. Self-supervised:
   no large human annotations

2. *General*:
   should decide *when* and
   which API to call

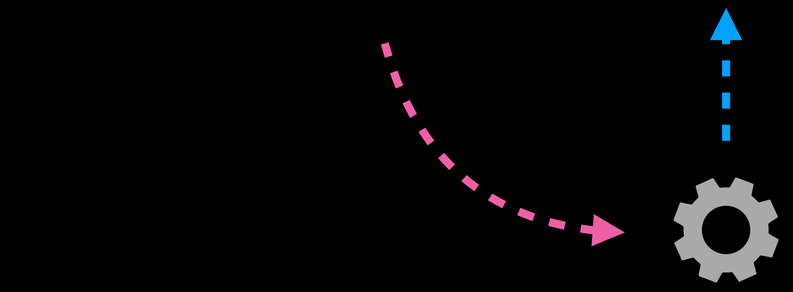
Boring LMs

“The name derives from “la tortuga”, the Spanish word for turtle."

Bocconi

Invoking an API as a language generation phenomenon

“The name derives from “la tortuga”, the Spanish word for <API> MT(tortuga) -> turtle </API> turtle.”



Bocconi

Toolformer: nuts and bolts

A three-step process

(1) Sampling API calls in a large corpus

(2) Filtering most promising API calls

(3) Fine-tuning the model

Sampling API calls

- Sampling: in-context learning
  - LM decides where to put an API call
    - $p_i = p_{LM}(< API > | x_{0,i-1})$
- Executing the calls
  - Another neural network (MT)
    Retrieval system
    Calculator
    Calendar
  - I/O must be a text

Your task is to add calls to a Question Answering API to a piece of text. The questions should help you get information required to complete the text. You can call the API by writing "[QA(question)]" where "question" is the question you want to ask. Here are some examples of API calls:

Input: Joe Biden was born in Scranton, Pennsylvania.

Output: Joe Biden was born in **[QA("Where was Joe Biden born?")]** Scranton, **[QA("In which state is Scranton?")]** Pennsylvania.

Input: Coca-Cola, or Coke, is a carbonated soft drink manufactured by the Coca-Cola Company.

Output: Coca-Cola, or **[QA("What other name is Coca-Cola known by?")]** Coke, is a carbonated soft drink manufactured by **[QA("Who manufactures Coca-Cola?")]** the Coca-Cola Company.

Input: x

Output:

Filtering the API

A good call is a call that reduces the perplexity on trailing tokens

API text prefixed to not break the flow

(API on) 
<API> MT(tortuga) -> turle </API>
The name derives from “la tortuga”, the Spanish word for turtle.

(No response)
<API> MT(tortuga) </API>
The name derives from “la tortuga”, the Spanish word for turtle.

(No API)
The name derives from “la tortuga”, the Spanish word for turtle.

Bocconi

Model Finetuning

- Augment the original dataset with API calls

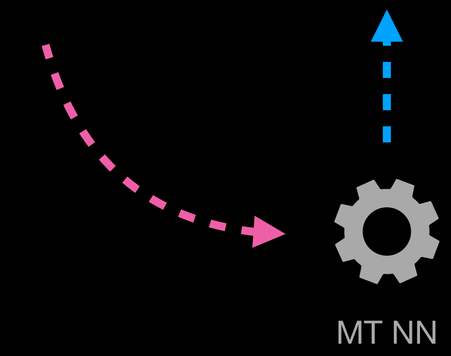
> The name derives from “la tortuga”, the Spanish word for <API> MT(tortuga) -> turtle </API> turtle.

- Standard language modeling objective
  - LM learns when and which API call to call

Inference

- Generating the sequence: *“<API> api name ->”* triggers the invoking external tools

“The name derives from “la cama”, the Spanish word for <API> MT(cama) -> bed </API> bed.”



Bocconi

APIs: External Systems

- Question Answering: Atlas (retrieval-augmented LM)
- Calculator
- Wikipedia Search: BM25 retriever
- Machine Translation System: NLLB any->Eng LM
- Calendar: current date

Evaluating Toolformer

Bocconi

Setup (1/2)

- Augmented Dataset: CCNet
- Models
  - GPT-J (fine-tuned on CCNet, CCNet augmented, and CCNet augmented but with API disabled during decoding)
  - OPT, GPT3 (pre-InstructGPT),
- Tasks
  - Factual Knowledge (LAMA, TempLAMA), Math (ASDiv, SVAMP, MAWPS), Question Answering (Web Questions, Natural Questions, TriviaQA, MLQA)

Disable means setting prob of the <API> token to zero

Setup 2/2

- Decoding
  - Greedy
  - If “<*API*>” is in the top-10, choose it

Factual Knowledge

Model	SQuAD	Google-RE	T-REx
GPT-J	17.8	4.9	31.9
GPT-J + CC	19.2	5.6	33.2
Toolformer (disabled)	22.1	6.3	34.9
Toolformer	**33.8**	**11.5**	**53.5**
OPT (66B)	21.6	2.9	30.1
GPT-3 (175B)	26.8	7.0	39.8

API usage:
98.1%

Math

Model	ASDiv	SVAMP	MAWPS
GPT-J	7.5	5.2	9.9
GPT-J + CC	9.6	5.0	9.3
Toolformer (disabled)	14.8	6.3	15.0
Toolformer	**40.4**	**29.4**	**44.0**
OPT (66B)	6.0	4.9	7.9
GPT-3 (175B)	14.0	10.0	19.8

API usage:
97.9%

QA

Model	WebQS	NQ	TriviaQA
GPT-J	18.5	12.8	43.9
GPT-J + CC	18.4	12.2	45.6
Toolformer (disabled)	18.9	12.6	46.7
Toolformer	**26.3**	**17.7**	**48.8**
OPT (66B)	18.6	11.4	45.7
GPT-3 (175B)	29.0	22.6	65.9

API usage:
99.3%

Multilingual QA

Model	Es	De	Hi	Vi	Zh	Ar
GPT-J	15.2	**16.5**	1.3	8.2	**18.2**	**8.2**
GPT-J + CC	15.7	14.9	0.5	8.3	13.7	4.6
Toolformer (disabled)	19.8	11.9	1.2	10.1	15.0	3.1
Toolformer	**20.6**	13.5	**1.4**	**10.6**	16.8	3.7
OPT (66B)	0.3	0.1	1.1	0.2	0.7	0.1
GPT-3 (175B)	3.4	1.1	0.1	1.7	17.7	0.1
GPT-J (All En)	24.3	27.0	23.9	23.3	23.1	23.6
GPT-3 (All En)	24.7	27.2	26.1	24.9	23.6	24.0

API usage* (min-max):
63.8%-94.9%

**MT API on Hindi: 7.3%*

DATESET:
"What day of the week was it 30 days ago?"

Temporal Datasets

Model	TEMPLAMA	DATESET
GPT-J	13.7	3.9
GPT-J + CC	12.9	2.9
Toolformer (disabled)	12.7	5.9
Toolformer	**16.3**	**27.3**
OPT (66B)	14.5	1.3
GPT-3 (175B)	15.5	0.8

API usage:
Calendar 0.2% on TempLAMA, 54.8% on DATESET

Bocconi

My take (and, partially, the authors')

Thanks!

- Making beyond-language-only models is cool
- Invoking APIs via language generation is a cool idea
  - It's self-supervised
    - In-context learning gets better and better
  - Paves the way to new LM-other systems interactions
- One-pass decoding prevents multi-step API interaction
- To be seen how it fits in the post GPT-4 era

Bocconi