**CS388: Natural Language Processing**

**Lecture 22:**

**Question Answering 1**

Greg Durrett

---

**Recall: Erasure Methods**

- This secondary classifier’s **weights** now give us **highlights** on the input.

<table>
<thead>
<tr>
<th>Movie Description</th>
<th>Negative</th>
<th>Positive</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mediocre, maybe even bad.</td>
<td>99.8%</td>
<td>63.4%</td>
</tr>
<tr>
<td>Mediocre, maybe even bad.</td>
<td>98.0%</td>
<td>74.5%</td>
</tr>
<tr>
<td>Mediocre, maybe even bad.</td>
<td>98.7%</td>
<td>74.5%</td>
</tr>
<tr>
<td>Mediocre, maybe even bad.</td>
<td>97.9%</td>
<td>74.5%</td>
</tr>
</tbody>
</table>

- The movie is mediocre, maybe even bad.

---

**Recall: Gradient-based Methods**

- Originally used for images

\[
S_c = \text{score of class } c \\
I_0 = \text{current image} \\
\frac{\partial S_c}{\partial I} \bigg|_{I_0}
\]

- Higher gradient magnitude = small change in pixels leads to large change in prediction

- For words: “pixels” are coordinates of each word’s vector, sum these up to get the importance of that word

---

Administrivia

- Jason Baldridge guest lecture next Thursday
- P2 back soon

---

Wallace, Gardner, Singh
Interpretability Tutorial at EMNLP 2020

Simonyan et al. (2013)
This Lecture

- Types of question answering/reading comprehension
- Span-based question answering on SQuAD
- SQuAD results

Reading Comprehension

Classical 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.)

- How to deal with open-domain data/relations? Need data to learn how to ground every predicate or need to be able to produce predicates in a zero-shot way

QA from Open IE

(a) CCG parse builds an underspecified semantic representation of the sentence.

<table>
<thead>
<tr>
<th>Former</th>
<th>municipalities</th>
<th>in</th>
<th>Brandenburg</th>
</tr>
</thead>
<tbody>
<tr>
<td>λf(z(x)) \land \text{former}(x)</td>
<td>λz.municipalities(x)</td>
<td>N \land \text{in}(y, x) \land \text{in}(y, \text{Brandenburg})</td>
<td>N \land \text{in}(y, \text{Brandenburg})</td>
</tr>
<tr>
<td>N \land \text{in}(y, x) \land \text{in}(y, \text{Brandenburg})</td>
<td></td>
<td>N \land \text{in}(y, \text{Brandenburg})</td>
<td></td>
</tr>
</tbody>
</table>

(b) Constant matches replace underspecified constants with Freebase concepts

\[
\begin{align*}
  b_0 &= \lambda z. \text{former}(z) \land \text{municipalities}(z) \land \text{in}(z, \text{Brandenburg}) \\
  b_1 &= \lambda z. \text{former}(z) \land \text{municipalities}(z) \land \text{in}(z, \text{Brandenburg}) \\
  b_2 &= \lambda z. \text{former}(z) \land \text{municipalities}(z) \land \text{location.containedBy}(z, \text{Brandenburg}) \\
  b_3 &= \lambda z. \text{OpenRel}(z, \text{Municipality}) \land \text{location.containedBy}(z, \text{Brandenburg}) \\
  b_4 &= \lambda z. \text{OpenType}(z) \land \text{OpenRel}(z, \text{Municipality}) \land \text{location.containedBy}(z, \text{Brandenburg})
\end{align*}
\]

- Why use the KB at all? Why not answer questions directly from text? Choi et al. (2015)
QA is very broad

- Factoid QA: what states border Mississippi?, when was Barack Obama born?
- Lots of this could be handled by QA from a knowledge base, if we had a big enough knowledge base
- “Question answering” as a term is so broad as to be meaningless
  - What is the meaning of life?
  - What is 4+5?
  - What is the translation of [sentence] into French? [McCann et al., 2018]

What are the limits of QA?

- Focus on questions where the answer might appear in text — still hard!
  - 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
- How long should I soak dry pinto beans? — could be written down in a KB but probably isn’t
- Today: QA when it requires retrieving the answer from a passage

Reading Comprehension

- “AI challenge problem”: answer question given context
- Recognizing Textual Entailment (2006)
- MCTest (2013): 500 passages, 4 questions per passage
- Two 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

```
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

What did James pull off of the shelves in the grocery store? Pudding
```

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

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)

Reading Comprehension

<table>
<thead>
<tr>
<th>ngram sliding window</th>
<th>MC160 Test</th>
<th>MC500 Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline (SW+D)</td>
<td>66.25</td>
<td>56.67</td>
</tr>
<tr>
<td>RTE</td>
<td>59.79^</td>
<td>53.52</td>
</tr>
<tr>
<td>Combined</td>
<td>67.60</td>
<td>60.83^</td>
</tr>
</tbody>
</table>

- 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)

Richardson (2013)

Better Systems

- Match an AMR (abstract meaning representation) of the question against the original text
- 70% accuracy (roughly 10% better than baseline)

Sachan and Xing (2016)

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
- “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
**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? Commonsense knowledge?

**Children’s Book Test**

- 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)
  - Hill et al. (2015)

**LAMBADA**

- *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

**Multiple-Choice**

- SWAG dataset was constructed to be difficult for ELMo
- BERT subsequently got 20% accuracy improvements and achieved human-level performance
- Problem: distractors too easy

- Let’s focus on architectures for retrieval from a passage

**Paperno et al. (2016)**

**Zellers et al. (2018)**
Span-based Question Answering

SQuAD

- Single-document question-answering task where the answer is always a substring of the passage (= a paragraph from Wikipedia)
- Predict start and end indices of the answer in the passage

Rajpurkar et al. (2016)

SQuAD

What was Marie Curie the first female recipient of?

START   END

- Like a tagging problem over the sentence (not multiclass classification), but we need some way of attending to the query

Rajpurkar et al. (2016)

Architectures

- Predict a distributions over start and end points of the answer

P(end | q, p) computed similarly

P(start = i | q, p) = softmax(p_T W q)

Who was the first female recipient of the Nobel Prize?

BiLSTM encoder

Marie Curie was the first...
Training and Inference

› Train on labeled data with start and end points, maximize likelihood of correct decisions:
\[
\log \sum_{i \in \text{gold starts}} p(\text{start} = i | p, q) + \log \sum_{i \in \text{gold ends}} p(\text{end} = i | p, q)
\]

What was the name of the first successful credit card?

› Inference: maximize \(P(\text{start}) + P(\text{end})\) with the constraint that (start, end) isn’t too big a span

In September 1958, Bank of America launched a new product called BankAmerican in Fresno. After a troubled gestation during which its creator resigned, BankAmerican went on to become the first successful credit card — that is, a financial instrument that was usable across a large number of merchants and also allowed cardholders to revolve a balance (earlier financial products could do one or the other but not both). In 1976, BankAmerican was renamed and spun off into a separate company known today as Visa Inc.

What do these models do?

Question: how many victorians are non-religious?
Answer: 20%

about 61.8% of victorians describe themselves as christian. roman catholics form the single largest religious group in the state with 26.7% of the victorian population, followed by anglicans and members of the uniting church. buddhism is the state’s largest non-christian religion, with 180,377 members as of the most recent census. victoria is also home of 152,775 muslims and 45,150 jews. hinduism is the fastest growing religion. around 20% of victorians claim no religion. amongst those who declare a religious affiliation, church attendance is low.

Why did this take off?

› SQuAD was big: >100,000 questions at a time when deep learning was exploding

› SQuAD was pretty easy: year-over-year progress for a few years until the dataset was essentially solved

› SQuAD had room to improve: “50% performance from a logistic regression baseline (classifier with 180M features over constituents)
Bidirectional Attention Flow

- 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.

\[
\tilde{u}_i = \sum_j \alpha_{ij} u_j
\]

query “specialized” to the \( i \)th word

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

dist over query

\[
S_{ij} = h_i \cdot u_j
\]

query \( U \)

passage \( H \)

Seo et al. (2016)

Bidirectional Attention Flow

QA with BERT

- Predict start and end positions in passage
- No need for cross-attention mechanisms!

Devlin et al. (2019)

QA with BERT

- How does this work?

What was Marie Curie the first female recipient of? [SEP] Marie Curie was the first female recipient of ...

Devlin et al. (2019)
SQuAD Results

SQuAD SOTA: Fall 18

- 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

SQuAD SOTA: Fall 19

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

- Totally figuring this out is very challenging
- Coref: the failed campaign movie of the same name
- Lots of surface clues: 1961, campaign, etc.
- Systems can do well without really understanding the text

**Question:** The Dodecanese [Campaign of WWII] that was an attempt by the Allied forces to capture islands in the Aegean Sea was the inspiration for which acclaimed 1961 commando film?

**Answer:** The Guns of Navarone

**Excerpt:** The Dodecanese Campaign of World War II was an attempt by Allied forces to capture the Italian-held Dodecanese islands in the Aegean Sea following the surrender of Italy in September 1943, and use them as bases against the German-controlled Balkans. The [failed campaign] and in particular the Battle of Leros, inspired the 1957 novel *The Guns of Navarone* and the successful [1961 movie of the same name].

Joshi et al. (2017)

---

**What are these models learning?**

- “Who...”: knows to look for people
- “Which film...”: can identify movies and then spot keywords that are related to the question
- Unless questions are made super tricky (target closely-related entities who are easily confused), they’re usually not so hard to answer

De Cao et al. (2020)

---

**What are these models learning?**

- Are these good explanations?

Sundararajan et al., 2017

---

**What are these models learning?**

Pairwise explanation method: explains predictions in terms of associations between words

Ye, Nair, Durrett (2021)
What are these models learning?

Ye, Nair, Durrett (2021)

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

Takeaways

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

‣ Memory networks let you reference input in an attention-like way, useful for generalizing language models to long-range reasoning

‣ Complex attention schemes can match queries against input texts and identify answers

‣ Next time: more complex datasets / QA settings