Recall: SRL

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

![Figure from He et al. (2017)](image)

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...

  Score by matching expected answer phrase (EAP) against answer candidate (AC)

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
- \( \lambda x. \text{type}(x, \text{Location}) \land \text{born_in}(\text{Barack Obama}, x) \) (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

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]

QA is very broad

- Why use the KB at all? Why not answer questions directly from text? Like information retrieval! [Choi et al., 2015]
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
- Recognizing Textual Entailment (2006)
- MCTest (2013): 500 passages, 4 questions per passage
- Two questions per passage explicitly require cross-sentence reasoning

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

Reading Comprehension

- 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)
Better Systems

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

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?

Children’s Book Test

- 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.

Hill et al. (2015)
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 masals of his own to send soon. Esther felt relieved. She thought that had exaggerated matters a little.

Hill et al. (2015)

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

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

Paperno et al. (2016)

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

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

Sukhbaatar et al. (2015)

bAbI

- 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”

Weston et al. (2014)

Evaluation: bAbI

- 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>MemNNs (window memory)</td>
<td>0.493</td>
</tr>
<tr>
<td>MemNNs (sentential memory + PE)</td>
<td>0.318</td>
</tr>
<tr>
<td>MemNNs (window memory + self-sup.)</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

CNN/Daily Mail

- Passage

  if you feel a ripple in the force today, it may be the news that the official is getting its first gay character, according to the sci-fi website . the upcoming novel * will feature a capable but flawed official named who also happens to be a lesbian. * the character is the first gay figure in the official — the movies, television shows, comics and books approved by franchise owner — according to .

- Question

  characters in * * movies have gradually become more diverse

- Answer

  @entity6

Hermann et al. (2015), Chen et al. (2016)
CNN/Daily Mail

- LSTM reader: encode question, encode passage, predict entity

  ![Diagram](image1)

- Can also use textual entailment-like models

  ![Diagram](image2)

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

CNN/Daily Mail

- 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

Hermann et al. (2015)

CNN/Daily Mail

- 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

SQuAD: Bidirectional Attention Flow

- Stanford Attentive Reader

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

**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

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

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

query “specialized” to the $i$th word

dist over query

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

- BiDAF: 73 EM / 81 F1
- ninet, 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: Today**

- 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

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