Understanding natural text with neural networks

Rudolf Kadlec, Martin Schmid, Ondřej Bajgar and Jan Kleindienst

IBM Watson, Prague

http://arxiv.org/abs/1603.01547

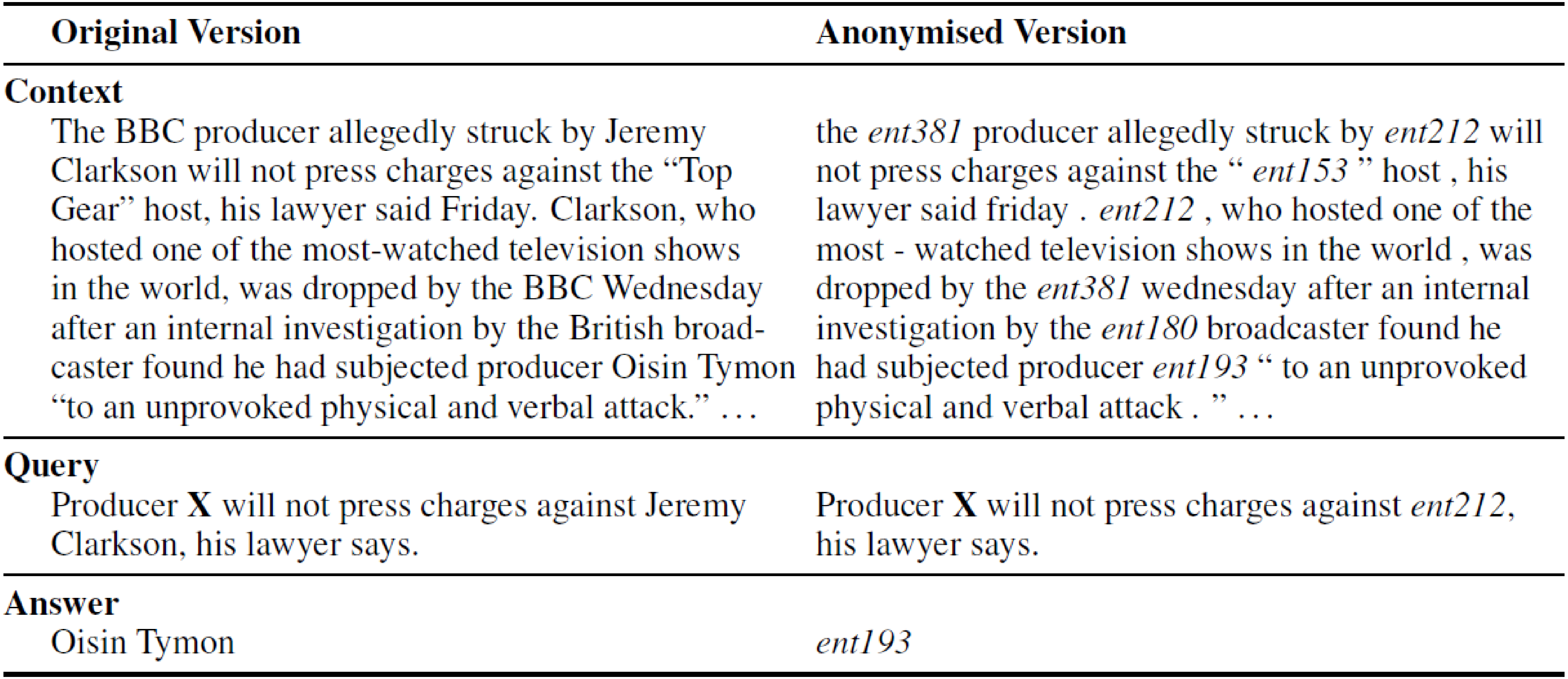
Motivation

- Unstructured text is rich source of information
  - As demonstrated by IBM Watson Jeopardy system
- Watson did not use any deep learning
- Here we show how to apply NNs to text understanding

- How to test text comprehension?
- Let the NN read an article and then ask questions about it
  - Cloze style questions can be used to generate these „questions“ automatically

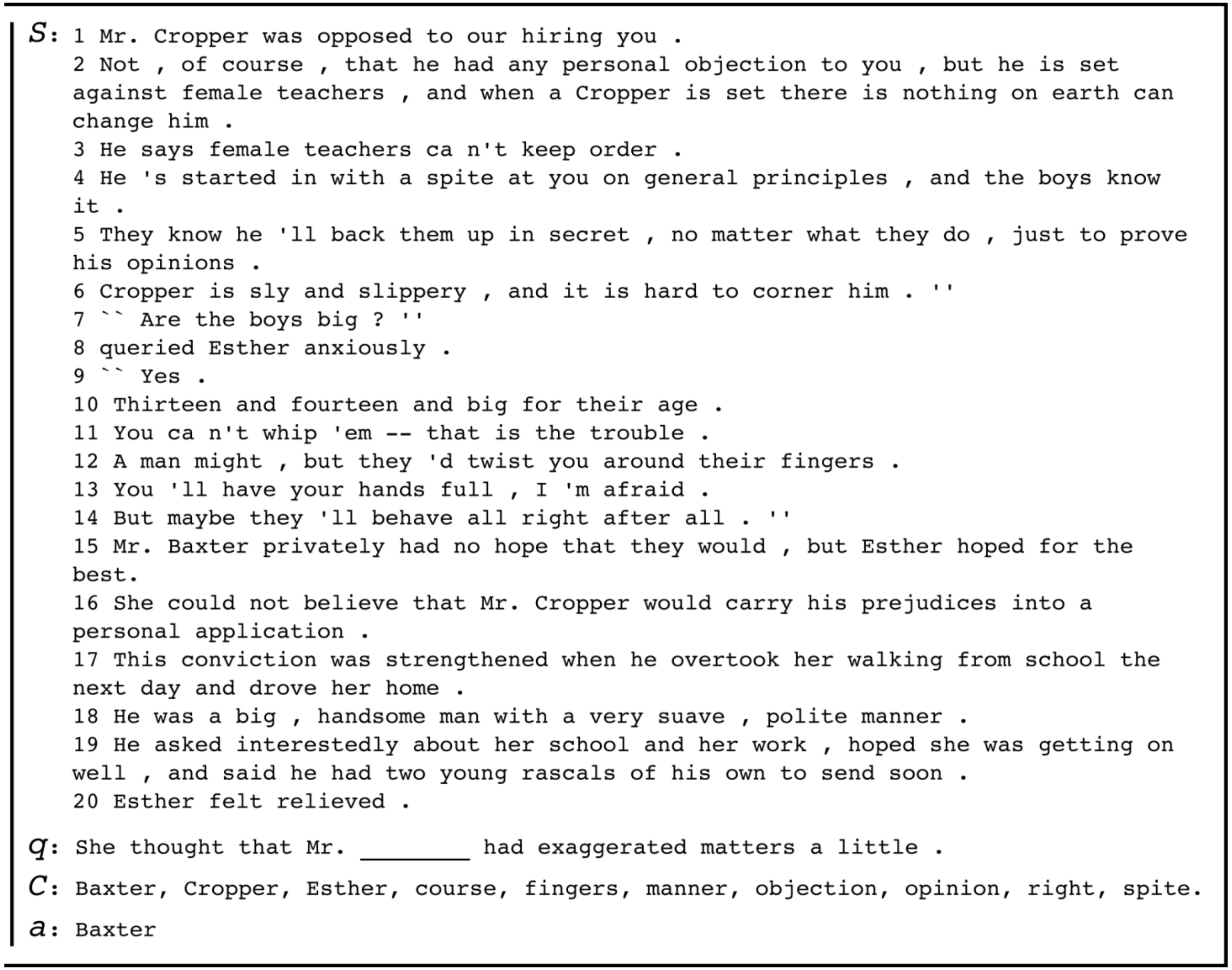
Datasets

CNN and Daily Mail (DeepMind)



	Original Version	Anonymised Version
Context		
	The BBC producer allegedly struck by Jeremy Clarkson will not press charges against the “Top Gear” host, his lawyer said Friday. Clarkson, who hosted one of the most-watched television shows in the world, was dropped by the BBC Wednesday after an internal investigation by the British broadcaster found he had subjected producer Oisin Tymon “to an unprovoked physical and verbal attack.” …	the *ent381* producer allegedly struck by *ent212* will not press charges against the “ *ent153* ” host , his lawyer said friday . *ent212* , who hosted one of the most - watched television shows in the world , was dropped by the *ent381* wednesday after an internal investigation by the *ent180* broadcaster found he had subjected producer *ent193* “ to an unprovoked physical and verbal attack . ” …
Query		
	Producer **X** will not press charges against Jeremy Clarkson, his lawyer says.	Producer **X** will not press charges against *ent212*, his lawyer says.
Answer		
	Oisin Tymon	*ent193*

Children Book Test (Facebook AI)



S: 1 Mr. Cropper was opposed to our hiring you .
2 Not , of course , that he had any personal objection to you , but he is set against female teachers , and when a Cropper is set there is nothing on earth can change him .
3 He says female teachers ca n't keep order .
4 He 's started in with a spite at you on general principles , and the boys know it .
5 They know he 'll back them up in secret , no matter what they do , just to prove his opinions .
6 Cropper is sly and slippery , and it is hard to corner him . ''
7 `` Are the boys big ? ''
8 queried Esther anxiously .
9 `` Yes .
10 Thirteen and fourteen and big for their age .
11 You ca n't whip 'em -- that is the trouble .
12 A man might , but they 'd twist you around their fingers .
13 You 'll have your hands full , I 'm afraid .
14 But maybe they 'll behave all right after all . ''
15 Mr. Baxter privately had no hope that they would , but Esther hoped for the best.
16 She could not believe that Mr. Cropper would carry his prejudices into a personal application .
17 This conviction was strengthened when he overtook her walking from school the next day and drove her home .
18 He was a big , handsome man with a very suave , polite manner .
19 He asked interestedly 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 .
20 Esther felt relieved .

q: She thought that Mr. ________ had exaggerated matters a little .

C: Baxter, Cropper, Esther, course, fingers, manner, objection, opinion, right, spite.

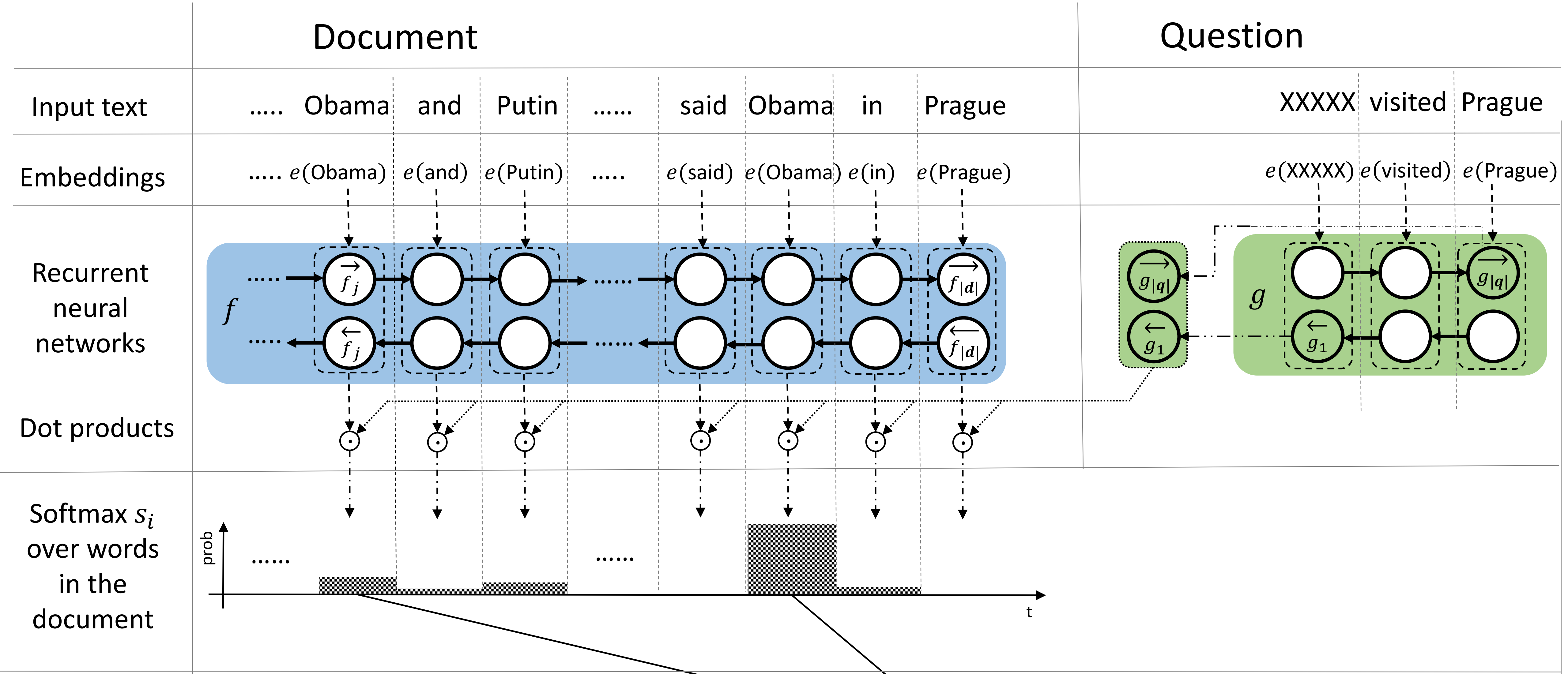
a: Baxter

	CNN			**Daily Mail**			**CBT CN**			**CBT NE**		
	train	valid	test	train	valid	test	train	valid	test	train	valid	test
# queries	380,298	3,924	3,198	879,450	64,835	53,182	120,769	2,000	2,500	108,719	2,000	2,500
Max # options	527	187	396	371	232	245	10	10	10	10	10	10
Avg # options	26.4	26.5	24.5	26.5	25.5	26.0	10	10	10	10	10	10
Avg # tokens	762	763	716	813	774	780	470	448	461	433	412	424
Vocab size	118,497			208,045			53,185			53,063		

Table 1: Statistics on the 4 data sets used to evaluate the model. CBT CN stands for CBT Common Nouns and CBT NE stands for CBT Named Entites. CBT had a fixed number of 10 options for answering each question. Statistics were taken from (Hermann et al., 2015) and the statistics provided with the CBT data set.

Attention Sum Reader

Our NN architecture


Document
Question
Input text
..... Obama and Putin ...... said Obama in Prague
XXXXX visited Prague
Embeddings
..... e(Obama) e(and) e(Putin) ..... e(said) e(Obama) e(in) e(Prague)
e(XXXXX) e(visited) e(Prague)
Recurrent neural networks
f
g
$\overrightarrow{f_j}$
$\overleftarrow{f_j}$
$\overrightarrow{f_{|d|}}$
$\overleftarrow{f_{|d|}}$
$\overrightarrow{g_{|q|}}$
$\overleftarrow{g_1}$
Dot products
Softmax s_i over words in the document
prob
t
......

Probability of the answer

$$\mathrm{P(Obama|q,d)} = \sum_{i\in I(Obama,\mathrm{d})} s_i = s_j + s_{j+5}$$

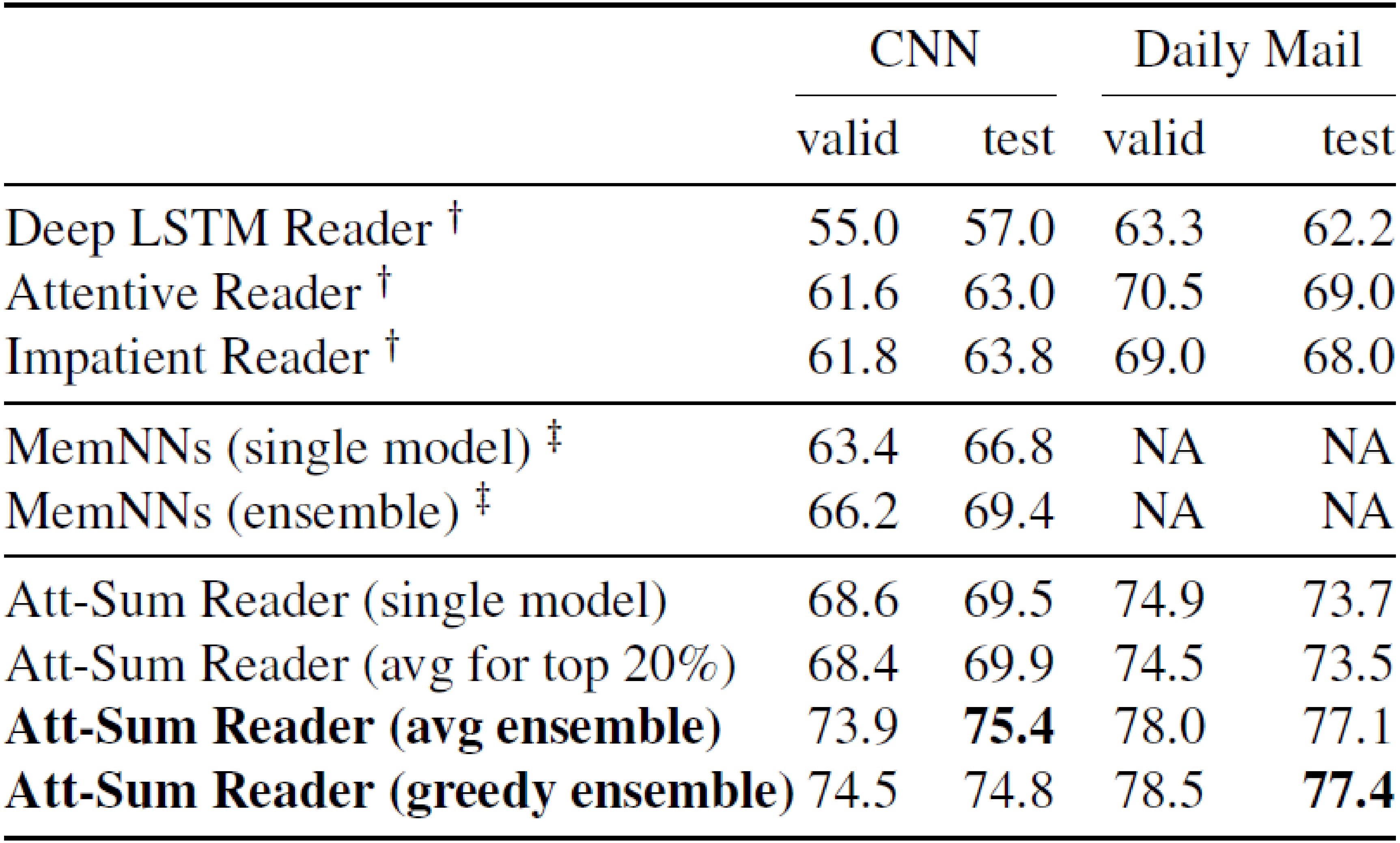
Results

CNN and Daily Mail dataset

(Hermann et al., 2015)

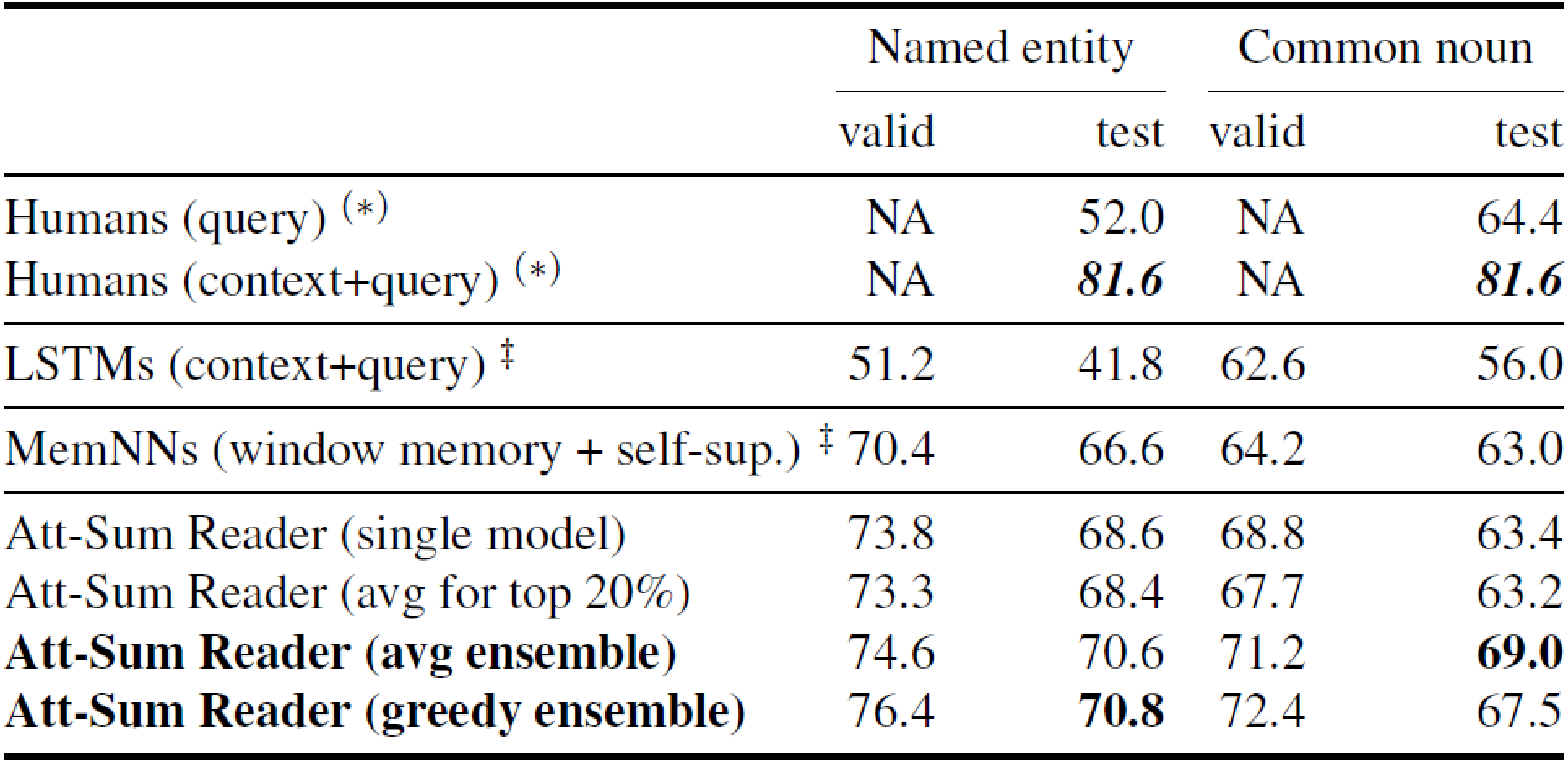
(Hill et al., 2016)

(Kadlec et al., 2016)



	CNN		Daily Mail	
	valid	test	valid	test
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 (single model) ‡	63.4	66.8	NA	NA
MemNNs (ensemble) ‡	66.2	69.4	NA	NA
Att-Sum Reader (single model)	68.6	69.5	74.9	73.7
Att-Sum Reader (avg for top 20%)	68.4	69.9	74.5	73.5
Att-Sum Reader (avg ensemble)	73.9	**75.4**	78.0	77.1
Att-Sum Reader (greedy ensemble)	74.5	74.8	78.5	**77.4**

Children Book Test



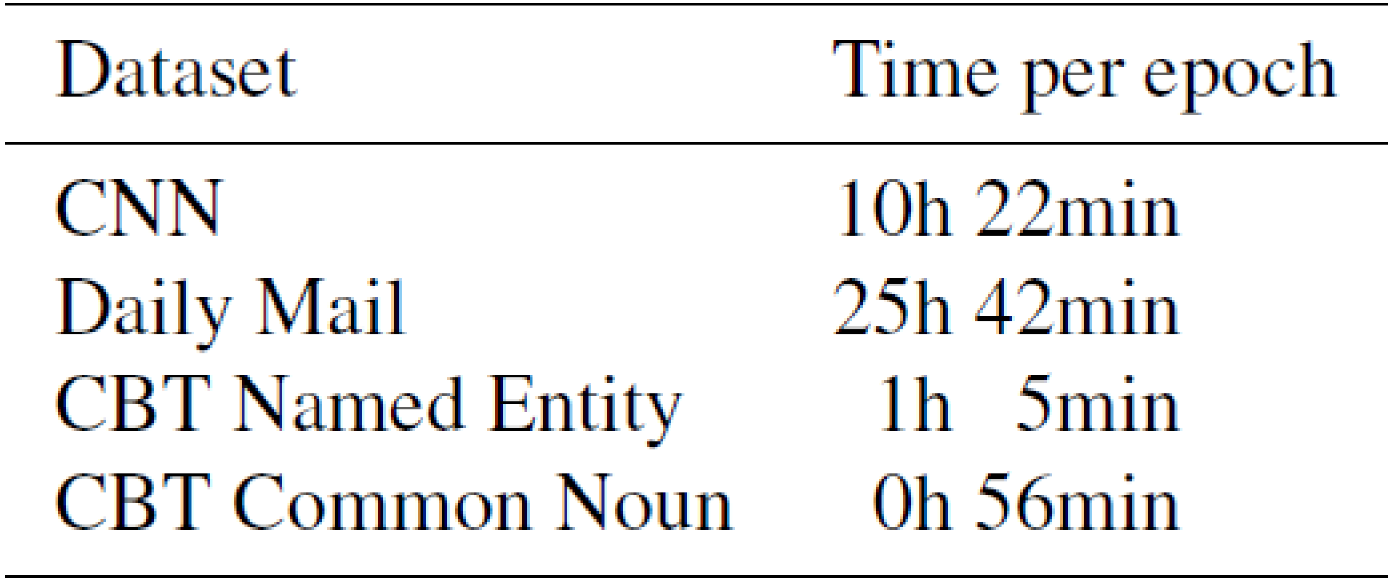
	Named entity		Common noun	
	valid	test	valid	test
Humans (query) $^{(*)}$	NA	52.0	NA	64.4
Humans (context+query) $^{(*)}$	NA	***81.6***	NA	***81.6***
LSTMs (context+query) ‡	51.2	41.8	62.6	56.0
MemNNs (window memory + self-sup.) ‡	70.4	66.6	64.2	63.0
Att-Sum Reader (single model)	73.8	68.6	68.8	63.4
Att-Sum Reader (avg for top 20%)	73.3	68.4	67.7	63.2
Att-Sum Reader (avg ensemble)	74.6	70.6	71.2	**69.0**
Att-Sum Reader (greedy ensemble)	76.4	**70.8**	72.4	67.5

(Hill et al., 2016)

(Kadlec et al., 2016)

Training times

- We used Nvidia K40 GPUs
- Models converged after 2 epochs



Dataset	Time per epoch
CNN	10h 22min
Daily Mail	25h 42min
CBT Named Entity	1h 5min
CBT Common Noun	0h 56min

Analysis

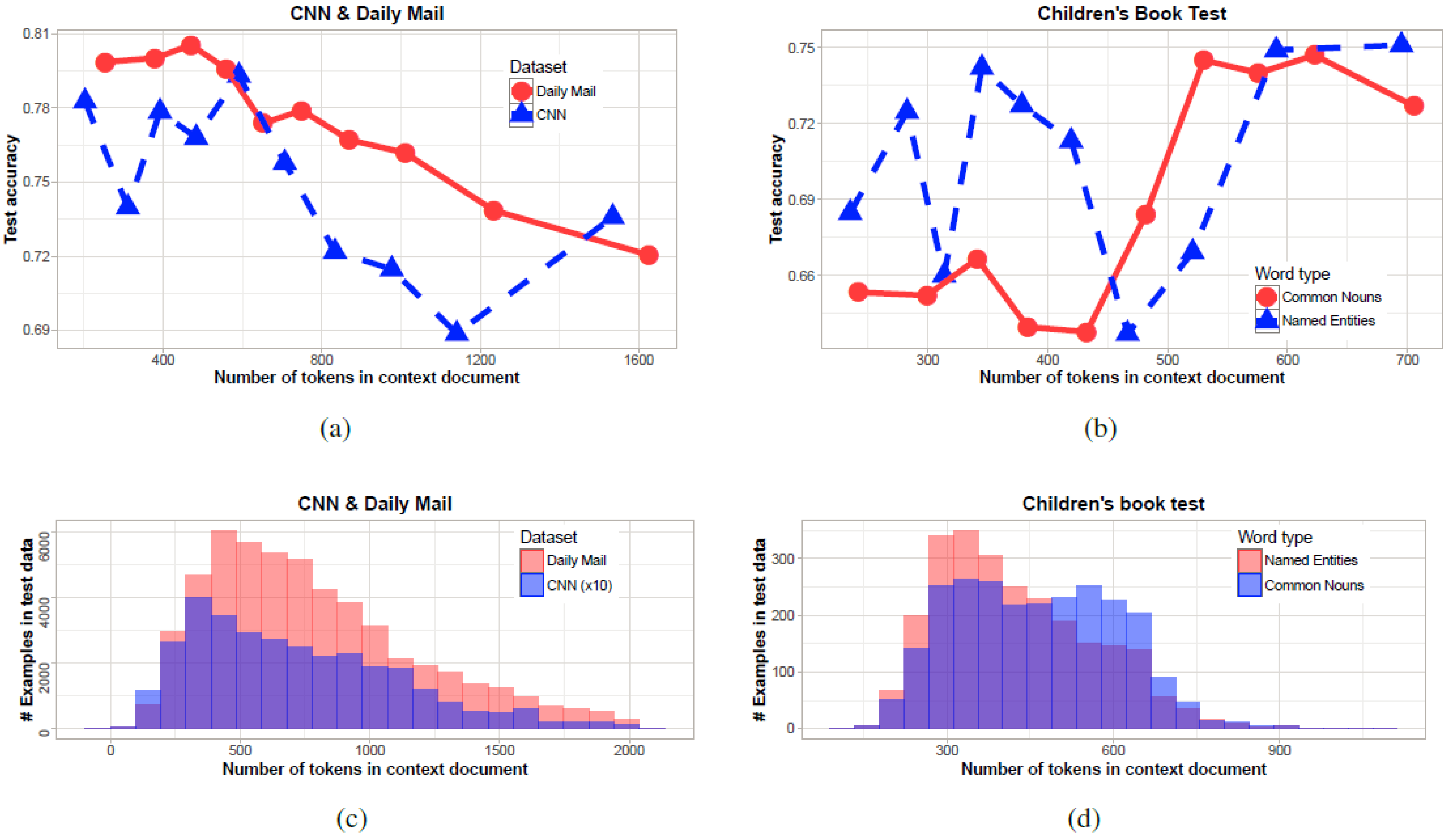
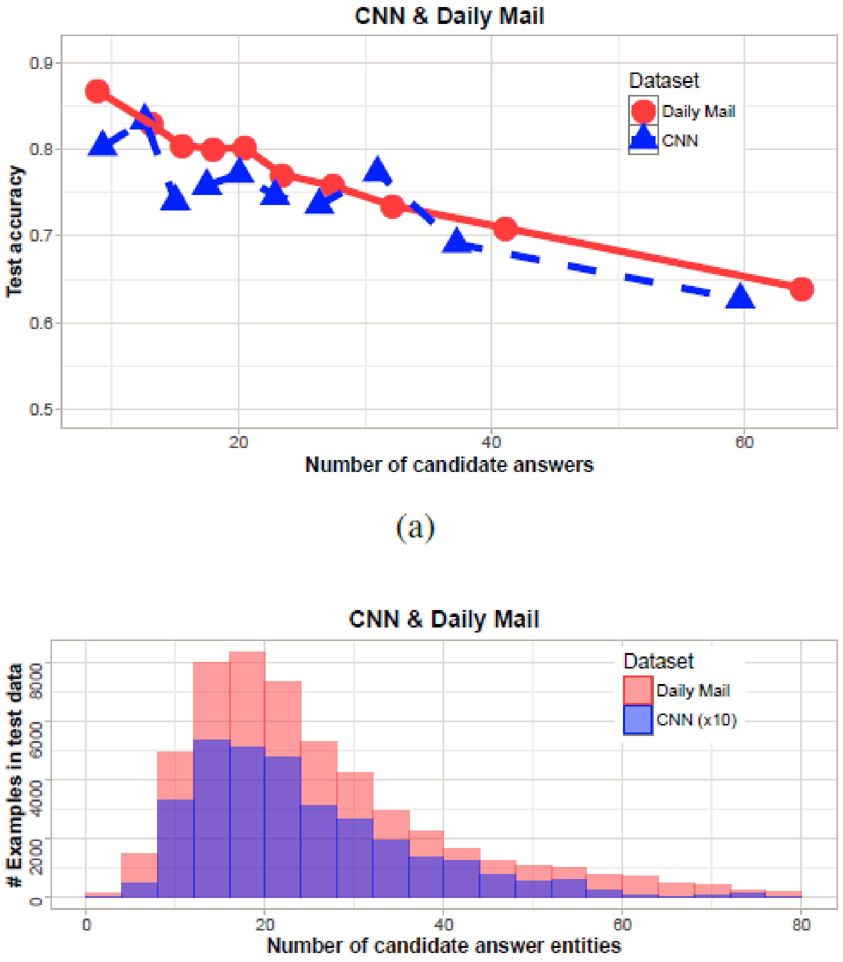


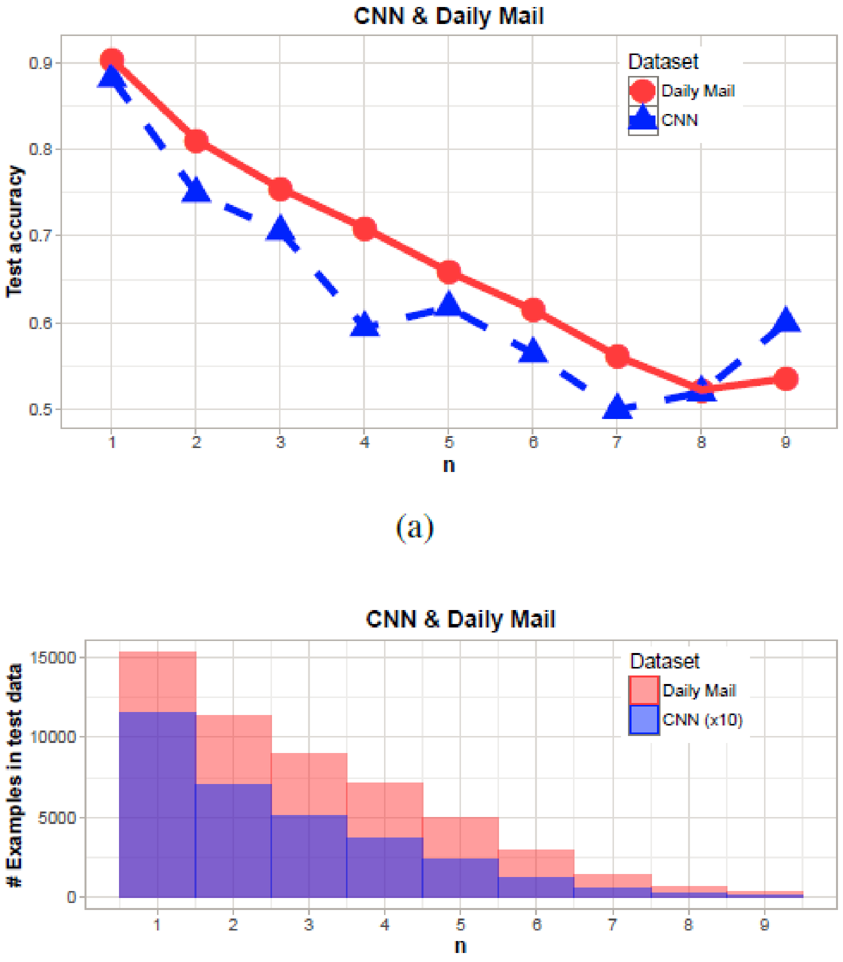
Figure 2: Sub-figures (a) and (b) plot the test accuracy against the length of the context document (for CNN the count was multiplied by 10). The examples were split into ten buckets of equal size by their context length. Averages for each bucket are plotted on each axis. Sub-figures (c) and (d) show distributions of context lengths in the four datasets. The number of examples was multiplied by 10 for the CNN dataset.



(a)

(b)

Figure 3: Subfigure (a) illustrates how the model accuracy decreases with an increasing number of candidate named entities. Subfigure (b) shows the overall distribution of the number of candidate answers in the news datasets. The number of examples was multiplied by 10 for the CNN dataset.

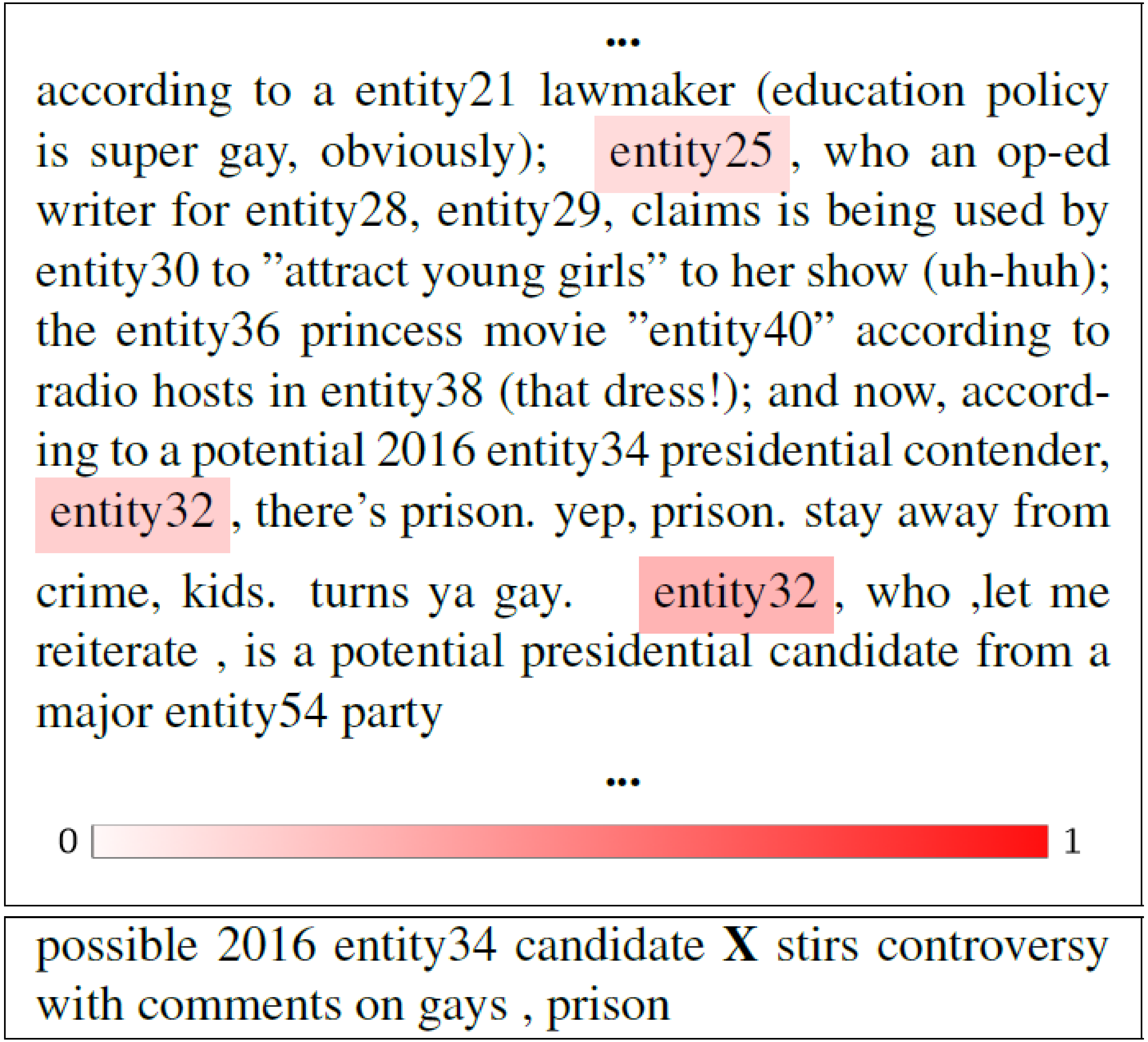


(a)

(b)

Figure 4: Subfigure (a) shows the model accuracy when the correct answer is among n most frequent named entities for $n \in [1, 10]$. Subfigure (b) shows the number of test examples for which the correct answer was the n–th most frequent entity. The number of examples was multiplied by 10 for the CNN dataset.

Example



...

according to a entity21 lawmaker (education policy is super gay, obviously); entity25 , who an op-ed writer for entity28, entity29, claims is being used by entity30 to ”attract young girls” to her show (uh-huh); the entity36 princess movie ”entity40” according to radio hosts in entity38 (that dress!); and now, according to a potential 2016 entity34 presidential contender, entity32 , there’s prison. yep, prison. stay away from crime, kids. turns ya gay. entity32 , who ,let me reiterate , is a potential presidential candidate from a major entity54 party

...

possible 2016 entity34 candidate **X** stirs controversy with comments on gays , prison

References

- DeepMind’s model:
  - Hermann, K. M., Kočiský, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M., & Blunsom, P. (2015). Teaching Machines to Read and Comprehend. *NIPS*.
- Facebook’s model:
  - Hill, F., Bordes, A., Chopra, S., & Weston, J. (2016). The Goldilocks Principle: Reading Children’s Books with Explicit Memory Representations. *ICLR*.
- IBM’s model
  - Kadlec, R., Schmid, M., Bajgar, O., & Kleindienst, J. (2016). Neural Text Understanding with Attention Sum Reader. http://arxiv.org/abs/1603.01547

IBM Watson

Improved Deep Learning Baselines for Ubuntu Corpus Dialogs

Rudolf Kadlec

Martin Schmid

Jan Kleindienst

IBM

$$P(\text{machine response} \mid \text{dialog context})$$

Some available datasets

- bAbI
  - “Unit tests”
    - Weston, J., Bordes, A., Chopra, S., & Mikolov, T. (2014). Towards AI-Complete Question Answering: A Set of Prerequisite Toy Tasks.
- Movie subtitles
  - Chit chat
- Ubuntu Dialog Corpus
  - IT support
    - Lowe, R. et. al. (2015): The Ubuntu Dialogue Corpus: A Large Dataset for Research in Unstructured Multi-Turn Dialogue Systems. *SIGdial*.

Our contribution

- We improved performance on utterance ranking task in Ubuntu Dialog Corpus compared to Lowe et al. 2015

- Huge ensemble of diverse models does the job

=> now we can test more interesting models

Ubuntu IRC Chats

nik90	bzoltan_: looks like you are getting a lot of test results for your static chroots :P	11:55
bzoltan_	nik90: thanks to you :)	11:55
nik90	zsombi: hey, I tested your alarms fix on vivid btw, and commented on the bug.	11:55
zsombi	nik90: thx!	11:56
nik90	bzoltan_: well hopefully you found it useful. Looking forward to minimized chroot creation time :)	11:56
zsombi	nik90: about the crashes, please check if you are using teh proper alarm object when creating the alarm. If you call reset() on an object which comes from the model you may get in trouble	11:57
nik90	zsombi: hmm let me check	11:58
nik90	zsombi: I don't seem to have any reset() function calls at all..I searched the entire project	12:09

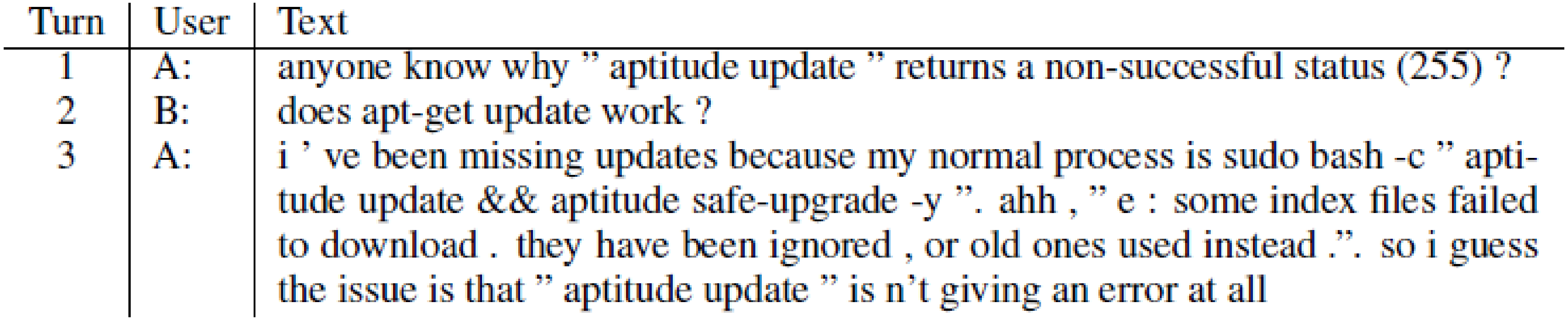
Trivia about Ubuntu Dialog Corpus

- 930000 Human-human dialogs
- First public problem solving dataset of this size
- The goal
  - Learn how to automatically respond
- How to measure quality of a predictive model for dialog:
  - Sampling the next utterance word by word from seq2seq models usually leads to low BLEU scores
  - Dialog is noisy
  - Use ranking instead of BLEU score

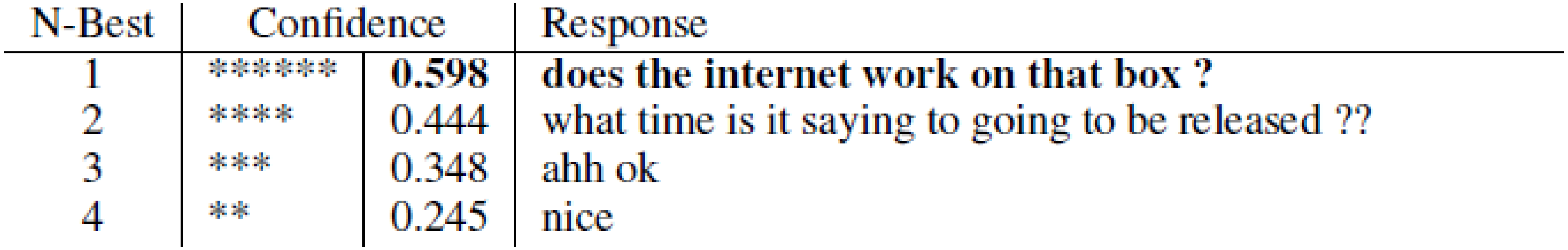
Lowe et. al. 2015: The Ubuntu Dialogue Corpus: A Large Dataset for Research in Unstructured Multi-Turn Dialogue Systems. *SIGdial*.

Task

- Given N utterances of a dialog context rank candidates for (N+1) th utterance



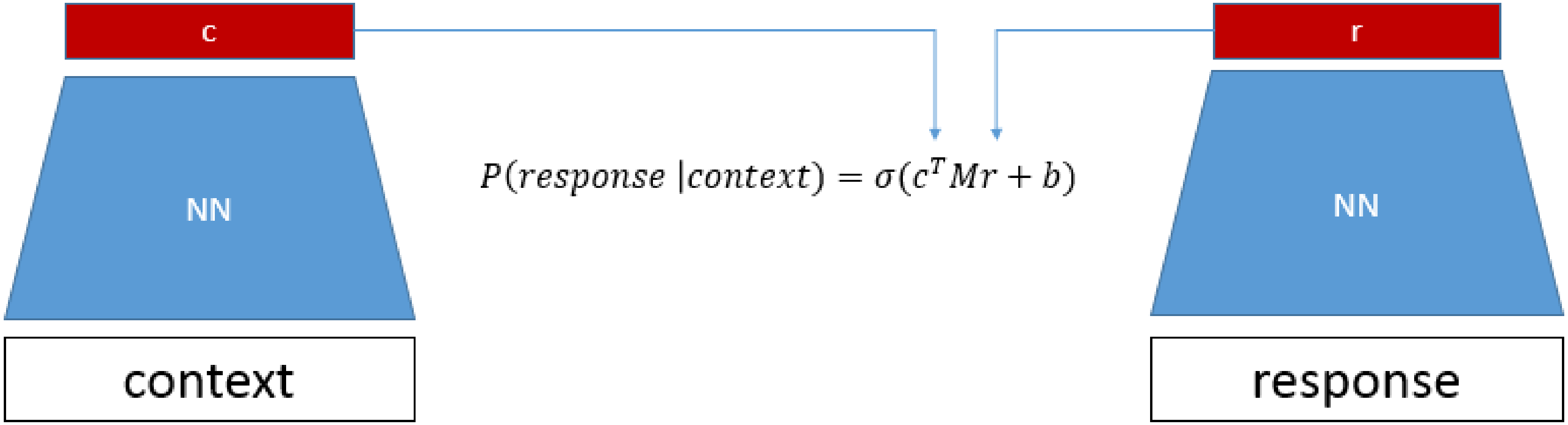
Turn	User	Text
1	A:	anyone know why ” aptitude update ” returns a non-successful status (255) ?
2	B:	does apt-get update work ?
3	A:	i ’ ve been missing updates because my normal process is sudo bash -c ” aptitude update && aptitude safe-upgrade -y ”. ahh , ” e : some index files failed to download . they have been ignored , or old ones used instead .”. so i guess the issue is that ” aptitude update ” is n’t giving an error at all



N-Best	Confidence		Response
1	******	**0.598**	**does the internet work on that box ?**
2	****	0.444	what time is it saying to going to be released ??
3	***	0.348	ahh ok
4	**	0.245	nice

Machine learning model

- Twin network architecture
- Input
  - Context
  - Response



$$P(response \mid context) = \sigma(c^T M r + b)$$

Training

- Training with positive and negative examples

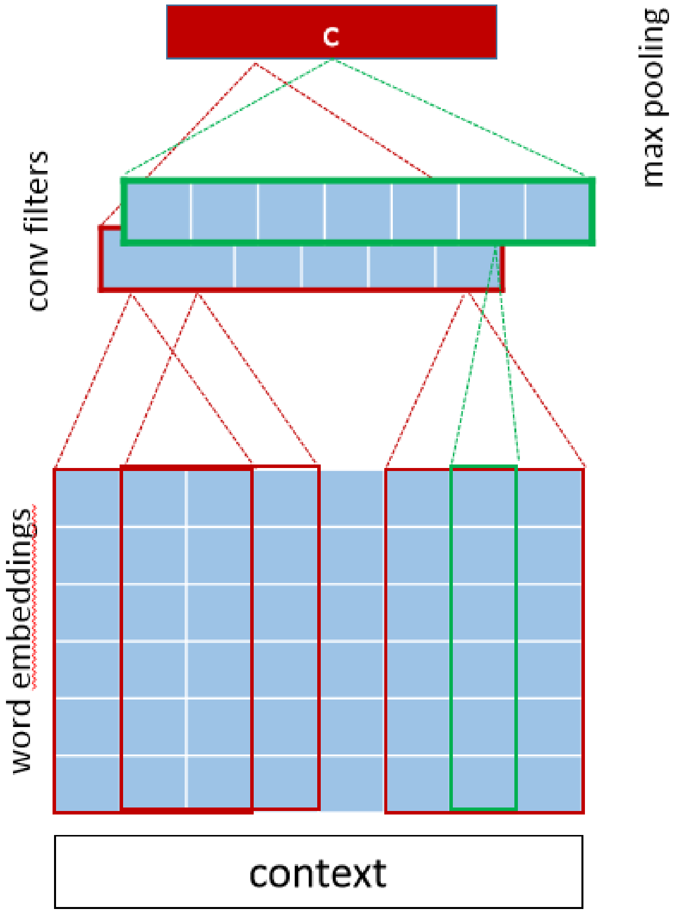
Context:

how to copy a file?

Responses:

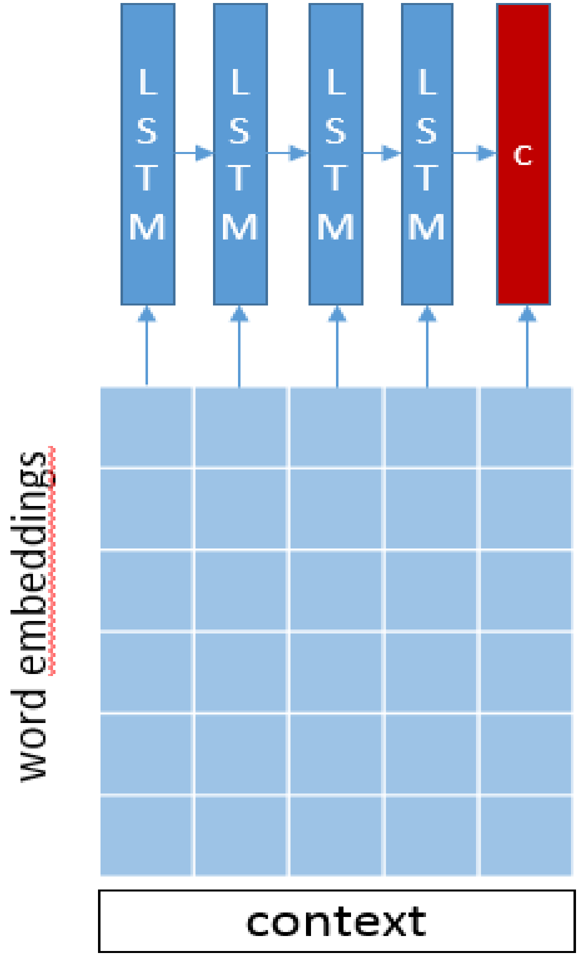
cp source target	1
sudo rm –rf /	0

CNN



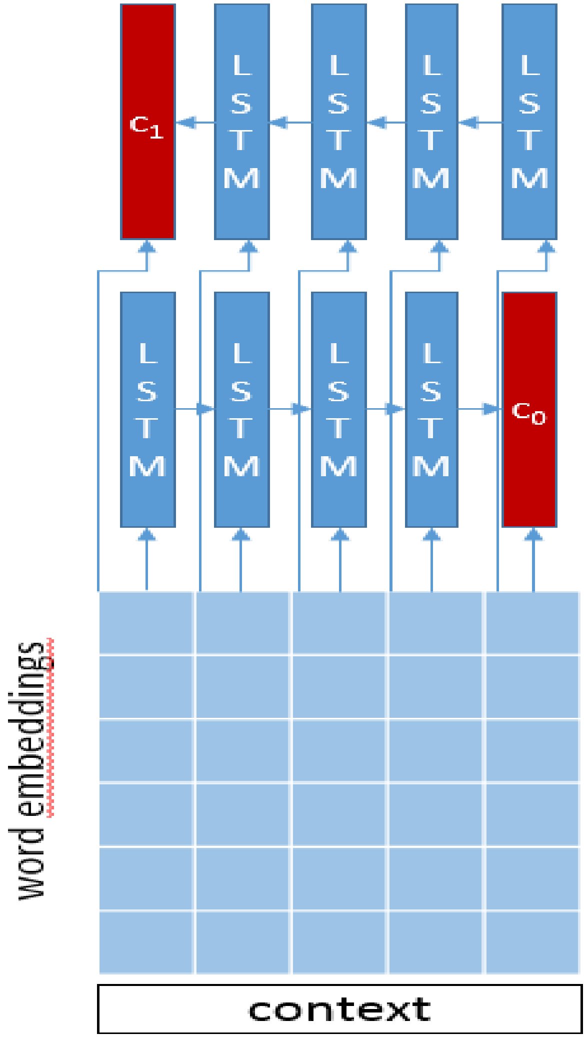
- CNN filters of various lengths
- Maxpooling over time

LSTM



- Sentence embedding is a vector with cell states in the last time step

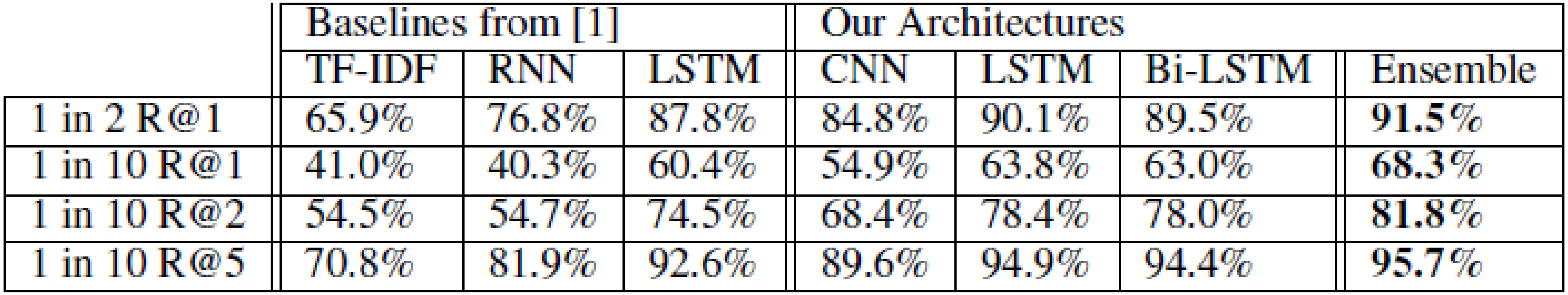
Bi-LSTM



- Sentence embedding is a concatenation of vectors with cell states from the last time step

Results

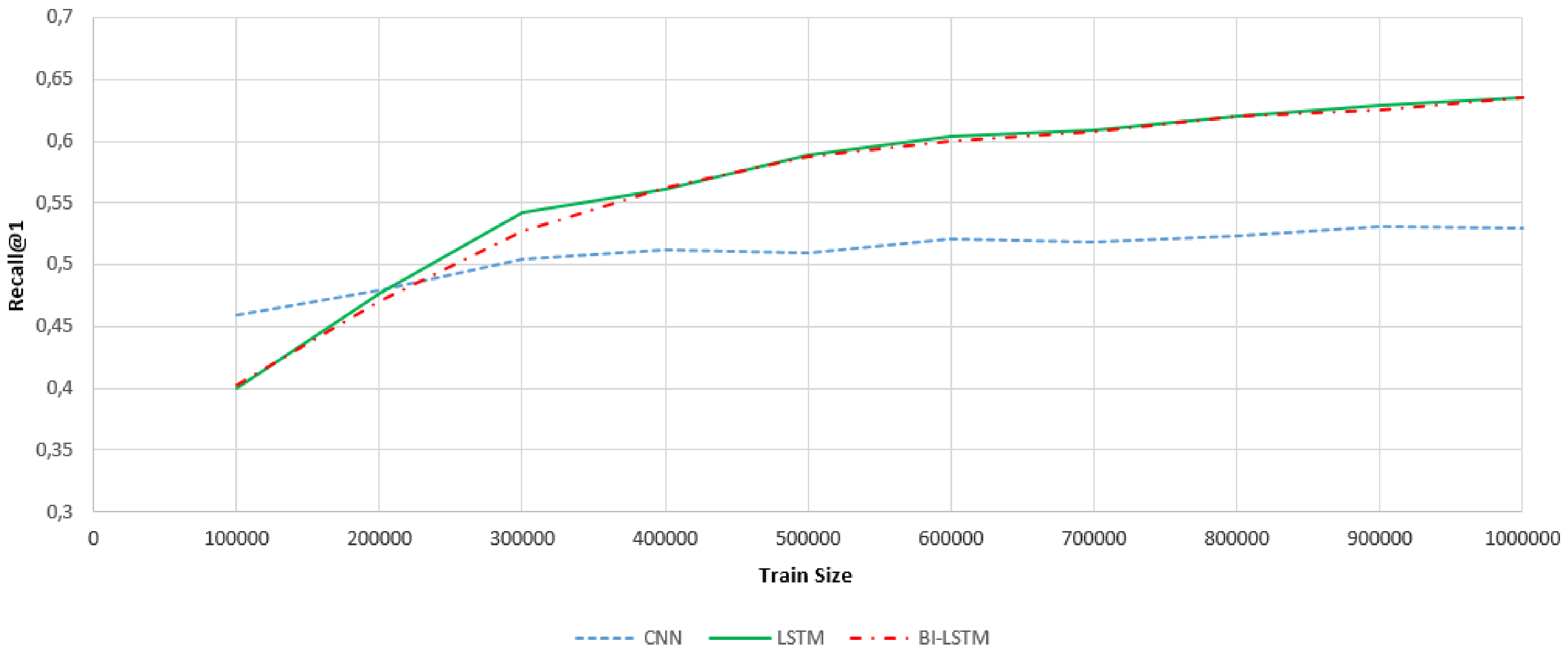
- The best results were achieved by ensemble of 28 models (11 LSTMs, 7 Bidir-LSTMs, 10 CNNs)
- Ensemble without CNNs has R@1 only 66.8% compared to 68.3% even though CNNs on their own are inferior to LSTMs



	Baselines from [1]			Our Architectures			
	TF-IDF	RNN	LSTM	CNN	LSTM	Bi-LSTM	Ensemble
1 in 2 R@1	65.9%	76.8%	87.8%	84.8%	90.1%	89.5%	**91.5%**
1 in 10 R@1	41.0%	40.3%	60.4%	54.9%	63.8%	63.0%	**68.3%**
1 in 10 R@2	54.5%	54.7%	74.5%	68.4%	78.4%	78.0%	**81.8%**
1 in 10 R@5	70.8%	81.9%	92.6%	89.6%	94.9%	94.4%	**95.7%**

How number of training examples affects accuracy

- CNNs better on small dataset
- LSTMs superior on larger sets



Summary

- We have improved baseline performance with “old” techniques
- Now we can try innovative architectures
  - For instance, external memory
    - Poster
      - Incorporating Unstructured Textual Knowledge Sources into Neural Dialogue Systems
- There are some issues with the dataset
  - Training examples has longer contexts that valid and test examples
  - Binary dataset is different than text dataset

References

- Original Ubuntu Dataset paper
  - Lowe, R., Pow, N., Serban, I., & Pineau, J. (2015). The Ubuntu Dialogue Corpus: A Large Dataset for Research in Unstructured Multi-Turn Dialogue Systems. *SIGdial*. Retrieved from http://arxiv.org/abs/1506.08909
- This work
  - Kadlec, R., Schmid, M., & Kleindienst, J. (2015). Improved Deep Learning Baselines for Ubuntu Corpus Dialogs. *Machine Learning for SLU&Interaction NIPS 2015 Workshop*. Retrieved from http://arxiv.org/abs/1510.03753
- New version of the Ubuntu dataset
  - https://github.com/rkadlec/ubuntu-ranking-dataset-creator

Thank you for your attention

Understanding natural text with neural networks

Rudolf Kadlec, Martin Schmid, Ondřej Bajgar and Jan Kleindienst

IBM Watson, Prague

http://arxiv.org/abs/1603.01547

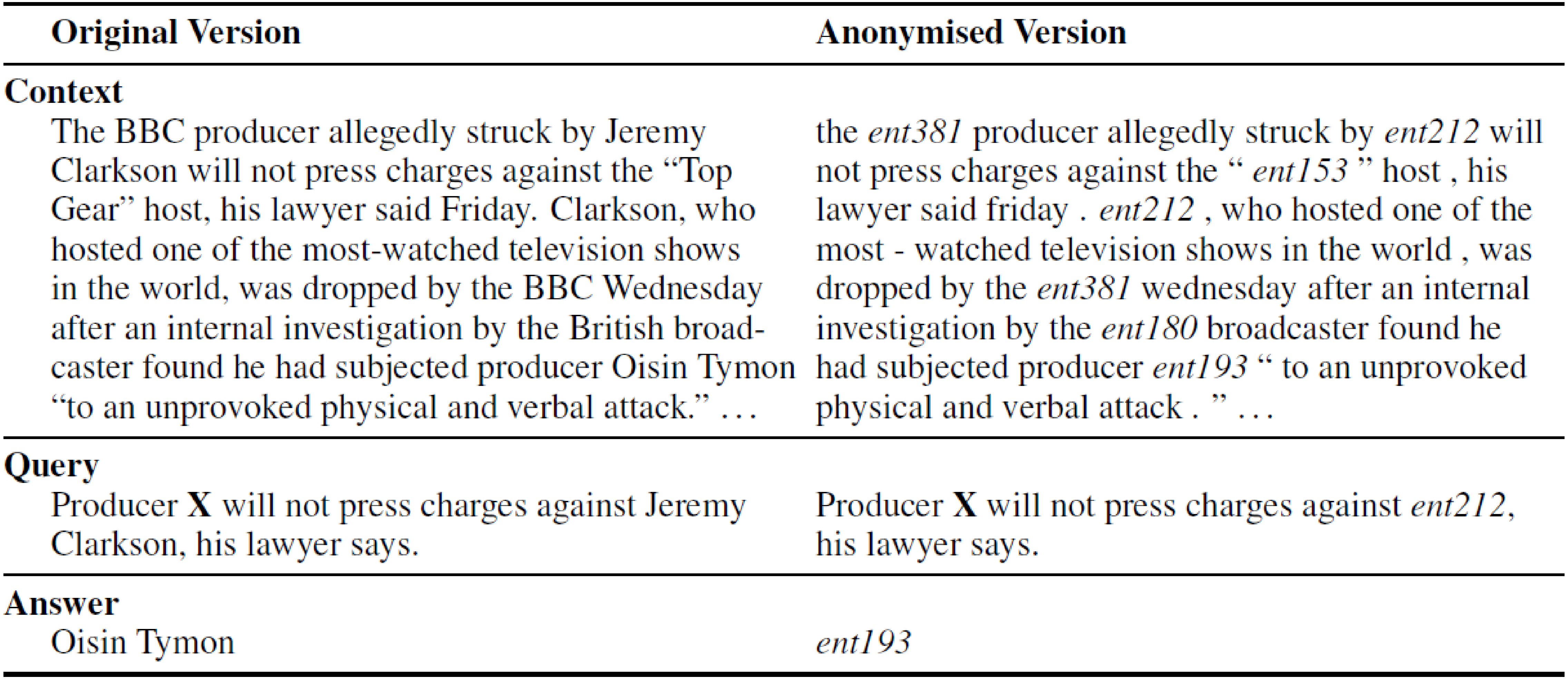
Motivation

- Unstructured text is rich source of information
  - As demonstrated by IBM Watson Jeopardy system
- Watson did not use any deep learning
- Here we show how to apply NNs to text understanding

- How to test text comprehension?
- Let the NN read an article and then ask questions about it
  - Cloze style questions can be used to generate these „questions“ automatically

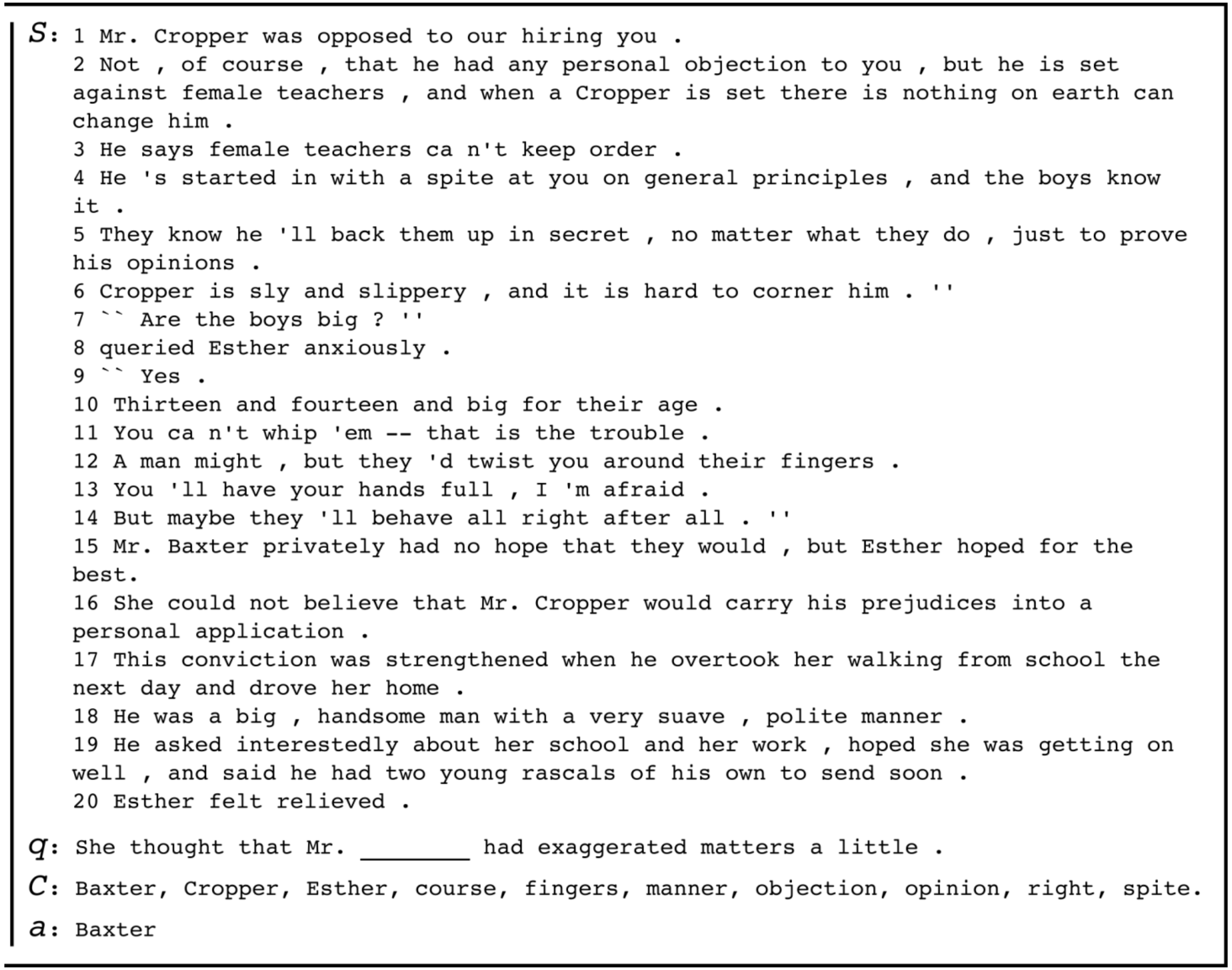
Datasets

CNN and Daily Mail (DeepMind)



	Original Version	Anonymised Version
Context	The BBC producer allegedly struck by Jeremy Clarkson will not press charges against the "Top Gear" host, his lawyer said Friday. Clarkson, who hosted one of the most-watched television shows in the world, was dropped by the BBC Wednesday after an internal investigation by the British broadcaster found he had subjected producer Oisin Tymon "to an unprovoked physical and verbal attack." . . .	the *ent381* producer allegedly struck by *ent212* will not press charges against the " *ent153* " host , his lawyer said friday . *ent212* , who hosted one of the most - watched television shows in the world , was dropped by the *ent381* wednesday after an internal investigation by the *ent180* broadcaster found he had subjected producer *ent193* " to an unprovoked physical and verbal attack . " . . .
Query	Producer **X** will not press charges against Jeremy Clarkson, his lawyer says.	Producer **X** will not press charges against *ent212*, his lawyer says.
Answer	Oisin Tymon	*ent193*

Children Book Test (Facebook AI)



S: 1 Mr. Cropper was opposed to our hiring you .
2 Not , of course , that he had any personal objection to you , but he is set against female teachers , and when a Cropper is set there is nothing on earth can change him .
3 He says female teachers ca n't keep order .
4 He 's started in with a spite at you on general principles , and the boys know it .
5 They know he 'll back them up in secret , no matter what they do , just to prove his opinions .
6 Cropper is sly and slippery , and it is hard to corner him . ''
7 `` Are the boys big ? ''
8 queried Esther anxiously .
9 `` Yes .
10 Thirteen and fourteen and big for their age .
11 You ca n't whip 'em -- that is the trouble .
12 A man might , but they 'd twist you around their fingers .
13 You 'll have your hands full , I 'm afraid .
14 But maybe they 'll behave all right after all . ''
15 Mr. Baxter privately had no hope that they would , but Esther hoped for the best.
16 She could not believe that Mr. Cropper would carry his prejudices into a personal application .
17 This conviction was strengthened when he overtook her walking from school the next day and drove her home .
18 He was a big , handsome man with a very suave , polite manner .
19 He asked interestedly 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 .
20 Esther felt relieved .

q: She thought that Mr. ________ had exaggerated matters a little .

C: Baxter, Cropper, Esther, course, fingers, manner, objection, opinion, right, spite.

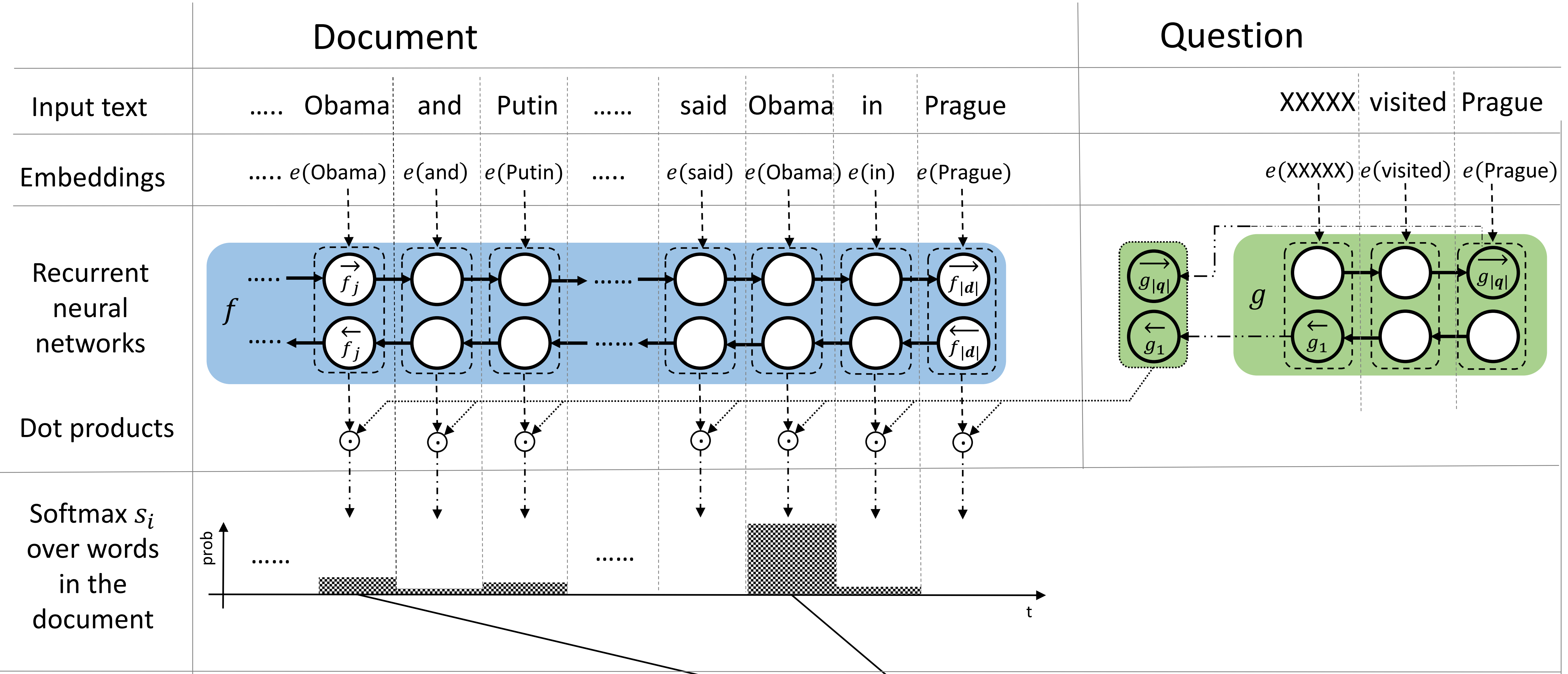
a: Baxter

	CNN			**Daily Mail**			**CBT CN**			**CBT NE**		
	train	valid	test	train	valid	test	train	valid	test	train	valid	test
# queries	380,298	3,924	3,198	879,450	64,835	53,182	120,769	2,000	2,500	108,719	2,000	2,500
Max # options	527	187	396	371	232	245	10	10	10	10	10	10
Avg # options	26.4	26.5	24.5	26.5	25.5	26.0	10	10	10	10	10	10
Avg # tokens	762	763	716	813	774	780	470	448	461	433	412	424
Vocab size	118,497			208,045			53,185			53,063		

Table 1: Statistics on the 4 data sets used to evaluate the model. CBT CN stands for CBT Common Nouns and CBT NE stands for CBT Named Entites. CBT had a fixed number of 10 options for answering each question. Statistics were taken from (Hermann et al., 2015) and the statistics provided with the CBT data set.

Attention Sum Reader

Our NN architecture

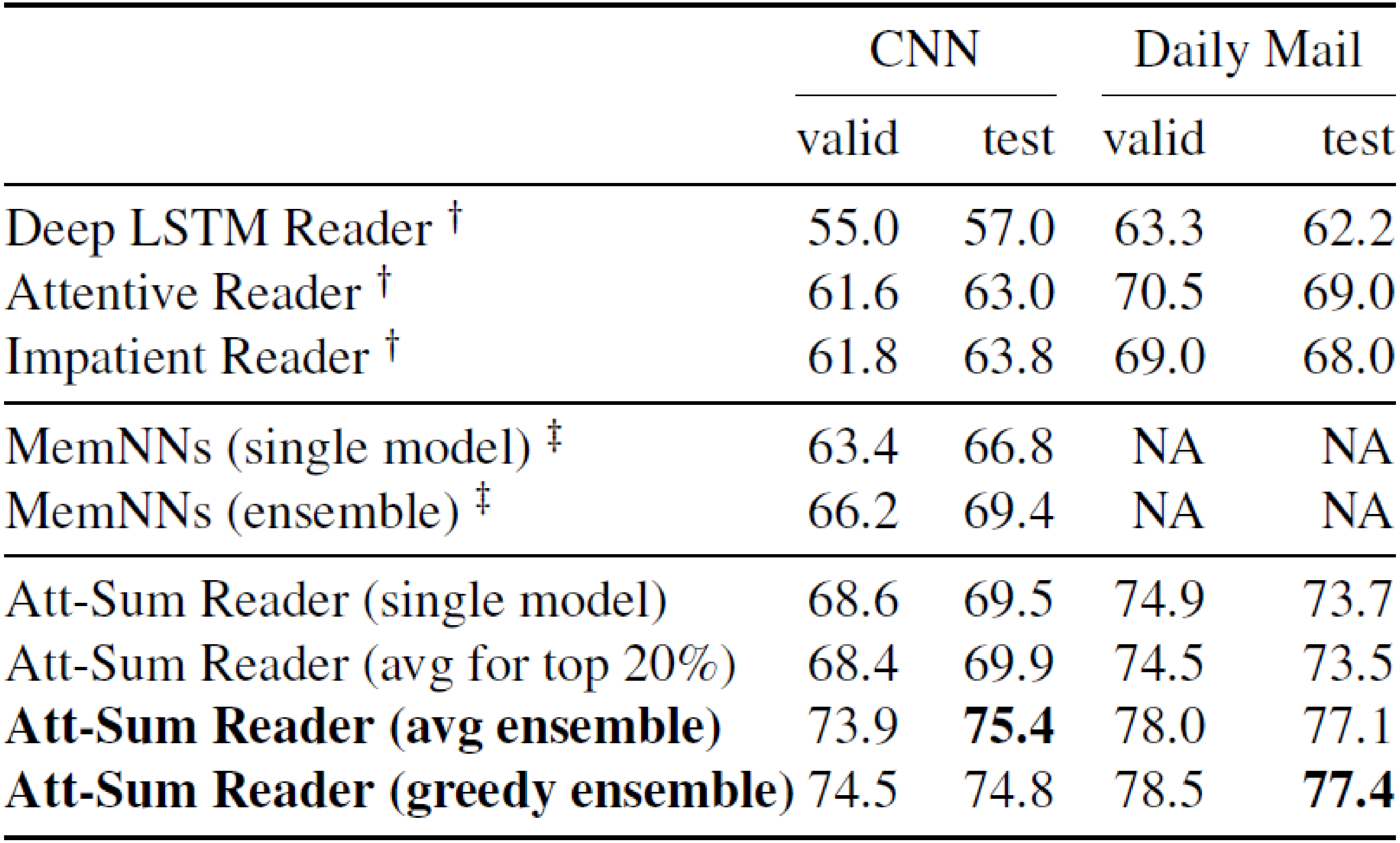

Document
Question
Input text
….. Obama and Putin …… said Obama in Prague
XXXXX visited Prague
Embeddings
….. e(Obama) e(and) e(Putin) ….. e(said) e(Obama) e(in) e(Prague)
e(XXXXX) e(visited) e(Prague)
Recurrent neural networks
f
$\overrightarrow{f_j}$
$\overleftarrow{f_j}$
$\overrightarrow{f_{|d|}}$
$\overleftarrow{f_{|d|}}$
g
$\overrightarrow{g_{|q|}}$
$\overleftarrow{g_1}$
Dot products
Softmax s_i over words in the document
prob
……
……
t

Probability of the answer

$$\mathrm{P}(\mathrm{Obama}|\mathrm{q},\mathrm{d}) = \sum_{i \in I(Obama,\mathrm{d})} s_i = s_j + s_{j+5}$$

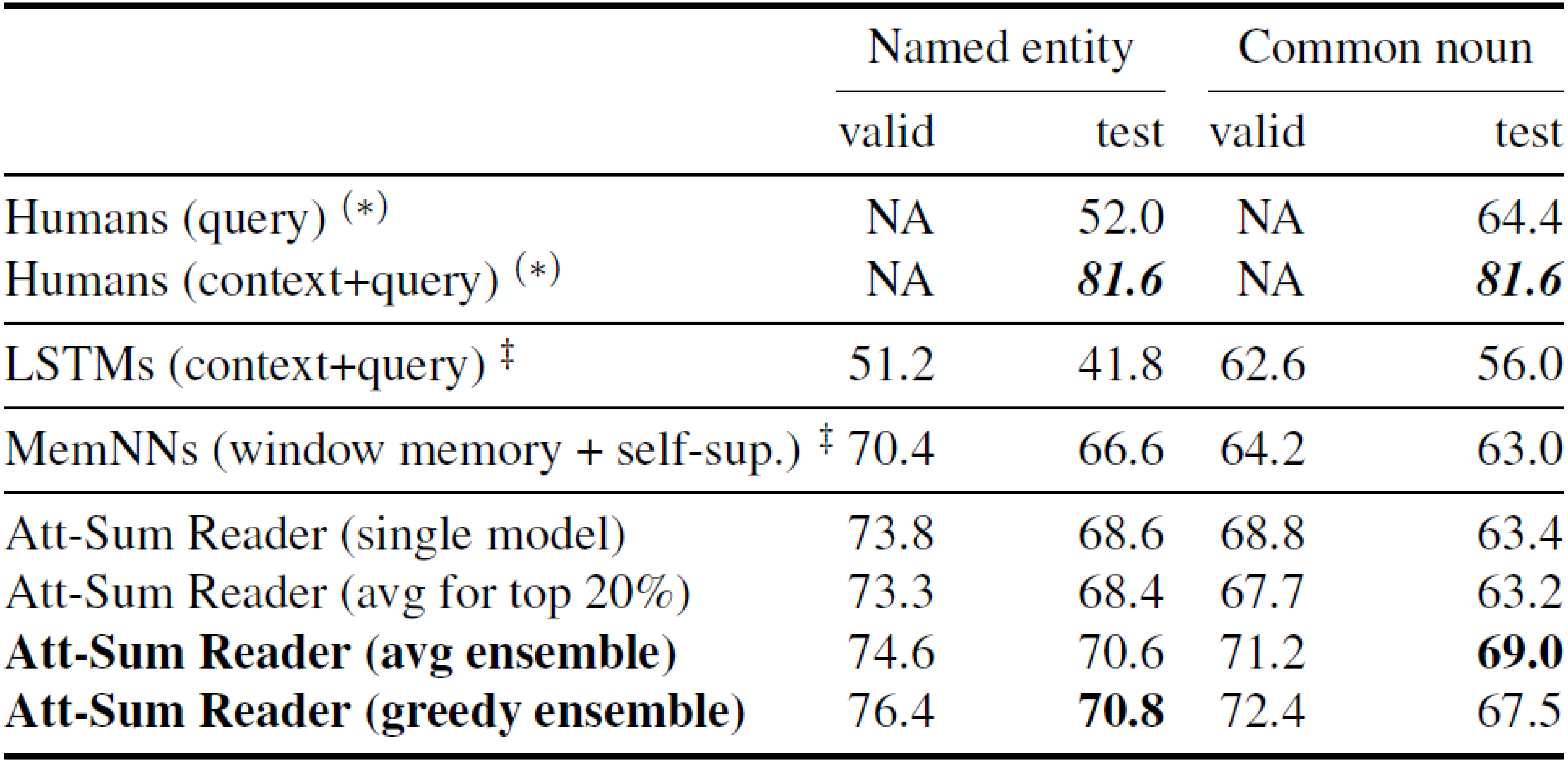
Results

CNN and Daily Mail dataset



		CNN		Daily Mail	
		valid	test	valid	test
(Hermann et al., 2015)	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
(Hill et al., 2016)	MemNNs (single model) ‡	63.4	66.8	NA	NA
	MemNNs (ensemble) ‡	66.2	69.4	NA	NA
(Kadlec et al., 2016)	Att-Sum Reader (single model)	68.6	69.5	74.9	73.7
	Att-Sum Reader (avg for top 20%)	68.4	69.9	74.5	73.5
	Att-Sum Reader (avg ensemble)	73.9	**75.4**	78.0	77.1
	Att-Sum Reader (greedy ensemble)	74.5	74.8	78.5	**77.4**

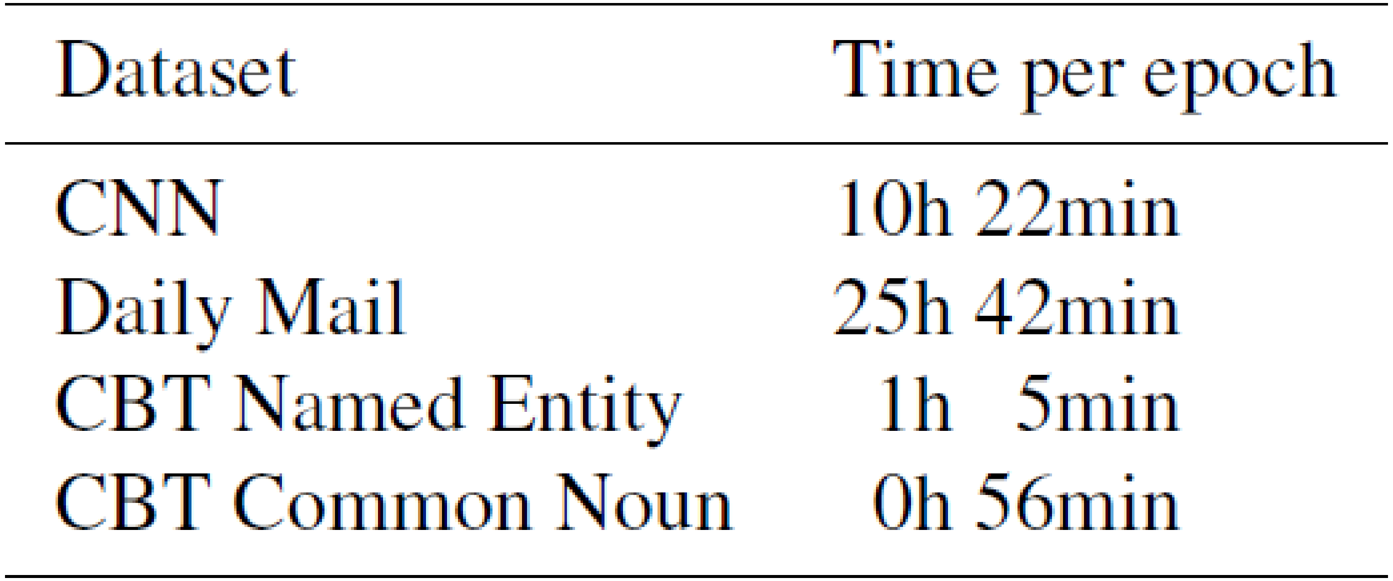
Children Book Test



		Named entity		Common noun	
		valid	test	valid	test
(Hill et al., 2016)	Humans (query) $^{(*)}$	NA	52.0	NA	64.4
	Humans (context+query) $^{(*)}$	NA	***81.6***	NA	***81.6***
	LSTMs (context+query) ‡	51.2	41.8	62.6	56.0
	MemNNs (window memory + self-sup.) ‡	70.4	66.6	64.2	63.0
(Kadlec et al., 2016)	Att-Sum Reader (single model)	73.8	68.6	68.8	63.4
	Att-Sum Reader (avg for top 20%)	73.3	68.4	67.7	63.2
	Att-Sum Reader (avg ensemble)	74.6	70.6	71.2	**69.0**
	Att-Sum Reader (greedy ensemble)	76.4	**70.8**	72.4	67.5

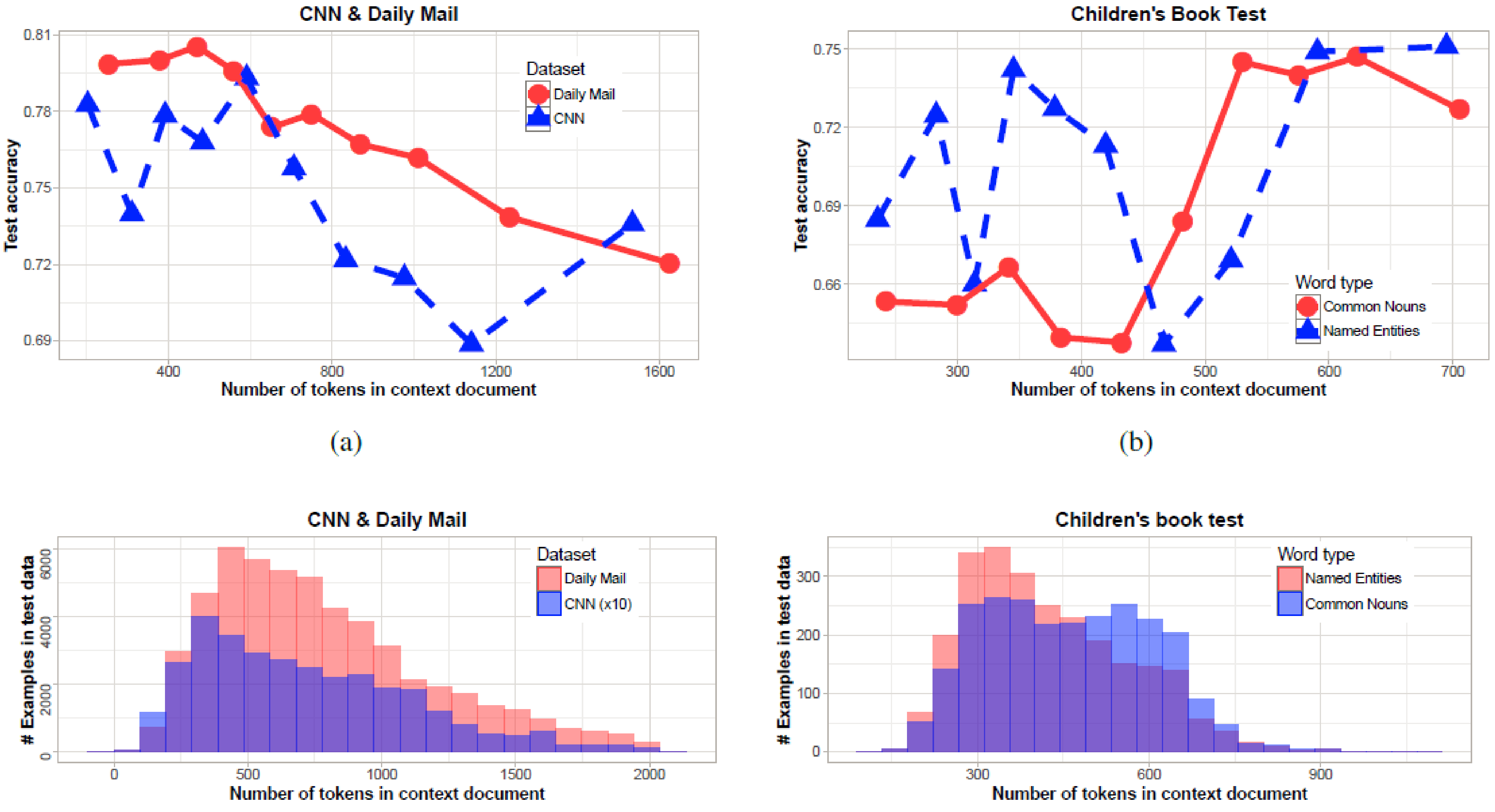
Training times

- We used Nvidia K40 GPUs
- Models converged after 2 epochs



Dataset	Time per epoch
CNN	10h 22min
Daily Mail	25h 42min
CBT Named Entity	1h 5min
CBT Common Noun	0h 56min

Analysis



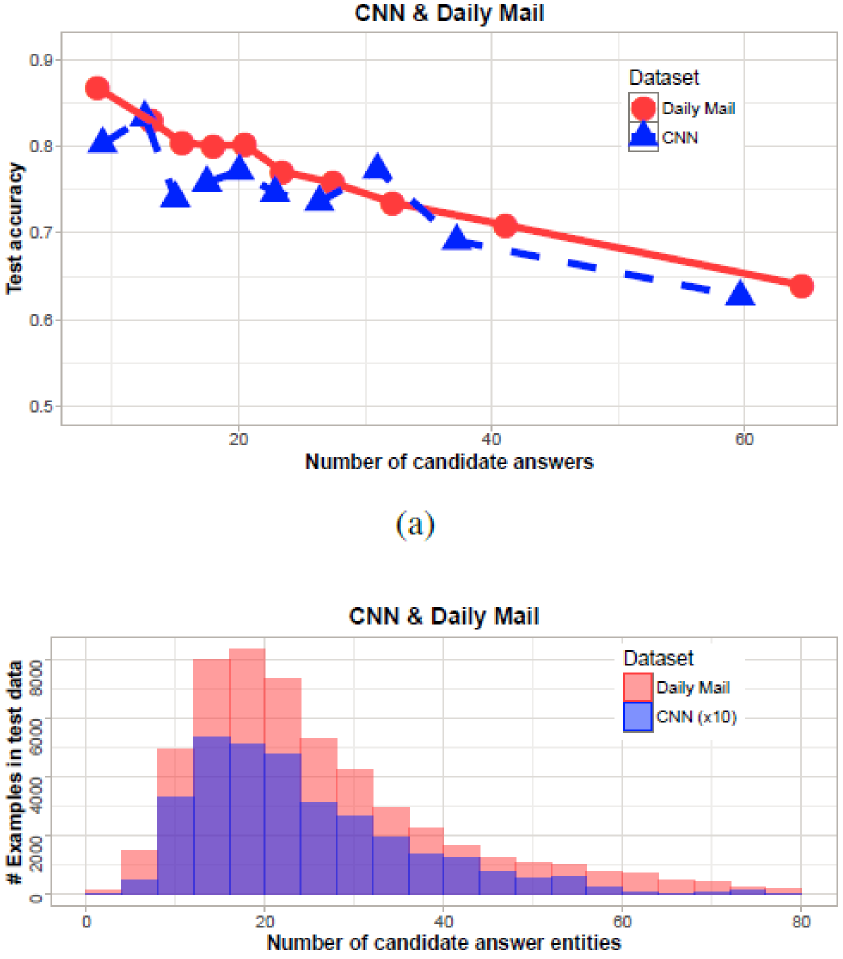
(a)

(b)

(c)

(d)

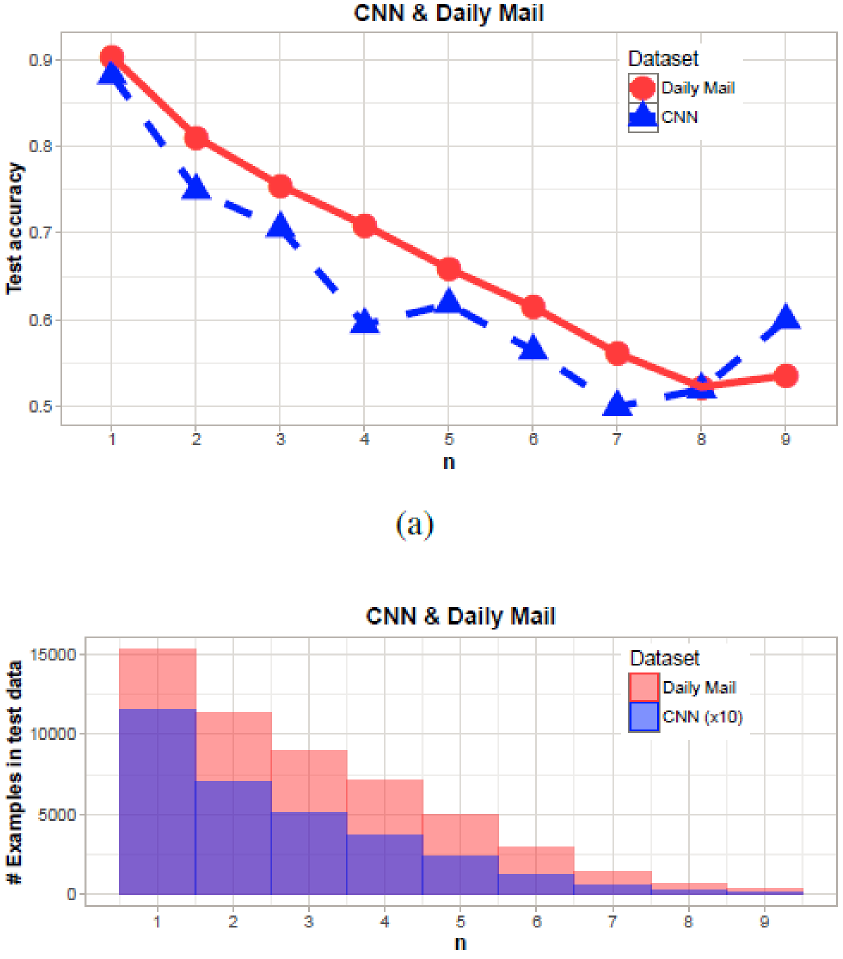
Figure 2: Sub-figures (a) and (b) plot the test accuracy against the length of the context document (for CNN the count was multiplied by 10). The examples were split into ten buckets of equal size by their context length. Averages for each bucket are plotted on each axis. Sub-figures (c) and (d) show distributions of context lengths in the four datasets. The number of examples was multiplied by 10 for the CNN dataset.



(a)

(b)

Figure 3: Subfigure (a) illustrates how the model accuracy decreases with an increasing number of candidate named entities. Subfigure (b) shows the overall distribution of the number of candidate answers in the news datasets. The number of examples was multiplied by 10 for the CNN dataset.

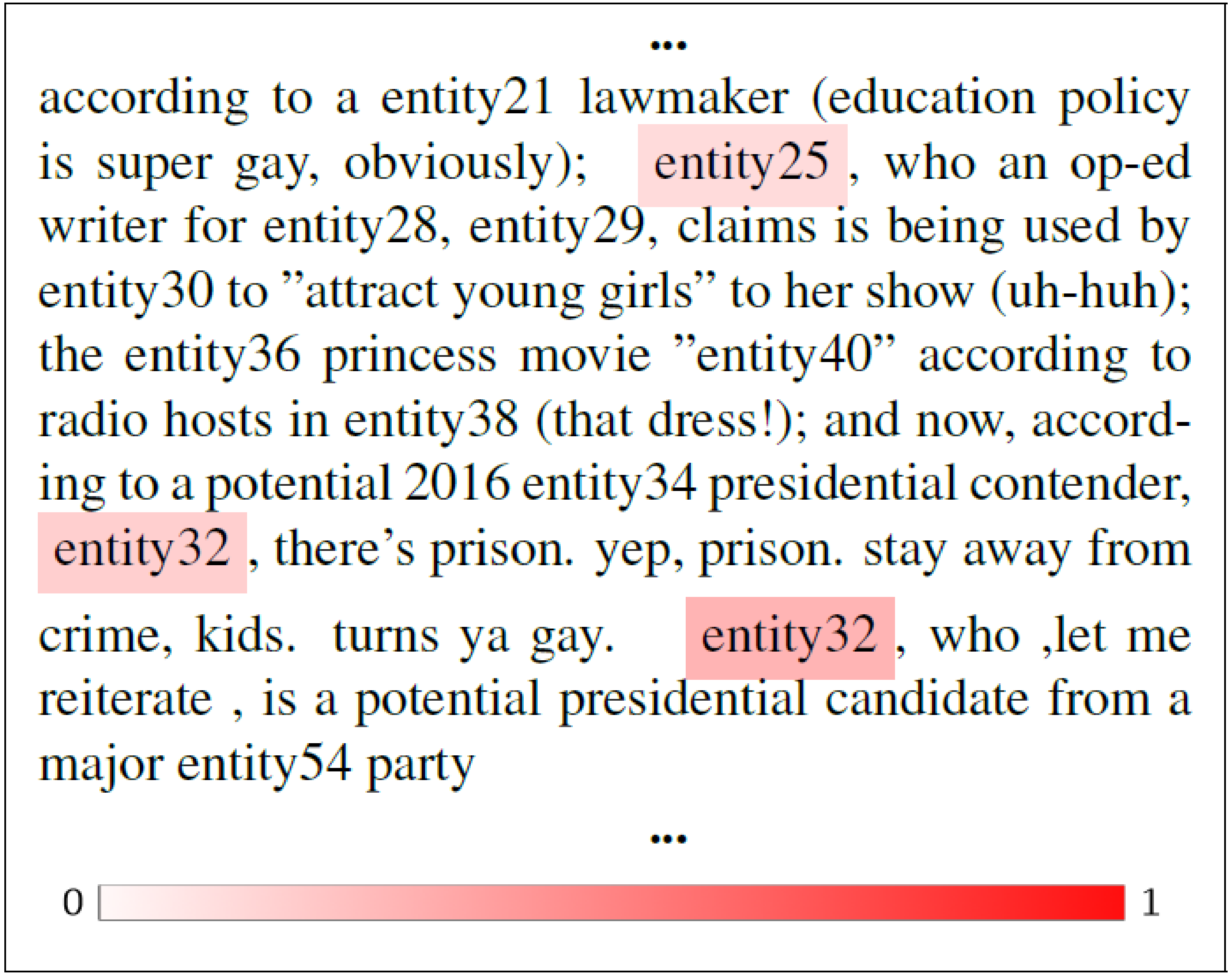


(a)

(b)

Figure 4: Subfigure (a) shows the model accuracy when the correct answer is among n most frequent named entities for $n \in [1, 10]$. Subfigure (b) shows the number of test examples for which the correct answer was the n–th most frequent entity. The number of examples was multiplied by 10 for the CNN dataset.

Example



...

according to a entity21 lawmaker (education policy is super gay, obviously); entity25 , who an op-ed writer for entity28, entity29, claims is being used by entity30 to ”attract young girls” to her show (uh-huh); the entity36 princess movie ”entity40” according to radio hosts in entity38 (that dress!); and now, according to a potential 2016 entity34 presidential contender, entity32 , there’s prison. yep, prison. stay away from crime, kids. turns ya gay. entity32 , who ,let me reiterate , is a potential presidential candidate from a major entity54 party

...

possible 2016 entity34 candidate **X** stirs controversy with comments on gays , prison

References

- DeepMind's model:
  - Hermann, K. M., Kočiský, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M., & Blunsom, P. (2015). Teaching Machines to Read and Comprehend. *NIPS*.
- Facebook's model:
  - Hill, F., Bordes, A., Chopra, S., & Weston, J. (2016). The Goldilocks Principle: Reading Children's Books with Explicit Memory Representations. *ICLR*.
- IBM's model
  - Kadlec, R., Schmid, M., Bajgar, O., & Kleindienst, J. (2016). Neural Text Understanding with Attention Sum Reader. http://arxiv.org/abs/1603.01547