Implicit Knowledge Base: Pre-trained Language Models

Tianxiang Sun
Fudan University
txsun19@fudan.edu.cn

December 15, 2019

Outline

1. Background
2. Related Topics
3. Language Models as Knowledge Bases?
4. How Can We Know What Language Models Know?
5. Language Models are Not Knowledge Bases (Yet)
6. Conclusion

Background

Language Models

Unidirectional Language Models (e.g., GPT)

$$P(x_t|x_{<t}, \theta).$$

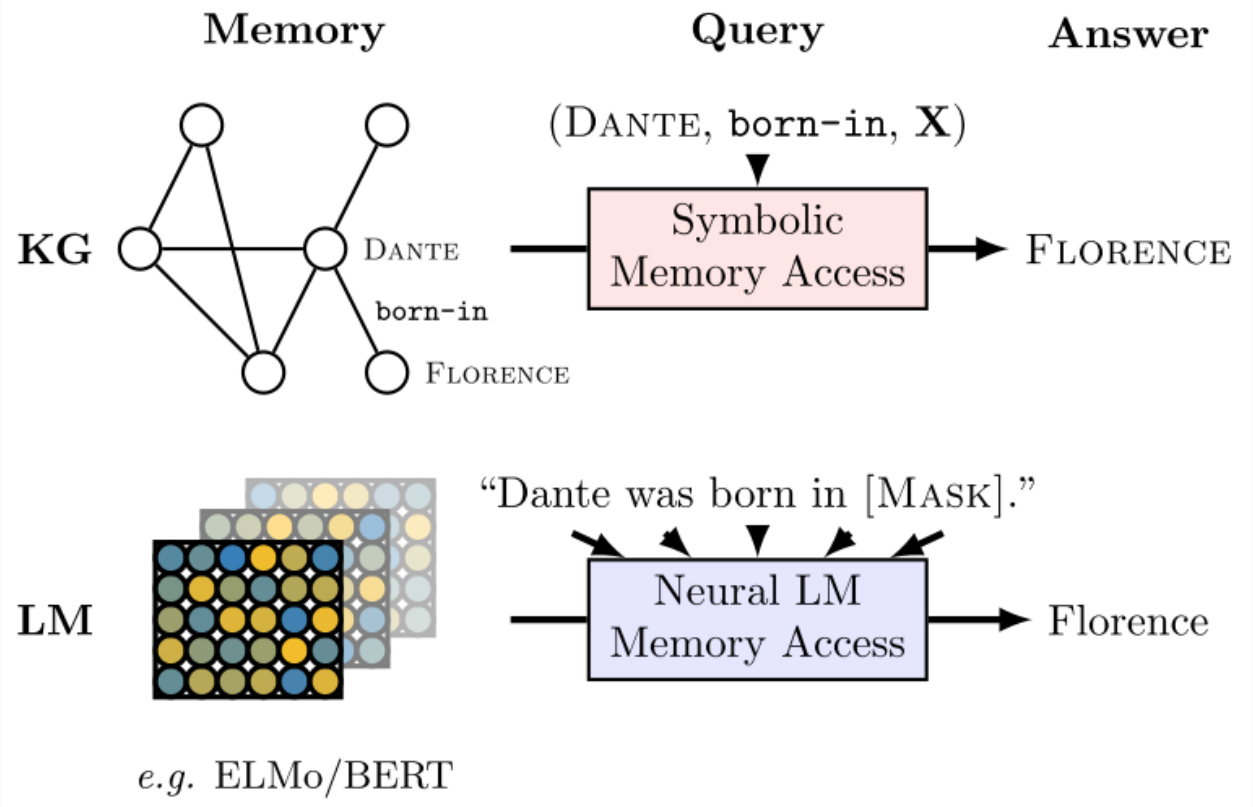
Bidirectional Language Models (e.g., ELMo, BERT)

$$P(x_t|x_{\setminus t}, \theta).$$

Pre-trained language models have led to a surge of improvements for downstream NLP tasks, which suggests vast amount of knowledge (linguistic, semantic, fact, commonsense, etc.) may be stored in language models.

Language Models as Knowledge Bases?

We can query knowledge bases and language models for factual knowledge like this [11]:



Related Topics

Probing Knowledge in Representation

Non-contextual representation (e.g., word2vec, GloVe, etc.)

- Anaologies [9]
- Properties / attributes [4, 13]

Contextualized representation (e.g., ELMo, GPT, BERT, etc.)

- Linguistic knowledge [1, 8, 14]
- **Factual knowledge** [2, 3, 7, 11]

Knowledge-aware Language Models (LMs)

Entity-enhanced LMs

- ERNIE [15]
- KnowBERT [10]

KG conditioned LMs

- KGLM [6]
- Latent relation LM [5]

Language Models as Knowledge Bases?

LAnguage Model Analysis (LAMA)

LAMA probe is to test factual and commonsense knowledge in language models via "fill-in-the-blank" cloze statement.

```
Dante was born in [MASK]
```

The authors manually creat single-token cloze statements from the following knowledge sources:

- Google-RE (from Wikipedia, 3 relations, 5527 facts)
- T-REx (from Wikidata, 41 relations, 34039 facts)
- ConceptNet (16 relation, 11458 facts, commonsense)
- SQuAD (305 facts)

Results

Withoud fine-tuning, BERT contains relational knowledge competitive with traditional IE methods that have some access to oracle knowledge. [11]

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.

Examples Generated by BERT-Large

	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]

How Can We Know What Language Models Know?

Motivation

LAMA is measuring a *lower* bound for what language models know, and there may be better ways to query language models.

`DirectX` is developed by `[MASK]`
`DirectX` is created by `[MASK]`
`[MASK]` released the `DirectX`

A related topic in information retrieval: Automatic Query Expansion.

Can we generate more efficient queries for language models?

Query Generation

Our task is to generate a set of queries $\{q_{r,i}\}_{i=1}^{T}$ for each relation r, where at least some of the queries effectively trigger LMs to predict ground-truth objects.

- **Mining-based Generation**
  First find a set of subject-object pairs, (x, y), for each relation r in Wikipedia, then use the top T most frequent words in the middle of x and y to construct queries for r.
  Barack Obama was born in Hawaii $\rightarrow$ *x was born in y*.
- **Paraphrasing-based Generation**
  Back translation: $1 \xrightarrow{forward} \sqrt{T} \xrightarrow{backward} T$.

Query Selection and Ensembling

- **Top-1 Selection.** Select the query $q_{r,i}$ with the highest accuracy

$$A(q_{r,i}) = \frac{\sum_{<x,y>\in r} \delta(y = \arg\max_{y'} P_{LM}(y'|x, q_{r,i}))}{|r|}.$$

- **Rank-based Ensemble.** Use the average log probabilities from the top K queries

$$s(y|x, r) = \frac{1}{K} \sum_{i=1}^{K} \log P(y|x, q_{r,i}),$$

$$P(y|x, r) = \mathrm{Softmax}(s(\cdot|x, r))_y.$$

- **Optimized Ensemble.**

$$s(y|x, r) = \sum_{i=1}^{T} P_{\theta_r}(q_{r,i}|r) \log P(y|x, q_{r,i}).$$

Note that selection and ensembling are done on a held-out train set newly created by the authors.

Experiments on T-REx

Mine: Mining-based generation. **Man**: Manually generation.
Para: Paraphrasing-based generation. **Opti.**: Optimized Ensemble.

Prompts	Top1	Top3	Top5	Opti.	Oracle
*BERT-base (**Man**=31.1)*					
Mine	30.7	32.7	31.2	36.9	45.1
Mine+Man	31.9	34.5	33.8	**38.1**	47.9
Mine+Para	30.7	33.0	33.7	33.6	45.0
Man+Para	*34.1*	35.8	36.6	37.3	47.9
*BERT-large (**Man**=32.3)*					
Mine	34.4	33.8	33.1	40.4	47.9
Mine+Man	*36.0*	38.6	37.1	**41.9**	50.8
Mine+Para	32.1	35.0	36.1	37.0	47.3
Man+Para	35.9	37.3	38.0	38.8	50.0

(a) Micro-averaged accuracy

Prompts	Top1	Top3	Top5	Opti.	Oracle
*BERT-base (**Man**=22.8)*					
Mine	21.2	22.1	21.4	24.0	32.2
Mine+Man	22.0	24.0	23.4	**25.2**	34.6
Mine+Para	20.2	22.1	22.6	22.6	32.2
Man+Para	*22.8*	23.8	24.6	25.0	34.9
*BERT-large (**Man**=25.7)*					
Mine	24.8	25.0	24.1	27.7	36.4
Mine+Man	*27.0*	27.6	26.8	**29.5**	38.9
Mine+Para	23.4	24.8	25.7	25.8	36.2
Man+Para	25.9	27.8	28.3	28.0	39.3

(b) Macro-averaged accuracy

Figure 1: Main results.

Some Interesting Findings

ID	Relations	Manual Prompts	Mined Prompts	Acc. Gain
P140	religion	x is affiliated with the y religion	x who converted to y	+60.0
P159	headquarters location	The headquarter of x is in y	x is based in y	+4.9
P20	place of death	x died in y	x died at his home in y	+4.6
P264	record label	x is represented by music label y	x recorded for y	+17.2
P279	subclass of	x is a subclass of y	x is a type of y	+22.7
P39	position held	x has the position of y	x is elected y	+7.9

Figure 2: Micro-accuracy gain of the mined prompts over the manual prompts.

ID	Modifications	Acc. Gain
P413	x plays in→at y position	+23.2
P495	x was created→made in y	+10.8
P495	x was→is created in y	+10.0
P361	x is a part of y	+2.7
P413	x plays ~~in~~ y position	+2.2

Figure 3: Small modifications (update, insert, and delete) in paraphrase lead to large accuracy gain.

Language Models are Not Knowledge Bases (Yet)

Factual Knowledge vs. Name-Based Reasoning

BERT could cheat: the impressive performance of BERT is partly due to reasoning about (the surface form of) entity names. Take the relation `native_language` as an examples, we query BERT by "The native language of [X] is [MASK]" and would get results:

[X]	BERT-base	Anwser
Jean Marais	French	French
Daniel Ceccaldi	Italian	French
Orane Demazis	Albanian	French
Kad Merad	Kurdish	French

It is often possible to guess properties of an entity from its name, with zero factual knowledge of the entity itself.

So is LM good at reasoning about names, good at memorizing facts, or both?

LAMA-UHN

Pörner et al. [12] add 2 filters to creat LAMA-UHN (UnHelpfuleNames), a subset of LAMA-Google-RE and LAMA-T-REx.

- **Filter 1: string match filte.** Deletes all KB triples where the correct answer is a case-insensitive substring of the subject entity name. For instance,

  `[IBM AIX]` is developed by `[IBM]`.

- **Filter 2: person name filte.** Uses cloze-style questions to elicit name associations inherent in BERT, and deletes KB triples that correlate with them. For instance,

  `[Harumi Inoue]` is a `[Japan]` citizen.

BERT's precision drops dramatically when we filter these certain easy-to-guess facts.

E-BERT: An Extension of BERT

BERT. Vocabulary: $\mathbb{L}_b$. Embedding function: $\mathcal{E}_\mathcal{B} : \mathbb{L}_b \rightarrow \mathbb{R}^{d_\mathcal{B}}$.

Wikipedia2vec. Wikipedia2vec embeds words and entities into a common space. Embedding function: $\mathcal{F} : \mathbb{L}_w \cup \mathbb{L}_e \rightarrow \mathbb{R}^{d_\mathcal{F}}$.

E-BERT. E-BERT aims to transform the output space of $\mathcal{F}$ in such a way that $\mathcal{B}$ is fooled into accepting entity embeddings in lieu of its native subword embeddings. To do this, E-BERT employs a linear projection $\mathcal{W}$ to minimize the squared distance of transformed wikipedia2vec word vectors and BERT subword vectors:

$$\arg\min_\mathcal{W} \mathbb{E}_{x \in \mathbb{L}_b \cap \mathbb{L}_w} \|\mathcal{W}(\mathcal{F}(x)) - \mathcal{E}_\mathcal{B}(x)\|_2^2.$$

Ensemble BERT and E-BERT. (a) AVG (b) CONCAT

Experiments

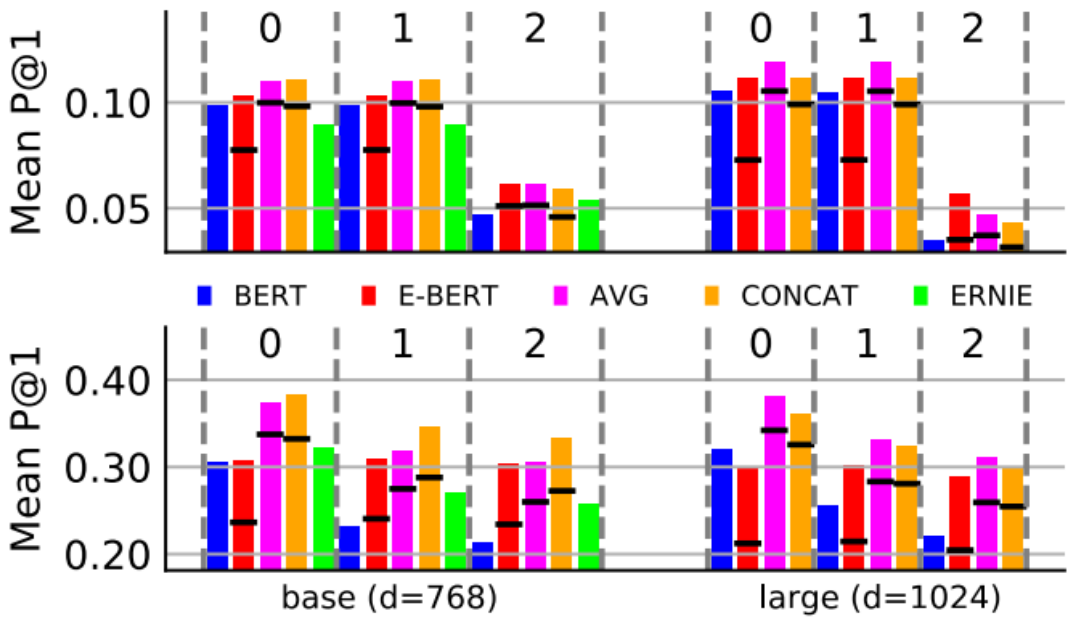


Figure 1: Mean P@1 on LAMA-Google-RE (top) and LAMA-T-REx (bottom). 0: unfiltered, 1: string match filter, 2: person name filter. Filters are applied sequentially. Black horizontal bars: Performance of wikipedia2vec without link graph loss.

Conclusion

Problems and Limitations

- How can we know what LMs know?
- How to construct better query prompts?
- Single token prediction.
- LAMA only query objects, as opposed to subjects and relations. It is trivial to query subjects, but not relations.
- How to incorporate factual and commonsense knowledge into LMs?

Conclusion

NLP tasks are not independent: solving some tasks, such as summarization, open-domain QA, data-to-text, needs external knowledge.

Knowledge Graph (explicit, symbolism)

- Hard to construct
- Hard to use: pipline (entity linking, relation extraction, etc.) brings complexity and error propagation
- Limited relations and units (triplet)
- Interpretability (human enjoys graphs)

Language Model (implicit, connectionism)

- Unexplainable
- Free of schema engineering and human annotations
- Open set of queries

Thanks for Listening!

References

[1] Anonymous. Are pre-trained language models aware of phrases? simple but strong baselines for grammar induction. In *Submitted to International Conference on Learning Representations*, 2020. under review.

[2] Zied Bouraoui, Jose Camacho-Collados, and Steven Schockaert. Inducing relational knowledge from bert. *arXiv preprint arXiv:1911.12753*, 2019.

[3] Joe Davison, Joshua Feldman, and Alexander M. Rush. Commonsense knowledge mining from pretrained models. In *EMNLP*, 2019.

References ii

[4] Abhijeet Gupta, Gemma Boleda, Marco Baroni, and Sebastian Padó. Distributional vectors encode referential attributes. In *EMNLP*, 2015.

[5] Hiroaki Hayashi, Zecong Hu, Chenyan Xiong, and Graham Neubig. Latent relation language models. *arXiv preprint arXiv:1908.07690*, 2019.

[6] Robert L. Logan IV, Nelson F. Liu, Matthew E. Peters, Matt Gardner, and Sameer Singh. Barack's wife hillary: Using knowledge graphs for fact-aware language modeling. In *ACL*, 2019.

[7] Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. How can we know what language models know? *arXiv preprint arXiv:1911.12543*, 2019.

References iii

[8] Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. Linguistic knowledge and transferability of contextual representations. In *NAACL*, 2019.

[9] Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In *NAACL*, 2013.

[10] Matthew E. Peters, Mark Neumann, Robert L. Logan IV, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A. Smith. Knowledge enhanced contextual word representations. In *EMNLP*, 2019.

[11] Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander H. Miller. Language models as knowledge bases? In *EMNLP*, 2019.

References iv

[12] Nina Pörner, Ulli Waltinger, and Hinrich Schütze. BERT is not a knowledge base (yet): Factual knowledge vs. name-based reasoning in unsupervised QA. *arXiv preprint arXiv:1911.12753*, 2019.

[13] Dana Rubinstein, Effi Levi, Roy Schwartz, and Ari Rappoport. How well do distributional models capture different types of semantic knowledge? In *ACL*, 2015.

[14] Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R. Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R. Bowman, Dipanjan Das, and Ellie Pavlick. What do you learn from context? probing for sentence structure in contextualized word representations. In *ICLR*, 2019.

[15] Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. ERNIE: enhanced language representation with informative entities. In *ACL*, 2019.