CNN/Daily Mail Reading Comprehension Task

Danqi Chen, Jason Bolton, Christopher D. Manning
Presented By: Jianqiu Kong
Overview

• Introduction
• Data
• Models and Systems
• Experiment
• Conclusion
Introduction

• Reading comprehension (RC) is the ability to read text, process it, and understand its meaning.

• Genuine reading comprehension involves interpretation of the text and making complex inferences.

• This paper provides an in-depth analysis of CNN/DailyMail dataset and what level of natural language understanding is needed
Data

• made from articles on the news websites CNN and Daily Mail

• It it tokenized, lowercased, and named entity recognition and coreference resolution have been run

• consists of a passage $p$, a question $q$ and an answer
Data

- the question is a cloze-style task, in which one of the article’s bullet points has had one entity replaced by a placeholder

- the answer is this questioned entity

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
Reading Comprehension Task

- The goal is to infer the missing entity (answer a) from all the possible entities which appear in the passage.

- suffer when either of entity recognition and coreference fails
Models and Systems

- conventional entity-centric classifier
- end-to-end neural network
Entity-Centric Classifier

Main Idea: design a feature vector $f_{p,q}(e)$ for each candidate entity $e$, and to learn a weight vector $\theta$ such that the correct answer $a$ is expected to rank higher than all other candidate entities.

$$\theta^T f_{p,q}(a) > \theta^T f_{p,q}(e), \forall e \in E \cap p \setminus \{a\}$$
Entity-Centric Classifier

feature templates employed:
• Whether entity e occurs in the passage
• Whether entity e occurs in the question
• The frequency of entity e in the passage
• The first position of occurrence of entity e in the passage
Entity-Centric Classifier

feature templates employed:

• n-gram exact match: between the text surrounding the placeholder and the text surrounding entity $e$
• Word distance: average minimum distance of each non-stop question word from the entity in the passage
• Sentence co-occurrence
• Dependency parse match:

\[
\begin{align*}
w & \xrightarrow{r} \text{@placeholder} \quad w \xrightarrow{r} e
\end{align*}
\]
End-to-end Neural Network

Encoding:

- all the words are mapped to d-dimensional vectors

- shallow bi-directional LSTM to encode contextual embeddings of each word in the passage

\[
\overrightarrow{h}_i = \text{LSTM}(\overrightarrow{h}_{i-1}, p_i), i = 1, \ldots, m \\
\overleftarrow{h}_i = \text{LSTM}(\overleftarrow{h}_{i+1}, p_i), i = m, \ldots, 1
\]

- use another bi-directional LSTM to map the question
End-to-end Neural Network

Attention:

\[ \alpha_i = \text{softmax}_i \mathbf{q}^T \mathbf{W_s} \mathbf{p}_i \]
\[ \mathbf{o} = \sum_i \alpha_i \mathbf{p}_i \]

\( \mathbf{o} \): a weighted combination of all contextual embeddings
\( \mathbf{W_s} \): in bilinear form
End-to-end Neural Network

Prediction:

\[ a = \arg \max_{a \in p \cap E} W_a^T o \]
End-to-end Neural Network

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
Experiments

Training the conventional classifier:

Stanford’s neural network dependency parser

ranking algorithm: LambdaMART
Experiments

**Training the neural network:**
keep most frequent 50k words, others as `<unk>` token
100-dimensional pretrained GloVe word embeddings
attention and output parameters initialized from $U(-0.01,0.01)$
LSTM weights initialized from $N(0,0.1)$
hidden size 128 for CNN and 256 for Daily Mail
## Experiments

### Main Results:

<table>
<thead>
<tr>
<th>Model</th>
<th>CNN Dev</th>
<th>CNN Test</th>
<th>Daily Mail Dev</th>
<th>Daily Mail Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frame-semantic model †</td>
<td>36.3</td>
<td>40.2</td>
<td>35.5</td>
<td>35.5</td>
</tr>
<tr>
<td>Word distance model †</td>
<td>50.5</td>
<td>50.9</td>
<td>56.4</td>
<td>55.5</td>
</tr>
<tr>
<td>Deep LSTM Reader †</td>
<td>55.0</td>
<td>57.0</td>
<td>63.3</td>
<td>62.2</td>
</tr>
<tr>
<td>Attentive Reader †</td>
<td>61.6</td>
<td>63.0</td>
<td>70.5</td>
<td>69.0</td>
</tr>
<tr>
<td>Impatient Reader †</td>
<td>61.8</td>
<td>63.8</td>
<td>69.0</td>
<td>68.0</td>
</tr>
<tr>
<td>MemNNs (window memory) ‡</td>
<td>58.0</td>
<td>60.6</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>MemNNs (window memory + self-sup.) ‡</td>
<td>63.4</td>
<td>66.8</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>MemNNs (ensemble) ‡</td>
<td>66.2*</td>
<td>69.4*</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Ours: Classifier</td>
<td>67.1</td>
<td>67.9</td>
<td>69.1</td>
<td>68.3</td>
</tr>
<tr>
<td>Ours: Neural net</td>
<td>72.4</td>
<td>72.4</td>
<td>76.9</td>
<td>75.8</td>
</tr>
</tbody>
</table>
Feature ablation analysis of entity centric classifier

<table>
<thead>
<tr>
<th>Features</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full model</td>
<td>67.1</td>
</tr>
<tr>
<td>– whether ( e ) is in the passage</td>
<td>67.1</td>
</tr>
<tr>
<td>– whether ( e ) is in the question</td>
<td>67.0</td>
</tr>
<tr>
<td>– frequency of ( e )</td>
<td>63.7</td>
</tr>
<tr>
<td>– position of ( e )</td>
<td>65.9</td>
</tr>
<tr>
<td>– ( n )-gram match</td>
<td>60.5</td>
</tr>
<tr>
<td>– word distance</td>
<td>65.4</td>
</tr>
<tr>
<td>– sentence co-occurrence</td>
<td>66.0</td>
</tr>
<tr>
<td>– dependency parse match</td>
<td>65.6</td>
</tr>
</tbody>
</table>

A low number indicates an important feature.
Breakdown of examples

- Exact match
- Sentence-level paraphrasing
- Partial clue
- Multiple sentences
- Coreference errors
- Ambiguous or very hard
## Breakdown of examples

<table>
<thead>
<tr>
<th>Category</th>
<th>Classifier</th>
<th>Neural net</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exact match</td>
<td>13 (100.0%)</td>
<td>13 (100.0%)</td>
</tr>
<tr>
<td>Paraphrasing</td>
<td>32 (78.1%)</td>
<td>39 (95.1%)</td>
</tr>
<tr>
<td>Partial clue</td>
<td>14 (73.7%)</td>
<td>17 (89.5%)</td>
</tr>
<tr>
<td>Multiple sentences</td>
<td>1 (50.0%)</td>
<td>1 (50.0%)</td>
</tr>
<tr>
<td>Coreference errors</td>
<td>4 (50.0%)</td>
<td>3 (37.5%)</td>
</tr>
<tr>
<td>Ambiguous / hard</td>
<td>2 (11.8%)</td>
<td>1 (5.9%)</td>
</tr>
<tr>
<td><strong>All</strong></td>
<td>66 (66.0%)</td>
<td>74 (74.0%)</td>
</tr>
</tbody>
</table>

Neural networks are better capable of learning semantic matches involving paraphrasing or lexical variation between the two sentences.
Conclusion

- the CNN/Daily Mail datasets is still quite noisy due to its method of data creation and coreference errors

- current neural networks have almost reached a performance ceiling on this dataset

- the required reasoning and inference level of this dataset is still quite simple
Thanks!