Feb 2020

Nasrin Mostafazadeh

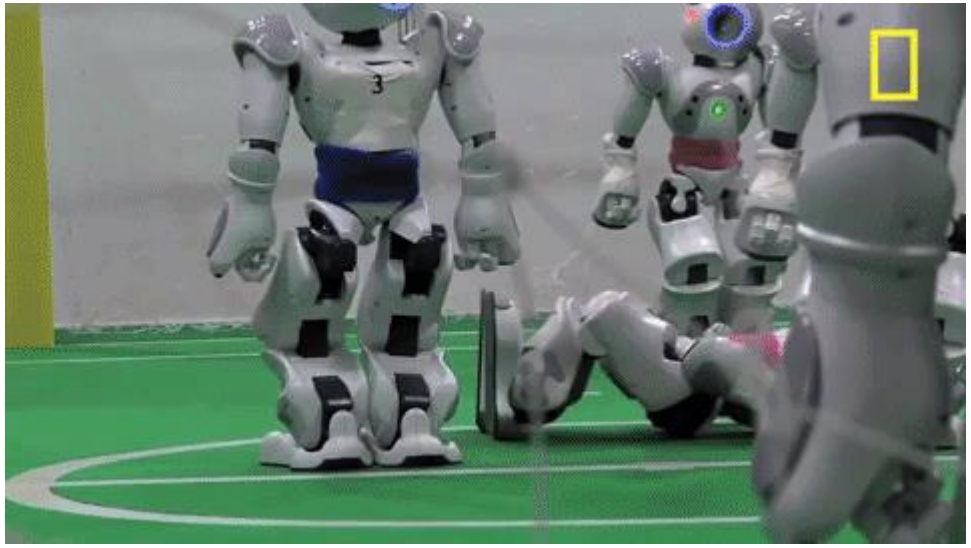
@nasrinmmm

ELEMENTAL cognition

Towards AI systems that can build coherent causal models of what they read!

State of Artificial Intelligence, ~15 years ago

RoboCup Competitions



https://www.youtube.com/watch?v=YPYVL5FpS6s

Classic Motivating NLU Problem

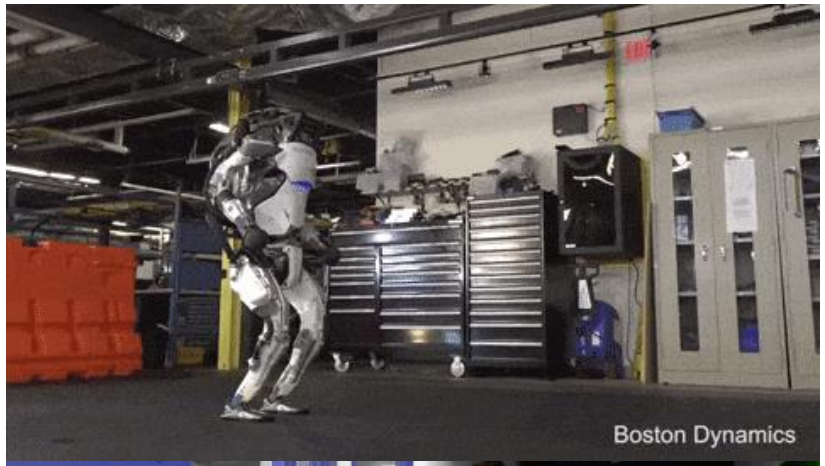
Deemed very challenging for AI systems at the time!

- The monkey ate the banana because it was hungry.
  - Question: What is it? Monkey or the banana?
  - Correct answer: Monkey

State of Artificial Intelligence, NOW!

Boston Dynamics' Robots

(2019)



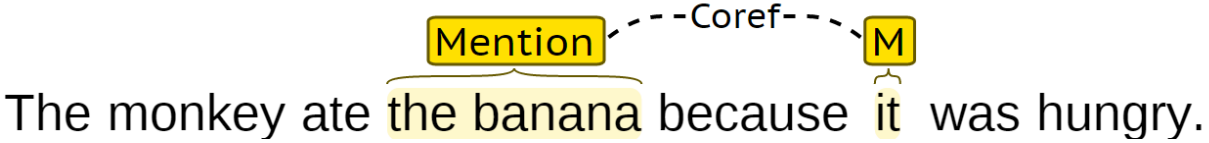
Slide credit: Omid Bakhshandeh

Stanford CoreNLP Coreference Resolver

(Feb 2020)

The Classic Example:

- The monkey ate the banana because it was hungry.
  - What is it? Monkey or the banana?



The paradigm shift in NLP, since 2015…

- ***2015-2017:***
  - ***What happened***: New SOTA established on various NLP benchmark
  - ***Recipe***: Encode the input text using BiLSTMs, decode with attention!
  - ***Shortcomings***: Could not tackle reading comprehension tasks that (supposedly) required:
    - Vast amount of background knowledge, or
    - Reasoning, or
    - Had long established contexts.

    e.g., Story Cloze Test (Mostafazadeh et al., 2016).

3. The BiLSTM Hegemony

**To a first approximation,
the de facto consensus in NLP in 2017 is
that no matter what the task,
you throw a BiLSTM at it, with
attention if you need information flow**

28

Chris Manning

Story Cloze Test (Mostafazadeh et al., 2016)

Narrative comprehension benchmark

Context:

Jim got his first credit card in college. He didn't have a job so he bought everything on his card. After he graduated he amounted a $10,000 debt. Jim realized that he was foolish to spend so much money.

Two alternative endings:

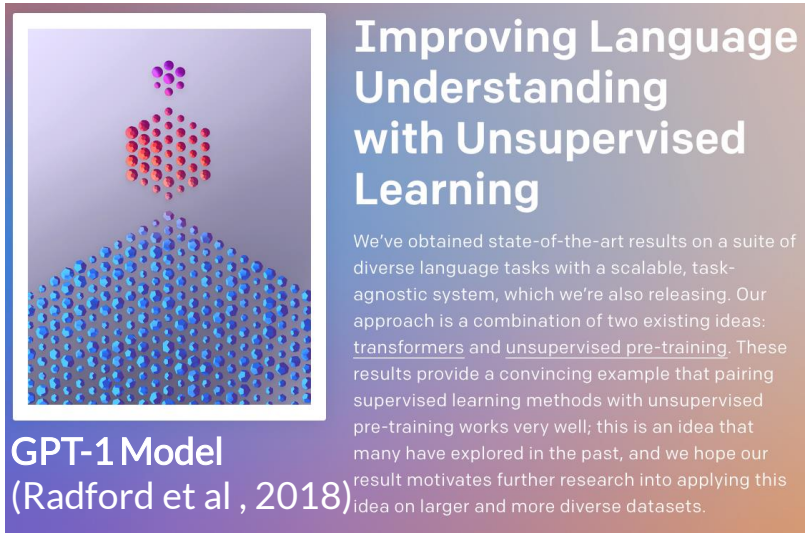
Jim decided to devised a plan for repayment. ✓

Jim decided to open another credit card.

A challenging commonsense reasoning task, where SOTA was ~65% for many months after release of the dataset.

Things got interesting in 2018!

- *Late 2017-2018:*
  - ***What happened***: The dawn of Attention is All you need (Vaswani et al., 2017), introducing transformers. Brand new established SOTA on various supposedly more complex reading comprehension tasks.
  - ***Recipe***: fine-tune large pretrained transformer-based models on downstream tasks (even with a small supervised data)!



GPT-1 Model (Radford et al , 2018)

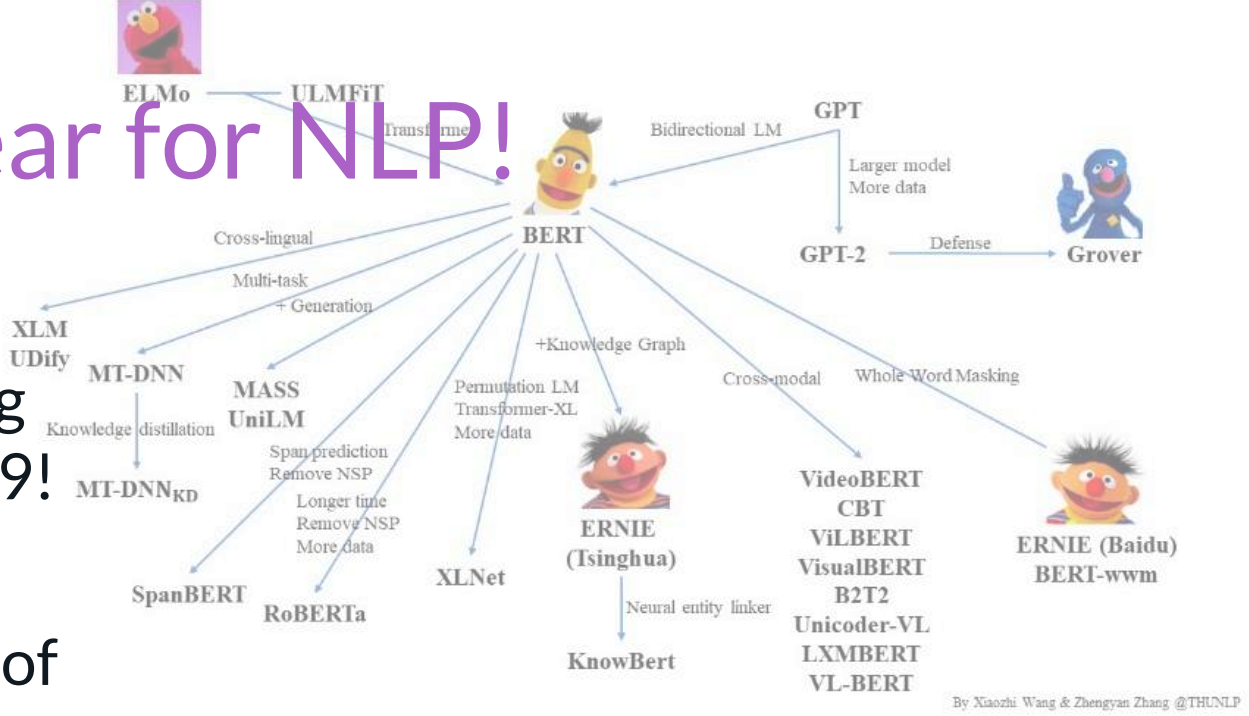
DATASET	TASK	SOTA	OURS
ROCStories	Commonsense Reasoning	77.6	**86.5**
COPA	Commonsense Reasoning	71.2	**78.6**

These results were on the Story Cloze Test v1, where there had been some stylistic biases (Sap et al., 2017).

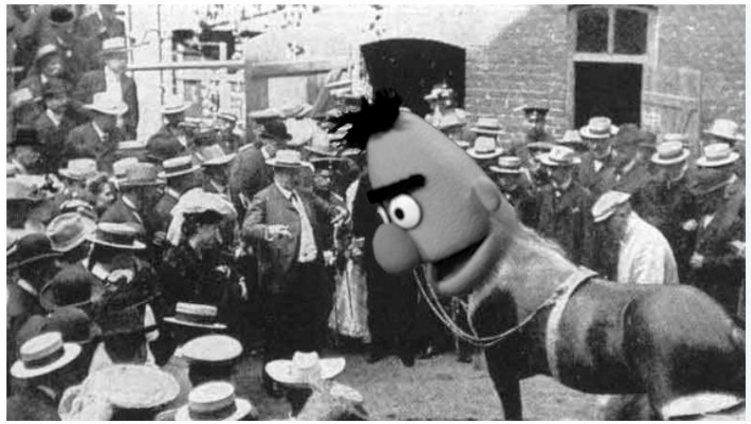
We tested a host of models on the new blind Story Cloze Test v 1.5 test set (Sharma et al., 2018).
The GPT-1 model was the only model still holding its rather high performance!

So, are these models actually learning to transfer various lexical, conceptual, and world knowledge?

2019 was an exciting year for NLP!



- The 2018's recipe of transfer learning was impressively in full bloom in 2019!
- The community has started to think about the problems and weaknesses of the emerging techniques.



NLP's Clever Hans Moment has Arrived

26.AUG.2019

It is now almost a cliché to find out that BERT (Devlin et al., 2019) performs "surprisingly well" on whatever dataset you throw at it.

Benjamin Heinzerling

Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference

R. Thomas McCoy,[1] Ellie Pavlick,[2] & Tal Linzen[1]
[1]Department of Cognitive Science, Johns Hopkins University
[2]Department of Computer Science, Brown University
tom.mccoy@jhu.edu, ellie_pavlick@brown.edu, tal.linzen@jhu.edu

So have we come far enough?

Our moonshot at ELEMENTAL cognition

Machines as thought partners!

We are working building AI systems that **build a shared understanding** with human and **explain** their answers well enough to eventually teach humans!

Building AI systems that can **build coherent causal models** of what they **read**!

When humans, even young children, read, they make countless implicit commonsense inferences that frame their understanding of the unfolding narrative!

Peppa was riding her bike. A car turned in front of her. Peppa turned her bike sharply. She fell off of her bike. Peppa skinned her knee.

While reading, humans construct a coherent representation of what happened and ***why***, combining information from the text with relevant background knowledge

Humans can construct the causal chain that describes how the sequence of events led to a particular outcome!

A car turned in front of Peppa *causes*→

Peppa to turn her bike sharply causes→

Peppa fell off of her bike

causes→

Peppa skinned her knee

causes→

(likely) she asks for help!

Humans can also describe how characters' different states, such as emotions and location, changed throughout the story.

Peppa was on her bike throughout riding it.

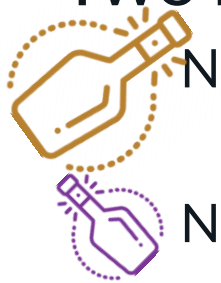
Then after falling, Peppa was on the ground.

eppa went from feeling (likely) happy to feeling in pain after falling.

Though humans build such mental models of situations with ease (Zwaan et al., 1995), **AI systems** for tasks such as reading comprehension and dialogue **remain far from exhibiting similar commonsense reasoning capabilities**!

Why?

- Two major bottlenecks in the AI research:



Not having ways of acquiring (often-implicit) commonsense knowledge at scale.

Not having ways to incorporate knowledge into the state-of-the-art AI systems.

GLUCOSE: GeneraLized and COntextualized Story Explanations!

A new commonsense reasoning framework for tackling both those bottlenecks at **scale**!


Jennifer Chu-Carroll


Lori Moon


Aditya Kalyanpur


David Buchanan


Lauren Berkowitz

GLUCOSE Commonsense Reasoning Framework

- Given a short story *S* and a selected sentence *X* in the story, GLUCOSE defines **ten dimensions of commonsense causal explanations** related to X, **inspired by human cognitive psychology**.

GLUCOSE framework through an Example

Peppa was riding her bike. A car turned in front of her. Peppa turned her bike sharply She fell off of her bike. Peppa skinned her knee.

Semi-structured Inference Rule = antecedent *connective* consequent

Dim #1

Is there an event that directly causes or enables X?

A car turned in front of her *Causes/Enables* Peppa turned her bike

subject verb preposition object — subject verb object

Contextualized: Specific statements exemplify how a general rule could be grounded in a particular context

Sth_A turns in front of Sth_B (that is Someone_A's vehicle) *Causes/Enables* Someone_A turns Sth_B away from Sth_A

subject verb preposition object — subject verb object1 preposition object2

Dim #2

Is there an emotion or basic human drive that motivates X?

Peppa wants safety *Causes/Enables* Peppa turned her bike

subject verb object — subject verb object

Generalized: General rules provide general mini-theories about the world!

Someone_A wants safety *Causes/Enables* Someone_A moves away from Something_A (that is dangerous)

subject verb object — subject verb preposition object1

Dim #3

Is there a location state that enables X?

Peppa was close to a car *Enables* Peppa turned her bike away from the car

subject verb preposition object — subject verb object1 preposition object2

Someone_A is close to Something_A *Enables* Someone_A moves away from Something_A

subject verb preposition object — subject verb preposition object2

GLUCOSE framework through an Example

Peppa was riding her bike. A car turned in front of her. Peppa turned her bike sharply. She fell off of her bike. Peppa skinned her knee.

Is there a possession state that enables X?

Peppa (subject) possesses (verb) a bike (object) *Enables* Peppa (subject) turned (verb) her bike (object)

Someone$_A$ (subject) possesses (verb) Something$_A$ (object) *Enables* Someone$_A$ (subject) moves (verb) Something$_A$ (object)

Are there any other attributes enabling X?

N/A (the dimension is not applicable for this example)

GLUCOSE is a unique perspective on commonsense reasoning for presenting often-implicit commonsense knowledge in the form of ***semi-structured general*** **inference rules** that are also ***grounded*** in the **context of a specific story**!

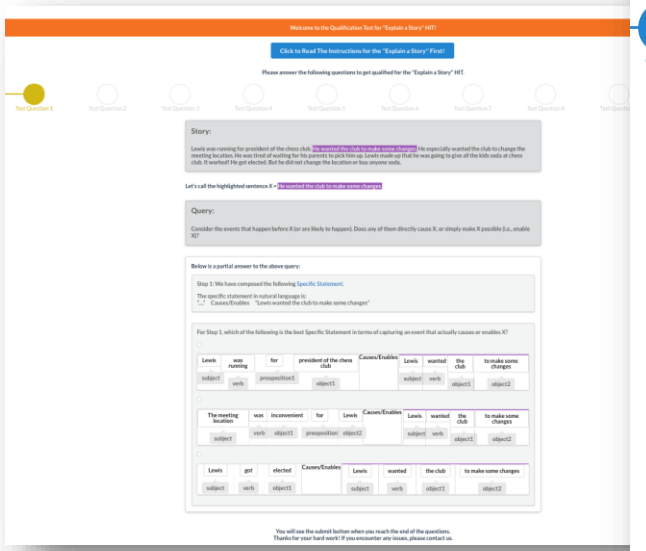
GLUCOSE captures **mini causal theories about the world** focused around **events, states** (location, possession, emotion, etc), **motivations**, and **naive human psychology**.

How to address the problem of implicit knowledge acquisition at scale?

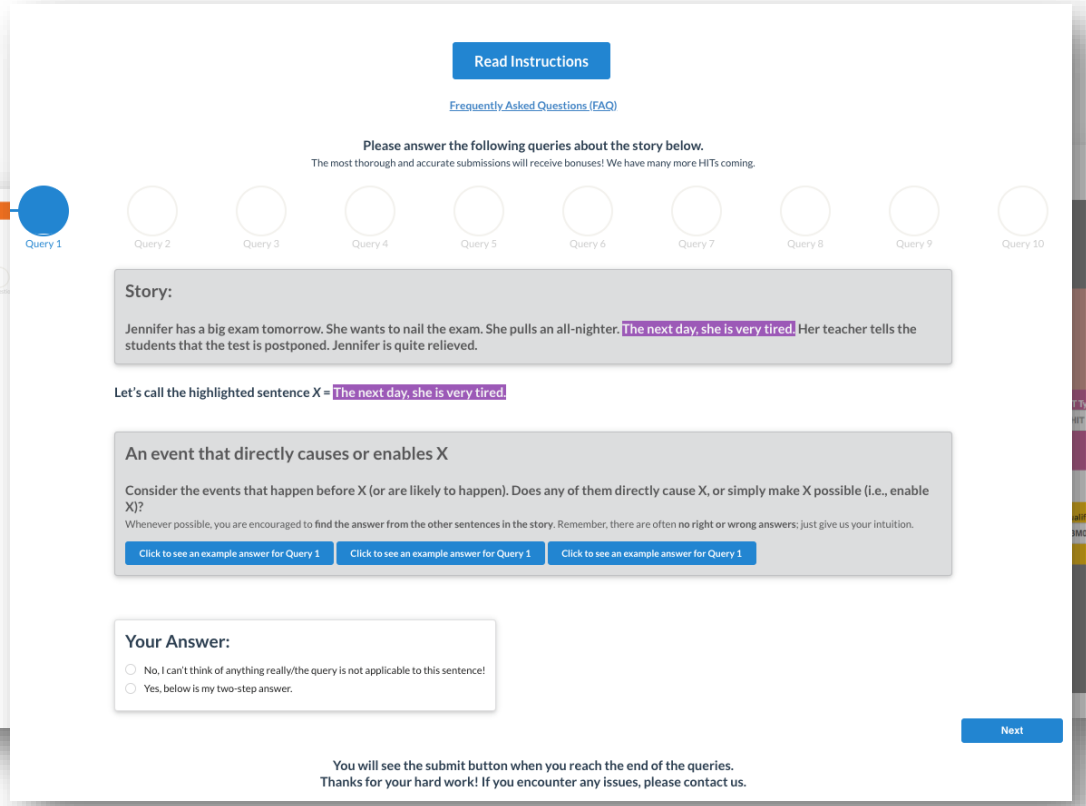
Filling in the GLUCOSE dimensions is **cognitively a complex task** for **lay workers**, since it **requires grasping the concepts of causality and generalization** and to write **semi-structured inference rules**!

An effective multi-stage crowdsourcing platform

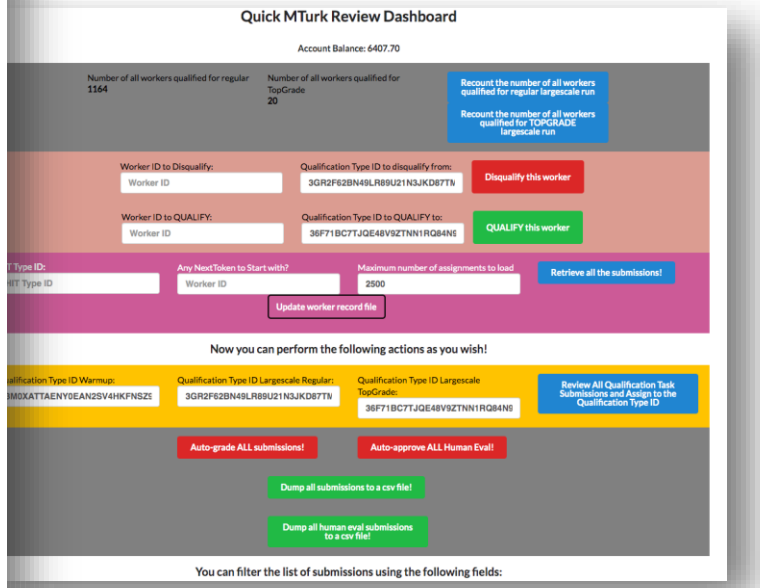
After **many rounds of pilot studies**, we successfully designed an **effective platform** for collecting GLUCOSE data that is **cognitively accessible to laypeople**!



GLUCOSE Qualification UI



GLUCOSE Main UI



GLUCOSE Review Dashboard

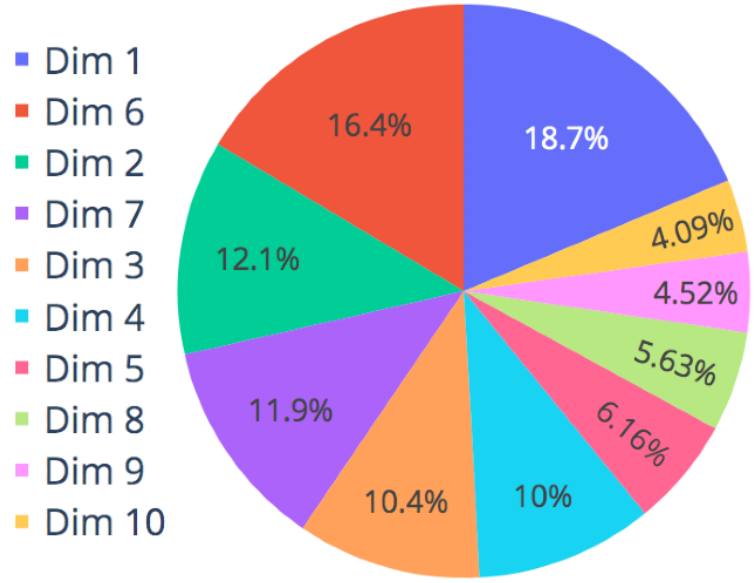
Statistics and Examples

Various implicit and script-like mini-theories:

- Someone_A gives Someone_B Something_A *Results in* Someone_B possess(es) Something_A
- Someone_A is Somewhere_A *Enables* Someone_A forgets Something_A Somewhere_A
- Someone_A is careless *Enables* Someone_A forgets Something_A Somewhere_A
- Someone_A forgets Something_A Somewhere_A *Results in* Something_A is Somewhere_A
- Someone_A feel(s) tired *Enables* Someone_A sleeps
- Someone_A is in bed *Enables* Someone_A sleeps
- Someone_A runs into Someone_B (who Someone_A has not seen for a long time) *Causes* Someone_A feel(s) surprised
- Someone_A asks Someone_B a question *Causes/Enables* Someone_B answers the question

# total inference rules	620K
# total unique stories	4700
# workers participated.	372
# mins per HIT on avg.	4.6min

To our knowledge, GLUCOSE is among the few cognitively-challenging AI tasks to have been successfully crowdsourced!



GLUCOSE captures extensive commonsense knowledge that is unavailable in the existing resources

Ceiling overlap between GLUCOSE and other resources based on best-effort mapping of relations.

GLUCOSE Dim	1	2	5	6	7	10
ConceptNet	1.2%	0.3%	0%	1.9%	0%	0%
ATOMIC	7.8%	1.2%	2.9%	5.3%	1.8%	4.9%

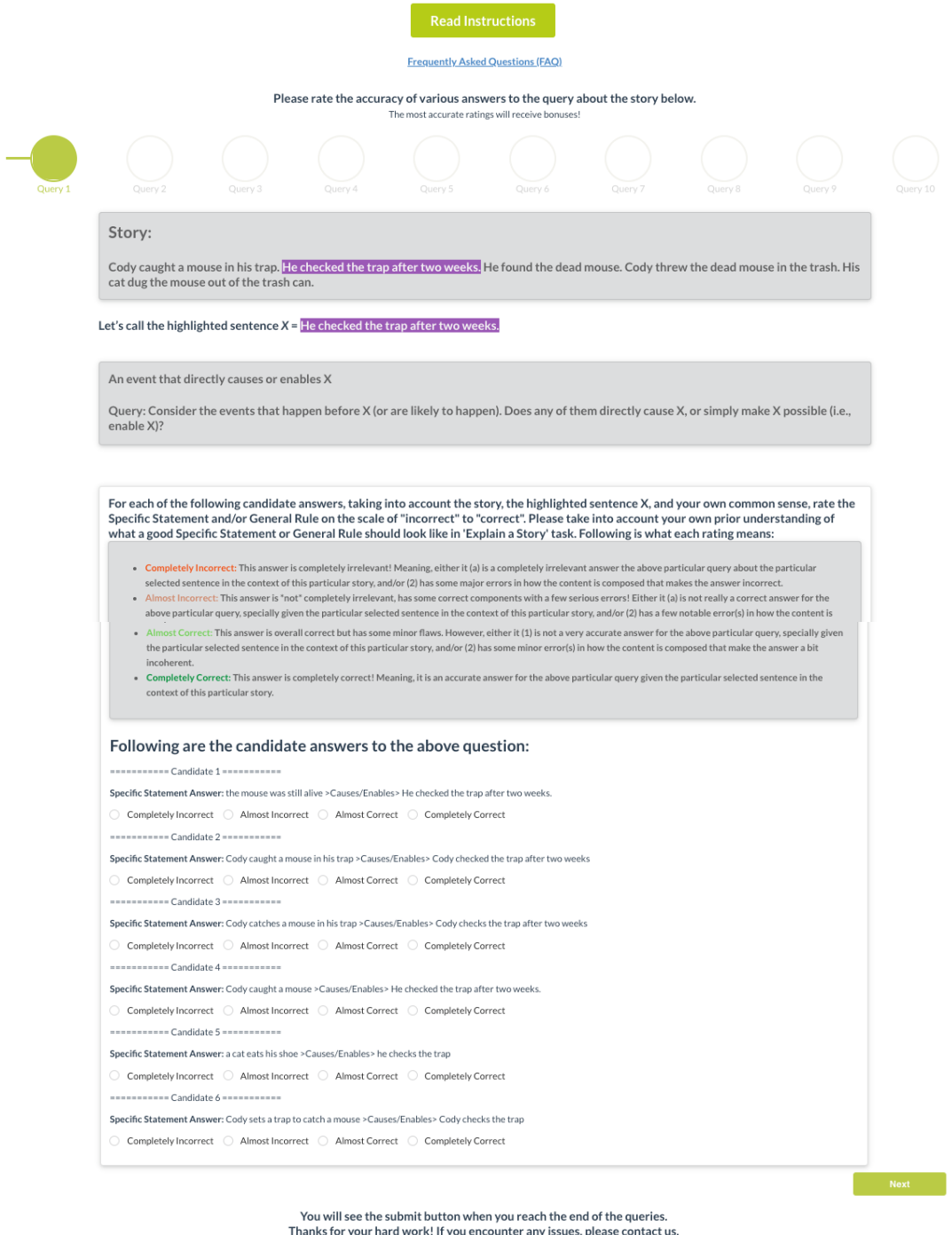
How to incorporate commonsense knowledge into the state-of-the-art AI systems?

GLUCOSE Commonsense Reasoning Benchmark

A testbed for evaluating models that can incorporate such commonsense knowledge and show inferential capabilities

- **Task**: Given a story *S*, the sentence *X*, and dimension *d*, the GLUCOSE specific and general rules should be predicted.
- **Test Set**: We carefully curated a **doubly vetted** test set, based on previously **unseen** stories and on which our **most reliable annotators** had **high** agreement. Our vetting process resulted in a test set of 500 GLUCOSE story/sentence pairs, each with 1-5 dimensions answered.
- **Evaluation Metrics**: Human and Automatic

We designed a specialized Human Evaluation UI for collecting reliable, reproducible, and calibrated ratings!



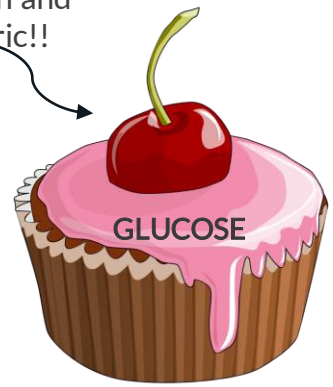
Automatic Evaluation
of natural language generations

- A majority of commonsense reasoning frameworks have been in **multiple-choice form, as opposed to natural language generation**, due to **ease of evaluation**
  - Multiple-choice tests are inherently easier to be **gamed**!
- Automatic evaluation for tasks involving natural language generation with diverse possibilities has been a major bottleneck for research
- BLEU's ease of replicability has made it a popular automated metric, but its correlation with human judgement has proven weak in various tasks.

Automatic Evaluation
of natural language generations in GLUCOSE

- We found very strong pairwise correlation between human and ScareBLEU corpus-level scores on our test set.
  - Spearman = 0.891, Pearson = 0.855, and Kendall's = 0.705, all with p-value < 0.001.
- This is accomplished through various design choices in GLUCOSE:
  1) GLUCOSE **semi-structured inference rules** are **designed to be evaluable**, where the **structure naturally limits the format** of the generated rules
  2) We curated our test set to eliminate cases with a wide range of correct responses where humans cannot agree, making the **limited number of gold references sufficient for automatic evaluation**
  3) We designed a systematic human evaluation process that can **collect calibrated ratings from judges** who are well educated about what constitutes a correct GLUCOSE rule.

Strong correlation between human and automatic metric!!



GLUCOSE task has a systematic evaluation that is fast and easily replicable!

Notable Models & Results

Baselines

KNN

GPT-2

Avg: g 0.5/3

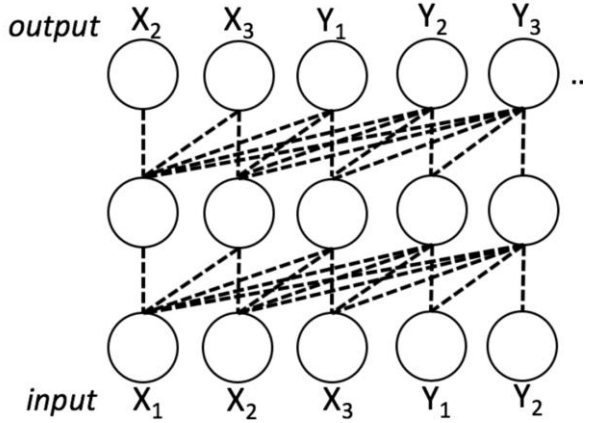
Avg: s 0.8/3

We show that:
1) The KNN model performs the worst, highlighting the importance of generalizing beyond the training data.

2) Pre-trained language model perform very poorly at the task and does not show basic commonsense inference!

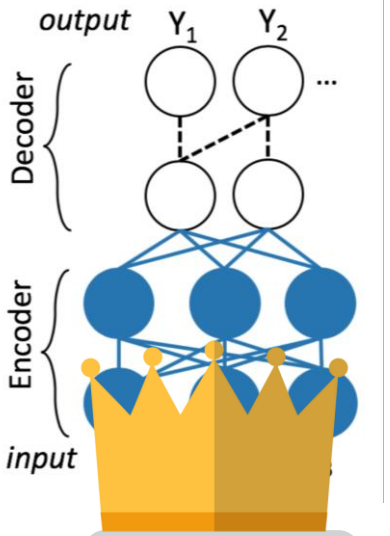
3) When the pre-trained neural models are fine-tuned on the rich GLUCOSE data, they achieve very high performance in making commonsense predictions on ***unseen*** stories.

Trained Models



Full-LM

Avg: s 1.9/3 g 1.7/3



Enc-Dec

Avg: s 2.6/3 g 2.3/3

Human!

Avg: s 2.8/3 g 2.6/3

Example Predictions

Dimension 3; a location enabling X.

- Input:
  - *Karen made a pan of lasagna. She brought it to the party. Nobody wanted to eat lasagna. Karen ate it for a week. She became tired of lasagna.*

GPT-2	She was **in front of a TV** >Enables> Karen made a pan of lasagna. N/A
Full-LM	Karen is at **home** >Enables> Karen made a pan of lasagna and brought it to the party $Someone_A$ is in $Somewhere_A$ >Enables> $Someone_A$ makes $Something_A$ (that is edible)
Enc-Dec	Karen is in **the kitchen** >*Enables*> Karen makes a pan of lasagna $Someone_A$ is in a kitchen >Enables> $Someone_A$ cooks $Something_A$
Human	Karen is in **the kitchen** >*Enables*> Karen made a pan of lasagna $Someone_A$ is in a kitchen >Enables> $Someone_A$ prepares $Something_A$ (that is a dish)

Example Predictions

Dimension 6; an event that X Causes/Enables.

- Input:
  - *Karen made a pan of lasagna. She brought it to the party. Nobody wanted to eat lasagna. Karen ate it for a week. She became tired of lasagna.*

Enc-Dec

Karen makes a pan of lasagna >Causes/Enables> Karen eats it for a week
$Someone_A$ makes $Something_A$ (that is food) >Causes/Enables> $Someone_A$ eats $Something_A$

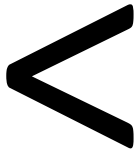
Human

Karen makes a pan of lasagna >Causes/Enables> Karen brought it to the party
$Someone_A$ prepares $Something_A$ (that is a dish) >Causes/Enables> $Someone_A$ takes $Something_A$ to SomethingB (that is an event)

We proved our following hypothesis

A promising new recipe for giving machines commonsense is to use **high-quality commonsense knowledge** as the **seed data** for training **neural models that have pre-existing lexical and conceptual knowledge**.

Static commonsense knowledge base with GLUCOSE mini-theories authored by humans



value

GLUCOSE-Trained model that can generate GLUCOSE dimensions for any novel input

Old-school commonsense knowledge base is static

Modern commonsense knowledge base is dynamic

To conclude:

We've come a rather long way in the last decade in NLP with lots of exciting progress.
My hope for our directions in 2020 is to work on tackling the following issues which we are still grappling with…

- Our amazing models sometimes make glaringly stupid mistakes, being brittle! This makes it hard to deploy these models into real-world products.
- We yet don't know the implications of establishing SOTA on various benchmarks. Are we making any real progress? Do these models work outside of our narrow lab settings in the real world?
- We still cannot tackle tasks that have little to no annotated data. Better knowledge transfer across domains and incorporating prior knowledge and world models is essential.
- Handful of top industry players get to pay the costs for building ever-larger (and not-green) models. Where are we going with this paradigm?
- And we yet don't have an AI system that has commonsense of perhaps even a dog (?), let alone a 5-year-old kid…

Thanks for listening!