

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

There is some world knowledge in LMs

Sentence: Tokyo is the capital of [MASK].

Mask 1 Predictions:

- 96.1% Japan
- 1.6% Asia

Sentence: Sendai is the capital of [MASK].

Mask 1 Predictions:

- 87.6% Japan
- 3.1% it
- 1.7% Asia
- 1.0% China
- 0.5% Tokyo

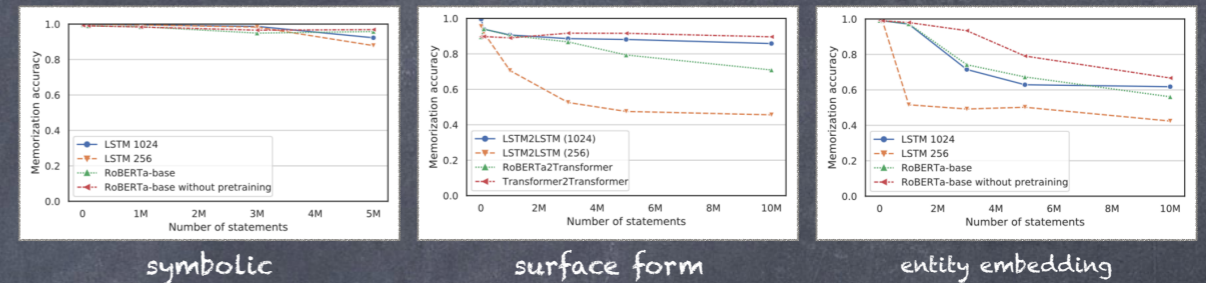
Sendai

From Wikipedia, the free encyclopedia

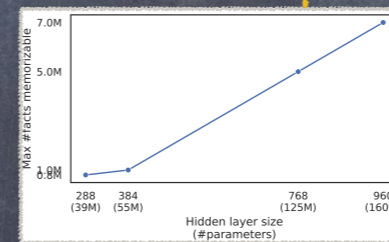
This article is about the capital city of Miyagi Prefecture. For other uses, see [Sendai \(disambiguation\)](#).

Sendai (Japanese: 仙台市, *Sendai-shi*; Japanese pronunciation: *[sɛ̃ːnai̯]*) is the capital city of Miyagi Prefecture, Japan, the largest

Fact Memorization with different **entity representations**

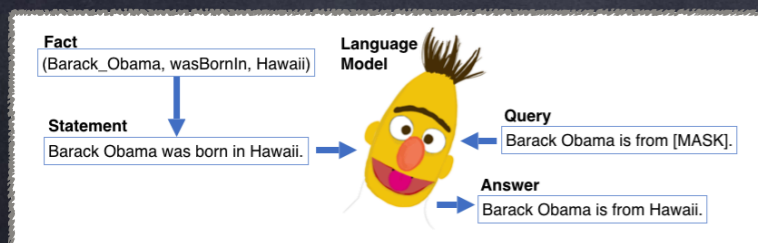


Storage capacity



- 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**

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**

LM pretraining enables **paraphrased queries**

