Machine Reading Comprehension

Furu Wei, Lead Researcher
Natural Language Computing Group, MSRA
July 2017

Joint work with Nan Yang, Lei Cui, Ming Zhou,
Wenhui Wang, Chaunqi Tan, Qingyu Zhou
Machine Reading Comprehension

- Summarize the key points
- Answer questions
- Ask questions
- Reply and comment
Our paper on machine reading comprehension for SQuAD (R-NET) has been accepted by ACL 2017 as a long paper.
Super Bowl 50 is set to take place on Sunday, February 7, at 6:30 PM (ET) on CBS. The game will take place at Levi’s Stadium (home of the San Francisco 49ers) in Santa Clara, Calif., in the Bay Area.
My phone number is +4911002233. I live in Munich, Germany. You can pay with your credit card, we accept: Visa, Mastercard, American Express, Maestro, Visa Debit. The delivery is twice per week on Tuesday and Saturday. If the purchase or a product is not good or you are unsatisfied please return the product with the receipt within 30 days to the driver or call us on +4911002233.
Microsoft Surface Pro 4 (128 GB, 4 GB RAM, Intel Core i5)

by Microsoft

Price: $775.00 & FREE Shipping

In Stock.

Get it as soon as Wednesday, March 1 when you choose Two-Day Shipping at checkout. Ships from and sold by Pcbase USA.

Style: Device Only
Size: Intel Core i5, 4GB RAM, 4GB

Used & new (78) from $629.99 & FREE shipping.

Customer Questions & Answers

Question: I think if you’re a regular 128gb card in for $60. I. g getting a newer model
Answer: I love it, but you might not.
By Joshua Ol on October 28, 2015

Question: I am using SP4 with point, for now, unless you this feature in the future.
Answer: Yes, the S.Book is almost twice the size of the SP4. It is particularly heavy and thick.
By WX on October 30, 2015

Customer Reviews

Top Customer Reviews

Rating: 5 Stars

Customer Reviews

The screen resolution is 2736 x 1824 for Microsoft Surface Pro 4 (128 GB, 4 GB RAM, Intel Core i5).

Does it have a headphone port?

It has a mini display port, but through the use of an adapter this will allow an HDMI connection.
Machine Reading:
Extraction

Machine Reading:
Synthesis & Generation

Machine Reading:
Reasoning & Inference
The Rhine (Romansh: Rein, German: Rhein, French: le Rhin, Dutch: Rijn) is a European river that begins in the Swiss canton of Graubünden in the southeastern Swiss Alps, forms part of the Swiss-Austrian, Swiss-Liechtenstein border, Swiss-German and then the Franco-German border, then flows through the Rhineland and eventually empties into the North Sea in the Netherlands. The biggest city on the river Rhine is Cologne, Germany with a population of more than 1,050,000 people. It is the second-longest river in Central and Western Europe (after the Danube), at about 1,230 km (760 mi), with an average discharge of about 2,900 m3/s (100,000 cu ft/s).

Question: What river is larger than the Rhine?

Answer: Danube

* Example from the SQuAD dataset
P1: Impact on Japan's Economy. The Triple Disaster devastated Japan's economy in four ways. First, it destroyed 138,000 buildings and cost $360 billion in economic damage. This is more than the $250 billion cost estimate for Hurricane Katrina.

Passage

P2: Japan’s 2011 Earthquake and Tsunami: Economic Effects and Implications for the U.S. Congressional Research Service 3 construction supplies. If imports of certain products from Japan become scarce, China, South Korea, or other nations may gain at Japan’s expense.

Question

What is the economic impact of the Japan earthquake

Answer

Japan have $360 billion economic damage because of earthquake.

* Example from the MS MARCO dataset
During the construction of the Quebec Bridge in 1907, the bridge’s designer, Theodore Cooper, received word that the suspended span being built out from the bridge’s cantilever was deflecting downward by a fraction of an inch (2.54 centimeters). Before he could telegraph to freeze the project, the whole cantilever arm broke off and plunged, along with seven dozen workers, into the St. Lawrence River. It was the worst bridge construction disaster in history. As a direct result of the inquiry that followed, the engineering “rules of thumb” by which thousands of bridges had been built around the world went down with the Quebec Bridge. Twentieth-century bridge engineers would thereafter depend on far more rigorous applications of mathematical analysis.

**Reasoning & Inference**

Which one of the following statements can be properly inferred from the passage?

(A): Bridges built before about 1907 were built without thorough mathematical analysis and, therefore, were unsafe for the public to use.

(B): Cooper’s absence from the Quebec Bridge construction site resulted in the breaking off of the cantilever.

(C): Nineteenth-century bridge engineers relied on their rules of thumb because analytical methods were inadequate to solve their design problems.

(D): Only a more rigorous application of mathematical analysis to the design of the Quebec Bridge could have prevented its collapse.

(E): Prior to 1907 the mathematical analysis incorporated in engineering rules of thumb was insufficient to completely assure the safety of bridges under construction.

* Example from the LSAT dataset (Logical Reasoning & Reading Comprehension)
Machine Reading: Extraction
Dataset | Question source | Formulation source | Size
---|---|---|---
SQuAD | crowdsourced | RC, spans in passage | 100K

MCTest (Richardson et al., 2013) | crowdsourced | RC, multiple choice | 2640
Algebra (Kushman et al., 2014) | standardized tests | computation | 514
Science (Clark and Ezioni, 2016) | standardized tests | reasoning, multiple choice | 855

WikiQA (Yang et al., 2015) | query logs | IR, sentence selection | 3847
TRECVQA (Voorhees and Tice, 2000) | query logs + human editor | IR, free form | 1479
CNN/Daily Mail (Hermann et al., 2015) | summary # + clone | RC, fill in 1.4M single entity | 1.4M
CBT (Hill et al., 2015) | clone | RC, fill in 688K single word | 688K

Table 1: A survey of several reading comprehension and question answering datasets. SQuAD is much larger than all datasets except the semi-synthetic clone-style datasets, and it is similar to TRECVQA in the open-endedness of the answers.

**NEWSQA: A MACHINE COMPREHENSION DATASET**

Adam Trischler, Tong Wang, Xingliang Yuan, Justin Harris, Alexandros Stordoni, Philip Bachman, Kalen Soneman

(abadam.trischler, tong.wang, eric.yuan, justin.harris, alexandroa.stordoni, phil.bachman, k.suman@maluuba.com

Maluuba Research
Montreal, Quebec, Canada

**ABSTRACT**

We present NewsQA, a challenging machine comprehension dataset of over 100,000 human-generated question-answer pairs. Crowdsources supply questions and answers based on a set of over 10,000 news articles from CNN, with answers consisting of spans of text from the corresponding articles. We collect this dataset through a four-stage process designed to solve exploratory questions that require reasoning. A thorough analysis confirms that NewsQA demands abilities beyond simple word matching and recognizing textual entailment. We measure human performance on the dataset and compare it to several strong neural models. The performance gap between humans and machines (0.199 in F1) indicates that significant progress can be made on NewsQA through future research. The dataset is freely available at https://datasets.maluuba.com/NewsQA.

**RACE: Large-scale ReAding Comprehension Dataset From Examinations**

Guokun Lai, Qing He, and Ruijin Luo, and Yiming Yang and Eduard Hovy

guokun, qing, harvard, yiming, hovy@cs.cmu.edu

Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213

**TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension**

Mandar Joishi, Emanuel Haid, Daniel S. Weld, and Luke Zettlemoyer

Mandar Joishi

uemand, weil, lzw@cs.washington.edu

1 Paul G. Allen School of Computer Science & Engineering, Unv. of Washington, Seattle, WA
2 Allen Institute for Artificial Intelligence, Seattle, WA
Before 2015:
- MCTest (Richardson et al, 2013): 2600 questions
- ProcessBank (Berant et al, 2014): 500 questions

After 2015:
- CNN/Daily Mail
- Children Book Test
- WikiReading
- LAMBADA
- SQuAD
- Who did What
- Maluuba
- NewsQA
- MS MARCO

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”.

**SQuAD: 100,000+ Questions for Machine Comprehension of Text**

Best Resource Paper in EMNLP 2016

Pranav Rajpurkar and Jian Zhang and Konstantin Lopyrev and Percy Liang
{pranavsr, zjian, klopyrev, pliang}@cs.stanford.edu
Computer Science Department
Stanford University

<table>
<thead>
<tr>
<th>Dataset</th>
<th># of questions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training</td>
<td>87,599</td>
</tr>
<tr>
<td>Dev</td>
<td>10,570</td>
</tr>
<tr>
<td>Test (not available for ~10K participants)</td>
<td></td>
</tr>
</tbody>
</table>

**ImageNet-style competition for machine reading comprehension**
Microsoft MAchine Reading COmprehension Dataset

A Reading Comprehension Dataset for the Artificial Intelligence research community

Join us in building intelligent agents which amplify the effectiveness of every person and business.

Read the MS MARCO Paper  View the MS MARCO Leaderboard  Download the MS MARCO Dataset
<table>
<thead>
<tr>
<th>Rank</th>
<th>Model</th>
<th>EM</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>nert (ensemble)</td>
<td>77.688</td>
<td>84.566</td>
</tr>
<tr>
<td>5</td>
<td>nert (single model)</td>
<td>76.614</td>
<td>82.438</td>
</tr>
<tr>
<td>6</td>
<td>Memetic Reader (ensemble)</td>
<td>75.738</td>
<td>82.658</td>
</tr>
<tr>
<td>7</td>
<td>CTH-BDAF (ensemble)</td>
<td>73.064</td>
<td>81.551</td>
</tr>
<tr>
<td>8</td>
<td>PPT (ensemble)</td>
<td>73.044</td>
<td>81.517</td>
</tr>
<tr>
<td>9</td>
<td>Multi-Perspective Matching (ensemble)</td>
<td>73.763</td>
<td>81.257</td>
</tr>
<tr>
<td>10</td>
<td>Dynamic Co-action Networks (ensemble)</td>
<td>71.625</td>
<td>80.383</td>
</tr>
<tr>
<td>11</td>
<td>Interactive AoA Reader (single model)</td>
<td>71.153</td>
<td>79.937</td>
</tr>
<tr>
<td>12</td>
<td>Interactive AoA Reader (ensemble)</td>
<td>70.601</td>
<td>79.821</td>
</tr>
<tr>
<td>13</td>
<td>Jiekit (single model)</td>
<td>70.657</td>
<td>79.549</td>
</tr>
<tr>
<td>14</td>
<td>Jiekit (ensemble)</td>
<td>70.657</td>
<td>79.549</td>
</tr>
<tr>
<td>15</td>
<td>FastQA (single model)</td>
<td>70.555</td>
<td>79.364</td>
</tr>
<tr>
<td>16</td>
<td>FastQA (ensemble)</td>
<td>70.733</td>
<td>79.303</td>
</tr>
<tr>
<td>17</td>
<td>Memetic Reader (ensemble)</td>
<td>69.863</td>
<td>79.207</td>
</tr>
<tr>
<td>18</td>
<td>PDMN (single model)</td>
<td>68.478</td>
<td>77.971</td>
</tr>
<tr>
<td>19</td>
<td>DeepQA (single model)</td>
<td>68.321</td>
<td>77.783</td>
</tr>
<tr>
<td>20</td>
<td>DeepQA (ensemble)</td>
<td>68.321</td>
<td>77.659</td>
</tr>
</tbody>
</table>

Human Performance: 82.304 91.321
# MS MARCO Leaderboard

The MS MARCO dataset was released at NIPS 2016. We appreciate the more than 1,000 downloads from the Research community in less than one month and the progress so far! Its exciting to see the research community coming together to solve this difficult problem. Here are the BLEU-1 and ROUGE-L scores for the best models we evaluated to date on the MS MARCO v1.1 test set. We will regularly update our leaderboard as we get submissions. Follow us on Twitter for updates.

![Twitter handle](https://twitter.com/MSMarcoAI)

<table>
<thead>
<tr>
<th>Rank</th>
<th>Model</th>
<th>Rouge-L</th>
<th>Bleu-1</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>R-Net</td>
<td>42.89</td>
<td>42.22</td>
</tr>
<tr>
<td></td>
<td>Microsoft AI and Research [Wei et al. '16]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>ReasoNet</td>
<td>38.81</td>
<td>39.86</td>
</tr>
<tr>
<td></td>
<td>Microsoft AI and Research [Shen et al. '16]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Prediction</td>
<td>37.33</td>
<td>40.72</td>
</tr>
<tr>
<td></td>
<td>Singapore Management University [Wang et al. '18]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>FastQA Ext</td>
<td>33.67</td>
<td>33.93</td>
</tr>
<tr>
<td></td>
<td>DFKI German Research Center for AI [Weissenborn et al. '17]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>FastQA</td>
<td>32.09</td>
<td>33.99</td>
</tr>
<tr>
<td></td>
<td>DFKI German Research Center for AI [Weissenborn et al. '17]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>ReasoNet Baseline</td>
<td>19.20</td>
<td>14.83</td>
</tr>
<tr>
<td></td>
<td>Trained on SQuAD, Microsoft AI &amp; Research [Shen et al. '16]</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

## Human Performance

Can your model read, comprehend, and answer questions better than humans? The below is current human performance on the MS MARCO task (which we will improve in future versions). This was ascertained by having two judges answer the same question and measuring our metrics over their responses.

<table>
<thead>
<tr>
<th>Model</th>
<th>Rouge-L</th>
<th>Bleu-1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human Performance</td>
<td>47</td>
<td>46</td>
</tr>
</tbody>
</table>
Method

**Before 2015:**
- Lexical matching
- Logistic regression

  - Sachan et al., ACL 2015
  - Wang et al., ACL 2015
  - Chen et al., ACL 2016

**After 2015:**
- Attentive Reader
- Memory Networks
- Gated-attention Reader
- ReasoNet
- Match-LSTM
- Attention Sum Reader
- Attention-over-Attention Reader
- Iterative Attentive Reader
- Dynamic coattention networks
- Bi-directional Attention Flow Network
- Multi-Perspective Context Matching

Deep Learning for Machine Reading

<table>
<thead>
<tr>
<th>CNN/Daily Mail</th>
<th>CBTTest</th>
<th>SQuAD</th>
<th>MARCO</th>
</tr>
</thead>
<tbody>
<tr>
<td>Attention-over-Attention (Cui et al. ACL 2017)</td>
<td>EpiReader (Trischler et al. EMNLP 2016)</td>
<td>BiDAF (Seo et al. ICLR 2017)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Dynamic Co-attention (Xiong et al. ICLR 2017)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>ReasoNet (Shen et al. KDD 2017)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>R-Net (Wang et al. ACL 2017)</td>
<td></td>
</tr>
</tbody>
</table>
Ferdinand III had started out as a contested king of Castile. By the time of his death in 1252, Ferdinand III had delivered to his son and heir, Alfonso X, a massively expanded kingdom. The boundaries of the new Castilian state established by Ferdinand III would remain nearly unchanged until the late 15th century.

When did Castilian border change after Ferdinand III’s death?
Span Prediction / Ranking

Challenges

• Span detection (answer candidate generation) \((n*(n-1)/2)\)
• Span/Ngram-level feature engineering or learning
Boundary Prediction / Ranking

Challenges

• Matching & passage-level evidence aggregation
• Prior knowledge & reasoning
R-NET
When did Castilian border change after Ferdinand III's death?

Ferdinand III had started out as a contested king of Castile. By the time of his death in 1252, Ferdinand III had delivered to his son and heir, Alfonso X, a massively expanded kingdom. The boundaries of the new Castilian state established by Ferdinand III would remain nearly unchanged until the late 15th century.

Answer: late 15th century
When did Castilian border change after Ferdinand III's death?

Ferdinand III had started out as a contested king of Castile. By the time of his death in 1252, Ferdinand III had delivered to his son and heir, Alfonso X, a massively expanded kingdom. The boundaries of the new Castilian state established by Ferdinand III would remain nearly unchanged until the late 15th century.

Answer: late 15th century
Probability of each word as start of the answer

Matching Networks

Self-Matching Networks

Passage

Question

Representation Networks
R-Net Technical Report

Skip

Answer Pointer Networks

Probability of each word as end of the answer

Self-Matching Networks

Matching Networks

Representation Networks
Representation Networks: 3-layer bidirectional RNN over question and passage

Pretrained word embeddings (Glove) and are fixed in training (300d)

Updated word embeddings from char embedding and composition with RNN (Bi-GRU) (200d)

For each word

75*6 = 450d
When did Castilian death...
Self-Matching Networks: 2-layer bidirectional RNN over passage
Police in Sanvape smoking device

Answer Pointer Networks
Police in San vape smoking device

Answer Pointer Networks
## Ablation Results

<table>
<thead>
<tr>
<th></th>
<th>Dev Set</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>EM</td>
</tr>
<tr>
<td><strong>R-NET (06/02/2017)</strong></td>
<td>71.1</td>
</tr>
<tr>
<td>- Char embedding</td>
<td>70.3</td>
</tr>
<tr>
<td>- Gating</td>
<td>67.9</td>
</tr>
<tr>
<td>- Self-Matching</td>
<td>66.9</td>
</tr>
<tr>
<td>- Gating, Self-Matching</td>
<td>65.2</td>
</tr>
</tbody>
</table>

**Major improvements**
- Self-Matching ~3 points
- Gated attention ~3 points
- New network structures
- Deep, training strategy, system tricks etc.
Positive Cases

- **Type**
- **Lexical approximation & variance**
- **Boundary (readable)**
- **Evidence aggregation**
<table>
<thead>
<tr>
<th>Reasoning</th>
<th>Description</th>
<th>Example</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lexical variation (synonymy)</td>
<td>Major correspondences between the question and the answer sentence are synonyms.</td>
<td>Q: What is the Rankine cycle sometimes <strong>called</strong>? Sentence: The Rankine cycle is sometimes referred to as a practical Carnot cycle.</td>
<td>33.3%</td>
</tr>
<tr>
<td>Lexical variation (world knowledge)</td>
<td>Major correspondences between the question and the answer sentence require world knowledge to resolve.</td>
<td>Q: Which governing bodies have veto power? Sen.: <strong>The European Parliament and the Council of the European Union</strong> have powers of amendment and veto during the legislative process.</td>
<td>9.1%</td>
</tr>
<tr>
<td>Syntactic variation</td>
<td>After the question is paraphrased into declarative form, its syntactic dependency structure does not match that of the answer sentence even after local modifications.</td>
<td>Q: What Shakespeare scholar is currently on the faculty? Sen.: <strong>Current faculty include</strong> the anthropologist Marshall Sahlins, ..., Shakespeare scholar David Bevington.</td>
<td>64.1%</td>
</tr>
<tr>
<td>Multiple sentence reasoning</td>
<td>There is anaphora, or higher-level fusion of multiple sentences is required.</td>
<td>Q: What collection does <strong>the V&amp;A Theatre &amp; Performance galleries</strong> hold? Sen.: The V&amp;A Theatre &amp; Performance galleries opened in March 2009. ... They hold the UK’s biggest national collection of material about live performance.</td>
<td>13.6%</td>
</tr>
<tr>
<td>Ambiguous</td>
<td>We don’t agree with the crowdworkers’ answer, or the question does not have a unique answer.</td>
<td>Q: What is the main goal of criminal punishment? Sen.: <strong>Achieving crime control via incapacitation and deterrence</strong> is a major goal of criminal punishment.</td>
<td>6.1%</td>
</tr>
</tbody>
</table>

Table 3: We manually labeled 192 examples into one or more of the above categories. Words relevant to the corresponding reasoning type are bolded, and the crowdsourced answer is underlined.

*Source: Rajpurkar et al. EMNLP 2016*
Future Work

Synthesis & Generation
S-NET

Reasoning & Inference
L-NET
THANKS