Question Answering

Alon Talmor
12/4/2019
What is Question Answering?

Which papers on Question Answering were published last month?

Find me hotels, near a coastline, that are dog friendly and sort the results by price

How do I renovate my home?
Turing test

During the Turing test, the human questioner asks a series of questions to both respondents. After the specified time, the questioner tries to decide which terminal is operated by the human respondent and which terminal is operated by the computer.

- Red: Question to respondents
- Black: Answers to questioner

Computer respondent

Human questioner

Human respondent
What is Intelligence?
<table>
<thead>
<tr>
<th>Researcher</th>
<th>Quotation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alfred Binet</td>
<td>Judgment, otherwise called &quot;good sense&quot;, &quot;practical sense&quot;, &quot;initiative&quot;, the faculty of adapting one's self to circumstances ... auto-critique.</td>
</tr>
<tr>
<td>David Wechsler</td>
<td>The aggregate or global capacity of the individual to act purposefully, to think rationally, and to deal effectively with his environment.</td>
</tr>
<tr>
<td>Lloyd Humphreys</td>
<td>...the resultant of the process of acquiring, storing in memory, retrieving, combining, comparing, and using in new contexts information and conceptual skills.</td>
</tr>
<tr>
<td>Howard Gardner</td>
<td>To my mind, a human intellectual competence must entail a set of skills of problem solving — enabling the individual to resolve genuine problems or difficulties that he or she encounters and, when appropriate, to create an effective product — and must also entail the potential for finding or creating problems — and thereby laying the groundwork for the acquisition of new knowledge.</td>
</tr>
<tr>
<td>Linda Gottfredson</td>
<td>The ability to deal with cognitive complexity.</td>
</tr>
<tr>
<td>Sternberg &amp; Salter</td>
<td>Goal-directed adaptive behavior.</td>
</tr>
<tr>
<td>Reuven Feuerstein</td>
<td>The theory of Structural Cognitive Modifiability describes intelligence as &quot;the unique propensity of human beings to change or modify the structure of their cognitive functioning to adapt to the changing demands of a life situation&quot;.</td>
</tr>
<tr>
<td>Legg &amp; Hutter</td>
<td>A synthesis of 70+ definitions from psychology, philosophy, and AI researchers: &quot;Intelligence measures an agent's ability to achieve goals in a wide range of environments&quot;, which has been mathematically formalized.</td>
</tr>
<tr>
<td>Alexander Wissner-Gross</td>
<td>( F = T \lor S _ \lor S )</td>
</tr>
</tbody>
</table>

From "Mainstream Science on Intelligence" (1994), an op-ed statement in the Wall Street Journal signed by fifty-two researchers (out of 131 total invited to sign):  

A very general mental capability that, among other things, involves the ability to reason, plan, solve problems, think abstractly, comprehend complex ideas, learn quickly and learn from experience. It is not merely book learning, a narrow academic skill, or test-taking smarts. Rather, it reflects a broader and deeper capability for comprehending our surroundings — "catching on," "making sense" of things, or "figuring out" what to do.  

From *Intelligence: Knowns and Unknowns* (1995), a report published by the Board of Scientific Affairs of the American Psychological Association:  

Individuals differ from one another in their ability to understand complex ideas, to adapt effectively to the environment, to learn from experience, to engage in various forms of reasoning, to overcome obstacles by taking thought. Although these individual differences can be substantial, they are never entirely consistent: a given person's intellectual performance will vary on different occasions, in different domains, as judged by different criteria. Concepts of "intelligence" are attempts to clarify and organize this complex set of phenomena. Although considerable clarity has been achieved in some areas, no such conceptualization has yet answered all the important questions, and none commands universal assent. Indeed, when two dozen prominent theorists were recently asked to define intelligence, they gave two dozen, somewhat different, definitions.
What we’ll talk about

• Commonsense Question Answering
• Reading Comprehension
• Complex Questions
• Reasoning and Probing
• Multi-Modal Question Answering
• What’s are we still missing?
Commonsense Question Answering

John opened the door for Dan and he entered, who entered?
A) John,  B) Dan

A person starts tickling a stranger, how does the stranger respond?
A) Laugh, B) Cry,  C) Surprised
CommonsenseQA: A Question Answering Challenge Targeting Commonsense Knowledge

Alon Talmor*, Jonathan Herzig*, Nicholas Lourie, Jonathan Berant

* Equal contribution
The trophy would not fit in the brown suitcase because it was too big.

What was too big?
A. The trophy
B. The suitcase

Winograd Schema Challenge, T. Winograd, 1972

VCR Zellers et al., 2019

ATOMIC Sap et al., 2019
Problem: Reporting Bias

Conversational maxim of quantity: communication should be as informative as necessary but no more than is required. (H. Paul Grice, 1975)

→ common sense is usually not stated.

“I took my fork and carefully picked up my food slowly moving it towards my mouth. I then opened my mouth and started chewing my food, chewing, chewing, chewing ...”
Possible Solutions to reporting bias
Can we **test** if machines can **overcome** the reporting bias?

Human annotation

- Expensive
- Does not scale well

e.g. the Cyc project
(D. B. Lenat, 1995)

"I hadn’t slept in days." → A person normally sleeps at least daily.
(Gordon and Van Durme, 2013)

Machine implicitly learning

Or other methods?
Desiderata

Create a commonsense challenge that will:

1. Tap onto reporting bias, using human annotators
2. Have enough examples for fine-tuning large models
3. Have broad coverage
Broad coverage? But human annotators are biased!

Our Solution: Use ConceptNet as a "seed"

(Speer et al., 2017)
Multi-choice question answering task

Where on a river can you hold a cup upright to catch water on a sunny day?

- [ ] river
- [x] waterfall
- [x] bridge
- [x] valley

Correct answer: river
Contributions

• A new method for generating broad domain commonsense questions with human annotators.

• A new QA dataset, CommonsenseQA, containing 12,102 examples.

• An empirical evaluation of state-of-the-art NLU models on CommonsenseQA.
Dataset Generation and Analysis
Same relation → additional commonsense relations

Dataset Generation – Subgraph Sampling

river

AtLocation

bridge

AtLocation

waterfall

AtLocation

valley

AtLocation

question concept

answer concept
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
- ✗ bank
- ✗ island

I'm crossing the **river**, where am I?

- ✗ waterfall
- ✗ bridge
- ✗ valley
- ✗ stream
- ✗ bank
- ✗ island

Free-form distractor answer.

Choices overall.

Prevents memorization.

Overall **12,102** examples were generated.
<table>
<thead>
<tr>
<th>Relation</th>
<th>Formulated question example</th>
<th>Correct ans</th>
<th>Distractor</th>
</tr>
</thead>
<tbody>
<tr>
<td>AtLocation</td>
<td>Where would I not want a fox?</td>
<td>Hen house</td>
<td>England, Mountains</td>
</tr>
<tr>
<td>CausesDesire</td>
<td>What do parents encourage kids to do when they experience boredom?</td>
<td>Sleep, Travel</td>
<td></td>
</tr>
<tr>
<td>CapableOf</td>
<td>What can a child do to quietly pass the time?</td>
<td>Read book, clean room, ask questions</td>
<td></td>
</tr>
</tbody>
</table>
Q. Where are Rosebushes typically found outside of large buildings?

Building → Has parts 23% → Courtyard → Spatial 41% → Flowers → Is member of 17% → Rosebushes

Q. I want to use string to keep something from moving, how should I do it?

Something → Spatial 41% → String → Activity 8% → Tie around 23% → Keep from moving
Experiments and baselines
Baselines

• **VecSim**: Cosine Similarity based on GLoVe
• **ESIM** (Chen et al., 2016): Natural Language Inference +
• **BIDAF++** (Seo et al., 2016): Reading Comprehension +
• **GPT** (Radford et al., 2018): Generative Pretrained Transformer
• **BERT** (Devlin et al., 2018): Pre-training of Deep Bidirectional Transformers
• **Human Performance**
BERT/GPT Setup

Softmax

logit1

logit2

logit3

[cls] question [sep] answer1 [sep]  [cls] question [sep] answer2 [sep]  [cls] question [sep] answer3 [sep]
Results

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rand/Maj</td>
<td>20</td>
</tr>
<tr>
<td>VecSim+</td>
<td>29.1</td>
</tr>
<tr>
<td>BiDAF++</td>
<td>32</td>
</tr>
<tr>
<td>ESIM</td>
<td>34.1</td>
</tr>
<tr>
<td>GPT</td>
<td>45.5</td>
</tr>
<tr>
<td>BERT</td>
<td>55.9</td>
</tr>
<tr>
<td>Human</td>
<td>88.9</td>
</tr>
</tbody>
</table>
Learning curves BERT-large

Human Performance

70-75%
<table>
<thead>
<tr>
<th>Category</th>
<th>Formulated question example</th>
<th>Correct answer</th>
<th>Distractor</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Surface clues</td>
<td>If someone laughs after surprising them they have a good sense of what?</td>
<td>humor</td>
<td>laughter</td>
<td>77.7</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>55%</td>
</tr>
<tr>
<td>Negation / Antonym</td>
<td>Where might the stapler be if I cannot find it?</td>
<td>desk drawer</td>
<td>desktop</td>
<td>42.8</td>
</tr>
<tr>
<td>Conjunction</td>
<td>On a hot day what can you do to enjoy something cool and sweet?</td>
<td>eat ice cream</td>
<td>fresh cake</td>
<td>23.8</td>
</tr>
</tbody>
</table>

Baseline analysis - BERT
CommonsenseQA

tau-nlp.org/commonsenseqa

CommonsenseQA

Where on a river can you hold a cup upright to catch water in a sunny day?
- waterfall, bridge, valley, hillside, mountain

Where can I stand on a river to see a waterfall without getting wet?
- waterfall, bridge, valley, cliff, button

Can we drive up a river next to a waterfall?
- waterfall, bridge, valley, cliff, button

News:
10/02/2019: Version 1.1 is out! 12,247 examples, and 5 answer choices instead of 3.

Leaderboard, Version 1.11

<table>
<thead>
<tr>
<th>Model</th>
<th>Affiliation</th>
<th>Date</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td></td>
<td>03/10/2019</td>
<td>86.9</td>
</tr>
<tr>
<td>XINet + Graph Reasoning (single model)*</td>
<td>Microsoft Research Asia and Bing</td>
<td>08/24/2019</td>
<td>73.3</td>
</tr>
<tr>
<td>RoBERTa + KE (single model)</td>
<td>Alibaba DAMO NLP</td>
<td>10/18/2019</td>
<td>73.3</td>
</tr>
<tr>
<td>DREAM (ensemble model)</td>
<td>Microsoft Research Asia and Bing</td>
<td>10/11/2019</td>
<td>73.3</td>
</tr>
<tr>
<td>FreeB. RoBERTa (ensemble model)</td>
<td>Microsoft Dynamics 365 AI Research &amp; UVD</td>
<td>10/03/2019</td>
<td>73.1</td>
</tr>
<tr>
<td>RoBERTa (ensemble model)</td>
<td>Facebook AI</td>
<td>08/13/2019</td>
<td>72.5</td>
</tr>
<tr>
<td>FreeB. RoBERTa (single model)</td>
<td>Microsoft Dynamics 365 AI Research &amp; UVD</td>
<td>10/03/2019</td>
<td>72.2</td>
</tr>
<tr>
<td>RoBERTa + IR (single model)</td>
<td>Microsoft STCA-NLP team</td>
<td>08/23/2019</td>
<td>72.1</td>
</tr>
<tr>
<td>RoBERTa (single model)</td>
<td>Facebook AI</td>
<td>08/13/2019</td>
<td>72.1</td>
</tr>
<tr>
<td>RoBERTa + CSPT (single model)</td>
<td>Anonymous</td>
<td>08/13/2019</td>
<td>69.6</td>
</tr>
<tr>
<td>RoBERTa + KE (single model)</td>
<td>Anonymous</td>
<td>08/24/2019</td>
<td>68.4</td>
</tr>
<tr>
<td>DREAM (single model)</td>
<td>Microsoft Research Asia and Bing</td>
<td>08/08/2019</td>
<td>66.9</td>
</tr>
<tr>
<td>CSRKG (AI2 IR, single model)</td>
<td>Microsoft Dynamics 365 AI Research</td>
<td>07/19/2019</td>
<td>65.3</td>
</tr>
<tr>
<td>AristoteL07 (single model)</td>
<td>Arista team at Allen Institute for AI</td>
<td>07/18/2019</td>
<td>64.6</td>
</tr>
<tr>
<td>H+KAS (single model)</td>
<td>Bosch Research and Technology Center (Pittsburgh)</td>
<td>08/20/2019</td>
<td>62.5</td>
</tr>
<tr>
<td>BERT+OMICS (single model)</td>
<td>Allen Institute for Artificial Intelligence - Israel</td>
<td>07/30/2019</td>
<td>62.5</td>
</tr>
<tr>
<td>BERT+AMS (single model)</td>
<td>Alibaba DAMO Speech Lab</td>
<td>07/10/2019</td>
<td>62.2</td>
</tr>
<tr>
<td>CSRKG (single model)</td>
<td>Microsoft Dynamics 365 AI Research</td>
<td>06/28/2019</td>
<td>61.8</td>
</tr>
<tr>
<td>BECON (ensemble)</td>
<td>Singapore University of Technology and Design</td>
<td>07/01/2019</td>
<td>59.6</td>
</tr>
<tr>
<td>KagNet (single model**)</td>
<td>INK lab @ USC/ISI</td>
<td>05/14/2019</td>
<td>58.2</td>
</tr>
<tr>
<td>CoSE (single model)</td>
<td>Salesforce Research</td>
<td>04/12/2019</td>
<td>57.9</td>
</tr>
<tr>
<td>BECON (single model)</td>
<td>Singapore University of Technology and Design</td>
<td>06/17/2019</td>
<td>57.9</td>
</tr>
</tbody>
</table>
CrowdSense - crowdsense.apps.allenai.org

Help the AI learn common sense!

Question:
What is True? Artificial Intelligence knows:

Answer 1:
some common sense

Answer 2:
no common sense

Answer 3:
all common sense

AI answer: some common sense
But what if we want to ask questions about a specific paragraph?
Marilyn Monroe was found dead of a barbiturate overdose in the early morning hours of Sunday, August 6, 1962, at her 12305 Fifth Helena Drive home in Los Angeles, California.

Answer: barbiturate overdose
BiDAF
Bi-Directional Attention Flow for Machine Comprehension
BiDAF
Bi-Directional Attention Flow for Machine Comprehension

Output Layer
Modeling Layer
Attention Flow Layer
Phrase Embed Layer
Word Embed Layer
Character Embed Layer

Query2Context
Context2Query

Query
Context

Word Embedding
Character Embedding

GLOVE
Char-CNN
SQuAD

Leaderboard

SQuAD2.0 tests the ability of a system to not only answer reading comprehension questions, but also abstain when presented with a question that cannot be answered based on the provided paragraph.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Model</th>
<th>EM</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>ALBERT + DAAF + Verifier (ensemble) PINGAN Omni-SinTic</td>
<td>90.002</td>
<td>92.425</td>
</tr>
<tr>
<td>2</td>
<td>ALBERT (ensemble model) Google Research &amp; TTIC</td>
<td>89.731</td>
<td>92.215</td>
</tr>
<tr>
<td>3</td>
<td>XLNet + DAAF + Verifier (ensemble) PINGAN Omni-SinTic</td>
<td>88.592</td>
<td>90.859</td>
</tr>
<tr>
<td>3</td>
<td>albert+verifier (single model) Ping An Life Insurance Company AI Team</td>
<td>88.355</td>
<td>91.019</td>
</tr>
<tr>
<td>4</td>
<td>ALBERT (single model) Google Research &amp; TTIC</td>
<td>88.107</td>
<td>90.902</td>
</tr>
<tr>
<td>4</td>
<td>UPM (ensemble) Anonymous</td>
<td>88.231</td>
<td>90.713</td>
</tr>
</tbody>
</table>

In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under *gravity*. The main forms of precipitation include drizzle, rain, sleet, snow, *graupel* and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals within a cloud. Short, intense periods of rain in scattered locations are called “showers”.

What causes precipitation to fall? *gravity*

What is another main form of precipitation besides drizzle, rain, snow, sleet and hail? *graupel*

Where do water droplets collide with ice crystals to form precipitation? within a cloud
RC is so many things!

- **SQuAD, NewsQA**
  - Predicate/argument structure

- **TriviaQA, SearchQA**
  - Quiz style question with larger contexts

- **HotpotQA, WikiHop**
  - ComplexWebQuestions
  - Multi-hop reasoning

- **DROP**
  - Arithmetic reasoning

- **NarrativeQA**
  - Narrative understanding
There is a surge of RC datasets, how do they differ? Can we make use of all of them?
MultiQA
An Empirical Investigation of Generalization and Transfer in Reading Comprehension

Alon Talmor and Jonathan Berant
Desiderata

In this work we would like to test the **generalization** and **transfer** of models trained on these different datasets.

**Questions of interest:**

• How does performance change when generalizing to new datasets?

• Can fine-tuning on large RC datasets improve results on smaller datasets (even in the presence of )?
Contributions

• An **analysis** of generalization and transfer on two popular RC models

• A new **multi-task model**, MultiQA, that improves state-of-the-art on several datasets.

• **Infrastructure** for easily performing experiments on multiple RC datasets.
<table>
<thead>
<tr>
<th>Dataset</th>
<th>Size</th>
<th>Context</th>
<th>Question</th>
<th>Multi-hop</th>
</tr>
</thead>
<tbody>
<tr>
<td>SQuAD</td>
<td>108K</td>
<td>Wikipedia</td>
<td>crowd</td>
<td>No</td>
</tr>
<tr>
<td>NewsQA</td>
<td>120K</td>
<td>Newswire</td>
<td>crowd</td>
<td>No</td>
</tr>
<tr>
<td>SearchQA</td>
<td>140K</td>
<td>Snippets</td>
<td>trivia</td>
<td>No</td>
</tr>
<tr>
<td>TriviaQA</td>
<td>95K</td>
<td>Snippets</td>
<td>trivia</td>
<td>No</td>
</tr>
<tr>
<td>HotpotQA</td>
<td>113K</td>
<td>Wikipedia</td>
<td>crowd</td>
<td>Yes</td>
</tr>
<tr>
<td>ComplexQuestions</td>
<td>2K</td>
<td>Snippets</td>
<td>Web queries/KB</td>
<td>No</td>
</tr>
<tr>
<td>ComplexWebQuestions</td>
<td>35K</td>
<td>Snippets</td>
<td>crowd/KB</td>
<td>Yes</td>
</tr>
<tr>
<td>ComQA</td>
<td>11K</td>
<td>Snippets</td>
<td>WikiAnswers</td>
<td>No</td>
</tr>
<tr>
<td>WikiHop</td>
<td>51K</td>
<td>Wikipedia</td>
<td>KB</td>
<td>Yes</td>
</tr>
<tr>
<td>DROP</td>
<td>26K</td>
<td>Wikipedia</td>
<td>crowd</td>
<td>Yes</td>
</tr>
</tbody>
</table>
Models

We experiment with 2 RC models:

**DocumentQA** (Clark and Gardner, 2018)

- BiDAF (Seo et al., 2016) + Self-attention
- IR component

**BERT** (Devlin et al., 2019)

- BERT

```
[cls] question [sep] context [sep]
```
Generalization Experiments
1. Do models generalize well?

**Document-QA**

<table>
<thead>
<tr>
<th>Evaluation</th>
<th>Training</th>
<th>CQ</th>
<th>CWQ</th>
<th>ComQA</th>
<th>WikiHop</th>
<th>DROP</th>
</tr>
</thead>
<tbody>
<tr>
<td>SQuAD</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NewsQA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SearchQA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TriviaQA-Wiki</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HotpotQA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intra-Domain</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Average: 12.8EM

**BERTQA**

<table>
<thead>
<tr>
<th>Evaluation</th>
<th>Training</th>
<th>CQ</th>
<th>CWQ</th>
<th>ComQA</th>
<th>WikiHop</th>
<th>DROP</th>
</tr>
</thead>
<tbody>
<tr>
<td>SQuAD</td>
<td></td>
<td>23.6</td>
<td>12</td>
<td>20</td>
<td>4.6</td>
<td>5.5</td>
</tr>
<tr>
<td>NewsQA</td>
<td></td>
<td>24.1</td>
<td>12.4</td>
<td>18.9</td>
<td>7.1</td>
<td>4.4</td>
</tr>
<tr>
<td>SearchQA</td>
<td></td>
<td>30.3</td>
<td>16.5</td>
<td>26.6</td>
<td>12.6</td>
<td>3.1</td>
</tr>
<tr>
<td>TriviaQA-Wiki</td>
<td></td>
<td>30.3</td>
<td>16.5</td>
<td>26.6</td>
<td>12.6</td>
<td>3.1</td>
</tr>
<tr>
<td>HotpotQA</td>
<td></td>
<td>27.7</td>
<td>15.5</td>
<td>22.1</td>
<td>10.7</td>
<td>9.1</td>
</tr>
<tr>
<td>Intra-Domain</td>
<td></td>
<td>30.8</td>
<td>27.1</td>
<td>51.6</td>
<td>52.9</td>
<td>17.9</td>
</tr>
</tbody>
</table>

Average: 30.3EM

Average: 36.1EM

The finding confirms that models over-fit to the particular dataset and generalize poorly.

BERTQA improves generalization compared to DocumentQA.
2. Does generalization improve when training on multiple datasets?

<table>
<thead>
<tr>
<th>Training Set</th>
<th>Evaluation</th>
<th>CQ</th>
<th>CWQ</th>
<th>ComQA</th>
<th>WikiHop</th>
<th>DROP</th>
</tr>
</thead>
<tbody>
<tr>
<td>SQuAD</td>
<td></td>
<td>23.6</td>
<td>12</td>
<td>20</td>
<td>4.6</td>
<td>5.5</td>
</tr>
<tr>
<td>NewsQA</td>
<td></td>
<td>24.1</td>
<td>12.4</td>
<td>18.9</td>
<td>7.1</td>
<td>4.4</td>
</tr>
<tr>
<td>SearchQA</td>
<td></td>
<td>30.3</td>
<td>18.5</td>
<td>25.8</td>
<td>12.4</td>
<td>2.8</td>
</tr>
<tr>
<td>TriviaQA-Wiki</td>
<td></td>
<td>30.3</td>
<td>16.5</td>
<td>23.6</td>
<td>12.6</td>
<td>5.1</td>
</tr>
<tr>
<td>HotpotQA</td>
<td></td>
<td>27.7</td>
<td>15.5</td>
<td>22.1</td>
<td>10.2</td>
<td>9.1</td>
</tr>
<tr>
<td>Multi-75K</td>
<td></td>
<td>34</td>
<td>18.2</td>
<td>30.9</td>
<td>11.7</td>
<td>8.6</td>
</tr>
</tbody>
</table>

MULTI-75K performs almost as well as the best single training dataset, reducing the need for selecting a single training set.
3. Does generalization improve when training on more data?

More training data improves generalization.
A 2D-visualization of the similarity between different datasets

- Context type (shape)
  - Triangles: Web snippets
  - Squares: Newswire

- Task characteristics (color)
  - Red: Multi-hop reasoning datasets
  - Blue: Trivia datasets
  - Green: Factoid RC datasets

Datasets with similar context types are closer in terms of generalization.

Datasets with similar tasks are closer in terms of generalization.

Single Edge: BERT ➔ NewsQA ➔ WikiHop

Generalization
## MultiQA – state of the art

**MultiQA**: Training BERTLarge on SQuAD, NewsQA, HotpotQA, TriviaQA-Bing, SearchQA

<table>
<thead>
<tr>
<th>Dataset</th>
<th>BERT-large Dev.</th>
<th>MultiQA Dev.</th>
<th>MultiQA Test.</th>
<th>SOTA</th>
</tr>
</thead>
<tbody>
<tr>
<td>NewsQA</td>
<td>51.5</td>
<td>53.9</td>
<td>52.3</td>
<td>53.1</td>
</tr>
<tr>
<td>SearchQA</td>
<td>59.2</td>
<td>60.7</td>
<td>59</td>
<td>58.8</td>
</tr>
<tr>
<td>TriviaQA-unfiltered</td>
<td>56.8</td>
<td>58.4</td>
<td>52</td>
<td></td>
</tr>
<tr>
<td>CWQ</td>
<td>30.8</td>
<td>35.4</td>
<td>34.9</td>
<td>34.2</td>
</tr>
<tr>
<td>HotpotQA</td>
<td>27.9</td>
<td>30.6</td>
<td>30.7</td>
<td>37.1</td>
</tr>
<tr>
<td>COMQA</td>
<td>45.8</td>
<td>51.9</td>
<td>44.4</td>
<td>21.2</td>
</tr>
</tbody>
</table>
Cavities in Reading Comprehension

What is the blade hobol?

The blade hobol is a **balk banana**, walking throw the baldana bloop.

Answer: balk banana
Does multi-dataset training help adversarial attacks?

Adversarial Squad (Jia and Liang, 2017)

- **SOTA (Hu et al., 2018)**
  - EM: 53
  - F1: 58.5

- **BERT-large**
  - EM: 60.4
  - F1: 66.3

- **MultiQA**
  - EM: 66.7
  - F1: 73.1
The MultiQA Project – models use AllenNLP

python multiqa.py train -datasets TriviaQA,SQuAD1-1,HotpotQA

Python multiqa.py evaluate -model TriviaQA -datasets SQuAD1-1,HotpotQA

Python multiqa.py generalize -datasets TriviaQA,SQuAD1-1,HotpotQA

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Train</th>
<th>Dev</th>
<th>BERT-Base</th>
<th>MultiQA BERT-Base (model)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SQuAD-1.1</td>
<td>data</td>
<td>data</td>
<td>80.1 / 87.5 (model)</td>
<td>81.7 / 88.8</td>
</tr>
<tr>
<td>NewsQA</td>
<td>data</td>
<td>data</td>
<td>47.5 / 62.9 (model)</td>
<td>48.3 / 64.7</td>
</tr>
<tr>
<td>HotpotQA</td>
<td>data</td>
<td>data</td>
<td>50.1 / 63.2 (model)</td>
<td>-</td>
</tr>
<tr>
<td>TriviaQA-unfiltered</td>
<td>data</td>
<td>data</td>
<td>59.4 / 65.2 (model)</td>
<td>59.0 / 64.7</td>
</tr>
<tr>
<td>SearchQA</td>
<td>data</td>
<td>data</td>
<td>58.7 / 65.2 (model)</td>
<td>58.8 / 65.3</td>
</tr>
</tbody>
</table>
The Web as a Knowledge-Base for answering complex questions

Alon Talmor, Jonathan Berant
Complex Questions

List the U2 concerts that will be held in Europe on a weekend near a Marriott.

U2 > Tours > U2 eXPERIENCE + iNNOCEANCE Tour 2018

www.u2.com/tour

eXPERIENCE & iNNOCEANCE Tour heading to North America from May to July and Europe from August to November. #U2eLTour #U2SongsOfExperience 'Augment Your Reality.' Download the U2 AR eXPERIENCE app before reaching the venue and trigger ..... Email List - Terms + Conditions - Privacy - Help - Translate.

Missing: marriott

Evaluating Natural Language Understanding Services for Conversational Question Answering Systems.

https://www.researchgate.net/publication/319127909_Evaluating...
Existing approaches
Marilyn Monroe was found dead of a barbiturate overdose in the early morning hours of Sunday, August 6, 1962, at her 12305 Fifth Helena Drive home in Los Angeles, California.

Answer: barbiturate overdose

Existing approaches: Reading comprehension

What did Marilyn Monroe die from?

Advantages:
• Does not require schema
• Wide coverage

Disadvantages:
• Assume paragraph given
• Struggles with compositionality
Existing approaches: Semantic parsing

How many subsidiaries does Apple have?

\[
\text{Count}(\lambda x. \text{Type}(x, \text{Subsidiary}) \land (\text{Parent}(x, \text{Apple})))
\]

(Execute on Knowledge-base)

Answer: 3

Advantages:
• High compositionality

Disadvantages:
• Low coverage
• Requires schema

Complex Coverage

Semantic Parsing

Reading Comprehension

Present
Approach overview
Observations

Thus, the missing piece is broad and complex question understanding.

Assumption: Answering simple questions is possible with a search engine and a reading comprehension model.

Apply reading comprehension.

Thus, the missing piece is broad and complex question understanding.
Approach Example

“What is the population of the largest economy in Europe?”

Decompose

“largest economy in Europe”

Query
Approach Example

Europe's largest national economy is that of Germany, which ranks fourth globally in nominal GDP, and fifth in purchasing power parity (PPP) GDP; followed by the United Kingdom, ranking fifth globally in nominal GDP, followed by France, ranking sixth globally in nominal GDP, followed by Italy, which ranks seventh...

Economy of Europe - Wikipedia

Answer

Germany
Approach Example

Compose

“What is the population of Germany?”

Query

Answer 80.62M
Contributions

- A **framework** for answering broad and complex questions through question decomposition
- Pointer-generator network for question decomposition
- New **dataset**: ComplexWebQuestions
Framework
Task definition

What tau-nlp papers were presented in the recent NAACL?

\[
\text{Conj \{naacl18 tau-pub1, naacl18 tau-pub2 .. \}}
\]

\[
\text{Comp \{pub1, pub2 .. \}}
\]

Papers presented in VAR

\[
\text{SimpQA \{Naacl 2018\}}
\]

Recent NAACL

\[
\text{SimpQA \{tau-pub1, tau-pub2 .. \}}
\]

What tau-nlp papers

computation tree
- Leaves: strings - not necessarily in the input!
- Inner nodes: semantic functions
- Denotations: \( \llbracket t \rrbracket = f(\llbracket c_1(t) \rrbracket, \ldots, \llbracket c_k(t) \rrbracket) \)
Dataset Generation

1. Seed Question, WebQuestionsSP

   Where is the euro 2012 tournament?
   ?x ns:sports.event.host 2012_tournament.

   Add compositionality (auto)
   ?x ns:sports.event.host 2012_tournament.
   ?x ns:location.location.people_born_here ?y.

2. Add compositionality using Freebase

   Lviv born here Adam Zagajewski books published Without end

3. Generate machine questions

   Where is the euro 2012 and is the place where the publisher of the book Without End was born?

   Generate pseudo language (auto, Wang et al. 2015)

4. Paraphrase using Amazon Mechanical Turk

   What city is the birthplace of the author of Without End and hosted Euro 2012?

   Convert to natural language (human)

35,000 examples: questions, answers, SPARQL programs, 9,000,000 web snippets
Human Paraphrasing

<table>
<thead>
<tr>
<th>Compositionality</th>
<th>Example (Natural Language)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conjunction</td>
<td>“What films star Taylor Lautner <strong>and</strong> have costume designs by Nina Proctor?”</td>
</tr>
<tr>
<td>Superlative</td>
<td>“Which school that Sir Ernest Rutherford attended has the <strong>latest</strong> founding date?”</td>
</tr>
<tr>
<td>Comparative</td>
<td>“Which of the countries bordering Mexico have an army size of <strong>less than 1050</strong>?”</td>
</tr>
<tr>
<td>Composition</td>
<td>“Where is the end of the river that originates in Shannon Pot?”</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Type</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>WH-Modifier</td>
<td>18%</td>
</tr>
<tr>
<td>Major Reorder</td>
<td>22%</td>
</tr>
<tr>
<td>Minor Reorder</td>
<td>42%</td>
</tr>
<tr>
<td>Synonyms</td>
<td>54%</td>
</tr>
<tr>
<td>Skip Words</td>
<td>27%</td>
</tr>
<tr>
<td>Add Words</td>
<td>29%</td>
</tr>
</tbody>
</table>

Prevalent phenomena in paraphrasing

Machine question vs Natural language edit distance
Model
Training Data

Option 1

Question

Answer

Google

Hard to train with few samples

Option 2

Question

Hard for humans, expensive
Training Data – Noisy Supervision

Generate Machine Questions in pseudo language

What do we know?
1. Question Compositionality
2. Best split point of the machine generated question

Paraphrase using Amazon Mechanical Turk

Human

How to split the paraphrased questions?

Paraphrased using Amazon Mechanical Turk

Human

What do we know?
1. Question Compositionality
2. Best split point of the machine generated question

What films star Taylor Lautner and have costume designs by Nina Proctor?
Training data: alignment

what films have taylor lautner starred in and in which costumes where designed by Nina Proctor

Conj(What, films, star, Taylor, Lautner, and, “,”, have, costume, designs, by, Nina, Proctor, )
Output Representation

Conjunction

“What films star Taylor Lautner and have costume designs by Nina Proctor?”

Conj( What, films, star, Taylor, Lautner, and, “,”, costume, designs, by, Nina, Proctor, )

Composition

“Where is the end of the river that originates in Shannon Pot?”

Comp( the, river, that, originates, in, Shannon, Pot, “,”, Where, is, the, end, of, %VAR, )

Simple

“Who is the director of the Matrix”

Simp( Who, is, the, director, of, the, Matrix, )
Pointer Generator

Conj( , What)

$P_{\text{copy}} = f(\text{attention context vector}, \text{decoder input}, \text{decoder state})$

Attention Distribution

Vocabulary Distribution

See et al. 2017
Baselines

- Google Box – answer at the top of google results
- SimpQA – Simple reading comprehension (Talmor et al., 2017), question as is
- SplitQA – Our question decomposition model
- RCQA – DocumentQA, question as is
- SplitRCQA – Our Splitting applied on DocumentQA
- Human – 100 questions answering using search engine
Results

<table>
<thead>
<tr>
<th>Method</th>
<th>p@1 Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>GoogleBox</td>
<td>2.5</td>
</tr>
<tr>
<td>SimpQA</td>
<td>19.9</td>
</tr>
<tr>
<td>SplitQA</td>
<td>25.9</td>
</tr>
<tr>
<td>RCQA</td>
<td>30.6</td>
</tr>
<tr>
<td>SplitRCQA</td>
<td>35.6</td>
</tr>
<tr>
<td>Human</td>
<td>63.6</td>
</tr>
</tbody>
</table>
"Computer: Analyse the distribution of the pieces that we have, correcting for changes in star configurations over four billion years, then extrapolate for the missing piece" (Star Trek, The Chase)
Can QA models reason?
oLMpics

On what Language Model pre-training captures
Arithmetic

How much is 37 minus eleven?

27
26
Twenty Five

How much is 10 plus twenty, and divide the result by 2?

Ten
15
20
Conjunction

what is usually located at jar **and** is usually located at pocket?

- mustard
- change
- transit ticket

BERT filtered ConceptNet triplets
Composition

Pseudo Functional Relations

what is the adopting father name of the person that founded apple?

- Steve Jobs
- Paul Reinhold Jobs
- Steve Wozniak
Quantifiers (Coffee-Cats)

Probe BERT for (subject, predict \(\rightarrow\) object)

“Cats Drink [MASK] . \(\rightarrow\) Coffee!

(“Knowledge of animal appearance among sighted and blind adults” Kim et al., 2019)

A quantifier challenge set

• [Cats, Drink, Coffee, Never] \(\rightarrow\) “Cat’s [MASK] drink coffee” (never)
• [Men, Have, Beard, Sometimes]
'The size of a [cat] is usually much [MASK] than the size of a [mouse].'

<table>
<thead>
<tr>
<th></th>
<th>germ</th>
<th>mouse</th>
<th>cat</th>
<th>horse</th>
<th>house</th>
<th>train</th>
<th>mountain</th>
<th>planet</th>
</tr>
</thead>
<tbody>
<tr>
<td>germ</td>
<td>-</td>
<td>larger</td>
<td>smaller</td>
<td>smaller</td>
<td>smaller</td>
<td>smaller</td>
<td>smaller</td>
<td>smaller</td>
</tr>
<tr>
<td>mouse</td>
<td>smaller</td>
<td>-</td>
<td>smaller</td>
<td>smaller</td>
<td>smaller</td>
<td>smaller</td>
<td>smaller</td>
<td>smaller</td>
</tr>
<tr>
<td>cat</td>
<td>smaller</td>
<td>larger</td>
<td>-</td>
<td>smaller</td>
<td>smaller</td>
<td>smaller</td>
<td>smaller</td>
<td>smaller</td>
</tr>
<tr>
<td>horse</td>
<td>larger</td>
<td>larger</td>
<td>larger</td>
<td>-</td>
<td>smaller</td>
<td>smaller</td>
<td>smaller</td>
<td>smaller</td>
</tr>
<tr>
<td>house</td>
<td>larger</td>
<td>larger</td>
<td>larger</td>
<td>larger</td>
<td>-</td>
<td>larger</td>
<td>smaller</td>
<td>smaller</td>
</tr>
<tr>
<td>train</td>
<td>larger</td>
<td>larger</td>
<td>larger</td>
<td>larger</td>
<td>larger</td>
<td>larger</td>
<td>-</td>
<td>smaller</td>
</tr>
<tr>
<td>mountain</td>
<td>larger</td>
<td>larger</td>
<td>larger</td>
<td>larger</td>
<td>larger</td>
<td>larger</td>
<td>larger</td>
<td>-</td>
</tr>
<tr>
<td>planet</td>
<td>larger</td>
<td>larger</td>
<td>larger</td>
<td>larger</td>
<td>larger</td>
<td>larger</td>
<td>larger</td>
<td>-</td>
</tr>
</tbody>
</table>
A Control: "24 [MASK] 55", answers "blah" and "ya"
Synesthesia and Multi-Modality

V. S. Ramachandran
Visual QA

Q: What is the shape of the large item, mostly occluded by the metallic cube? A: sphere

Q: What color is the object that is a different size? A: purple

Q: What color ball is close to the small purple cylinder? A: gray

GQA

Is the bowl to the right of the green apple? What type of fruit in the image is round? What color is the fruit on the right side, red or green? Is there any milk in the bowl to the left of the apple?
Why is [person4] pointing at [person1]?

a) He is telling [person3] that [person1] ordered the pancakes.
b) He just told a joke.
c) He is feeling accusatory towards [person1].
d) He is giving [person1] directions.
In botany, a fruit is the seed-bearing structure in flowering plants (also known as angiosperms) formed from the ovary after flowering.

Fruits are the means by which angiosperms disseminate seeds. Edible fruits, in particular, have propagated with the movements of humans and animals in a symbiotic relationship as a means for seed dispersal and nutrition; in fact, humans and many animals have become dependent on fruits as a source of food. Accordingly, fruits account for a substantial fraction of the world's agricultural output, and some (such as the apple and the pomegranate) have acquired extensive cultural and symbolic meanings.
<table>
<thead>
<tr>
<th>Name</th>
<th>Image</th>
<th>Location</th>
<th>County</th>
<th>Initiated</th>
<th>Completed</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abbey House</td>
<td><img src="image" alt="Abbey House" /></td>
<td>Barrow-in-Furness</td>
<td>Cumbria</td>
<td>1910</td>
<td>1914</td>
<td>Guest-house built in the [Tudor Revival] style, of red ashlar and slate, for [Vickers Ltd.][3]</td>
</tr>
<tr>
<td>Abbotswod</td>
<td><img src="image" alt="Abbotswod" /></td>
<td>Lower Swell</td>
<td>Gloucestershire</td>
<td>1901</td>
<td>1901</td>
<td>Alterations to an existing property, and the design of gardens, for Mark Fenwick, a banker and mine owner.[4]</td>
</tr>
<tr>
<td>Castle Drogo</td>
<td><img src="image" alt="Castle Drogo" /></td>
<td>Drewsteignton</td>
<td>Devon</td>
<td>1911</td>
<td>1930</td>
<td>[English country house][5] borrowing styles of castle-building from the [medieval] and [Tudor] periods, along with more [minimalist][6] contemporary approaches.</td>
</tr>
<tr>
<td>Deanery Garden</td>
<td><img src="image" alt="Deanery Garden" /></td>
<td>Sonning</td>
<td>Berkshire</td>
<td>1899</td>
<td>1901</td>
<td>[Arts and Crafts style][7] house with garden laid out by Lutyens and planted by garden designer [Gertrude Jekyll][8], one of the several commissions from [Edward Hudson][9], founder of [Country Life magazine][10].</td>
</tr>
<tr>
<td>Folly Farm</td>
<td><img src="image" alt="Folly Farm" /></td>
<td>Sulhamstead</td>
<td>Berkshire</td>
<td>1906</td>
<td>1912</td>
<td>Built around a 17th-century farmhouse, to which Lutyens made extensions in a [neoclassical] style around 1906, and then in a [vernacular] style around 1912, for metals trader Zachary Merton.</td>
</tr>
<tr>
<td>Goddards</td>
<td><img src="image" alt="Goddards" /></td>
<td>Abinger</td>
<td>Surrey</td>
<td>1898</td>
<td>1900</td>
<td>In a [Tudor] style, with gardens by [Gertrude Jekyll], commissioned for charitable purposes by shipping magnate Frederick Mirrielees.</td>
</tr>
</tbody>
</table>
So what types of reasoning are we missing?
Pavlovian conditioning - Inductive Reasoning

Before Conditioning

During Conditioning

After Conditioning
Abductive Reasoning
Learning from few examples

IQ-TEST

1 + 4 = 5
2 + 5 = 12
3 + 6 = 21
8 + 11 = ?

CHOOSE ANSWER

1 2 3
4 5 6
Planning, Causality, Creativity
Reasoning and Question Answering

**Reasoning**
- Inducting Reasoning
- Deductive Reasoning
- Abductive Reasoning

**“Semantic functions”**
- Math
- Quantifiers
- Logic
- Superlatives
- Composition
- Invention

- Common sense
- Cause and effect
- Structured answer generation
- Planning
- Association and symbolic KB
Thanks! Questions?