Machine Reading Framework

The Attentive Reader

Mary went to England

X visited England
CNN/Daily Mail Datasets


Hermann et al. Teaching machines to read and comprehend. NIPS 2015
CNN/Daily Mail Datasets

Lexicalised ...

(CNN) New Zealand are on course for a first ever World Cup title after a thrilling semifinal victory over South Africa, secured off the penultimate ball of the match.

Chasing an adjusted target of 298 in just 43 overs after a rain interrupted the match at Eden Park, Grant Elliott hit a six right at the death to confirm victory and send the Auckland crowd into raptures. It is the first time they have ever reached a world cup final.

Question:
_____ reach cricket Word Cup final?

Answer:
New Zealand
CNN/Daily Mail Datasets

**Good**

we aimed to factor out world knowledge through entity anonymisation so models could not rely on correlations rather than understanding.

**Bad**

The generation process and entity anonymisation reduced the task to multiple choice and introduced additional noise.
Good

posing reading comprehension as a large scale conditional modelling task made it accessible to machine learning researchers, generating a great deal of subsequent research.

Bad

while this approach is reasonable for building applications, it is entirely the wrong way to develop and evaluate natural language understanding.
Narrative QA

Narrative QA: examples

**Question:** How is Oscar related to Dana?
**Answer:** He is her son

**Summary snippet:** ...Peter's former girlfriend Dana Barrett has had a son, Oscar...

**Story snippet:**

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

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

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

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

Good
A challenging evaluation that tests a range of language understanding, particularly temporal aspects of narrative, and also scalability as current models cannot represent and reason over full narratives

Bad
Performing well on this task is clearly well beyond current models, both representationally and computationally. As such it will be hard for researchers to hill climb on this evaluation.
Narrative QA

**Good**
The relatively small number of narratives for training models forces researchers to approach this task from a transfer learning perspective.

**Bad**
The relatively small number of narratives means that this dataset is not of immediate use for those wanting to build supervised models for applications.
**MS Marco**

Questions are mined from a search engine and matched with candidate answer passages using IR techniques.

---

### MS MARCO V2 Leaderboard

**Follows MSMarcoAI**

First released at MPS 2018, the MS/MARCO dataset is an ambitious, real-world Machine Reading Comprehension Dataset. Based on feedback from the community, we designed and released the v2 dataset and its related challenges ranked by difficulty (easiest to hardest). Can your model read, comprehend, and answer questions better than humans?

1. Given a query and 10 passages provide the best answer available based on context.
2. Given a query and 10 passages provide the best answer available in natural language that could be used by a smart device (digital assistant) (intermediate).
3. T90 (Expert)

Models are ranked by ROUGE-L Score.

#### Novice Task

<table>
<thead>
<tr>
<th>Rank</th>
<th>Model</th>
<th>Submission Date</th>
<th>Rouge-L</th>
<th>Bleu-1</th>
<th>P1</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Human Performance</td>
<td>April 29th, 2016</td>
<td>50.67</td>
<td>48.50</td>
<td>94.72</td>
</tr>
<tr>
<td>2</td>
<td>BERT Base/LR</td>
<td>June 19th, 2018</td>
<td>46.72</td>
<td>50.49</td>
<td>72.98</td>
</tr>
<tr>
<td>3</td>
<td>BERT JT Zhao</td>
<td>June 30th, 2018</td>
<td>42.28</td>
<td>48.14</td>
<td>72.98</td>
</tr>
<tr>
<td>4</td>
<td>DNET++ QA Geeks</td>
<td>June 1st, 2018</td>
<td>41.91</td>
<td>48.88</td>
<td>72.83</td>
</tr>
<tr>
<td>5</td>
<td>BERT tao@Taiwan</td>
<td>June 14th, 2018</td>
<td>39.82</td>
<td>42.27</td>
<td>72.98</td>
</tr>
<tr>
<td>6</td>
<td>DNET QA Geeks</td>
<td>May 29th, 2018</td>
<td>32.38</td>
<td>29.12</td>
<td>74.38</td>
</tr>
<tr>
<td>7</td>
<td>BIBAF+youth/wen Yi</td>
<td>May 29th, 2018</td>
<td>27.59</td>
<td>29.64</td>
<td>75.98</td>
</tr>
</tbody>
</table>
**MS Marco**

**Good**

The reliance on real queries creates a much more useful resource for those interested in applications.

**Bad**

People rarely ask interesting questions of search engines, and the use of IR techniques to collect candidate passages limits the usefulness of this dataset for evaluating language understanding.
<table>
<thead>
<tr>
<th>Good</th>
<th>Bad</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unrestricted answers allow a greater range of questions.</td>
<td>How to evaluate freeform answers is an unsolved problem. Bleu is not the answer!</td>
</tr>
</tbody>
</table>
SQuAD

What is SQuAD?

Stanford Question Answering Dataset (SQuAD) is a reading comprehension dataset, consisting of questions posed by crowdworkers on a set of Wikipedia articles, where the answer to every question is a segment of text, or span, from the corresponding reading passage, or the question might be unanswerable.

How SQuAD2.0 combines the 100,000 questions in SQuAD1.1 with over 50,000 new, unanswerable questions written adversarially by crowdworkers to look similar to answerable ones. To do well on SQuAD2.0, systems must not only answer questions when possible, but also abstain when presented with a question that cannot be answered based on the provided paragraph. How will your system compare to humans on this task?

Getting Started

We've built a few resources to help you get started with the dataset. Download a copy of the dataset (distributed under the CC BY-SA 4.0 license):

- SQuAD2.0 paper (Rajpurkar et al., 2018)
- SQuAD1.0 paper (Rajpurkar et al., 2016)

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. How will your system compare to humans on this task?

<table>
<thead>
<tr>
<th>Rank</th>
<th>Model</th>
<th>EM</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>V5*3-NET (single model)</td>
<td>68.438</td>
<td>71.282</td>
</tr>
<tr>
<td></td>
<td>Kangwon National University in South Korea</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>KACTEL-MRC3FDN (single model)</td>
<td>68.224</td>
<td>70.871</td>
</tr>
<tr>
<td></td>
<td>Kangwon National University, Natural Language Processing Lab.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>KakaNet2 (single model)</td>
<td>65.708</td>
<td>69.369</td>
</tr>
<tr>
<td></td>
<td>Kaka NLP Team</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>abcNet (single model)</td>
<td>65.256</td>
<td>69.398</td>
</tr>
<tr>
<td></td>
<td>Fudan University &amp; Laihuan AI Lab</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>BiDAF + Self Attention + ELMo (single model)</td>
<td>63.382</td>
<td>66.262</td>
</tr>
<tr>
<td></td>
<td>Allen Institute for Artificial Intelligence (modified by Stanford)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>BiDAF + Self Attention (single model)</td>
<td>59.302</td>
<td>62.335</td>
</tr>
<tr>
<td></td>
<td>Allen Institute for Artificial Intelligence (modified by Stanford)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
In the 1960s, a series of discoveries, the most important of which was seafloor spreading, showed that the Earth's lithosphere, which includes the crust and rigid uppermost portion of the upper mantle, is separated into a number of tectonic plates that move across the plastically deforming, solid, upper mantle, which is called the asthenosphere. There is an intimate coupling between the movement of the plates on the surface and the convection of...

**Question:**
Which parts of the Earth are included in the lithosphere?
### SQuAD

<table>
<thead>
<tr>
<th>Good</th>
<th>Bad</th>
</tr>
</thead>
<tbody>
<tr>
<td>Very scalable annotation process that can cheaply generate large numbers of questions per article.</td>
<td>Annotating questions directly from the context passages strongly skews the data distribution. The task then becomes reverse engineering the annotators, rather than language understanding.</td>
</tr>
</tbody>
</table>
Good

The online leaderboard allows easy benchmarking of systems and motivates competition.

Bad

Answers as spans reduces the task to multiple choice, and doesn’t allow questions with answers latent in the text.
## SQuAD

<table>
<thead>
<tr>
<th>Good</th>
<th>Bad</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human upperbound sets reasonable goal.</td>
<td>Allows mischaracterization of what it means to “read”.</td>
</tr>
</tbody>
</table>
Quiz Bowl
Quiz Bowl

**Good**
Free data from experts

**Bad**
Sometimes can be trivially solved with pattern matching
Quiz Bowl

**Good**

Based on already known knowledge

**Bad**

Not tied to readable data
Quiz Bowl

<table>
<thead>
<tr>
<th>Good</th>
<th>Bad</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human comparison makes sense</td>
<td>More cumbersome computer evaluation</td>
</tr>
</tbody>
</table>
Possible Projects

- Improve selection of answer spans
- Improve IR search for context
- Visualizing reader spans
- Domain adaptation (Wikipedia/Questions/Books)