A Thorough Examination of the CNN/Daily Mail Reading Comprehension Task

Danqi Chen, Jason Bolton, Christopher D. Manning

Stanford University

August 10, 2016

A Thorough Examination of the CNN/Daily Mail Reading Comprehension Task

Danqi Chen, Jason Bolton, Christopher D. Manning
Stanford University

August 10, 2016

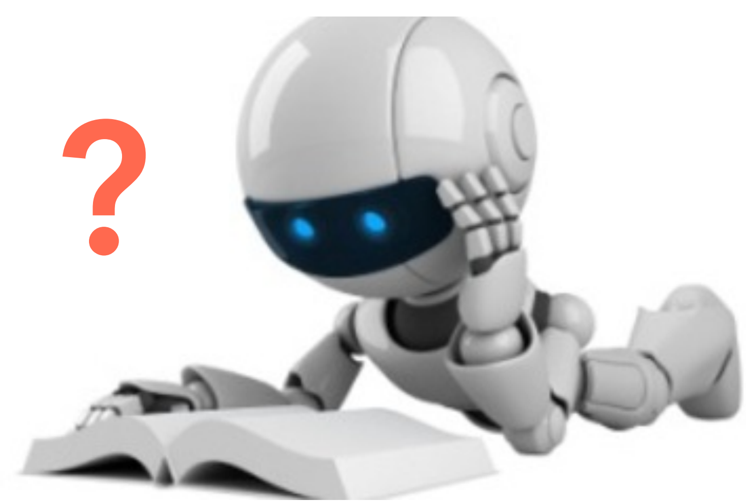
Reading Comprehension

Reading comprehension is the ability to **read text**, **process it**, and **understand its meaning**.

Reading Comprehension

WIKIPEDIA
The Free Encyclopedia

Reading comprehension is the ability to **read text**, **process it**, and **understand its meaning**.



Reading Comprehension

Passage (P) + Question (Q) $\longrightarrow$ Answer (A)

Reading Comprehension

Passage (P) + Question (Q) $\longrightarrow$ Answer (A)

P

Alyssa got to the beach after a long trip. She's from Charlotte. She traveled from Atlanta. She's now in Miami. She went to Miami to visit some friends. But she wanted some time to herself at the beach, so she went there first. After going swimming and laying out, she went to her friend Ellen's house. Ellen greeted Alyssa and they both had some lemonade to drink. Alyssa called her friends Kristin and Rachel to meet at Ellen's house…….

Q What city is Alyssa in?

A Miami

Data is a bottleneck

- People have attempted to collect **human-labeled** data for reading comprehension:
  - **MCTest** (Richardson et al, 2013): 660 x 4 questions
  - **ProcessBank** (Berant et al, 2014): 585 questions

Data is a bottleneck

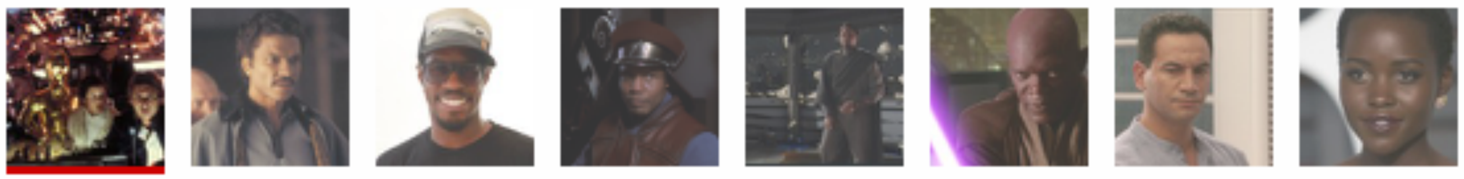
- People have attempted to collect **human-labeled** data for reading comprehension:
  - **MCTest** (Richardson et al, 2013): 660 x 4 questions
  - **ProcessBank** (Berant et al, 2014): 585 questions

- Small, expensive
- Difficult to learn statistical models

(Hermann et al, 2015)

CNN/Daily Mail Datasets

CNN Entertainment » 'Star Wars' universe gets its first gay character



Story highlights

Official "Star Wars" universe gets its first gay character, a lesbian governor

The character appears in the upcoming novel "Lords of the Sith"

Characters in "Star Wars" movies have gradually become more diverse

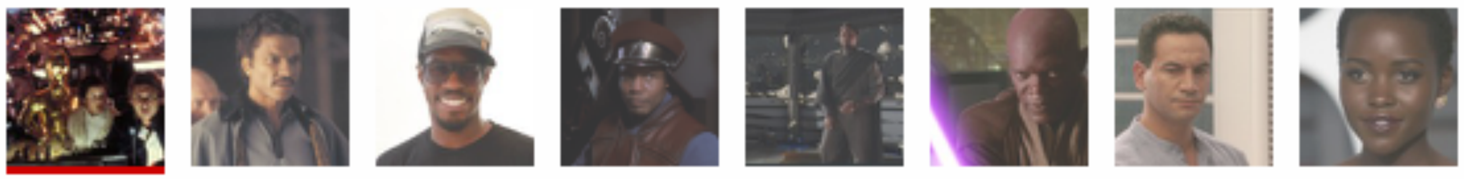
(CNN) — If you feel a ripple in the Force today, it may be the news that the official Star Wars universe is getting its first gay character.

According to the sci-fi website Big Shiny Robot, the upcoming novel "Lords of the Sith" will feature a capable but flawed Imperial official named Moff Mors who "also happens to be a lesbian."

The character is the first gay figure in the official Star Wars universe -- the movies, television shows, comics and books approved by Star Wars franchise owner Disney -- according to Shelly Shapiro, editor of "Star Wars" books at Random House imprint Del Rey Books.

CNN/Daily Mail Datasets

CNN Entertainment » 'Star Wars' universe gets its first gay character



Story highlights

Official "Star Wars" universe gets its first gay character, a lesbian governor

The character appears in the upcoming novel "Lords of the Sith"

Characters in "Star Wars" movies have gradually become more diverse

(CNN) — If you feel a ripple in the Force today, it may be the news that the official Star Wars universe is getting its first gay character.

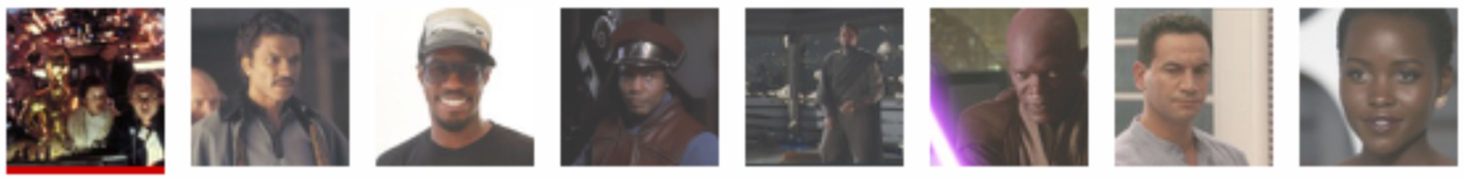
According to the sci-fi website Big Shiny Robot, the upcoming novel "Lords of the Sith" will feature a capable but flawed Imperial official named Moff Mors who "also happens to be a lesbian."

The character is the first gay figure in the official Star Wars universe -- the movies, television shows, comics and books approved by Star Wars franchise owner Disney -- according to Shelly Shapiro, editor of "Star Wars" books at Random House imprint Del Rey Books.

(Hermann et al, 2015)

CNN/Daily Mail Datasets

CNN Entertainment » 'Star Wars' universe gets its first gay character



Story highlights

Official "Star Wars" universe gets its first gay character, a lesbian governor

The character appears in the upcoming novel "Lords of the Sith"

Characters in ______ movies have gradually become more diverse

(CNN) — If you feel a ripple in the Force today, it may be the news that the official Star Wars universe is getting its first gay character.

According to the sci-fi website Big Shiny Robot, the upcoming novel "Lords of the Sith" will feature a capable but flawed Imperial official named Moff Mors who "also happens to be a lesbian."

The character is the first gay figure in the official Star Wars universe -- the movies, television shows, comics and books approved by Star Wars franchise owner Disney -- according to Shelly Shapiro, editor of "Star Wars" books at Random House imprint Del Rey Books.

(Hermann et al, 2015)

CNN/Daily Mail Datasets

P

(@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 "

Q

characters in " @placeholder " movies have gradually become more diverse

A

@entity6

(Hermann et al, 2015)

CNN/Daily Mail Datasets

P

(@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 "

Q

characters in " @placeholder " movies have gradually become more diverse

A

@entity6

CNN: 380k, Daily Mail: 879k training - free!

What is this paper about?

What is this paper about?

System **Lower Bound**

Our **simple systems** work quite well.

What is this paper about?

System **Lower Bound**

Our **simple systems** work quite well.

Analysis **Upper Bound**

The task might be not that hard.
We are **almost done**.

What is this paper about?

System **Lower Bound**

Our **simple systems** work quite well.

Analysis **Upper Bound**

The task might be not that hard.
We are **almost done**.

Discussion: what's next?

What is this paper about?

System **Lower Bound**

Our **simple systems** work quite well.

Analysis **Upper Bound**

The task might be not that hard.
We are **almost done**.

Discussion: what's next?

System I: Entity-Centric Classifier

- For each candidate entity e, we build a symbolic feature vector:

$$f_{P,Q}(e)$$

System I: Entity-Centric Classifier

- For each candidate entity *e*, we build a symbolic feature vector:

$$f_{P,Q}(e)$$

- The goal is to learn feature weights such that the correct answer ranks higher than the other entities (we used **LambdaMart** algorithm).

System I: Entity-Centric Classifier

- For each candidate entity e, we build a symbolic feature vector:

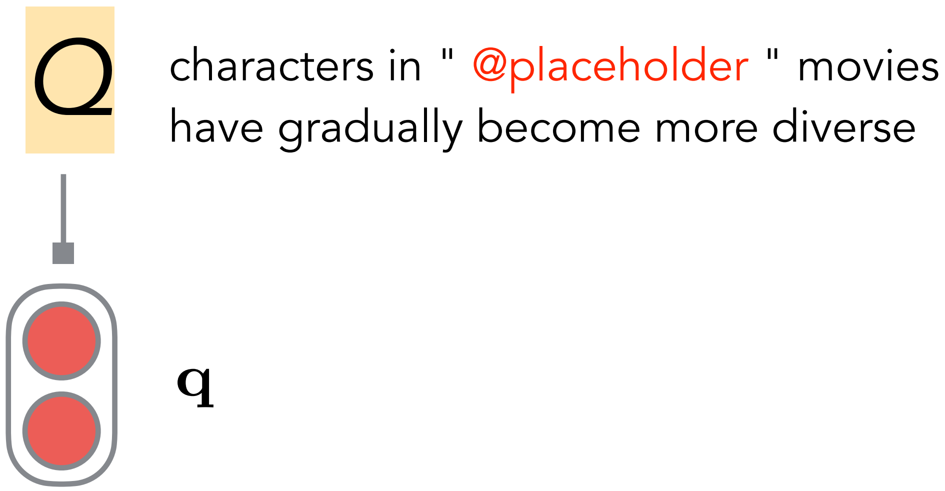
$$f_{P,Q}(e)$$

1. Whether *e* occurs in *P*
2. Whether *e* occurs in *Q*
3. Frequency of *e* in *P*
4. First position of *e* in *P*

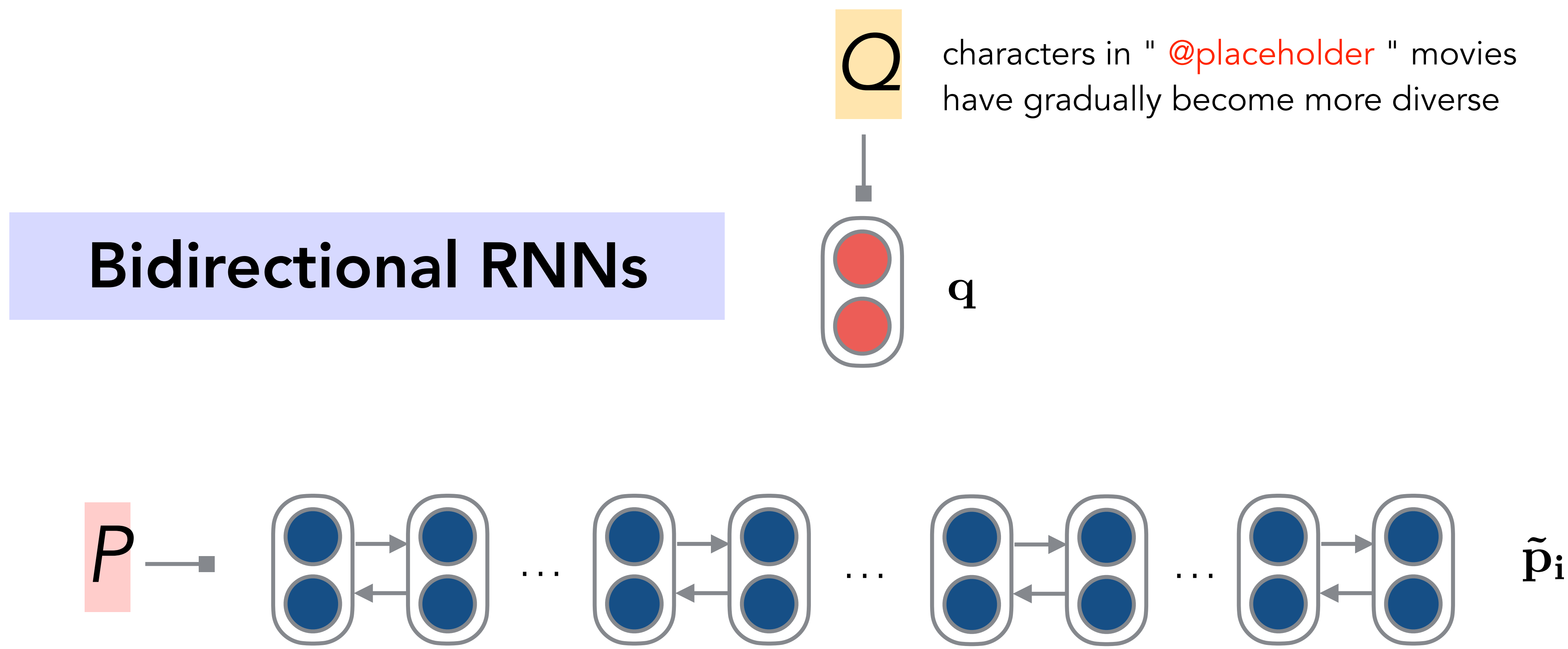
5. *Whether e co-occurs with another Q word in P.*
6. word **distance**
7. **n-gram** exact match
8. **dependency parse** match

System II: End-to-end Neural Network

Bidirectional RNNs

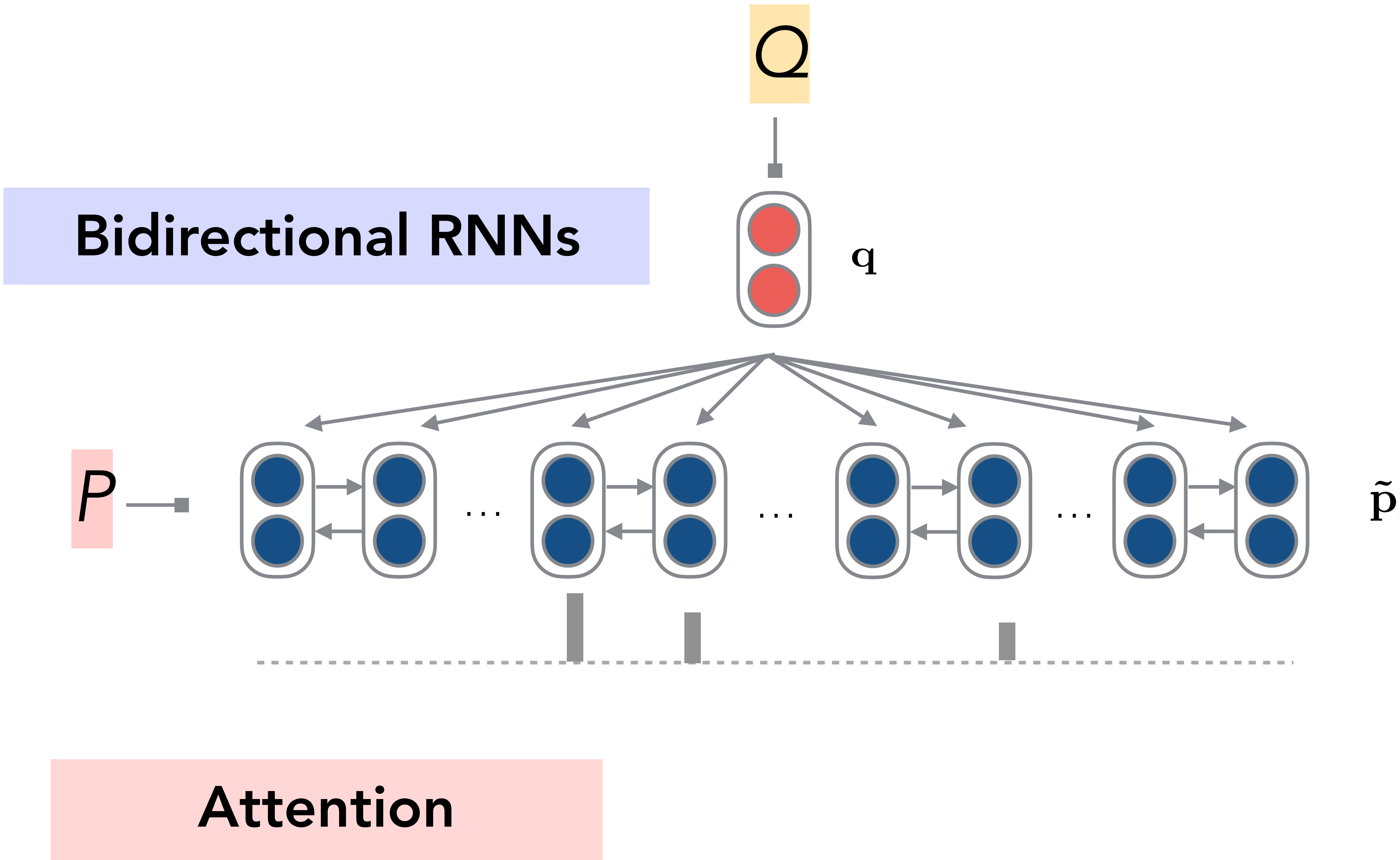


System II: End-to-end Neural Network



(@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 "

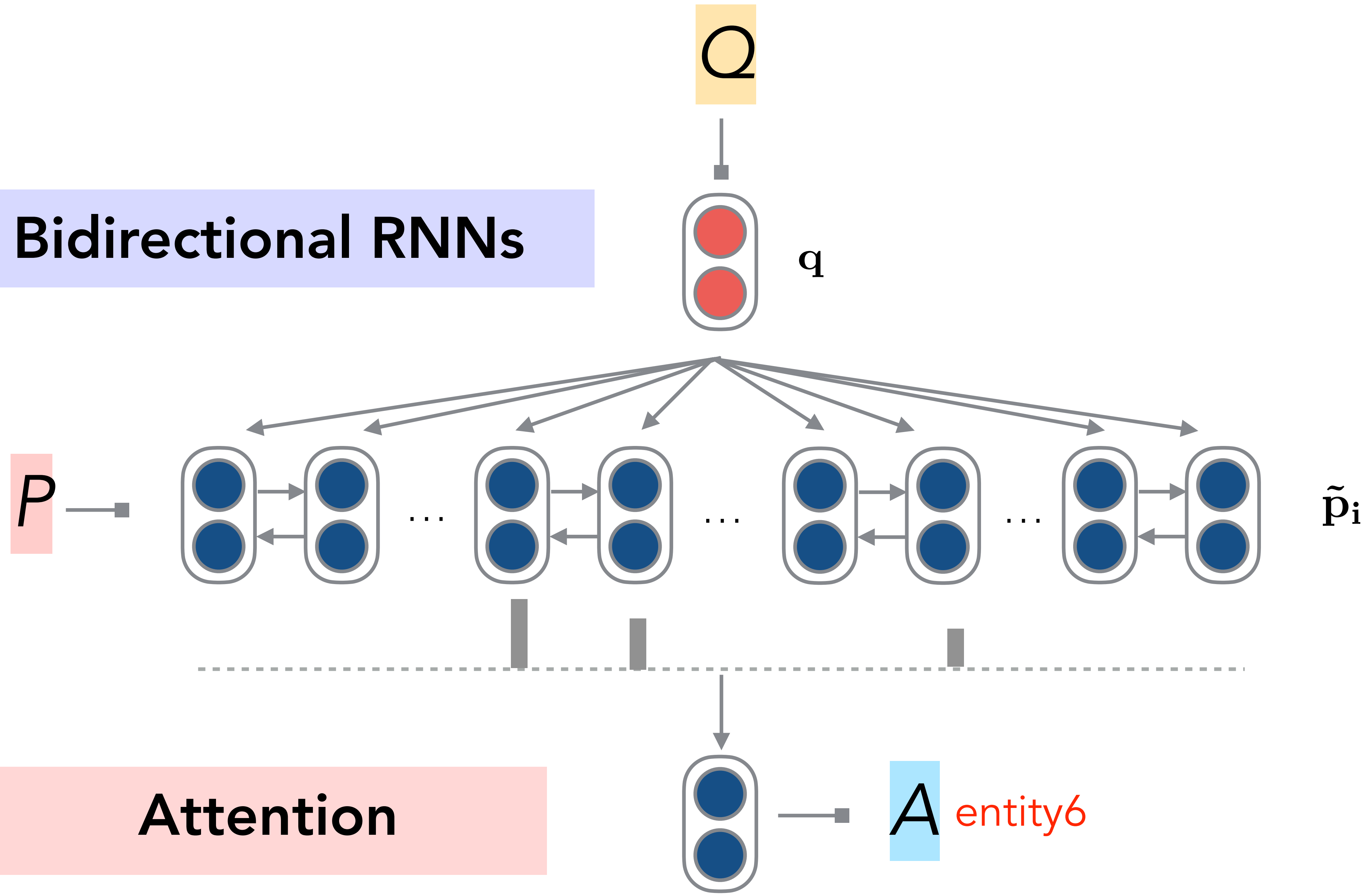
System II: End-to-end Neural Network



$$\alpha_i = \boxed{\operatorname*{softmax}_i \left(\mathbf{q}^\top \mathbf{W}_s \tilde{\mathbf{p}}_i\right)}$$

System II: End-to-end Neural Network

Bidirectional RNNs



Attention

$$\alpha_i = \boxed{\operatorname{softmax}_i \left(\mathbf{q}^\top \mathbf{W}_s \tilde{\mathbf{p}}_i\right)}$$

$$\mathbf{o} = \sum_i \alpha_i \tilde{\mathbf{p}}_i$$

System II: End-to-end Neural Network

System II: End-to-end Neural Network

- Pretty standard (popular) architecture in ACL16?

System II: End-to-end Neural Network

- Pretty standard (popular) architecture in ACL16?
- **Details**: GRU, 100d Glove, SGD, Dropout (0.2), batch size = 32, hidden size = 128 or 256….. **No magic!**

Results

- **Baselines:** (Hermann et al, 2015) (Hill et al, 2016)

Results

- **Baselines:** (Hermann et al, 2015) (Hill et al, 2016)

	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

Results

- **Baselines:** (Hermann et al, 2015) (Hill et al, 2016)

	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
NN: Attentive Reader	61.6	63.0	70.5	69.0
NN: Impatient Reader	61.8	63.8	69.0	68.0

Results

- **Baselines:** (Hermann et al, 2015) (Hill et al, 2016)

	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
NN: Attentive Reader	61.6	63.0	70.5	69.0
NN: Impatient Reader	61.8	63.8	69.0	68.0
MemNNs	63.4	66.8	N/A	N/A
MemNNs (ensemble)	66.2	69.4	N/A	N/A

Results

- **Baselines:** (Hermann et al, 2015) (Hill et al, 2016)

	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
NN: Attentive Reader	61.6	63.0	70.5	69.0
NN: Impatient Reader	61.8	63.8	69.0	68.0
MemNNs	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

Results

7-10% improvement!

- **Baselines:** (Hermann et al, 2015) (Hill et al, 2016)

	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
NN: Attentive Reader	61.6	63.0	70.5	69.0
NN: Impatient Reader	61.8	63.8	69.0	68.0
MemNNs	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	**73.8**	**73.6**	**77.6**	**76.6**

Results

7-10% improvement!

- **Baselines:** (Hermann et al, 2015) (Hill et al, 2016)

	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
NN: Attentive Reader	61.6	63.0	70.5	69.0
NN: Impatient Reader	61.8	63.8	69.0	68.0
MemNNs	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	**73.8**	**73.6**	**77.6**	**76.6**
Ours: neural net (ensemble)	**77.2**	**77.6**	**80.2**	**79.2**

*updated results / ensemble: 5 models

Results

	CNN		Daily Mail	
	Dev	Test	Dev	Test
NN: Attentive Reader	61.6	63.0	70.5	69.0
Ours: neural net	**73.8**	**73.6**	**77.6**	**76.6**

- Differences from **Attentive Reader** (Hermann et al, 2015):

Results

	CNN		Daily Mail	
	Dev	Test	Dev	Test
NN: Attentive Reader	61.6	63.0	70.5	69.0
Ours: neural net	**73.8**	**73.6**	**77.6**	**76.6**

- Differences from **Attentive Reader** (Hermann et al, 2015):
  - **Bilinear** attention
  - Remove a redundant layer before prediction
  - Predict among entities only, not all words

Results

	CNN		Daily Mail	
	Dev	Test	Dev	Test
NN: Attentive Reader	61.6	63.0	70.5	69.0
Ours: neural net	**73.8**	**73.6**	**77.6**	**76.6**

- Differences from **Attentive Reader** (Hermann et al, 2015):
  - **Bilinear** attention
  - Remove a redundant layer before prediction
  - Predict among entities only, not all words

Maybe we did better at hyper-parameter tuning? ◕‿◕

Results until 2016/8

		CNN		Daily Mail	
		Dev	Test	Dev	Test
(Hermann et al, 2015)	NIPS'15	61.8	63.8	69.0	68.0
(Hill et al, 2016)	ICLR'16	63.4	66.8	N/A	N/A
(Kobayashi et al, 2016)	NAACL'16	71.3	72.9	N/A	N/A
(Kadlec et al, 2016)	ACL'16	68.6	69.5	75.0	73.9
(Dhingra et al, 2016)	2016/6/5	73.0	73.8	76.7	75.7
(Sodorni et al, 2016)	2016/6/7	72.6	73.3	N/A	N/A
(Trischler et al, 2016)	2016/6/7	73.4	74.0	N/A	N/A
(Weissenborn, 2016)	2016/7/12	N/A	73.6	N/A	**77.2**
(Cui et al, 2016)	2016/7/15	73.1	**74.4**	N/A	N/A
Ours: neural net	ACL'16	**73.8**	73.6	**77.6**	76.6
Ours: neural net (ensemble)	ACL'16	**77.2**	**77.6**	**80.2**	**79.2**

What is this paper about?

System **Lower Bound**

Our **simple models** work quite well.

Analysis **Upper Bound**

The task might be not that hard.

We are **almost done**.

Discussion: what's next?

Our Classifier: Ablating individual features

	Accuracy
Full model	67.1
- whether e is in the passage	-0%
- whether e is in the question	-0.1%
- frequency of e	**-3.4%**
- position of e	-1.2%
- *whether e co-occurs with Q word in P.*	-1.1%
- n-gram match	**-6.6%**
- word distance	-1.7%
- dependency parse match	-1.5%

*on CNN dev set

Breakdown of the Examples

Exact match

Paraphrasing

Partial clue

Multiple sentences

Coreference errors

Ambiguous / hard

Exact Match

P … it 's clear @entity0 is leaning toward @entity60 …

Q " it 's clear @entity0 is leaning toward @placeholder ,
" says an expert who monitors @entity0

A @entity60

Paraphrasing

P

… @entity0 called me personally to let me know that he would n't be playing here at @entity23 , " @entity3 said …

Q

@placeholder says he understands why @entity0 wo n't play at his tournament

A

@entity3

Partial Clue

P @entity12 " @entity2 professed that his " @entity11 " is not a religious book . …

Q a tv movie based on @entity2 's book " @placeholder " casts a @entity76 actor as @entity5

A @entity11

Multiple sentences

P

… " we got some groundbreaking performances , here too , tonight , " @entity6 said . " we got @entity17 , who will be doing some musical performances . he 's doing a his - and - her duet all by himself . "…

Q

" he 's doing a his - and - her duet all by himself , " @entity6 said of @placeholder

A

@entity17

Coreference Error

P … hip - hop star @entity246 saying on @entity247 that he was canceling an upcoming show for the @entity249 …

Q rapper @placeholder " disgusted , " cancels upcoming show for @entity280

@entity280 = @entity249 = SAEs

A @entity246

Ambiguous / Hard

P

… a small aircraft carrying @entity5 , @entity6 and @entity7 " the @entity12 " @entity3 crashed …

Q

pilot error and snow were reasons stated for @placeholder plane crash

A

@entity5

Breakdown of the Examples

Exact match

Paraphrasing

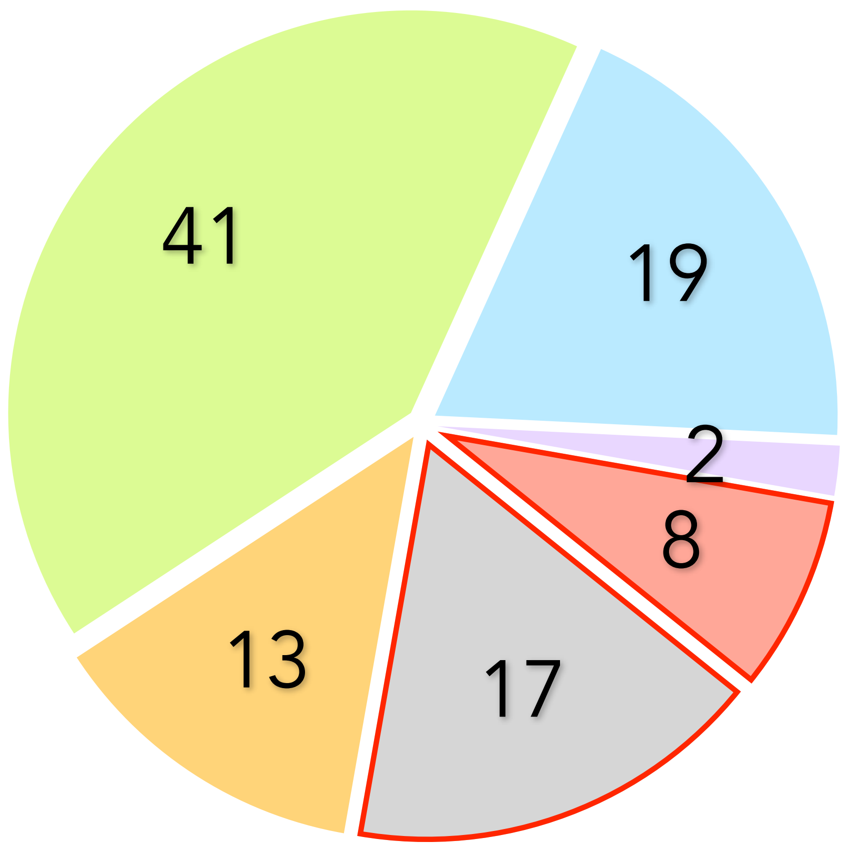
Partial clue

Multiple sentences

Coreference errors

Ambiguous / hard

CNN: 100 samples



Breakdown of the Examples

Exact match

Paraphrasing

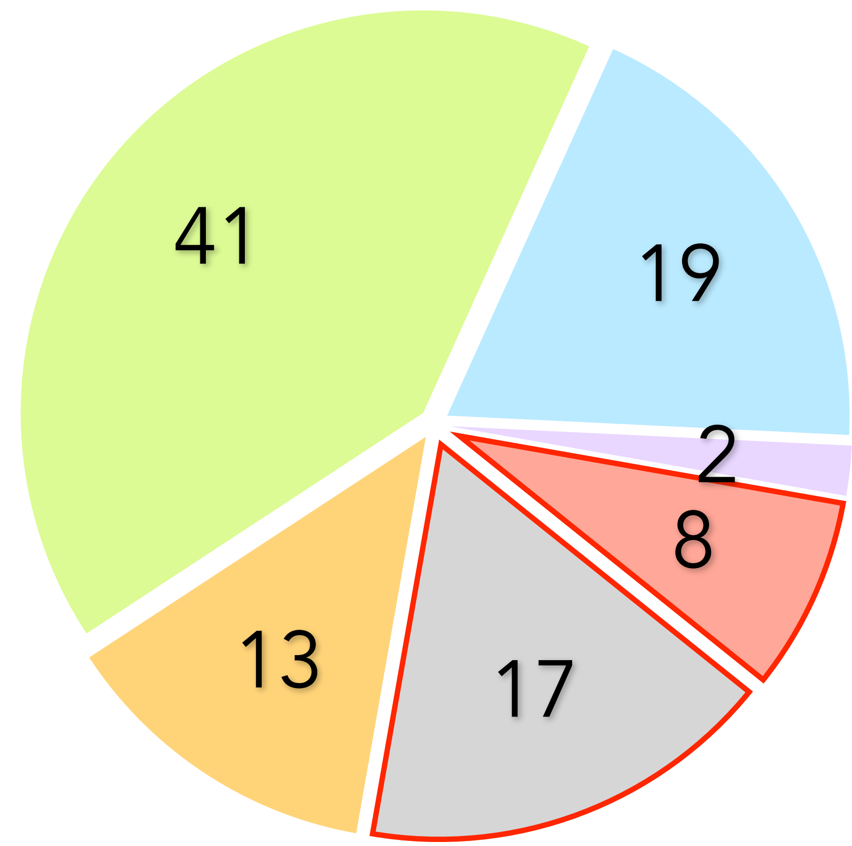
Partial clue

Multiple sentences

Coreference errors

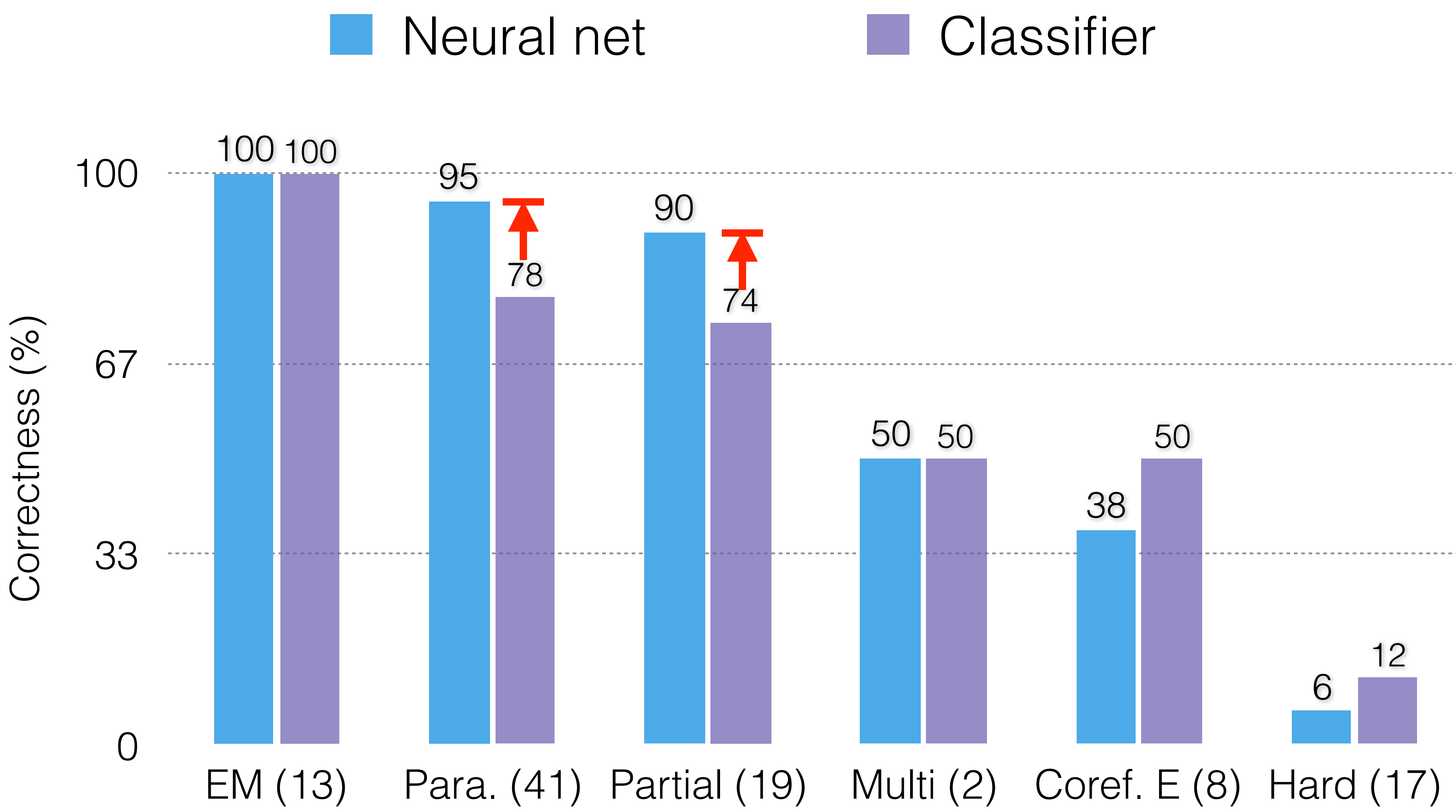
Ambiguous / hard

CNN: 100 samples

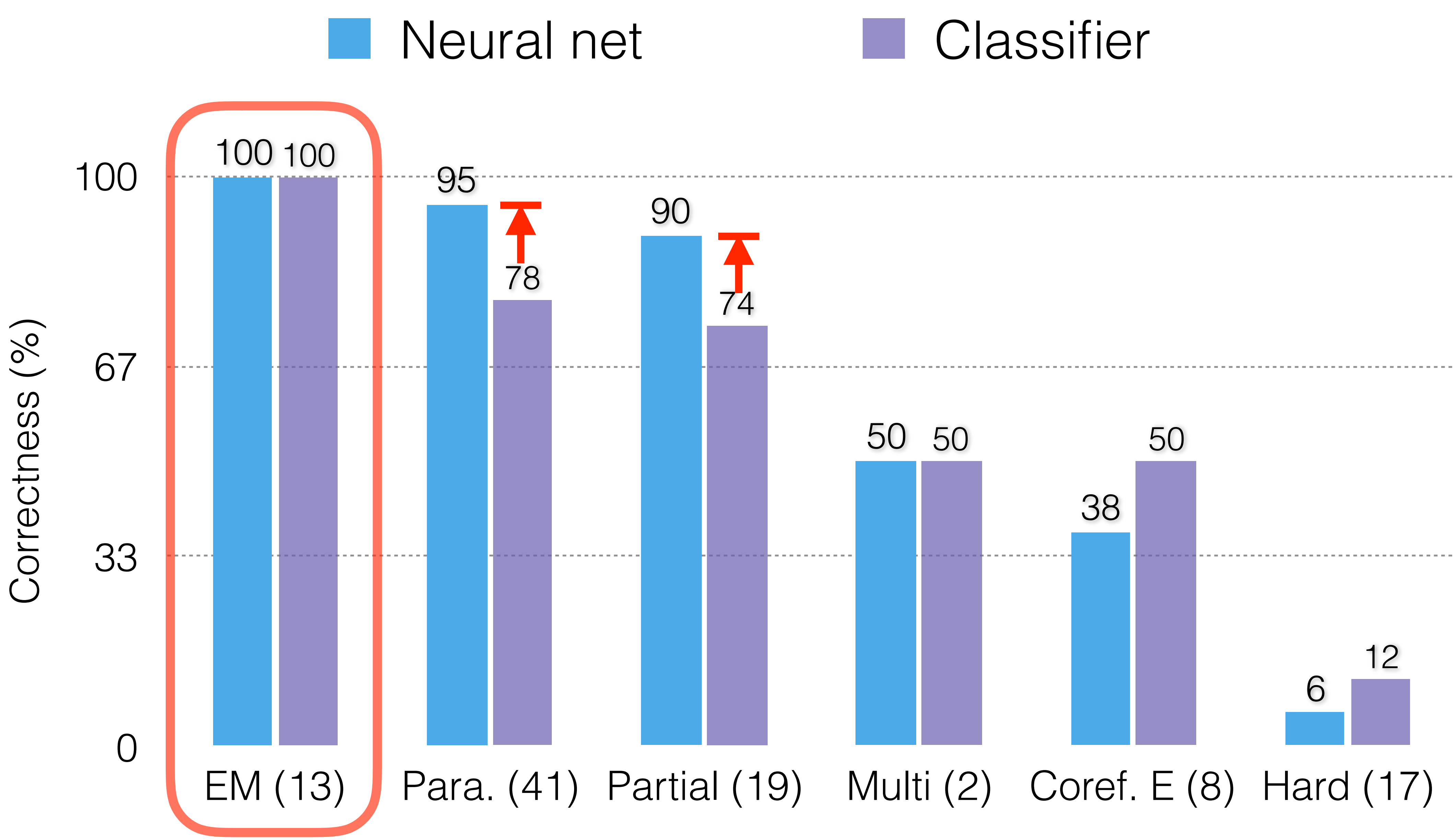


neural net	**73.8**	73.6
neural net (ensemble)	**77.2**	**77.6**

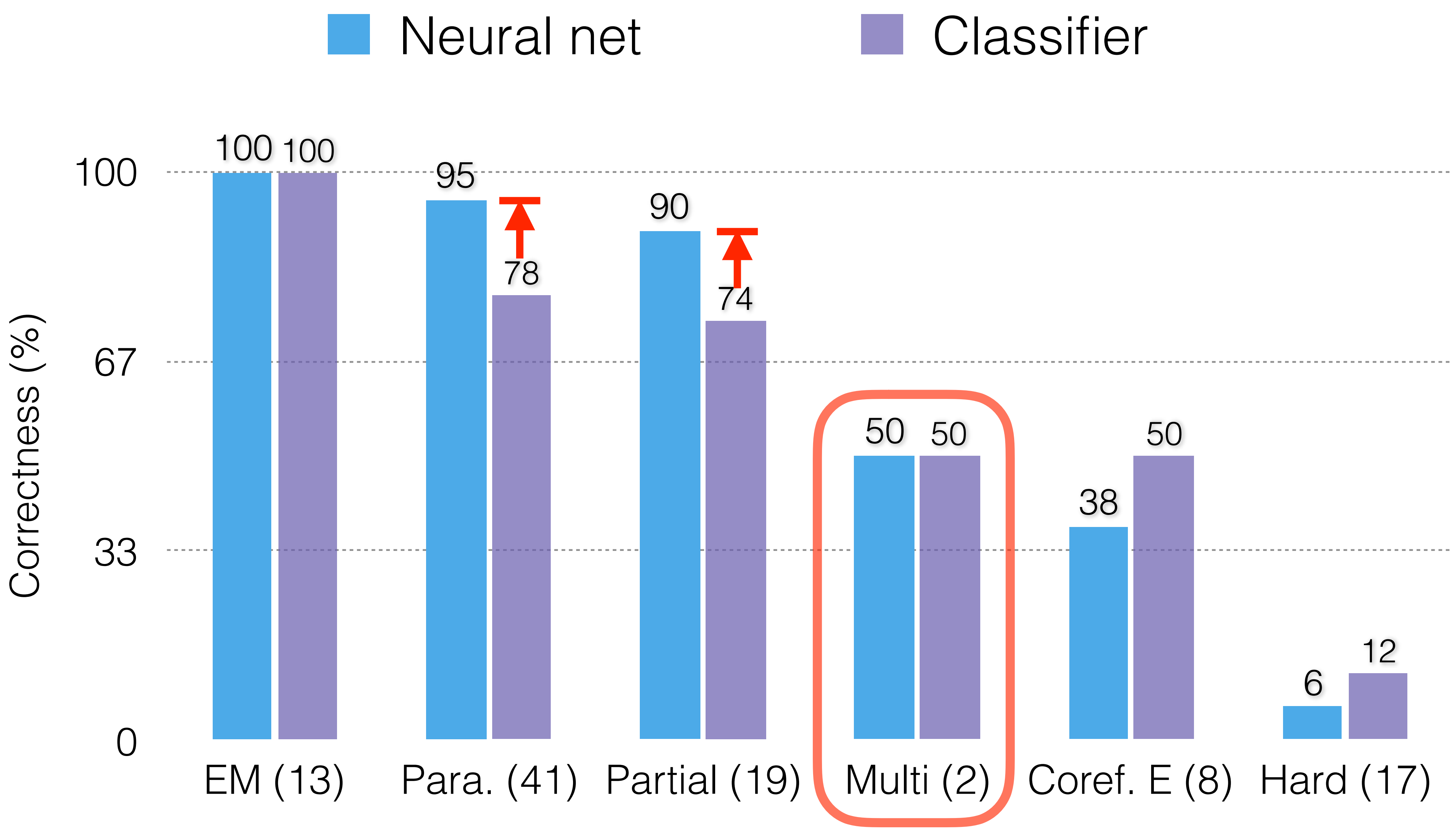
Per-category Accuracies



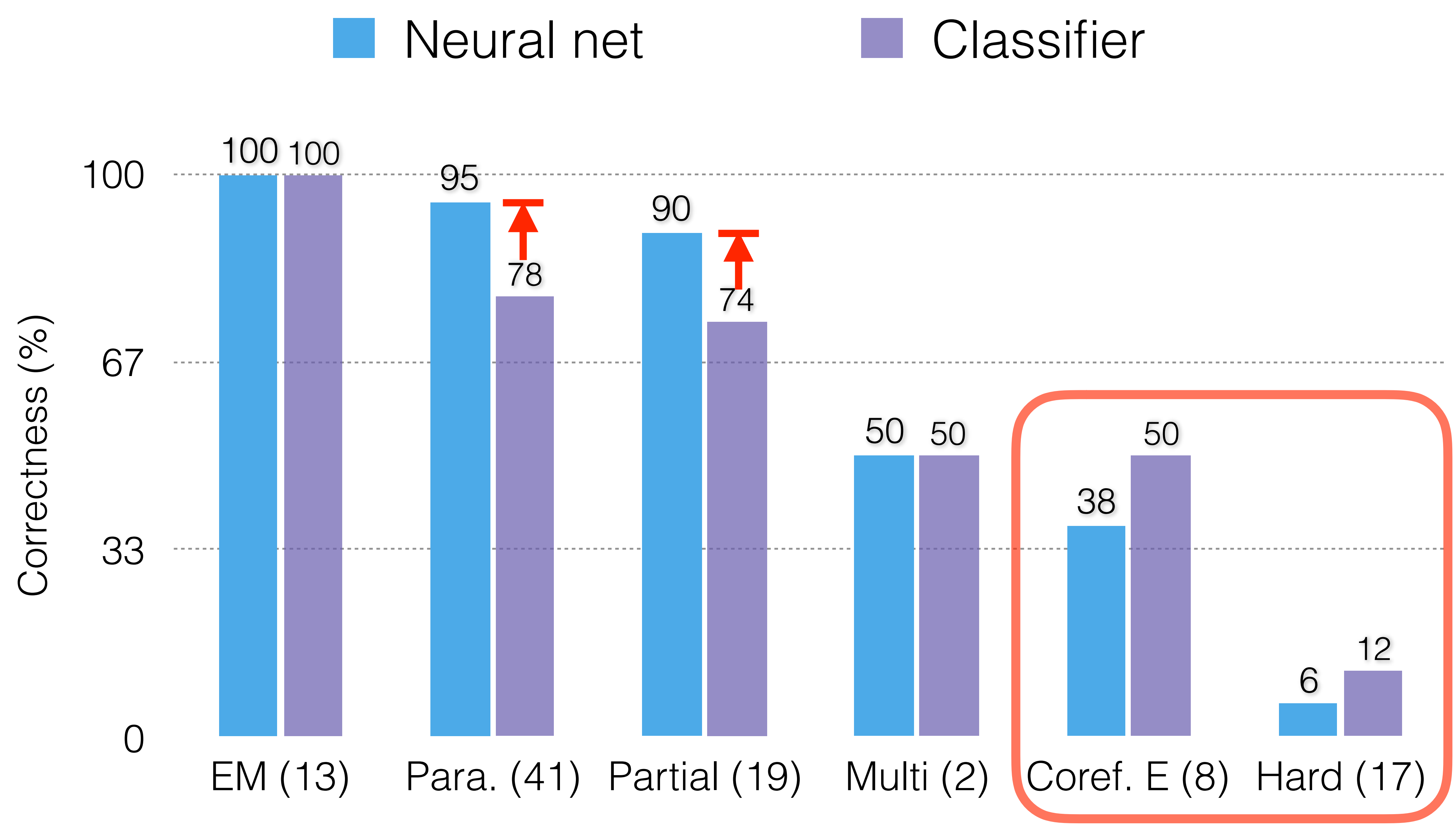
Per-category Accuracies



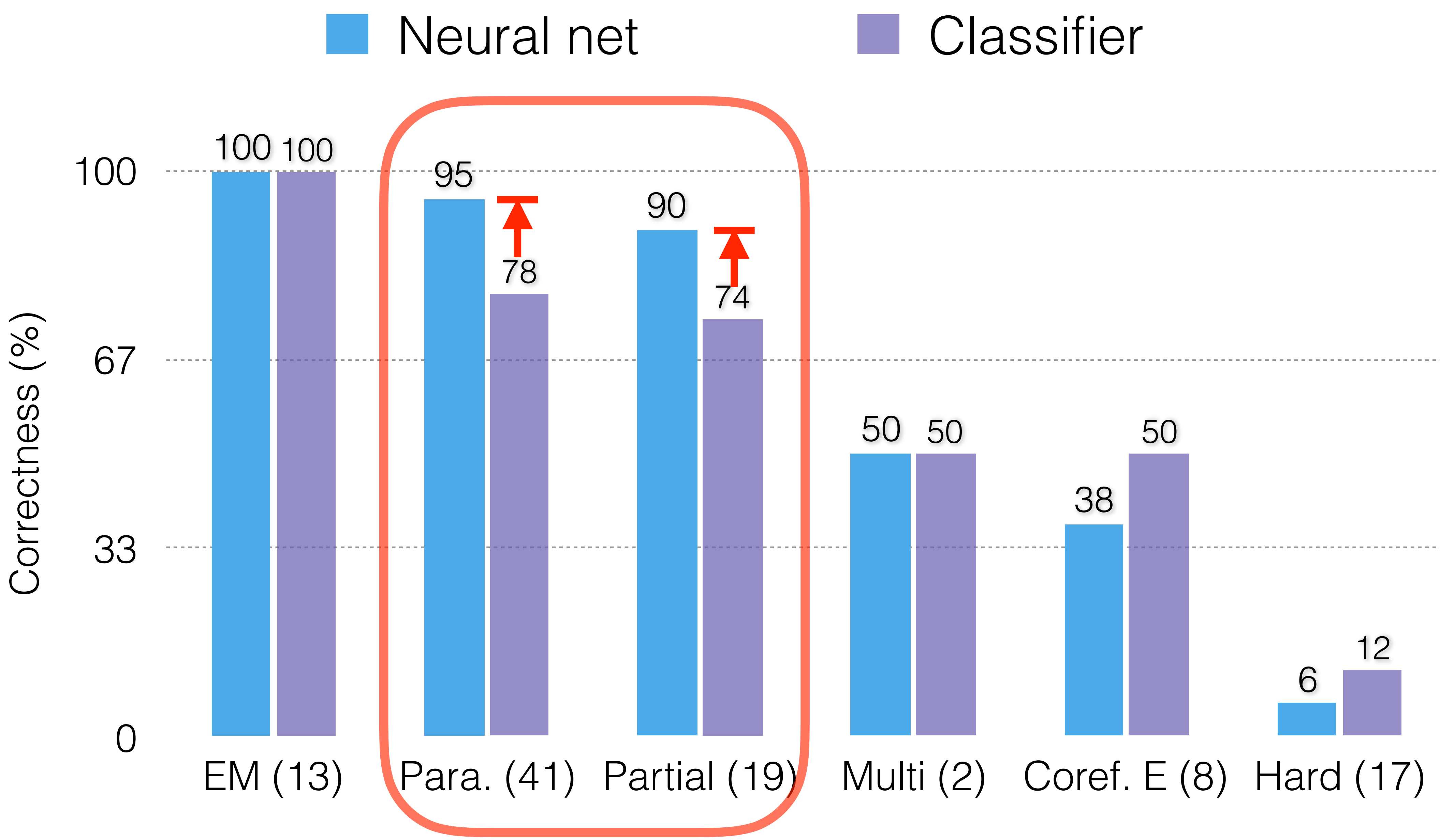
Per-category Accuracies



Per-category Accuracies



Per-category Accuracies



Conclusions

Conclusions

- Reminder: **Simple models** sometimes just work; **neural nets are great** for learning semantic matches.

Conclusions

- Reminder: **Simple models** sometimes just work; **neural nets are great** for learning semantic matches.
- **The CNN/Daily Mail task:** large but still noisy, and we almost have hit the capacity; not hard enough for reasoning and inference.

Conclusions

- Reminder: **Simple models** sometimes just work; **neural nets are great** for learning semantic matches.
- **The CNN/Daily Mail task:** large but still noisy, and we almost have hit the capacity; not hard enough for reasoning and inference.
- **Future:**

Conclusions

- Reminder: **Simple models** sometimes just work; **neural nets are great** for learning semantic matches.
- **The CNN/Daily Mail task:** large but still noisy, and we almost have hit the capacity; not hard enough for reasoning and inference.
- **Future:**
  - Leverage these datasets to solve more realistic RC tasks!

Conclusions

- Reminder: **Simple models** sometimes just work; **neural nets are great** for learning semantic matches.
- **The CNN/Daily Mail task:** large but still noisy, and we almost have hit the capacity; not hard enough for reasoning and inference.
- **Future:**
  - Leverage these datasets to solve more realistic RC tasks!
  - Complex models?

Conclusions

- Reminder: **Simple models** sometimes just work; **neural nets are great** for learning semantic matches.
- **The CNN/Daily Mail task:** large but still noisy, and we almost have hit the capacity; not hard enough for reasoning and inference.
- **Future:**
  - Leverage these datasets to solve more realistic RC tasks!
  - Complex models?
  - More datasets coming up: WikiReading, LAMBADA, SQuAD..

Conclusions

- Reminder: **Simple models** sometimes just work; **neural nets are great** for learning semantic matches.
- **The CNN/Daily Mail task:** large but still noisy, and we almost have hit the capacity; not hard enough for reasoning and inference.
- **Future:**
  - Leverage these datasets to solve more realistic RC tasks!
  - Complex models?
  - More datasets coming up: WikiReading, LAMBADA, SQuAD..

It is an exciting time for **reading comprehension**!

Code available at

https://github.com/danqi/rc-cnn-dailymail

Thanks!
Questions?