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	Question	Answer
(@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 .	characters in " @placeholder " movies have gradually become more diverse	@entity6

Reading Comprehesion 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 θ such that the correct answer a is expected to rank higher than all other candidate entities.

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

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 occurence 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:

$$w \xrightarrow{r} \text{@placeholder} \quad ; \quad w \xrightarrow{r} e$$

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

$$\overrightarrow{\mathbf{h}}_i = \text{LSTM}(\overrightarrow{\mathbf{h}}_{i-1}, \mathbf{p}_i), i = 1, \ldots, m$$
$$\overleftarrow{\mathbf{h}}_i = \text{LSTM}(\overleftarrow{\mathbf{h}}_{i+1}, \mathbf{p}_i), i = m, \ldots, 1$$

- use another bi-directional LSTM to map the question

End-to-end Neural Network

Attention:

$$\alpha_i = \text{softmax}_i \, \mathbf{q}^\intercal \mathbf{W}_s \tilde{\mathbf{p}}_i$$
$$\mathbf{o} = \sum_i \alpha_i \tilde{\mathbf{p}}_i$$

o: a weighted combination of all contextual embeddings

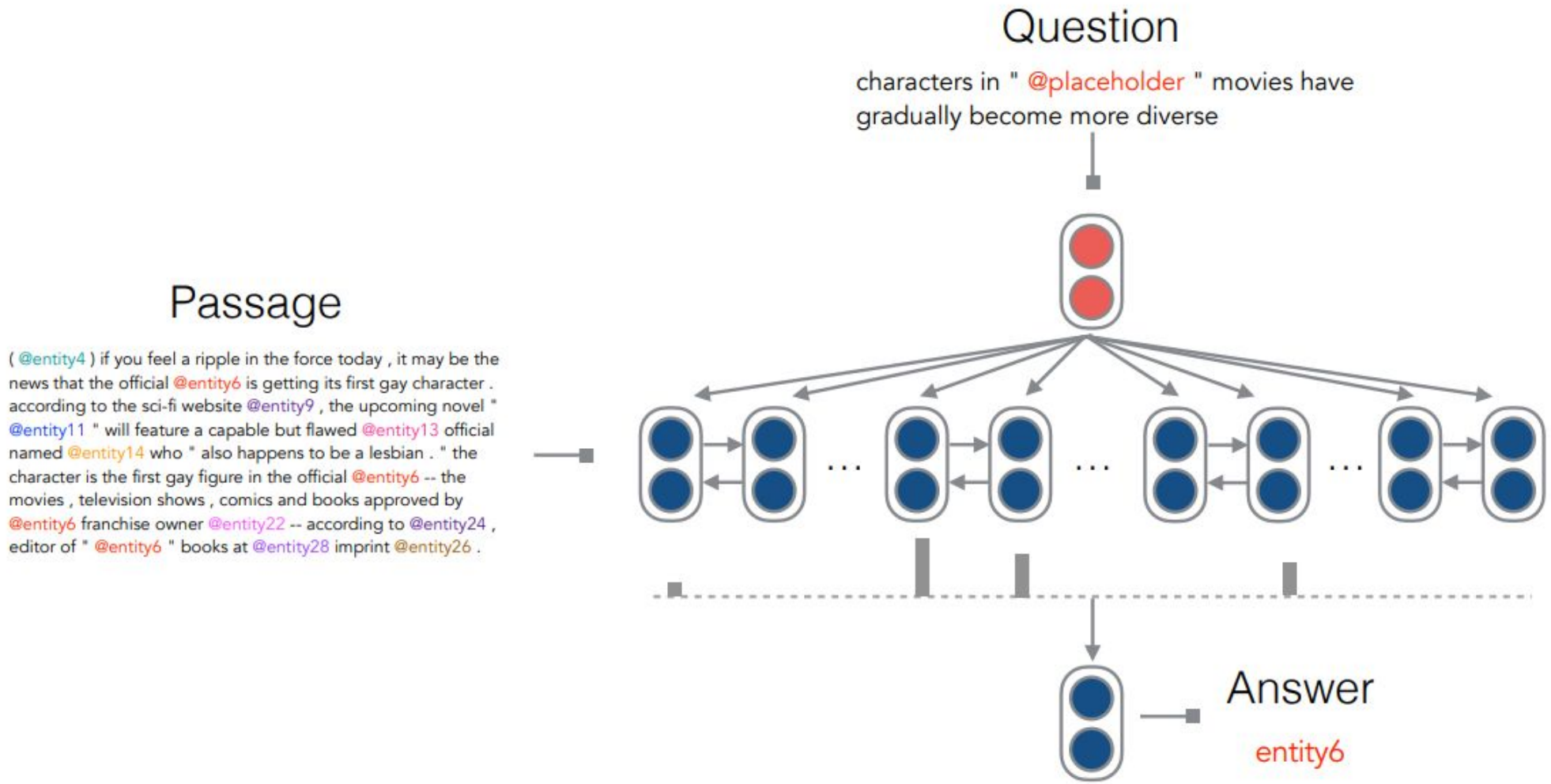
Ws in bilinear form

End-to-end Neural Network

Prediction:

$$a = \arg\max_{a \in p \cap E} W_a^{\intercal} \mathbf{o}$$

End-to-end Neural Network



Experiments

Training the conventional classifier:

Stanford’s neural network dependency parser

ranking algorith: 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:

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

Feature ablation analysis of entity centric classifier

Features	Accuracy
Full model	67.1
– whether *e* is in the passage	67.1
– whether *e* is in the question	67.0
– frequency of *e*	**63.7**
– position of *e*	65.9
– *n*-gram match	**60.5**
– word distance	65.4
– sentence co-occurrence	66.0
– dependency parse match	65.6

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

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

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!