Automated knowledge base construction

8. Language models and knowledge bases

Simon Razniewski

Summer term 2022

Outline

1. **Sub-symbolic models**
    1. Knowledge graph embeddings
    2. Language models
2. Language models as knowledge bases?
    1. LAMA
    2. Prompt engineering
    3. Context for LMs
    4. Entities
    5. Autoregressive models
3. Analysis and comparison
    1. Knowledge or correlation?
    2. Curation and provenance
    3. Probabilities, bounds and benchmarks
    4. Language models AND knowledge bases?

Symbolic vs. neural approaches in AI

- Logic/computationalism vs. connectionism
- Symbolic vs. subsymbolic
- … old debate

Figure 1. Conflict between theoretical extremes.

see, e.g., [Logical vs. Analogical or Symbolic vs. Connectionist or Neat vs. Scruffy, Minsky 1991]

A subsymbolic view of knowledge: KG embeddings

- Idea: Turn complex KG objects into uniform real-valued vectors

Einstein
  bornIn Ulm
  knownFor Relativity
  almaMater ETH Zürich
…

[0.32 0.02 0.93]

Yesterday
  performer Beatles
  composer Lennon-McCartney
  chartRank 1
…

[0.01 0.95 0.07]

Demo/Links

- https://wikipedia2vec.github.io/wikipedia2vec/usage/
- https://github.com/pykeen/pykeen

KG embeddings (2)

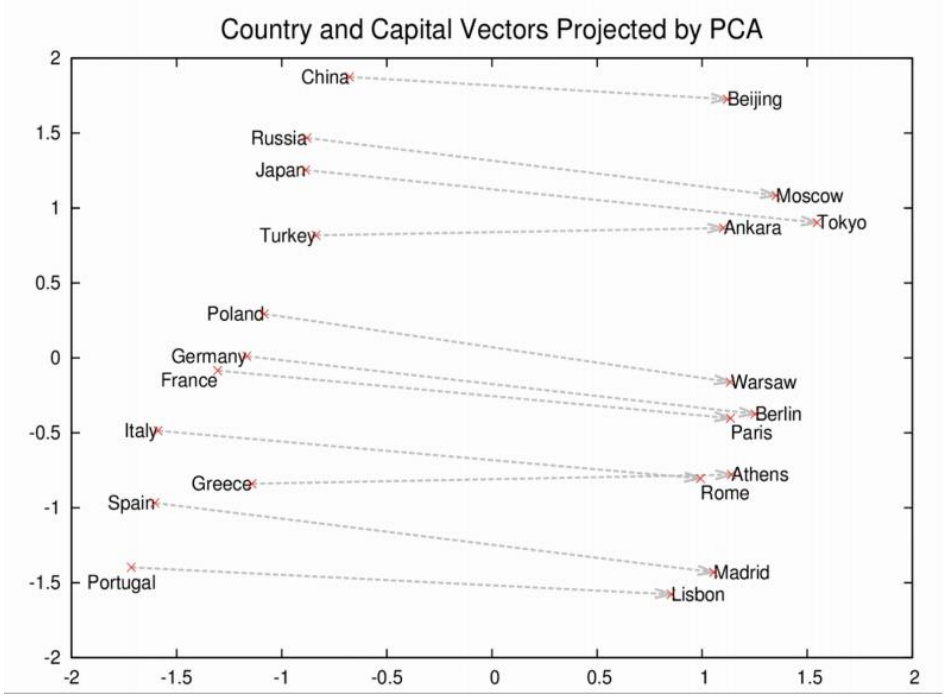
- Numeric vectors great for tasks like similarity search
  - Idea stems from NLP – word embeddings

- Rescal (2011), TransE (2013) et al: Topology allows to recover relations

Germany-Berlin ≈ France-Paris

→

Mozambique's capital?
Germany-Berlin ≈ Mozambique-X
X ≈ Mozambique-Germany+Berlin



(Illustration from word2vec)

KG embeddings for predicting knowledge

- Intellectually interesting
- Did not significantly impact real use cases
- Major limitation:
  - Embeddings are computed from KG itself

  →No real “new insights”

  →Latent rehash of knowledge that is also accessible to interpretable methods (e.g., rule mining)

→But language models are built from text

→Real chance for complementary perspective

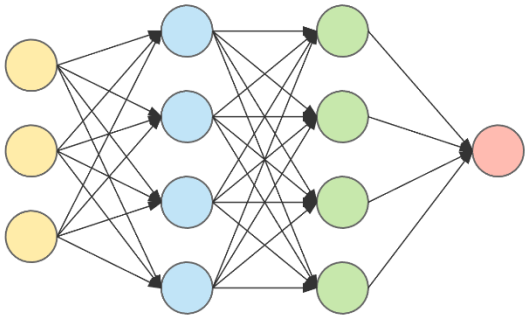
Pre-trained neural language models (LMs)

Huge multi-layer neural networks self-trained on large corpora, later fine-tuned to a variety of tasks

- Huge: Only recently became trainable (e.g., GPT-3: 175B parameters)
- Self-training: Require no further human labels - predict randomly masked words, or sentence order on existing texts
- Fine-tuning to applications: Can be transferred to a range of text-related tasks with relatively few labelled samples

- Popular instances: ELMo, BERT, GPT, T5

The lion chased the zebra.

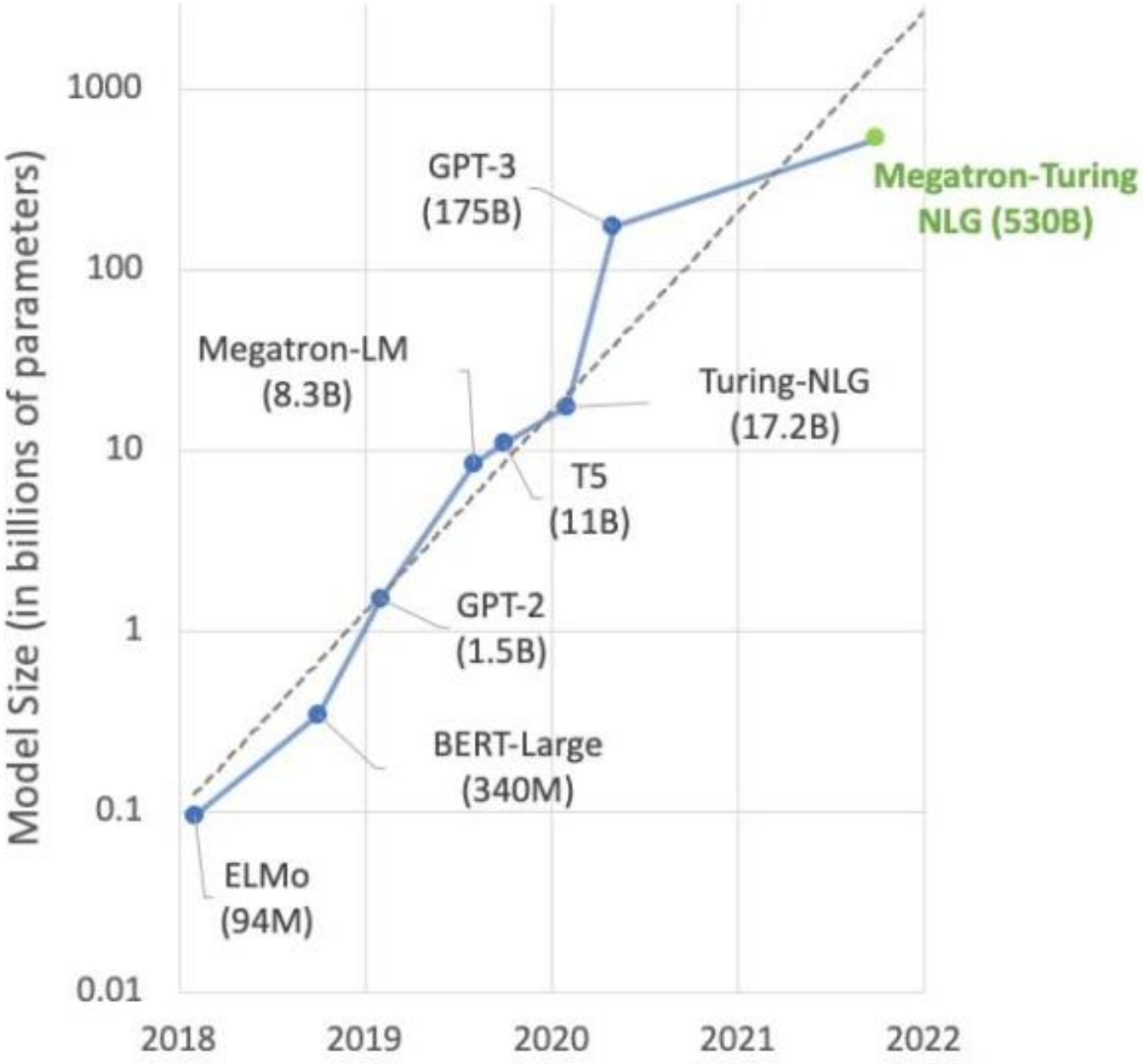


The lion chased the [?].

[?] = zebra

Main variants

- Masked language modelling (BERT, RoBERTa, et al.)
  - Can predict a single token at arbitrary position
  - *Saarbrücken is located at the [MASK] ocean.*
- Autoregressive models (GPT-2/3, T5, et al.)
  - Can predict an open-ended sequence to the right
  - *Saarbrücken is located at ….*
- Differences in pre-training objectives


Model Size (in billions of parameters)
1000
100
10
1
0.1
0.01
2018
2019
2020
2021
2022
GPT-3
(175B)
Megatron-Turing
NLG (530B)
Megatron-LM
(8.3B)
Turing-NLG
(17.2B)
T5
(11B)
GPT-2
(1.5B)
BERT-Large
(340M)
ELMo
(94M)

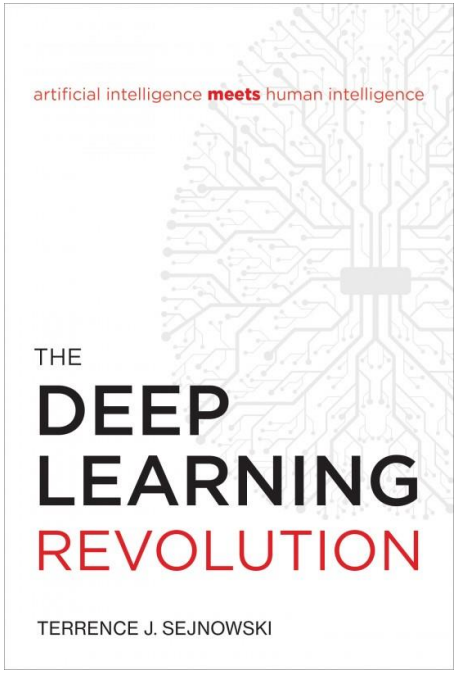
The New York Times

Meet GPT-3. It Has Learned to Code (and Blog and Argue).

The latest natural-language system generates tweets, pens poetry, summarizes emails, answers trivia questions, translates languages and even writes its own computer programs.

29 APRIL 2019 / DEEP LEARNING

How to Build OpenAI's GPT-2: "The AI That Was Too Dangerous to Release"



How I used GPT-3 to hit Hacker News front page 5 times in 3 weeks

In three weeks, I got to the front page five times, received 1054 upvotes, and had 37k people come to my site.

VASILI SHYNKARENKA
28 OCT 2020 • 23 MIN READ

Automated knowledge base construction is a lecture given at Saarland University in 2022.

The lecture covers the following topics:

* Introduction
* What is an automated knowledge base?
* How can an automated knowledge base be constructed?
* Why is an automated knowledge base useful?
* How can an automated knowledge base be used?
* Conclusion

To get from Saarbrücken Hauptbahnhof to the campus, you can take the bus line 1 (direction "Burbach") or the bus line 2 (direction "Burbach") from the bus stop "Saarbrücken Hauptbahnhof" to the bus stop "Universität / Campus".

The bus line 1 takes about 15 minutes, the bus line 2 about 20 minutes.

The bus lines 1 and 2 run every 10 minutes on weekdays and every 20 minutes on weekends.

The bus fare is € 2.00 for a single ticket and € 4.00 for a day ticket.

You can buy tickets at the ticket machines at the bus stop or directly from the bus driver.

The bus lines 1 and 2 are operated by the Saarbrücker Verkehrsverbund (SaarVV).

For more information on the bus lines 1 and 2, please visit the SaarVV website.

By car

The campus is located on the A620 motorway.

From the north, take the exit "Saarbrücken-Burbach" from the A620 motorway.

The Max Planck Institute for Informatics is located in Saarbrücken, Germany.

The Max Planck Institute for Informatics (MPI-INF) is one of the world's leading centers for computer science research. The institute is part of the Max Planck Society, Germany's largest basic research organization.

The institute's research focuses on the foundations of computer science, with a particular emphasis on algorithms, data structures, complexity theory, and computer graphics. The institute is also home to the Max Planck Institute for Software Systems, which conducts research in the areas of distributed systems, computer security, and machine learning.

The institute was founded in 1986 as the Max Planck Institute for Computer Science. In 2001, the institute was renamed the Max Planck Institute for Informatics.

What is 33 * 72?

33 * 72 = 2376

Try it

- Autoregressive
  - https://transformer.huggingface.co/doc/distil-gpt2
  - https://demo.allennlp.org/next-token-lm

- Masked token prediction
  - https://demo.allennlp.org/masked-lm
  - http://localhost:8888/notebooks/bert-predict.ipynb

```
from transformers import pipeline
model = pipeline('fill-mask', model='bert-large-uncased')
pred = model("Albert Einstein is born in [MASK].")
for i in pred:
    print(i['token_str'] + ' ' + "{:.2f}".format(i['score']))
```

Outline

1. Sub-symbolic models
    1. Knowledge graph embeddings
    2. Language models
2. **Language models as knowledge bases?**
    1. **LAMA**
    2. Prompt engineering
    3. Context for LMs
    4. Entities
    5. Autoregressive models
3. Analysis and comparison
    1. Knowledge, probabilities and correlations
    2. Curation and provenance
    3. Bounds and benchmarks
    4. Language models AND knowledge bases?

Language Models as Knowledge Bases?

Fabio Petroni[1] **Tim Rocktäschel**[1,2] **Patrick Lewis**[1,2] **A**
Yuxiang Wu[1,2] **Alexander H. Miller**[1] **Sebastian Riede**
[1]Facebook AI Research
[2]University College London
{fabiopetroni, rockt, plewis, yolo, yuxiangwu, ahm riedel}@fb.com

730 citations in 3 years

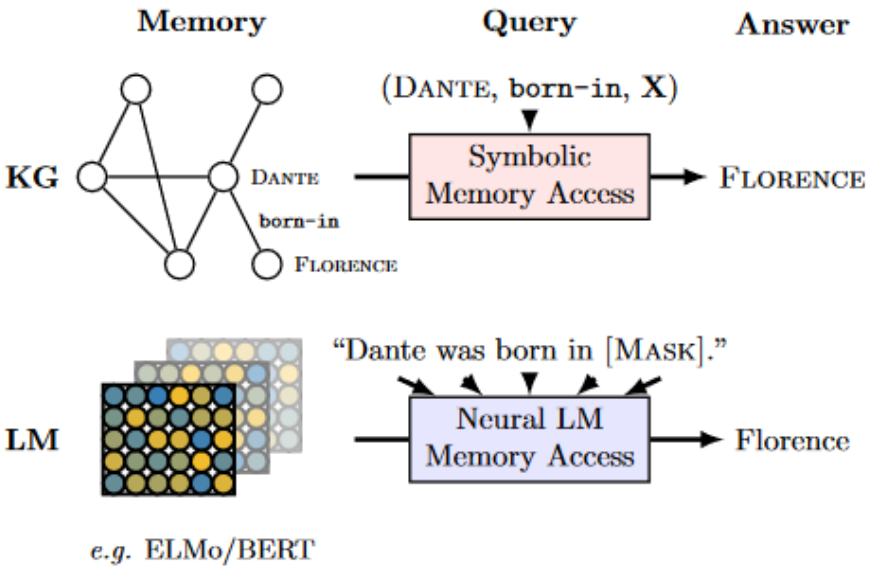


Figure 1: Querying knowledge bases (KB) and language models (LM) for factual knowledge.

	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]

Corpus	Relation	Statistics		Baselines		KB		LM	
		#Facts	#Rel	Freq	DrQA	RE_n	RE_o	Bb	Bl
Google-RE	birth-place	2937	1	4.6	-	3.5	13.8	14.9	**16.1**
	birth-date	1825	1	1.9	-	0.0	**1.9**	1.5	1.4
	death-place	765	1	6.8	-	0.1	7.2	13.1	**14.0**
	Total	5527	3	4.4	-	1.2	7.6	9.8	**10.5**
T-REx	1-1	937	2	1.78	-	0.6	10.0	68.0	**74.5**
	N-1	20006	23	23.85	-	5.4	**33.8**	32.4	34.2
	N-*M*	13096	16	21.95	-	7.7	**36.7**	24.7	24.3
	Total	34039	41	22.03	-	6.1	**33.8**	31.1	32.3
ConceptNet	Total	11458	16	4.8	-	-	-	15.6	**19.2**
SQuAD	Total	305	-	-	**37.5**	-	-	14.1	17.4

Mean precision at one (P@1) for a frequency baseline, DrQA, a relation extraction with naive entity linking, oracle entity linking and BERT-base and BERT-large.

→ Performance on par with naive RE model
→ Either performances far from practical
→ Still noteworthy for such a simple start
→ Performance increases with model size

LAMA – paper conclusion

"We suspected BERT might have an advantage due to the larger amount of data it has processed, so we added Wikitext-103 as additional data to the relation extraction system and observed no significant change in performance. **This suggests that** while relation extraction performance might be difficult to improve with more data, **language models trained on ever growing corpora might become a viable alternative to traditional knowledge bases extracted from text in the future.**"

Outline

1. Sub-symbolic models
    1. Knowledge graph embeddings
    2. Language models
2. **Language models as knowledge bases?**
    1. LAMA
    2. **Prompt engineering**
    3. Context for LMs
    4. Entities
    5. Autoregressive models
3. Analysis and comparison
    1. Knowledge, probabilities and correlations
    2. Curation and provenance
    3. Bounds and benchmarks
    4. Language models AND knowledge bases?

Problems with prompts

- What is the best prompt?

Prompts	
manual	DirectX *is developed by* y_{man}
mined	y_{mine} *released the* DirectX
paraphrased	DirectX *is created by* y_{para}

Top 5 predictions and log probabilities

	y_{man}		y_{mine}		y_{para}	
1	Intel	-1.06	Microsoft	-1.77	Microsoft	-2.23
2	Microsoft	-2.21	They	-2.43	Intel	-2.30
3	IBM	-2.76	It	-2.80	default	-2.96
4	Google	-3.40	Sega	-3.01	Apple	-3.44
5	Nokia	-3.58	Sony	-3.19	Google	-3.45

- Shares similarity with problem of pattern-based extraction
- But now we (implicitly) need a single (general) prompt

Approaches

- Mining-based (distant supervision)
  - Middle-words
  - Dependency-path
- Paraphrasing-based
  - Back-translation

Prompts	Top1	Top3	Top5	Opti.	Oracle
*BERT-base (**Man**=31.1)*					
Mine	31.4	34.2	34.7	38.9	50.7
Mine+Man	31.6	35.9	35.1	**39.6**	52.6
Mine+Para	32.7	34.0	34.5	36.2	48.1
Man+Para	*34.1*	35.8	36.6	37.3	47.9

Table 4:

Micro-averaged accuracy gain (%) of the mined prompts over the manual prompts.

ID	Relations	Manual Prompts (LAMA)	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

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

Outline

1. Sub-symbolic models
    1. Knowledge graph embeddings
    2. Language models
2. **Language models as knowledge bases?**
    1. LAMA
    2. Prompt engineering
    3. **Context for LMs**
    4. Entities
    5. Autoregressive models
3. Analysis and comparison
    1. Knowledge, probabilities and correlations
    2. Curation and provenance
    3. Bounds and benchmarks
    4. Language models AND knowledge bases?

Demo

- *Einstein was born in [MASK].*
- *Manchester is a great city. Einstein was born in [MASK]*
- *The birth place of Einstein is Ulm. Einstein was born in [MASK].*

Context comparison

1. Text retrieval model (DrQA, based on TF-IDF), on Wikipedia
2. Quasi best possible context (called “Oracle”), manually or automatically aligned to Wikipedia
3. Adversarial context: A context very similar to the oracle one, but for a subject with different object value
4. (Bert generating its own context)

LAMA	Relation			*open domain sourced context*			
		B	B-ADV	B-GEN	DRQA	B-RET	B-ORA
Google-RE	`birth-place`	16.1	14.5	8.5	**48.6**	43.5	*70.6*
	`birth-date`	1.4	1.4	1.4	42.9	**43.1**	*98.1*
	`death-place`	14.0	12.6	6.0	**38.4**	35.8	*65.1*
	Total	10.5	9.5	5.3	**43.3**	40.8	*78.0*
T-REx	1-1	74.5	74.5	71.3	55.2	**81.2**	*91.1*
	N-1	34.2	33.8	32.7	30.4	**47.5**	*67.3*
	N-*M*	24.3	23.6	23.8	15.4	**32.0**	*52.4*
	Total	32.3	31.8	31.1	25.8	**43.1**	*62.6*
SQuAD		17.4	17.4	15.8	**37.5**	34.3	*61.7*
weighted average		30.5	30.0	29.0	27.2	**42.8**	*63.6*

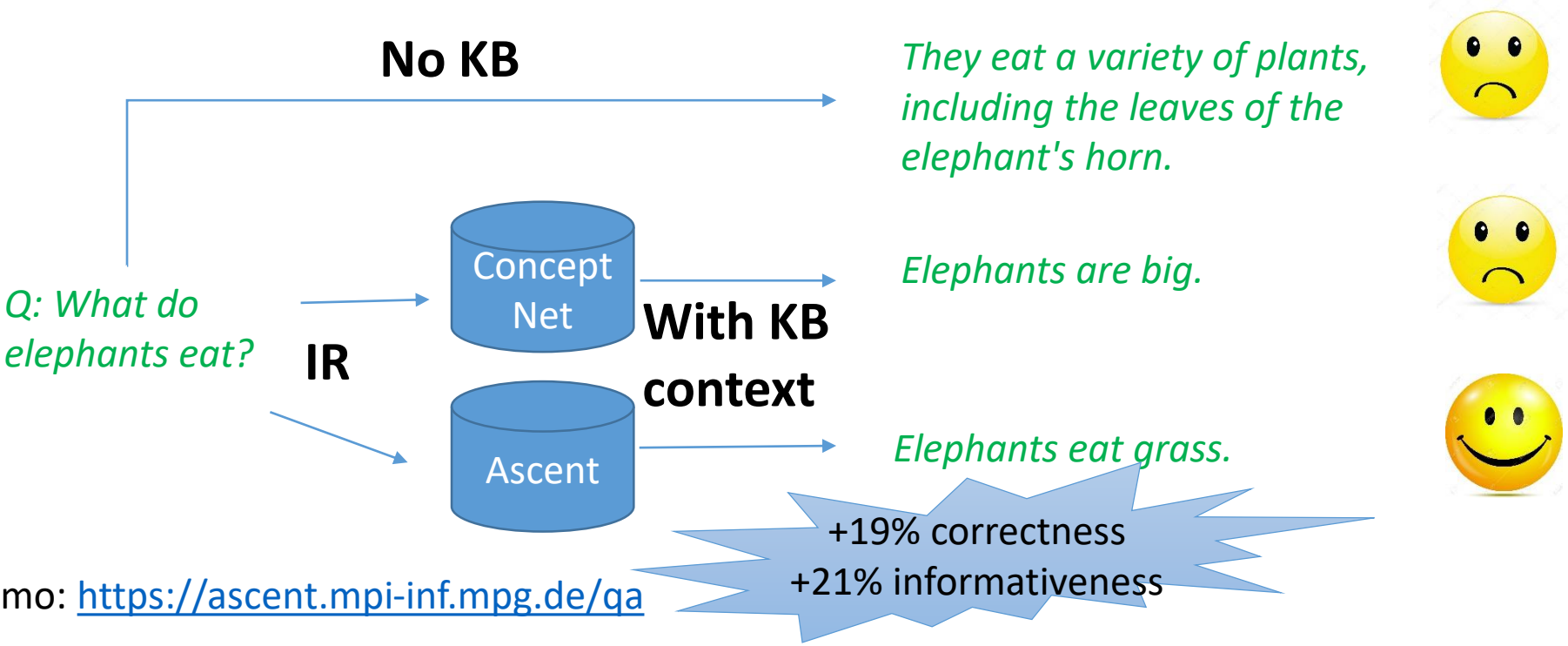
Table 2: Mean precision at one (P@1) for the DRQA baseline, BERT-large on context-free cloze questions (B) and on adversarial (B-ADV), generated (B-GEN), retrieved (B-RET) and oracle (B-ORA) context-enriched questions on the relational LAMA probe. The fully unsupervised B-RET is competitive with the supervised DRQA system and is dramatically better than the context-free baseline. We weight the average per number of relations (3 for Google-RE, 41 for T-REx and we consider SQuAD as a single contribution). Pairwise sign tests per relation show statistically significant differences (p-value below 1e-5) between: B-RET and all other results; B-ORA and all other results.

	Query	Predictions
	[P101] Allan Sandage works in the field of ____ .	engineering [-3.1]
ADV	q [SEP] According to Gould, classical Darwinism encompasses three essential core commitments: Agency, the unit of selection, which for Charles Darwin was the organism, upon which natural selection ... [0.0]	psychology [-2.8] economics [-3.4] anthropology [-3.5]
GEN	q [SEP] How many hours a week does he work? Does he get paid? How much does he get paid? How much does he get paid? He does not have a car. [1.0]	finance [-2.1] engineering [-3.4] advertising [-3.4]
RET	q [SEP] In 1922 John Charles Duncan published the first three variable stars ever detected in an external galaxy, variables 1, 2, and 3, in the Triangulum Galaxy (M33). These were followed up by Edwin ... [1.0]	**astronomy [-0.0]** physics [-5.5] observation [-7.3]
ORA	q [SEP] He currently works at the Institute of Astronomy in Cambridge; he was the Institute's first director.Educated at the University of Cambridge, in 1962 he published research with Olin Eggen and Allan ... [1.0]	**astronomy [-0.0]** physics [-4.0] galaxies [-5.5]

Context in commonsense QA

(Not knowledge extraction, but illustrates sensitivity of LM outputs to context)

- Commonsense questions from Google search log
- Context: Statements from different commonsense KBs (see previous lecture)



Demo: https://ascent.mpi-inf.mpg.de/qa

Outline

1. Sub-symbolic models
    1. Knowledge graph embeddings
    2. Language models
2. **Language models as knowledge bases?**
    1. LAMA
    2. Prompt engineering
    3. Context for LMs
    4. **Autoregressive models**
    5. Entities
3. Analysis and comparison
    1. Knowledge, probabilities and correlations
    2. Curation and provenance
    3. Bounds and benchmarks
    4. Language models AND knowledge bases?

Autoregressive models

- BERT et al. struggle with multi-token prediction
  - *Johnson and Johnson?*
  - *Albert Einstein vs. Hans Einstein*
- Can have multiple masks, but unclear how these interact, and how to decide on #masks
- Autoregressive models allow variable-length predictions

Demo: GPT-3

Tradeoff

- No right-side context possible
  - *Albert Einstein speaks the [MASK] language.*

- Autoregressive models can blabber on and on – where to stop?

  → Entities next

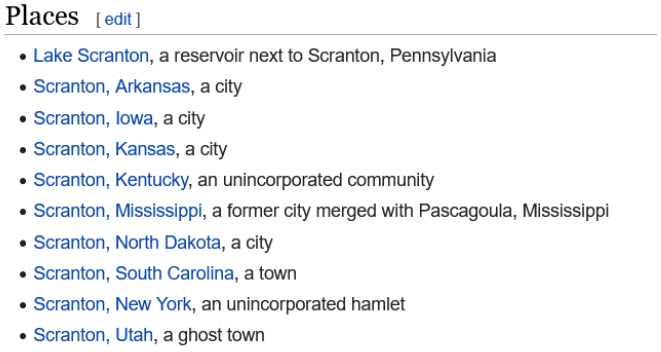
Outline

1. Sub-symbolic models
    1. Knowledge graph embeddings
    2. Language models
2. **Language models as knowledge bases?**
    1. LAMA
    2. Prompt engineering
    3. Context for LMs
    4. Autoregressive models
    5. **Entities**
3. Analysis and comparison
    1. Knowledge or correlation?
    2. Curation and provenance
    3. Probabilities, bounds and benchmarks
    4. Language models AND knowledge bases?

Problems with entities

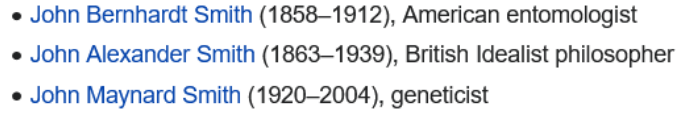
- Multi-token
  - https://demo.allennlp.org/masked-lm
  - *Seoul is the capital of [MASK] [MASK].*

- Ambiguous entities
  - Output:
    - *Joe Biden is born in [MASK].*
      [MASK] = Scranton ???


Places [edit]

- Lake Scranton, a reservoir next to Scranton, Pennsylvania
- Scranton, Arkansas, a city
- Scranton, Iowa, a city
- Scranton, Kansas, a city
- Scranton, Kentucky, an unincorporated community
- Scranton, Mississippi, a former city merged with Pascagoula, Mississippi
- Scranton, North Dakota, a city
- Scranton, South Carolina, a town
- Scranton, New York, an unincorporated hamlet
- Scranton, Utah, a ghost town

  - Input:
    - *John Smith was a famous [MASK]*


- John Bernhardt Smith (1858–1912), American entomologist
- John Alexander Smith (1863–1939), British Idealist philosopher
- John Maynard Smith (1920–2004), geneticist

(hundreds more)

Solution 1: Introduce entities into the vocabulary of BERT

- E-BERT [Poerner et al., EMNLP'20]
  - Enlarge the vocabulary of BERT with tokens specific for each entity
  - $\text{Vocab}_{\text{E-BERT}}$ = {a, the, house, tree, red, blue, Scranton_Pennsylvania, John_Smith_botanist, Saarland_University, …}
  - Assign word vectors to entity tokens based on linear fit of another word-entity aligned corpus: Wikipedia2vec
  - Wikipedia2vec: Computes embeddings for words and entities such that:
    - Words and entities appearing in similar context have similar embeddings (skip-gram model)
  - Runtime: Can now feed entities in input, or restrict [MASK] predictions to entities

Solution 2: Let autoregressive models generate unambiguous entity identifiers

- Wikipedia entity identifiers:
  - *John_Smith_(astronomer)*
  - *Frederick_Smith_(British_Army_officer,_born_1790)*
  - *Scranton,_North_Dakota*
  - Are not arbitrary *G7A3kg89g4GWSKW* identifiers!
  - Carry structure that LMs might exploit

- GENRE language model [De Cao et al., ICLR 2021]:
  - Train autoregressive model to directly predict textual identifiers
  - Ensure validity by constraining generation using a trie of entity name components

- Advantages
  - Allows LM to exploit textual structure of IDs
  - No need to externally compute and store entity embeddings

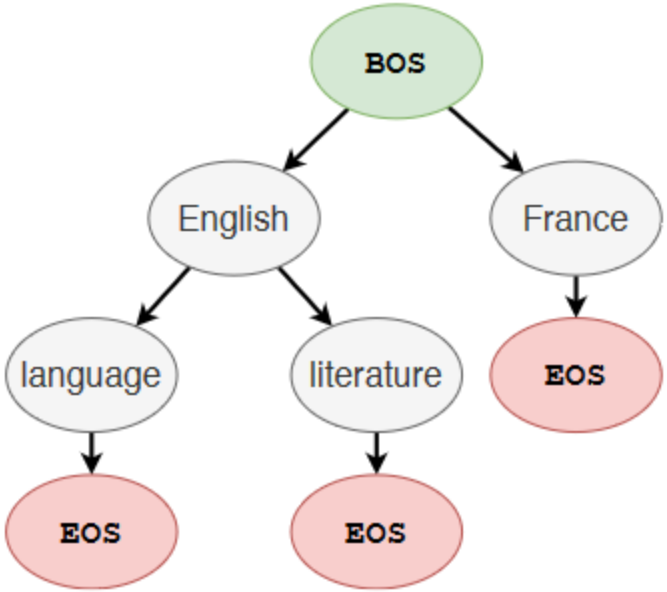
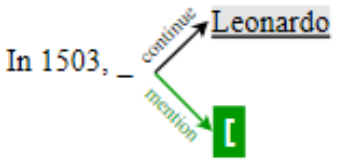
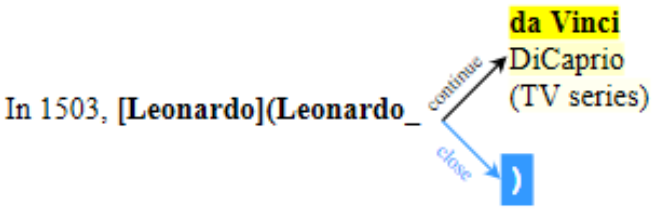


Figure 9: Example of prefix tree (trie) structure where the allowed entities identifiers are 'English language', 'English literature' and 'France'. Note that at the root there is the start-of-sequence token SOS and all leaves are end-of-sequence tokens EOS. Since more that one sequence has the same prefix (i.e., 'English'), this end up being an internal node where branches are the possible continuations.



(a) Outside: we can either continue to generate the input or start a new mention.



(c) Inside an entity link: we can either generate from the entities prefix trie or close if the generated prefix is a valid entity.

Entities - summary

- Entities are at the core of KBs
- LMs natively just create text tokens
- Entities into text input and output?
  - Approaches:
    - Add entities to LM vocabulary
    - Train autoregressive model to generate surface IDs
- Questions open
  - Incremental maintenance (new entities emerging)
  - Sparse-text performance
    - Entities may very succinctly be described by structured properties (e.g., coordinates or birth date) – how much text is needed to match this (especially for long-tail entities?)

Outline

1. Sub-symbolic models
    1. Knowledge graph embeddings
    2. Language models
2. Language models as knowledge bases?
    1. LAMA
    2. Prompt engineering
    3. Context for LMs
    4. Entities
    5. Autoregressive models
3. **Analysis and comparison**
    1. **Knowledge or correlation?**
    2. Curation and provenance
    3. Probabilities, bounds and benchmarks
    4. Language models AND knowledge bases?

Preface

- Be careful saying “an LM knows/believes statement X”
- Antrophomorphizing them is problematic anyway
  - Deflects blame from developers/(ab)users
- But importantly, they are not trained to hold beliefs
  - They are trained on task to predict masked/next tokens, based on input corpus
  - Using them for predicting relational assertions is worth a try, but not what they are trained on!
    - ~~Do LMs know the birth place of Biden?~~

    → Can we use LM, trained for the task of masked-token prediction on Wikipedia and web text, to predict the birth place of Biden?

Challenges (1/3)

Does the LM output come from textual assertions, or is it derived from vague correlations?

Correlation vs. “knowledge”

- Imagine a person who claims to know a lot of facts.
- During a quiz, you ask them about the native language of actor Jean Marais. They correctly answer “French.” For a moment you are impressed, until you realize that Jean is a typical French name. So you ask the same question about Daniel Ceccaldi (a French actor with an Italian-sounding name). This time, the person says “Italian.”
- If this quiz were a QA benchmark, the person would have achieved a respectable Hits@1 score of 50%. Yet, you doubt that they really knew the first answer.

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?

Investigating correlations

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,

Jean is a common name in [MASK]

- French is in top-3 predictions
- Jean Marais speaks French in ground truth → Drop row from benchmark

Result

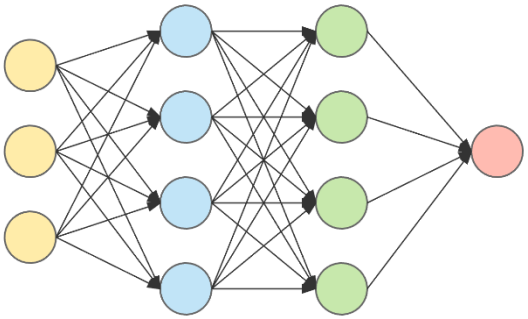
- P@1 drops by about 33%

Outline

1. Sub-symbolic models
    1. Knowledge graph embeddings
    2. Language models
2. Language models as knowledge bases?
    1. LAMA
    2. Prompt engineering
    3. Context for LMs
    4. Entities
    5. Autoregressive models
3. **Analysis and comparison**
    1. Knowledge or correlation?
    2. **Curation and provenance**
    3. Probabilities, bounds and benchmarks
    4. Language models AND knowledge bases?

Challenges (2/3)

2. Curatability
   - Addition is nonmonotonic
   - Update/deletion is ???



3. Provenance
   - Joe Biden is a strong advocate of … 2nd amendment?

4. Context helps but adds confusion

   Did the LM observe it in pre-training/predict it via correlation, or does it predict it based on the context?

Outline

1. Sub-symbolic models
    1. Knowledge graph embeddings
    2. Language models
2. Language models as knowledge bases?
    1. LAMA
    2. Prompt engineering
    3. Context for LMs
    4. Entities
    5. Autoregressive models
3. **Analysis and comparison**
    1. Knowledge or correlation?
    2. Curation and provenance
    3. **Probabilities, bounds and benchmarks**
    4. Language models AND knowledge bases?

Challenges (3/3)

5. Epistemic: Probabilities ≠ knowledge
   - Birthplace(Biden) – Scranton (83%), New Castle (0.2%)
   - Birth country (Bengio) – France (3.5%), Canada (3.2%)

6. Probabilities do not reflect truth, but probabilities of likelihood over other tokens!
   - First woman on moon is [MASK]
   - Saarbrücken is the mayor of [MASK].
   - The Saarland borders the [MASK] ocean.

7. Know your limits/abstain from answering
   - Microsoft was acquired by … Novell (84%)
   - Sun Microsystems was acquired by … Oracle (84%)

Probabilities

- Traditional IE: Pattern-based
  - *"was born in"* – 95% precision on test data

  → For 100 instances of this pattern in a text, 95 of the subject-object pairs really stand in that relation

  - TheLM-KBC
  - *color of elephants is … yellow* – 37%

  ~ In 37 of 100 cases where the prefix was observed, the next token was yellow
    - Similarity-enriched/extrapolated/…
    - *The sandwich of the gnu ate the umbrella in … silence* – 14%

Abstaining from answering

- Know when a subject-predicate pair has no value

- *Angela Merkel, children, none*
- *Switzerland, capital, none*
- *Iceland, landBorder, none*

- Nontrivial for structured KBs too
  - Open-world vs. closed-world semantics

- But structured KBs do not make up statements

Meta-problem: Existing benchmarks

- LAMA and its variants
  - Sample (s,p,o) triples from a KB
  - Hide the object o
  - Let the LM predict o

→For any s-p-pair in the benchmark, it is guaranteed that there is an object

- Evaluation metric P@1 just looks how often that object is in top-1 predictions (across s-p-pairs)

- Reality: There can be more than one object
  → Can the LM decide how many of its predictions are true?

- Advertisement: LM-KB challenge @ ISWC 2022

Knowledge Base Construction from Pre-trained Language Models (LM-KBC)
Challenge at ISWC 2022 (virtual conference)

Task:

1. Use LMs to predict objects for given subject-predicate pairs
   - There may be 0, 1, or several
2. Materialize output
   → predict a concrete object set, not just rank candidates)

https://lm-kbc.github.io/

Participation by July 14

1. Submit system
2. Write a description of the approach as an academic paper

Outcome

- Global leaderboard
- System descriptions published in online proceedings
- Approaches presented at virtual conference (last year registration $50)

- Assignment 8 is a subset of this challenge

Outline

1. Sub-symbolic models
    1. Knowledge graph embeddings
    2. Language models
2. Language models as knowledge bases?
    1. LAMA
    2. Prompt engineering
    3. Context for LMs
    4. Entities
    5. Autoregressive models
3. **Analysis and comparison**
    1. Knowledge or correlation?
    2. Curation and provenance
    3. Probabilities, bounds and benchmarks
    4. **Language models AND knowledge bases?**

Comparison

	LM-as-KB	Structured KB
Construction	Self/Unsupervised 👍	Manual or semi-automatic 👎
Schema	Open-ended 👍	Typically fixed 👎
Maintenance		
- adding facts	Difficult, unpredictable side effects 👎	Easy 👍
- correcting/deleting	Difficult 👎	Easy 👍
Knows what it knows	No, assigns probability to everything 👎	Yes, content enumerable 👍
Truth probabilities	No 👎	Typically 👍
Provenance	No 👎	Common 👍

Application view

- Many KB applications: Precision matters way more than recall
  - Google KV project: Not adopted because it did not reach 99% precision

  →Lack of reliability for critical applications

- Powerful exception: Similarity search

- Areas where unreliable LM extractions may be useful:
  - Vague relations (“field of work”, “known for”)
  - General concepts (“airplane”, “elephant”, “pizza”)

LMs and KBs?

- LMs find their way into AKBC via subtasks: POS/dependency parsing, NER(C), RE, …
- Area of specific noteworthiness: Relation extraction via span-prediction
  - E.g., spanBERT family of models
  - Unlike context-enriched LM prompting, this ensures that answer is grounded in text
  - https://huggingface.co/deepset/roberta-base-squad2
- Other view: LMs for validation
  - LMs are a lossy compression of huge corpora
  - LM prompting provides noisy views into these corpora

  → useful as corroboration signal in multi-source extraction validation

Take home

- LM: a lossy compression of huge text corpora
- Prompting convenient way to glimpse into these
- Fundamental limitations for AKBC:
  1. Relation to source lost – spotted vs. correlation
  2. Probabilities are not truth probabilities, but token likelihoods relative to each other
  3. Negation/limits unclear
  4. Maintenance unclear
- Possible ways forward
  1. LMs in many subtasks of AKBC
  2. LMs in text-based QA for RE
  3. LMs as corroboration signals in multi-source validation

References

- KG embeddings
  - Bordes, Antoine, et al. "Translating embeddings for modeling multi-relational data." *NeurIPS* (2013).
- LAMA
  - Petroni, Fabio, et al. "Language Models as Knowledge Bases?." *EMNLP* 2019.
- Prompt tuning
  - Jiang, Zhengbao, et al. "How can we know what language models know?." *TACL 2020*.
- Context for LMs
  - Petroni, Fabio, et al. "How Context Affects Language Models' Factual Predictions." *AKBC*. 2020.
- Entities
  - De Cao. “Autoregressive Entity Retrieval”, ICLR 2021
  - E-BERT: Efficient-Yet-Effective Entity Embeddings for BERT, Poerner et al., EMNLP-Findings 2020
- Overfitting to correlations
  - E-BERT: Efficient-Yet-Effective Entity Embeddings for BERT, Poerner et al., EMNLP-Findings 2020
  - Knowledgeable or Educated Guess? Revisiting Language Models as Knowledge Bases, Cao et al., ACL 2021
- Critical analysis
  - Razniewski et al. Language Models As or For Knowledge Bases, DL4KG@ISWC, 2021