Language Models as Knowledge Bases:

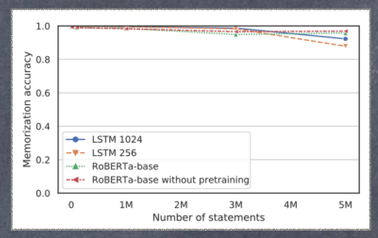
On Entity Representations, Storage Capacity, and Paraphrased Queries

Benjamin Heinzerling & Kentaro Inui (RIKEN AIP & Tohoku University)

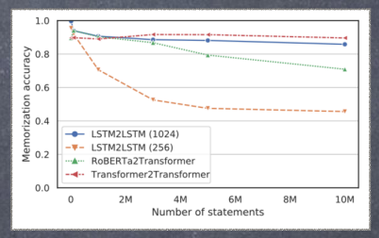
Takeaways

- LMs can produce text that looks like world knowledge -> "LM-as-KB"
- Compare entity representations: symbolic, surface form, embedding
- LMs can store millions of facts
- Pretrained LMs have ability for paraphrased queries, but it's brittle

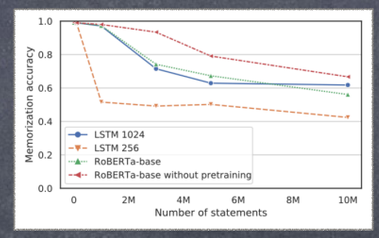
Fact Memorization with different entity representations



symbolic



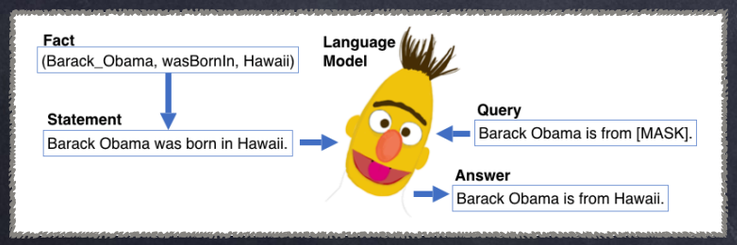
surface form



entity embedding

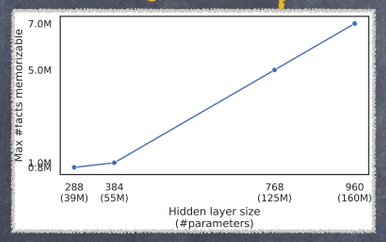
There is some world knowledge in LMs

Language model as knowledge base



- Actual BERT predictions:
- "Barack Obama was born in [MASK]." -> Chicago
- "Barack Obama was born on [MASK]." -> Earth
- "Barack Obama was born on the island of [MASK]." -> Hawaii

Storage capacity



- Setup: Store facts in LM by memorizing statements
- Generate statements from Wikidata: "S was born in O"
- Training objective: Overfit LM to statements
- Evaluation method: Predict relation object
- Evaluation metric: Accuracy of object prediction

LM pretraining enables paraphrased queries