Engineered AI Still Matters for Question Answering

J. William Murdock, Lin Pan, Chung-Wei Hang, Mary Swift, Zhiguo Wang, Chris Nolan, Prathyusha Peddi, Nisarga Markandaiah, Eunyoung Ha, Kazi Hasan, Yang Yu, Wei Zhang

IBM Watson
Stanford Question Answering Dataset (SQuAD)

- Reading comprehension data
  - Passages
  - Questions about the passages
  - Answers in the passages
- Wikipedia passages
- Crowd workers saw the passages, wrote questions, and selected answers
- Very popular for statistical reading comprehension research

... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals **within a cloud**. ...

Where do water droplets collide with ice crystals to form precipitation?

Why should anyone care about SQuAD?

– Answering reading comprehension questions is an interesting AI challenge
– Not a particularly useful capability by itself
  • Users do not want to provide a passage + a question and ask for an answer from that passage
– Important subtask of factoid question answering
– Combine a system built for SQuAD with a passage search capability
A system that excels at SQuAD will also excel at factoid question answering.
Factoid-1527

– Factoid question answering data
  • Answers are typically entities or numbers
– Fairly small (1,527 questions total)
– Questions written without being tied to a specific piece of text
– We use Wikipedia and Wiktionary as sources
– IBM confidential

In what year did William Bligh arrive in Tahiti?

1788

(not a real example)
DDQA: Multi-Strategy Factoid Question Answering

- DDQA = Discovery DeepQA
- Simpler version of IBM Watson 1.0, designed for cloud
- Engineered features, knowledge bases, and rules
- Expensive to create

DDQA for SQuAD

Question Analysis

SQuAD Passage

Answer Generation

Answer Scoring

KBs

Statistical Models

Answer Merging & Ranking

Rules

Statistical Models

Question Analysis

Text Search index
Single-Strategy Statistical Factoid Question Answering

- No manually engineered features, knowledge bases, or rules
- One statistical model does everything
Single-Strategy Statistical Question Answering for SQuAD
Integrated Question Answering

Question Analysis

Passage Retrieval

Answer Generation

Answer Scoring

Answer Merging & Ranking

Questions

KBs

Rules

Statistical Models

Statistical Models

Text Search index

Statistical Model

Reading Comprehension

Answers
YAGO TyCor: A DDQA example component

- DBpedia*: knowledge-base of structured information from Wikipedia
- YAGO**: semi-automatically constructed taxonomy from WordNet, Wikipedia, etc.
- YAGO Disjointness Axioms***: axioms built for IBM Watson 1.0 specifying disjoint types, e.g., person, location

What Kansas city has the highest population?

YAGO TyCor is very expensive!

– Dbpedia and YAGO are free for us because they already exist
– WordNet and Wikipedia already existed
– WordNet and Wikipedia both took huge investments of effort
– If we wanted to do something similar for copper mining it would cost a lot
– We created the disjointness axioms
– And the logic to reason about entities and types
– And this is just one of dozens of components!
Bidirectional Attention Flow

- Off-the-shelf Deep Neural Net
- Some engineering on structure
- No manually engineered features

IBM would like statistics alone to be best

- Recall: YAGO TyCor is very expensive and one of many
- Recall: Bidirectional Attention Flow is relatively cheap
- We would like to hear that the cheap system is also the best
Metrics

– Exact Match
  • % of questions where the top-ranked answer exactly matches the answer-key

– Mean Reciprocal Rank
  • Average across all questions of the reciprocal rank of the highest ranked correct answer
  • First answer correct gets 1, second gets \( \frac{1}{2} \), third gets \( \frac{1}{3} \), etc.

– Other metrics: See paper
## SQuAD Results

<table>
<thead>
<tr>
<th></th>
<th>Exact Match</th>
<th>Mean Rec. Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Statistical</td>
<td>66%</td>
<td>71%</td>
</tr>
<tr>
<td>DDQA + Statistical</td>
<td>67%</td>
<td>73%</td>
</tr>
</tbody>
</table>

- The statistical system alone provides nearly all of the power.
- Adding DDQA provides very little benefit despite all of its great cost.
The statistical system alone provides very little power.

Adding DDQA provides enormous benefit.
A system that excels at SQuAD will also excel at factoid question answering.

Hypothesis: Not Confirmed

Needs more evidence
Is SQuAD a toy problem?

– Recall: Not a particularly useful capability by itself

– Is a statistical system learning reading comprehension? Or just statistical trends in how people write questions for given passages?

– (Not talking about the metaphysical question)

– Are the systems learning to identify answers given \((\text{passage}, \text{question})\) or given \((\text{passage}, \text{question-written-for-that-passage})\)?

... Various species of poison dart frogs secrete **lipophilic alkaloid toxins** through their flesh ...

What are dart frogs are known to secrete?

Amphibians secrete a wide diversity of chemicals from skin glands as defense against predators, parasites, and pathogens. Most defensive chemicals are produced endogenously through biosynthesis, but poison frogs sequester **lipophilic alkaloids** from dietary arthropods. *

I'm looking for my quarter I dropped!

DID YOU DROP IT HERE?

No, I dropped it two blocks down the street!

THEN WHY ARE YOU LOOKING FOR IT HERE?

Because the light is better here!
Is factoid a toy problem?

– Users do not really want a system that only answers factoid questions
– *Most* information needs require more than just an entity or quantity to address
– *Sometimes* users do ask factoid questions
– When they do, it is nice to provide a (correct) factoid answer
– The complete factoid question answering problem seems *closer* to a real-world use case than SQuAD reading comprehension.
– This will only be *proven* when we show that it adds value as part of a comprehensive information finding system.
– Google and WolframAlpha work very well for some kinds of factoid questions

– To our knowledge, nobody has a big commercial success doing narrow-domain factoid question answering using customer supplied content.
Engineered AI still matters

– Mounting evidence from data like SQuAD: single-strategy deep neural networks are the state-of-the-art for answering questions

– Might be an artifact of the limitations of SQuAD and similar data sets

– Need more experiments with more data

– We believe that there is still a significant role for multi-strategy systems that make extensive use of engineered knowledge and rules.