

Language Model as Knowledge Bases?

Fabio Petroni, Tim Rocktaschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu,
Alexander H. Miller, Sebastian Riedel

Facebook AI Research
University College London

EMNLP 2019

Motivation

- Does pretrained deep language models store structured relational information?

Methodology

- Create cloze test using pre-defined template for each relation
- Create sentences from randomly sampled triples
- Mask out objects mention. Let model predict the result

A was born in ____

Methodology

- Restricted to **single token** mention only
- Need extra decoder for predicting multiple words
 - Additional tuneable parameters

Recap: Language Model

- Given a sequence of tokens, predict next token
- Given a sequence of tokens, assign a probability to the sequence

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

Recap: Language Model

- Deep language model
 - HUGE corpus, unsupervised training, large parameter size
- ELMo
 - BiLSTM
- GPT
 - Transformer decoder
- BERT
 - Transformer encoder
 - **Pretraining task:** Masking, next sentence prediction

Recap: Knowledge Base

- Structured information
- Usually stored in the form of triples

Subject -- Relation -- Object

Dataset

- Google-RE
 - Manually extracted facts from Wikipedia
- T-Rex
 - Automatically aligned Wikidata triples with Wikipedia
- ConceptNet
 - OMCS
- SQuAD

Baseline

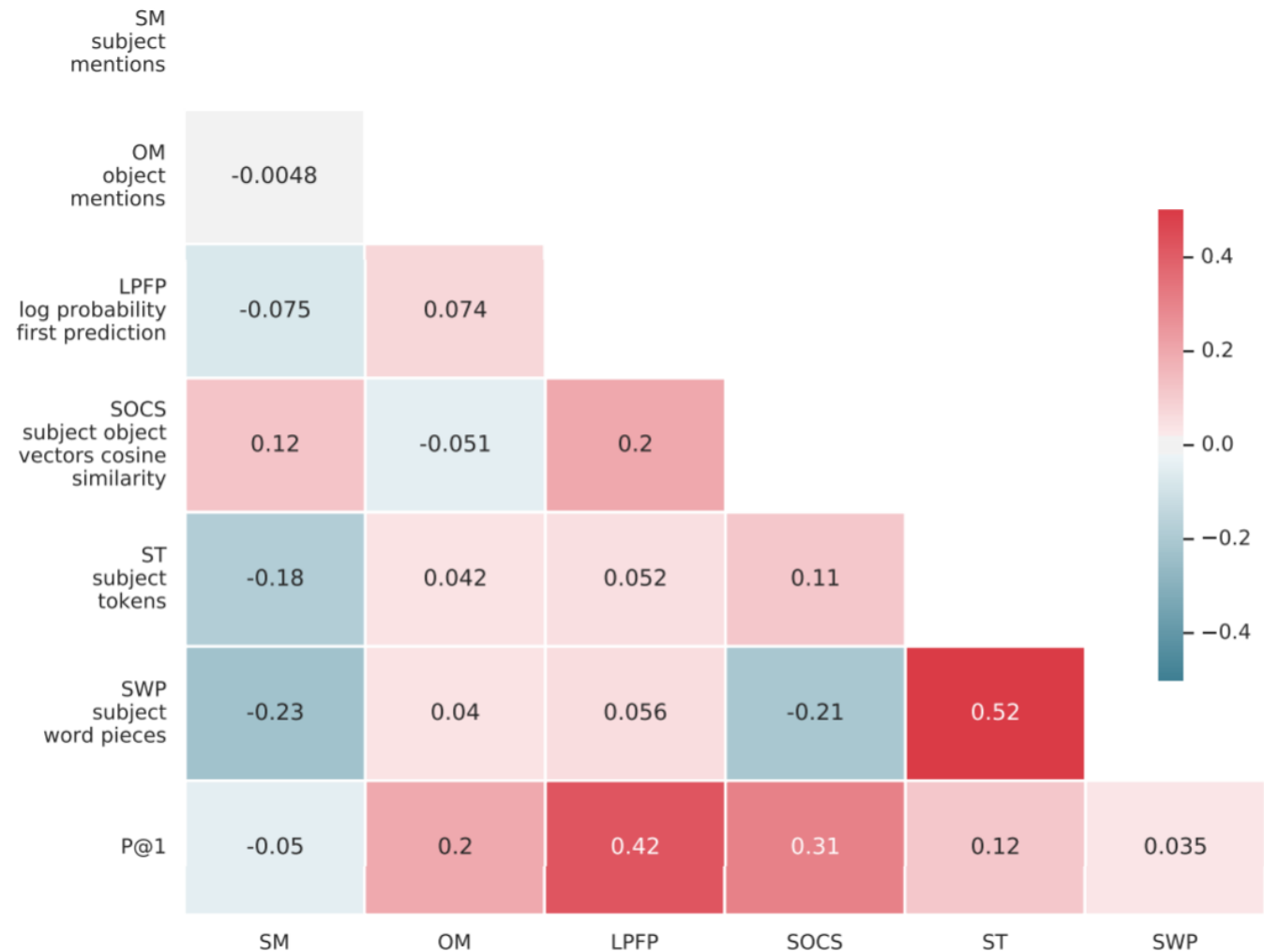
- Freq:
 - Rank word based on frequency
- RE:
 - Automatically extract KB first
 - Rank all objects linked to correct relation
- DrQA
 - TF/IDF retrieval for relevant articles
 - Extract answers from topK articles using neural models

	Relation	Query	Answer	Generation
T-Rex	P19	Francesco Bartolomeo Conti was born in ____.	Florence	Rome [-1.8], Florence [-1.8], Naples [-1.9], Milan [-2.4], Bologna [-2.5]
	P20	Adolphe Adam died in ____.	Paris	Paris [-0.5], London [-3.5], Vienna [-3.6], Berlin [-3.8], Brussels [-4.0]
	P279	English bulldog is a subclass of ____.	dog	dogs [-0.3], breeds [-2.2], dog [-2.4], cattle [-4.3], sheep [-4.5]
	P37	The official language of Mauritius is ____.	English	English [-0.6], French [-0.9], Arabic [-6.2], Tamil [-6.7], Malayalam [-7.0]
	P413	Patrick Oboya plays in ____ position.	midfielder	centre [-2.0], center [-2.2], midfielder [-2.4], forward [-2.4], midfield [-2.7]
	P138	Hamburg Airport is named after ____.	Hamburg	Hess [-7.0], Hermann [-7.1], Schmidt [-7.1], Hamburg [-7.5], Ludwig [-7.5]
	P364	The original language of Mon oncle Benjamin is ____.	French	French [-0.2], Breton [-3.3], English [-3.8], Dutch [-4.2], German [-4.9]
	P54	Dani Alves plays with ____.	Barcelona	Santos [-2.4], Porto [-2.5], Sporting [-3.1], Brazil [-3.3], Portugal [-3.7]
	P106	Paul Toungui is a ____ by profession .	politician	lawyer [-1.1], journalist [-2.4], teacher [-2.7], doctor [-3.0], physician [-3.7]
	P527	Sodium sulfide consists of ____.	sodium	water [-1.2], sulfur [-1.7], sodium [-2.5], zinc [-2.8], salt [-2.9]
	P102	Gordon Scholes is a member of the ____ political party.	Labor	Labour [-1.3], Conservative [-1.6], Green [-2.4], Liberal [-2.9], Labor [-2.9]
	P530	Kenya maintains diplomatic relations with ____.	Uganda	India [-3.0], Uganda [-3.2], Tanzania [-3.5], China [-3.6], Pakistan [-3.6]
	P176	iPod Touch is produced by ____.	Apple	Apple [-1.6], Nokia [-1.7], Sony [-2.0], Samsung [-2.6], Intel [-3.1]
	P30	Bailey Peninsula is located in ____.	Antarctica	Antarctica [-1.4], Bermuda [-2.2], Newfoundland [-2.5], Alaska [-2.7], Canada [-3.1]
	P178	JDK is developed by ____.	Oracle	IBM [-2.0], Intel [-2.3], Microsoft [-2.5], HP [-3.4], Nokia [-3.5]
	P1412	Carl III used to communicate in ____.	Swedish	German [-1.6], Latin [-1.9], French [-2.4], English [-3.0], Spanish [-3.0]
	P17	Sunshine Coast, British Columbia is located in ____.	Canada	Canada [-1.2], Alberta [-2.8], Yukon [-2.9], Labrador [-3.4], Victoria [-3.4]
	P39	Pope Clement VII has the position of ____.	pope	cardinal [-2.4], Pope [-2.5], pope [-2.6], President [-3.1], Chancellor [-3.2]
	P264	Joe Cocker is represented by music label ____.	Capitol	EMI [-2.6], BMG [-2.6], Universal [-2.8], Capitol [-3.2], Columbia [-3.3]
	P276	London Jazz Festival is located in ____.	London	London [-0.3], Greenwich [-3.2], Chelsea [-4.0], Camden [-4.6], Stratford [-4.8]
	P127	Border TV is owned by ____.	ITV	Sky [-3.1], ITV [-3.3], Global [-3.4], Frontier [-4.1], Disney [-4.3]
	P103	The native language of Mammootty is ____.	Malayalam	Malayalam [-0.2], Tamil [-2.1], Telugu [-4.8], English [-5.2], Hindi [-5.6]
	P495	The Sharon Cuneta Show was created in ____.	Philippines	Manila [-3.2], Philippines [-3.6], February [-3.7], December [-3.8], Argentina [-4.0]
ConceptNet	AtLocation	You are likely to find a overflow in a ____.	drain	sewer [-3.1], canal [-3.2], toilet [-3.3], stream [-3.6], drain [-3.6]
	CapableOf	Ravens can ____.	fly	fly [-1.5], fight [-1.8], kill [-2.2], die [-3.2], hunt [-3.4]
	CausesDesire	Joke would make you want to ____.	laugh	cry [-1.7], die [-1.7], laugh [-2.0], vomit [-2.6], scream [-2.6]
	Causes	Sometimes virus causes ____.	infection	disease [-1.2], cancer [-2.0], infection [-2.6], plague [-3.3], fever [-3.4]
	HasA	Birds have ____.	feathers	wings [-1.8], nests [-3.1], feathers [-3.2], died [-3.7], eggs [-3.9]
	HasPrerequisite	Typing requires ____.	speed	patience [-3.5], precision [-3.6], registration [-3.8], accuracy [-4.0], speed [-4.1]
	HasProperty	Time is ____.	finite	short [-1.7], passing [-1.8], precious [-2.9], irrelevant [-3.2], gone [-4.0]
	MotivatedByGoal	You would celebrate because you are ____.	alive	happy [-2.4], human [-3.3], alive [-3.3], young [-3.6], free [-3.9]
	ReceivesAction	Skills can be ____.	taught	acquired [-2.5], useful [-2.5], learned [-2.8], combined [-3.9], varied [-3.9]
	UsedFor	A pond is for ____.	fish	swimming [-1.3], fishing [-1.4], bathing [-2.0], fish [-2.8], recreation [-3.1]

Table 3: Examples of generation for BERT-large. The last column reports the top five tokens generated together with the associated log probability (in square brackets).

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

- Pearson correlation
- The number of times an object in training data is positively correlates with performance
 - Same not true for subject
- Log probability of prediction is strongly correlates to P@1
 - If BERT has a high confidence, the predictions are usually correct



Pearson correlation coefficient for the P@1 of the BERT-large model on T-REx and a set of metrics: SM and OM refer to the number of times a subject and an object are mentioned in the BERT training corpus⁴ respectively; LPFP is the log probability score associated with the first prediction; SOCS is the cosine similarity between subject and object vectors (we use spaCy⁵); ST and SWP are the number of tokens in the subject with a standard tokenization and the BERT WordPiece tokenization respectively.

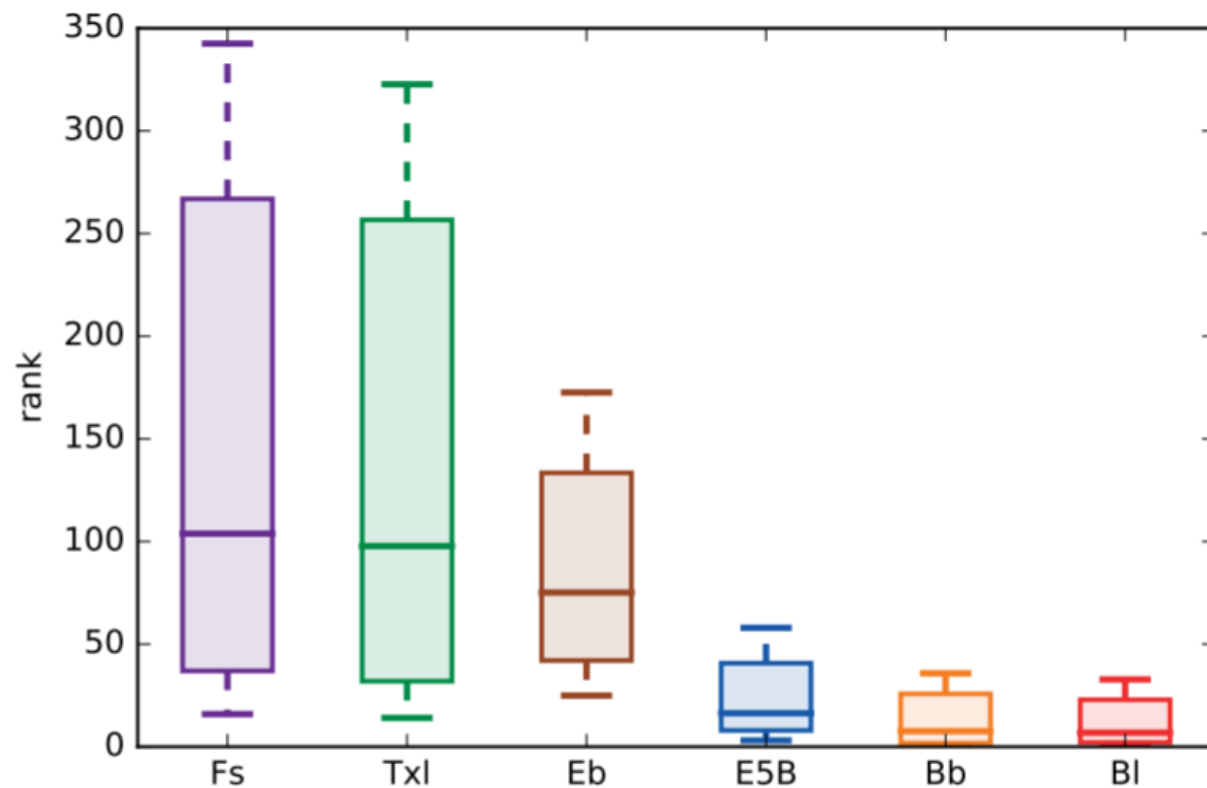


Figure 4: Average rank distribution for 10 different mentions of 100 random facts per relation in T-REx. ELMo 5.5B and both variants of BERT are least sensitive to the framing of the query but also are the most likely to have seen the query sentence during training.

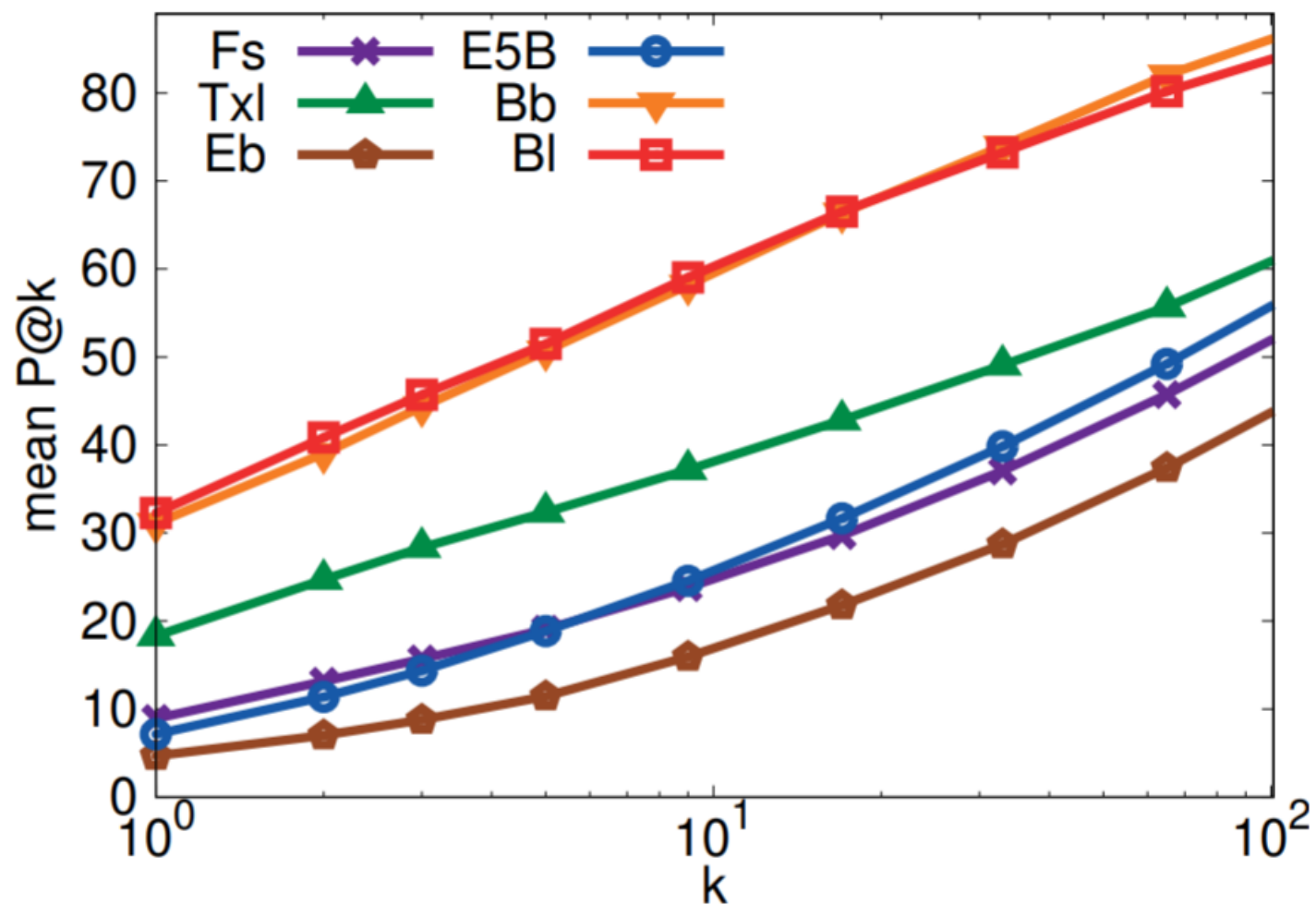


Figure 2: Mean P@k curve for T-REx varying k. Base-10 log scale for X axis.