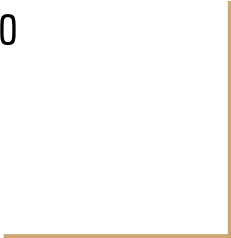


Knowledge and Language Models

Hyunjun “John” Yoo



Overview -- The Two Papers

- Language Models as Knowledge Bases?
- How Much Knowledge Can You Pack Into the Parameters of a Language Model?



Language Models as Knowledge Bases?



Introduction

- Pretraining language models on large text corpora led to progress in NLP tasks.
- Able to store *relational* knowledge and answer cloze statements

“Actions speak louder than _____”


Advantages over traditional knowledge bases?

<u>Traditional Knowledge Base</u>	<u>Language Models</u>
Requires complex NLP pipelines and schema engineering to extract relational data.	Potentially can answer relational queries directly.
Human supervision needed during training.	No human supervision needed during training.

Background -- Unidirectional Language Models

- These models predict the probability of a sequence of words by considering each word in the context of the words that precede it.

$$p(w) = \prod_t p(w_t | w_{t-1}, \dots, w_1)$$


$$p(w_t | w_{t-1}, \dots, w_1) = \text{softmax}(Wh_t + b)$$

Unidirectional Language Models Used In Study

- **fairseq-fconv**

- Convolutional sequence-to-sequence model implemented in Meta's fairseq library

- **Transformer-XL**

- Based on Transformer (Attention Is All You Need reference) but can "take into account a longer history by caching previous outputs and by using relative instead of absolute positional encoding"

Bidirectional Language Models Used In Study

- **ELMo**

- Estimates probabilities in a bidirectional context using LSTM (long short term memory) networks

- **BERT (This is most important)**

- Uses Transformer architecture and “samples positions in the input sequence randomly and learn to fill the word at the masked position”

Model	Base Model	#Parameters	Training Corpus	Corpus Size
fairseq-fconv (Dauphin et al., 2017)	ConvNet	324M	WikiText-103	103M Words
Transformer-XL (large) (Dai et al., 2019)	Transformer	257M	WikiText-103	103M Words
ELMo (original) (Peters et al., 2018a)	BiLSTM	93.6M	Google Billion Word	800M Words
ELMo 5.5B (Peters et al., 2018a)	BiLSTM	93.6M	Wikipedia (en) & WMT 2008-2012	5.5B Words
BERT (base) (Devlin et al., 2018a)	Transformer	110M	Wikipedia (en) & BookCorpus	3.3B Words
BERT (large) (Devlin et al., 2018a)	Transformer	340M	Wikipedia (en) & BookCorpus	3.3B Words

Table 1: Language models considered in this study.

Purpose of This Study

- Studies on pretrained models have primarily focused on their linguistic and semantic capabilities and their performance on NLP tasks.
- **Rather than studying linguistic knowledge in pre-trained models, we want to know how much factual and commonsense knowledge is stored instead.**

The LAMA (LAnguage Model Analysis) Probe

A corpus of facts that test the factual and commonsense knowledge in language models.

<https://github.com/facebookresearch/LAMA>

What are these “facts”?

- **Subject-relation-object triples**
 - “Paris is the capital of France”

- **Question-answer pairs**
 - “What is the capital of Missouri? Jefferson City.”

Facts converted to cloze statements

- Each fact is converted into a “fill-in-the-blank” aka cloze statement.

“Paris is the capital of _____”

- Models are then asked to fill in the missing token based on the knowledge learnt from their training data.

Knowledge Sources Used In Study

- **Google-RE**
 - 60k facts manually extracted from Wikipedia
 - Focuses on three relations: place of birth, date of birth, place of death
- **T-REx**
 - Contains larger set of data than Google-RE from Wikidata
 - Includes wider range of relationships
- **ConceptNet**
 - Knowledge base built on top of OMCS (Open Mind Common Sense) sentences
- **SQuAD**
 - Commonly used for training/evaluating question-answering models

Baselines for Results Comparison -- Freq

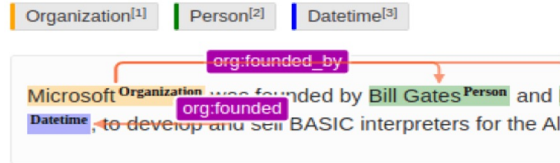
- Ranks possible answers based on how often they appear in a certain context of test data

“Indicates the upper bound performance of a model that always predicts the same objects for a particular relation”

Baselines -- RE (Relation Extraction)

- RE model uses LSTM-based neural networks and attention mechanisms to extract relational triples from sentences.

RE_m	RE_o
Uses exact string matching to link extracted entities to the subjects and objects of interest.	Uses oracle for entity linking as well as string matching for more accurate matching.



Baselines -- DrQA

- Question-answering system designed for open-domain queries.

Information Retrieval

Uses TF/IDF (Term Frequency/Inverse Document Frequency) to find relevant Wikipedia articles.



Reading Comprehension

A neural model then reads and extracts answers from the top k articles found.

Results! Google-RE and T-REx

Corpus	Relation	Statistics		Baselines		KB		LM					
		#Facts	#Rel	Freq	DrQA	RE _n	RE _o	Fs	Txl	Eb	E5B	Bb	Bl
Google-RE	birth-place	2937	1	4.6	-	3.5	13.8	4.4	2.7	5.5	7.5	14.9	16.1
	birth-date	1825	1	1.9	-	0.0	1.9	0.3	1.1	0.1	0.1	1.5	1.4
	death-place	765	1	6.8	-	0.1	7.2	3.0	0.9	0.3	1.3	13.1	14.0
	Total	5527	3	4.4	-	1.2	7.6	2.6	1.6	2.0	3.0	9.8	10.5
T-REx	1-1	937	2	1.78	-	0.6	10.0	17.0	36.5	10.1	13.1	68.0	74.5
	<i>N</i> -1	20006	23	23.85	-	5.4	33.8	6.1	18.0	3.6	6.5	32.4	34.2
	<i>N</i> - <i>M</i>	13096	16	21.95	-	7.7	36.7	12.0	16.5	5.7	7.4	24.7	24.3
	Total	34039	41	22.03	-	6.1	33.8	8.9	18.3	4.7	7.1	31.1	32.3
ConceptNet	Total	11458	16	4.8	-	-	-	3.6	5.7	6.1	6.2	15.6	19.2
SQuAD	Total	305	-	-	37.5	-	-	3.6	3.9	1.6	4.3	14.1	17.4

Table 2: Mean precision at one (P@1) for a frequency baseline (Freq), DrQA, a relation extraction with naïve entity linking (RE_n), oracle entity linking (RE_o), fairseq-fconv (Fs), Transformer-XL large (Txl), ELMo original (Eb), ELMo 5.5B (E5B), BERT-base (Bb) and BERT-large (Bl) across the set of evaluation corpora.

Fs: fairseq-fconv, **Txl**: Transformer XL large, **Eb**: ELMo, **E5B**: ELMo 5.5B, **Bb**: BERT, **Bl**: BERT-large

T-REx Scores for N-1 Relations

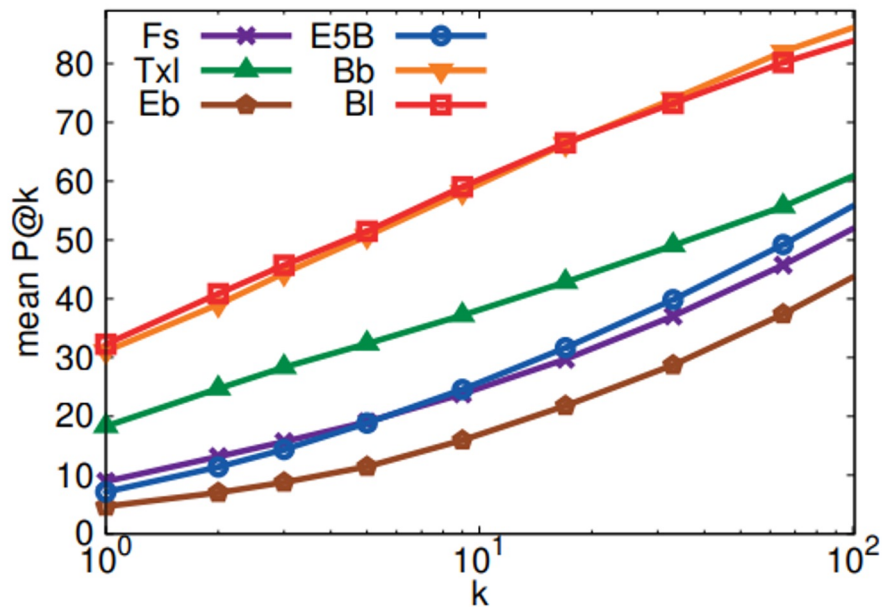


Figure 2: Mean P@k curve for T-REx varying k. Base-10 log scale for X axis.

Results! ConceptNet and SQuAD

Corpus	Relation	Statistics		Baselines		KB		LM					
		#Facts	#Rel	Freq	DrQA	RE _n	RE _o	Fs	Txl	Eb	E5B	Bb	Bl
Google-RE	birth-place	2937	1	4.6	-	3.5	13.8	4.4	2.7	5.5	7.5	14.9	16.1
	birth-date	1825	1	1.9	-	0.0	1.9	0.3	1.1	0.1	0.1	1.5	1.4
	death-place	765	1	6.8	-	0.1	7.2	3.0	0.9	0.3	1.3	13.1	14.0
	Total	5527	3	4.4	-	1.2	7.6	2.6	1.6	2.0	3.0	9.8	10.5
T-REx	1-1	937	2	1.78	-	0.6	10.0	17.0	36.5	10.1	13.1	68.0	74.5
	N-1	20006	23	23.85	-	5.4	33.8	6.1	18.0	3.6	6.5	32.4	34.2
	N-M	13096	16	21.95	-	7.7	36.7	12.0	16.5	5.7	7.4	24.7	24.3
	Total	34039	41	22.03	-	6.1	33.8	8.9	18.3	4.7	7.1	31.1	32.3
ConceptNet	Total	11458	16	4.8	-	-	-	3.6	5.7	6.1	6.2	15.6	19.2
SQuAD	Total	305	-	-	37.5	-	-	3.6	3.9	1.6	4.3	14.1	17.4

Table 2: Mean precision at one (P@1) for a frequency baseline (Freq), DrQA, a relation extraction with naïve entity linking (RE_n), oracle entity linking (RE_o), fairseq-fconv (Fs), Transformer-XL large (Txl), ELMo original (Eb), ELMo 5.5B (E5B), BERT-base (Bb) and BERT-large (Bl) across the set of evaluation corpora.

In terms of P@10, BERT-Large performs almost as well as DrQA (57.1 vs 63.5)

Conclusions

- BERT-large shown to have superior factual knowledge compared to its competitors
 - Competitive with traditional, supervised non-neural methods
- Directly extracting a knowledge base from text that is on par with BERT-large is non-trivial.
 - Even when given data likely to express target facts and aided by a generous entity-linking oracle, it did not perform as well as BERT-large.

Verdict

Language models provide a sufficient alternative to knowledge bases!



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



Introduction

- Large pre-trained neural language models excel in NLP tasks and can act as implicit knowledge bases.



Purpose of Study

- Unlike prior studies, this work assesses these models on open-domain question answering without external data.
- It also investigates if larger models with more parameters can store and retrieve more information.
- The study uses the "T5" model series, including a model with about 11 billion parameters, to see how model size impacts knowledge retrieval.

Transfer Learning

- LM is first trained on a large, unstructured text dataset (pre-training) and then fine-tuned for a specific NLP task.
- The current favored models for NLP transfer learning are Transformer-based, especially "encoder-only" models like BERT, but...
 - Encoder-only models don't work for closed-book QA, which provides no context.
 - Instead, "encoder-decoder" Transformer models where every NLP problem is treated as a text-to-text problem are applicable because they generate answers directly

Experiment -- Datasets Used

- **Natural Questions**

- A dataset containing real web search questions, with each one linked to a Wikipedia article containing the answer.

- **WebQuestions**

- Consists of web search questions associated with answers from FreeBase, a structured database of common facts.

- **TriviaQA**

- Features trivia questions from quizzes, with each question paired with documents from web and Wikipedia searches that might contain the answer.

How Do We Evaluate?

- **WebQuestions and TriviaQA**

- Answers are compared to ground truth after normalization (lowercasing, removing articles, punctuation, and extra spaces)

- **Natural Questions**

- Open-domain format that requires a single normalized answer (like one above)
- Multi-answer reading comprehension format

“Where did the spaceship launch? (multiple places)”



Models Used In Training

- Experiments were conducted with different sizes of T5 models to see how performance scales with size. The sizes range from Base (220M parameters) to Large (770M), 3B (3 billion), and the largest at 11B (11 billion) parameters.
- Additionally, the study also uses the T5.1.1 checkpoints that were pre-trained only on **unlabeled** data.

Fine-tuning Procedure

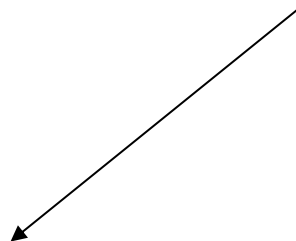
- The T5 models were fine-tuned following the protocol of the original T5 study without hyperparameter adjustments.
 - AdaFactor optimizer
 - Constant learning rate of 0.001
 - 10% dropout rate
 - Batch size of 196608 tokens
- For WebQuestions, we halve the batch size and double the dropout rate due to being a smaller dataset.
- T5.1.1 models have a 5% dropout rate instead.

Validation Step

10% of training set
used for validation
for each dataset



Train for 20000
steps, but usually
validation accuracy
plateaued after a
couple hundred
steps.

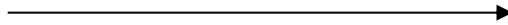


Best-performing
checkpoint used for final
evaluation.

Salient Span Masking (SSM)

- SSM is a pre-training objective where the model is trained to identify and reconstruct important information (salient spans) like named entities and dates from text sentences.

The approach begins with using BERT to find sentences in Wikipedia that contain these important pieces of information.



The model is then trained to fill in the blanks where this information has been masked out.

Results!

Table 1: Scores achieved by fine-tuning T5 on the open-domain Natural Questions (NQ), WebQuestions (WQ), and TriviaQA (TQA) tasks.

	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

Comparison With Open-Book Systems

- The models, particularly T5-11B and T5.1.1-XXL with SSM, outperform or are competitive with most existing open-book question answering systems.
- These systems traditionally rely on retrieving information from an external database before generating an answer, which involves computational and memory overhead.
- The efficiency of T5 models in directly answering questions without this retrieval step could be attributed to their ability to internalize a vast amount of information during the pre-training phase.

False Negatives Example

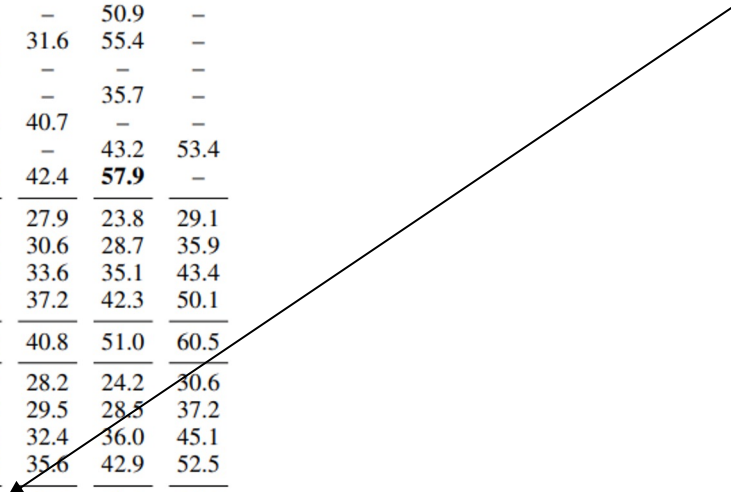
Difference in phrasing

Instances where the answer's wording differed from the ground truth but retained the same meaning (e.g., "April 15" vs. "April 15th").

Table 1: Scores achieved by fine-tuning T5 on the open-domain Natural Questions (NQ), WebQuestions (WQ), and TriviaQA (TQA) tasks.

	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

After removing unanswerable questions, the score for NQ became **57.8**



Conclusion

- More parameters mean more knowledge stored within!
 - However, the size of these models presents challenges due to the high computational resources required, which may not be feasible in settings with limited resources.
- Unlike "open-book" models that show the source of their information, the studied LLMs distribute knowledge across their parameters in a way that is not easily interpretable.

Verdict

In summary, while LLMs show promise in question answering, there are significant challenges related to their size and interpretability that need to be addressed in future research, especially to improve their reasoning capabilities.

Questions?