Lecture 21: Question Answering 1
Recall: SRL

- Identify predicate, disambiguate it, identify that predicate’s arguments
- Verb roles from Propbank (Palmer et al., 2005)

**Gold**

```
Housing starts are expected to quicken a bit from August's pace
```

**quicken:**

- **Arg0-PAG:** *causer of speed-up*
- **Arg1-PPT:** *thing becoming faster* (vnrole: 45.4-patient)
- **Arg2-EXT:** *EXT*
- **Arg3-DIR:** *old speed*
- **Arg4-PRD:** *new speed*

Figure from He et al. (2017)
Recall: SRL for QA

- Question and several answer candidates

Q: *Who discovered prions?*

AC1: *In 1997, Stanley B. Prusiner, a scientist in the United States, discovered prions...*

AC2: *Prions were researched by...*

Shen and Lapata (2007)
This Lecture

- Types of question answering/reading comprehension
- Memory networks
- CNN/Daily Mail task: Attentive Reader
- SQuAD task: Bidirectional Attention Flow
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
Why use the KB at all? Why not answer questions directly from text? Like information retrieval!

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
  - *Is P=NP?*
  - *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 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?
“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
- 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).

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)
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?
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)
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 interestedly 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
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 look at architectures for retrieval from a passage

Zellers et al. (2018)
Memory Networks
Memory Networks

- Memory networks let you reference input with attention
- Encode input items into two vectors: a key and a value
- Keys compute attention weights given a query, weighted sum of values gives the output

\[
q = \alpha_i v_i \\
\alpha = \text{softmax}(e) \\
e_i = q \cdot k_i
\]

Sukhbaatar et al. (2015)
Memory Networks

- Three layers of memory network where the query representation is updated additively based on the memories at each step.

- How to encode the sentences?
  - Bag of words (average embeddings)
  - Positional encoding: multiply each word by a vector capturing position in sentence

Sukhbaatar et al. (2015)
Evaluation on 20 tasks proposed as building blocks for building “AI-complete” systems

- Various levels of difficulty, exhibit different linguistic phenomena
- Small vocabulary, language isn’t truly “natural”

<table>
<thead>
<tr>
<th>Task 1: Single Supporting Fact</th>
<th>Task 2: Two Supporting Facts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mary went to the bathroom.</td>
<td>John is in the playground.</td>
</tr>
<tr>
<td>John moved to the hallway.</td>
<td>John picked up the football.</td>
</tr>
<tr>
<td>Mary travelled to the office.</td>
<td>Bob went to the kitchen.</td>
</tr>
<tr>
<td>Where is Mary? A: office</td>
<td>Where is the football? A: playground</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Task 13: Compound Coreference</th>
<th>Task 14: Time Reasoning</th>
</tr>
</thead>
<tbody>
<tr>
<td>Daniel and Sandra journeyed to the office.</td>
<td>In the afternoon Julie went to the park.</td>
</tr>
<tr>
<td>Then they went to the garden.</td>
<td>Yesterday Julie was at school.</td>
</tr>
<tr>
<td>Sandra and John travelled to the kitchen.</td>
<td>Julie went to the cinema this evening.</td>
</tr>
<tr>
<td>After that they moved to the hallway.</td>
<td>Where did Julie go after the park? A: cinema</td>
</tr>
<tr>
<td>Where is Daniel? A: garden</td>
<td>Where was Julie before the park? A: school</td>
</tr>
</tbody>
</table>

Weston et al. (2014)
### Evaluation: bAbI

<table>
<thead>
<tr>
<th>Task</th>
<th>Baseline</th>
<th>MemN2N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean error (%)</td>
<td>6.7</td>
<td>25.8</td>
</tr>
<tr>
<td>Failed tasks (err. &gt; 5%)</td>
<td>4</td>
<td>17</td>
</tr>
</tbody>
</table>

3-hop memory network does pretty well, better than LSTM at processing these types of examples.
Evaluation: Children’s Book Test

<table>
<thead>
<tr>
<th>Methods</th>
<th>Named Entities</th>
</tr>
</thead>
<tbody>
<tr>
<td>Humans (query) (*)</td>
<td>0.520</td>
</tr>
<tr>
<td>Humans (context+query) (*)</td>
<td><strong>0.816</strong></td>
</tr>
<tr>
<td>Maximum frequency (corpus)</td>
<td>0.120</td>
</tr>
<tr>
<td>Maximum frequency (context)</td>
<td>0.335</td>
</tr>
<tr>
<td>Sliding window</td>
<td>0.168</td>
</tr>
<tr>
<td>Word distance model</td>
<td>0.398</td>
</tr>
<tr>
<td>Kneser-Ney language model</td>
<td>0.390</td>
</tr>
<tr>
<td>Kneser-Ney language model + cache</td>
<td>0.439</td>
</tr>
<tr>
<td>LSTMs (query)</td>
<td>0.408</td>
</tr>
<tr>
<td>LSTMs (context+query)</td>
<td>0.418</td>
</tr>
<tr>
<td>Contextual LSTMs (window context)</td>
<td>0.436</td>
</tr>
<tr>
<td>MemNNs (lexical memory)</td>
<td>0.431</td>
</tr>
<tr>
<td><strong>MemNNs (window memory)</strong></td>
<td><strong>0.493</strong></td>
</tr>
<tr>
<td>MemNNs (sentential memory + PE)</td>
<td>0.318</td>
</tr>
<tr>
<td><strong>MemNNs (window memory + self-sup.)</strong></td>
<td><strong>0.666</strong></td>
</tr>
</tbody>
</table>

- Outperforms LSTMs substantially with the right supervision
Memory Network Takeaways

- Memory networks provide a way of attending to abstractions over the input.
- Useful model for attending to multiple parts of an input.
- What can we do with more basic attention?
CNN/Daily Mail: Attentive Reader
Single-document, (usually) single-sentence cloze task

Formed based on article summaries — information should mostly be present, makes it easier than Children’s Book Test

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

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: Bidirectional Attention Flow
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
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.

Rajpurkar et al. (2016)
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.
- Dist over query.

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

Seo et al. (2016)
Bidirectional Attention Flow

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!

Devlin et al. (2019)
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>2</td>
<td>BERT (ensemble)</td>
<td>87.433</td>
<td>93.160</td>
</tr>
<tr>
<td></td>
<td>Google AI Language</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><a href="https://arxiv.org/abs/1810.04805">Link</a></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>Google AI Language</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><a href="https://arxiv.org/abs/1810.04805">Link</a></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>Microsoft Research Asia</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>Microsoft Research Asia</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>Google Brain &amp; CMU</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>Microsoft Research Asia</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>Google Brain &amp; CMU</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...)*
<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></td>
<td>Joint Laboratory of HIT and iFLYTEK Research</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>BERT + ConvLSTM + MTL + Verifier (ensemble)</td>
<td>86.730</td>
<td>89.286</td>
</tr>
<tr>
<td></td>
<td>Layer 6 AI</td>
<td></td>
<td></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></td>
<td>Google AI Language</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><a href="https://github.com/google-research/bert">https://github.com/google-research/bert</a></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>SemBERT (ensemble)</td>
<td>86.166</td>
<td>88.886</td>
</tr>
<tr>
<td></td>
<td>Shanghai Jiao Tong University</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>BERT + DAE + AoA (single model)</td>
<td>85.884</td>
<td>88.621</td>
</tr>
<tr>
<td></td>
<td>Joint Laboratory of HIT and iFLYTEK Research</td>
<td></td>
<td></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></td>
<td>Google AI Language</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><a href="https://github.com/google-research/bert">https://github.com/google-research/bert</a></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>BERT + MMFT + ADA (ensemble)</td>
<td>85.082</td>
<td>87.615</td>
</tr>
<tr>
<td></td>
<td>Microsoft Research Asia</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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

- 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>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><em>Rajpurkar &amp; Jia et al. ‘18</em></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>ALBERT (ensemble model)</td>
<td>89.731</td>
<td>92.215</td>
</tr>
<tr>
<td></td>
<td><em>Google Research &amp; TTIC</em></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><em>PINGAN Omni-Sinitic</em></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><em>Google Research &amp; TTIC</em></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><em>Anonymous</em></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><em>Shanghai Jiao Tong University &amp; CloudWalk</em></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><em>Shanghai Jiao Tong University &amp; CloudWalk</em></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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 *The Guns of Navarone*. 

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
Latest Datasets

- DROP
- SQuAD 2.0
- SQuAD 2.0
- Multi-hop: next time
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.