

[EMNLP20]How Much Knowledge Can You Pack Into the Parameters of a Language Model?

*Adam Roberts,[|] Colin Raffel,[|] Noam Shazeer,[|]
[|]Google*

Speaker: 杨晰

xyang41@stu.ecnu.edu.cn

Outline

- Motivation
- Background
 - Question Answering (QA)
 - Transfer Learning with Language Models (LMs)
- Experiments
 - Pretrained LM:T5
 - QA Datasets
 - Results
- Conclusions
- Future work

Motivation

- Neural language models(LMs) that pre-trained on unlabeled text can **internalize** a sort of implicit “knowledge base” → useful

Motivation

- Neural language models(LMs) that pre-trained on unlabeled text can **internalize** a sort of implicit “knowledge base” → useful
- unstructured and unlabeled text data is freely available in huge quantities on the Internet
- it's possible to retrieve information using informal natural language queries

Motivation

- Neural language models(LMs) that pre-trained on unlabeled text can **internalize** a sort of implicit “knowledge base” → useful
- unstructured and unlabeled text data is freely available in huge quantities on the Internet
- it's possible to retrieve information using informal natural language queries



Evaluate the capability of language models on the practical task of **open domain closed book question answering** (without access to any external knowledge or context)

BG: Question Answering

- Reading comprehensive
 - input: the **question** + **context** (containing the answer)
 - output: the **span** that contains the answer / the **text** of answer


BG: Question Answering

- Reading comprehensive
 - input: the **question** + **context** (containing the answer)
 - output: the **span** that contains the answer / the **text** of answer
- Open-domain open-book QA
 - input: the **question** (context-independent)+ **an external collection of knowledge**
 - output: the **text** of answer

BG: Question Answering

- Reading comprehensive
 - input: the **question** + **context** (containing the answer)
 - output: the **span** that contains the answer / the **text** of answer
- Open-domain open-book QA
 - input: the **question** (context-independent)+ **an external collection of knowledge**
 - output: the **text** of answer
- ➡ Open-domain closed book QA
 - input: the **question** (context-independent)
 - output: the **text** of answer

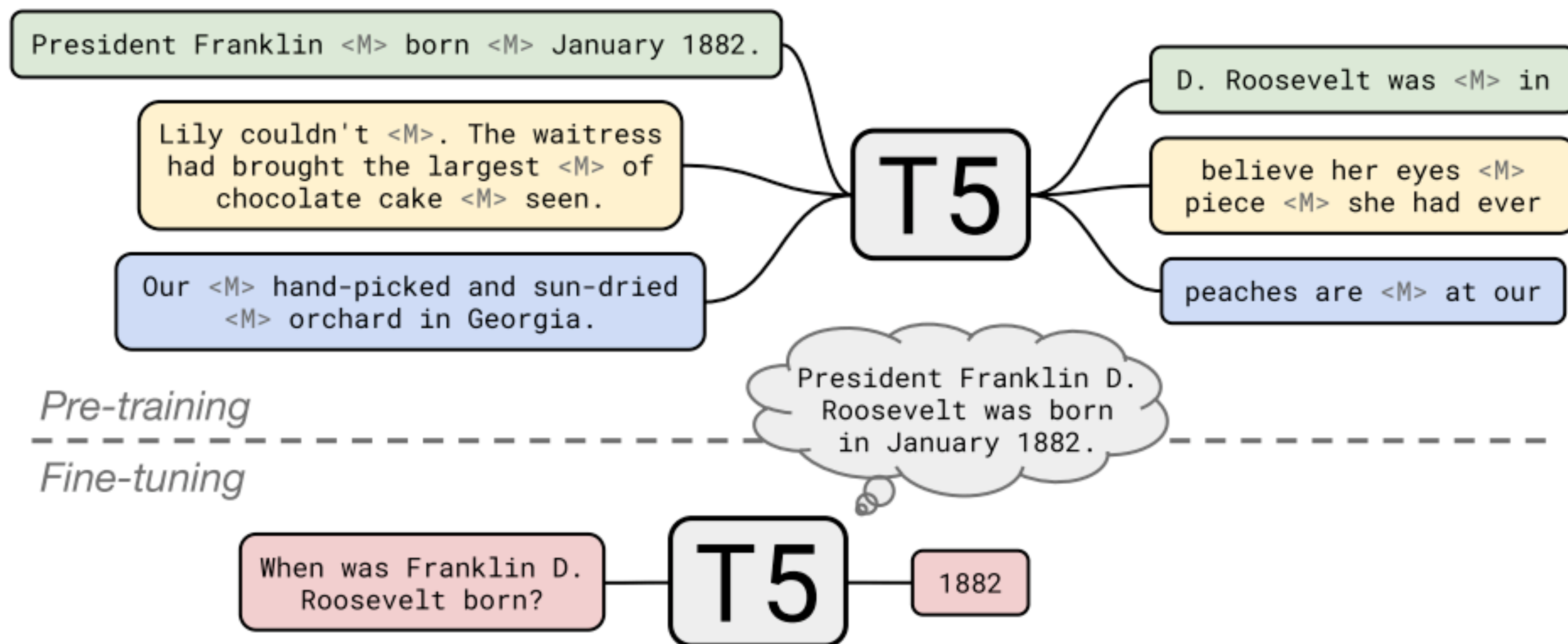
BG: Transfer Learning with LMs

- Pre-train LMs, fine-tune them on downstream tasks
- Transformer-based pre-trained LMs
 - Encoder-only: BERT → not applicable to closed-book QA
 - Encoder-Decoder: Text-to-Text Transfer Transformer (T5) 

Motivation

- Evaluate the capability of T5 language model on the practical task of open domain closed book question answering (without access to any external knowledge or context)

Pre-trained Model:T5



Pre-trained Model: T5

- T5
 - unsupervised task: “span corruption”
 - supervised tasks: translation, summarization, classification, reading comprehension
- T5.I.I
 - unsupervised task only

QA Datasets

- Natural Questions example

Example 1

Question: what color was john wilkes booth's hair

Wikipedia Page: John_Wilkes_Booth

Long answer: Some critics called Booth “the handsomest man in America” and a “natural genius”, and noted his having an “astounding memory”; others were mixed in their estimation of his acting. He stood 5 feet 8 inches (1.73 m) tall, had jet-black hair, and was lean and athletic. Noted Civil War reporter George Alfred Townsend described him as a “muscular, perfect man” with “curling hair, like a Corinthian capital”.

Short answer: jet-black

QA Datasets

- TriviaQA example

Question: The Dodecanese Campaign of WWII that was an attempt by the Allied forces to capture islands in the Aegean Sea was the inspiration for which acclaimed 1961 commando film?

Answer: The Guns of Navarone

Excerpt: The Dodecanese Campaign of World War II was an attempt by Allied forces to capture the Italian-held Dodecanese islands in the Aegean Sea following the surrender of Italy in September 1943, and use them as bases against the German-controlled Balkans. The failed campaign, and in particular the Battle of Leros, inspired the 1957 novel **The Guns of Navarone** and the successful 1961 movie of the same name.

QA Datasets

- Natural Questions
 - input: the **question** + **Wikipedia page** (containing the answer)
- WebQuestions
 - input: the **question** + **Freebase** (external KB)
- TriviaQA
 - input: the **question** + **excerpt** (containing the answer)

QA Datasets

- Natural Questions
 - input: the **question** + ~~Wikipedia page~~ (containing the answer)
- WebQuestions
 - input: the **question** + ~~Freebase~~ (external KB)
- TriviaQA
 - input: the **question** + ~~excerpt~~ (containing the answer)

Experiments

- Results

220 million parameters

770 million parameters

3 billion parameters

11 billion parameters



	NQ	WQ	TQA	
			dev	test
Chen et al. (2017)	–	20.7	–	–
Lee et al. (2019)	33.3	36.4	47.1	–
Min et al. (2019a)	28.1	–	50.9	–
Min et al. (2019b)	31.8	31.6	55.4	–
Asai et al. (2019)	32.6	–	–	–
Ling et al. (2020)	–	–	35.7	–
Guu et al. (2020)	40.4	40.7	–	–
Férvy et al. (2020)	–	–	43.2	53.4
Karpukhin et al. (2020)	41.5	42.4	57.9	–
T5-Base	25.9	27.9	23.8	29.1
T5-Large	28.5	30.6	28.7	35.9
T5-3B	30.4	33.6	35.1	43.4
T5-11B	32.6	37.2	42.3	50.1
T5-11B + SSM	34.8	40.8	51.0	60.5
T5.1.1-Base	25.7	28.2	24.2	30.6
T5.1.1-Large	27.3	29.5	28.5	37.2
T5.1.1-XL	29.5	32.4	36.0	45.1
T5.1.1-XXL	32.8	35.6	42.9	52.5
T5.1.1-XXL + SSM	35.2	42.8	51.9	61.6

SSM: Salient Span Masking

Experiments

- Human Evaluation

Table 2: A breakdown of the 150 hand-evaluated examples from Natural Questions where the T5 predictions were labelled as incorrect by the automatic procedure. We found only 62% of these to be true negatives

Category	Percentage	Example		
		Question	Target(s)	T5 Prediction
True Negative	62.0%	what does the ghost of christmas present sprinkle from his torch	little warmth, warmth	confetti
Phrasing Mismatch	13.3%	who plays red on orange is new black	kate mulgrew	katherine kiernan maria mulgrew
Incomplete Annotation	13.3%	where does the us launch space shuttles from	florida	kennedy lc39b
Unanswerable	11.3%	who is the secretary of state for northern ireland	karen bradley	james brokenshire

Experiments

- Human Evaluation

Table 2: A breakdown of the 150 hand-evaluated examples from Natural Questions where the T5 predictions were labelled as incorrect by the automatic procedure. We found only 62% of these to be true negatives

Category	Percentage	Example		
		Question	Target(s)	T5 Prediction
True Negative	62.0%	what does the ghost of christmas present sprinkle from his torch	little warmth, warmth	confetti
Phrasing Mismatch	13.3%	who plays red on orange is new black	kate mulgrew	katherine kiernan maria mulgrew
Incomplete Annotation	13.3%	where does the us launch space shuttles from	florida	kennedy lc39b
Unanswerable	11.3%	who is the secretary of state for northern ireland	karen bradley	james brokenshire

Performance is underestimated

Conclusions

- large LMs pre-trained on unstructured text can attain **competitive results on open-domain QA benchmarks** without any access to external knowledge.

Future work

- close-book models on more efficient LM
- This model distributes knowledge in its parameters in an **inexplicable** way
- Maximum-likelihood objective cannot teach the model to learn the fact, **cannot explicitly update** or **remove** knowledge from a pre-trained model
- Measure on QA tasks that need reasoning capabilities

Thanks

Q & A