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

Recall: Bidirectional Attention Flow

- Each passage word now "knows about" the query

Recall: QA with BERT

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

<table>
<thead>
<tr>
<th>Rank</th>
<th>Model</th>
<th>EM</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>ALBERT (ensemble model)</td>
<td>89.79</td>
<td>92.03</td>
</tr>
<tr>
<td>2</td>
<td>XLNet + DAAN + VeriVer (ensemble)</td>
<td>88.59</td>
<td>90.89</td>
</tr>
<tr>
<td></td>
<td>FANGAN (Open-Src)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>ALBERT (single model)</td>
<td>88.107</td>
<td>90.902</td>
</tr>
<tr>
<td>2</td>
<td>LPM (ensemble)</td>
<td>88.221</td>
<td>90.753</td>
</tr>
<tr>
<td></td>
<td>Anonymous</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>XLNet + SG-Net Verifier (ensemble)</td>
<td>88.174</td>
<td>90.702</td>
</tr>
<tr>
<td>4</td>
<td>XLNet + 5G-Net Verifier (single model)</td>
<td>87.238</td>
<td>90.073</td>
</tr>
</tbody>
</table>

- Performance is very saturated
- Harder QA settings are needed

This Lecture

- Problems in QA, especially related to answer type overfitting
- Retrieval-based QA / multi-hop QA
- New QA frontiers

Problems in QA

Adversarial SQuAD

- SQuAD questions are often easy: "what was she the recipient of?" passage: “… recipient of Nobel Prize…"
Adversarial SQuAD

What was Marie Curie the first female recipient of?

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

SQuAD questions are often easy: “what was she the recipient of?” passage: “… first female recipient of the 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

Weakness to Adversaries

- 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

Adversarial SQuAD

Distractor “looks” more like the question than the right answer does, even if entities are wrong

Performance

<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>RsSOR</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>

Jia and Liang (2017)
Universal Adversarial “Triggers”

Wallace et al. (2019)

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 to Jia and Liang, but instead add the same adversary to every passage.

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

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?

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...

Retrieval Models
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)
- 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 +bigrams</th>
</tr>
</thead>
<tbody>
<tr>
<td>SQuAD</td>
<td>62.7</td>
<td>76.1 77.8</td>
</tr>
<tr>
<td>CuratedTREC</td>
<td>81.0</td>
<td>85.2 86.0</td>
</tr>
<tr>
<td>WebQuestions</td>
<td>73.7</td>
<td>75.5 74.4</td>
</tr>
<tr>
<td>WikiMovies</td>
<td>61.7</td>
<td>54.4 70.3</td>
</tr>
</tbody>
</table>

Retrieval with BERT

- 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

\[
h_q = W_q \text{BERT}_Q(q)[CLS]
\]
\[
h_b = W_b \text{BERT}_B(b)[CLS]
\]
\[
S_{rer} = h_q^T h_b
\]

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
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  
  Kwiatkowski et al. (2019)

Multi-Hop Question 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

Multi-Hop Question Answering

- 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

WikiHop

- Figure from Welbl et al. (2018)
Meat Corliss Archer is an American television sitcom that aired on CBS.

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.

- Much longer and more convoluted questions

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.

Model can ignore the bridging entity and directly predict the answer

High lexical overlap

No simple lexical overlap.

...but only one government position appears in the context!
**Investigation**

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

Chen and Durrett (2019)

**Finding the answer directly**

**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...

**Answer:** Chief of Protocol, businesswoman, actress

Kaushik and Lipton (2018)

**No Context Baseline**

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

**Answer:** Chief of Protocol

Chen and Durrett (2019)

**Results on WikiHop**

- 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
**Investigation**

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*

---

**Sentence Factored Model**

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?

**Answer prediction**: Delhi

Softmax over all sentences is the *only* cross-sentence interaction

---

**Sentence Factored Model**

*Answer prediction: Delhi*

软性选择所有句子是唯一的跨句交互

---

**Results on HotpotQA**

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

- Min et al. ACL 2019 “Compositional Questions do not Necessitate Multi-hop Reasoning”
- Focused just on HotpotQA
- Additionally tries to adversarially harden Hotpot against these attacks. Some limited success, but doesn’t solve the problem

Question Answering with Chains

- 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:
  - In Doc
  - Coref
  - Other
  - Shared Entity

  Chain Extractor

  Final Answer Span

- "Kiss and Tell" is a comedy 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

Question Answering with Chains

- 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 2:
  - In Doc
  - Coref
  - Other
  - Shared Entity

  Chain Extractor

  Final Answer Span

- "Kiss and Tell" is a comedy 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...

- More speculative than the other chain but still leads to the answer

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.**

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.

QA Results

- High performance on WikiHop (*past systems didn’t use BERT) and Hotpot.
- **Also large gains on hard examples in HotpotQA** (our model from part 1 could not find answers in a single hop).
- Ongoing work: how can reasoning chains be taken below the sentence level and be more strongly tied to interpretable logical inference?

New Types of QA
One thread of research: let’s build QA datasets to help the community focus on modeling particular things

**Passage (some parts shortened)**

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

**Question types:** subtraction, comparison (*which did he visit first*), counting and sorting (*which kicker kicked more field goals*), invites ad hoc solutions (structure the model around predicting differences between numbers)

Dua et al. (2019)

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

**MultiQA**

<table>
<thead>
<tr>
<th>CQ</th>
<th>CWQ</th>
<th>ComQA</th>
<th>WikiHop</th>
<th>DROP</th>
<th>SQuAD</th>
<th>NewsQA</th>
<th>SQuAD</th>
<th>TQA-G</th>
<th>TQA-W</th>
<th>HIPQA</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>
</tr>
<tr>
<td>NewsQA</td>
<td>24.1</td>
<td>12.4</td>
<td>18.9</td>
<td>7.1</td>
<td>6.4</td>
<td>60.4</td>
<td>-</td>
<td>10.1</td>
<td>37.6</td>
<td>28.4</td>
</tr>
<tr>
<td>SEARCHQA</td>
<td>30.3</td>
<td>18.5</td>
<td>23.8</td>
<td>12.4</td>
<td>2.8</td>
<td>23.3</td>
<td>12.7</td>
<td>-</td>
<td>53.2</td>
<td>58.4</td>
</tr>
</tbody>
</table>

... Talmor and Berant (2019)

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

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

We still don’t have (solvable) QA settings which seem to require really complex reasoning as opposed to surface-level pattern recognition

Kočiský et al. (2017)