Gabriele Dominici
gd489@cam.ac.uk

Language Models as Knowledge Bases?

Fabio Petroni, Tim Rocktaschel, Patrick Lewis, Anton Bakhtin,
Yuxiang Wu, Alexander H. Miller, Sebastian Riedel

Pitch

Agenda

- 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

Pitch

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?")

Pitch

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

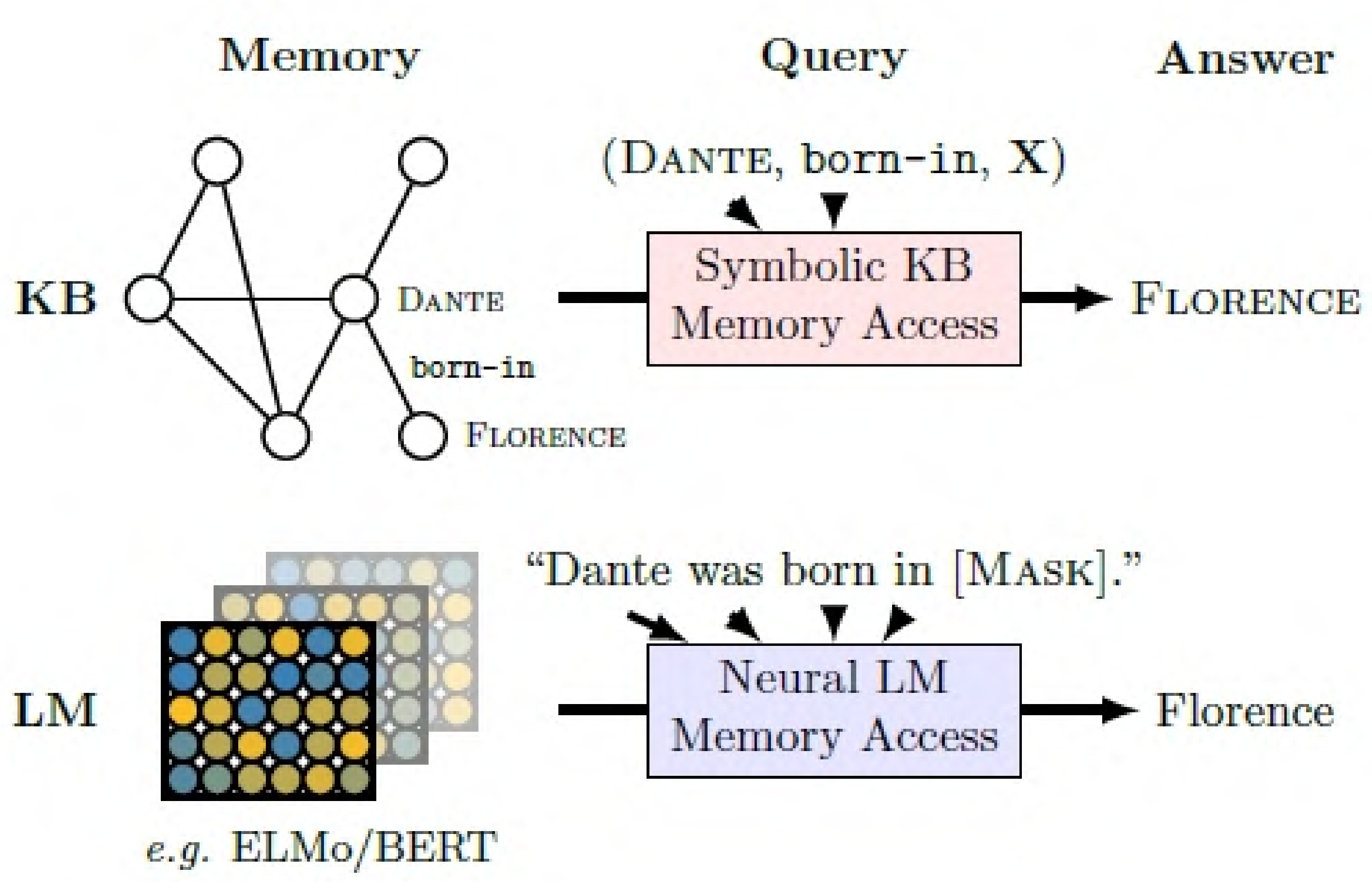

Image from "Language Model as Knowledge Bases?" Petroni et al.

Pitch

Authors' questions

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

How much relational
knowledge do LM store?

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

$$p(\mathbf{w}) = \prod_t p(w_t \mid w_{t-1}, \ldots, w_1).$$

Pitch

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

$$p(w_i) = p(w_i \mid w_1, \ldots, w_{i-1}, w_{i+1}, \ldots, w_N)$$

Pitch

Results

Table with all results

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.

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

Additional takeaways

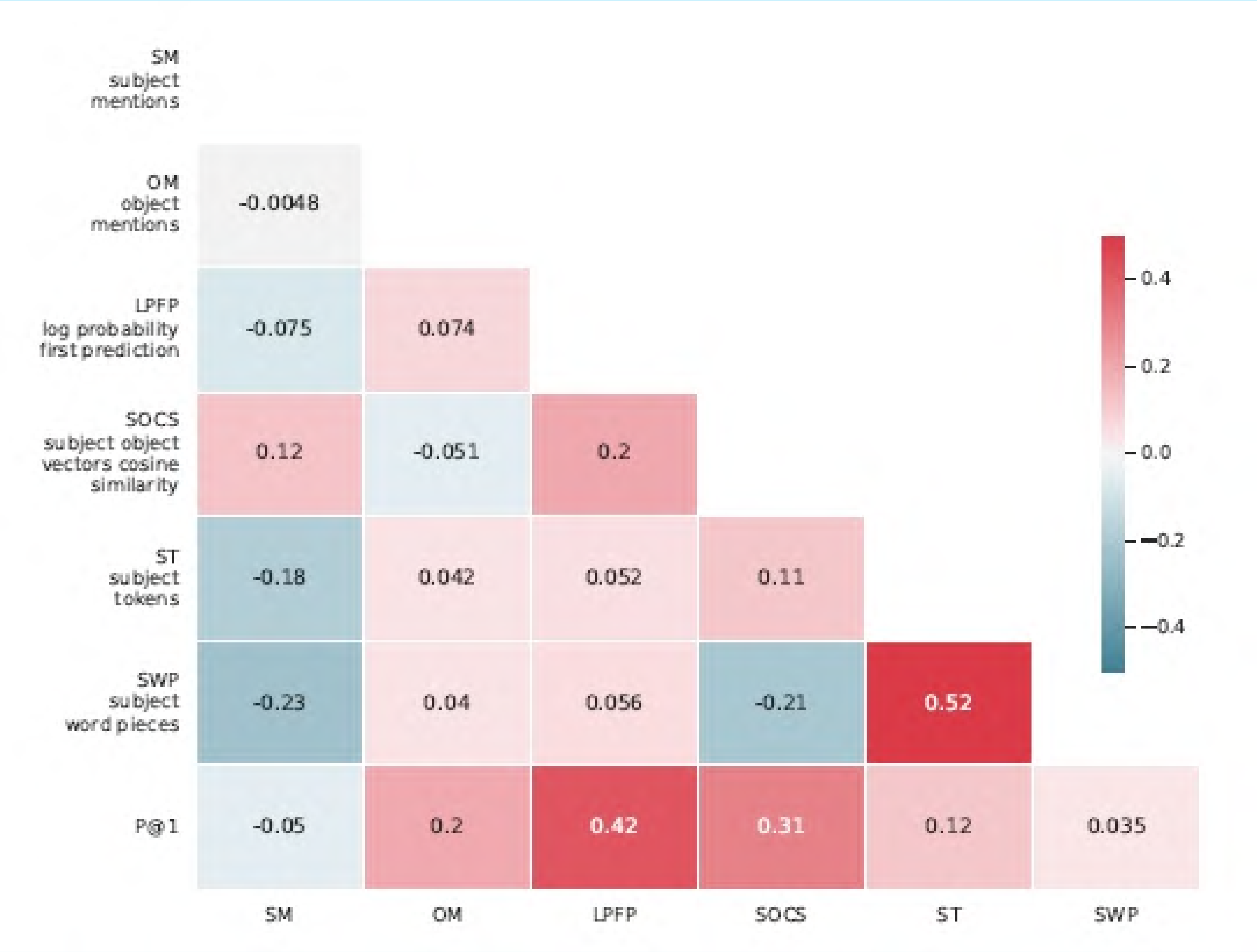

Chart from "Language Model as Knowledge Bases?" Petroni et al.

T-REX

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

Additional takeaways

Dataset	Query	Answer	Generation
T-Rex	Dani Alves plays with ____ .	Barcelona	Santos, Porto, Sporting, Brazil, Portugal
ConceptNet	Time is ____.	finite	short , passing , precious, irrelevant, gone

Pitch

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?

Pitch

References

[1] Mihai Surdeanu and Heng Ji. 2014. Overview of the English Slot Filling Track at the TAC2014 Knowledge Base Population Evaluation.

[2] Hady Elsahar, Pavlos Vougiouklis, Arslen Remaci, Christophe Gravier, Jonathon Hare, Frederique Laforest, and Elena Simperl. 2018. T-rex: A large scale alignment of natural language with knowledge base triples.

[3] Robert Speer and Catherine Havasi. 2012. Representing general relational knowledge in conceptnet 5.

[4] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text.

[5] Daniil Sorokin and Iryna Gurevych. 2017. Contextaware representations for knowledge base relation extraction.

[6] Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading wikipedia to answer opendomain questions.

[7] Yann N. Dauphin, Angela Fan, Michael Auli, and David Grangier. 2017. Language modeling with gated convolutional networks.

[8] Zihang Dai, Zhilin Yang, Yiming Yang, Jaime G. Carbonell, Quoc V. Le, and Ruslan Salakhutdinov. 2019. Transformer-xl: Attentive language models beyond a fixed-length context.

[9] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018a. Deep contextualized word representations.

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018a. BERT: pre-training of deep bidirectional transformers for language understanding.