Broad Context Language Modeling as Reading Comprehension

Zewei Chu    Hai Wang    Kevin Gimpel    David McAllester
LAMBADA: Word prediction requiring a broad discourse context
(Paperno et al., 2016)

he bent down and searched the large container, trying to find anything else hidden in it other than the _____
he turned to one of the cops beside him. “search the entire coffin.” the man nodded and bustled forward towards the coffin.

he bent down and searched the large container, trying to find anything else hidden in it other than the ______
he turned to one of the cops beside him. “search the entire coffin.” the man nodded and bustled forward towards the coffin.

he bent down and searched the large container, trying to find anything else hidden in it other than the body.

**only answerable given discourse context!**
he turned to one of the cops beside him. “search the entire coffin.” the man nodded and bustled forward towards the coffin.

he bent down and searched the large container, trying to find anything else hidden in it other than the body.

In >80% of instances, answer is in context (though not in this example)
Overview

- We view LAMBADA as reading comprehension and apply off-the-shelf models

- We improve state-of-the-art from 7.3% to 49%

- Manual analysis shows a variety of phenomena:
  - Easy ones solved by comprehension models
  - Hard ones need more semantics, coreference, external knowledge
Document:
actress @entity1 has entered a rehab facility for her addictions, a spokesman said. "@entity1 has valiantly battled substance abuse over the years and whenever she has needed to seek treatment she has done so," said spokesman @entity5 … @entity1 won an @entity15 in 1973 for her performance in "cabaret." …

Question:
XXXXX won an @entity15 for her performance in "cabaret"
Document:
actress @entity1 has entered a rehab facility for her addictions, a spokesman said. "@entity1 has valiantly battled substance abuse over the years and whenever she has needed to seek treatment she has done so," said spokesman @entity5 ...
@entity1 won an @entity15 in 1973 for her performance in "cabaret."

Question:
XXXXX won an @entity15 for her performance in "cabaret"
Neural Reading Comprehension Models

- lots of recent activity here!
  - Hermann et al. (2015)
  - Hill et al. (2016)
  - Chen et al. (2016)
  - Kadlec et al. (2016)
  - Dhingra et al. (2016)
  - *inter alia*

- we will describe the Attention Sum Reader (Kadlec et al., 2016) because it is simple and works well
# Attention Sum Reader

(Adlec et al., 2016)

<table>
<thead>
<tr>
<th>Document</th>
<th>Question</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input text</td>
<td>..... Obama and Putin ..... said Obama in Prague</td>
</tr>
</tbody>
</table>
Attention Sum Reader
(Kadlec et al., 2016)

- Encode document using bidirectional RNN
- Encode question using another bidirectional RNN
Attention Sum Reader
(Kadlec et al., 2016)

Compute attention over positions of document using question representation
Attention Sum Reader
(Kadlec et al., 2016)

<table>
<thead>
<tr>
<th>Document</th>
<th>Question</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input text</td>
<td>..... Obama and Putin ..... said Obama in Prague</td>
</tr>
<tr>
<td>Embeddings</td>
<td>..... e(Obama) e(and) e(Putin) ..... e(said) e(Obama) e(in) e(Prague)</td>
</tr>
</tbody>
</table>

- Normalize over positions of document
### Attention Sum Reader (Kadlec et al., 2016)

<table>
<thead>
<tr>
<th>Input text</th>
<th>Obama and Putin said Obama in Prague</th>
</tr>
</thead>
</table>

#### Embeddings

- e(Obama)
- e(and)
- e(Putin)
- e(said)
- e(Obama)
- e(in)
- e(Prague)

#### Recurrent neural networks

$f$ and $g$

#### Dot products

Softmax $s_i$ over words in the document

$P(Obama|q, d) = \sum_{i \in I(Obama,d)} s_i = s_j + s_{j+5}$
Gated Attention Reader
(Dhingra et al., 2016)
Training Data?

- we need data to train these comprehension models

- LAMBADA only includes dev/test sets
Automatic Training Data Creation

- **Observation:** In >80% of LAMBADA, answer word is in context

- We automatically create training instances where answer word is in context
  - Each instance has 4-5 sentences and >=50 words
  - Total of 1.8 million instances for training

- Training data is available (see my web page)
Results: LAMBADA Test Set

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>LSTM</td>
<td>0</td>
</tr>
<tr>
<td>n-gram</td>
<td>0.1</td>
</tr>
<tr>
<td>random capitalized word in context</td>
<td>7.3</td>
</tr>
<tr>
<td>Human</td>
<td>86</td>
</tr>
</tbody>
</table>

Paperno et al. (2016)
Results: LAMBADA Test Set

Paperno et al. (2016)

0 0.1 7.3 11.7 86

Accuracy

LSTM n-gram random capitalized word in context most frequent word in context Human
Results: LAMBADA Test Set

Paperno et al. (2016)
Results: Neural Readers on LAMBADA

Accuracy

- random capitalized word in context: 7.3
- most frequent word in context: 11.7
- Attention Sum Reader
- Gated Attention Reader
- Human: 86
Results: Neural Readers on LAMBADA

- **Random capitalized word in context**: 7.3
- **Most frequent word in context**: 11.7
- **Attention Sum Reader**: 44.5
- **Gated Attention Reader**: 49
- **Human**: 86

Accuracy
Results: Neural Readers on LAMBADA

Dhingra, Yang, Cohen, Salakhutdinov. “Linguistic Knowledge as Memory for Recurrent Neural Networks” arXiv 2017

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>random capitalized word in context</td>
<td>7.3</td>
</tr>
<tr>
<td>most frequent word in context</td>
<td>11.7</td>
</tr>
<tr>
<td>Attention Sum Reader</td>
<td>44.5</td>
</tr>
<tr>
<td>Gated Attention Reader</td>
<td>49</td>
</tr>
<tr>
<td>Gated Att. + Ling. Knowledge</td>
<td>51.6</td>
</tr>
<tr>
<td>Human</td>
<td>86</td>
</tr>
</tbody>
</table>
Manual Analysis

- 100 LAMBADA instances
- 86 of 100 correct
- labeled instances with types of understanding needed to solve them
... it meant that dog and cat could contribute to the conversation.

“hey, dog, ever been in one of these things?” asked cat.

“no.” replied dog sadly.

“you say that so glumly,” said ____
Example Label: Simple Speaker Tracking

... it meant that dog and cat could contribute to the conversation.

“hey, dog, ever been in one of these things?” asked cat.

“no.” replied dog sadly.

“you say that so glumly,” said ____

- Answerable by tracking who is speaking without understanding what they are saying
- 19 of 100
- these are also easy (16/19 correct)
Example Label 2: Coreference Resolution

instead, he danced, turned circles under the spider's body, and crouched close to the spider's abdomen to avoid its bite.

   brian joined thomas. the two boys fought together, just as they had done in our prison, just as the spiders were doing. now both were stabbing and dodging, double-teaming against the ______

- 21 of 100
- these are hard (8/21 correct)
Example Label 3: External Knowledge

...he turned to one of the cops beside him. “search the entire coffin.”

the man nodded and bustled forward towards the coffin. he bent down and searched the large container, trying to find anything else hidden in it other than the ______

- Involves knowledge possessed by human readers but not contained in context
- 24 of 100
- these are also hard (5/24 correct)
Conclusions

- reading comprehension models improve state-of-the-art on LAMBADA from 7.3% to 49% (since improved further)
- but the last 50% are difficult!
- manual analysis reveals several categories of phenomena, including coreference and external knowledge
Thanks!
“no worries; she's fine,” said sheila, opening the door for him to enter.

“oh, i can't stay. i just came to offer my services. if you need anything, don't hesitate to ask. i'm right across the street.”

“actually, we need someone to fix our plumbing,” said _____
Example Label 1: Unambiguous Name Cue

“no worries; she's fine,” said sheila, opening the door for him to enter.

“oh, i can't stay. i just came to offer my services. if you need anything, don't hesitate to ask. i'm right across the street.”

“actually, we need someone to fix our plumbing,” said _____

- answer is clearly a name based on local cues, and context only contains a single name
- 9 of 100
- these are easy (8/9 correct by GA Reader)
Results: Neural Readers on LAMBADA

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>random capitalized word in context</td>
<td>7.3</td>
</tr>
<tr>
<td>Stanford Reader</td>
<td>21.7</td>
</tr>
<tr>
<td>Modified Stanford Reader</td>
<td>32.1</td>
</tr>
<tr>
<td>Attention Sum Reader</td>
<td>44.5</td>
</tr>
<tr>
<td>Gated Attention Reader</td>
<td>49</td>
</tr>
<tr>
<td>Human</td>
<td>86</td>
</tr>
</tbody>
</table>
Results: Neural Readers on LAMBADA

Dhingra, Yang, Cohen, Salakhutdinov. “Linguistic Knowledge as Memory for Recurrent Neural Networks” arXiv 2017

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>random capitalized word in context</td>
<td>7.3</td>
</tr>
<tr>
<td>Stanford Reader</td>
<td>21.7</td>
</tr>
<tr>
<td>Modified Stanford Reader</td>
<td>32.1</td>
</tr>
<tr>
<td>Attention Sum Reader</td>
<td>44.5</td>
</tr>
<tr>
<td>Gated Attention Reader</td>
<td>49</td>
</tr>
<tr>
<td>Gated Attention + Linguistic Knowledge</td>
<td>51.6</td>
</tr>
<tr>
<td>Human</td>
<td>86</td>
</tr>
</tbody>
</table>