Commonsense Reasoning: Models and New Challenges

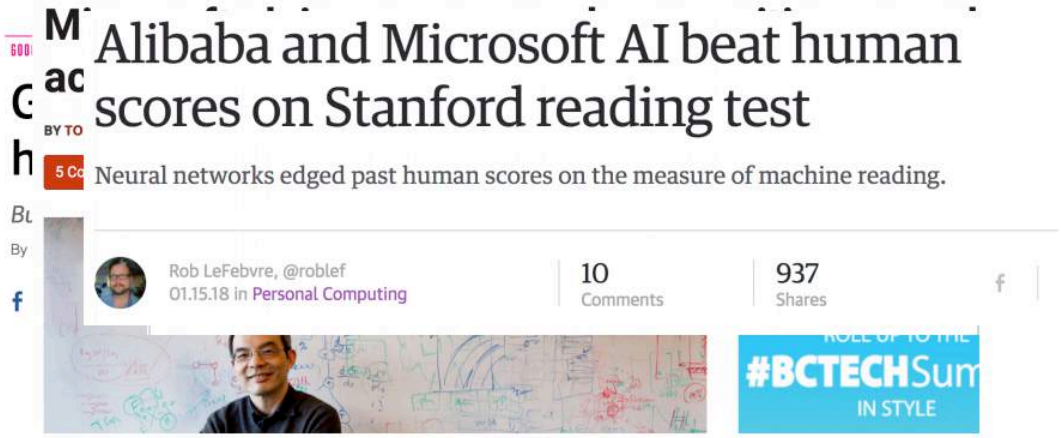
Sean (Xiang) Ren

Department of Computer Science

Information Science Institute

USC

http://inklab.usc.edu



2018 human-level performance on **reading comprehension** on SQuAD (Stanford QA dataset)

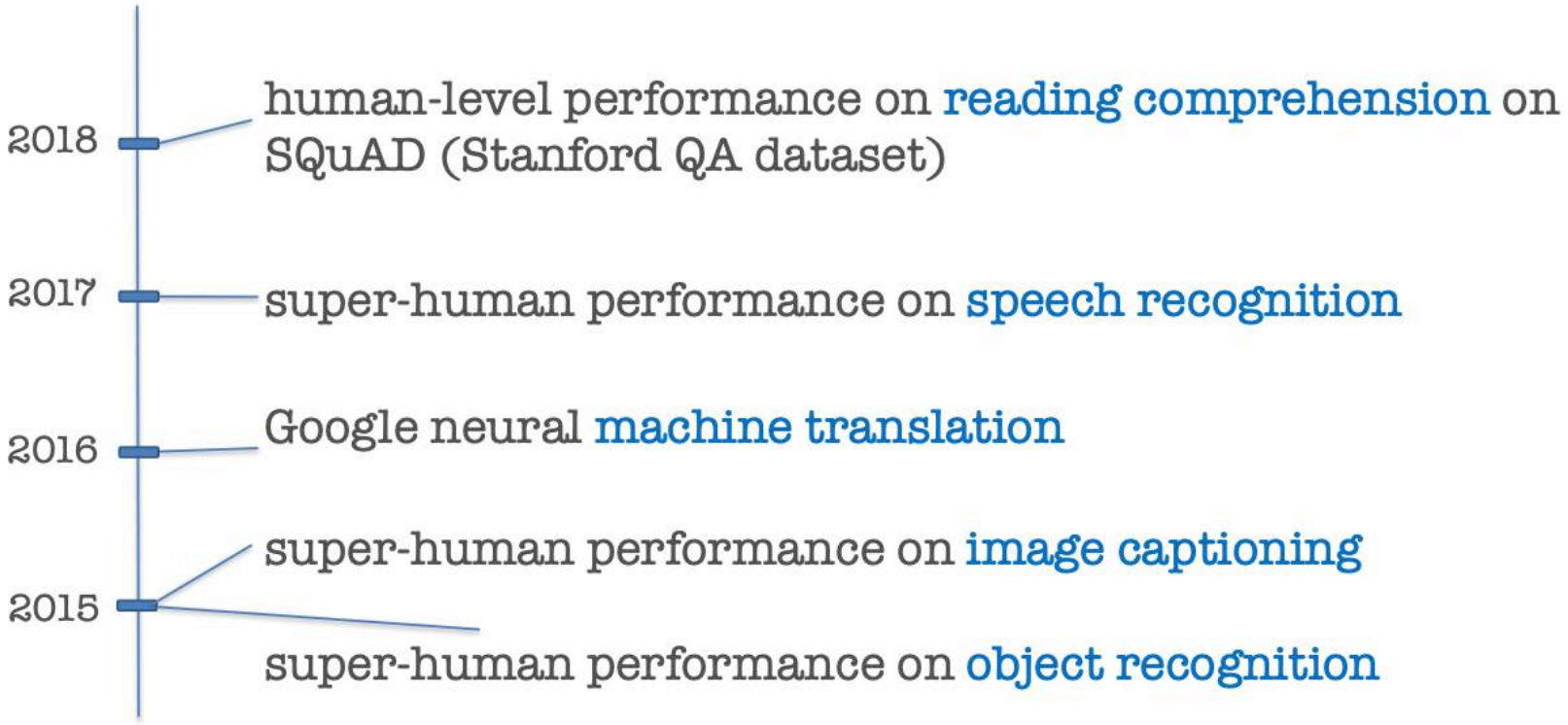
2017 super-human performance on **speech recognition**

2016 Google neural **machine translation**

2015 super-human performance on **image captioning**

super-human performance on **object recognition**

Done Solving AI?



Solving a “dataset” vs the underlaying “task”

Image credit: Yejin Choi

Why Commonsense Knowledge?

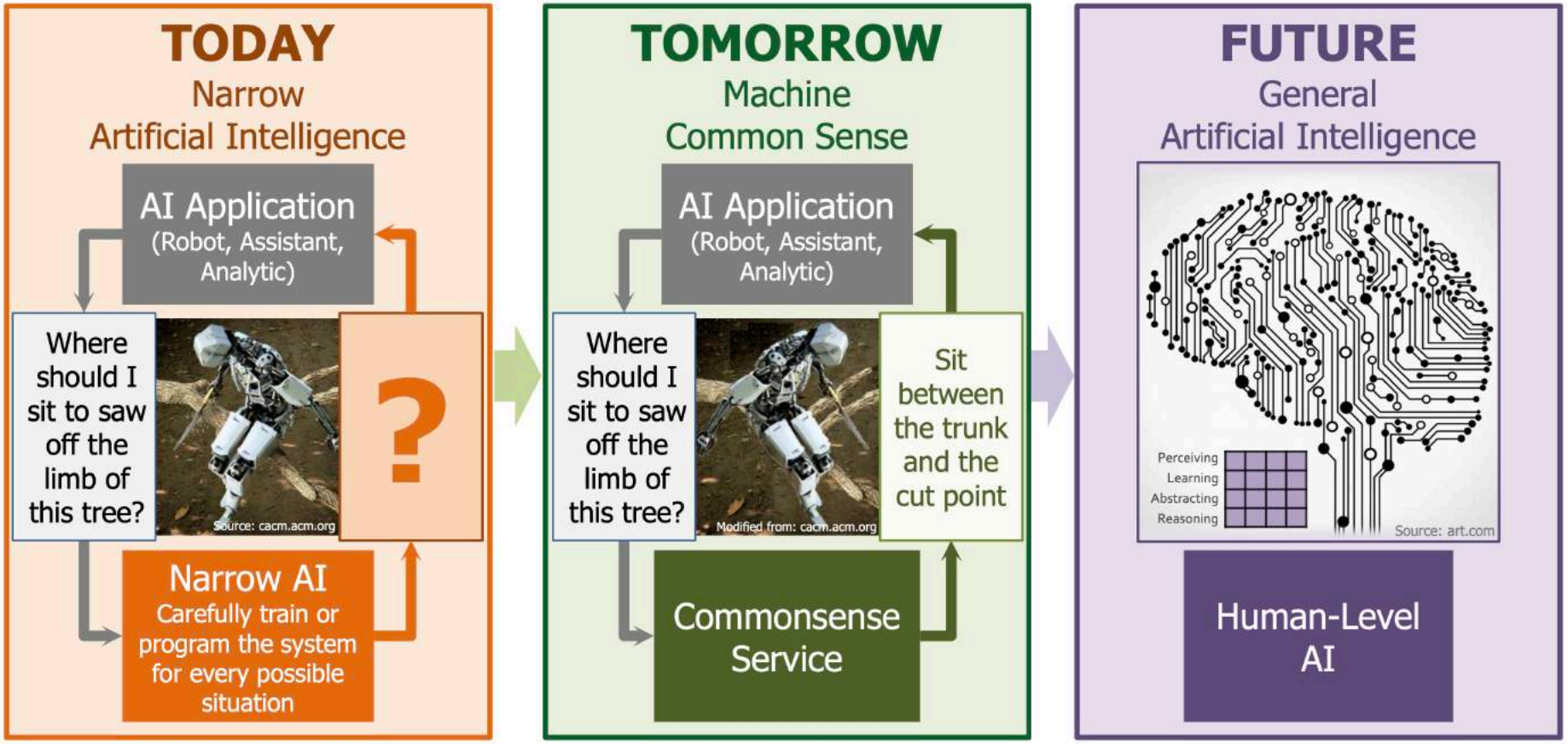


Image source: https://www.darpa.mil/program/machine-common-sense

Commonsense problems in NLP

NLU: Multi-choice QA (w/o context)

Where do adults usually use glue sticks?

A: classroom **B:** *office* **C:** desk drawer

NLG: Constrained Sentence Generation (w/ a set of keywords)

Generate a daily-life scene about a concept-set: {apple, bag, tree}

*A boy picks some **apples** from a **tree** and puts them into a **bag**.*

Commonsense Reasoning (CSR)?

- Definition of Common Sense: the basic level of practical knowledge and reasoning
  - Physical objects, properties, laws
  - Human behaviors / social conventions
  - Temporal commonsense

- The human-like ability to understand and generate everyday scenarios (situations, events)

- The computation process of manipulating commonsense knowledge to make compositional logical inference.

This Talk

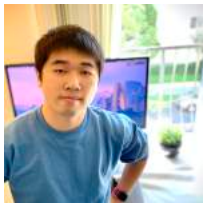
- Part I: Discriminative Commonsense Reasoning
  - Improving language understanding with commonsense
  - Models: KagNet and multi-hop relational network

- Part II: Generative Commonsense Reasoning
  - Imposing commonsense to text generation
  - A new task & dataset: CommonGen
  - Methods and Evaluation

Part I

KagNet: Knowledge-Aware Graph Networks for Commonsense Reasoning

Bill Yuchen Lin *Xinyue Chen* *Jamin Chen* *Xiang Ren*



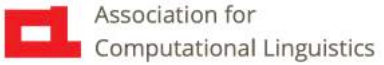

University of Southern California - Information Science Institution

INK Lab @ USC-ISI

http://inklab.usc.edu


EMNLP-IJCNLP 2019

Hong Kong, China



Commonsense Question Answering

Where do adults usually use glue sticks?

A: classroom **B:** office **C:** desk drawer

What do you need to fill with ink to write notes on an A4 paper?

A: fountain pen **B:** printer **C:** pencil

Can you choose **the most plausible answer** based on daily life **commonsense** knowledge?

(Bill Yuchen Lin et al. 2019)

Commonsense Question Answering

Where do adults usually use glue sticks?

A: classroom **B:** office **C:** desk drawer

What do you need to fill with ink to write notes on an A4 paper?

A: fountain pen **B:** printer **C:** pencil

From the **CommonsenseQA** dataset (Talmor et al. NAACL 2019)

Research question:

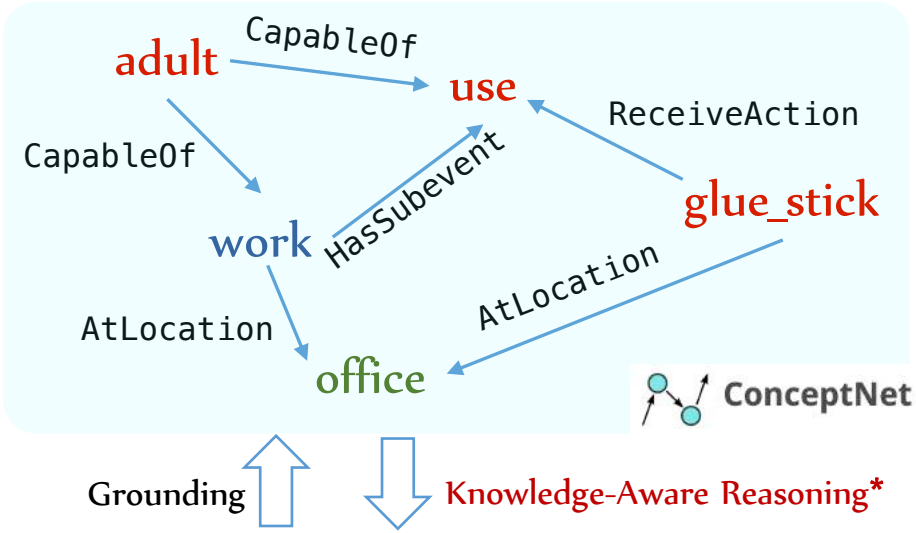
How can we impose commonsense in NLU models?

(Bill Yuchen Lin et al. 2019)

KagNet: Knowledge-Aware Graph Networks

Knowledge-Aware Reasoning

Symbol Space



A **Schema Graph** for the choice **B**

Semantic Space

Where do adults use glue sticks?

A: classroom B: office C: desk drawer

Question

Answer Candidates

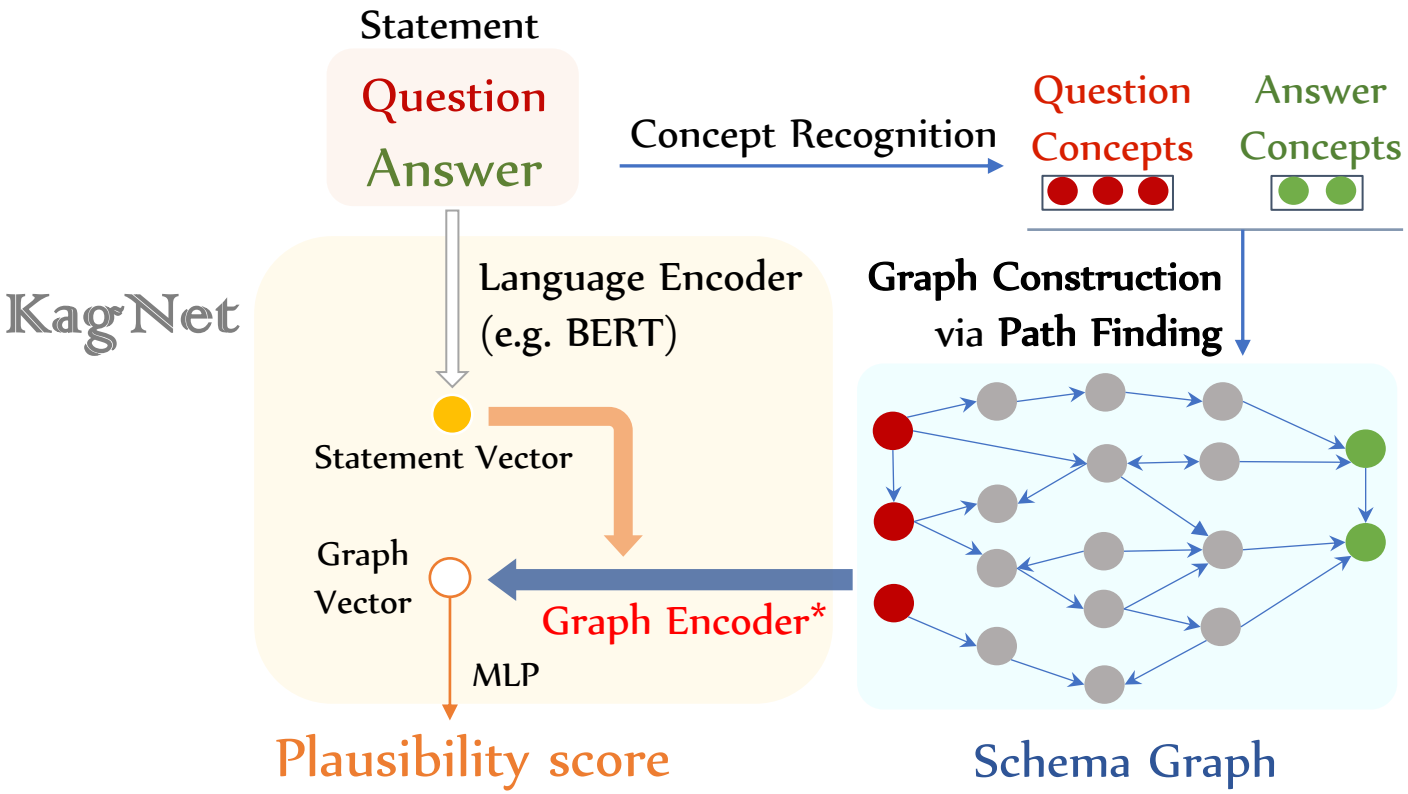
(Bill Yuchen Lin et al. 2019)

Challenges in knowledge-aware reasoning

- How can we find the schema graphs?
  - Noisy and Incomplete
  - Numerous graphs; how to select the most related ones
- How do we encode these graphs for **reasoning**?
  - Complex multi-relational graph structures
  - **NO supervision in aligning** graphs and question-answer pairs
  - Need to be compatible with neural sentence encoders

(Bill Yuchen Lin et al. 2019)

Proposed Framework Overview

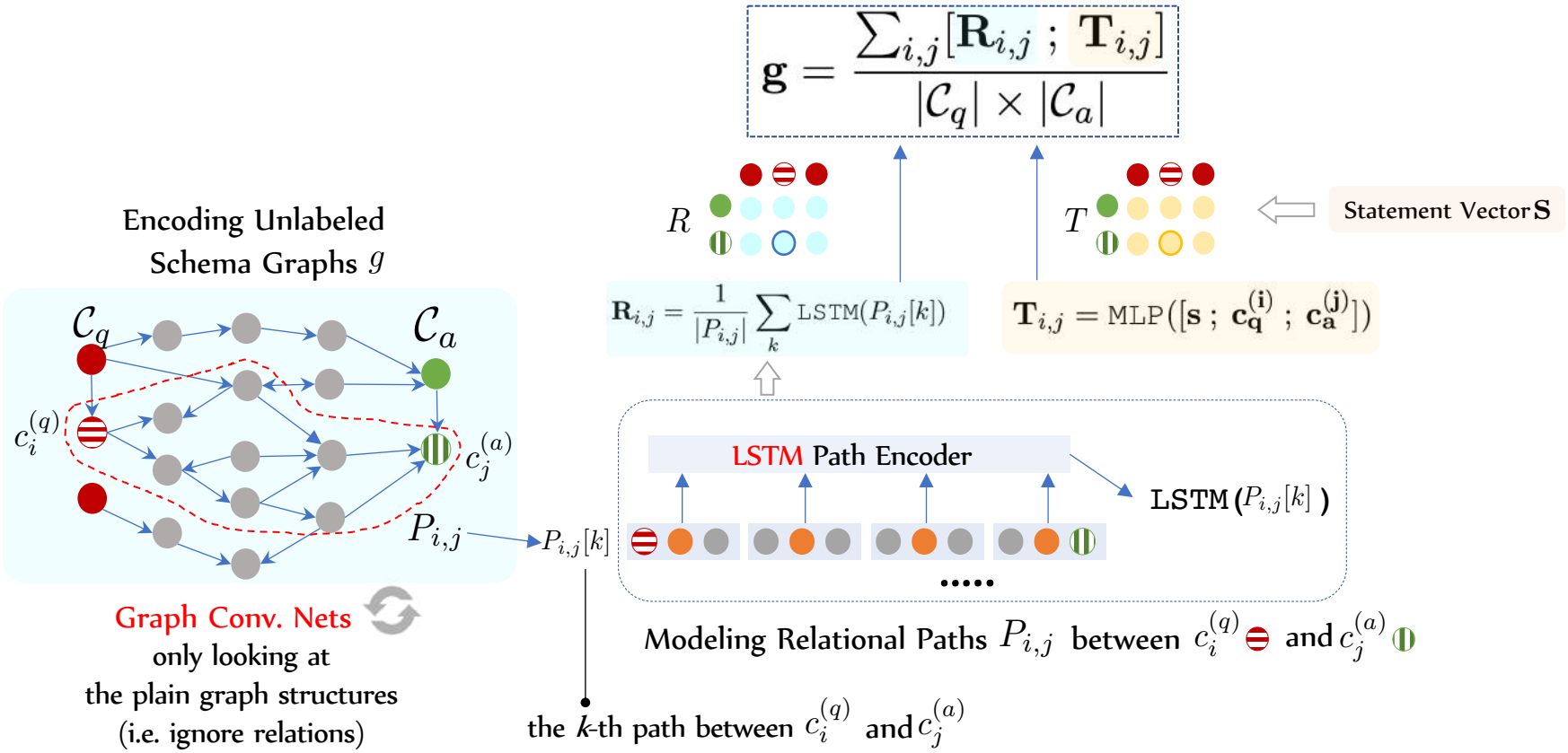


(Bill Yuchen Lin et al. 2019)

(1) Schema Graph Construction

- **Concept Recognition**
  - Tokenization / Lemmatization
  - Match ConceptNet vocabulary
  - Merge multiple smaller concepts into a longer one
    - e.g. `` fountain", ``pen" --> ``fountain pen"
  - Question Concepts C_q and Answer Concepts C_a
- **Path Finding**
  - Find paths between each QA-concept pair (one from C_q and one from C_a)
    - denotes the set of paths between i-th question concept and j-th answer concept $c_i^{(q)}$, $c_j^{(a)}$
  - Path **pruning** by length (<= 5 nodes) and embedding-based metric.

(2) Path-based Relational Graph Encoder



(3) w/ Hierarchical Path-based Attention

- Two average pooling:
  - Assuming all QA-concept pairs are equally important $\mathbf{g} = \frac{\sum_{i,j}[\mathbf{R}_{i,j}\ ;\ \mathbf{T}_{i,j}]}{|\mathcal{C}_q| \times |\mathcal{C}_a|}$
  - Assuming all paths are equally relevant $\mathbf{R}_{i,j} = \frac{1}{|P_{i,j}|}\sum_k \texttt{LSTM}(P_{i,j}[k])$

- Modeling the two-level importance as latent weights:

$$\alpha_{(i,j,k)} = \mathbf{T}_{i,j}\ \mathbf{W}_1\ \texttt{LSTM}(P_{i,j}[k]),$$
$$\hat{\alpha}_{(i,j,\cdot)} = \texttt{SoftMax}(\alpha_{(i,j,\cdot)}),$$
$$\hat{\mathbf{R}}_{i,j} = \sum_k \hat{\alpha}_{(i,j,k)} \cdot \texttt{LSTM}(P_{i,j}[k])$$

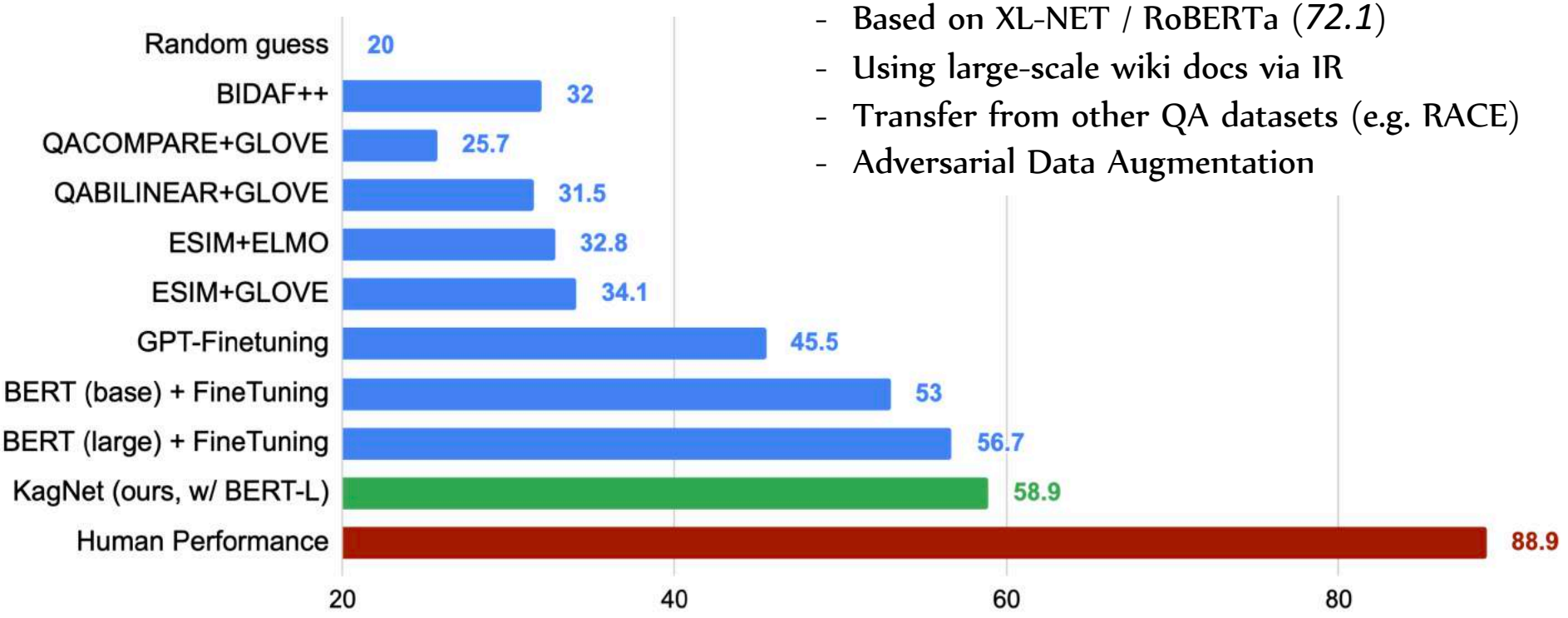
Path-Level Attention
(attending on semantic space)

$$\beta_{(i,j)} = \mathbf{s}\ \mathbf{W}_2\ \mathbf{T}_{i,j}$$
$$\hat{\beta}_{(\cdot,\cdot)} = \texttt{SoftMax}(\beta_{(\cdot,\cdot)})$$
$$\hat{\mathbf{g}} = \sum_{i,j} \hat{\beta}_{(i,j)}[\hat{\mathbf{R}}_{i,j}\ ;\ \mathbf{T}_{i,j}]$$

ConceptPair-Level Attention
(attending on statement)

(Bill Yuchen Lin et al. 2019)

Experiments



Recent follow-up submissions:

- Based on XL-NET / RoBERTa (*72.1*)
- Using large-scale wiki docs via IR
- Transfer from other QA datasets (e.g. RACE)
- Adversarial Data Augmentation

More Performance on Official Test Set: https://www.tau-nlp.org/csqa-leaderboard

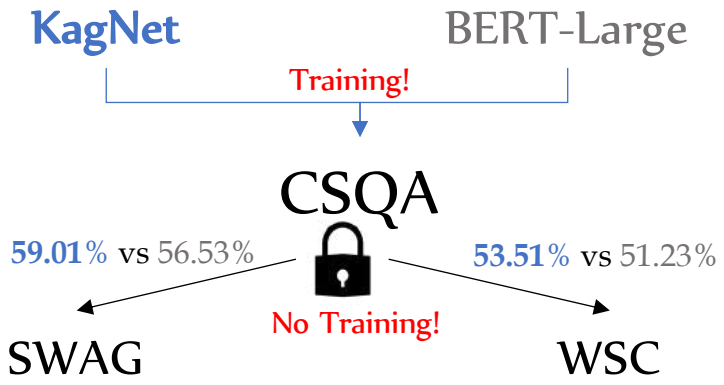
(Bill Yuchen Lin et al. 2019)

Interpretability

What do you **fill** with **ink** to **write** on an **A4 paper**?

A: fountain pen ✓ (KagNet); B: printer (BERT);
C: squid D: pencil case (GPT); E: newspaper

Transferability



(Bill Yuchen Lin et al. 2019)

Conclusion

- A novel framework for knowledge-aware commonsense QA
- A graph neural network for relational reasoning.
  - GCN + Path-based LSTM + Hierarchical Attention
  - Promising for other reasoning tasks over graphs (e.g. GQA)
- Future directions in commonsense reasoning:
  - Towards Learnable Graph Construction (instead of heuristic algs.)
  - Explicitly deal with negations ("not", "but", etc.) and comparisons ("largest", "most", etc.).
    - Logical forms, executable semantic parsing.
  - Interactively reasoning over a sequence of questions
- Our code is at https://github.com/INK-USC/KagNet

Multi-Hop Graph Relation Networks for Knowledge-Aware Question Answering

https://arxiv.org/abs/2005.00646

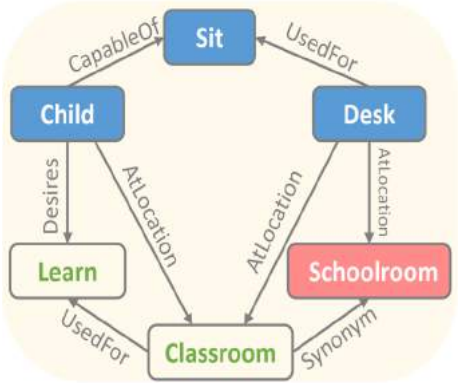
Yanlin Feng[♣*] **Xinyue Chen**[♠*] **Bill Yuchen Lin**[♥] **Peifeng Wang**[♥] **Jun Yan**[♥] **Xiang Ren**[♥]
fengyanlin@pku.edu.cn, kiwisher@sjtu.edu.cn,
{yuchen.lin, peifengw, yanjun, xiangren}@usc.edu
[♥]University of Southern California
[♣]Peking University [♠]Shanghai Jiao Tong University

Motivation

- **KG-Augmented Commonsense QA**:

Leverage KG to provide knowledge which is not stated explicitly in the context.

1. Extract the paths/subgraph localized at the entities mentioned in the context from KG.
2. Encode the paths/subgraph.



Where does a child likely sit at a desk?

A. Schoolroom *
B. Furniture store
C. Patio
D. Office building
E. Library

- Previous works on encoding paths/sub-graph
  - **Path-based Modeling**

1. Model the relational paths with sequence model.
2. Use attention to aggregate the paths.

Interpretable, but not scalable.

  - **Relational Graph NN**

Model the subgraph with message passing.

Scalable, but lack transparency

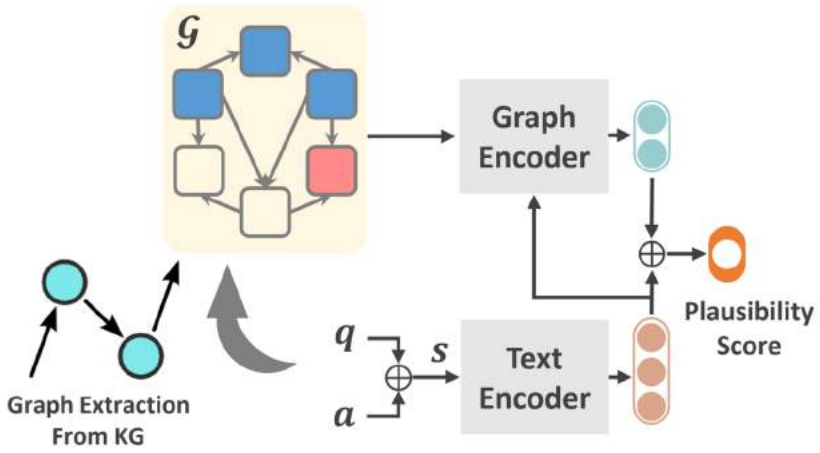
Key idea: Modeling All Paths Directly in Graph Networks!

Reasoning Pipeline

1. **Text Encoder**: Understand the textual input (question + answer choice).

2. **Graph Encoder**: Reason over the contextual subgraphs.

3. **Classifier**: Integrate the output from text/graph encoder to give a plausibility score.



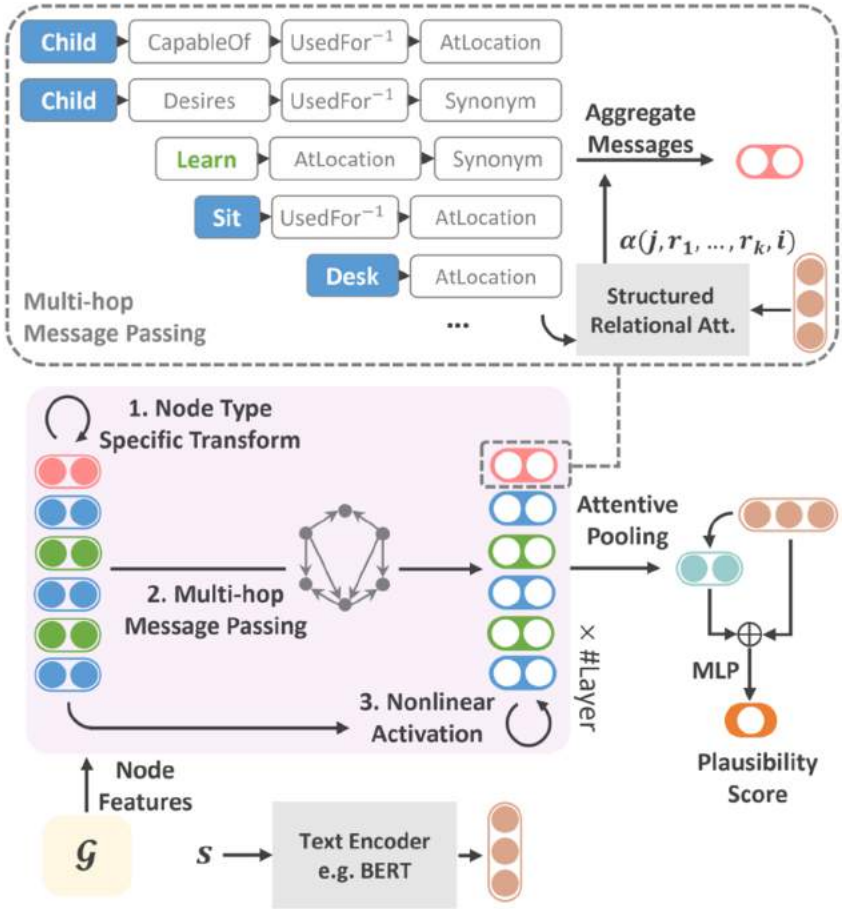
Our Method for Encoding KG

Goal: To combine both interpretability (path-based modeling) and scalability (GNN).

How: Endow GNN with the capability to model paths directly.

1. Multi-Hop Message Passing
- We extend message passing in GNN to k-hop paths modeling.

2. Structured Relational Attention
- Incoming message for a node is aggregated by attention mechanism.



Results

Methods	Single	Ensemble
RoBERTa†	72.1	72.5
RoBERTa + KEDGN†	72.5	74.4
RoBERTa + KE†	73.3	-
RoBERTa + HyKAS 2.0† (Ma et al., 2019)	73.2	-
RoBERTa + FreeLB† (Zhu et al., 2020)	72.2	73.1
XLNet + DREAM†	66.9	73.3
XLNet + GR† (Lv et al., 2019)	75.3	-
ALBERT† (Lan et al., 2019)	-	**76.5**
RoBERTa + MHGRN (K = 2)	75.4	**76.5**

CommonsenseQA's Leaderboard

Methods	Dev (%)	Test (%)
T5-3B† (Raffel et al., 2019)	-	83.20
UnifiedQA† (Khashabi et al., 2020)	-	**87.20**
RoBERTa-Large (w/o KG)	66.76 (±1.14)	64.80 (±2.37)
+ RGCN	64.65 (±1.96)	62.45 (±1.57)
+ GconAttn	66.85 (±1.82)	64.75 (±1.48)
+ RN (1-hop)	64.85 (±1.11)	63.65 (±2.31)
+ RN (2-hop)	67.00 (±0.71)	65.20 (±1.18)
+ MHGRN (K = 3)	68.10 (±1.02)	**66.85** (±1.19)
AristoRoBERTaV7†	79.2	77.8
+ MHGRN (K = 3)	78.6	**80.6**

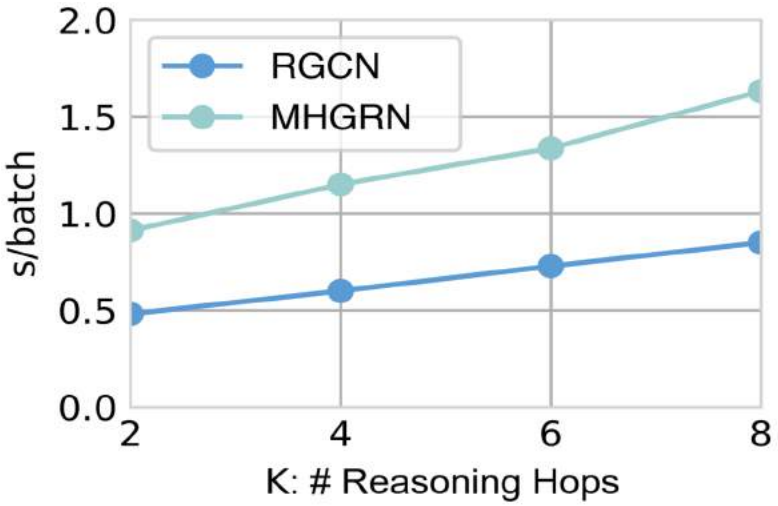
OpenBookQA's Leaderboard

Code: https://github.com/INK-USC/MHGRN

Results

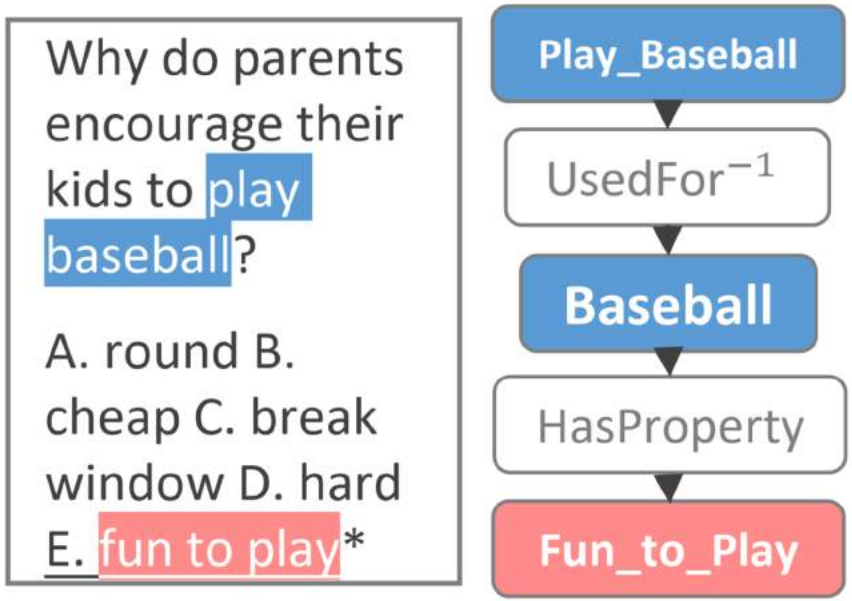
Scalability

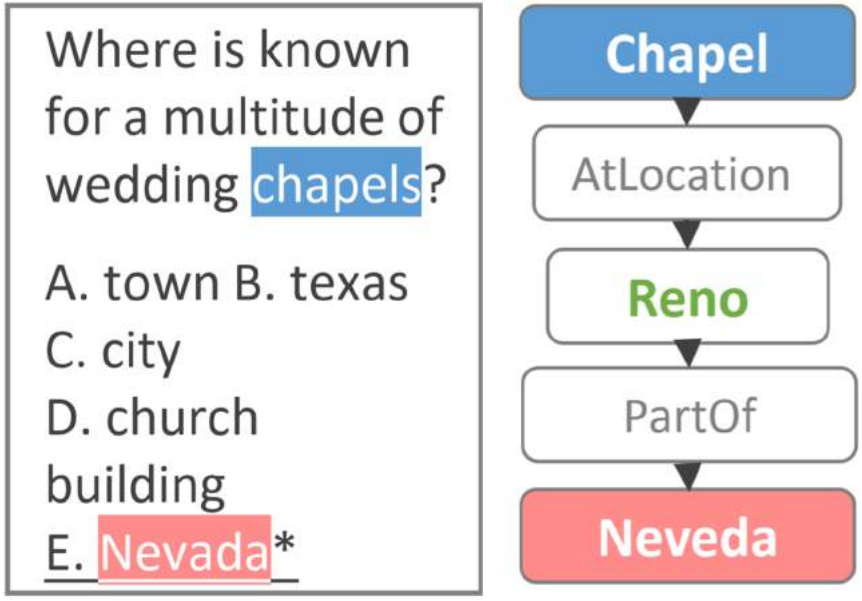
Model	Time	Space
	$\mathcal{G}$ is a dense graph	
K-hop KagNet	$\mathcal{O}\left(m^K n^{K+1} K\right)$	$\mathcal{O}\left(m^K n^{K+1} K\right)$
K-layer RGCN	$\mathcal{O}\left(mn^2 K\right)$	$\mathcal{O}\left(mnK\right)$
MHGRN	$\mathcal{O}\left(m^2 n^2 K\right)$	$\mathcal{O}\left(mnK\right)$
$\mathcal{G}$ is a sparse graph with maximum node degree $\Delta \ll n$		
K-hop KagNet	$\mathcal{O}\left(m^K nK\Delta^K\right)$	$\mathcal{O}\left(m^K nK\Delta^K\right)$
K-layer RGCN	$\mathcal{O}\left(mnK\Delta\right)$	$\mathcal{O}\left(mnK\right)$
MHGRN	$\mathcal{O}\left(m^2 nK\Delta\right)$	$\mathcal{O}\left(mnK\right)$



Results

Interpretability





Part II

CommonGen: A Constrained Text Generation Challenge for Generative Commonsense Reasoning

https://inklab.usc.edu/CommonGen/

Bill Yuchen Lin[♥] **Wangchunshu Zhou**[♥] **Ming Shen**[♥] **Pei Zhou**[♥]
Chandra Bhagavatula[♠] **Yejin Choi**[♠♦] **Xiang Ren**[♥]
[♥]University of Southern California [♠]Allen Institute for Artificial Intelligence
[♦]Paul G. Allen School of Computer Science & Engineering, University of Washington

What is CommonGen?

- Most current tasks for machine commonsense focus on discriminative reasoning.
  - CommonsenseQA, SWAG.

- Humans not only use commonsense knowledge for understanding text, but also for generating sentences.

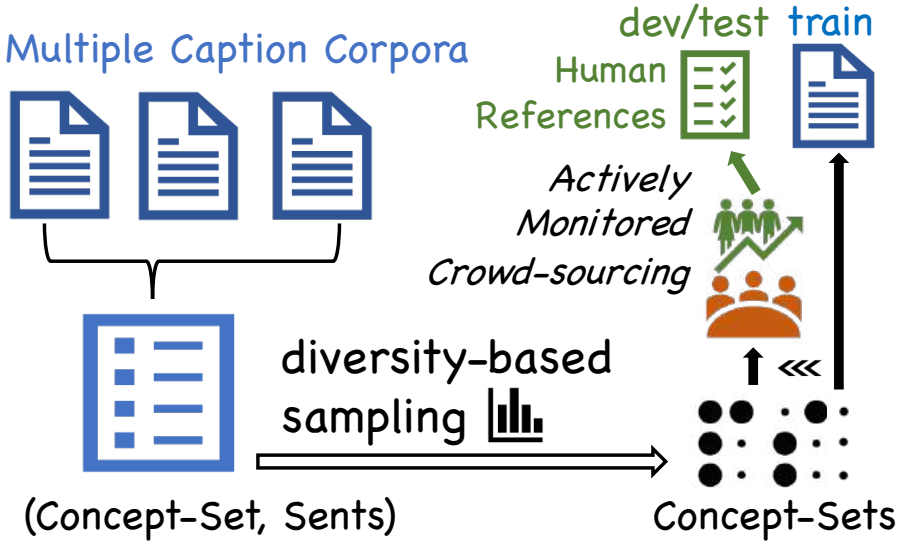
Input:

-A set of common concepts (actions & objects)

Output:

-A sentence that describes an everyday scenario the given concepts.

Construction



Statistics	Train	Dev	Test
# Concept-Sets	**32,651**	**993**	**1,497**
-Size = 3	25,020	493	-
-Size = 4	4,240	250	747
-Size = 5	3,391	250	750
# Sentences	67,389	4,018	6,042
per Concept-Set	2.06	4.04	4.04
Average Length	10.54	11.55	13.34
# Unique Concepts	4,697	766	1,248
# Unique Concept-Pairs	59,125	3,926	8,777
# Unique Concept-Triples	50,713	3,766	9,920
% Unseen Concepts	-	6.53%	8.97%
% Unseen Concept-Pairs	-	96.31%	100.00%
% Unseen Concept-Triples	-	99.60%	100.00%

Why is it hard? Two key Challenges of CommonGen

(1) Relational knowledge are **latent** and **compositional**.

{ exercise, rope, wall, tie, wave }

Underlying Relational Commonsense Knowledge

(exercise, HasSubEvent , releasing energy)

(rope, UsedFor, tying something)

(releasing energy, HasPrerequisite, motion)

(wave, IsA, motion) ; (rope, UsedFor, waving)

The motion costs more energy if ropes are tied to a wall.

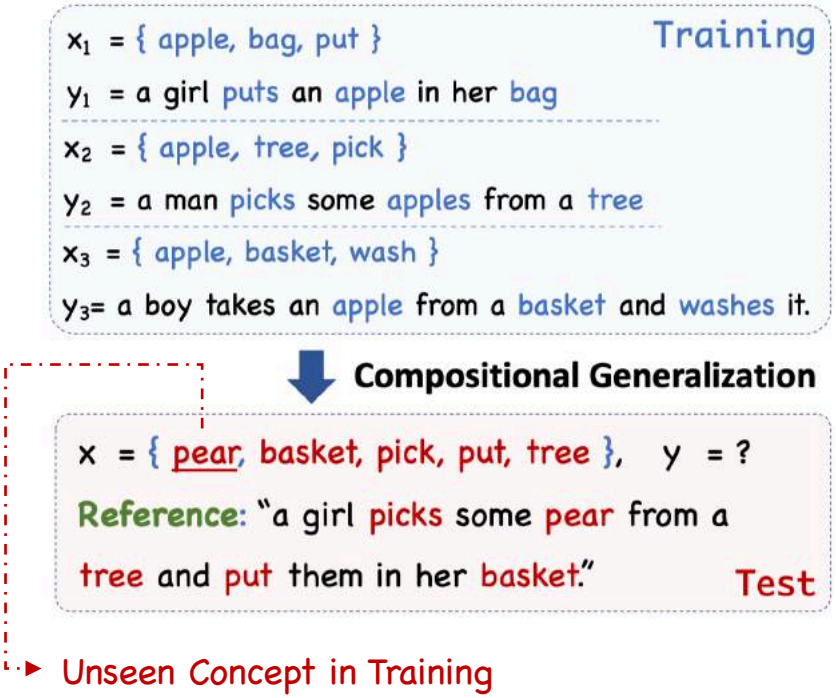
Relational Reasoning for Generation

A woman in a gym exercises by waving ropes tied to a wall.

Category	Relations	1-hop	2-hop
Spatial knowledge	AtLocation, LocatedNear	9.40%	39.31%
Object properties	UsedFor,CapableOf,PartOf, ReceivesAction,MadeOf, FormOf, HasProperty,HasA	9.60%	44.04%
Human behaviors	CausesDesire,MotivatedBy, Desires,NotDesires,Manner	4.60%	19.59%
Temporal knowledge	Subevent, Prerequisite, First/Last-Subevent	1.50%	24.03%
General	RelatedTo, Synonym, DistinctFrom, IsA, HasContext,SimilarTo	74.89%	69.65%

Why is it hard? Two key Challenges of CommonGen

(2) Compositional Generalization for unseen concept compounds.



Statistics	Train	Dev	Test
# Concept-Sets	**32,651**	**993**	**1,497**
-Size = 3	25,020	493	-
-Size = 4	4,240	250	747
-Size = 5	3,391	250	750
# Sentences	67,389	4,018	6,042
per Concept-Set	2.06	4.04	4.04
Average Length	10.54	11.55	13.34
# Unique Concepts	4,697	766	1,248
# Unique Concept-Pairs	59,125	3,926	8,777
# Unique Concept-Triples	50,713	3,766	9,920
% Unseen Concepts	-	6.53%	8.97%
% Unseen Concept-Pairs	-	96.31%	100.00%
% Unseen Concept-Triples	-	99.60%	100.00%

Experimental Results

Model \ Metrics	ROUGE-2/L		BLEU-3/4		METEOR	CIDEr	SPICE	Coverage
bRNN-CopyNet (Gu et al., 2016)	7.61	27.79	10.70	5.70	15.80	4.79	15.00	51.15
Trans-CopyNet	8.78	28.08	11.90	7.10	15.50	4.61	14.60	49.06
MeanPooling-CopyNet	9.66	31.14	10.70	6.10	16.40	5.06	17.20	55.70
LevenTrans. (Gu et al., 2019)	10.58	32.23	19.70	11.60	20.10	7.54	19.00	63.81
ConstLeven. (Susanto et al., 2020)	11.82	33.04	18.90	10.10	24.20	10.51	22.20	94.51
GPT-2 (Radford et al., 2019)	17.18	39.28	30.70	21.10	26.20	12.15	25.90	79.09
BERT-Gen (Bao et al., 2020)	18.05	40.49	30.40	21.10	27.30	12.49	27.30	86.06
UniLM (Dong et al., 2019)	21.48	**43.87**	38.30	27.70	29.70	14.85	30.20	89.19
UniLM-v2 (Bao et al., 2020)	18.24	40.62	31.30	22.10	28.10	13.10	28.10	89.13
BART (Lewis et al., 2019)	**22.23**	41.98	36.30	26.30	**30.90**	13.92	30.60	**97.35**
T5-Base (Raffel et al., 2019)	14.57	34.55	26.00	16.40	23.00	9.16	22.00	76.67
T5-Large (Raffel et al., 2019)	22.01	42.97	**39.00**	**28.60**	30.10	**14.96**	**31.60**	95.29
Human Performance	48.88	63.79	48.20	44.90	36.20	43.53	63.50	99.31

(1) Seq2seq models

(2) Fine-tuning pre-trained LMs

(3) Agreement

Manual Eval.

	C.Leven	GPT	BERT-G.	UniLM	BART	T5
Hit@1	3.2	21.5	22.3	21.0	26.3	**26.8**
Hit@3	18.2	63.0	59.5	69.0	69.0	**70.3**
Hit@5	51.4	95.5	95.3	96.8	96.3	**97.8**

Case Study & Transfer Learning

Concept-Set: { hand, sink, wash, soap }

[bRNN-CopyNet]: a hand works in the sink .

[MeanPooling-CopyNet]: the hand of a sink being washed up

[ConstLeven]: a hand strikes a sink to wash from his soap.

[GPT-2]: hands washing soap on the sink.

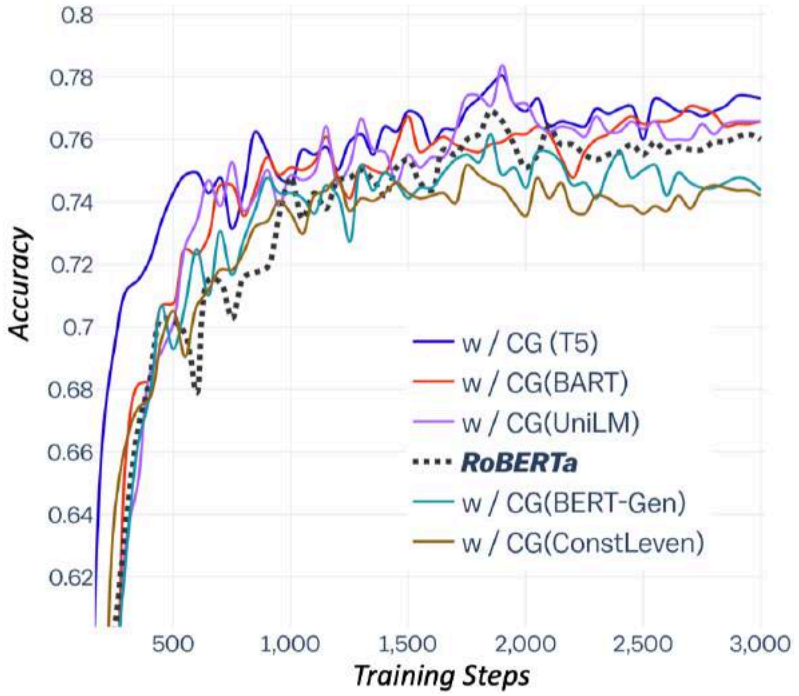
[BERT-Gen]: a woman washes her hands with a sink of soaps.

[UniLM]: hands washing soap in the sink

[BART]: a man is washing his hands in a sink with soap and washing them with hand soap.

[T5]: hand washed with soap in a sink.

1. A girl is washing her hands with soap in the bathroom sink.
2. I will wash each hand thoroughly with soap while at the sink.
3. The child washed his hands in the sink with soap.
4. A woman washes her hands with hand soap in a sink.
5. The girl uses soap to wash her hands at the sink.



Learning curve for the transferring study (acc on dev). We use trained CommonGen models to generate choice-specific context for the CommonsenseQA task.

Learning with Natural Language Explanations

Sentiment on ENT is positive or negative?

x_1: There was a long wait for a table outside, but it was a little too hot in the sun anyway so our ENT was very nice.

Users' natural language explanations

Positive, because the words "*very nice*" is within 3 words after the ENT.

Relation between ENT1 and ENT2?

x_2: Officials in Mumbai said that the two suspects, David Headley, and ENT1, who was born in Pakistan but is a ENT2 citizen, both visited Mumbai before the attacks.

per: nationality, because the words "*is a*" appear right before ENT2 and the word "*citizen*" is right after ENT2.

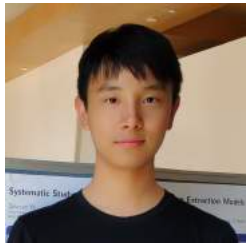
(Wang et al., ICLR'20) http://inklab.usc.edu/project-NExT

Students







Research Partnership

Microsoft Academic

SPARK AT STANFORD

Semantic Scholar

BioPortal

Collaborators

Dan MacFarland, Sociology, Stanford University
Jure Leskovec, Computer Science, Stanford University
Dan Jurafsky, Computer Science, Stanford University
Jiawei Han, Computer Science, UIUC
Morteza Dehghani, Psychology, USC
Kennth Yates, Clinical Education, USC
Craig Knoblock, USC ISI
Curt Langlotz, Bioinformatics, Stanford University
Kuansan Wang, Microsoft Academic
Leonardo Neves, Snap Research
Mark Musen, Bioinformatics, Stanford University

Funding

NSF

DARPA

IARPA
BE THE FUTURE

J.P.Morgan

Google

SCHMIDT FAMILY FOUNDATION

amazon

Adobe

Thank you!

- USC Intelligence and Knowledge Discovery (INK) Lab
  - http://inklab.usc.edu/

- Code: https://github.com/INK-USC
  - xiangren@usc.edu
    - @xiangrenNLP