COS 484: Natural Language Processing

Question Answering

Fall 2019
Announcements

- Final project presentation: January 13, 10am-12pm
- Revised project proposal: November 22
  - Come meet us during OHs!
## Course planning

<table>
<thead>
<tr>
<th>Date</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nov 19</td>
<td>Question answering</td>
</tr>
<tr>
<td>Nov 21</td>
<td>Coreference resolution</td>
</tr>
<tr>
<td>Nov 26</td>
<td>Guest lecture: Jason Weston</td>
</tr>
<tr>
<td>Nov 28</td>
<td>NO CLASS (Thanksgiving)</td>
</tr>
<tr>
<td>Dec 3</td>
<td>Guest lecture: Tom Kwiatkowski</td>
</tr>
<tr>
<td>Dec 5</td>
<td>Information extraction</td>
</tr>
<tr>
<td>Dec 10</td>
<td>Grounding</td>
</tr>
<tr>
<td>Dec 12</td>
<td>Project advice</td>
</tr>
</tbody>
</table>

- **Transformers + Question Answering**
- **Dialogue**
- **Advanced topics in QA and others**
RNNs vs Transformers

The movie was terribly exciting!
Transformers

Key concepts:
- (scaled) dot-product attention
- Self-attention
- Multi-head self-attention

(Vaswani et al, 2017): Attention is all you need
Recap: seq2seq with attention

\[ e_i^t = g(h_i^{enc}, h_t^{dec}) \]

\[ \alpha_i^t = \frac{\exp(e_i^t)}{\sum_{j=1}^n \exp(e_j^t)} \]

\[ a^t = \sum_{i=1}^n \alpha_i^t h_i^{enc} \]
Generalized Attention

- A query \( q \) and a set of key-value \((k_i, v_i)\) pairs to an output

- Dot-product attention:
  \[
  A(q, \{k_i, v_i\}) = \sum_i \frac{e^{q \cdot k_i}}{\sum_j e^{q \cdot k_j}} v_i
  \]
  \( k_i, v_i, q \in \mathbb{R}^d \)

- If we have multiple queries:
  \[
  A(Q, K, V) = \text{softmax}(QK^\top)V
  \]
  \( Q \in \mathbb{R}^{n_Q \times d}, K, V \in \mathbb{R}^{n \times d} \)

- Scaled dot-product attention:
  \[
  A(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d}}\right)V
  \]
Self-attention

- Input: $x_1, x_2, \ldots, x_n \in \mathbb{R}^{d_{in}}$
- Output: $h_1, h_2, \ldots, h_n \in \mathbb{R}^d$
- Key idea: let’s use each word as query and compute the attention with all the other words
- Input: $X \in \mathbb{R}^{n \times d_{in}}$

\[
A(XW^Q, XW^K, XW^V) \in \mathbb{R}^{n \times d}
\]

$W^Q, W^K, W^V \in \mathbb{R}^{d_{in} \times d}$
Multi-head self-attention

One head is not expressive enough. Let’s have multiple heads!

$$A(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O$$

$$\text{head}_i = A(XW^Q_i, XW^K_i, XW^V_i)$$

In practice, $h = 8$, $d = d_{out}/h$, $W^O \in \mathbb{R}^{d_{out} \times d_{out}}$

https://github.com/jessevig/bertviz
Putting it all together

- Each Transformer block has two sub-layers
  - Multi-head attention
  - 2-layer feedforward NN (with ReLU)
- Each sublayer has a residual connection and a layer normalization
  \[ \text{LayerNorm}(x + \text{SubLayer}(x)) \]
- Input layer has a positional encoding

- BERT\_base: 12 layers, 12 heads, hidden size = 768, 110M parameters
- BERT\_large: 24 layers, 16 heads, hidden size = 1024, 340M parameters

(Ba et al, 2016): Layer Normalization
Encoder-decoder architecture

(Vaswani et al, 2017): Attention is all you need
Question Answering

- Goal: build computer systems to answer questions

<table>
<thead>
<tr>
<th>Question</th>
<th>Answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>When were the first pyramids built?</td>
<td>2630 BC</td>
</tr>
<tr>
<td>What’s the weather like in Princeton?</td>
<td>42 F</td>
</tr>
<tr>
<td>Where is Einstein’s house?</td>
<td>112 Mercer St, Princeton, NJ 08540</td>
</tr>
<tr>
<td>Why do we yawn?</td>
<td>When we’re bored or tired we don’t breathe as deeply as we normally do. This causes a drop in our blood-oxygen levels and yawning helps us counter-balance that.</td>
</tr>
</tbody>
</table>
Question Answering

- You can easily find these answers in Google today!

2630 BC
Most were built as tombs for the country's pharaohs and their consorts during the Old and Middle Kingdom periods. The earliest known Egyptian pyramids are found at Saqqara, northwest of Memphis. The earliest among these is the Pyramid of Djoser (constructed 2630 BC–2611 BC) which was built during the third dynasty.

The First Pyramids Built - Timeline Index
www.timelineindex.com/content/view

where is einstein's house?

112 Mercer St, Princeton, NJ 08540
Question Answering

- People ask lots of questions to Digital Personal Assistants:

<table>
<thead>
<tr>
<th>Smart Speaker Use Case Frequency January 2018</th>
<th>Use daily</th>
<th>Use monthly</th>
<th>Have ever tried</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ask a question</td>
<td>91%</td>
<td>72.9%</td>
<td>33.3%</td>
</tr>
<tr>
<td>Listen to streaming music service</td>
<td>69.5%</td>
<td>78.2%</td>
<td>41.9%</td>
</tr>
<tr>
<td>Check the weather</td>
<td>85.2%</td>
<td>69.1%</td>
<td>41.4%</td>
</tr>
<tr>
<td>Set a timer</td>
<td>71.4%</td>
<td>51.8%</td>
<td>24.1%</td>
</tr>
<tr>
<td>Listen to radio</td>
<td>68.8%</td>
<td>47.6%</td>
<td>25.5%</td>
</tr>
<tr>
<td>Set an alarm</td>
<td>65.7%</td>
<td>48%</td>
<td>25.3%</td>
</tr>
<tr>
<td>Listen to news / sports</td>
<td>58.1%</td>
<td>39.4%</td>
<td>14.8%</td>
</tr>
<tr>
<td>Play game or answer trivia</td>
<td>52.3%</td>
<td>31.2%</td>
<td>11.1%</td>
</tr>
<tr>
<td>Find a recipe or cooking instructions</td>
<td>40.5%</td>
<td>25.5%</td>
<td>6.1%</td>
</tr>
<tr>
<td>Use a favorite skill or assistant app</td>
<td>46.5%</td>
<td>28.8%</td>
<td>14.7%</td>
</tr>
<tr>
<td>Check traffic</td>
<td>41.2%</td>
<td>25.8%</td>
<td>7.7%</td>
</tr>
<tr>
<td>Call someone</td>
<td>40.7%</td>
<td>22.7%</td>
<td>10.3%</td>
</tr>
<tr>
<td>Listen to podcasts and other talk formats</td>
<td>40.7%</td>
<td>24.1%</td>
<td>10.1%</td>
</tr>
<tr>
<td>Control smart home devices</td>
<td>38.1%</td>
<td>29.9%</td>
<td>20.8%</td>
</tr>
<tr>
<td>Access my calendar</td>
<td>35.1%</td>
<td>10.6%</td>
<td>8.2%</td>
</tr>
<tr>
<td>Message someone</td>
<td>34.2%</td>
<td>17.9%</td>
<td>8.2%</td>
</tr>
<tr>
<td>Made a purchase</td>
<td>26%</td>
<td>11.5%</td>
<td>2.1%</td>
</tr>
</tbody>
</table>

Source: Voicebot Smart Speaker Consumer Adoption Report January 2018
Question Answering

IBM Watson defeated two of Jeopardy's greatest champions in 2011
Why care about question answering?

- Lots of immediate applications: search engines, dialogue systems
- Question answering is an important testbed for evaluating how well compute systems understand human language

"Since questions can be devised to query **any aspect** of text comprehension, the ability to answer questions is the **strongest possible demonstration of understanding.**"
QA Taxonomy

- Factoid questions vs non-factoid questions
- Answers
  - A short span of text
  - A paragraph
  - Yes/No
  - A database entry
  - A list
- Context
  - A passage, a document, a large collection of documents
  - Knowledge base
  - Semi-structured tables
  - Images
The first recorded travels by Europeans to China and back date from this time. The most famous traveler of the period was the Venetian Marco Polo, whose account of his trip to "Cambaluc," the capital of the Great Khan, and of life there astounded the people of Europe. The account of his travels, Il milione (or, The Million, known in English as the Travels of Marco Polo), appeared about the year 1299. Some argue over the accuracy of Marco Polo's accounts due to the lack of mentioning the Great Wall of China, tea houses, which would have been a prominent sight since Europeans had yet to adopt a tea culture, as well the practice of foot binding by the women in capital of the Great Khan. Some suggest that Marco Polo acquired much of his knowledge through contact with Persian traders since many of the places he named were in Persian.

How did some suspect that Polo learned about China instead of by actually visiting it?

**Answer:** through contact with Persian traders
James the Turtle was always getting in trouble. Sometimes he'd reach into the freezer and empty out all the food. Other times he'd sled on the deck and get a splinter. His aunt Jane tried as hard as she could to keep him out of trouble, but he was sneaky and got into lots of trouble behind her back.

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.

His aunt was waiting for him in his room. She told James that she loved him, but he would have to start acting like a well-behaved turtle.

After about a month, and after getting into lots of trouble, James finally made up his mind to be a better turtle.

1) What is the name of the trouble making turtle?
A) Fries  
B) Pudding  
C) James  
D) Jane

2) What did James pull off of the shelves in the grocery store?
A) pudding  
B) fries  
C) food  
D) splinters
The Virginia governor’s race, billed as the marquee battle of an otherwise anticlimactic 2013 election cycle, is shaping up to be a foregone conclusion. Democrat Terry McAuliffe, the longtime political fixer and moneyman, hasn’t trailed in a poll since May. Barring a political miracle, Republican Ken Cuccinelli will be delivering a concession speech on Tuesday evening in Richmond. In recent ...

Q: What are the candidates running for? A: Governor
Q: Where? A: Virginia
Q: Who is the democratic candidate? A: Terry McAuliffe
Q: Who is his opponent? A: Ken Cuccinelli
Q: What party does he belong to? A: Republican
Q: Which of them is winning?
Long-form Question Answering

 questões: How do jellyfish function without a brain and a nervous system?

Abstractive

Jellyfish don't have brains. Jellyfish have a nervous system. They know how to swim, they don't even know what to do with their bodies. They have a “nervous system”, which is basically a bunch of nerve cells that connect to the body. The nervous system is made up of neurons, neurons, nerves, etc. Each neuron doesn't have a specific set of "neurons", it just has a different set of nerves. It has a set of sensory neurons, which are all connected to each other. It just knows that it does react to the stimulus, and it knows that the stimulus is there. They have a bunch of nervous systems that act like a filter to get information back.

Extractive

But none of these hypotheses make much sense for a jellyfish because they don't have brains at all. They just have a nerve net — a loose ring of neurons that runs around the rim of their pulsating bells. They have an unusual nervous system, 451 because jellyfish are not bilaterally symmetrical — that is, they don't have a left side and a right side. Jellyfish don't have brains, but their nervous systems detect smells, light and other stimuli, and they coordinate their physical responses.

https://ai.facebook.com/blog/longform-qa/
(Fan et al, 2019): ELI5: Long Form Question Answering
Open-domain Question Answering

Q: How many of Warsaw's inhabitants spoke Polish in 1933?

DrQA

Wikipedia
The Free Encyclopedia

Document Retriever

Document Reader

833,500

>>> process('What is the answer to life, the universe, and everything?')

Top Predictions:

<table>
<thead>
<tr>
<th>Rank</th>
<th>Answer</th>
<th>Doc</th>
<th>Answer Score</th>
<th>Doc Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>42</td>
<td>Phrases from The Hitchhiker's Guide to the Galaxy</td>
<td>47242</td>
<td>141.26</td>
</tr>
</tbody>
</table>
Knowledge Base Question Answering

100M entities (nodes)  1B assertions (edges)

Which states’ capitals are also their largest cities by area?

\[
u \cdot \text{Type.USState} \cap \text{Capital}.\text{argmax} (\text{Type.City} \cap \text{ContainedBy}.u, \text{Area})
\]

Arizona, Hawaii, Idaho, Indiana, Iowa, Oklahoma, Utah

(Berant et al, 2013): Semantic Parsing on Freebase from Question-Answer Pairs
Table-based Question Answering


<table>
<thead>
<tr>
<th>Year</th>
<th>City</th>
<th>Country</th>
<th>Nations</th>
</tr>
</thead>
<tbody>
<tr>
<td>1896</td>
<td>Athens</td>
<td>Greece</td>
<td>14</td>
</tr>
<tr>
<td>1900</td>
<td>Paris</td>
<td>France</td>
<td>24</td>
</tr>
<tr>
<td>1904</td>
<td>St. Louis</td>
<td>USA</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>2004</td>
<td>Athens</td>
<td>Greece</td>
<td>201</td>
</tr>
<tr>
<td>2008</td>
<td>Beijing</td>
<td>China</td>
<td>204</td>
</tr>
<tr>
<td>2012</td>
<td>London</td>
<td>UK</td>
<td>204</td>
</tr>
</tbody>
</table>

$x = \text{Greece held its last Summer Olympics in which year?}$

$y = 2004$
Visual Question Answering

What color are her eyes?
What is the mustache made of?

How many slices of pizza are there?
Is this a vegetarian pizza?

(Antol et al, 2015): Visual Question Answering
Stanford Question Answering Dataset (SQuAD)

Passage

Super Bowl 50 was an American football game to determine the champion of the National Football League (NFL) for the 2015 season. The American Football Conference (AFC) champion Denver Broncos defeated the National Football Conference (NFC) champion Carolina Panthers 24–10 to earn their third Super Bowl title. The game was played on February 7, 2016, at Levi's Stadium in the San Francisco Bay Area at Santa Clara, California.

Question: Which NFL team won Super Bowl 50?
Answer: Denver Broncos

Question: What does AFC stand for?
Answer: American Football Conference

Question: What year was Super Bowl 50?
Answer: 2016

- (passage, question, answer) triples
- Passage is from Wikipedia, question is crowd-sourced
- Answer must be a span of text in the passage (aka. “extractive question answering”)
- SQuAD 1.1: 100k answerable questions, SQuAD 2.0: another 50k unanswerable questions

https://stanford-qa.com
(Rajpurkar et al, 2016): SQuAD: 100,000+ Questions for Machine Comprehension of Text
Stanford Question Answering Dataset (SQuAD)

SQuAD 1.1 evaluation:

- 3 gold answers are collected for each answer
- Two metrics: exact match (EM) and F1
- Exact match: 1/0 accuracy on whether you match one of the three answers
- F1: take each gold answer and system output as bag of words, compute precision, recall and harmonic mean. Take the max of the three scores.

Q: Rather than taxation, what are private schools largely funded by?

A: {tuition, charging their students tuition, tuition}

(Rajpurkar et al, 2016): SQuAD: 100,000+ Questions for Machine Comprehension of Text
Feature-based models

- Generate a list of candidate answers \( \{a_1, a_2, \ldots, a_M\} \)
  - Considered only the constituents in parse trees

- Define a feature vector \( \phi(p, q, a_i) \in \mathbb{R}^d \):
  - Word/bigram frequencies
  - Parse tree matches
  - Dependency labels, length, part-of-speech tags

- Apply a (multi-class) logistic regression model

(Rajpurkar et al, 2016): SQuAD: 100,000+ Questions for Machine Comprehension of Text
BiLSTM-based models

BiDAF

(Seo et al, 2017): Bidirectional Attention Flow for Machine Comprehension
BiLSTM-based models

- Encode the question using word/character embeddings; pass to an biLSTM encoder
- Encode the passage similarly
- Passage-to-question and question-to-passage attention
- Modeling layer: another BiLSTM layer
- Output layer: two classifiers for predicting start and end points
- The entire model can be trained in an end-to-end way

(Seo et al, 2017): Bidirectional Attention Flow for Machine Comprehension
BERT-based models

Use the output of the masked word's position to predict the masked word

Possible classes:
- All English words
- Improvisation (10%)
- Zyzzyva (0%) ...

Randomly mask 15% of tokens

Input:
[CLS] Let's stick to improvisation in this skit

Pre-training
BERT-based models

- Concatenate question and passage as one single sequence separated with a [SEP] token, then pass it to the BERT encoder
- Train two classifiers on top of the passage tokens
Experiments on SQuAD v1.1

<table>
<thead>
<tr>
<th>Model</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic Regression</td>
<td>51.0</td>
</tr>
<tr>
<td>BiDAF++</td>
<td>81.1</td>
</tr>
<tr>
<td>+ Human Performance</td>
<td>90.9</td>
</tr>
<tr>
<td>state-of-the-art (as of Nov 2019)</td>
<td>95.1</td>
</tr>
</tbody>
</table>

*: single model only
Is Reading Comprehension solved?

AI systems are beating humans in reading comprehension

By Associated Press January 24, 2018 2:25pm

Nope, maybe the SQuAD dataset is solved.
Is Reading Comprehension solved?

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

(Jia et al, 2017): Adversarial Examples for Evaluating Reading Comprehension Systems
SQuAD Limitations

- SQuAD has a number of limitations:
  - Only span-based answers (no yes/no, counting, implicit why)
  - Questions were constructed looking at passages
    - Not genuine information needs
    - Generally greater lexical and syntactic matching between question and answer span
  - Barely any multi-fact/sentence inference beyond coreference

- Nevertheless, it is a well-targeted, well-structured, clean dataset
  - The most used and competed QA dataset
  - A useful starting point for building systems in industry (although in-domain data always really helps!)
DrQA Demo
https://github.com/facebookresearch/DrQA

Hello! Please ask a question.

What is question answering?

A computer science discipline within the fields of information retrieval and natural language processing.

Who was the winning pitcher in the 1956 World Series?

Don Larsen

What is the answer to life, the universe, and everything?

42