Goal: build an end-to-end question answering system that can use full Wikipedia to answer any factoid question.
Large-scale QA + Machine comprehension of Text “Machine Reading at Scale” (MRS)

Our system DrQA:
Q What is question answering? A a computer science discipline within the fields of information retrieval and natural language processing
Q Who was the winning pitcher in the 1956 World Series? A Don Larsen
Q What is the answer to life, the universe, and everything? A 42
Try it out yourself! https://github.com/facebookresearch/DrQA

Document Retriever + Document Reader
• Document retriever: finding relevant articles from 5 million Wikipedia articles
• Document reader (reading comprehension system): identifying the answer spans from those articles
  Q: How many of Warsaw’s inhabitants spoke Polish in 1933?

Data: SQuAD + Distantly Supervised Data
(Q, A) —— (P, Q, A) if P is retrieved and A can be found in P
Q: What part of the atom did Chadwick discover?
A: neutron

Atom
From Wikipedia, the free encyclopedia
The atomic mass of these isotopes varied by integer amounts, called the whole number rule. The explanation for these different isotopes awaited the discovery of the neutron, an uncharged particle with a mass similar to the proton, by the physicist James Chadwick in 1932. Isotopes were then explained as elements with the same number of protons, but different numbers of neutrons within the nucleus.

Results

Finding Relevant Articles

<table>
<thead>
<tr>
<th></th>
<th>Wiki Search</th>
<th>unigram</th>
<th>+bigram</th>
</tr>
</thead>
<tbody>
<tr>
<td>SQuAD</td>
<td>62.7</td>
<td>76.1</td>
<td>77.8</td>
</tr>
<tr>
<td>TREC</td>
<td>81.0</td>
<td>85.2</td>
<td>86.0</td>
</tr>
<tr>
<td>WebQuestions</td>
<td>73.7</td>
<td>75.5</td>
<td>74.4</td>
</tr>
<tr>
<td>WikiMovies</td>
<td>61.7</td>
<td>54.4</td>
<td>70.3</td>
</tr>
</tbody>
</table>

70-86% of questions we have that the answer segment appears in the top 5 articles

Performance on SQuAD (single model, Feb 2017)

<table>
<thead>
<tr>
<th></th>
<th>EM</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic regression</td>
<td>40.4</td>
<td>51.0</td>
</tr>
<tr>
<td>Fine-Grained Gating (Carnegie)</td>
<td>62.5</td>
<td>73.3</td>
</tr>
<tr>
<td>Match-LSTM (Singapore)</td>
<td>64.7</td>
<td>73.7</td>
</tr>
<tr>
<td>DCN (Salesforce)</td>
<td>66.2</td>
<td>75.9</td>
</tr>
<tr>
<td>BiDAF (UW &amp; Allen Institute)</td>
<td>68.0</td>
<td>77.3</td>
</tr>
<tr>
<td>Ours</td>
<td>70.7</td>
<td>79.4</td>
</tr>
<tr>
<td>r-net (MSR Asia)</td>
<td>71.3</td>
<td>79.7</td>
</tr>
<tr>
<td>State-of-the-art (July 2017)</td>
<td>75.7</td>
<td>83.5</td>
</tr>
<tr>
<td>Human performance</td>
<td>82.3</td>
<td>91.2</td>
</tr>
</tbody>
</table>

Exact match features are important!

Full Results

- Pre-trained SQuAD model
- SQuAD + fine-tuning on DS data
- Multi-task learning

Exact match (top-1 prediction)