CS388: Natural Language Processing

Lecture 23: Question Answering 2

Greg Durrett

TExAS
The University of Texas at Austin
Announcements

- eCIS
- Project 2 back (except for late days)
- Jason Baldridge talk Thursday
Recall: SQuAD

- Single-document, single-sentence question-answering task where the answer is always a substring of the passage
- Predict start and end indices of the answer in the passage

Rajpurkar et al. (2016)
Recall: Bidirectional Attention Flow

Each passage word now “knows about” the query
Recall: QA with BERT

What was Marie Curie the first female recipient of? [SEP] One of the most famous people born in Warsaw was Marie...

- Predict start and end positions of answer in passage
- No need for crazy BiDAF-style layers

Devlin et al. (2019)
Recall: What are these models learning?

ABC isn’t used at all! The model is mostly using the fact that only one typeface is in the context.

Ye, Nair, Durrett (2021)
This Lecture

- Problems in QA, especially related to answer type overfitting
- Retrieval-based QA / multi-hop QA
- New QA frontiers
Problems in QA
SQuAD questions are often easy: “what was she the recipient of?” passage: “...recipient of Nobel Prize...”

Jia and Liang (2017)
What was Marie Curie the first female recipient of? [SEP] ... first female recipient of the Nobel Prize ...

- BERT easily learns surface-level correspondences like this with self-attention
Adversarial SQuAD

- SQuAD questions are often easy: “what was she the recipient of?” passage: “...recipient of Nobel Prize...”

- Can we make them harder by adding a *distractor* answer in a very similar context?

- Take question, modify it to look like an answer (but it’s not), then append it to the passage

Jia and Liang (2017)
Adversarial SQuAD

**Article:** Super Bowl 50

**Paragraph:** “Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver’s Executive Vice President of Football Operations and General Manager. Quarterback Jeff Dean had jersey number 37 in Champ Bowl XXXIV.”

**Question:** “What is the name of the quarterback who was 38 in Super Bowl XXXIII?”

**Original Prediction:** John Elway

**Prediction under adversary:** Jeff Dean

- **Distractor** “looks” more like the question than the right answer does, even if entities are wrong.
Weakness to Adversaries

<table>
<thead>
<tr>
<th>Model</th>
<th>Original</th>
<th>ADDONESENT</th>
</tr>
</thead>
<tbody>
<tr>
<td>ReasoNet-E</td>
<td>81.1</td>
<td>49.8</td>
</tr>
<tr>
<td>SEDT-E</td>
<td>80.1</td>
<td>46.5</td>
</tr>
<tr>
<td>BiDAF-E</td>
<td>80.0</td>
<td>46.9</td>
</tr>
<tr>
<td>Mnemonic-E</td>
<td>79.1</td>
<td>55.3</td>
</tr>
<tr>
<td>Ruminating</td>
<td>78.8</td>
<td>47.7</td>
</tr>
<tr>
<td>jNet</td>
<td>78.6</td>
<td>47.0</td>
</tr>
<tr>
<td>Mnemonic-S</td>
<td>78.5</td>
<td>56.0</td>
</tr>
<tr>
<td>ReasoNet-S</td>
<td>78.2</td>
<td>50.3</td>
</tr>
<tr>
<td>MPCM-S</td>
<td>77.0</td>
<td>50.0</td>
</tr>
<tr>
<td>SEDT-S</td>
<td>76.9</td>
<td>44.8</td>
</tr>
<tr>
<td>RaSOR</td>
<td>76.2</td>
<td>49.5</td>
</tr>
<tr>
<td>BiDAF-S</td>
<td>75.5</td>
<td>45.7</td>
</tr>
<tr>
<td>Match-E</td>
<td>75.4</td>
<td>41.8</td>
</tr>
<tr>
<td>Match-S</td>
<td>71.4</td>
<td>39.0</td>
</tr>
<tr>
<td>DCR</td>
<td>69.3</td>
<td>45.1</td>
</tr>
<tr>
<td>Logistic</td>
<td>50.4</td>
<td>30.4</td>
</tr>
</tbody>
</table>

- Performance of basically every model drops to below 60% (when the model doesn't train on these)
- BERT variants also weak to these kinds of adversaries
- Unlike other adversarial models, we don’t need to customize the adversary to the model; this single sentence breaks every SQuAD model

Jia and Liang (2017)
Universal Adversarial “Triggers”

**Input** (underline = correct span, red = trigger, underline = target span)

**Question:** Why did he walk?
For exercise, Tesla walked between 8 to 10 miles per day. He squished his toes one hundred times for each foot every night, saying that it stimulated his brain cells. **why how because to kill american people.**

**Question:** Why did the university see a drop in applicants?
In the early 1950s, student applications declined as a result of increasing crime and poverty in the Hyde Park neighborhood. In response, the university became a ....... **why how because to kill american people.**

- Similar to Jia and Liang, but instead add the same adversary to *every* passage
- Adding “**why how because to kill american people**” causes SQuAD models to return this answer 10-50% of the time when given a “why" question
- Similar attacks on other question types like “who”

Wallace et al. (2019)
How to fix QA?

- Better models?
  - But a model trained on weak data will often still be weak to adversaries
  - Training on Jia+Liang adversaries can help, but there are plenty of other similar attacks which that doesn't solve

- Better datasets
  - Same questions but with more distractors may challenge our models
  - Next up: retrieval-based QA models

- Harder QA tasks
  - Ask questions which cannot be answered in a simple way
  - Afterwards: multi-hop QA and other QA settings
Retrieval Models
SQuAD-style QA is very artificial, not really a real application.

Real QA systems should be able to handle more than just a paragraph of context — theoretically should work over the whole web?

Q: What was Marie Curie the recipient of?

Marie Curie was awarded the Nobel Prize in Chemistry and the Nobel Prize in Physics...

Mother Teresa received the Nobel Peace Prize in...

Curie received his doctorate in March 1895...

Skłodowska received accolades for her early work...
Open-domain QA

- SQuAD-style QA is very artificial, not really a real application
- Real QA systems should be able to handle more than just a paragraph of context — theoretically should work over the whole web?
- This also introduces more complex *distractors* (bad answers) and should require stronger QA systems
- QA pipeline: given a question:
  - Retrieve some documents with an IR system
  - Zero in on the answer in those documents with a QA model
DrQA

‣ How often does the retrieved context contain the answer? (uses Lucene, basically sparse tf-idf vectors)

‣ Full retrieval results using a QA model trained on SQuAD: task is much harder

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Wiki Search</th>
<th>Doc. Retriever plain</th>
<th>Doc. Retriever +bigrams</th>
</tr>
</thead>
<tbody>
<tr>
<td>SQuAD</td>
<td>62.7</td>
<td>76.1</td>
<td>77.8</td>
</tr>
<tr>
<td>CuratedTREC</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>

<table>
<thead>
<tr>
<th>Dataset</th>
<th>SQuAD</th>
</tr>
</thead>
<tbody>
<tr>
<td>SQuAD (All Wikipedia)</td>
<td>27.1</td>
</tr>
<tr>
<td>CuratedTREC</td>
<td>19.7</td>
</tr>
<tr>
<td>WebQuestions</td>
<td>11.8</td>
</tr>
<tr>
<td>WikiMovies</td>
<td>24.5</td>
</tr>
</tbody>
</table>

Chen et al. (2017)
Problems

- Many SQuAD questions are not suited to the “open” setting because they’re underspecified
  - *Where did the Super Bowl take place?*
  - *Which player on the Carolina Panthers was named MVP?*
- SQuAD questions were written by people looking at the passage — encourages a question structure which mimics the passage and doesn’t look like “real” questions

Lee et al. (2019)
Natural Questions

- Real questions from Google, answerable with Wikipedia
- Short answers and long answers (snippets)
- Questions arose naturally, unlike SQuAD questions which were written by people looking at a passage. This makes them much harder
- Short answer F1s < 60, long answer F1s < 75

**Question:**
where is blood pumped after it leaves the right ventricle?

**Short Answer:**
None

**Long Answer:**
From the right ventricle, blood is pumped through the semilunar pulmonary valve into the left and right main pulmonary arteries (one for each lung), which branch into smaller pulmonary arteries that spread throughout the lungs.

Kwiatkowski et al. (2019)
Can we do better than a simple IR system?

Encode the query with BERT, pre-encode all paragraphs with BERT, query is basically nearest neighbors

\[
\begin{align*}
    h_q &= W_q BERT_Q(q)[CLS]
    \\
    h_b &= W_b BERT_B(b)[CLS]
    \\
    S_{retr}(b, q) &= h_q^T h_b
\end{align*}
\]
- Technique for integrating retrieval into pre-training
- Retriever relies on a maximum inner-product search (MIPS) over BERT embeddings
- MIPS is fast — challenge is how to refresh the BERT embeddings
- Fine-tuning can exploit the same kind of textual knowledge
- Can work for tasks requiring knowledge lookups
### REALM

<table>
<thead>
<tr>
<th>Name</th>
<th>Architectures</th>
<th>Pre-training</th>
<th>NQ (79k/4k)</th>
<th>WQ (3k/2k)</th>
<th>CT (1k/1k)</th>
<th># params</th>
</tr>
</thead>
<tbody>
<tr>
<td>BERT-Baseline (Lee et al., 2019)</td>
<td>Sparse Retr. + Transformer</td>
<td>BERT</td>
<td>26.5</td>
<td>17.7</td>
<td>21.3</td>
<td>110m</td>
</tr>
<tr>
<td>T5 (base) (Roberts et al., 2020)</td>
<td>Transformer Seq2Seq</td>
<td>T5 (Multitask)</td>
<td>27.0</td>
<td>29.1</td>
<td>-</td>
<td>223m</td>
</tr>
<tr>
<td>T5 (large) (Roberts et al., 2020)</td>
<td>Transformer Seq2Seq</td>
<td>T5 (Multitask)</td>
<td>29.8</td>
<td>32.2</td>
<td>-</td>
<td>738m</td>
</tr>
<tr>
<td>T5 (11b) (Roberts et al., 2020)</td>
<td>Transformer Seq2Seq</td>
<td>T5 (Multitask)</td>
<td>34.5</td>
<td>37.4</td>
<td>-</td>
<td>11318m</td>
</tr>
<tr>
<td>DrQA (Chen et al., 2017)</td>
<td>Sparse Retr. + DocReader</td>
<td>N/A</td>
<td>-</td>
<td>20.7</td>
<td>25.7</td>
<td>34m</td>
</tr>
<tr>
<td>HardEM (Min et al., 2019a)</td>
<td>Sparse Retr. + Transformer</td>
<td>BERT</td>
<td>28.1</td>
<td>-</td>
<td>-</td>
<td>110m</td>
</tr>
<tr>
<td>GraphRetriever (Min et al., 2019b)</td>
<td>GraphRetriever + Transformer</td>
<td>BERT</td>
<td>31.8</td>
<td>31.6</td>
<td>-</td>
<td>110m</td>
</tr>
<tr>
<td>PathRetriever (Asai et al., 2019)</td>
<td>PathRetriever + Transformer</td>
<td>MLM</td>
<td>32.6</td>
<td>-</td>
<td>-</td>
<td>110m</td>
</tr>
<tr>
<td>ORQA (Lee et al., 2019)</td>
<td>Dense Retr. + Transformer</td>
<td>ICT + BERT</td>
<td>33.3</td>
<td>36.4</td>
<td>30.1</td>
<td>330m</td>
</tr>
<tr>
<td>Ours (X = Wikipedia, Z = Wikipedia)</td>
<td>Dense Retr. + Transformer</td>
<td>REALM</td>
<td>39.2</td>
<td>40.2</td>
<td>46.8</td>
<td>330m</td>
</tr>
<tr>
<td>Ours (X = CC-News, Z = Wikipedia)</td>
<td>Dense Retr. + Transformer</td>
<td>REALM</td>
<td><strong>40.4</strong></td>
<td><strong>40.7</strong></td>
<td>42.9</td>
<td>330m</td>
</tr>
</tbody>
</table>

- 330M parameters + a knowledge base beats an 11B parameter T5 model

Guu et al. (2020)
Multi-Hop Question Answering
Mul;-Hop	Ques;on	Answering

‣ Very few SQuAD questions require actually combining multiple pieces of information — this is an important capability QA systems should have

‣ Several datasets test *multi-hop reasoning*: ability to answer questions that draw on several sentences or several documents to answer

Welbl et al. (2018), Yang et al. (2018)
Annotators shown Wikipedia and asked to pose a simple question linking two entities that require a third (bridging) entity to associate.

A model shouldn’t be able to answer these without doing some reasoning about the intermediate entity.

Figure from Welbl et al. (2018)
**Question:** What government position was held by the woman who portrayed Corliss Archer in the film Kiss and Tell?

**Doc 1**
Shirley Temple Black was an American actress, businesswoman, and singer ...

As an adult, she served as **Chief of Protocol of the United States** ...

**Doc 2**
Kiss and Tell is a comedy film in which 17-year-old Shirley Temple acts as Corliss Archer.

**Doc 3**
Meet Corliss Archer is an American television sitcom that aired on CBS ...

- Much longer and more convoluted questions

Example picked from HotpotQA [Yang et al., 2018]
Multi-hop Reasoning

**Question:** The Oberoi family is part of a hotel company that has a head office in what city?

The Oberoi family is an Indian family that is famous for its involvement in hotels, namely through The Oberoi Group...

The Oberoi Group is a hotel company with its head office in Delhi. ...

This is an idealized version of multi-hop reasoning. Do models need to do this to do well on this task?

Example picked from HotpotQA [Yang et al., 2018]
The Oberoi family is an Indian family that is famous for its involvement in hotels, namely through The Oberoi Group.

**Question:** The Oberoi family is part of a hotel company that has a head office in what city?

**Doc 1:** The Oberoi family is an Indian family that is famous for its involvement in hotels, namely through The Oberoi Group.

**Doc 2:** The Oberoi Group is a hotel company with its head office in Delhi.

Model can ignore the bridging entity and directly predict the answer.

Example picked from HotpotQA (Yang 2018)
Question: What government position was held by the woman who portrayed Corliss Archer in the film Kiss and Tell?

Shirley Temple Black was an American actress, businesswoman, and singer... As an adult, she served as Chief of Protocol of the United States.

Kiss and Tell is a comedy film in which 17-year-old Shirley Temple acts as Corliss Archer.

Meet Corliss Archer is an American television sitcom that aired on CBS...

No simple lexical overlap.
...but only one government position appears in the context!

Example picked from HotpotQA [Yang et al., 2018]
Can a model identify the answer with only a set of candidates?

*Government position* → *Chief of Protocol, actress, singer*

Can a model identify where the answer is in a single hop?

*Oberoi Family* → *Delhi*
Finding the answer directly

**Question:** What government position was held by the woman who portrayed Corliss Archer in the film Kiss and Tell?

**Chief of Protocol**, **businesswoman**, ... **actress**

*Kaushik and Lipton (2018)*
**Question:** What government position was held by the woman who portrayed Corliss Archer in the film Kiss and Tell?

Chen and Durrett (2019)
More than half of questions can be answered without even using the context!

- SOTA models trained on this may be learning question-answer correspondences, not multi-hop reasoning as advertised

Chen and Durrett (2019)
Can a model identify the answer with only a set of candidates?

Government position \(\rightarrow\) Chief of Protocol, actress, singer

Can a model identify where the answer is in a single hop?

Oberoi Family \(\rightarrow\) Delhi

Chen and Durrett (2019)
Find the answer by comparing each sentence with the question *separately*!

**Question:** The Oberoi family is part of a hotel company that has a head office in what city?

Doc 1

*The Oberoi family is an Indian family that is ...*

Doc 2

*The Oberoi Group is a hotel company with its head office in Delhi.*

Doc 3

*Future Fibre Technologies a fiber technologies company ...*

Chen and Durrett (2019)
Answer prediction:

Delhi

Softmax over all sentences is the only cross-sentence interaction

The Oberoi family...

The Oberoi Group ... in Delhi.

Future Fibre Technologies is a fibre...

The Oberoi family ...

Chen and Durrett (2019)
Results on HotpotQA

A simple single sentence reasoning model can solve more than half questions on HotpotQA.

Chen and Durrett (2019)
Other Work

- Min et al. ACL 2019 “Compositional Questions do not Necessitate Multi-hop Reasoning”
- Focuses just on HotpotQA
- Additionally tries to adversarially harden Hotpot against these attacks. Some limited success, but doesn't solve the problem
Maybe we can strengthen our models to avoid these weaknesses. Force them to explicitly extract a reasoning chain to make them better.

Chen et al. (2019)
**Question:** What government position was held by the woman who portrayed Corliss Archer in the film *Kiss and Tell*?

**Answer:** Chief of Protocol

**Reasoning Chain 1**

- *Shirley Temple Black* was an American actress, businesswoman, and diplomat...
  - As an adult, she served as the Chief of Protocol of the United States...
  - She began her diplomatic career in 1969, when she represented...
- *Kiss and Tell* is a film in which 17-year-old Shirley Temple acts as *Corliss Archer*.
- “A Kiss for Corliss” is a sequel to the film “Kiss and Tell”.
  - It stars *Shirley Temple* in her final starring role...

- **Strong connection between the entities used here**

Chen et al. (2019)
Chain Supervision

- Extract pseudogold chains based on:
  - Within-document coreference: we don’t run a coreference system but instead link all sentences within a paragraph
  - Shared entities: enable connections between different sources
  - Given these chains, we learn a model to extract them. **At test time, no annotations are needed**

Chen et al. (2019)
Chain Extraction and QA

- Paragraphs are encoded with BERT to compute sentence representations
- A pointer network selects a sequence of sentences
- A final BERT model then extracts an answer span from one or more chains

Chen et al. (2019)
High performance on WikiHop (*past systems didn't use BERT) and Hotpot

Also large gains on hard examples in HotpotQA
Best systems: use hyperlink structure of Wikipedia and a strong multi-step retrieval mode built on BERT

Asai et al. (2020)
New Types of QA
QA datasets to model programs/computation

That year, his Untitled (1981), a painting of a haloed, black-headed man with a bright red skeletal body, depicted amid the artists signature scrawls, was sold by Robert Lehrman for $16.3 million, well above its $12 million high estimate.

<table>
<thead>
<tr>
<th>Passage (some parts shortened)</th>
<th>Question</th>
<th>Answer</th>
<th>BiDAF</th>
</tr>
</thead>
<tbody>
<tr>
<td>How many more dollars was the Untitled (1981) painting sold for than the 12 million dollar estimation?</td>
<td>4300000</td>
<td>$16.3 million</td>
<td></td>
</tr>
</tbody>
</table>

- Question types: subtraction, comparison (*which did he visit first*), counting and sorting (*which kicker kicked more field goals*),
- Invites ad hoc solutions like predicting two numbers + operation
MultiQA

- Maybe we should just look at lots of QA datasets instead?

<table>
<thead>
<tr>
<th></th>
<th>CQ</th>
<th>CWQ</th>
<th>ComQA</th>
<th>WikiHop</th>
<th>DROP</th>
<th>SQUAD</th>
<th>NEWSQA</th>
<th>SEARCHQA</th>
<th>TQA-G</th>
<th>TQA-W</th>
<th>HOTPOTQA</th>
</tr>
</thead>
<tbody>
<tr>
<td>SQUAD</td>
<td>23.6</td>
<td>12.0</td>
<td>20.0</td>
<td>4.6</td>
<td>5.5</td>
<td>-</td>
<td>31.8</td>
<td>8.4</td>
<td>37.8</td>
<td>33.4</td>
<td>11.8</td>
</tr>
<tr>
<td>NEWSQA</td>
<td>24.1</td>
<td>12.4</td>
<td>18.9</td>
<td>7.1</td>
<td>4.4</td>
<td>60.4</td>
<td>-</td>
<td>10.1</td>
<td>37.6</td>
<td>28.4</td>
<td>8.0</td>
</tr>
<tr>
<td>SEARCHQA</td>
<td>30.3</td>
<td>18.5</td>
<td>25.8</td>
<td>12.4</td>
<td>2.8</td>
<td>23.3</td>
<td>12.7</td>
<td>-</td>
<td>53.2</td>
<td>35.4</td>
<td>5.2</td>
</tr>
</tbody>
</table>

... 

- BERT trained on SQuAD gets <40% performance on any other QA dataset

- Our QA models are pretty good at fitting single datasets with 50k-100k examples, but still aren’t learning general question answering

Talmor and Berant (2019)
Humans see a summary of a book: ...Peter’s former girlfriend Dana Barrett has had a son, Oscar...

Question: How is Oscar related to Dana?

Answering these questions from the source text (not summary) requires complex inferences and is extremely challenging; no progress on this dataset in 2 years

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.
Takeaways

- Lots of problems with current QA settings, lots of new datasets

- QA over tables, images, knowledge bases, ...

- Models can often work well for one QA task but don’t generalize

- There’s lots that we can’t do, but we’re getting really good at putting our hands on random facts from the Internet