

Language models as knowledge bases?

Huaisheng Zhu

04/05/2023

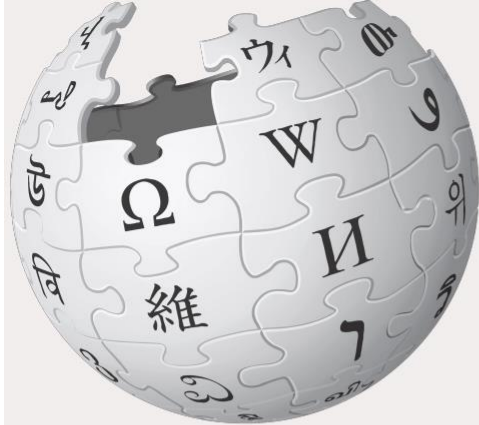


PennState

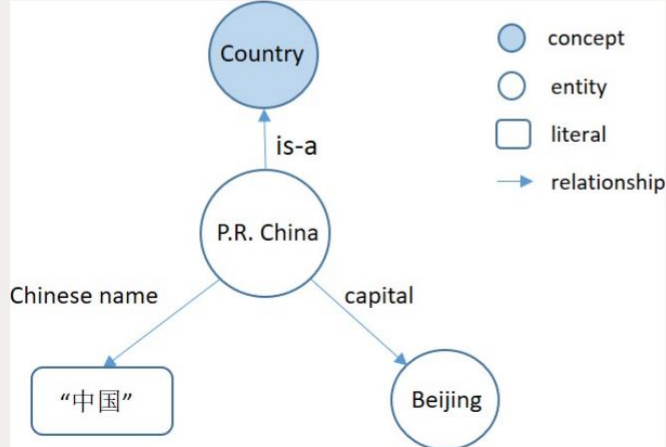
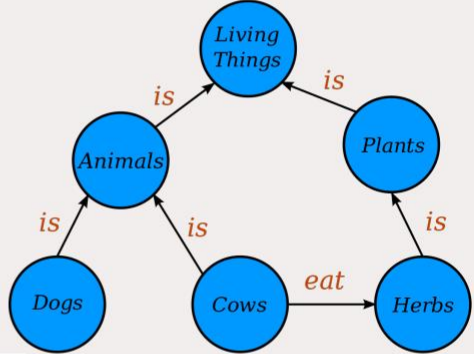
Petroni, Fabio, et al. "Language Models as Knowledge Bases?." EMNLP(2019)

Knowledge Bases

- A knowledge base allows for rapid search, retrieval, and reuse
- Stores information as answers to questions or solutions to problems
- Can be fed into a language model



WIKIPEDIA
The Free Encyclopedia



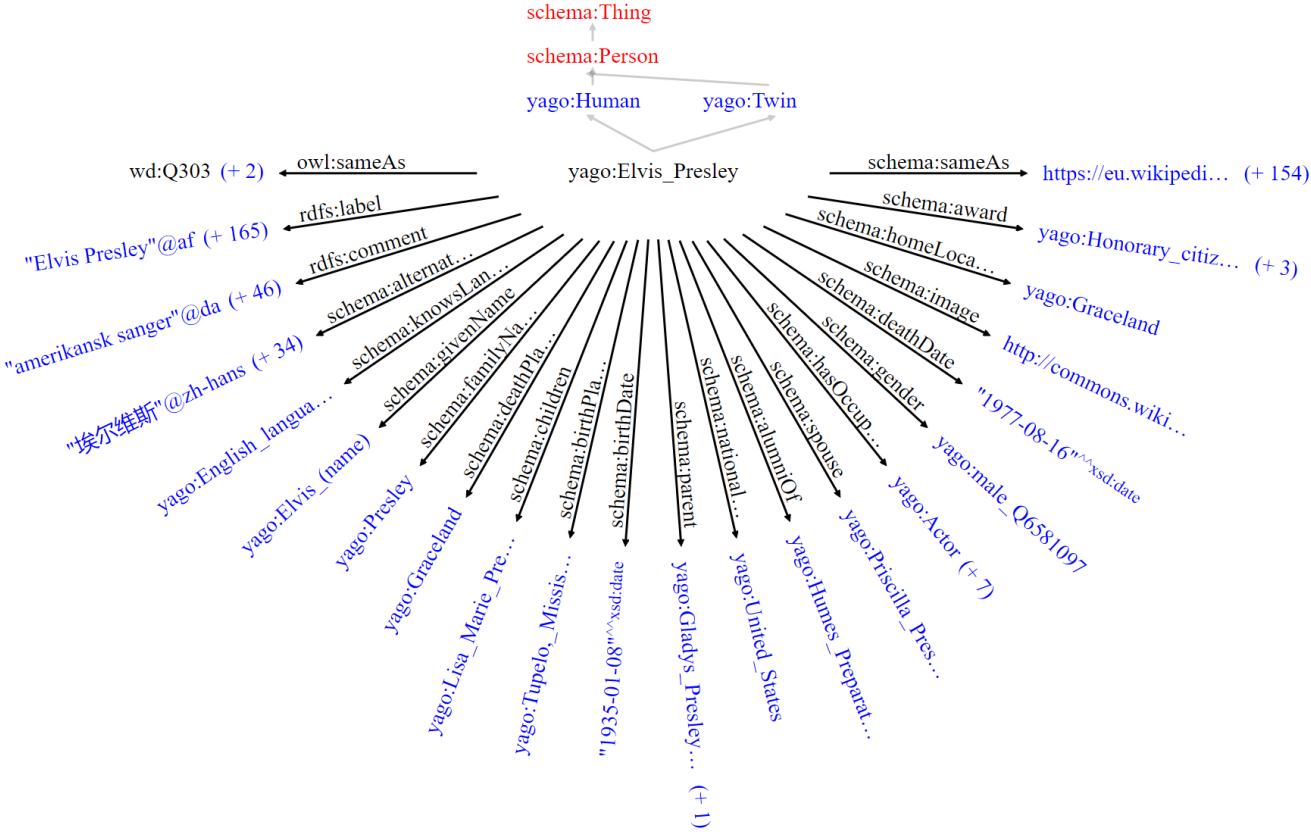
Legend:

- concept
- entity
- literal
- relationship

Examples of Knowledge Bases

YAGO is a large knowledge base with general knowledge about people, cities, countries, movies, and organizations.

Visualization as a graph



Knowledge Bases

- Concepts like classes and individuals are modeled as nodes
- Relations as edges of graphs
- Classes – concepts like documents, events, or subjects
- Individuals – instances of a class or an object
- Relations – capture relationships between classes and individuals
 - is-type-of, is-instance-of, and has-attribute

WordNet Search - 3.1

- [WordNet home page](#) - [Glossary](#) - [Help](#)

Word to search for:

Display Options:

Key: "S:" = Show Synset (semantic) relations, "W:" = Show Word (lexical) relations

Display options for sense: (gloss) "an example sentence"

Noun

- [S:](#) (n) **smile**, [smiling](#), [grin](#), [grinning](#) (a facial expression characterized by turning up the corners of the mouth; usually shows pleasure or amusement)

Verb

- [S:](#) (v) **smile** (change one's facial expression by spreading the lips, often to signal pleasure)
- [S:](#) (v) **smile** (express with a smile) "*She smiled her thanks*"

How knowledge bases are used in NLP models:

- Entity extraction – replace or augment entity occurrences in text

In fact, the **Chinese** **NORP** market has the **three** **CARDINAL** most influential names of the retail and tech space – **Alibaba** **GPE**, **Baidu** **ORG**, and **Tencent** **PERSON** (collectively touted as **BAT** **ORG**), and is betting big in the global **AI** **GPE** in retail industry space. The **three** **CARDINAL** giants which are claimed to have a cut-throat competition with the **U.S.** **GPE** (in terms of resources and capital) are positioning themselves to become the 'future **AI** **PERSON** platforms'. The trio is also expanding in other **Asian** **NORP** countries and investing heavily in the **U.S.** **GPE** based **AI** **GPE** startups to leverage the power of **AI** **GPE**. Backed by such powerful initiatives and presence of these conglomerates, the market in APAC AI is forecast to be the fastest-growing **one** **CARDINAL**, with an anticipated **CAGR** **PERSON** of **45%** **PERCENT** over **2018 - 2024** **DATE**.

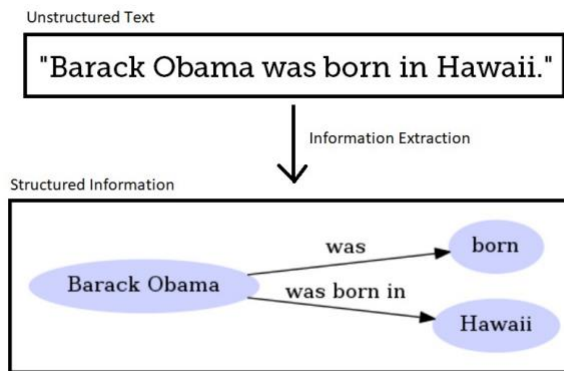
To further elaborate on the geographical trends, **North America** **LOC** has procured **more than 50%** **PERCENT** of the global share in **2017** **DATE** and has been leading the regional landscape of **AI** **GPE** in the retail market. The **U.S.** **GPE** has a significant credit in the regional trends with **over 65%** **PERCENT** of investments (including M&As, private equity, and venture capital) in artificial intelligence technology. Additionally, the region is a huge hub for startups in tandem with the presence of tech titans, such as **Google** **ORG**, **IBM** **ORG**, and **Microsoft** **ORG**.

How knowledge bases are used in NLP models:

- Coreference resolution
 - Coreference resolution is the task of finding all expressions that refer to the same entity in a text.

"I voted for Nader because he was most aligned with my values," she said.

- Entity Linking



Querying knowledge bases from these methods need supervised data. Moreover, errors can easily propagate and accumulate throughout the pipeline.

Proposed solution for the knowledge bases

- Language Models
 - Ask the model to fill in masked tokens
 - “Alex was born in [MASK]”
 - Pre-trained high-capacity models such as ELMo and BERT store vast amounts of linguistic knowledge useful for downstream tasks

The Pros:

- Requires no schema engineering
- No need for human annotations
- Supports a more diverse/open set of inquiries

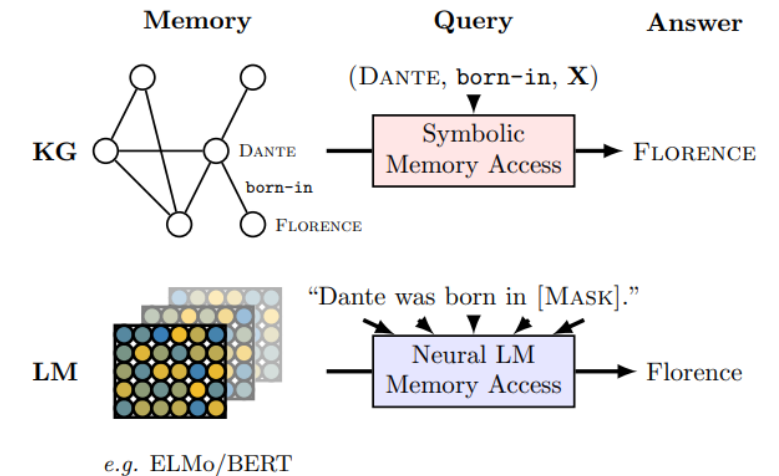


Figure 1: Querying knowledge bases (KB) and language models (LM) for factual knowledge.

Research Questions

- How much relational knowledge do they store?
- How does this differ for different types of knowledge such as facts about entities, common sense, and general question answering?
- How does their performance without fine-tuning compare to symbolic knowledge bases automatically extracted from text?

LAMA (Language Model Analysis)

- Consisting of a set of knowledge sources, each comprised of a set of facts (subject, relation, object)
- Success depends on predicting masked objects such as “Dante was born in ____”
- Tested for a variety of types of knowledge: relations between entities stored in Wikidata, common sense relations between concepts from ConceptNet, and so on.

LAMA (Language Model Analysis)

- Steps for LAMA
 - Query each model for a missing token
 - Evaluate each model based on how highly they rank the ground truth token against every word in a fixed candidate vocabulary

Knowledge Sources and datasets used

- Google-RE – contains ~60K facts manually extracted from Wikipedia
 - Only utilized 3 relations: “place of birth”, “date of birth” and “place of death”
 - manually defined a template for each considered relation, e.g., “[Adam] was born in [Illinois]” for “place of birth”
- T-Rex – is a subset of Wikidata triples
 - consider 41 Wikidata relations and subsample at most 1000 facts per relation.
 - Facts were automatically aligned to Wikipedia (can be noisy)
 - e.g. ‘Adolphe Adam died in _.’ for ‘Paris’

Knowledge Sources and datasets used

- SQuAD
 - Question-answering dataset
 - a subset of 305 context-insensitive questions with single token answers
 - Example: rewriting “Who developed the theory of relativity?” as “The theory of relativity was developed by ___”.
- ConceptNet
 - Multilingual knowledge base, initially built on top of Open Mind Common Sense sentences
 - English parts that have single-token objects covering 16 relations
 - Example: Time is ___”.

Language Models used

- Fairseq-fconv
 - Multiple layers of gated convolutions
 - Trained on the WikiText-103 corpus
- Transformer-XL
 - Large-scale LM based on the Transformer
 - Takes into account a longer history
 - Used relative instead of absolute positional encoding
 - Trained on the WikiText-103 corpus

Language Models used

- ELMO:
 - ELMo: Forward and backward LSTM, resulting in \vec{h}_i and \overleftarrow{h}_i
 - Trained on the Google Billion Word Dataset
 - ELMo 5.5B:
 - Trained on English Wikipedia and monolingual news crawl from WMT 2008-2012

Trained through averaged forward and backward probabilities from the corresponding softmax layers

Language Models used

- BERT:
 - Transformer architecture
 - Trained on the BookCorpus and English Wikipedia
 - language modelling (15% of tokens were masked and BERT was trained to predict them from context) and next sentence prediction (if a chosen next sentence was probable or not given the first sentence)
 - BERT-base (12 encoders with 12 bidirectional self-attention heads)
 - BERT-large (24 encoders with 16 bidirectional self-attention heads)

Masked the token at position t , fed output to vector corresponding to masked token (h_t) into softmax layer

Baselines

- **Freq:**
 - subject and relation pair, this baseline ranks words based on how frequently they appear as objects for the given relation in the test data
- **Relation Extraction (RE):**
 - extracts relation triples from a given sentence using an LSTM-based encoder and an attention mechanism
 - constructs a knowledge graph of triples
 - At test time, they queried this graph by finding the subject entity and then rank all objects in in the correct relation based on the confidence scores by the RE

Baselines

- **DrQA**
 - a popular system for open-domain question answering
 - Two-step pipeline:
 - First, a TF/IDF information retrieval step is used to find relevant articles from a large store of documents (e.g. Wikipedia)
 - Secondly, on the retrieved top k articles, a neural reading comprehension model then extracts answers

Metrics

- Rank-based metrics
- For multiple valid objects for Subject-Relation pair, removed all other valid objects from the candidates when ranking at test time other than the ones they were testing
- Mean precision at k (P@k)
 - For a given fact, this value is 1 if the object is ranked among the top k results, 0 otherwise

Results

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-1</i>	20006	23	23.85	-	5.4	33.8	6.1	18.0	3.6	6.5	32.4	34.2
	<i>N-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

Results

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-1</i>	20006	23	23.85	-	5.4	33.8	6.1	18.0	3.6	6.5	32.4	34.2
	<i>N-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

RE_n – naïve entity linking, i.e. exact string matching

RE_o – uses an oracle for entity-linking, i.e. any given (s, r, o) in sentence x, if any other (s', r, o') has been extracted in the same sentence, s will be linked to s', and o to o'

Results on Google-RE

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

- From earlier example, “Adam was born in [MASK]”
- BERT-Large (last column) outperformed all models by a substantial margin

Results on Google-RE

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-1</i>	20006	23	23.85	-	5.4	33.8	6.1	18.0	3.6	6.5	32.4	34.2
	<i>N-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

- More facts and relations than Google-RE
- BERT-Large performed better on 1-to-1 relations, i.e. “capital-of”
- N-1: Multiple valid subjects-relations-> 1 correct object
- N-M relations: multiple objects for a subject-relation pair. i.e. “Brian owns [car, laptop, iPhone,etc]”

Results on ConceptNet and SQuAD

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

ConceptNet

- BERT-Large achieved best performance for ConceptNet
 - Able to retrieve commonsense knowledge at a similar level to factual knowledge

SQuAD

- Open domain cloze-style (fill in the blanks)
- Huge performance gap between BERT-Large and supervised DrQA
- Note: BERT and ELMo were both unsupervised and not fine-tuned for this task
- In terms of P@10 (Top-10 best answers), gap is remarkably small (57.1 for BI and 63.5 for DrQA)

Results on ConceptNet and SQuAD

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

ConceptNet

- BERT-Large achieved best performance for ConceptNet
 - Able to retrieve commonsense knowledge at a similar level to factual knowledge

SQuAD

- Open domain cloze-style (fill in the blanks)
- Huge performance gap between BERT-Large and supervised DrQA
- Note: BERT and ELMo were both unsupervised and not fine-tuned for this task
- In terms of P@10 (Top-10 best answers), gap is remarkably small (57.1 for BI and 63.5 for DrQA)

Other Results

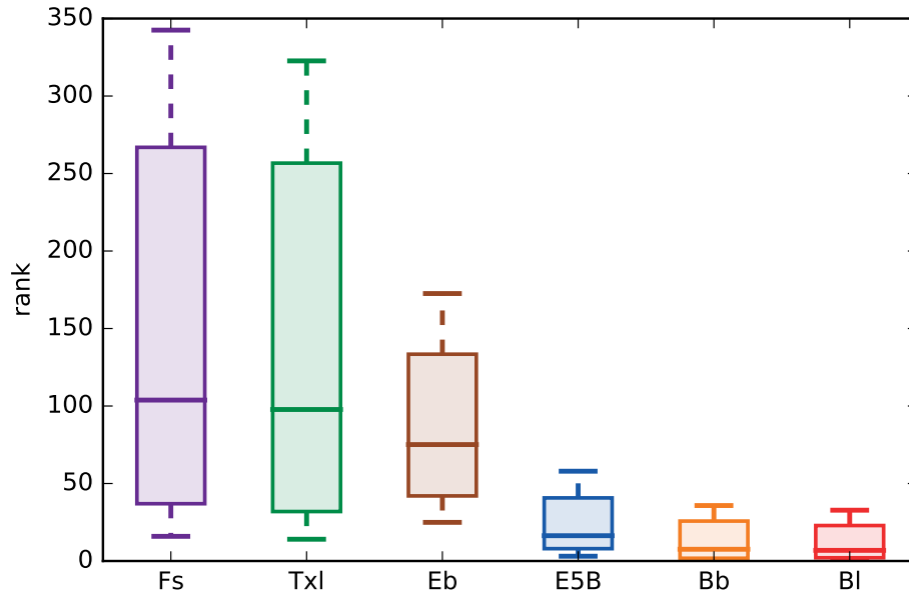


Figure 4: Average rank distribution for 10 different mentions of 100 random facts per relation in T-REx. ELMo 5.5B and both variants of BERT are least sensitive to the framing of the query but also are the most likely to have seen the query sentence during training.

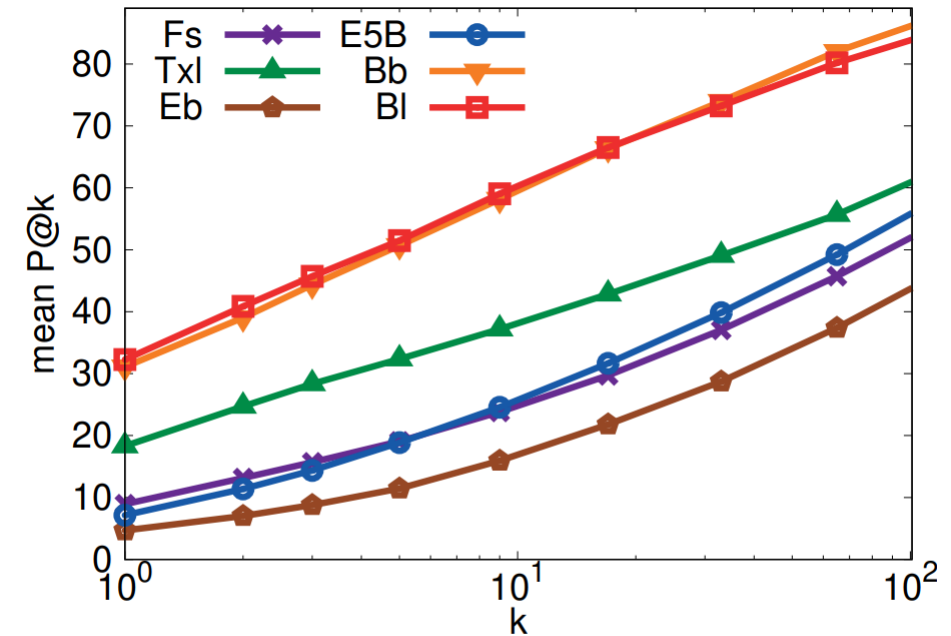


Figure 2: Mean P@k curve for T-REx varying k. Base-10 log scale for X axis.

Conclusion

- For an unsupervised, not fine-tuned, pre-trained model BERT-Large, it is possible to recall knowledge better than its competitors.
- The factual knowledge can be recovered surprisingly well from pretrained language models, however, for some relations (particularly N-to-M relations) performance is very poor.
- Language models trained on ever-growing corpora might become a viable alternative to traditional knowledge bases extracted from text in the future.

Limitations

- Only used Single-Token objects as prediction targets
- Chose only query objects in triples
- Still spent time manually defining templates for each relation

Thank you!