CS388: Natural Language Processing
Lecture 19: Reading Comprehension

Greg Durrett
Administrivia

- Project 2 due Friday at 5pm
- Project proposals due next Thursday
- Spec posted on course website — I’ll pitch some ideas/interesting papers from EMNLP on Tuesday
Recall: CNNs for Machine Translation

- “ByteNet”: operates over characters (bytes)
- Encode source sequence w/dilated convolutions
- Predict $n$th target character by looking at the $n$th position in the source and a dilated convolution over the $n-1$ target tokens so far
- To deal with divergent lengths, $t_n$ actually looks at $s_{n\alpha}$ where $\alpha$ is a heuristically-chosen parameter
- Assumes mostly monotonic translation

Kalchbrenner et al. (2016)
Recall: Transformers

- Encoder and decoder are both transformers
- Decoder consumes the previous generated token (and attends to input), but has no recurrent state

Vaswani et al. (2017)
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”

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

(other representations like SQL possible too...)

- 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)
What can’t KB QA systems do?

- 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
- 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

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)
MCTest State of the Art

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

Sachan and Xing (2016)
10+ QA datasets released since 2015

- Children’s Book Test, CNN/Daily Mail, SQuAD, TriviaQA are most well-known (others: SearchQA, MS Marco, RACE, WikiHop, ...)

Question answering: questions are in natural language

- Answers: multiple choice or require picking from the passage
- 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
Axis 1: QA vs. cloze

Axis 2: single-sentence vs. passage

- Often shallow methods work well because most answers are in a single sentence (SQuAD, MCTest)
- Some explicitly require linking between multiple sentences (MCTest)

Axis 3: single-document (datasets in this lecture) vs. multi-document (TriviaQA, WikiHop, HotPotQA, ...)
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)
LSTM Language Models

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

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.

Hill et al. (2015)
Children’s Book Test: Results

- Present 10 options drawn from the text (correct + 9 distractors), ask the model to pick among them

<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>0.816</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>
</tbody>
</table>

- Neural LMs aren’t better than n-gram LMs

Hill et al. (2015)
Children’s Book Test: Results

- Present 10 options drawn from the text (correct + 9 distractors), ask the model to pick among them

<table>
<thead>
<tr>
<th>METHODS</th>
<th>NAMED ENTITIES</th>
<th>COMMON NOUNS</th>
<th>VERBS</th>
<th>PREPOSITIONS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Humans (query) (*)</td>
<td>0.520</td>
<td>0.644</td>
<td>0.716</td>
<td>0.676</td>
</tr>
<tr>
<td>Humans (context+query) (*)</td>
<td><strong>0.816</strong></td>
<td><strong>0.816</strong></td>
<td><strong>0.828</strong></td>
<td>0.708</td>
</tr>
<tr>
<td>Maximum frequency (corpus)</td>
<td>0.120</td>
<td>0.158</td>
<td>0.373</td>
<td>0.315</td>
</tr>
<tr>
<td>Maximum frequency (context)</td>
<td>0.335</td>
<td>0.281</td>
<td>0.285</td>
<td>0.275</td>
</tr>
<tr>
<td>Sliding window</td>
<td>0.168</td>
<td>0.196</td>
<td>0.182</td>
<td>0.101</td>
</tr>
<tr>
<td>Word distance model</td>
<td>0.398</td>
<td>0.364</td>
<td>0.380</td>
<td>0.237</td>
</tr>
<tr>
<td>Knéser-Ney language model</td>
<td><strong>0.390</strong></td>
<td><strong>0.544</strong></td>
<td><strong>0.778</strong></td>
<td><strong>0.768</strong></td>
</tr>
<tr>
<td>Knéser-Ney language model + cache</td>
<td><strong>0.439</strong></td>
<td><strong>0.577</strong></td>
<td><strong>0.772</strong></td>
<td><strong>0.679</strong></td>
</tr>
<tr>
<td>LSTMs (query)</td>
<td><strong>0.408</strong></td>
<td><strong>0.541</strong></td>
<td><strong>0.813</strong></td>
<td><strong>0.802</strong></td>
</tr>
<tr>
<td>LSTMs (context+query)</td>
<td><strong>0.418</strong></td>
<td><strong>0.560</strong></td>
<td><strong>0.818</strong></td>
<td><strong>0.791</strong></td>
</tr>
</tbody>
</table>

- Why are these results so low?  

Hill et al. (2015)
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)
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)
**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”

<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>Strongly Supervised MemNN [22]</th>
<th>LSTM [22]</th>
<th>MemNN WSH</th>
<th>BoW</th>
<th>PE</th>
<th>MemN2N 1 hop PE LS joint</th>
<th>MemN2N 2 hops PE LS joint</th>
<th>MemN2N 3 hops PE LS joint</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean error (%)</td>
<td>6.7</td>
<td>51.3</td>
<td>40.2</td>
<td>25.1</td>
<td>20.3</td>
<td>25.8</td>
<td>15.6</td>
<td>13.3</td>
</tr>
<tr>
<td>Failed tasks (err. &gt; 5%)</td>
<td>4</td>
<td>20</td>
<td>18</td>
<td>15</td>
<td>13</td>
<td>17</td>
<td>11</td>
<td>11</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>0.493</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 for cloze tasks where far-back context is necessary
- 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
  
  \[
  \text{X visited England} \quad ||| \quad \text{Mary visited England}
  \]

• Can also use textual entailment-like models
  
  \[
  \text{X visited England} \quad ||| \quad \text{Mary visited England}
  \]

Multiclass classification problem over entities in the document

Hermann et al. (2015), Chen et al. (2016)
Attentive reader:
- u = encode query
- s = encode sentence
- r = attention(u -> s)
- prediction = f(candidate, u, r)

- Uses fixed-size representations for the final prediction, multiclass classification

Hermann et al. (2015)
- 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
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

One of the most famous people born in Warsaw was **Maria Skłodowska-Curie**, who achieved international recognition for her research on radioactivity and was the first **female recipient** of the **Nobel Prize**. Famous musicians include Władysław Szpilman and Frédéric Chopin. Though Chopin was born in the village of Żelazowa Wola, about 60 km (37 mi) from Warsaw, he moved to the city with his family when he was seven months old. Casimir Pulaski, a Polish general and hero of the American Revolutionary War, was born here in 1745.

What was Maria Curie the first female recipient of?
- **Ground Truth Answers:** Nobel Prize  Nobel Prize  Nobel Prize

What year was Casimir Pulaski born in Warsaw?
- **Ground Truth Answers:** 1745  1745  1745

Who was one of the most famous people born in Warsaw?
- **Ground Truth Answers:** Maria Skłodowska-Curie  Maria Skłodowska-Curie

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.
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.

- $\tilde{u}_i = \sum_j \alpha_{ij} u_j$
- $\alpha_{ij} = \text{softmax}_j(S_{ij})$
- query “specialized” to the $i$th word
- dist over query

$S_{ij} = h_i \cdot u_j$

Seo et al. (2016)
Bidirectional Attention Flow

Seo et al. (2016)

Each passage word now “knows about” the query.
# SQuAD SOTA

<table>
<thead>
<tr>
<th>Rank</th>
<th>Model</th>
<th>EM</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>BERT (ensemble) Google AI Language</td>
<td>87.433</td>
<td>93.160</td>
</tr>
<tr>
<td>2</td>
<td>BERT (single model) Google AI Language</td>
<td>85.083</td>
<td>91.835</td>
</tr>
<tr>
<td>2</td>
<td>nlnet (ensemble) Microsoft Research Asia</td>
<td>85.356</td>
<td>91.202</td>
</tr>
<tr>
<td>3</td>
<td>QANet (ensemble) Google Brain &amp; CMU</td>
<td>84.454</td>
<td>90.490</td>
</tr>
<tr>
<td>4</td>
<td>r-net (ensemble) Microsoft Research Asia</td>
<td>84.003</td>
<td>90.147</td>
</tr>
<tr>
<td>5</td>
<td>QANet (ensemble) Google Brain &amp; CMU</td>
<td>83.877</td>
<td>89.737</td>
</tr>
</tbody>
</table>

- **BiDAF**: 73 EM / 81 F1
- nlnet, QANet, r-net — dueling super complex systems (much more than BiDAF...)
- **BERT**: transformer-based approach with pretraining on 3B tokens
But how well are these doing?

- Can construct adversarial examples that fool these systems: add one carefully chosen sentence and performance drops to below 50%

- Still “surface-level” matching, not complex understanding

- Other challenges: recognizing when answers aren’t present, doing multi-step reasoning

**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 and Liang (2017)
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.