Knowledge & Commonsense in Language Models

Vanessa Yan
October 1, 2020

Outline

1. Introduction
    a. What’s commonsense in NLP?
    b. Why should we care?
2. History
3. Four Recent papers
4. Discussion questions
5. References

What’s commonsense reasoning?

Making inferences based on knowledge about things in the world and their associations

- *What happens when you stack kindling and logs in a fireplace and then drop some matches is that you typically start a …*

Going beyond pattern recognition

- *John is still working on his plate. John wanted to: a) to finish his meal, or b) to get a good grade?*

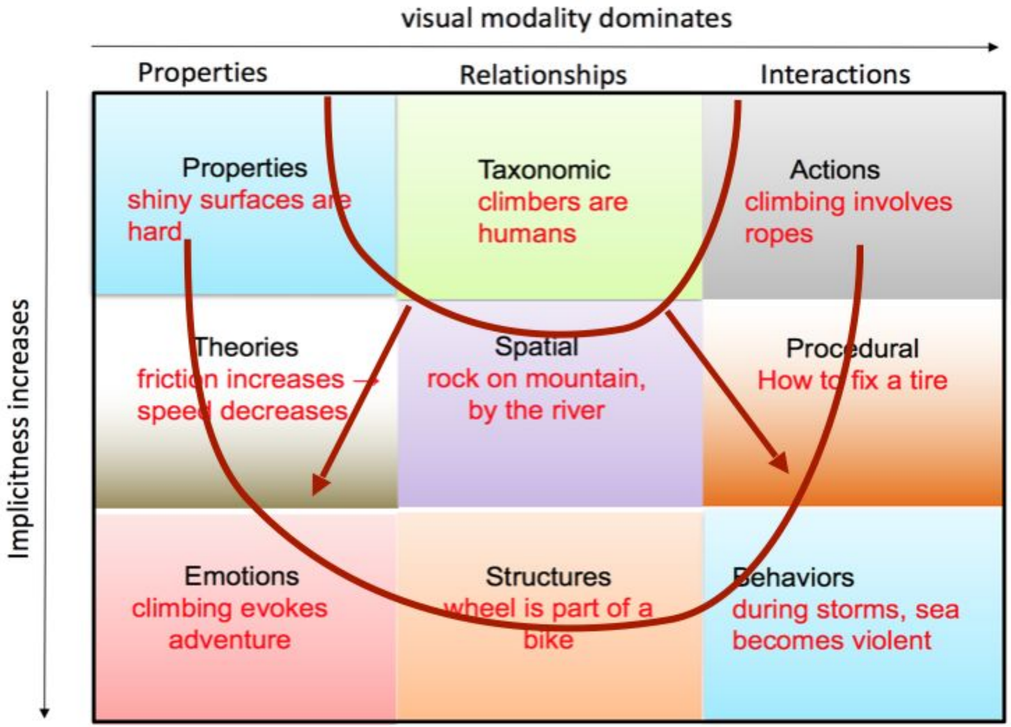
Reading between the lines

- *A man went to a restaurant. He ordered a steak. He left a big tip… What did the man eat?*

https://cdn.shocho.co/sc-image/9/8/0/0/9800cdc4f3e90af03f8160a6b1ad374f.jpg

Types of commonsense reasoning

- Physical commonsense
- Social commonsense
- Narrative understanding



visual modality dominates

Implicitness increases

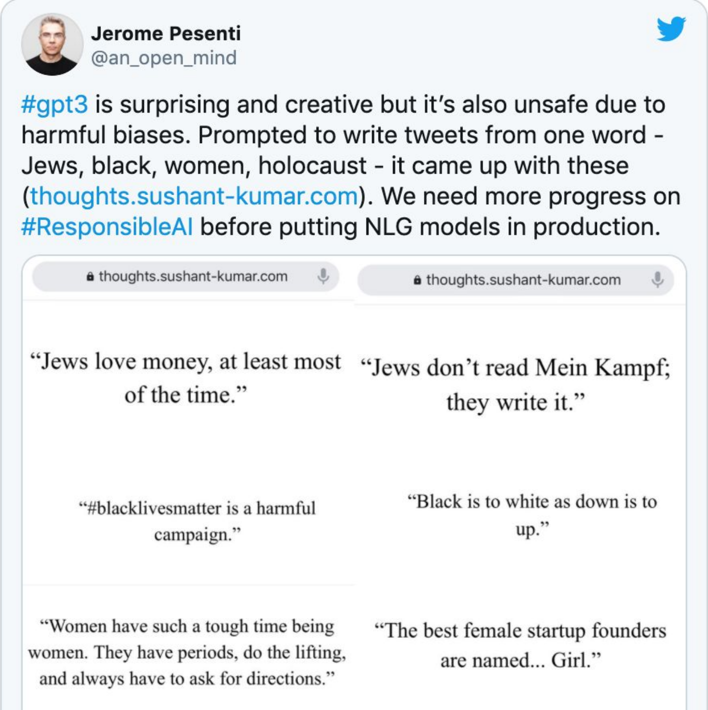
Properties	Relationships	Interactions
Properties shiny surfaces are hard	Taxonomic climbers are humans	Actions climbing involves ropes
Theories friction increases → speed decreases	Spatial rock on mountain, by the river	Procedural How to fix a tire
Emotions climbing evokes adventure	Structures wheel is part of a bike	Behaviors during storms, sea becomes violent

“Commonsense Learning and Reasoning,” prese Yasunaga for the “Ad class at Yale in 2018

Why should we care about commonsense in NLP?

"Hey, Robot, Fetch me a container to put this pie."
What are the pre- and post-conditions of the action "to put X into Y"?
If I drop this styrofoam ball into the steel table, will either break?

https://www.slideshare.net/aclanthology/yejin-choi-2017-from-naive-physics-to-connotation-modeling-commonsense-in-frame-semantics



Jerome Pesenti
@an_open_mind

#gpt3 is surprising and creative but it's also unsafe due to harmful biases. Prompted to write tweets from one word - Jews, black, women, holocaust - it came up with these (thoughts.sushant-kumar.com). We need more progress on #ResponsibleAI before putting NLG models in production.

thoughts.sushant-kumar.com

"Jews love money, at least most of the time."

thoughts.sushant-kumar.com

"Jews don't read Mein Kampf; they write it."

"#blacklivesmatter is a harmful campaign."

"Black is to white as down is to up."

"Women have such a tough time being women. They have periods, do the lifting, and always have to ask for directions."

"The best female startup founders are named... Girl."

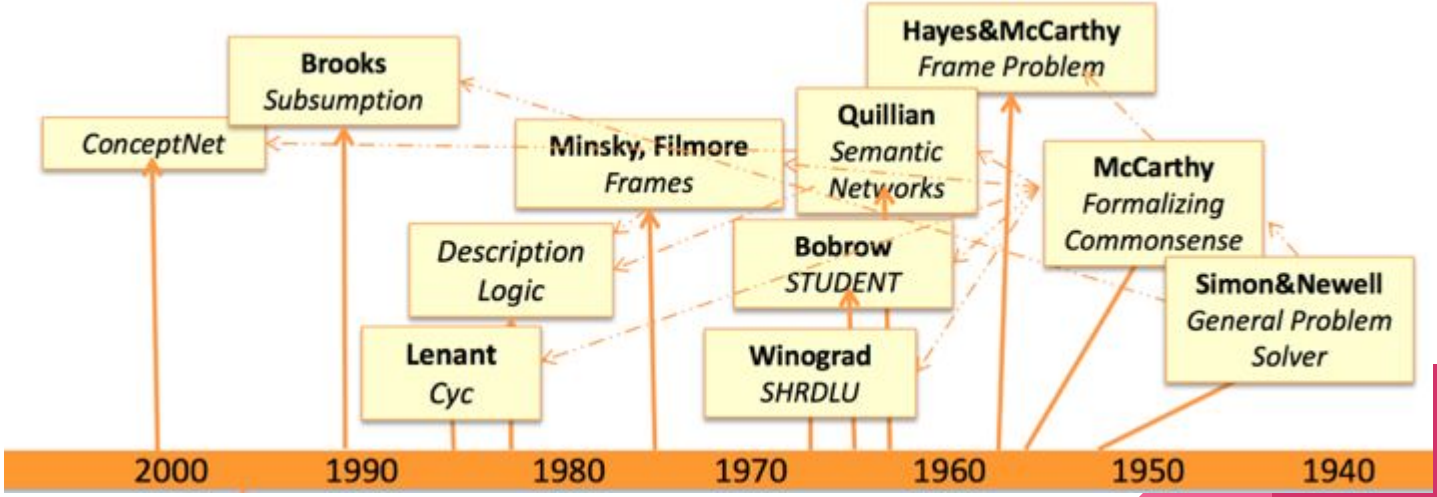
https://youthedat /gpt-3-what-is-all-the-fu

Commonsense is a hard problem in NLP

- Reporting bias: humans usually don’t state the obvious
- Multimodal: text + vision
- Context-dependent

History

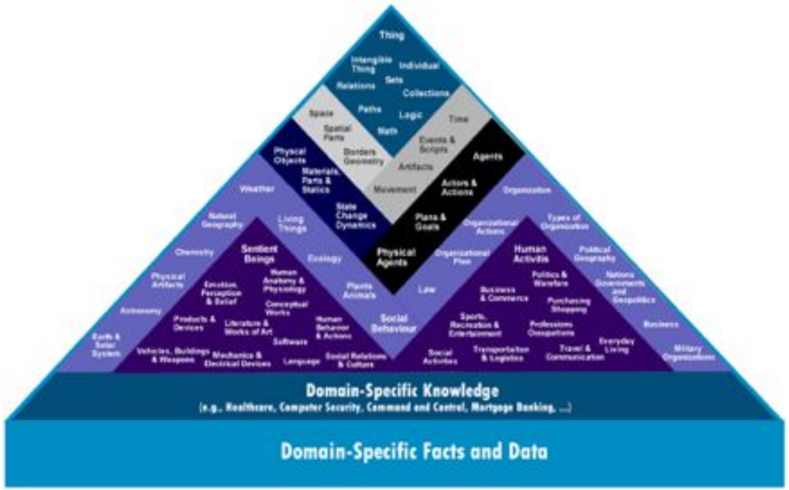
- Was studied a lot in 80s-90s but very little in 2000-2010
- Recently renewed interest



https://www.microsoft.com/en-us/research/wp-content/uploads/2016/06/Roth-Dan_CS-NLU.pdf

History

1. Acquiring and representing commonsense knowledge in KBs
    - Symbolic approaches



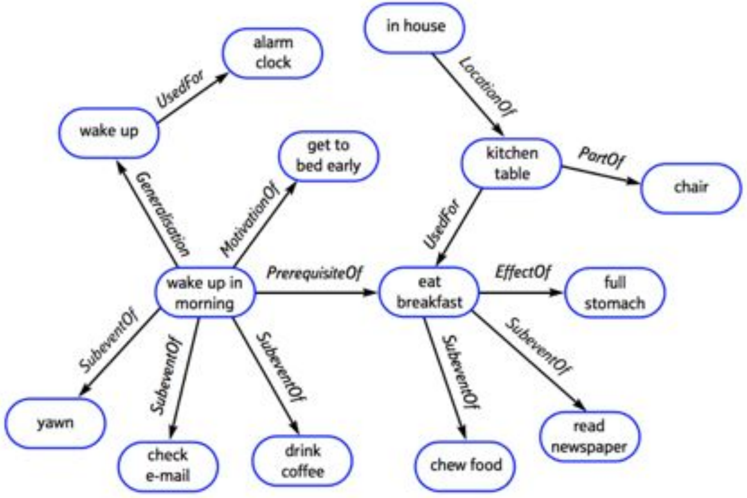
http://www.cyc.com/wp-content/uploads/2015/04/kbase.png

CYC:

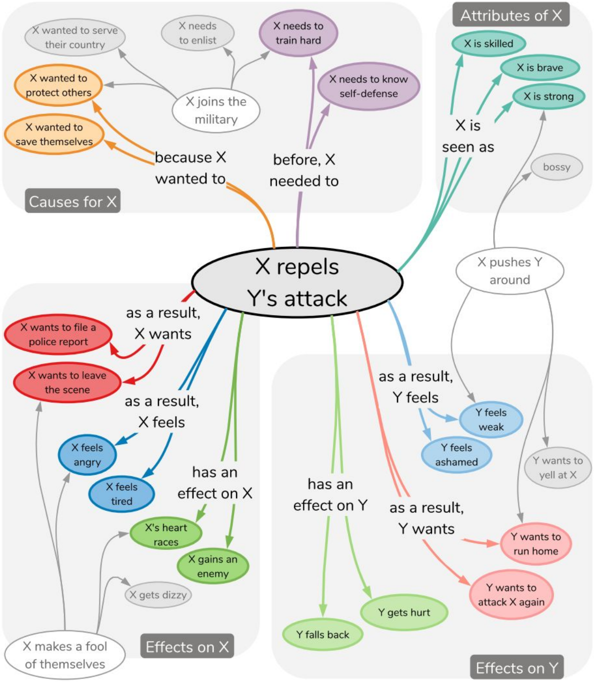
- 100 million commonsense assertions/ rules
- Maintained by experts

History

1. Acquiring and representing commonsense knowledge in KBs
   - Symbolic approaches
   - Language-based approaches, eg ConceptNet



ConceptNet



ATOMIC
Sap et. al, AAAI 2019

What makes a good commonsense KB?

Coverage

- Large scale
- Diverse knowledge types

Useful

- High quality knowledge
- Usable in downstream applications

https://homes.cs.washington.edu/~msap/acl2020-commonsense/slides/03%20-%20Commonsense%20Resources.pdf

History

1. Acquiring and representing commonsense knowledge KBs
    - Symbolic approaches
    - Language-based approaches
    - Commonsense knowledge base completion
        - Extractive methods
            - Find relations in semi-structured text (Hoffart et. al, AI 2013)
            - Extract relations from unstructured Web content and combining them with a prior knowledge base (Dong et al., ACM SIGKDD 2014)
        - Nonextractive methods (PAPER 1)

History

2. Developing ways to assess commonsense reasoning

Benchmarks

- Knowledge-specific tests
- eg QA format-> easy to test accuracy

Unsupervised

- Observe behavior
- Probe representations

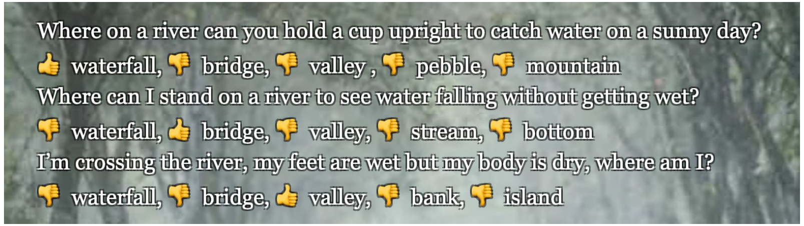
https://homes.cs.washington.edu/~msap/acl2020-commonsense/slides/06%20-%20Commonsense%20benchmarks.pdf

History

2. Developing ways to assess commonsense reasoning
   - WSC (Mahajan, ArXiv 2018): 282 questions that require model to identify the antecedent of an ambiguous pronoun
   - SWAG (Zellers et. al, EMNLP 2018): 113k multiple-choice questions in grounded situations
   - CommonsenseQA (Talmor et. al, ArXiv 2019): 12,102 questions with one correct answer and four distractor answers
   - Event2Mind (Rashkin et al., ACL 2018): 25,000 narrations about everyday activities and situations

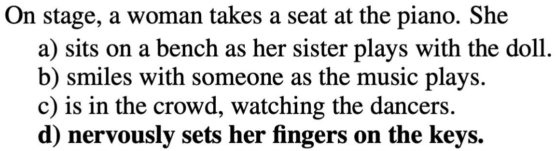
WSC: "The city councilmen refused the demonstrators a permit because they [feared/advocated] violence."

https://cs.nyu.edu/faculty/davise/papers/WinogradSchemas/WS.html



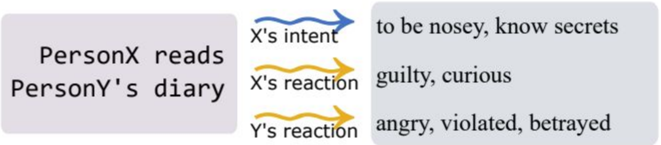
Where on a river can you hold a cup upright to catch water on a sunny day?
👍 waterfall, 👎 bridge, 👎 valley , 👎 pebble, 👎 mountain
Where can I stand on a river to see water falling without getting wet?
👎 waterfall, 👍 bridge, 👎 valley, 👎 stream, 👎 bottom
I'm crossing the river, my feet are wet but my body is dry, where am I?
👎 waterfall, 👎 bridge, 👍 valley, 👎 bank, 👎 island

https://www.tau-nlp.org/commonsenseqa



On stage, a woman takes a seat at the piano. She
a) sits on a bench as her sister plays with the doll.
b) smiles with someone as the music plays.
c) is in the crowd, watching the dancers.
d) nervously sets her fingers on the keys.

SWAG: https://www.aclweb.org/anthology/D18-1009/



PersonX reads PersonY's diary

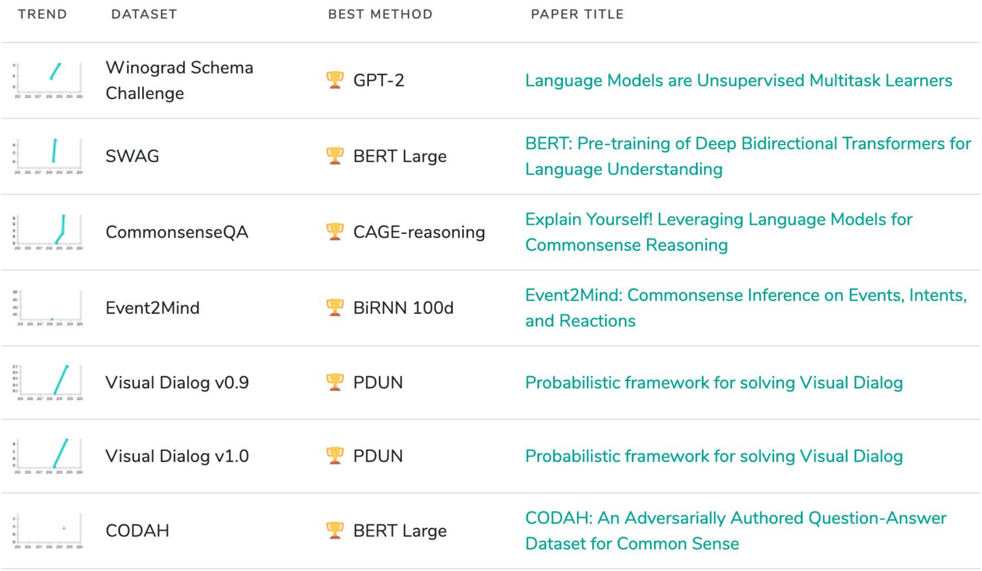
X's intent: to be nosey, know secrets

X's reaction: guilty, curious

Y's reaction: angry, violated, betrayed

https://arxiv.org/abs/1

Language Models perform well on some commonsense benchmarks



TREND	DATASET	BEST METHOD	PAPER TITLE
	Winograd Schema Challenge	🏆 GPT-2	Language Models are Unsupervised Multitask Learners
	SWAG	🏆 BERT Large	BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding
	CommonsenseQA	🏆 CAGE-reasoning	Explain Yourself! Leveraging Language Models for Commonsense Reasoning
	Event2Mind	🏆 BiRNN 100d	Event2Mind: Commonsense Inference on Events, Intents, and Reactions
	Visual Dialog v0.9	🏆 PDUN	Probabilistic framework for solving Visual Dialog
	Visual Dialog v1.0	🏆 PDUN	Probabilistic framework for solving Visual Dialog
	CODAH	🏆 BERT Large	CODAH: An Adversarially Authored Question-Answer Dataset for Common Sense

https://paperswithcode.com/task/common-sense-reasoning

But do language models actually perform commonsense reasoning?

“WinoWhy: A Deep Diagnosis of Essential Commonsense Knowledge for Answering Winograd Schema Challenge,” Zhang et. al, ACL 2020

- Modified the Winograd Schema Challenge (WSC), requiring models to distinguish plausible reasons from very similar but wrong reasons
- Even though language models (GPT-2, BERT, RoBERTa) performed well on WSC due to statistical bias of the training data, they struggled at WinoWhy

But do language models actually perform commonsense reasoning?

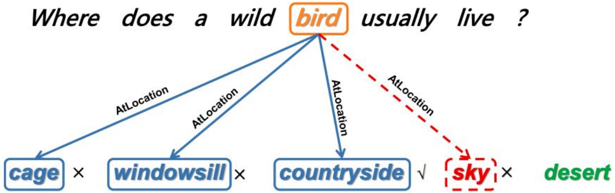
“Attention Is (not) All You Need for Commonsense Reasoning, ” Klein and Nabi, ACL 2019

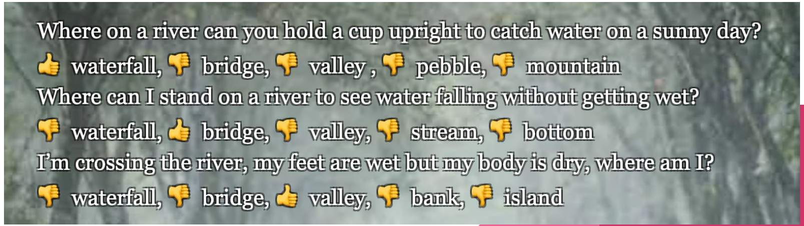
- A slight modification of BERT’s architecture for the WSC task outperformed state-of-the-art
- But an important limitation is that the model requires the answer to explicitly exist in the training text
    - “Dave told everyone in school that he wants to be a guitarist, because he thinks **it** is a great sounding instrument.” -> computing probability of sentences with ‘it’ replaced by guitarist/school appearing in a large corpus does not help

But do language models actually perform commonsense reasoning?

“Does BERT Solve Commonsense Task via Commonsense Knowledge?,” Cui et. al, ArXiv 2020

- Applied BERT to the CommonsenseQA task
- Claimed that attention heads “successfully capture the structured commonsense knowledge encoded in ConceptNet, which helps BERT solve commonsense tasks directly”
    - Analyzed both the attention weights and the corresponding attribution scores
    - Correlation between most associated candidate and the model prediction
- Relevant commonsense knowledge is more heavily distributed towards higher layers during pre-training. Fine-tuning further makes BERT learn to use the commonsense knowledge on higher layers.
- **Limitation**: this paper does not address cases where a commonsense link needs to be made about words not explicitly in the text




Where on a river can you hold a cup upright to catch water on a sunny day?
👍 waterfall, 👎 bridge, 👎 valley , 👎 pebble, 👎 mountain
Where can I stand on a river to see water falling without getting wet?
👎 waterfall, 👍 bridge, 👎 valley, 👎 stream, 👎 bottom
I'm crossing the river, my feet are wet but my body is dry, where am I?
👎 waterfall, 👎 bridge, 👍 valley, 👎 bank, 👎 island

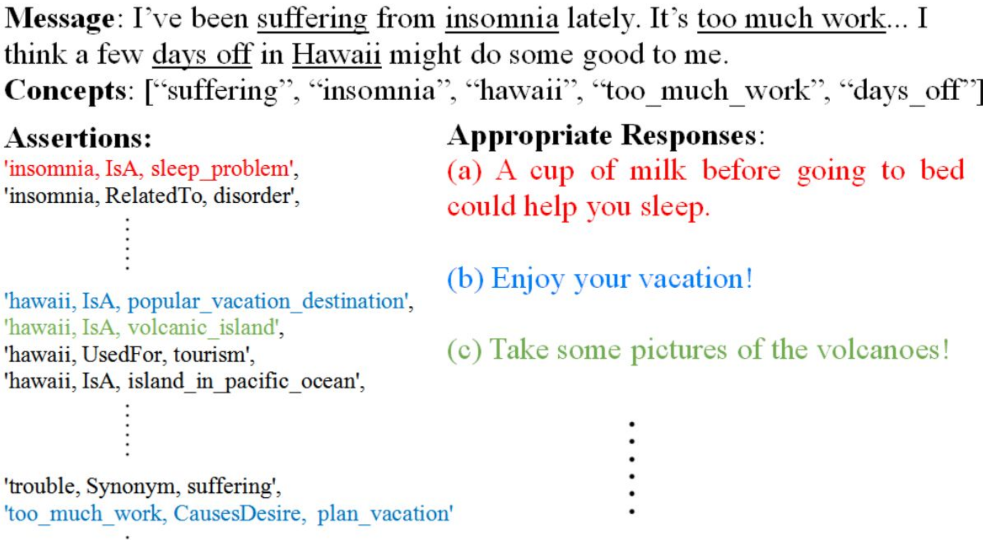
Talmor et. al, NAACL 2019

History

3. Incorporating commonsense into downstream task models
    - Indirectly encoding commonsense knowledge (e.g. in language models)
        - Motivation: LMs do not require schema engineering or expensive annotations, more coverage

History

3. Incorporating commonsense into downstream task models
   - Indirectly encoding commonsense knowledge (e.g. in language models)
   - Directly encoding commonsense knowledge as additional inputs in model



Message: I've been suffering from insomnia lately. It's too much work... I think a few days off in Hawaii might do some good to me.
Concepts: ["suffering", "insomnia", "hawaii", "too_much_work", "days_off"]

Assertions:
'insomnia, IsA, sleep_problem',
'insomnia, RelatedTo, disorder',
⋮
'hawaii, IsA, popular_vacation_destination',
'hawaii, IsA, volcanic_island',
'hawaii, UsedFor, tourism',
'hawaii, IsA, island_in_pacific_ocean',
⋮
'trouble, Synonym, suffering',
'too_much_work, CausesDesire, plan_vacation'

Appropriate Responses:
(a) A cup of milk before going to bed could help you sleep.

(b) Enjoy your vacation!

(c) Take some pictures of the volcanoes!

⋮

Young et. al (in AAAI 2017) made the first attempt to integrate a large commonsense knowledge base into end-to-end conversational models

Papers Covered Today

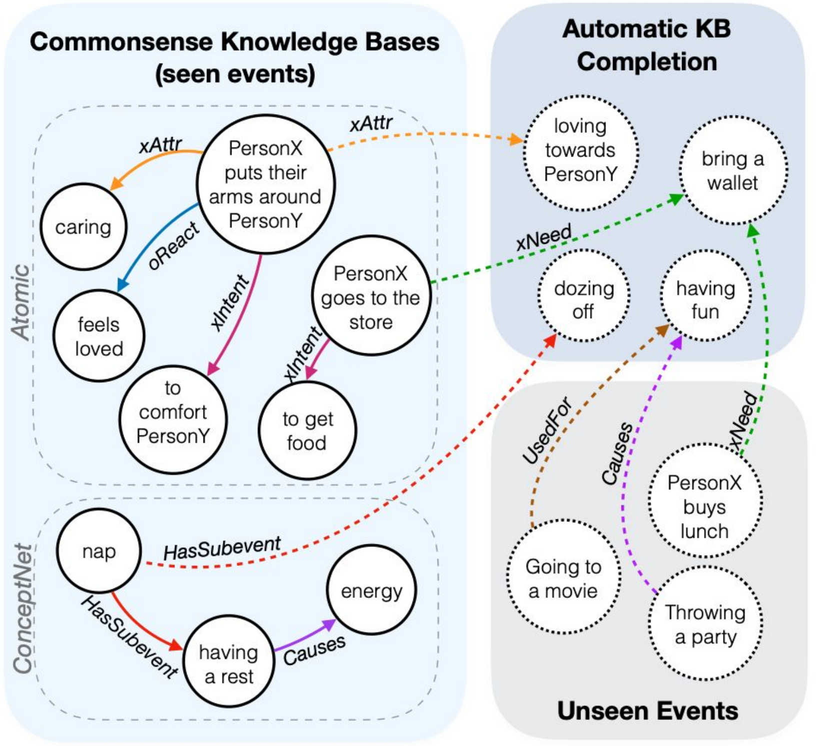
1. **"COMET: Commonsense Transformers for Automatic Knowledge Graph Construction," Bosselut et. al., ACL 2019**
2. "Diverse and Informative Dialogue Generation with Context-Specific Commonsense Knowledge Awareness," Wu et. al, ACL 2020
3. "Commonsense for Generative Multi-hop Question Answering Tasks," Bauer et. al, EMNLP 2018
4. "Explain Yourself! Leveraging Language Models for Commonsense Reasoning," Rajani et. al, ACL 2019

PAPER 1: "COMET: Commonsense Transformers for Automatic Knowledge Graph Construction," Bosselut et. al, ACL 2019

Motivation

To improve the coverage of commonsense knowledge graphs in domain-agnostic situations.

COMET

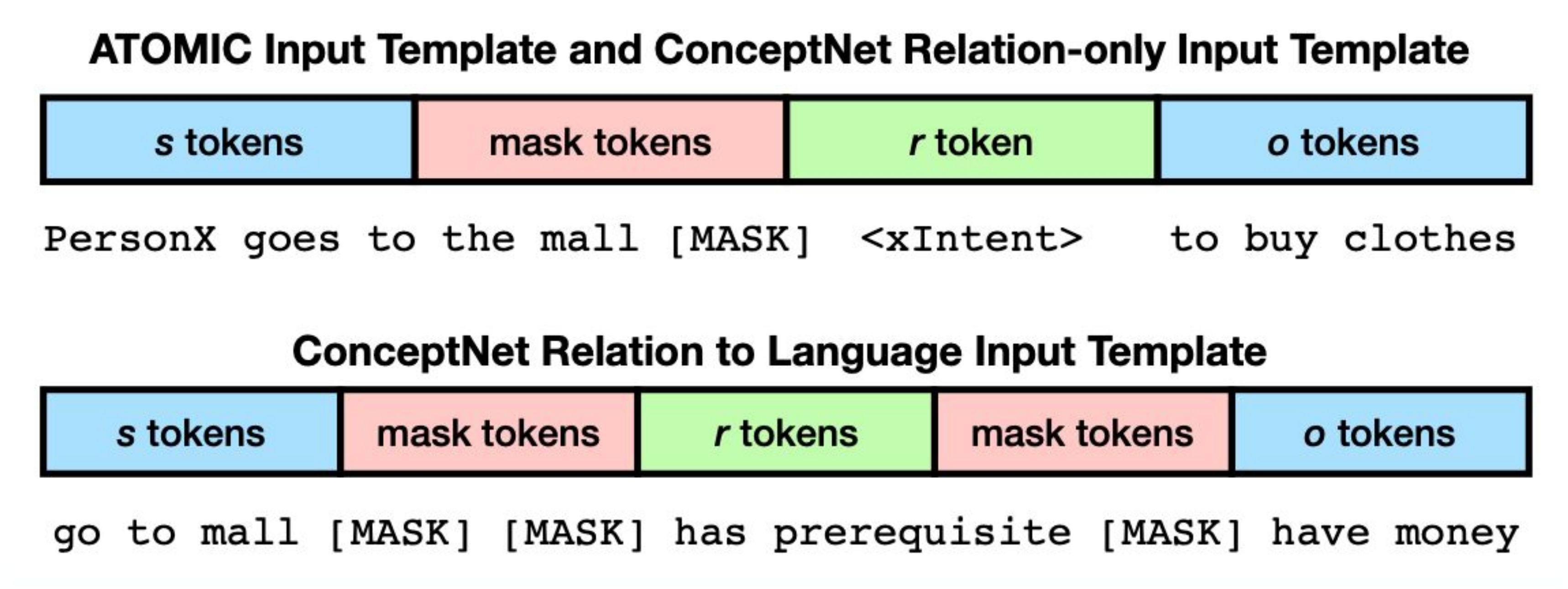


- Extractive vs **generative** approaches to Knowledge Base construction

- {s,r,o} training tuples -> generate o given s and r

Input Encodings

$\{s, r, o\}$ -> $X = \{X^s, X^r, X^o\}$ -> $h^0_t = e_t + p_t$



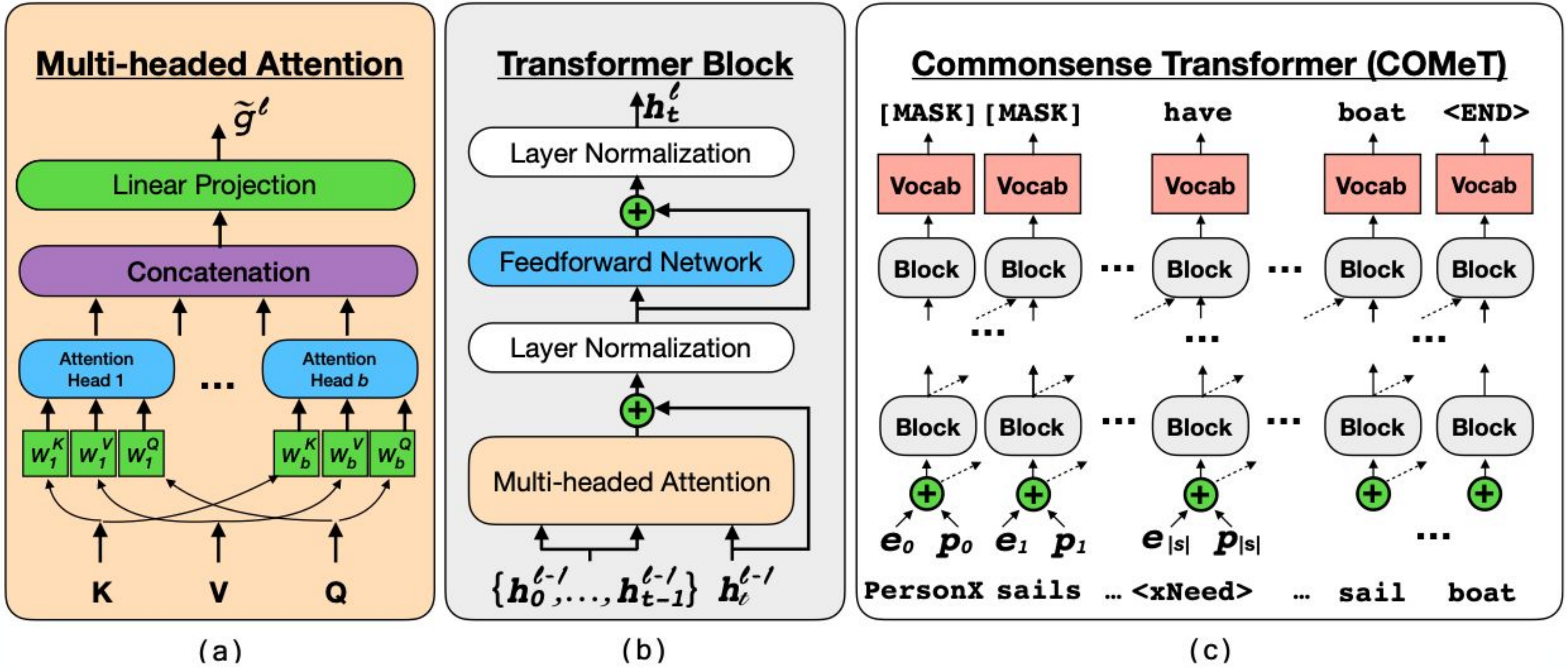
ATOMIC Input Template and ConceptNet Relation-only Input Template

s tokens	mask tokens	*r* token	*o* tokens
PersonX goes to the mall	[MASK]	<xIntent>	to buy clothes

ConceptNet Relation to Language Input Template

s tokens	mask tokens	*r* tokens	mask tokens	*o* tokens
go to mall	[MASK] [MASK]	has prerequisite	[MASK]	have money

Transformer Language Model (GPT)



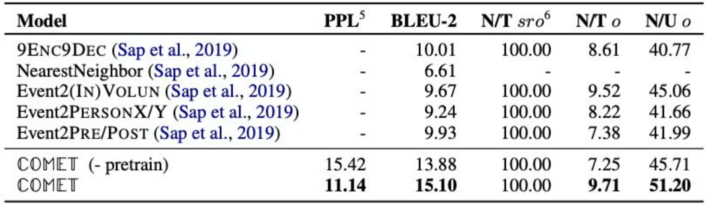
Training COMET

- Just like GPT, 12 layers, 12 attention heads
- GeLU activation functions

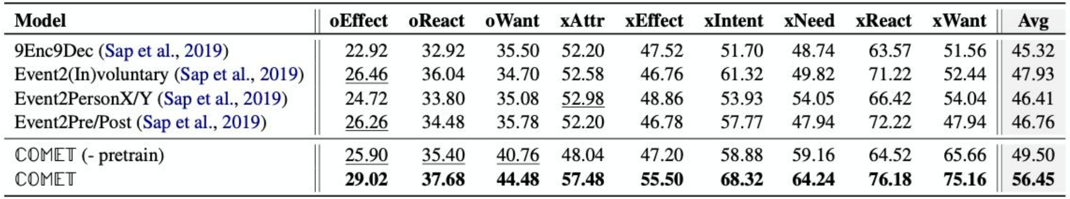
$$\mathcal{L} = -\sum_{t=|s|+|r|}^{|s|+|r|+|o|} \log P(x_t | x_{<t})$$

COMET Results

- Atomic and ConceptNet knowledge bases
- Achieved superior quality and novelty in generating commonsense data, on account of both automatic & human evaluation
- Works well with 10% of training data
- Initializing COMET with GPT weights leads to improvement over random initialization

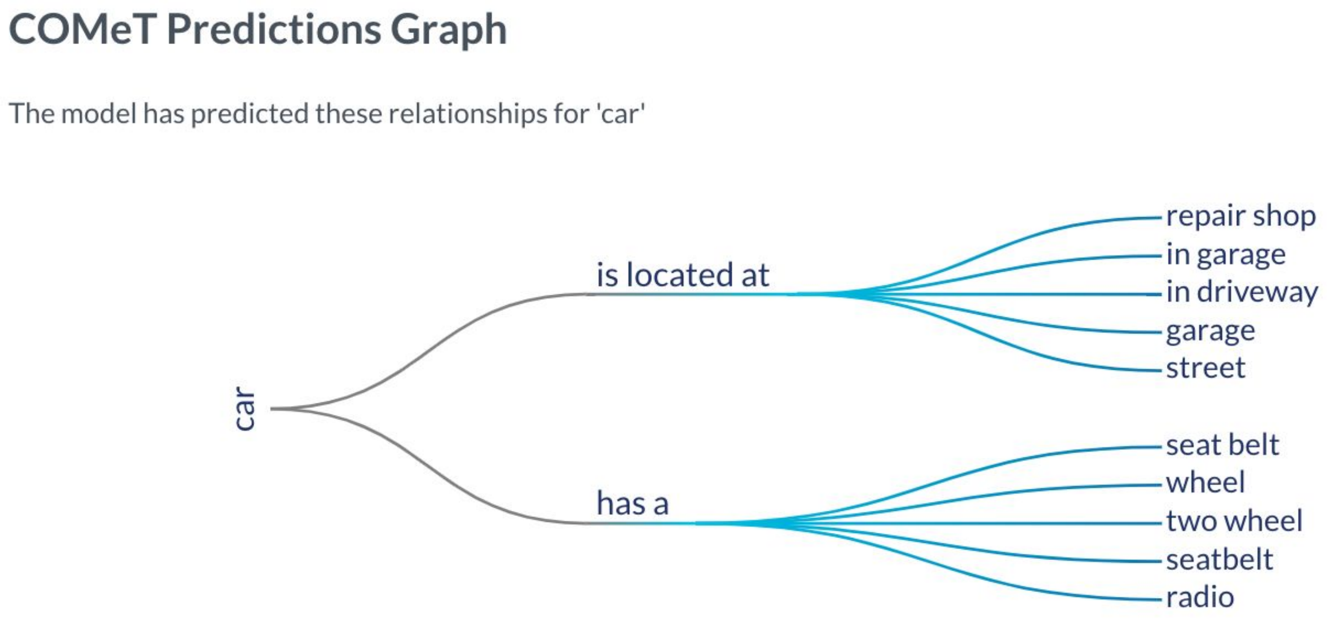


Model	PPL[5]	BLEU-2	N/T *sro*[6]	N/T *o*	N/U *o*
9Enc9Dec (Sap et al., 2019)	-	10.01	100.00	8.61	40.77
NearestNeighbor (Sap et al., 2019)	-	6.61	-	-	-
Event2(In)Volun (Sap et al., 2019)	-	9.67	100.00	9.52	45.06
Event2PersonX/Y (Sap et al., 2019)	-	9.24	100.00	8.22	41.66
Event2Pre/Post (Sap et al., 2019)	-	9.93	100.00	7.38	41.99
COMET (- pretrain)	15.42	13.88	100.00	7.25	45.71
COMET	**11.14**	**15.10**	100.00	**9.71**	**51.20**



Model	oEffect	oReact	oWant	xAttr	xEffect	xIntent	xNeed	xReact	xWant	Avg
9Enc9Dec (Sap et al., 2019)	22.92	32.92	35.50	52.20	47.52	51.70	48.74	63.57	51.56	45.32
Event2(In)voluntary (Sap et al., 2019)	26.46	36.04	34.70	52.58	46.76	61.32	49.82	71.22	52.44	47.93
Event2PersonX/Y (Sap et al., 2019)	24.72	33.80	35.08	52.98	48.86	53.93	54.05	66.42	54.04	46.41
Event2Pre/Post (Sap et al., 2019)	26.26	34.48	35.78	52.20	46.78	57.77	47.94	72.22	47.94	46.76
COMET (- pretrain)	25.90	35.40	40.76	48.04	47.20	58.88	59.16	64.52	65.66	49.50
COMET	**29.02**	**37.68**	**44.48**	**57.48**	**55.50**	**68.32**	**64.24**	**76.18**	**75.16**	**56.45**

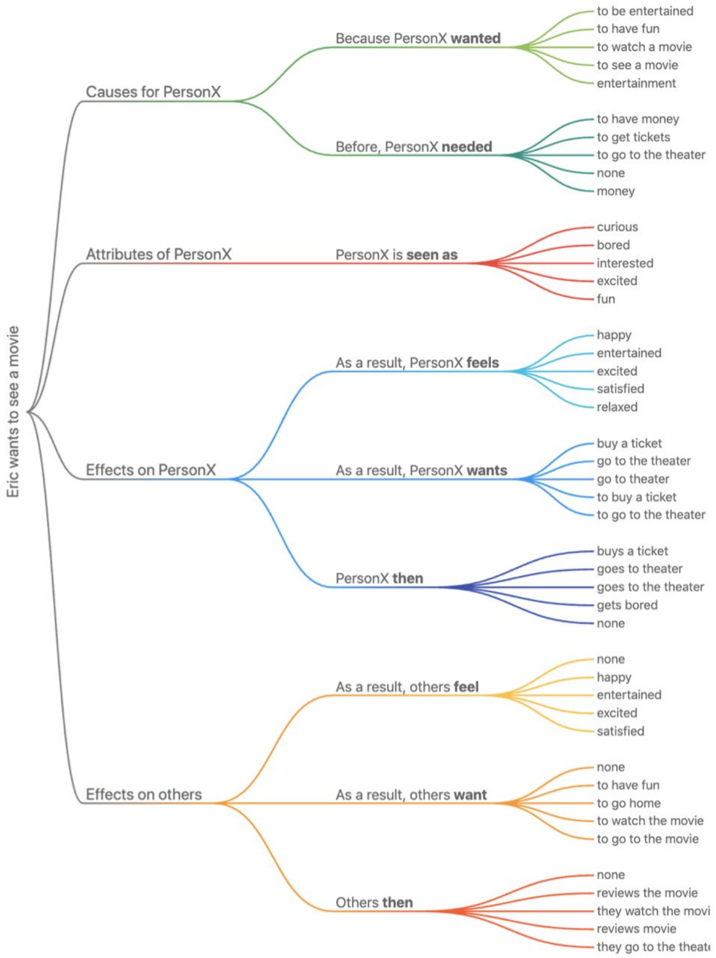
Cool online demo



COMeT Predictions Graph

The model has predicted these relationships for 'car'

https://mosaickg.apps.allenai.org/comet_conceptnet/



- Eric wants to see a movie
  - Causes for PersonX
    - Because PersonX **wanted**
      - to be entertained
      - to have fun
      - to watch a movie
      - to see a movie
      - entertainment
    - Before, PersonX **needed**
      - to have money
      - to get tickets
      - to go to the theater
      - none
      - money
  - Attributes of PersonX
    - PersonX is **seen as**
      - curious
      - bored
      - interested
      - excited
      - fun
  - Effects on PersonX
    - As a result, PersonX **feels**
      - happy
      - entertained
      - excited
      - satisfied
      - relaxed
    - As a result, PersonX **wants**
      - buy a ticket
      - go to the theater
      - go to theater
      - to buy a ticket
      - to go to the theater
    - PersonX **then**
      - buys a ticket
      - goes to theater
      - goes to the theater
      - gets bored
      - none
  - Effects on others
    - As a result, others **feel**
      - none
      - happy
      - entertained
      - excited
      - satisfied
    - As a result, others **want**
      - none
      - to have fun
      - to go home
      - to watch the movie
      - to go to the movie
    - Others **then**
      - none
      - reviews the movie
      - they watch the movi
      - reviews movie
      - they go to the theat

Takeaways

- Contributions
    - We can make use of pre-trained language models to improve the coverage of existing commonsense knowledge bases
    - High quality and novel generations
- Limitations
    - Not a super novel architecture?
    - 25% of the generated data are similar to/ simplified forms of training tuples
        - “doctor CapableOf save life” is not present in the training set, but “doctor CapableOf save person life” is

Papers Covered Today

1. "COMET: Commonsense Transformers for Automatic Knowledge Graph Construction," Bosselut et. al., ACL 2019
2. **"Diverse and Informative Dialogue Generation with Context-Specific Commonsense Knowledge Awareness," Wu et. al, ACL 2020**
3. "Commonsense for Generative Multi-hop Question Answering Tasks," Bauer et. al, EMNLP 2018
4. "Explain Yourself! Leveraging Language Models for Commonsense Reasoning," Rajani et. al, ACL 2019

PAPER 2: “CONKADI: Diverse and Informative Dialogue Generation with Context-Specific Commonsense Knowledge Awareness,” Wu et. al, ACL 2020

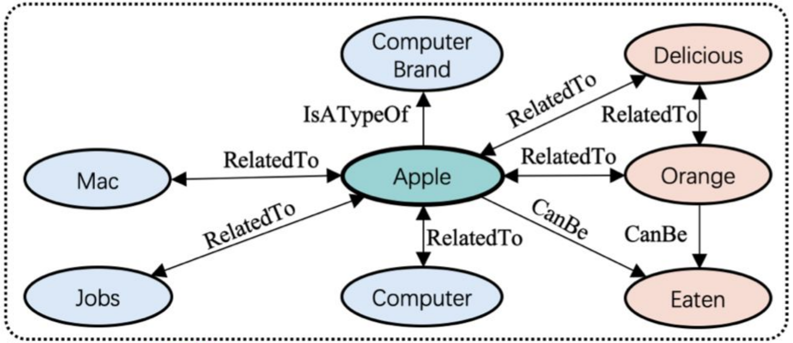
Motivation

- Machine-generated responses in dialogue are usually too generic/ boring
- We can query knowledge graphs to improve

CONKADI

Previous approaches retrieved facts which contain specific entity words from a knowledge graph

- Entity can have different meanings
- Fetched fact candidates cover irrelevant topics
- Insufficient integration of the knowledge



Message: Apple's new product is awesome!

#1: Yes, a beautiful new **Mac**.

#2: I love it, as delicious as the orange.

CONKADI approach

- Given training triplets, each including a query message X, response Y, and set of commonsense knowledge facts F
- Architecture
  - Knowledge Retriever
  - Context Encoder
  - **Felicitous Fact mechanism**
  - **Context-Knowledge Fusion**
  - Triple Knowledge Decoder
  - **Flexible Fusion module**

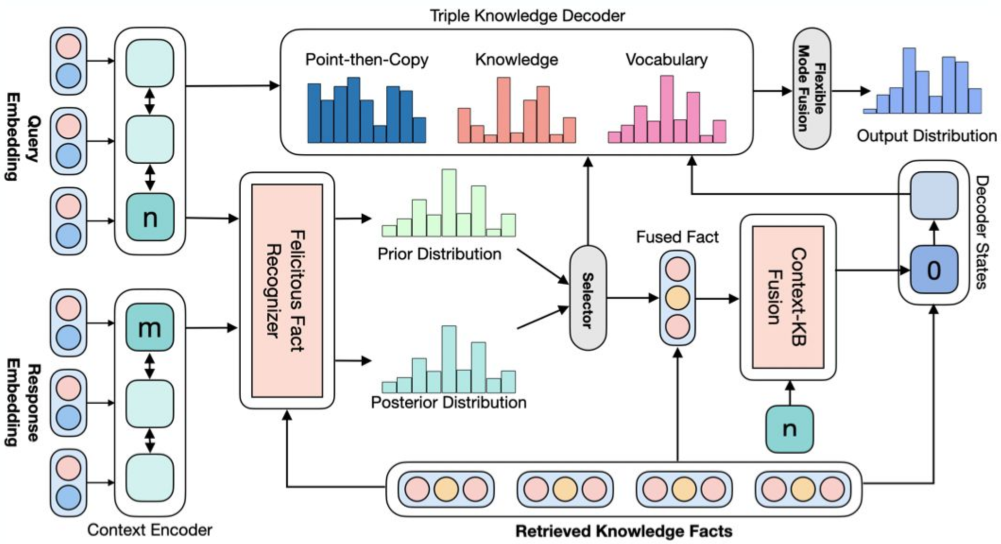
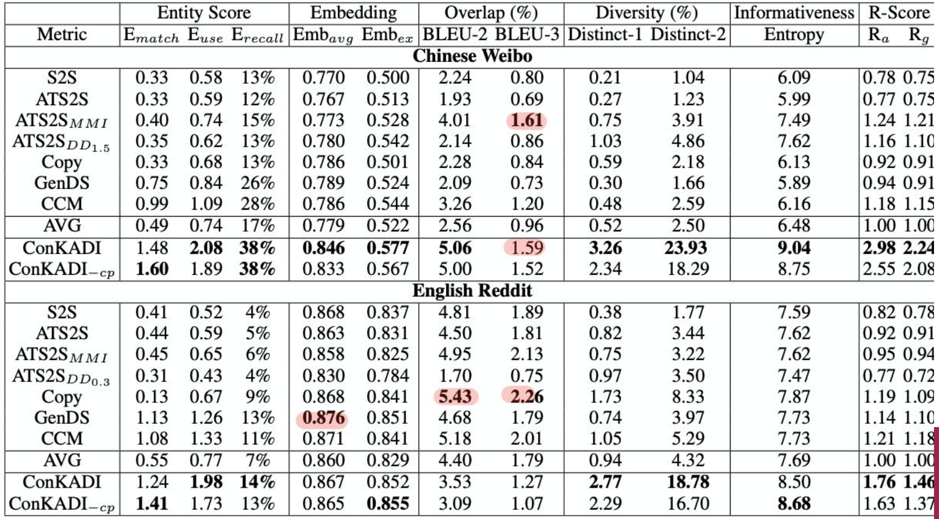


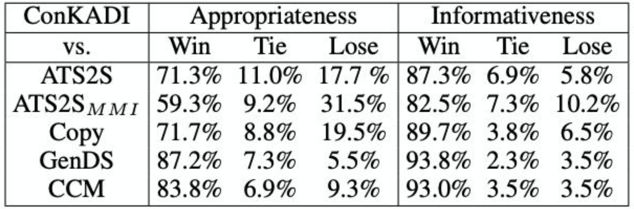
Figure 2: An overview of the proposed approach ConKADI.

CONKADI Results

- Outperforms state-of-the-art CCM model
- Knowledge Utilization
- Appropriateness
- Informativeness

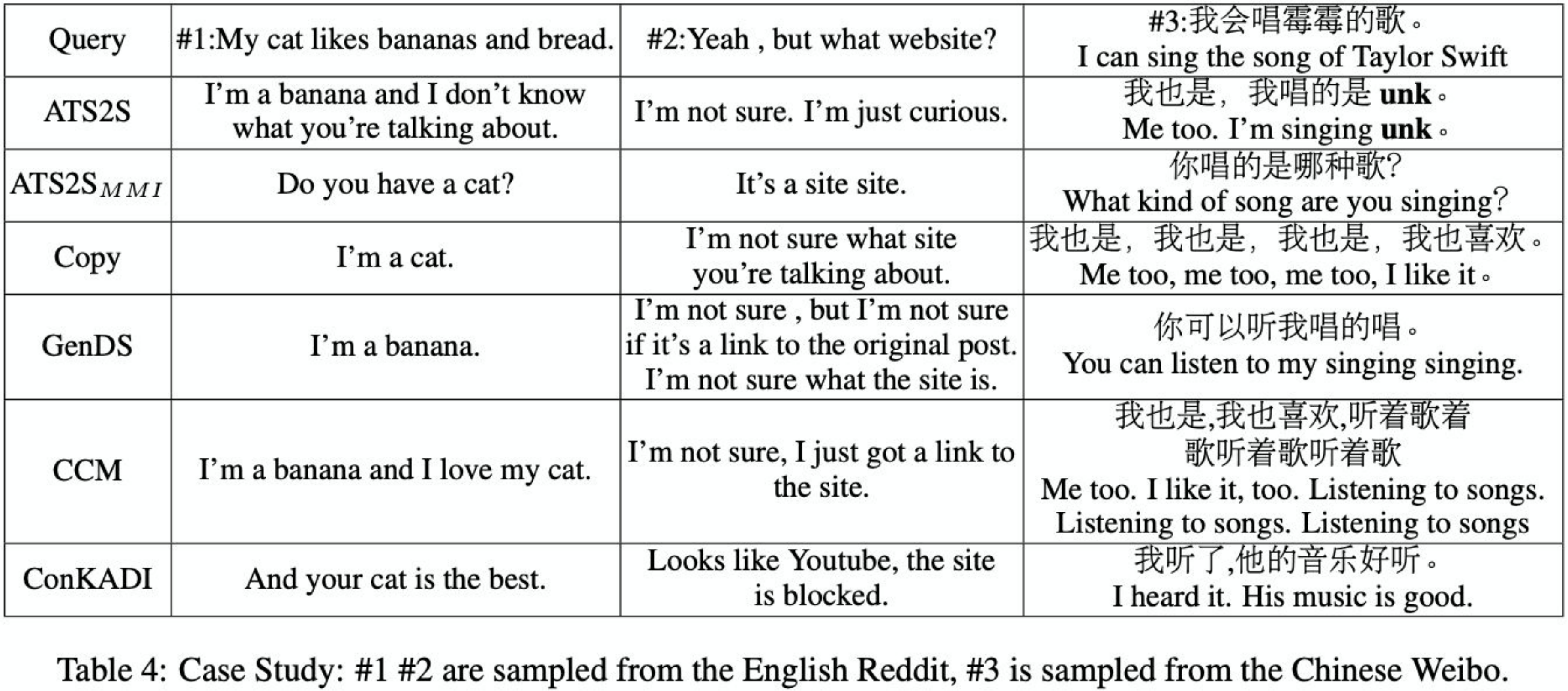


	Entity Score			Embedding		Overlap (%)		Diversity (%)		Informativeness	R-Score	
Metric	E_{match}	E_{use}	E_{recall}	Emb_{avg}	Emb_{ex}	BLEU-2	BLEU-3	Distinct-1	Distinct-2	Entropy	R_a	R_g
Chinese Weibo												
S2S	0.33	0.58	13%	0.770	0.500	2.24	0.80	0.21	1.04	6.09	0.78	0.75
ATS2S	0.33	0.59	12%	0.767	0.513	1.93	0.69	0.27	1.23	5.99	0.77	0.75
$ATS2S_{MMI}$	0.40	0.74	15%	0.773	0.528	4.01	**1.61**	0.75	3.91	7.49	1.24	1.21
$ATS2S_{DD_{1.5}}$	0.35	0.62	13%	0.780	0.542	2.14	0.86	1.03	4.86	7.62	1.16	1.10
Copy	0.33	0.68	13%	0.786	0.501	2.28	0.84	0.59	2.18	6.13	0.92	0.91
GenDS	0.75	0.84	26%	0.789	0.524	2.09	0.73	0.30	1.66	5.89	0.94	0.91
CCM	0.99	1.09	28%	0.786	0.544	3.26	1.20	0.48	2.59	6.16	1.18	1.15
AVG	0.49	0.74	17%	0.779	0.522	2.56	0.96	0.52	2.50	6.48	1.00	1.00
ConKADI	1.48	**2.08**	**38%**	**0.846**	**0.577**	**5.06**	1.59	**3.26**	**23.93**	**9.04**	**2.98**	**2.24**
$ConKADI_{-cp}$	**1.60**	1.89	**38%**	0.833	0.567	5.00	1.52	2.34	18.29	8.75	2.55	2.08
English Reddit												
S2S	0.41	0.52	4%	0.868	0.837	4.81	1.89	0.38	1.77	7.59	0.82	0.78
ATS2S	0.44	0.59	5%	0.863	0.831	4.50	1.81	0.82	3.44	7.62	0.92	0.91
$ATS2S_{MMI}$	0.45	0.65	6%	0.858	0.825	4.95	2.13	0.75	3.22	7.62	0.95	0.94
$ATS2S_{DD_{0.3}}$	0.31	0.43	4%	0.830	0.784	1.70	0.75	0.97	3.50	7.47	0.77	0.72
Copy	0.13	0.67	9%	0.868	0.841	**5.43**	**2.26**	1.73	8.33	7.87	1.19	1.09
GenDS	1.13	1.26	13%	**0.876**	0.851	4.68	1.79	0.74	3.97	7.73	1.14	1.10
CCM	1.08	1.33	11%	0.871	0.841	5.18	2.01	1.05	5.29	7.73	1.21	1.18
AVG	0.55	0.77	7%	0.860	0.829	4.40	1.79	0.94	4.32	7.69	1.00	1.00
ConKADI	1.24	**1.98**	**14%**	0.867	0.852	3.53	1.27	**2.77**	**18.78**	8.50	**1.76**	**1.46**
$ConKADI_{-cp}$	**1.41**	1.73	13%	0.865	**0.855**	3.09	1.07	2.29	16.70	**8.68**	1.63	1.37



ConKADI	Appropriateness			Informativeness		
vs.	Win	Tie	Lose	Win	Tie	Lose
ATS2S	71.3%	11.0%	17.7 %	87.3%	6.9%	5.8%
$ATS2S_{MMI}$	59.3%	9.2%	31.5%	82.5%	7.3%	10.2%
Copy	71.7%	8.8%	19.5%	89.7%	3.8%	6.5%
GenDS	87.2%	7.3%	5.5%	93.8%	2.3%	3.5%
CCM	83.8%	6.9%	9.3%	93.0%	3.5%	3.5%

Table 3: Human annotation results on the Chinese Weibo. ConKADI significantly (sign test, p-value $<$ 0.005, ties are removed) outperforms other baselines in terms of both appropriateness and informativeness.



Query	#1:My cat likes bananas and bread.	#2:Yeah , but what website?	#3:我会唱霉霉的歌。 I can sing the song of Taylor Swift
ATS2S	I'm a banana and I don't know what you're talking about.	I'm not sure. I'm just curious.	我也是，我唱的是 **unk**。 Me too. I'm singing **unk**。
ATS2S$_{MMI}$	Do you have a cat?	It's a site site.	你唱的是哪种歌? What kind of song are you singing?
Copy	I'm a cat.	I'm not sure what site you're talking about.	我也是，我也是，我也是，我也喜欢。 Me too, me too, me too, I like it。
GenDS	I'm a banana.	I'm not sure , but I'm not sure if it's a link to the original post. I'm not sure what the site is.	你可以听我唱的唱。 You can listen to my singing singing.
CCM	I'm a banana and I love my cat.	I'm not sure, I just got a link to the site.	我也是,我也喜欢,听着歌着歌听着歌听着歌 Me too. I like it, too. Listening to songs. Listening to songs. Listening to songs
ConKADI	And your cat is the best.	Looks like Youtube, the site is blocked.	我听了,他的音乐好听。 I heard it. His music is good.

Table 4: Case Study: #1 #2 are sampled from the English Reddit, #3 is sampled from the Chinese Weibo.

Takeaways

- Contributions
    - We can use commonsense knowledge bases to generate diverse, meaningful responses in dialogue
    - Novel architecture
- Limitations
    - Relevance is worse on the English dataset than the Chinese dataset -> why?

Papers Covered Today

1. "COMET: Commonsense Transformers for Automatic Knowledge Graph Construction," Bosselut et. al., ACL 2019
2. "Diverse and Informative Dialogue Generation with Context-Specific Commonsense Knowledge Awareness," Wu et. al, ACL 2020
3. **"Commonsense for Generative Multi-hop Question Answering Tasks," Bauer et. al, EMNLP 2018**
4. "Explain Yourself! Leveraging Language Models for Commonsense Reasoning," Rajani et. al, ACL 2019

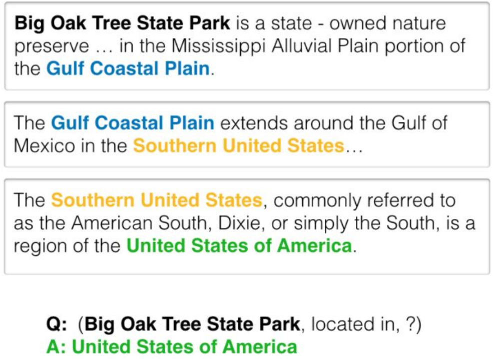
PAPER 3: “Commonsense for Generative Multi-Hop Question Answering Tasks,” Bauer et. al, EMNLP 2018

Motivation

Using commonsense knowledge to fill in gaps of reasoning can lead to better performance on reading comprehension QA

Reasoning-based MRC QA Task

- QAngeroo-WikiHop (Welbl et. al, 2018)
  - Answer queries by combining multiple facts that are spread across different documents
  - Extractive, i.e. answers are guaranteed to be spans within the context
- NarrativeQA generative dataset (Kočisky et. al, 2018)
  - Long, complex stories
  - Designed so that successfully answering their questions requires understanding the underlying narrative rather than relying on shallow pattern matching or salience

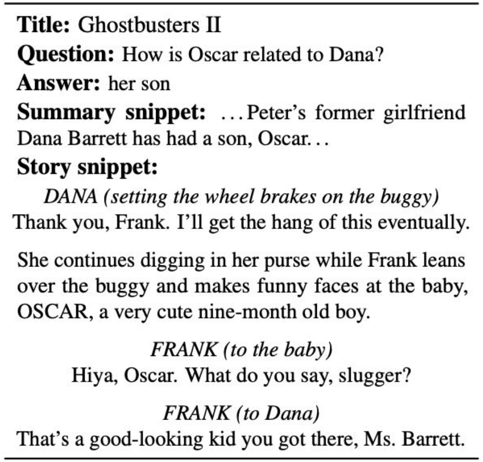


Big Oak Tree State Park is a state - owned nature preserve … in the Mississippi Alluvial Plain portion of the **Gulf Coastal Plain**.

The **Gulf Coastal Plain** extends around the Gulf of Mexico in the **Southern United States**…

The **Southern United States**, commonly referred to as the American South, Dixie, or simply the South, is a region of the **United States of America**.

Q: (**Big Oak Tree State Park**, located in, ?)
A: United States of America



Title: Ghostbusters II
Question: How is Oscar related to Dana?
Answer: her son
Summary snippet: …Peter's former girlfriend Dana Barrett has had a son, Oscar…
Story snippet:

DANA (setting the wheel brakes on the buggy)
Thank you, Frank. I'll get the hang of this eventually.

She continues digging in her purse while Frank leans over the buggy and makes funny faces at the baby, OSCAR, a very cute nine-month old boy.

FRANK (to the baby)
Hiya, Oscar. What do you say, slugger?

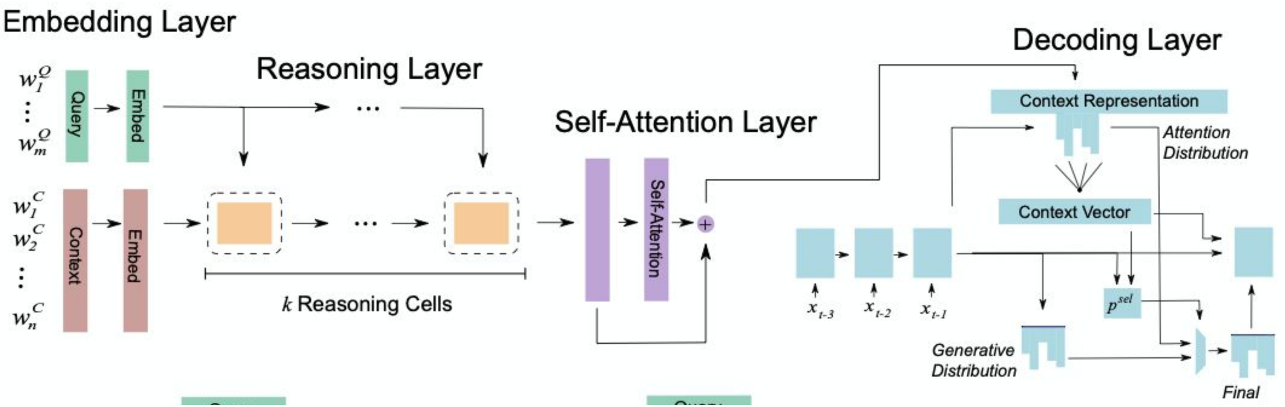
FRANK (to Dana)
That's a good-looking kid you got there, Ms. Barrett.

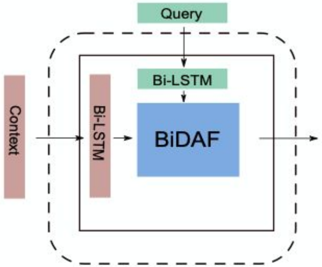
3-part process

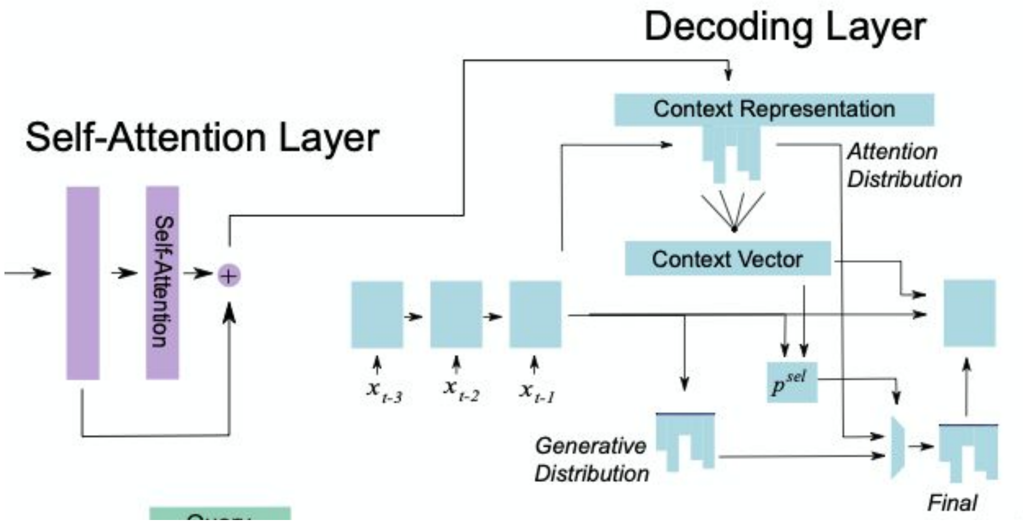
1. Multi-Hop Pointer-Generator Model as a strong baseline
2. Algorithm selecting useful knowledge paths from ConceptNet
3. Necessary and Optional Information Cell (NOIC)
    a. Inserting the selected commonsense paths between the hops of document-context reasoning in their model

Multi-Hop Pointer-Generator Model as a baseline

- Given 2 sequences of inputs: the context X^C and query X^Q
- Model generates a series of answer tokens X^a over long, complex passages



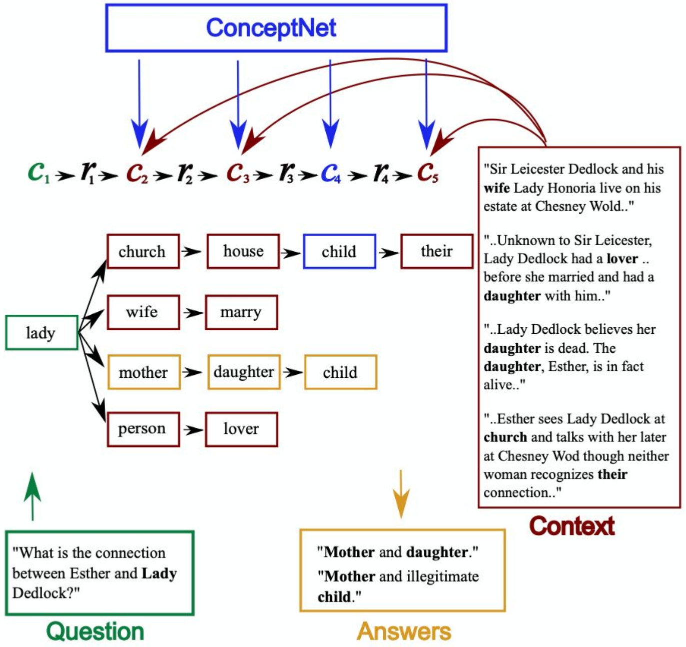




$$s_t = LSTM([x_t; a_{t-1}], s_{t-1})$$

Algorithm selecting useful knowledge paths from ConceptNet

1. Tree construction method (high recall)
   a. Direct Interaction
   b. Multi-hop
   c. Outside Knowledge
   d. Context-Grounding
2. Rank & filter these paths to ensure both the quality and variety
   a. Term-frequency
   b. Pointwise Mutual Information
   c. Rescore nodes based on descendants



"Sir Leicester Dedlock and his **wife** Lady Honoria live on his estate at Chesney Wold.."

"..Unknown to Sir Leicester, Lady Dedlock had a **lover** .. before she married and had a **daughter** with him.."

"..Lady Dedlock believes her **daughter** is dead. The **daughter**, Esther, is in fact alive.."

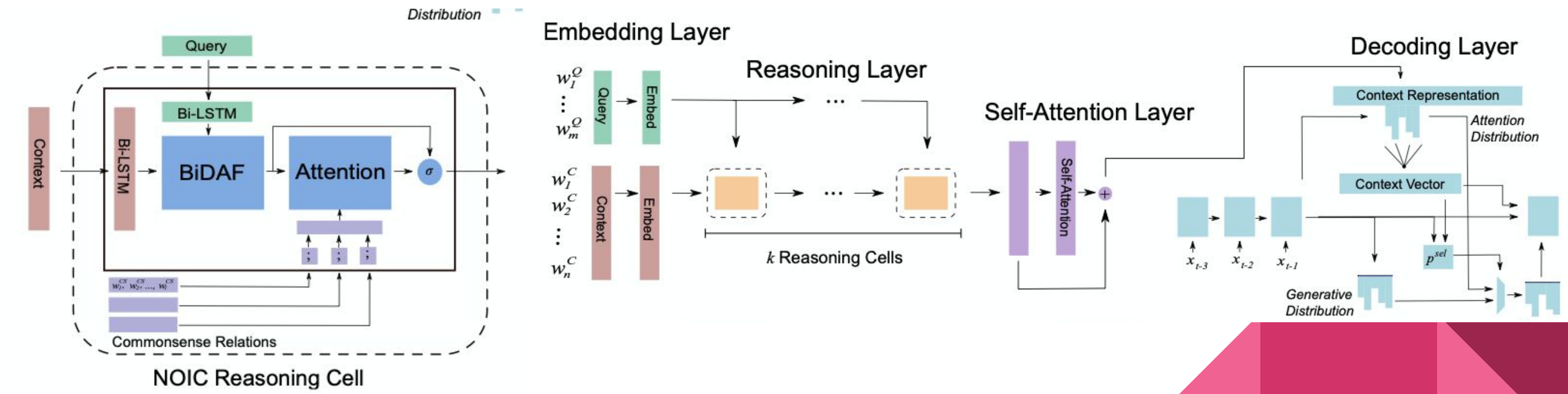
"..Esther sees Lady Dedlock at **church** and talks with her later at Chesney Wod though neither woman recognizes **their** connection.."

"What is the connection between Esther and **Lady** Dedlock?"

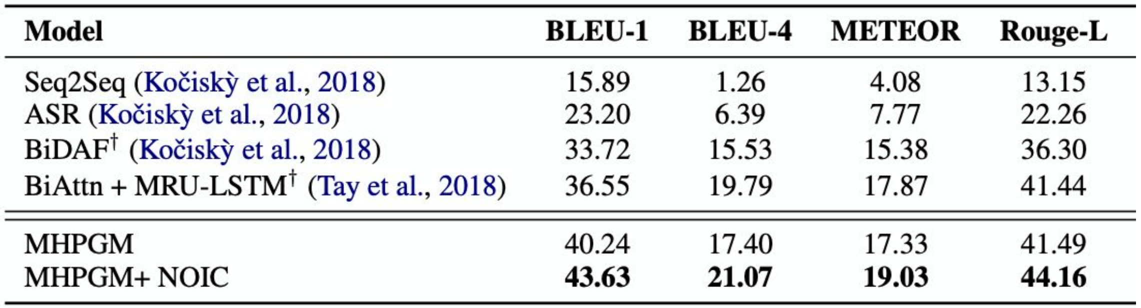
"**Mother** and **daughter**."
"**Mother** and illegitimate **child**."

Necessary and Optional Information Cell (NOIC)

- Use selectively-gated attention mechanism to fill in gaps of reasoning with commonsense from ConceptNet

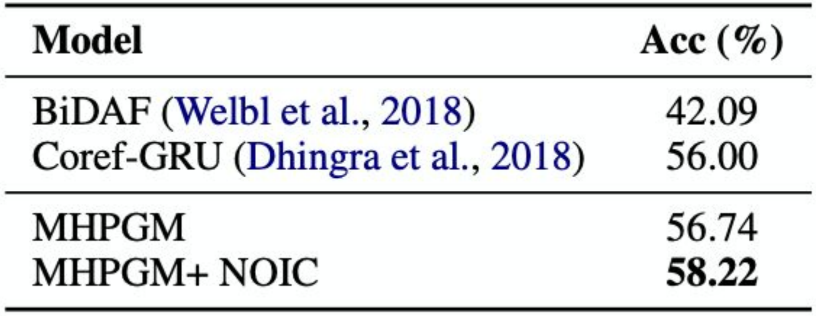


Results



Model	BLEU-1	BLEU-4	METEOR	Rouge-L
Seq2Seq (Kočiskỳ et al., 2018)	15.89	1.26	4.08	13.15
ASR (Kočiskỳ et al., 2018)	23.20	6.39	7.77	22.26
BiDAF[†] (Kočiskỳ et al., 2018)	33.72	15.53	15.38	36.30
BiAttn + MRU-LSTM[†] (Tay et al., 2018)	36.55	19.79	17.87	41.44
MHPGM	40.24	17.40	17.33	41.49
MHPGM+ NOIC	**43.63**	**21.07**	**19.03**	**44.16**

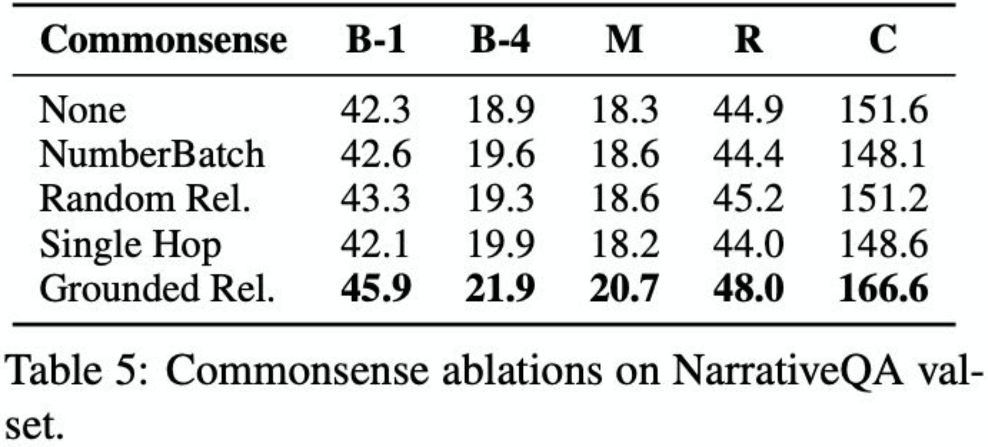
NarrativeQA



Model	Acc (%)
BiDAF (Welbl et al., 2018)	42.09
Coref-GRU (Dhingra et al., 2018)	56.00
MHPGM	56.74
MHPGM+ NOIC	**58.22**

WikiHop

Results (II)



Commonsense	B-1	B-4	M	R	C
None	42.3	18.9	18.3	44.9	151.6
NumberBatch	42.6	19.6	18.6	44.4	148.1
Random Rel.	43.3	19.3	18.6	45.2	151.2
Single Hop	42.1	19.9	18.2	44.0	148.6
Grounded Rel.	**45.9**	**21.9**	**20.7**	**48.0**	**166.6**

Table 5: Commonsense ablations on NarrativeQA val-set.

Takeaways

- Contributions
  - Incorporating commonsense achieves SOTA performance on reading-comprehension QA
  - Novel way of incorporating commonsense knowledge into the model
- Limitations
  - Hard to train?
- Any comments?

Papers Covered Today

1. "COMET: Commonsense Transformers for Automatic Knowledge Graph Construction," Bosselut et. al., ACL 2019
2. "Diverse and Informative Dialogue Generation with Context-Specific Commonsense Knowledge Awareness," Wu et. al, ACL 2020
3. "Commonsense for Generative Multi-hop Question Answering Tasks," Bauer et. al, EMNLP 2018
4. **"Explain Yourself! Leveraging Language Models for Commonsense Reasoning," Rajani et. al, ACL 2019**

PAPER 4: “Explain Yourself! Leveraging Language Models for Commonsense Reasoning, Rajani et. al,” ACL 2019

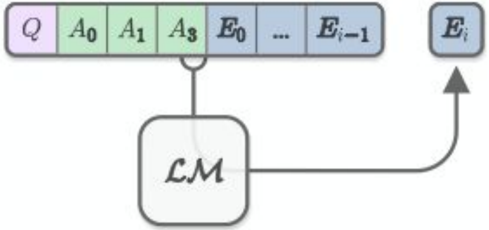
Motivation

Explore how language models can explicitly perform commonsense reasoning

3-part process

- CoS-E
  - Collected human explanations for commonsense reasoning on top of a multiple-choice QA dataset (CommonsenseQA/CQA)
  - Both natural language and highlighted annotations
- CAGE: use a pretrained language model to generate explanations that are close to human-generated ones
- Commonsense reasoning model that uses output of CAGE

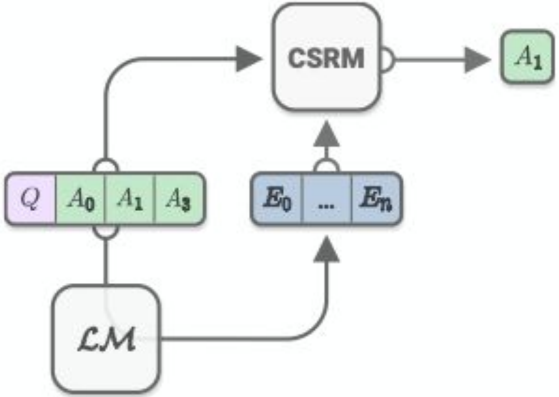
More about CAGE



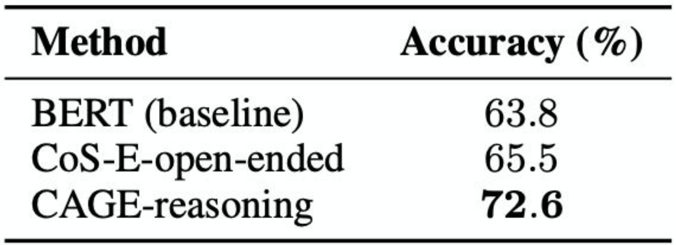
- Fine-tune GPT on CQA & COS-E
- Supervised training: learn to generate an explanation close to the human-generated CoS-E explanation
    - Explain-and-then-predict/Reasoning approach
- Generate explanations for the rest of the CQA training and validation data

More about the commonsense reasoning model

- Add a binary classifier to BERT to perform multiple-choice QA
- For each question, create 3 input sequences, each is a
    - Concatenation of the question, CAGE explanation, and one of the 3 answer choices

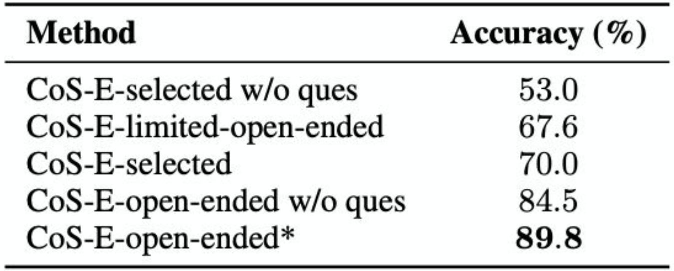


Results

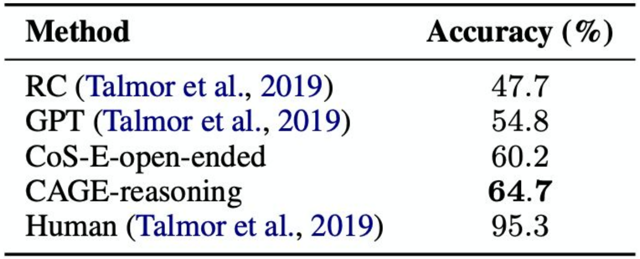


Method	Accuracy (%)
BERT (baseline)	63.8
CoS-E-open-ended	65.5
CAGE-reasoning	**72.6**

During training



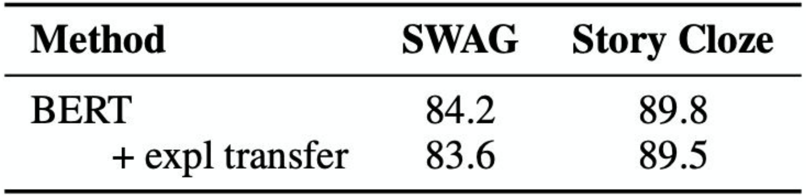
Method	Accuracy (%)
CoS-E-selected w/o ques	53.0
CoS-E-limited-open-ended	67.6
CoS-E-selected	70.0
CoS-E-open-ended w/o ques	84.5
CoS-E-open-ended*	**89.8**



Method	Accuracy (%)
RC (Talmor et al., 2019)	47.7
GPT (Talmor et al., 2019)	54.8
CoS-E-open-ended	60.2
CAGE-reasoning	**64.7**
Human (Talmor et al., 2019)	95.3

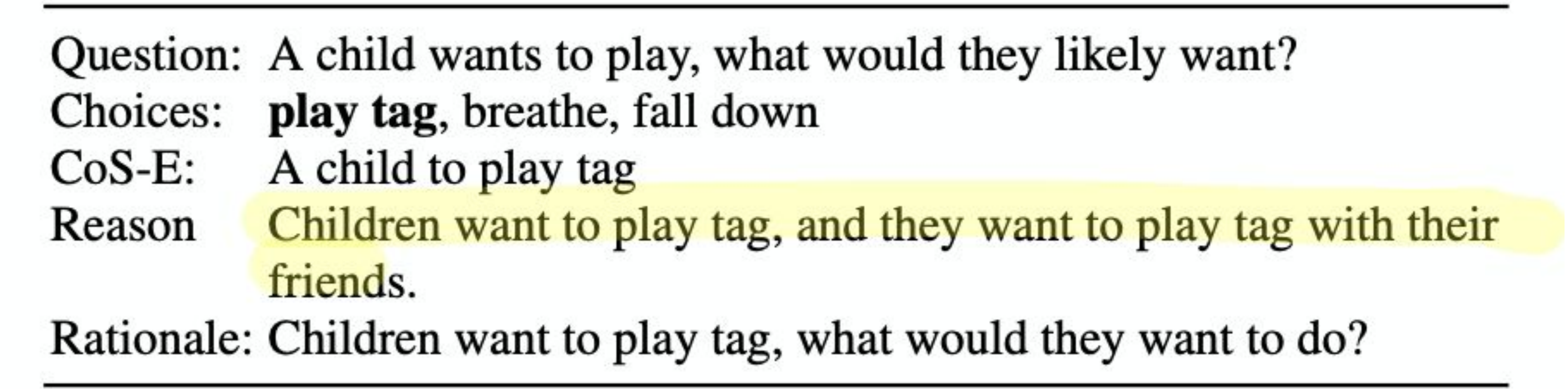
During testing

Domain transfer



Method	SWAG	Story Cloze
BERT	84.2	89.8
+ expl transfer	83.6	89.5

Example



Question:	A child wants to play, what would they likely want?
Choices:	**play tag**, breathe, fall down
CoS-E:	A child to play tag
Reason	Children want to play tag, and they want to play tag with their friends.
Rationale:	Children want to play tag, what would they want to do?

Takeaways

- Contributions
  - Generating commonsense explanations achieves SOTA performance on multiple-choice QA
  - Explanations can be transferred to other datasets and still perform comparably to the baseline
- Limitations
  - Explanations might reinforce biased reasoning into downstream tasks, eg gender bias in CQA
  - In newer, more difficult version of CQA, concatenation approach for the BERT classifier doesn't work well
- How might you incorporate explanations into the BERT classifier instead of using simple concatenation?

Discussion Questions

1. We covered some examples of how commonsense knowledge can improve performance in dialogue generation and QA. What downstream tasks might you want to apply commonsense learning to? Any potential applications to your own research?

Discussion Questions

2. Which of the methods of representing/ incorporating commonsense knowledge do you find most promising/ interesting?

- Non-extractive knowledge graph construction
- Felicitious Fact mechanism to select context-specific knowledge from ConceptNet
- Forcing LMs to generate explanations
- Building tree-based knowledge paths from ConceptNet

Discussion Questions

3. The CAGE framework was built on GPT. We now have even larger models like GPT-3 which has 175 billion parameters. To what extent do you think these large-scaled pretrained language models capture commonsense?

Discussion Questions

4. Do you see any limitations of current evaluation benchmarks/ datasets used for commonsense learning?

Discussion Questions

5. How do you think current evaluation benchmarks/ datasets could be further improved?

Discussion Questions

6. Any future directions/ unsolved ideas you’d like to explore further in this area?

References

Papers

1. "COMET: Commonsense Transformers for Automatic Knowledge Graph Construction," Bosselut et. al., ACL 2019
2. "Diverse and Informative Dialogue Generation with Context-Specific Commonsense Knowledge Awareness," Wu et. al, ACL 2020
3. "Commonsense for Generative Multi-hop Question Answering Tasks," Bauer et. al, EMNLP 2018
4. "Explain Yourself! Leveraging Language Models for Commonsense Reasoning," Rajani et. al, ACL 2019
5. “ATOMIC: An Atlas of Machine Commonsense for If-Then Reasoning,” Sap et. al, AAAI 2019
6. “YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia,” Hoffart et. al, Artificial Intelligence 2013
7. “Knowledge vault: a web-scale approach to probabilistic knowledge fusion,” Dong et al., ACM SIGKDD 2014
8. “SWAG: A Large-Scale Adversarial Dataset for Grounded Commonsense Inference," Zellers et. al, EMNLP 2018
9. “Event2Mind: Commonsense Inference on Events, Intents, and Reactions,” Rashkin et. al, ACL 2018
10. “WinoWhy: A Deep Diagnosis of Essential Commonsense Knowledge for Answering Winograd Schema Challenge,” Zhang et. al, ACL 2020
11. “Attention Is (not) All You Need for Commonsense Reasoning, ” Klein and Nabi, ACL 2019
12. “Does BERT Solve Commonsense Task via Commonsense Knowledge?,” Cui et. al, ArXiv 2020
13. "Augmenting End-to-End Dialogue Systems with Commonsense Knowledge," Young et. al, AAAI 2017
14. “Constructing Datasets for Multi-hop Reading Comprehension Across Documents,” Welbl et. al, ACL 2018
15. “The NarrativeQA Reading Comprehension Challenge," Kočisky et. al, ArXiv 2018
16. “ConceptNet — a practical commonsense reasoning tool-kit,” Liu and Singh, BT Technology Journal 2014

References

Past presentations

1. “Natural Language Understanding with Common Sense Reasoning,” presented by Dan Roth for Microsoft Research Faculty Summit 2015
2. “From Naive Physics to Connotation: Modeling Commonsense in Frame Semantics,” presented by Yejin Choi for StarSem 2017
3. “Commonsense Learning and Reasoning,” presented by Michihiro Yasunaga for the “Advanced NLP” class at Yale in 2018
4. “Commonsense Resources,” presented by Maarten Sap for ACL 2020
5. “Commonsense Benchmarks,” presented by Maarten Sap for ACL 2020

Webpages

1. “GPT3 What is All the Fuss About,” Fiona J McEvoy in You The Data August 2020
2. “Commonsense Inferences about Concepts (COMmonsensE Transformers on ConceptNet),” Allen Institute for AI
3. “The Winograd Schema Challenge,” Ernest Davis, Leora Morgenstern, and Charles Ortiz
4. “CommonsenseQA,” Alan Talmor, Tel-Aviv University Natural Language Processing Group
5. “Cyc platform,” Cycorp Inc.
6. “Common Sense Reasoning Benchmarks,” Paperswithcode.com
7. An image of a fireplace