Commonsense Knowledge in Pre-trained Language Models

Vered Shwartz

July 5th, 2020

Commonsense Knowledge in Pre-trained Language Models

Vered Shwartz

July 5th, 2020

Commonsense Knowledge in Pre-trained Language Models

Vered Shwartz

July 5th, 2020

Commonsense Knowledge in Pre-trained Language Models

Vered Shwartz

July 5th, 2020

Do pre-trained LMs *already* capture commonsense knowledge?

To fine-tune or not to fine-tune, that is the question

To fine-tune or not to fine-tune,

~~that is the question~~

Out-of-the box

Knowledge-base Completion

Converting KB relations to natural language templates and using LMs to query / score

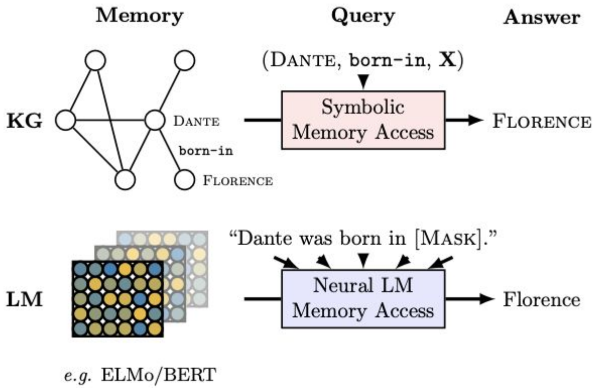
LMs:
Templates:
KBs:
Conclusion:

Knowledge-base Completion

Converting KB relations to natural language templates and using LMs to query / score

- **Petroni et al. (2019)**:

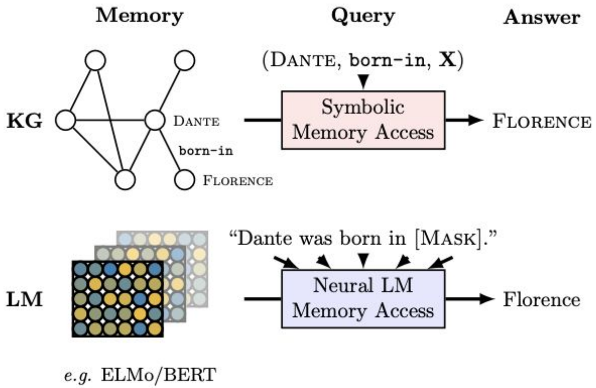
LMs:	ELMo / BERT
Templates:	Hand-crafted templates
KBs:	ConceptNet and Wikidata
Conclusion:	BERT performs well but all models perform poorly on many-to-many relations



Knowledge-base Completion

Converting KB relations to natural language templates and using LMs to query / score

- **Petroni et al. (2019)**:
  - LMs: ELMo / BERT
  - Templates: Hand-crafted templates
  - KBs: ConceptNet and Wikidata
  - Conclusion: BERT performs well but all models perform poorly on many-to-many relations



- **Feldman et al. (2019)**:
  - BERT
  - Hand-crafted templates scored by GPT2
  - ConceptNet, mining from Wikipedia
  - Performs worse than supervised methods on ConceptNet but is more likely to generalize to different domains

Candidate Sentence S_i	$\log p(S_i)$
"musician can playing musical instrument"	-5.7
"musician can be play musical instrument"	-4.9
"musician often play musical instrument"	-5.5
"a musician can play a musical instrument"	$\mathbf{-2.9}$

Table 1: Example of generating candidate sentences. Several enumerated sentences for the triple `(musician, CapableOf, play musical instrument)`. The sentence with the highest log-likelihood according to a pretrained language model is selected.

Properties of Concepts (Weir et al., 2020)

1) Do pre-trained LMs correctly distinguish concepts associated with a given set of assumed properties?

Properties of Concepts (Weir et al., 2020)

1) Do pre-trained LMs correctly distinguish concepts associated with a given set of assumed properties?

A ____ has fur.

Properties of Concepts (Weir et al., 2020)

1) Do pre-trained LMs correctly distinguish concepts associated with a given set of assumed properties?

A ____ has fur.

Properties of Concepts (Weir et al., 2020)

1) Do pre-trained LMs correctly distinguish concepts associated with a given set of assumed properties?

A ____ has fur.
A ____ has fur, is big, and has claws.

Properties of Concepts (Weir et al., 2020)

1) Do pre-trained LMs correctly distinguish concepts associated with a given set of assumed properties?

A ____ has fur.
A ____ has fur, is big, and has claws.
A ____ has fur, is big, and has claws, has teeth, is an animal, …

Properties of Concepts (Weir et al., 2020)

1) Do pre-trained LMs correctly distinguish concepts associated with a given set of assumed properties?

- Good performance, RoBERTa > BERT

Properties of Concepts (Weir et al., 2020)

1) Do pre-trained LMs correctly distinguish concepts associated with a given set of assumed properties?

- Good performance, RoBERTa > BERT
- Perceptual (e.g. visual) < non-perceptual (e.g. encyclopaedic or functional) - can't be learned from texts alone

Properties of Concepts (Weir et al., 2020)

1) Do pre-trained LMs correctly distinguish concepts associated with a given set of assumed properties?

- Good performance, RoBERTa > BERT
- Perceptual (e.g. visual) < non-perceptual (e.g. encyclopaedic or functional) - can't be learned from texts alone
- Highly-ranked incorrect answers typically apply to a subset of properties

Properties of Concepts (Weir et al., 2020)

2) Can pre-trained LMs be used to list the properties associated with given concepts?

Context	Human Response	PF	RoBERTa-L Response	p_{LM}
Everyone knows that a bear has ___	fur	27	teeth	.36
	claws	15	claws	.18
	teeth	11	eyes	.05
	cubs	7	ears	.03
	paws	7	horns	.02

Properties of Concepts (Weir et al., 2020)

2) Can pre-trained LMs be used to list the properties associated with given concepts?

Context	**Human Response**	**PF**	**RoBERTa-L Response**	p_{LM}
Everyone knows that a bear has ___	fur	27	teeth	.36
	claws	15	claws	.18
	teeth	11	eyes	.05
	cubs	7	ears	.03
	paws	7	horns	.02

Low correlation with human elicited properties, but coherent and mostly "verifiable by humans".

Can we trust knowledge from LMs?

How well do LMs handle mutual exclusivity?*

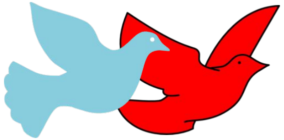
Sentence:

The color of the dove who was sitting on the bench was [MASK].

Mask 1 Predictions:

15.0%	red
9.8%	blue
7.0%	different
5.7%	yellow
5.3%	purple

How well do LMs handle mutual exclusivity?*



Sentence:

The color of the dove who was sitting on the bench was [MASK].

Mask 1 Predictions:

15.0%	**red**
9.8%	**blue**
7.0%	**different**
5.7%	**yellow**
5.3%	**purple**

*Disclaimer: anecdotal example https://demo.allennlp.org/masked-lm

LMs also generate fictitious facts!

LMs also generate fictitious facts!

Distributionally-related:

Barack's Wife Hillary:
Using Knowledge Graphs for Fact-Aware Language Modeling

Robert L. Logan IV* **Nelson F. Liu†§** **Matthew E. Peters§**
Matt Gardner§ **Sameer Singh***

* University of California, Irvine, CA, USA
† University of Washington, Seattle, WA, USA
§ Allen Institute for Artificial Intelligence, Seattle, WA, USA

{rlogan, sameer}@uci.edu, {mattg, matthewp}@allenai.org, nfliu@cs.washington.edu

LMs also generate fictitious facts!

Distributionally-related:

Barack's Wife Hillary:
Using Knowledge Graphs for Fact-Aware Language Modeling

Robert L. Logan IV[*] **Nelson F. Liu**[†§] **Matthew E. Peters**[§]
Matt Gardner[§] **Sameer Singh**[*]

[*] University of California, Irvine, CA, USA
[†] University of Washington, Seattle, WA, USA
[§] Allen Institute for Artificial Intelligence, Seattle, WA, USA
{rlogan, sameer}@uci.edu, {mattg, matthewp}@allenai.org, nfliu@cs.washington.edu

Syntactically-similar:

Negated LAMA: Birds cannot fly

Nora Kassner, Hinrich Schütze
Center for Information and Language Processing (CIS)
LMU Munich, Germany
kassner@cis.lmu.de

Zero-shot LM-based Models for commonsense tasks

Zero-shot setup

Zero-shot setup

PLM(The answer is **answer_choice_1**)

PLM(The answer is **answer_choice_2**)

...

PLM(The answer is **answer_choice_k**)

Language Model

Zero-shot setup

P_{LM}(The answer is **answer_choice_1**)
P_{LM}(The answer is **answer_choice_2**)
…
P_{LM}(The answer is **answer_choice_k**)

Language Model

P_{LM}(**answer_choice_1** | The answer is [MASK])
P_{LM}(**answer_choice_2** | The answer is [MASK])
…
P_{LM}(**answer_choice_k** | The answer is [MASK])

Masked Language Model

Unsupervised Commonsense Question Answering with Self-Talk
(Shwartz et al., 2020)

Can we use LMs to generate required, missing or implicit knowledge for multiple choice commonsense question answering tasks?

Generating Clarifications

Question Generation

What do professors primarily do?

teach courses

Generating Clarifications

Question Generation

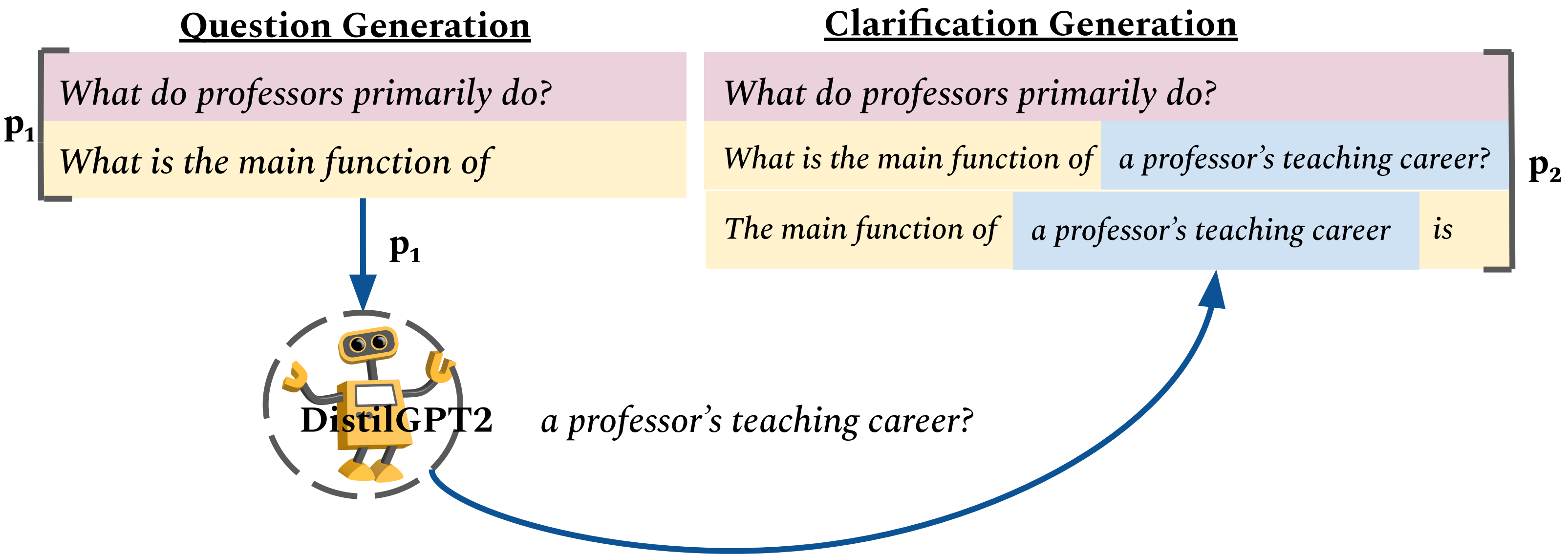
$\mathbf{p_1}$

What do professors primarily do?

What is the main function of

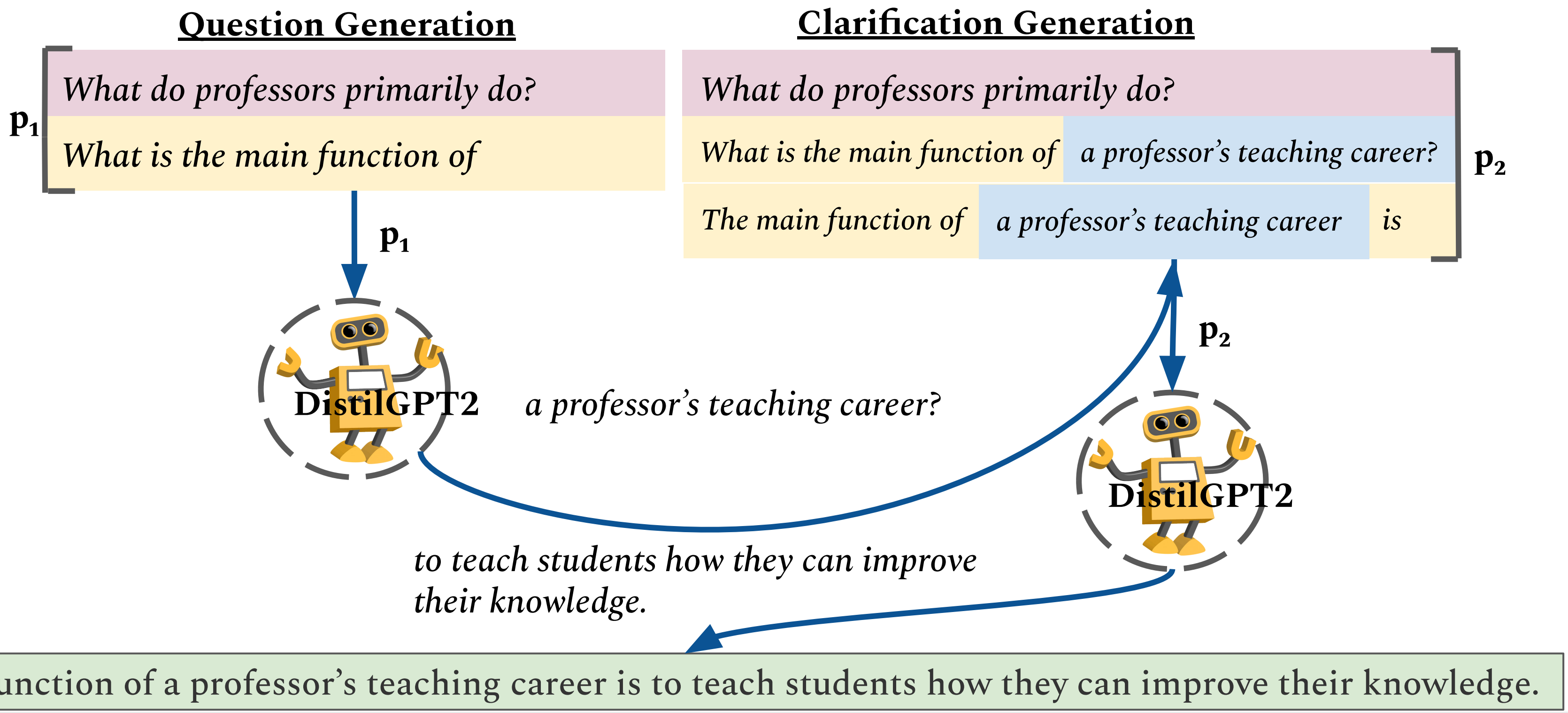
$\mathbf{p_1}$

Generating Clarifications



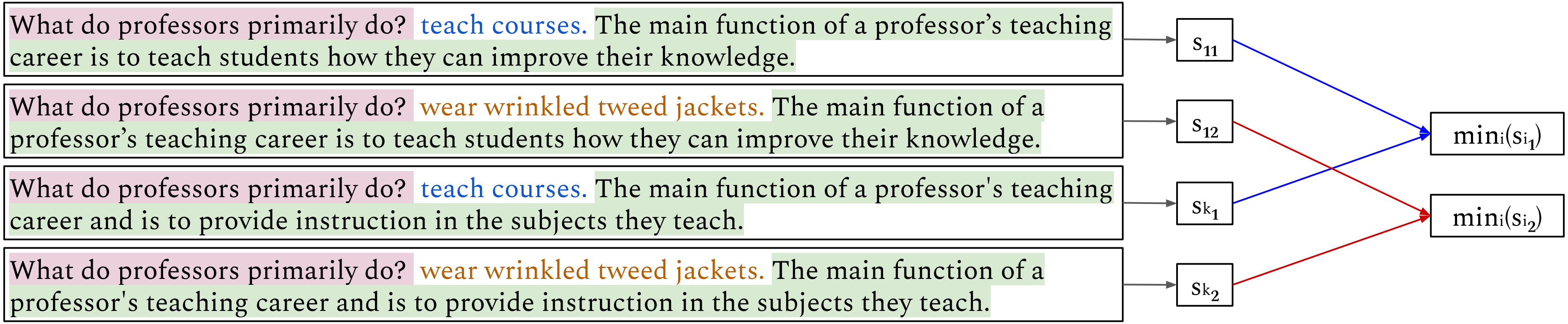
teach courses

Generating Clarifications



teach courses

Model


What do professors primarily do? teach courses. The main function of a professor’s teaching career is to teach students how they can improve their knowledge.
What do professors primarily do? wear wrinkled tweed jackets. The main function of a professor’s teaching career is to teach students how they can improve their knowledge.
What do professors primarily do? teach courses. The main function of a professor's teaching career and is to provide instruction in the subjects they teach.
What do professors primarily do? wear wrinkled tweed jackets. The main function of a professor's teaching career and is to provide instruction in the subjects they teach.
s_{11}
s_{12}
s_{k_1}
s_{k_2}
$\min_i(s_{i_1})$
$\min_i(s_{i_2})$

Knowledge-informed Model

Generating clarifications from ConceptNet, Google Ngrams and COMET

Taylor was doing her job so she put the money in the drawer.

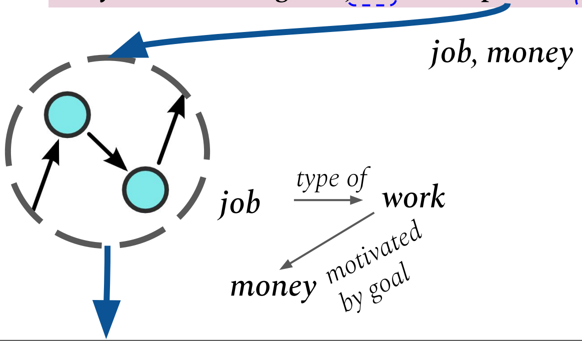
What will Taylor want to do next?

Knowledge-informed Model

Generating clarifications from ConceptNet, Google Ngrams and COMET

Taylor was doing her job so she put the money in the drawer.

What will Taylor want to do next?



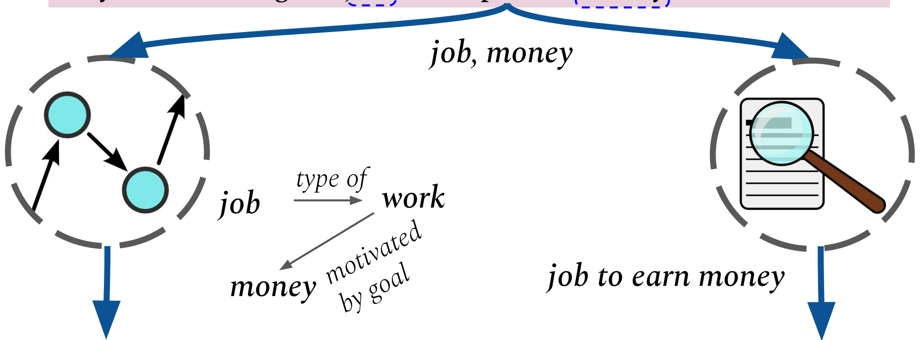
Job is a type of work. You would work because you want money.

Knowledge-informed Model

Generating clarifications from ConceptNet, Google Ngrams and COMET

Taylor was doing her job so she put the money in the drawer.

What will Taylor want to do next?

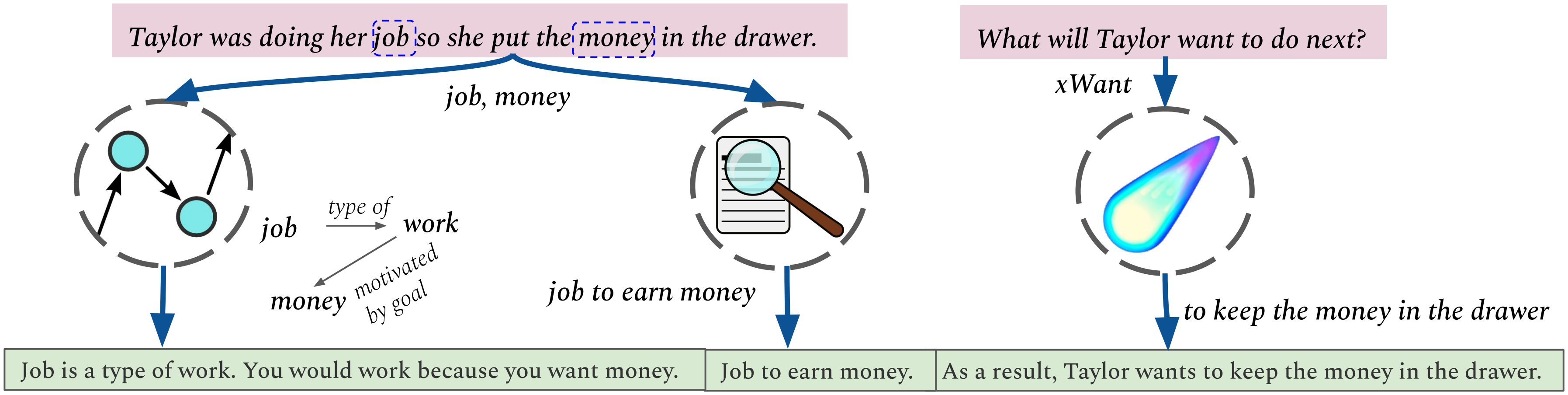


Job is a type of work. You would work because you want money.

Job to earn money.

Knowledge-informed Model

Generating clarifications from ConceptNet, Google Ngrams and COMET



Unsupervised Commonsense Question Answering with Self-Talk

- Generating knowledge with LMs improve upon the baseline and performs similarly to knowledge-informed models.

Unsupervised Commonsense Question Answering with Self-Talk

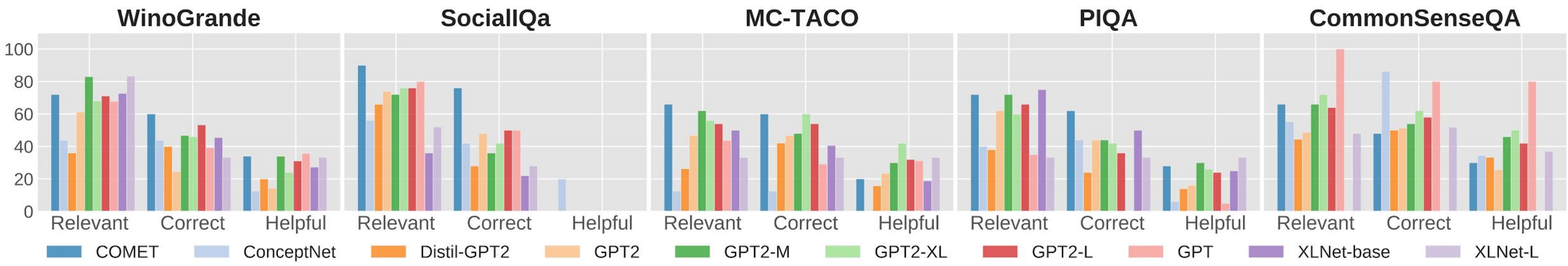
- Generating knowledge with LMs improve upon the baseline and performs similarly to knowledge-informed models.

- Generated clarifications don't align with what humans consider helpful.

Unsupervised Commonsense Question Answering with Self-Talk

- Generating knowledge with LMs improve upon the baseline and performs similarly to knowledge-informed models.

- Generated clarifications don't align with what humans consider helpful.



To fine-tune or not to fine-tune,

~~that is the question~~

LMs provide a good basis for commonsense task models

MC-TACO

+Create Submission

MC-TACO is a dataset of 13k question-answer pairs that require temporal commonsense comprehension... (More)

Public Submissions | Getting Started | About

Human Performance — Exact Match: 0.7580

Rank	Submission	Created	Exact Match	F1
1	T5 - 3B fine-tuned + number n... *Zakaria Kaddari, Youssef Mell...*	03/24/2020	**0.5908**	0.7946
2	T5 - 3B baseline *Zakaria Kaddari, Youssef Mell...*	03/15/2020	0.5758	0.7845

LMs provide a good basis for commonsense task models

Abductive NLI

aNLI

+Create Submission

Abductive Natural Language Inference (aNLI) is a new commonsense benchmark dataset designed to test... (More)

Public Submissions | Getting Started | About

Human Performance — Accuracy: 0.9290

Rank	Submission	Created	Run Time	Accuracy
1	L2R2 (RoBERTa Large + Likeli... *Anonymous Submission*	01/27/2020	33 minutes	**0.8681**
2	egel *dynamics*	01/20/2020	29 minutes	0.8595

LMs provide a good basis for commonsense task models

Physical IQA

We introduce Physical IQa: Physical Interaction QA, a new commonsense QA benchmark for naive... (More)

+Create Submission

Public Submissions | Getting Started | About

Human Performance — Accuracy: 0.9490

Rank	Submission	Created	Run Time	Accuracy
1	RoBERTa Large Finetuning *AI2*	11/19/2019	2 hours	**0.7940**
2	RoBERTa w. CNQA *USC ISI*	12/02/2019	an hour	0.7853

LMs provide a good basis for commonsense task models

WinoGrande

WinoGrande

WinoGrande is a new collection of 44k problems, inspired by Winograd Schema Challenge (Levesque... (More)

+ Create Submission

Public Submissions | Getting Started | About

Human Performance — AUC: 0.9400

Rank	Submission	Created	AUC	Acc (XS)	Acc (S)	Acc (M)	Acc (L)	Acc (XL)
1	TTTTT *University of Waterloo and Ac...*	03/13/2020	**0.7673**	**0.6825**	**0.7051**	**0.7759**	**0.8240**	**0.8461**
2	Roberta-large + G-DAug-Combo *anonymous*	02/23/2020	0.7146	0.6106	0.6712	0.7119	0.7736	0.7923
3	Roberta-large + G-DAug-Div *anonymous*	02/16/2020	0.7118	0.6163	0.6667	0.7046	0.7714	0.7929

LMs provide a good basis for commonsense task models

WinoGrande

WinoGrande is a new collection of 44k problems, inspired by Winograd Schema Challenge (Levesque... (More)

+Create Submission

Public Submissions | Getting Started | About

Human Performance — AUC: 0.9400

Rank	Submission	Created	AUC	Acc (XS)	Acc (S)	Acc (M)	Acc (L)	Acc (XL)
1	TTTTT *University of Waterloo and Ac...*	03/13/2020	**0.7673**	**0.6825**	**0.7051**	**0.7759**	**0.8240**	**0.8461**
2	Roberta-large + G-DAug-Combo *anonymous*	02/23/2020	0.7146	0.6106	0.6712	0.7119	0.7736	0.7923
3	Roberta-large + G-DAug-Div *anonymous*	02/16/2020	0.7118	0.6163	0.6667	0.7046	0.7714	0.7929

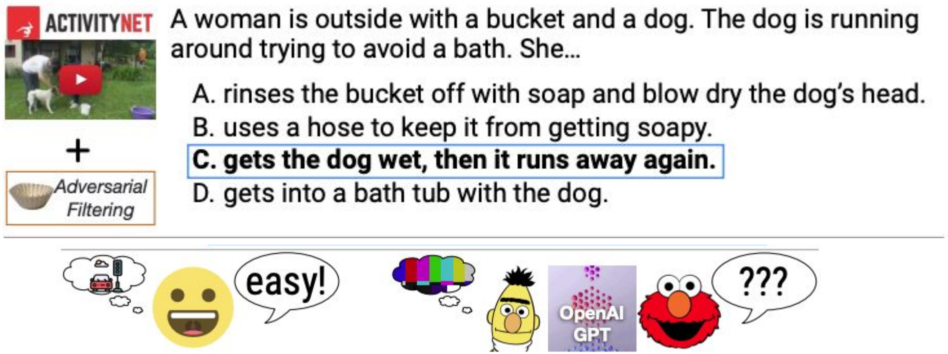
...but they need a "push in the right direction" (fine tuning)

Can good performance be attributed to knowledge in LMs or to training a large model on a large dataset?

HellaSwag (Zellers et al., 2019)

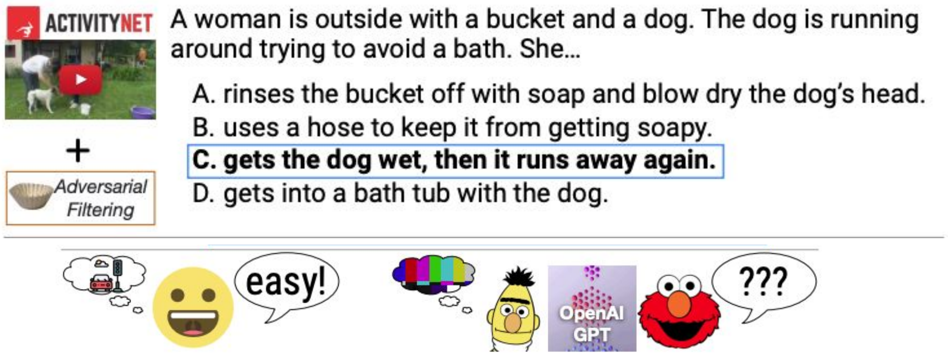
HellaSwag (Zellers et al., 2019)

- LMs mostly pick up ***lexical cues***
- No model actually solves commonsense reasoning to date.

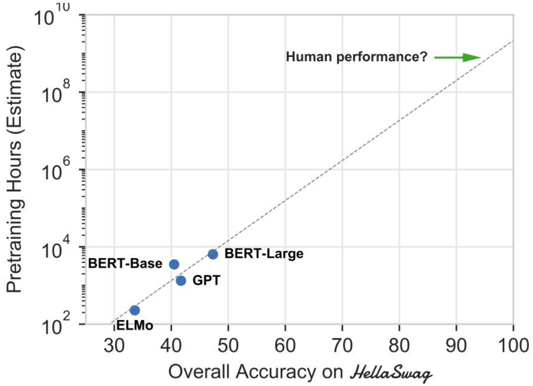


HellaSwag (Zellers et al., 2019)

- LMs mostly pick up ***lexical cues***
- No model actually solves commonsense reasoning to date.

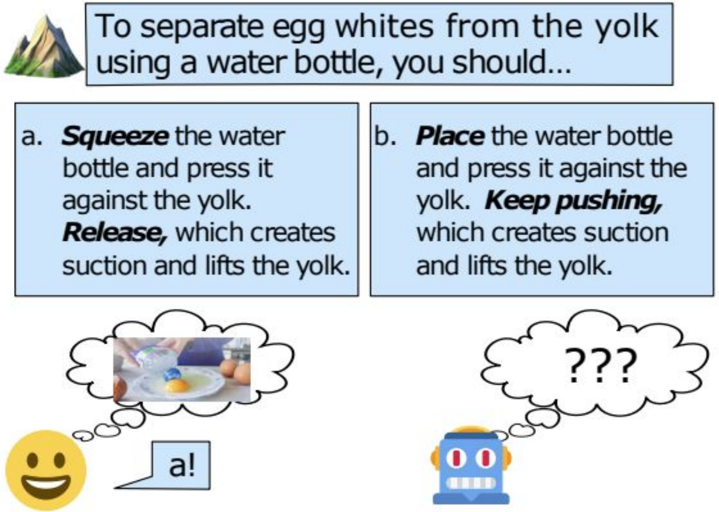


- If no algorithmic advance is made, it would take ***100k GPU hours*** to reach human performance on HellaSWAG!



PIQA (Bisk et al., 2020)

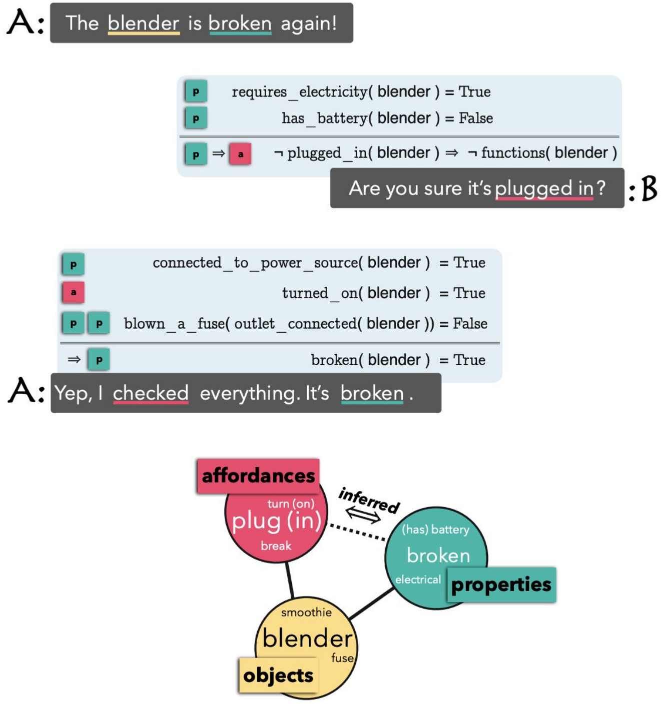
LMs lack an understanding of some of the most basic physical properties of the world.



Can you teach LMs commonsense?

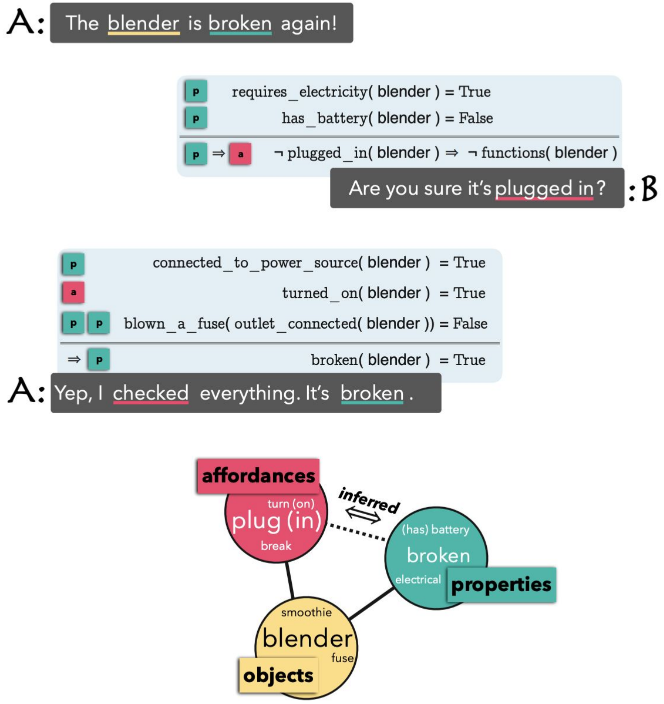
Do Neural Language Representations Learn Physical Commonsense?

Forbes et al. (2019): Fine-tune BERT to predict object properties ("uses electricity"), affordances ("plug in"), and the inferences between them (e.g. plug-in(x)⇒x uses electricity).



Do Neural Language Representations Learn Physical Commonsense?

Forbes et al. (2019): Fine-tune BERT to predict object properties ("uses electricity"), affordances ("plug in"), and the inferences between them (e.g. plug-in(x)⇒x uses electricity).



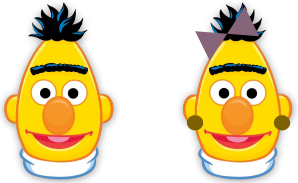
Best performance: functional properties (e.g. "uses electricity") given affordances.

Reasonable performance: encyclopedic (is an animal) and commonsense properties (comes in pairs).

Worst performance: perceptual properties (smooth) which are often not expressed by affordances

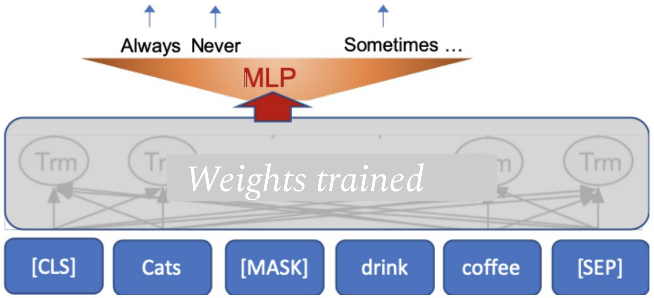
Can you teach LMs symbolic reasoning?

Talmor et al. (2019): oLMpics - testing BERT and RoBERTa on a set of symbolic reasoning tasks:



(RoBERTa)

answer choices



fine-tuning

Can you teach LMs symbolic reasoning?

Talmor et al. (2019): oLMpics - testing BERT and RoBERTa on a set of symbolic reasoning tasks:

Always-Never: A chicken [MASK] has horns. A. never B. rarely C. sometimes D. often E. always

Can you teach LMs symbolic reasoning?

Talmor et al. (2019): oLMpics - testing BERT and RoBERTa on a set of symbolic reasoning tasks:

Always-Never: A chicken [MASK] has horns. A. never B. rarely C. sometimes D. often E. always

Can you teach LMs symbolic reasoning?

Talmor et al. (2019): oLMpics - testing BERT and RoBERTa on a set of symbolic reasoning tasks:

Always-Never: A chicken [MASK] has horns. A. never B. rarely C. sometimes D. often E. always

Reporting bias: LMs are trained on texts describing things that ***do*** happen!

Can you teach LMs symbolic reasoning?

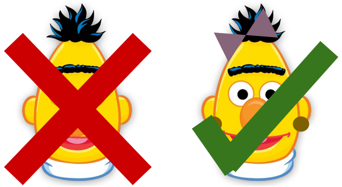
Talmor et al. (2019): oLMpics - testing BERT and RoBERTa on a set of symbolic reasoning tasks:

Age Comparison:
A 21 year old person age is [MASK] than a 35 year old person. A. younger B. older

Can you teach LMs symbolic reasoning?

Talmor et al. (2019): oLMpics - testing BERT and RoBERTa on a set of symbolic reasoning tasks:

Age Comparison:
A 21 year old person age is [MASK] than a 35 year old person. A. younger B. older

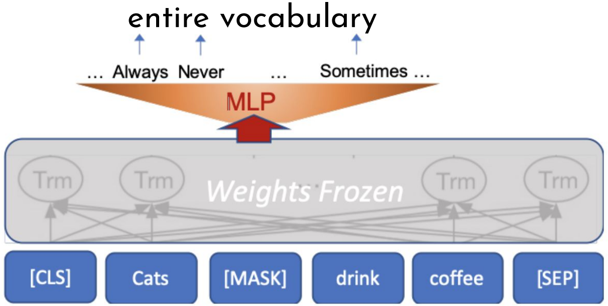


Can you teach LMs symbolic reasoning?

Talmor et al. (2019): oLMpics - testing BERT and RoBERTa on a set of symbolic reasoning tasks:

Age Comparison:
A 21 year old person age is [MASK] than a 35 year old person. A. younger B. older

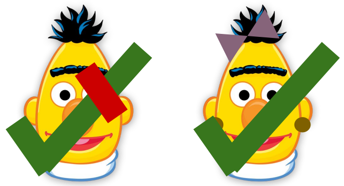
RoBERTa also performs well in a zero-shot setup:



Can you teach LMs symbolic reasoning?

Talmor et al. (2019): oLMpics - testing BERT and RoBERTa on a set of symbolic reasoning tasks:

Negation: It was [MASK] hot, it was really cold   A. really B. not



Can you teach LMs symbolic reasoning?

	RoBERTa-L	BERT-WWM	BERT-L	RoBERTa-B	BERT-B
ALWAYS-NEVER					
AGE COMPARISON	✓	✓			
OBJECTS COMPARISON	✓	✓✗			
ANTONYM NEGATION	✓		✓✗	✓✗	
PROPERTY CONJUNCTION		✓✗			
TAXONOMY CONJUNCTION	✓✗	✓✗		✓✗	
ENCYC. COMPOSITION					
MULTI-HOP COMPARISON					

Can you teach LMs symbolic reasoning?

RoBERTa > BERT

	RoBERTa-L	BERT-WWM	BERT-L	RoBERTa-B	BERT-B
Always-Never					
Age Comparison	✓	✓			
Objects Comparison	✓	✓✗			
Antonym Negation	✓		✓✗	✓✗	
Property Conjunction		✓✗			
Taxonomy Conjunction	✓✗	✓✗		✓✗	
Encyc. Composition					
Multi-hop Comparison					

Can you teach LMs symbolic reasoning?

RoBERTa > BERT

	RoBERTa-L	BERT-WWM	BERT-L	RoBERTa-B	BERT-B
Always-Never					
Age Comparison	✓	✓			
Objects Comparison	✓	✗			
Antonym Negation	✓		✗	✗	
Property Conjunction		✗			
Taxonomy Conjunction	✗	✗		✗	
Encyc. Composition					
Multi-hop Comparison					

Worse performance on compositionality tasks

Can you teach LMs symbolic reasoning?

RoBERTa > BERT

	RoBERTa-L	BERT-WWM	BERT-L	RoBERTa-B	BERT-B
ALWAYS-NEVER					
AGE COMPARISON	✓	✓			
OBJECTS COMPARISON	✓	✗			
ANTONYM NEGATION	✓		✗	✗	
PROPERTY CONJUNCTION		✗			
TAXONOMY CONJUNCTION	✗	✗		✗	
ENCYC. COMPOSITION					
MULTI-HOP COMPARISON					

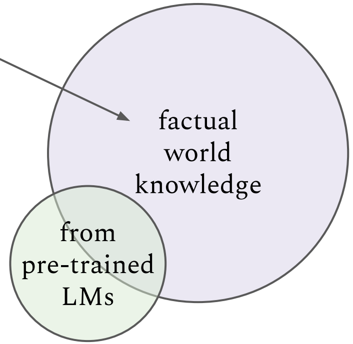
Worse performance on compositionality tasks

LMs are context-dependent and small changes to the input hurts their performance.

Summary

- Pre-trained language models some commonsense knowledge - but it is far from an exhaustive source.

Insufficient coverage (reporting bias; Gordon and Van Durme, 2013).



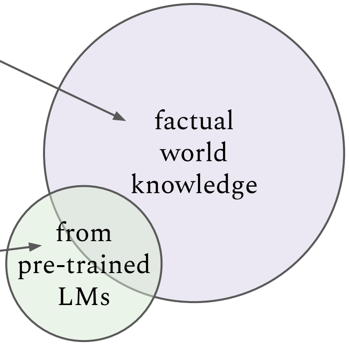
Summary

- Pre-trained language models some commonsense knowledge - but it is far from an exhaustive source.

- Use with caution! LMs also generate false facts.

Insufficient coverage
(reporting bias; Gordon and Van Durme, 2013).

Insufficient precision

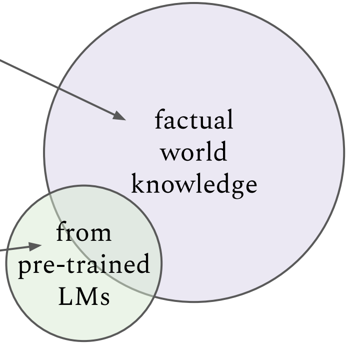


Summary

- Pre-trained language models some commonsense knowledge - but it is far from an exhaustive source.

- Use with caution! LMs also generate false facts.

Insufficient coverage
(reporting bias; Gordon and Van Durme, 2013).



Insufficient precision

Thank you!

vereds@allenai.org

References + Additional Reading

[1] Language Models as Knowledge Bases? Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H. Miller and Sebastian Riedel. EMNLP 2019.
[2] Commonsense Knowledge Mining from Pretrained Models. Joshua Feldman, Joe Davison, and Alexander M. Rush. EMNLP 2019.
[3] Barack's Wife Hillary: Using Knowledge Graphs for Fact-Aware Language Modeling. Robert Logan, Nelson F. Liu, Matthew E. Peters, Matt Gardner, and Sameer Singh. ACL 2019.
[4] Negated LAMA: Birds cannot fly. Nora Kassner and Hinrich Schütze. arXiv 2019.
[5] Do Neural Language Representations Learn Physical Commonsense? Maxwell Forbes, Ari Holtzman, and Yejin Choi. CogSci 2019.
[6] oLMpics -- On what Language Model Pre-training Captures. Alon Talmor, Yanai Elazar, Yoav Goldberg, and Jonathan Berant. arXiv 2019.
[7] On the Existence of Tacit Assumptions in Contextualized Language Models. Nathaniel Weir, Adam Poliak, Benjamin Van Durme. arXiv 2020.
[8] Deep Contextualized Word Representations. Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. NAACL 2018.
[9] BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. NAACL 2019.
[10] Roberta: A robustly optimized bert pretraining approach. Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. arXiv 2019.
[11] HellaSwag: Can a Machine Really Finish Your Sentence? Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. ACL 2019.
[12] PIQA: Reasoning about Physical Commonsense in Natural Language. Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, Yejin Choi. AAAI 2020.
[13] Unsupervised Commonsense Question Answering with Self-Talk. Vered Shwartz, Peter West, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. arXiv 2020.