A Supplemental Material

A.1 Details of the Implementation

We implemented our model in PyTorch and trained it on four Nvidia Tesla P100 GPUs. The RNN was a gated recurrent unit (GRU) (Cho et al., 2014). The optimizer was Adam (Kingma and Ba, 2014). The word-based word embeddings were fixed GloVe 300-dimensional vectors (Pennington et al., 2014). The character-based word embeddings were obtained using trainable eight-dimensional character embeddings and a 100-dimensional CNN and max pooling. Table 1 shows other hyper parameters.

In FEVER, if the model predicts A_T as 'Supports' or 'Refutes', the model extracts at least one sentence by removing the EOE sentence from the candidates to be extracted at $t = 1$.

A.2 Samples of QFE Outputs

The section describes some examples of QFE outputs. Table 2 shows examples on HotpotQA, and Table 3 shows examples on FEVER. We should note that QFE does not necessarily extract the sentence with the highest probability score at any step because QFE determines the evidence by using the beam search algorithm.

Three or four correct evidence sentences are extracted in the first and second examples in Table 2. The third example is a typical mistake of QFE; QFE extracts too few evidence sentences. In the fourth example, QFE extracts too many evidence sentences. The fifth and sixth questions are typical yes/no questions in HotpotQA. However, like other QA models, our model makes mistakes in answering such easy questions.

One or two evidence sentences are extracted correctly in the first, second, and third examples in Table 3. In FEVER, most claims requiring two evidence sentences can be verified by either of two correct evidence sentences, like in the second example. However, there are some claims that require both evidence sentences, like the third example. The fourth example is a typical mistake of QFE; QFE extracts too few evidence sentences. In the fifth and sixth example, the answers of the questions are 'Not Enough Info'. QFE unfortunately extracts evidence when the QA model predicts another label.

	HotpotQA	FEVER
size of word vectors: d_w	400	435
width of RNN: d_c	150	150
dropout keep ratio	0.8	0.8
batch size	72	96
learning rate	0.001	0.001
beam size	5	5

Table 1: Hyper Parameters.

References

Kyunghyun Cho, Bart Merrienboer, Caglar Gulcehre, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. In *EMNLP*, pages 1724–1734.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global vectors for word representation. In *EMNLP*, pages 1532–1543.

Q: What plant has about 40 species native to Asia , Manglietia or Abronia?
A_T: Manglietia, $\hat{A}_T$: Manglietia

gold	predicted	probability[%]	text
✓	1	**85.4**	Abronia ... is a genus of about 20 species of ....
✓	2	$14.6 \rightarrow$ **54.5**	Manglietia is a genus of flowering plants in the family Magnoliaceae.
✓	3	$0.0 \rightarrow 45.1 \rightarrow$ **61.4**	There are about 40 species native to Asia.
	4	$0.0 \rightarrow \ 0.0 \rightarrow 37.2 \rightarrow$ **99.2**	EOE sentence

Q:Ricky Martin's concert tour in 1999 featured an American heavy metal band formed in what year?
A_T:1991, $\hat{A}_T$: 1991

gold	predicted	probability[%]	text
✓	1	**100.0**	Formed on October 12, 1991, the group was founded by vocalist/guitarist Robb Flynn and bassist Adam Duce.
✓	2	$0.0 \rightarrow$ **98.8**	Other bands that were featured included Machine Head, Slipknot, and Amen.
✓	3	$0.0 \rightarrow \ 0.0 \rightarrow$ **97.7**	Machine Head is an American heavy metal band from Oakland, California.
✓	4	$0.0 \rightarrow \ 1.1 \rightarrow \ 1.4 \rightarrow$ **97.3**	Livin La Vida Loco ... by Ricky Martin, was a concert tour in 1999.
	5	$0.0 \rightarrow \ 0.0 \rightarrow \ 0.9 \rightarrow \ 2.0 \rightarrow$ **97.0**	EOE sentence

Q: Where is the singer of "B Boy" raised?
A_T: Philadelphia, $\hat{A}_T$: Philadelphia

gold	predicted	probability[%]	text
✓	1	**100.0**	Raised in Philadelphia, he embarked ....
✓	2	$0.0 \rightarrow$ **100.0**	"B Boy" is a song by American hip hop recording artist Meek Mill.
	3	$0.0 \rightarrow \ 0.0 \rightarrow$ **79.0**	EOE sentence
✓	—	$0.0 \rightarrow \ 0.0 \rightarrow 20.8$	Robert Rihmeek Williams ... known by his stage name, Meek Mill, ...

Q: Which comic series involves characters such as Nick Fury and Baron von Strucker?
A_T: Marvel, $\hat{A}_T$: Sgt. Fury

gold	predicted	probability[%]	text
✓	1	**70.7**	Andrea von Strucker ... characters appearing in American comic books published by Marvel Comics.
	2	$1.7 \rightarrow$ **41.6**	It is the first series to feature Nick Fury Jr. as its main character.
	3	$17.3 \rightarrow 38.2 \rightarrow$ **31.6**	Nick Fury is a 2017 ongoing comic book series published by Marvel Comics.
	4	$0.0 \rightarrow \ 0.0 \rightarrow 41.6 \rightarrow$ **92.0**	EOE sentence
✓	—	$0.0 \rightarrow \ 0.0 \rightarrow \ 0.0 \rightarrow \ 0.0$	Nick Fury: ... the Marvel Comics character Nick Fury.

Q: Are both "Cooking Light" and "Vibe" magazines?
A_T: yes, $\hat{A}_T$: yes

gold	predicted	probability[%]	text
✓	1	**89.0**	Cooking Light is an American monthly food and lifestyle magazine founded in 1987.
✓	2	$11.0 \rightarrow$ **97.4**	Vibe is an American music and entertainment magazine founded by producer Quincy Jones.
	3	$0.0 \rightarrow \ 0.0 \rightarrow$ **95.4**	EOE sentence

Q: Are Robert Philibosian and David Ignatius both politicians?
A_T: no, $\hat{A}_T$: yes

gold	predicted	probability[%]	text
✓	1	**100.0**	Robert Harry Philibosian (born 1940) is an American politician.
✓	2	$0.0 \rightarrow$ **98.7**	David R. Ignatius (May 26, 1950), is an American journalist and novelist.
	3	$0.0 \rightarrow \ 0.0 \rightarrow$ **97.4**	EOE sentence

Table 2: Outputs of QFE on HotpotQA. The sentences are extracted in the order shown in the predicted column. The extraction scores of the sentences at each step are in the probability column.

Q: Fox 2000 Pictures released the film Soul Food. A_T: Supports $\hat{A}_T$: Supports			
gold	predicted	probability[%]	text
✓	1	**98.0**	Soul Food is a 1997 American comedy-drama film ... and released by Fox 2000 Pictures.
	2	$0.0 \rightarrow$ **75.9**	EOE sentence
Q: Terry Crews was a football player. A_T: Supports $\hat{A}_T$: Supports			
gold	predicted	probability[%]	text
✓	1	**96.0**	Terry Alan Crews ... is an American actor , artist , and former American football player.
✓	2	$3.8 \rightarrow$ **56.4**	In football , Crews played as ....
	3	$0.0 \rightarrow 41.8 \rightarrow$ **86.4**	EOE sentence
Q: Jack Falahee is an actor and he is unknown. A_T: Refutes $\hat{A}_T$: Refutes			
gold	predicted	probability[%]	text
✓	1	**95.1**	Jack Ryan Falahee (born February 20 , 1989) is an American actor.
✓	2	$4.9 \rightarrow$ **67.1**	He is known for his role as Connor Walsh on ....
	3	$0.0 \rightarrow 32.9 \rightarrow$ **100.0**	EOE sentence
Q: Same Old Love is disassociated from Selena Gomez. A_T: Refutes $\hat{A}_T$: Refutes			
gold	predicted	probability[%]	text
✓	1	**75.0**	"Same Old Love" is a song by American singer Selena Gomez ....
	2	$0.0 \rightarrow$ **69.8**	EOE sentence
✓	–	$0.2 \rightarrow \ 0.5$	Gomez promoted "Same Old Love" ....
Q: Annette Badland was in the 2015 NBA Finals A_T: Not Enough Info $\hat{A}_T$: Not Enough Info			
gold	predicted	probability[%]	text
	1	**98.6**	EOE sentence
Q: Billboard Dad is a genre of music. A_T: Not Enough Info $\hat{A}_T$: Refutes			
gold	predicted	probability[%]	text
	1	**98.4**	Billboard Dad (film) is a 1998 American direct-to-video comedy film ....
	–	$0.00 \rightarrow$ **83.5**	EOE sentence

Table 3: Outputs of QFE (single model) on FEVER. The sentences are extracted in the order shown in the predicted column. The extraction scores of the sentences at each step are in the probability column.