CSE 5525: Foundations of Speech and Language Processing

Question Answering I
Huan Sun (CSE@OSU)

Many thanks to Prof. Greg Durrett @ UT Austin for sharing his slides. Some images/examples were from the two textbooks by (1) Jurafsky and Martin and (2) Eisenstein.
Previously, QA as Semantic Parsing

- Many ways to build these parsers

- One approach: run a “supertagger” (tagging the sentence with complex labels), then run the parser

\[
\begin{align*}
\text{What} & \quad \text{states} & \quad \text{border} & \quad \text{Texas} \\
\frac{(S/(S\backslash NP))/N}{\lambda f.\lambda g.\lambda x.f(x) \land g(x)} & \frac{N}{\lambda x.\text{state}(x)} & \frac{(S\backslash NP)/NP}{\lambda x.\lambda y.\text{borders}(y, x)} & \frac{NP}{\text{texas}}
\end{align*}
\]

- Parsing is easy once you have the tags, so we’ve reduced it to a (hard) tagging problem

CCG parsing

Zettlemoyer and Collins (2005)
Derivation $d$ on sentence $x$: 

- No more explicit syntax in these derivations like we had in CCG

Building the lexicon: more sophisticated process than GENLEX, but can handle thousands of predicates

Log-linear model with features on rules: 

$$ P(d| x) \propto \exp w^\top \left( \sum_{r \in d} f(r, x) \right) $$

Similar to CRF parsers
This Lecture

- Types of question answering/reading comprehension
- CNN/Daily Mail task: Attentive Reader
- SQuAD task: Bidirectional Attention Flow
Question Answering

- Form semantic representation from semantic parsing, execute against structured knowledge base

Q: where was Barack Obama born

\[ \lambda x. \text{type}(x, \text{Location}) \land \text{born_in}(\text{Barack\_Obama}, x) \]

(also Prolog / GeoQuery, etc.)
Question Answering

- Form semantic representation from semantic parsing, execute against structured knowledge base

Q: *where was Barack Obama born*

\[
\lambda x. \text{type}(x, \text{Location}) \land \text{born_in}(\text{Barack_Obama}, x)
\]

(also Prolog / GeoQuery, etc.)

- When do we formulate QA as a semantic parsing task?
QA is very broad

- Factoid QA: what states border Mississippi?, when was Barack Obama born?
  - Lots of this could be handled by QA over a knowledge base, if we had a big enough knowledge base.
- “Question answering”: many types; very broad
  - Is P=NP?
  - What is 4+5?
  - What is the translation of [sentence] into French? [McCann et al., 2018]
QA is very broad

- Factoid QA: *what states border Mississippi?*, *when was Barack Obama born?*
  - Lots of this could be handled by QA over a knowledge base, if we had a big enough knowledge base

- From the point of data source’s view:
  - Based on knowledge bases
  - Based on texts and tables (unstructured, or semi-structured data)
  - Community QA (such as Stack Overflow, WebMD, etc)
What are the limits of QA?

- Focus on questions where the answer might plausibly appear in text... but this is Still TOO BROAD

- *What were the main causes of World War II?* — requires summarization

- *Can you get the flu from a flu shot?* — want IR to provide an explanation of the answer, not just yes/no

- *What temperature should I cook chicken to?* — could be written down in a KB but probably isn’t
What are the limits of QA?

- Focus on questions where the answer might plausibly appear in text... but this is still too broad

- *What were the main causes of World War II?* — requires summarization

- *Can you get the flu from a flu shot?* — want IR to provide an explanation of the answer, not just yes/no

- *What temperature should I cook chicken to?* — could be written down in a KB but probably isn’t

- Today: can we do QA when it *requires retrieving the answer from a passage*?
Reading Comprehension
“AI challenge problem”: answer question given context

Questions per passage explicitly require cross-sentence reasoning

One day, James thought he would go into town and see what kind of trouble he could get into. He went to the grocery store and pulled all the pudding off the shelves and ate two jars. Then he walked to the fast food restaurant and ordered 15 bags of fries. He didn't pay, and instead headed home.

3) Where did James go after he went to the grocery store?
A) his deck
B) his freezer
C) a fast food restaurant
D) his room

Richardson (2013)
Baselines

- **N-gram matching**: append question + each answer, return answer which gives highest n-gram overlap with a sentence.

- **Parsing**: find direct object of “pulled” in the document where the subject is James.

- Don’t need any complex semantic representations

---

One day, James thought he would go into town and see what kind of trouble he could get into. He went to the grocery store and **pulled all the pudding off the shelves** and ate two jars. Then he walked to the fast food restaurant and ordered 15 bags of fries. He didn't pay, and instead headed home.

2) What did James **pull off of the shelves** in the grocery store?

   A) pudding
   B) fries
   C) food
   D) splinters

Richardson (2013)
Classic textual entailment systems don’t work as well as n-grams.

Scores are low partially due to questions spanning multiple sentences.

Unfortunately not much data to train better methods on (2000 questions).
Dataset Explosion

- 30+ QA datasets released since 2015
- Question answering: questions are in natural language
  - Answers: multiple choice, require picking from the passage, or generate freeform answer (last is pretty rare)
  - Require human annotation
30+ QA datasets released since 2015

Question answering: questions are in natural language

- Answers: multiple choice, require picking from the passage, or generate freeform answer (last is pretty rare)
- Require human annotation

“Cloze” task: word (often an entity) is removed from a sentence

- Answers: multiple choice, pick from passage, or pick from vocabulary
- Can be created automatically from things that aren’t questions
You can learn to analyze datasets along these axes
Dataset Properties

- **Axis 1**: cloze task (fill in blank) vs. multiple choice vs. span-based vs. freeform generation
- **Axis 2**: what’s the input?
  - One paragraph? One document? All of Wikipedia?
  - Some explicitly require linking between multiple sentences (MCTest, WikiHop, HotpotQA)

You can learn to analyze datasets along these axes
Dataset Properties

- **Axis 1**: cloze task (fill in blank) vs. multiple choice vs. span-based vs. freeform generation
- **Axis 2**: what’s the input?
  - One paragraph? One document? All of Wikipedia?
  - Some explicitly require linking between multiple sentences (MCTest, WikiHop, HotpotQA)
- **Axis 3**: what capabilities are needed to answer questions?
  - Finding simple information? Combining information across multiple sources?

You can learn to analyze datasets along these axes
children's book test: take a section of a children's story, block out an entity and predict it (one-doc multi-sentence cloze task)
Mr. Baxter privately had no hope that they would do best. She could not believe that Mr. Cropper would carry his prejudices into a personal application. This conviction was strengthened when he overtook her walking from school the next day and drove her home. He was a big, handsome man with a very suave, polite manner. He asked interestingly about her school and her work, hoped she was getting on well, and said he had two young rascals of his own to send soon. Esther felt relieved. She thought that had exaggerated matters a little.

What types of models can we develop?
LSTM Language Models

Mr. Baxter privately had no hope that they would, but Esther hoped for the best. She could not believe that Mr. Cropper would carry his prejudices into a personal application. This conviction was strengthened when he overtook her walking from school the next day and drove her home. He was a big, handsome man with a very suave, polite manner. He asked interestingly about her school and her work, hoped she was getting on well, and said he had two young rascals of his own to send soon. Esther felt relieved. She thought that had exaggerated matters a little.

- Predict next word with LSTM LM
- Context: either just the current sentence (query) or the whole document up to this point (query+context)

Hill et al. (2015)
**Context:** They tuned, discussed for a moment, then struck up a lively jig. Everyone joined in, turning the courtyard into an even more chaotic scene, people now dancing in circles, swinging and spinning in circles, everyone making up their own dance steps. I felt my feet tapping, my body wanting to move.

**Target sentence:** Aside from writing, I’ve always loved _____.

**Target word:** dancing

- GPT/BERT can in general do very well at cloze tasks because this is what they’re trained to do.

- Hard to come up with plausible alternatives: “cooking”, “drawing”, “soccer”, etc. don’t work in the above context.
Dataset was constructed to be difficult for ELMo (pretrained language model)

- BERT subsequently got 20+% accuracy improvements and achieved human-level performance

Zellers et al. (2018)

The person blows the leaves from a grass area using the blower. The blower...

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>a)</td>
<td>puts the trimming product over her face in another section.</td>
</tr>
<tr>
<td>b)</td>
<td>is seen up close with different attachments and settings featured.</td>
</tr>
<tr>
<td>c)</td>
<td>continues to blow mulch all over the yard several times.</td>
</tr>
<tr>
<td>d)</td>
<td>blows beside them on the grass.</td>
</tr>
</tbody>
</table>


Dataset was constructed to be difficult for ELMo (pretrained language model)

BERT subsequently got 20+% accuracy improvements and achieved human-level performance

Problem: distractors too easy

Let’s look at architectures for retrieval from a passage

The person blows the leaves from a grass area using the blower. The blower...

| a) puts the trimming product over her face in another section. |
| b) is seen up close with different attachments and settings featured. |
| c) continues to blow mulch all over the yard several times. |
| d) blows beside them on the grass. |
CNN/Daily Mail: Attentive Reader
Single-document, (usually) single-sentence cloze task

Formed based on article summaries — information should mostly be present

Need to process the question, can’t just use LSTM LMs

CNN/Daily Mail

Passage

( @entity4 ) if you feel a ripple in the force today, it may be the news that the official @entity6 is getting its first gay character, according to the sci-fi website @entity9, the upcoming novel "@entity11" will feature a capable but flawed @entity13 official named @entity14 who "also happens to be a lesbian." the character is the first gay figure in the official @entity6 -- the movies, television shows, comics and books approved by @entity6 franchise owner @entity22 -- according to @entity24, editor of "@entity6" books at @entity28 imprint @entity26.

Question

characters in "@placeholder"

movies have gradually become more diverse

Answer

@entity6

Hermann et al. (2015), Chen et al. (2016)
LSTM reader: encode question, encode passage, predict entity

- Can also use textual entailment-like models

Multiclass classification problem over entities in the document

Hermann et al. (2015), Chen et al. (2016)
Attentive reader:
\[ u = \text{encode query} \]
\[ s = \text{encode sentence} \]
\[ r = \text{attention}(u \rightarrow s) \]
\[ \text{prediction} = f(\text{candidate, } u, r) \]

- Uses fixed-size representations for the final prediction, multiclass classification.
Chen et al (2016): small changes to the attentive reader

Additional analysis of the task found that many of the remaining questions were unanswerable or extremely difficult

Stanford Attentive Reader  76.2  76.5  79.5  78.7

Hermann et al. (2015), Chen et al. (2016)
SQuAD
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

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”.

**What causes precipitation to fall?**
**gravity**

**What is another main form of precipitation besides drizzle, rain, snow, sleet and hail?**
**graupel**

**Where do water droplets collide with ice crystals to form precipitation?**
**within a cloud**

Rajpurkar et al. (2016)
What was Marie Curie the first female recipient of?

Like a tagging problem over the sentence (not multiclass classification), but we need some way of attending to the query.
Bidirectional Attention Flow (Offline Reading)

- Passage (context) and query are both encoded with BiLSTMs
- Context-to-query attention: compute softmax over columns of $S$, take weighted sum of $u$ based on attention weights for each passage word

$$S_{ij} = h_i \cdot u_j$$

$$\alpha_{ij} = \text{softmax}_j(S_{ij})$$

$$\tilde{u}_i = \sum_j \alpha_{ij} u_j$$

- query “specialized” to the $i$th word
- dist over query

Seo et al. (2016)
Each passage word now “knows about” the query.
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 in passage
- No need for cross-attention mechanisms!
### SQuAD SOTA: Fall 18

<table>
<thead>
<tr>
<th>Rank</th>
<th>Model</th>
<th>EM</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Human Performance</td>
<td>82.304</td>
<td>91.221</td>
</tr>
<tr>
<td></td>
<td>Stanford University</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(Rajpurkar et al. ’16)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>BERT (ensemble)</td>
<td>87.433</td>
<td>93.160</td>
</tr>
<tr>
<td></td>
<td><em>Google AI Language</em></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>BERT (single model)</td>
<td>85.083</td>
<td>91.835</td>
</tr>
<tr>
<td></td>
<td><em>Google AI Language</em></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>nlnet (ensemble)</td>
<td>85.356</td>
<td>91.202</td>
</tr>
<tr>
<td></td>
<td><em>Microsoft Research Asia</em></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>nlnet (ensemble)</td>
<td>85.954</td>
<td>91.677</td>
</tr>
<tr>
<td></td>
<td><em>Microsoft Research Asia</em></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>QANet (ensemble)</td>
<td>84.454</td>
<td>90.490</td>
</tr>
<tr>
<td></td>
<td><em>Google Brain &amp; CMU</em></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>r-net (ensemble)</td>
<td>84.003</td>
<td>90.147</td>
</tr>
<tr>
<td></td>
<td><em>Microsoft Research Asia</em></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>QANet (ensemble)</td>
<td>83.877</td>
<td>89.737</td>
</tr>
<tr>
<td></td>
<td><em>Google Brain &amp; CMU</em></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- BiDAF: 73 EM / 81 F1
- nlnet, QANet, r-net — dueling super complex systems (much more than BiDAF...)
SQuAD SOTA: Spring 19

- SQuAD 2.0: harder dataset because some questions are unanswerable
- Industry contest

<table>
<thead>
<tr>
<th>Rank</th>
<th>Model</th>
<th>EM</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>BERT + DAE + AoA (ensemble)</td>
<td>87.147</td>
<td>89.474</td>
</tr>
<tr>
<td>2</td>
<td>BERT + ConvLSTM + MTL + Verifier (ensemble)</td>
<td>86.730</td>
<td>89.286</td>
</tr>
<tr>
<td>3</td>
<td>BERT + N-Gram Masking + Synthetic Self-Training (ensemble)</td>
<td>86.673</td>
<td>89.147</td>
</tr>
<tr>
<td>4</td>
<td>SemBERT(ensemble)</td>
<td>86.166</td>
<td>88.886</td>
</tr>
<tr>
<td>5</td>
<td>BERT + DAE + AoA (single model)</td>
<td>85.884</td>
<td>88.621</td>
</tr>
<tr>
<td>6</td>
<td>BERT + N-Gram Masking + Synthetic Self-Training (single model)</td>
<td>85.150</td>
<td>87.715</td>
</tr>
<tr>
<td>7</td>
<td>BERT + MMFT + ADA (ensemble)</td>
<td>85.082</td>
<td>87.615</td>
</tr>
</tbody>
</table>
SQuAD SOTA: more recently

- Performance is very saturated
- Harder QA settings are needed!

<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>89.731</td>
<td>92.215</td>
</tr>
<tr>
<td></td>
<td>Google Research &amp; TTIC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>XLNet + DAAF + Verifier (ensemble)</td>
<td>88.592</td>
<td>90.859</td>
</tr>
<tr>
<td></td>
<td>PINGAN Omni-Sinitic</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></td>
<td>Google Research &amp; TTIC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>UPM (ensemble)</td>
<td>88.231</td>
<td>90.713</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></td>
<td>Shanghai Jiao Tong University &amp; CloudWalk</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>XLNet + SG-Net Verifier++ (single model)</td>
<td>87.238</td>
<td>90.071</td>
</tr>
<tr>
<td></td>
<td>Shanghai Jiao Tong University &amp; CloudWalk</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

[https://rajpurkar.github.io/SQuAD-explorer/](https://rajpurkar.github.io/SQuAD-explorer/)
Takeaways

- Many flavors of reading comprehension tasks: cloze or actual questions, single or multi-sentence

- Complex attention schemes can match queries against input texts and identify answers
Question Answering 2
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
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: SQuAD SOTA

<table>
<thead>
<tr>
<th>Rank</th>
<th>Model Description</th>
<th>EM</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Human Performance</td>
<td>86.831</td>
<td>89.452</td>
</tr>
<tr>
<td></td>
<td>Stanford University</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(Rajpurkar &amp; Jia et al. ‘18)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>ALBERT (ensemble model)</td>
<td>89.731</td>
<td>92.215</td>
</tr>
<tr>
<td></td>
<td>Google Research &amp; TTIC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>XLNet + DAAF + Verifier (ensemble)</td>
<td>88.592</td>
<td>90.859</td>
</tr>
<tr>
<td></td>
<td>PINGAN Omni-Sinitic</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></td>
<td>Google Research &amp; TTIC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>UPM (ensemble)</td>
<td>88.231</td>
<td>90.713</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></td>
<td>Shanghai Jiao Tong University &amp; CloudWalk</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>XLNet + SG-Net Verifier++ (single model)</td>
<td>87.238</td>
<td>90.071</td>
</tr>
<tr>
<td></td>
<td>Shanghai Jiao Tong University &amp; CloudWalk</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Performance is very saturated
- Harder QA settings are needed
This Part

- 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...”

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

Jia and Liang (2017)
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)
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
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
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
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
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...
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
  - Find 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</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>plain</td>
</tr>
<tr>
<td>SQuAD</td>
<td>62.7</td>
<td>76.1</td>
</tr>
<tr>
<td>CuratedTREC</td>
<td>81.0</td>
<td>85.2</td>
</tr>
<tr>
<td>WebQuestions</td>
<td>73.7</td>
<td>75.5</td>
</tr>
<tr>
<td>WikiMovies</td>
<td>61.7</td>
<td>54.4</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)
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

\[ h_q = W_q BERT_Q(q)[CLS] \]
\[ h_b = W_b BERT_B(b)[CLS] \]
\[ S_{retr}(b, q) = h_q^T h_b \]
Many SQuAD questions are not suited to the “open” setting because they’re underspecified.

Where did the Super Bowl take place?

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)
NaturalQuestions

- 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)

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.
Multi-Hop Question Answering
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.

Welbl et al. (2018), Yang et al. (2018)
WikiHop

- 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]
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]
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.

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]
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*?

**Answer:**
- 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.

**Keywords:** 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)
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

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)
Sentence Factored Model

Answer prediction:

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

Chen and Durrett (2019)
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.
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

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

- **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 ...

- 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**

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 (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 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.

**Question**

How many more dollars was the *Untitled (1981)* painting sold for than the 12 million dollar estimation?

**Answer**

4300000

**BiDAF**

$16.3$ million

- 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?

- 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

<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>-</td>
<td>60.4</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>

...
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

- 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