Commonsense Reasoning: Models and New Challenges

Sean (Xiang) Ren

Department of Computer Science
Information Science Institute
USC
http://inklab.usc.edu
human-level performance on reading comprehension on SQuAD (Stanford QA dataset)

super-human performance on speech recognition

Google neural machine translation

super-human performance on image captioning

super-human performance on object recognition
Done Solving AI?

- 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
- 2015: Super-human performance on object recognition
Solving a “dataset” vs the underlying “task”

Image credit: Yejin Choi
Why Commonsense Knowledge?

**TODAY**
Narrow Artificial Intelligence

- **AI Application** (Robot, Assistant, Analytic)
- Narrow AI
  Carefully train or program the system for every possible situation

**TOMORROW**
Machine Common Sense

- **AI Application** (Robot, Assistant, Analytic)
- Commonsense Service
  Where should I sit to saw off the limb of this tree?

**FUTURE**
General Artificial Intelligence

- Human-Level AI
  Sit between the trunk and the cut point

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)  KagNet: Knowledge-Aware Graph Networks
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)
Knowledge-Aware Reasoning

Where do adults use glue sticks?
A: classroom  B: office  C: desk drawer

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

KagNet: Knowledge-Aware Graph Networks

(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
    - Path pruning by length ($\leq 5$ nodes) and embedding-based metric.
(2) Path-based Relational Graph Encoder

\[ g = \sum_{i,j} \frac{R_{i,j} \cdot T_{i,j}}{|C_q| \times |C_a|} \]

- **Encoding Unlabeled Schema Graphs** $g$
- **Graph Conv. Nets** only looking at the plain graph structures (i.e. ignore relations)

**Statement Vector** $S$

**LSTM Path Encoder**

- Modeling Relational Paths $P_{i,j}$ between $c_i^{(q)}$ and $c_j^{(a)}$

- Encoding the $k$-th path between $c_i^{(q)}$ and $c_j^{(a)}$

**LSTM**($P_{i,j}[k]$)

**R**

$R_{i,j} = \frac{1}{|P_{i,j}|} \sum_k \text{LSTM}(P_{i,j}[k])$

**T**

$T_{i,j} = \text{MLP}([s; c_q^{(i)}; c_a^{(j)}])$

KagNet: Knowledge-Aware Graph Networks

(Bill Yuchen Lin et al. 2019)
(3) w/ Hierarchical Path-based Attention

• Two average pooling:
  • Assuming all QA-concept pairs are equally important
    \[ g = \frac{\sum_{i,j} [R_{i,j} ; T_{i,j}]}{|C_q| \times |C_a|} \]
  • Assuming all paths are equally relevant
    \[ R_{i,j} = \frac{1}{|P_{i,j}|} \sum_k \text{LSTM}(P_{i,j}[k]) \]

• Modeling the two-level importance as latent weights:

  \[ \alpha(i,j,k) = T_{i,j} W_1 \text{LSTM}(P_{i,j}[k]) \]
  \[ \hat{\alpha}(i,j,:) = \text{SoftMax}(\alpha(i,j,:)) \]
  \[ \hat{R}_{i,j} = \sum_k \hat{\alpha}(i,j,k) \cdot \text{LSTM}(P_{i,j}[k]) \]

  Path-Level Attention (attending on semantic space)

  \[ \beta(i,j) = s W_2 T_{i,j} \]
  \[ \hat{\beta}(:,:) = \text{SoftMax}(\beta(:,:)) \]
  \[ \hat{g} = \sum_{i,j} \hat{\beta}(i,j) \hat{R}_{i,j} ; T_{i,j} \]

  ConceptPair-Level Attention (attending on statement)
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
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

KagNet

BERT-Large

CSQA

59.01% vs 56.53%

SWAG

53.51% vs 51.23%

WSC

(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

(Bill Yuchen Lin et al. 2019)
Multi-Hop Graph Relation Networks for Knowledge-Aware Question Answering


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.

• Previous works on encoding paths/sub-graph
  o Path-based Modeling
    1. Model the relational paths with sequence model.
    2. Use attention to aggregate the paths.
    Interpretable, but not scalable.
  o 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

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

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

CommonsenseQA’s Leaderboard

OpenBookQA’s Leaderboard

Code: [https://github.com/INK-USC/MHGRN](https://github.com/INK-USC/MHGRN)
## Results

### Scalability

<table>
<thead>
<tr>
<th>Model</th>
<th>Time</th>
<th>Space</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\mathcal{G}$ is a dense graph</td>
<td></td>
</tr>
<tr>
<td>$K$-hop KAGNet</td>
<td>$O\left(m^K n^{K+1}K\right)$</td>
<td>$O\left(m^K n^{K+1}K\right)$</td>
</tr>
<tr>
<td>$K$-layer RGCN</td>
<td>$O\left(mn^2K\right)$</td>
<td>$O\left(mnK\right)$</td>
</tr>
<tr>
<td>MHGRN</td>
<td>$O\left(m^2n^2K\right)$</td>
<td>$O\left(mnK\right)$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model</th>
<th>Time</th>
<th>Space</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\mathcal{G}$ is a sparse graph with maximum node degree $\Delta \ll n$</td>
<td></td>
</tr>
<tr>
<td>$K$-hop KAGNet</td>
<td>$O\left(m^K n K^\Delta \Delta^K\right)$</td>
<td>$O\left(m^K n K^\Delta \Delta^K\right)$</td>
</tr>
<tr>
<td>$K$-layer RGCN</td>
<td>$O\left(mnK \Delta\right)$</td>
<td>$O\left(mnK\right)$</td>
</tr>
<tr>
<td>MHGRN</td>
<td>$O\left(m^2nK \Delta\right)$</td>
<td>$O\left(mnK\right)$</td>
</tr>
</tbody>
</table>

![Graph showing scalability](image-url)
Results

Interpretability

Why do parents encourage their kids to play baseball?
A. round  B. cheap  C. break window  D. hard  E. fun to play*

Where is known for a multitude of wedding chapels?
A. town  B. texas  C. city  D. church building  E. Nevada*
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.

**Concept-Set**: a collection of objects/actions.

- dog, frisbee, catch, throw

**Generative Commonsense Reasoning**

**Expected Output**: everyday scenarios covering all given concepts.

- A dog leaps to catch a thrown frisbee.
- The dog catches the frisbee when the boy throws it.
- A man throws away his dog’s favorite frisbee expecting him to catch it in the air.

**Input**:
- A set of common concepts (actions & objects)

**Output**:
- A sentence that describes an everyday scenario the given concepts.
Construction

Multiple Caption Corpora

Human References

Actively Monitored Crowd-sourcing

dev/test train

(Concept-Set, Sents)

Concept-Sets

diversity-based sampling

<table>
<thead>
<tr>
<th>Statistics</th>
<th>Train</th>
<th>Dev</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td># Concept-Sets</td>
<td>32,651</td>
<td>993</td>
<td>1,497</td>
</tr>
<tr>
<td>- Size = 3</td>
<td>25,020</td>
<td>493</td>
<td>-</td>
</tr>
<tr>
<td>- Size = 4</td>
<td>4,240</td>
<td>250</td>
<td>747</td>
</tr>
<tr>
<td>- Size = 5</td>
<td>3,391</td>
<td>250</td>
<td>750</td>
</tr>
<tr>
<td># Sentences per Concept-Set</td>
<td>67,389</td>
<td>4,018</td>
<td>6,042</td>
</tr>
<tr>
<td>Average Length</td>
<td>2.06</td>
<td>4.04</td>
<td>4.04</td>
</tr>
<tr>
<td></td>
<td>10.54</td>
<td>11.55</td>
<td>13.34</td>
</tr>
<tr>
<td># Unique Concepts</td>
<td>4,697</td>
<td>766</td>
<td>1,248</td>
</tr>
<tr>
<td># Unique Concept-Pairs</td>
<td>59,125</td>
<td>3,926</td>
<td>8,777</td>
</tr>
<tr>
<td># Unique Concept-Triples</td>
<td>50,713</td>
<td>3,766</td>
<td>9,920</td>
</tr>
<tr>
<td>% Unseen Concepts</td>
<td>-</td>
<td>6.53%</td>
<td>8.97%</td>
</tr>
<tr>
<td>% Unseen Concept-Pairs</td>
<td>-</td>
<td>96.31%</td>
<td>100.00%</td>
</tr>
<tr>
<td>% Unseen Concept-Triples</td>
<td>-</td>
<td>99.60%</td>
<td>100.00%</td>
</tr>
</tbody>
</table>
Why is it hard? 
Two key Challenges of CommonGen

(1) Relational knowledge are latent and compositional.

![Diagram showing relation between exercise, rope, wall, tie, and 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.

<table>
<thead>
<tr>
<th>Category</th>
<th>Relations</th>
<th>1-hop</th>
<th>2-hop</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spatial knowledge</td>
<td>AtLocation, LocatedNear</td>
<td>9.40%</td>
<td>39.31%</td>
</tr>
<tr>
<td>Object properties</td>
<td>UsedFor, CapableOf, PartOf, ReceivesAction, MadeOf, FormOf, HasProperty, HasA</td>
<td>9.60%</td>
<td>44.04%</td>
</tr>
<tr>
<td>Human behaviors</td>
<td>CausesDesire, MotivatedBy, Desires, NotDesires, Manner</td>
<td>4.60%</td>
<td>19.59%</td>
</tr>
<tr>
<td>Temporal knowledge</td>
<td>Subevent, Prerequisite, First/Last-Subevent</td>
<td>1.50%</td>
<td>24.03%</td>
</tr>
<tr>
<td>General</td>
<td>RelatedTo, Synonym, DistinctFrom, IsA, HasContext, SimilarTo</td>
<td>74.89%</td>
<td>69.65%</td>
</tr>
</tbody>
</table>
Why is it hard?
Two key Challenges of CommonGen

(2) Compositional Generalization for unseen concept compounds.

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

\[ x_1 = \{ \text{apple, bag, put} \} \quad \text{Training} \]
\[ y_1 = \text{a girl puts an apple in her bag} \]
\[ x_2 = \{ \text{apple, tree, pick} \} \]
\[ y_2 = \text{a man picks some apples from a tree} \]
\[ x_3 = \{ \text{apple, basket, wash} \} \]
\[ y_3 = \text{a boy takes an apple from a basket and washes it.} \]

Reference: "a girl picks some pear from a tree and put them in her basket."
# Experimental Results

<table>
<thead>
<tr>
<th>Model</th>
<th>ROUGE-2/L</th>
<th>BLEU-3/4</th>
<th>METEOR</th>
<th>CIDEr</th>
<th>SPICE</th>
<th>Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>bRNN-CopyNet (Gu et al., 2016)</td>
<td>7.61</td>
<td>27.79</td>
<td>10.70</td>
<td>5.70</td>
<td>15.80</td>
<td>4.79</td>
</tr>
<tr>
<td>Trans-CopyNet</td>
<td>8.78</td>
<td>28.08</td>
<td>11.90</td>
<td>7.10</td>
<td>15.50</td>
<td>4.61</td>
</tr>
<tr>
<td>MeanPooling-CopyNet</td>
<td>9.66</td>
<td>31.14</td>
<td>10.70</td>
<td>6.10</td>
<td>16.40</td>
<td>5.06</td>
</tr>
<tr>
<td>LevenTrans. (Gu et al., 2019)</td>
<td>10.58</td>
<td>32.23</td>
<td>19.70</td>
<td>11.60</td>
<td>20.10</td>
<td>7.54</td>
</tr>
<tr>
<td>ConstLeven. (Susanto et al., 2020)</td>
<td>11.82</td>
<td>33.04</td>
<td>18.90</td>
<td>10.10</td>
<td>24.20</td>
<td>10.51</td>
</tr>
<tr>
<td>GPT-2 (Radford et al., 2019)</td>
<td>17.18</td>
<td>39.28</td>
<td>30.70</td>
<td>21.10</td>
<td>26.20</td>
<td>12.15</td>
</tr>
<tr>
<td>BERT-Gen (Bao et al., 2020)</td>
<td>18.05</td>
<td>40.49</td>
<td>30.40</td>
<td>21.10</td>
<td>27.30</td>
<td>12.49</td>
</tr>
<tr>
<td>UniLM (Dong et al., 2019)</td>
<td>21.48</td>
<td><strong>43.87</strong></td>
<td>38.30</td>
<td>27.70</td>
<td>29.70</td>
<td>14.85</td>
</tr>
<tr>
<td>UniLM-v2 (Bao et al., 2020)</td>
<td>18.24</td>
<td>40.62</td>
<td>31.30</td>
<td>22.10</td>
<td>28.10</td>
<td>13.10</td>
</tr>
<tr>
<td>BART (Lewis et al., 2019)</td>
<td><strong>22.23</strong></td>
<td>41.98</td>
<td>36.30</td>
<td>26.30</td>
<td><strong>30.90</strong></td>
<td>13.92</td>
</tr>
<tr>
<td>T5-Base (Raffel et al., 2019)</td>
<td>14.57</td>
<td>34.55</td>
<td>26.00</td>
<td>16.40</td>
<td>23.00</td>
<td>9.16</td>
</tr>
<tr>
<td>T5-Large (Raffel et al., 2019)</td>
<td>22.01</td>
<td>42.97</td>
<td><strong>39.00</strong></td>
<td><strong>28.60</strong></td>
<td>30.10</td>
<td><strong>14.96</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Human Performance</th>
<th>C.Leven</th>
<th>GPT</th>
<th>BERT-G.</th>
<th>UniLM</th>
<th>BART</th>
<th>T5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hit@1</td>
<td>3.2</td>
<td>21.5</td>
<td>22.3</td>
<td>21.0</td>
<td>26.3</td>
<td><strong>26.8</strong></td>
</tr>
<tr>
<td>Hit@3</td>
<td>18.2</td>
<td>63.0</td>
<td>59.5</td>
<td>69.0</td>
<td>69.0</td>
<td><strong>70.3</strong></td>
</tr>
<tr>
<td>Hit@5</td>
<td>51.4</td>
<td>95.5</td>
<td>95.3</td>
<td><strong>96.8</strong></td>
<td>96.3</td>
<td><strong>97.8</strong></td>
</tr>
</tbody>
</table>
Case Study & Transfer Learning

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

\textbf{[bRNN-CopyNet]}: a hand works in the sink.
\textbf{[MeanPooling-CopyNet]}: the hand of a sink being washed up
\textbf{[ConstLeven]}: a hand strikes a sink to wash from his soap.
\textbf{[GPT-2]}: hands washing soap on the sink.
\textbf{[BERT-Gen]}: a woman washes her hands with a sink of soaps.
\textbf{[UniLM]}: hands washing soap in the sink
\textbf{[BART]}: a man is washing his hands in a sink with soap and washing them with hand soap.
\textbf{[T5]}: hand washed with soap in a sink.

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

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

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
Kenneth 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

Research Partnership
Thank you!

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

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