Language Models as Knowledge Bases?

Fabio Petroni, Tim Rocktaschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H. Miller, Sebastian Riedel
• Knowledge Bases (KB)
• Language Model and KB
• LAMA
• Models used
• Results
• Conclusion
Knowledge Bases
Knowledge Bases (KB)

- A KB is a technology to store information
- Effective solution for accessing annotated relational data
- It is possible to query them (Dante, born-in, X)

Disadvantages:
- It is difficult to populate KB
- Complex pipeline to populate KB automatically [1]
Language Model and KB
Language Model (LM)

• A model that represents the language domain

• Predict the next word in a sentence (e.g. "Dante was born in")
• Predict the masked word in a sentence (e.g. "Dante was born in [MASK] in 1265")
• Answer questions (e.g. "Where was Dante born?")
LM as KB

Similarities
- Contain knowledge
- Can be queried
- Can be updated / improved

Advantages
- No schema engineering
- No need for human annotations
- Open set of queries

Image from "Language Model as Knowledge Bases?" Petroni et al.
Authors' questions

How much relational knowledge do LM store?

How does this differ for different types of knowledge? (facts about entities, common sense, general question answering)

How does the performance of LM without fine-tuning compare to symbolic knowledge bases automatically extracted from the text?
LAMA
(Language Model Analysis)
LAMA probe

Test the factual and commonsense knowledge in LM

- Uses a set of knowledge sources (corpus of facts)
- Fact = (subject, relation, object) | (question, answer)
- Facts become cloze sentences used to query LM
- Evaluation: how highly LM ranks Ground Truth token
- P@k: 1 if the gold entity is in the top k results
- HYP: LM have more factual knowledge if they score high the Ground Truth
Knowledge Sources

Google-RE
- ~60K facts manually extracted from Wikipedia
- 3 relations used (place of birth, date of birth and place of death)
- template manually defined

T-Rex
- subset of Wikipedia triples derived from the T-Rex dataset [2]
- 41 relations
- 1000 facts per relations
- template manually defined

ConceptNet [3]
- multilingual KB
- commonsense relationship
- 16 English relationship
- object masked in the sentence

SQuAD
- question answer dataset
- 305 context insensitive questions with single token answers
- questions rewritten to cloze sentences
Baselines

**Freq**
- It ranks words on how frequently they appear as an object of a specific relation
- Predict the same object for each relation

**Relation Extraction (RE) [5]**
- LSTM model based on attention which extract triples
- Trained on Wikipedia subcorpus
- Create a Knowledge Graph
- \( RE_n \) = naive entity linking
- \( RE_o \) = oracle entity linking

**DrQA [6]**
- Open-domain question answering system
- First step: TF-IDF information retrieval
- Second step: neural model extracts answers
Models used
Unidirectional LM

**fairseq-conv (Fs) [7]**
- Multiple layers of gated convolution
- Pretrained on the Wikitex-103 corpus

**Transformers-XL (large Txl) [8]**
- Large-scale LM based on Transformer with no fixed input length
- Cache previous outputs
- Use relative position encoding

\[
p(w) = \prod_t p(w_t | w_{t-1}, \ldots, w_1).
\]
Bidirectional LM

ELMO (original Eb – 5.5B E5B) [9]
- Multi-layers BiLSTM

BERT (base Bb – large Bl) [10]
- Encoder module of a Transformers
- Pretraining: Masked LM – NSP

\[ p(w_i) = p(w_i | w_1, \ldots, w_{i-1}, w_{i+1}, \ldots, w_N) \]
Results
## Table with all results

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Relation</th>
<th>Statistics</th>
<th>Baselines</th>
<th>KB</th>
<th>LM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>#Facts</td>
<td>#Rel</td>
<td>Freq</td>
<td>DrQA</td>
</tr>
<tr>
<td>Google-RE</td>
<td>birth-place</td>
<td>2937</td>
<td>1</td>
<td>4.6</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>birth-date</td>
<td>1825</td>
<td>1</td>
<td>1.9</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>death-place</td>
<td>765</td>
<td>1</td>
<td>6.8</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>5527</td>
<td>3</td>
<td>4.4</td>
<td>-</td>
</tr>
<tr>
<td>T-REx</td>
<td>1-1</td>
<td>937</td>
<td>2</td>
<td>1.78</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>N-1</td>
<td>20006</td>
<td>23</td>
<td>23.85</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>N-M</td>
<td>13096</td>
<td>16</td>
<td>21.95</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>34039</td>
<td>41</td>
<td>22.03</td>
<td>-</td>
</tr>
<tr>
<td>ConceptNet</td>
<td>Total</td>
<td>11458</td>
<td>16</td>
<td>4.8</td>
<td>-</td>
</tr>
<tr>
<td>SQuAD</td>
<td>Total</td>
<td>305</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 2: Mean precision at one (P@1) for a frequency baseline (Freq), DrQA, a relation extraction with naive entity linking (RE\textsubscript{n}), oracle entity linking (RE\textsubscript{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.

Table from "Language Model as Knowledge Bases?" Petroni et al.
Additional takeaways

T-REX

- Object Mentions correlated with P@1
- Log probability correlated with P@1
- Cosine similarity SO correlated with P@1

Chart from "Language Model as Knowledge Bases?" Petroni et al.
## Additional takeaways

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Query</th>
<th>Answer</th>
<th>Generation</th>
</tr>
</thead>
<tbody>
<tr>
<td>T-Rex</td>
<td>Dani Alves plays with ____ .</td>
<td>Barcelona</td>
<td>Santos, Porto, Sporting, Brazil, Portugal</td>
</tr>
<tr>
<td>ConceptNet</td>
<td>Time is ____ .</td>
<td>finite</td>
<td>short, passing, precious, irrelevant, gone</td>
</tr>
</tbody>
</table>
Conclusion
Conclusion

• Systematic analysis of the factual and commonsense knowledge in publicly available pre-trained LM as is (LAMA probe)
• BERT large recall object of relationship consistently better than similar models
• BERT large is also competitive with other methods, which use oracles
• KB–RE models had not a significant improvement with an additional dataset
• Bigger corpus has an impact on the performance of BERT
• It will be easier to improve the performance of BERT rather than RE models
Questions?
References


