

Question: Can LM's be considered as KB's?[1]

Not There, Not Yet![2]

Pawan Kumar

Feb 5, 2020

CSE, IIT Delhi

Table of contents

1. LMs and KBs
2. LAMA Bench Mark
3. Experiments
4. Piazza Discussion

Intro

KB's: Knowledge Bases/Knowledge Graphs

- Facts are stored symbolically as triples (S, P, O).
- Effective way to store and access annotated gold-standard relational data.
- *Pipelined-KB's*

LM's: Language Models

- (May be) Facts are in a latent space (Model Weights)

Pipeline

Entity Extraction -> Co-reference Resolution -> Entity Linking ->
Relation Extraction

Error Propagation and Accumulation

- Learn language lexicon/Surface forms.
- Learn fallible Syntactic heuristics.
- (May be) learn Semantics (factual and Commonsense Knowledge).

Experimental Setup

Knowledge Sources e.g. Wikidata, ConceptNet.

Question-schema

(S,P,O): (DANTE, born-in, FLORENCE)

Cloze form: DANTE is born-in *[MASKED]*

Experiments

Experimental Setup-I

Uni-directional LM's:

fairseq-fconv: WikiText-103

Transformer-XL: WikiText-103

Bi-Directional LM's:

ELMo: Wikipedia, News crawl,

BERT: BookCorpus, English Wikipedia

KBs: off-the-shelf relation extractor and Oracle-based entity linker.

Freq : freq of X, based on how many (S, P, X)-triples exist.

RE : KG created with a pre-trained Relation Extraction(RE)

DrQA : open-domain QA systems which uses TF/IDF for ranking articles and then a neural reading comprehension model to extract answers.

Experimental Setup-II

LAMA-probe Questions are created from following Knowledge Sources:

Google-RE: uses only 4 relations.

T-REx: uses 41 relations, 1000 facts each relations.

ConceptNet: uses 16 relations.

SQuAD: use 305 context-insensitive questions

Assumptions:

1. a Common Vocabulary of 21K case-sensitive tokens.
2. Manually Defined Templates for close form questions
3. Single Token Answers only
4. Query only Object Slots

Results-I

Mean Precision at One

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

Figure 1: 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.

Results-II

Mean Precision at k

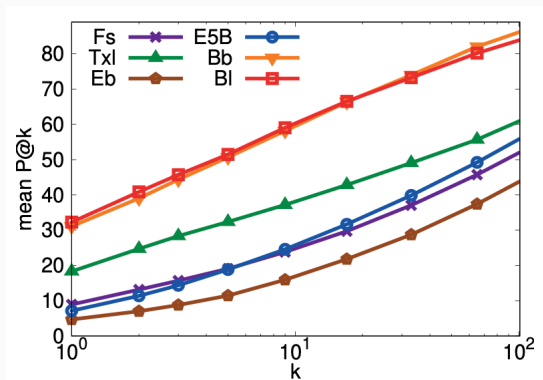


Figure 2: Mean P@k curve for T-REx varying k. Base-10 log scale for X axis.

Analysis

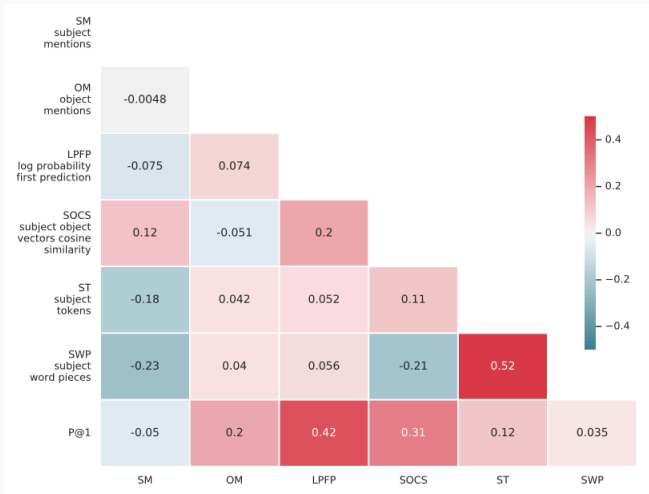


Figure 3: Pearson coeff for P@1 of the BERT-large on T-REx (SM: Subj. Mentions, OM: Obj. Mention in Training; LPFP:log prob. of first pred; SOCS: cosine sim. of Subj. and Obj. ; ST: standard tokenization, SWP: WordPiece.)

Piazza Discussion

Piazza Discussion For the Motion

Comments:

- "Language models are unsupervised multitask learners"(Radford et al., 2019)
- "knowledge is learned in a more robust manner"

Critical Comments:

Vipul: Multi-Token can be done with Span-predictions

Vipul: A combo of KB + GNN may be a viable model for most NLP Tasks.

Rajas: knowledge is learned in a more robust manner, e.g. born-in and birthed-in are similar.

Comments:

- Mere a Happy co-incidence, It's too Early to pass a verdict.

Critical Comments:

Vaibhav: 1. Temporal Facts, 2. Complex Query,

Shubham: 1. Don't store actual facts, not precise.

Atishya: 1. KBs have a lot of in-built reasoning into their structure,

Piazza Discussion On LM's Extensions

Siddhant: Make a Knowledge Graph with help from a pre-trained LM.

Sankalan: Relation Extraction by fine-tuning a pre-trained model on may be a T5 task.


Jigyasa, Rajas: Compositional queries (Obama -> mother -> birth place) may be formulated as "Obama's mother, ... was born in ...". i.e. A multi-hop is turned into Multiple single hops.

Atishya, Soumya: Require work on metric for spans. Answer Category can be checked.

Soumya: Answer Category can be checked for type. Augmenting text with entity-type and together creating a new concept-embedding space.

Pratyush: When preparing for a QA Task, use out-of-domain fine-tuning.

Thanks.

 F. Petroni, T. Rocktäschel, P. Lewis, A. Bakhtin, Y. Wu, A. H. Miller, and S. Riedel.

Language models as knowledge bases?

arXiv preprint arXiv:1909.01066, 2019.

 N. Poerner, U. Waltinger, and H. Schütze.

Bert is not a knowledge base (yet): Factual knowledge vs. name-based reasoning in unsupervised qa.

arXiv preprint arXiv:1911.03681, 2019.

LAMA-UHN

The result with benchmark LAMA Name Un-helpful is used