Attention-over-Attention Neural Networks for Reading Comprehension

Yiming Cui, Zhipeng Chen, Si Wei, Shijin Wang, Ting Liu and Guoping Hu

Joint Laboratory of HIT and iFLYTEK (HFL), Beijing

2017-04-22

Outline

- Introduction: Cloze-style Reading Comprehension
- Related Works
- Attention-over-Attention Reader (AoA Reader)
- Experiments & Analysis
- Conclusions & Future Works

PREFACE

- This work was first made publicly available on June, 2016 @arXiv
- Citations: 18 (based on Google Scholar)
- Propose a novel mechanism called "Attention-over-Attention"

Attention-over-Attention Neural Networks for Reading Comprehension

Yiming Cui[†], Zhipeng Chen[†], Si Wei[†], Shijin Wang[†], Ting Liu[‡] and Guoping Hu[†]
[†]Joint Laboratory of HIT and iFLYTEK, iFLYTEK Research, Beijing, China
[‡]Research Center for Social Computing and Information Retrieval,
Harbin Institute of Technology, Harbin, China
[†]{ymcui,zpchen,siwei,sjwang3,gphu}@iflytek.com
[‡]tliu@ir.hit.edu.cn

Introduction

- Recently, ***Reading Comprehension*** has become enormously popular in the community
- A main obstacle in the RC research
  - NO MUCH DATA !
- The related works are often started from providing the relevant corpus, and then proposing some technical insights in solving them

Introduction

- Key points in RC
  - →**Document**
  - Query
  - Candidates
  - Answer

James the Turtle was always getting in trouble. Sometimes he'd reach into the freezer and empty out all the food. Other times he'd sled on the deck and get a splinter. His aunt Jane tried as hard as she could to keep him out of trouble, but he was sneaky and got into lots of trouble behind her back.
One day, James thought he would go into town and see what kind of trouble he could get into. He went to the grocery store and pulled all the pudding off the shelves and ate two jars. Then he walked to the fast food restaurant and ordered 15 bags of fries. He didn't pay, and instead headed home.
His aunt was waiting for him in his room. She told James that she loved him, but he would have to start acting like a well-behaved turtle.
After about a month, and after getting into lots of trouble, James finally made up his mind to be a better turtle.

1) What is the name of the trouble making turtle?
A) Fries
B) Pudding
C) James
D) Jane

*Example is chosen from the MCTest dataset (Richardson et al., 2013)

Introduction

- Key points in RC
  - Document
  - →**Query**
  - Candidates
  - Answer

James the Turtle was always getting in trouble. Sometimes he'd reach into the freezer and empty out all the food. Other times he'd sled on the deck and get a splinter. His aunt Jane tried as hard as she could to keep him out of trouble, but he was sneaky and got into lots of trouble behind her back.
One day, James thought he would go into town and see what kind of trouble he could get into. He went to the grocery store and pulled all the pudding off the shelves and ate two jars. Then he walked to the fast food restaurant and ordered 15 bags of fries. He didn't pay, and instead headed home.
His aunt was waiting for him in his room. She told James that she loved him, but he would have to start acting like a well-behaved turtle.
After about a month, and after getting into lots of trouble, James finally made up his mind to be a better turtle.

1) What is the name of the trouble making turtle?
A) Fries
B) Pudding
C) James
D) Jane

*Example is chosen from the MCTest dataset (Richardson et al., 2013)

Introduction

- Key points in RC
  - Document
  - Query
  - →**Candidates**
  - Answer

James the Turtle was always getting in trouble. Sometimes he'd reach into the freezer and empty out all the food. Other times he'd sled on the deck and get a splinter. His aunt Jane tried as hard as she could to keep him out of trouble, but he was sneaky and got into lots of trouble behind her back.

One day, James thought he would go into town and see what kind of trouble he could get into. He went to the grocery store and pulled all the pudding off the shelves and ate two jars. Then he walked to the fast food restaurant and ordered 15 bags of fries. He didn't pay, and instead headed home.

His aunt was waiting for him in his room. She told James that she loved him, but he would have to start acting like a well-behaved turtle.

After about a month, and after getting into lots of trouble, James finally made up his mind to be a better turtle.

1) What is the name of the trouble making turtle?

A) Fries
B) Pudding
C) James
D) Jane

*Example is chosen from the MCTest dataset (Richardson et al., 2013)

Introduction

- Key points in RC
  - Document
  - Query
  - Candidates
  - **→Answer**

James the Turtle was always getting in trouble. Sometimes he'd reach into the freezer and empty out all the food. Other times he'd sled on the deck and get a splinter. His aunt Jane tried as hard as she could to keep him out of trouble, but he was sneaky and got into lots of trouble behind her back.

One day, James thought he would go into town and see what kind of trouble he could get into. He went to the grocery store and pulled all the pudding off the shelves and ate two jars. Then he walked to the fast food restaurant and ordered 15 bags of fries. He didn't pay, and instead headed home.

His aunt was waiting for him in his room. She told James that she loved him, but he would have to start acting like a well-behaved turtle.

After about a month, and after getting into lots of trouble, James finally made up his mind to be a better turtle.

1) What is the name of the trouble making turtle?

A) Fries

B) Pudding

C) James

D) Jane

*Example is chosen from the MCTest dataset (Richardson et al., 2013)

Introduction

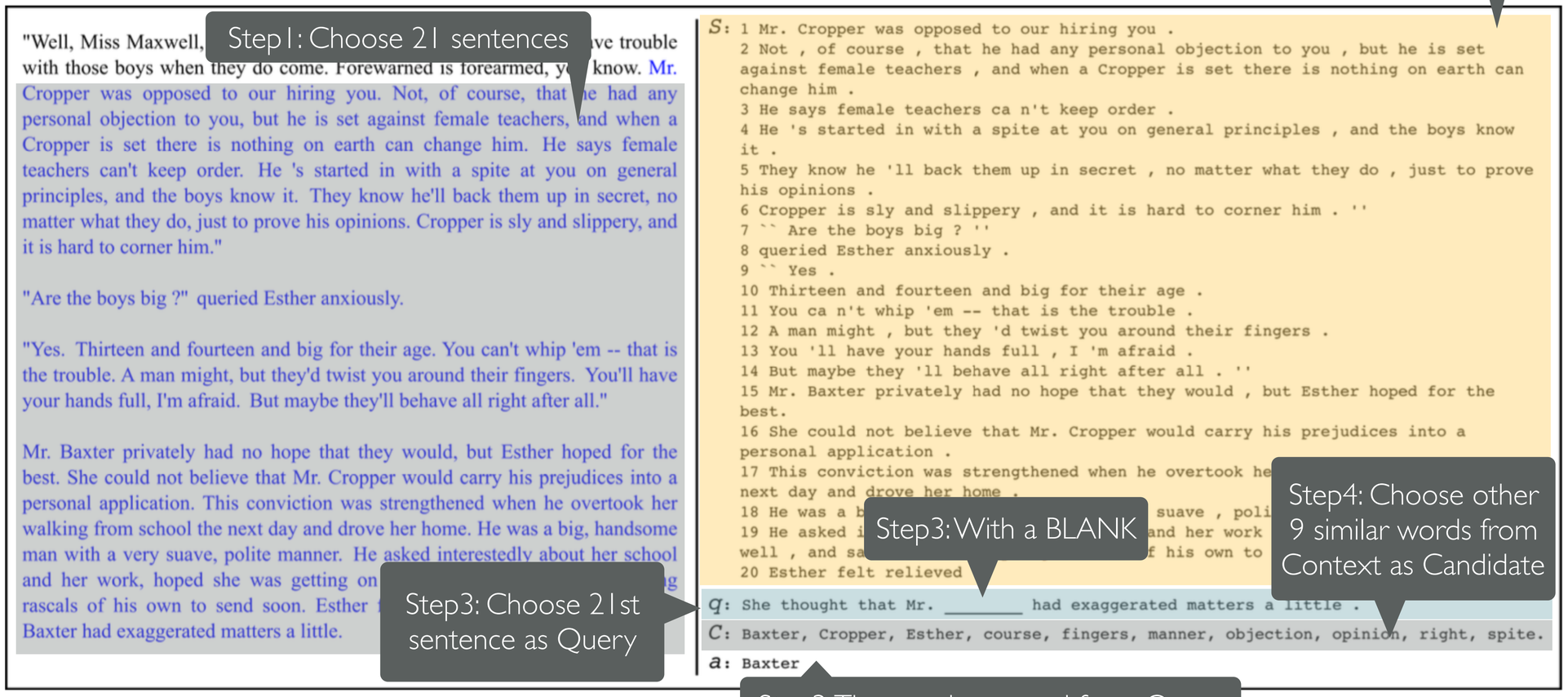
- Specifically, in cloze-style RC
  - Document: the same as the general RC
  - Query: a sentence with a blank
  - Candidate (optional): several candidates to fill in
  - Answer: a single word that exactly match the query (the answer word should appear in the document)

Original 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." ...
Query Producer **X** will not press charges against Jeremy Clarkson, his lawyer says.
Answer Oisin Tymon

*Example is chosen from the CNN dataset (Hermann et al., 2015)

Introduction

- CBT dataset (Hill et al., 2015)

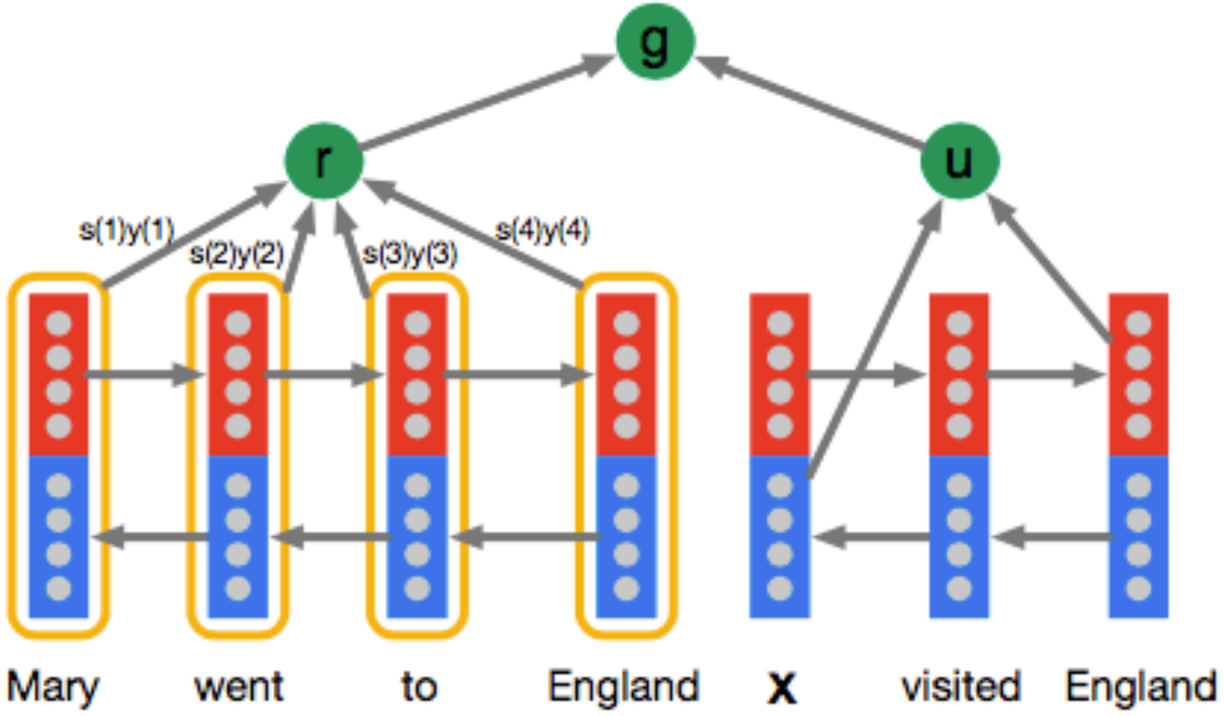


Related Works

- Attentive Reader (Hermann et al., NIPS2015)
- Attention Sum Reader (Kadlec et al., ACL2016)
- Consensus Attention Reader (Cui et al., COLING2016)
- Gated-attention Reader (Dhingra et al., ICLR2017)
- …

Attentive Reader

- Teaching Machines to Read and Comprehend (Hermann et al., 2015)

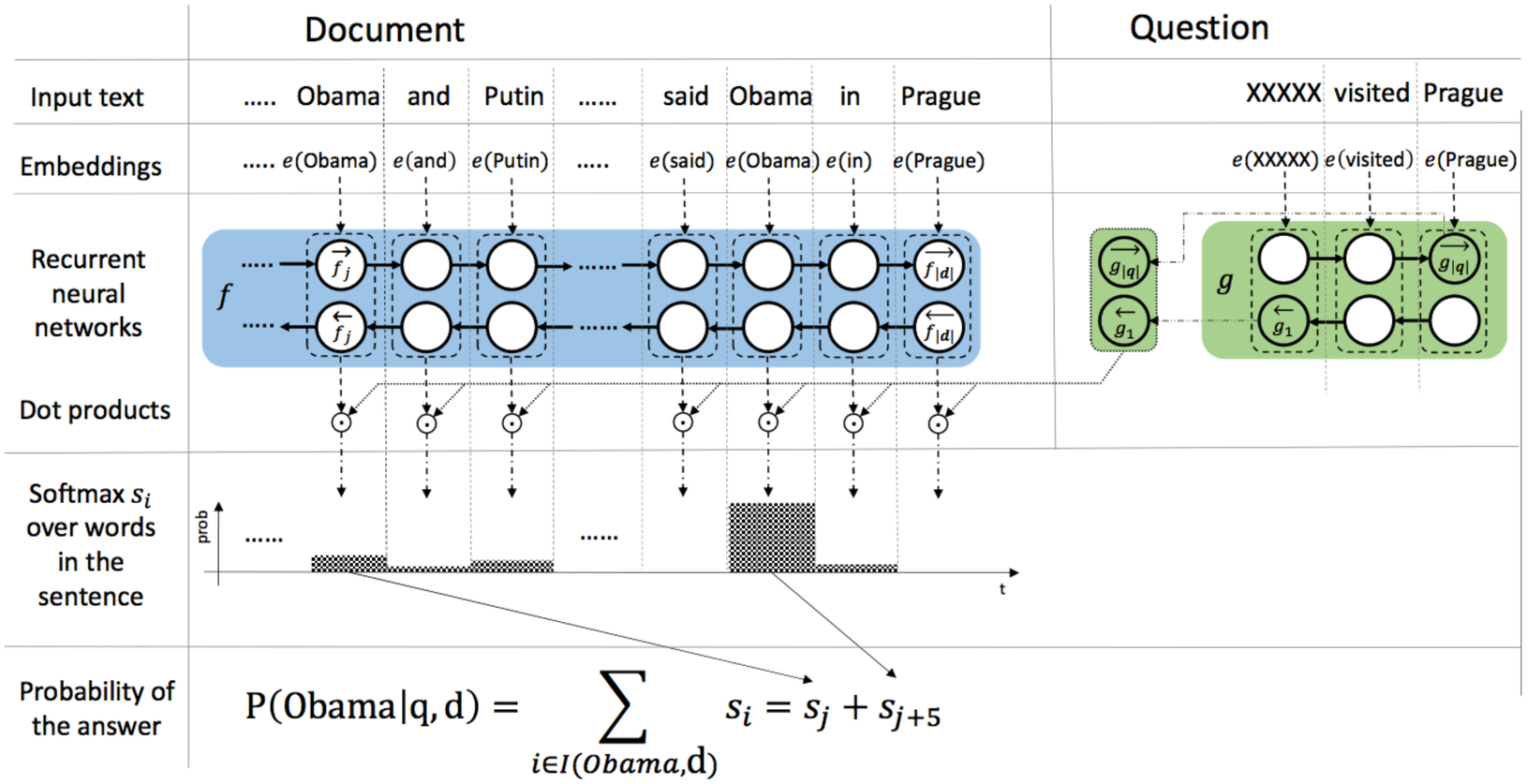


$$
\begin{aligned}
m(t) &= \tanh\left(W_{ym} y_d(t) + W_{um} u\right), \\
s(t) &\propto \exp\left(\mathrm{w}_{ms}^{\intercal} m(t)\right), \\
r &= y_d s,
\end{aligned}
$$

$$
g^{\mathrm{AR}}(d, q) = \tanh\left(W_{rg} r + W_{ug} u\right).
$$

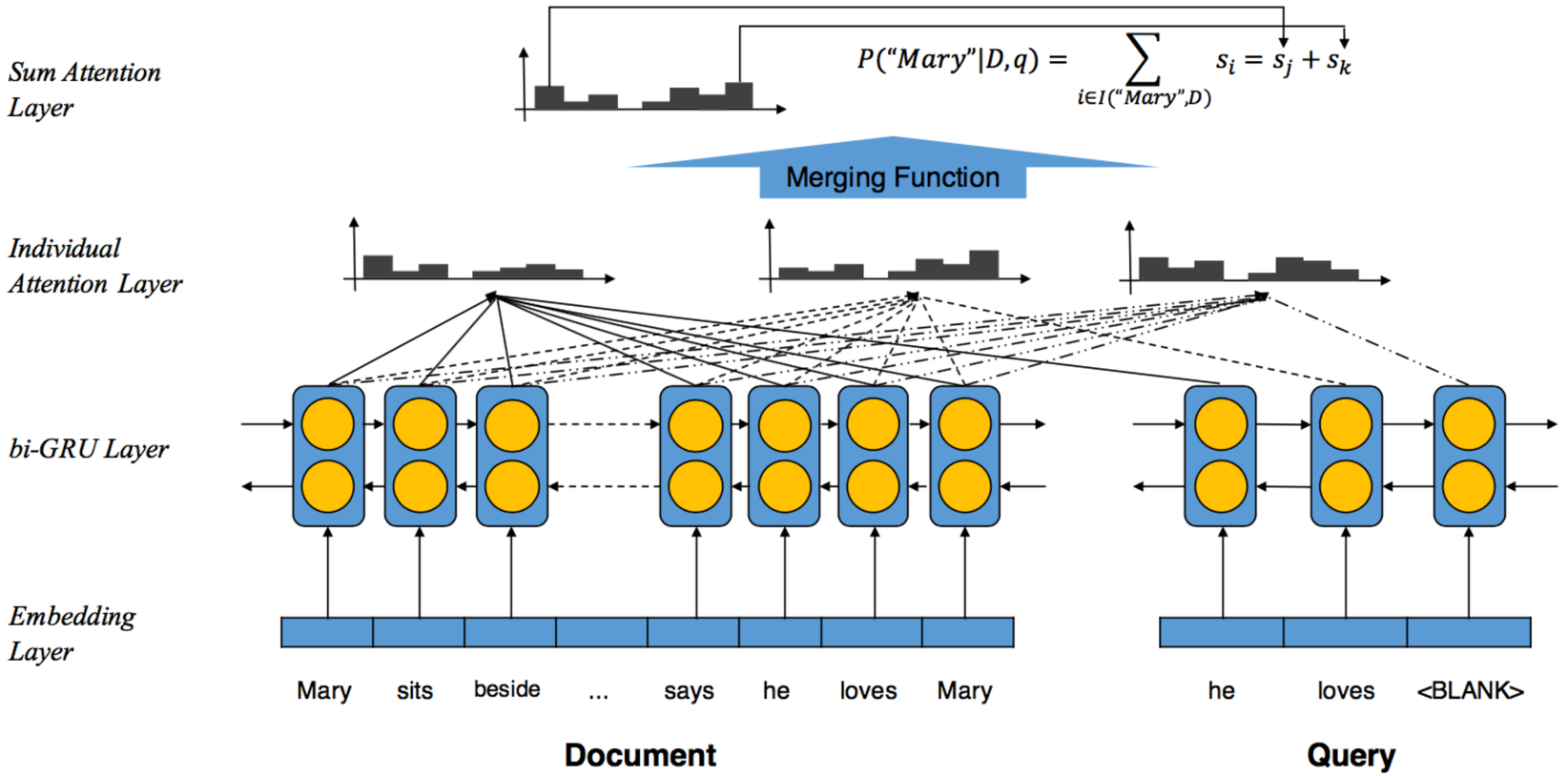
Attention Sum Reader

- Text Understanding with the Attention Sum Reader Network (Kadlec et al., 2016)



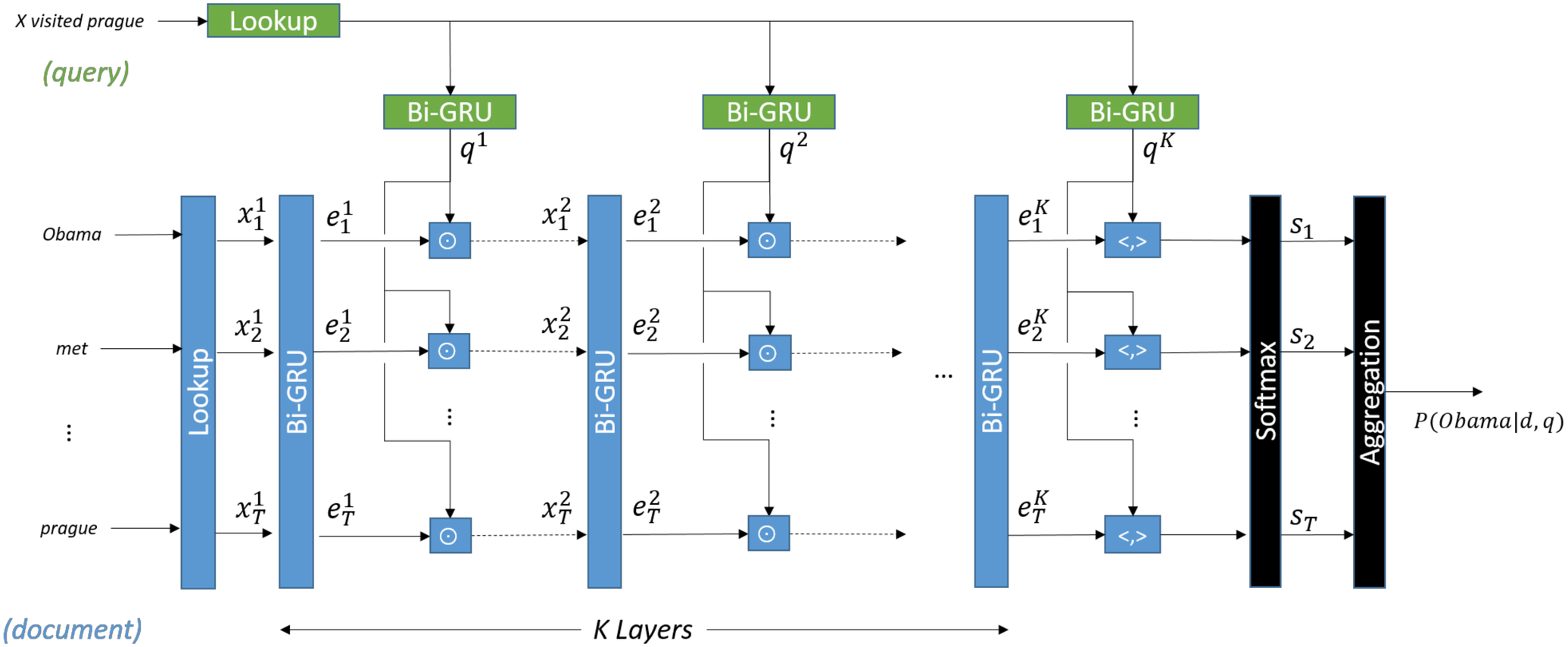
CONSENSUS ATTENTION READER

- Consensus Attention-based Neural Networks for Chinese Reading Comprehension (Cui et al., 2016)



GATED-ATTENTION READER

- Gated-Attention Reader for Text Comprehension (Dhingra et al., 2017)



AoA Reader

- Motivated by
  - AS Reader (Kadlec et al., ACL2016)
  - CAS Reader (Cui et al., COLING2016)
- Some of the components in AoA Reader has been widely adopted in the follow-up works (see the papers cited)

AoA Reader

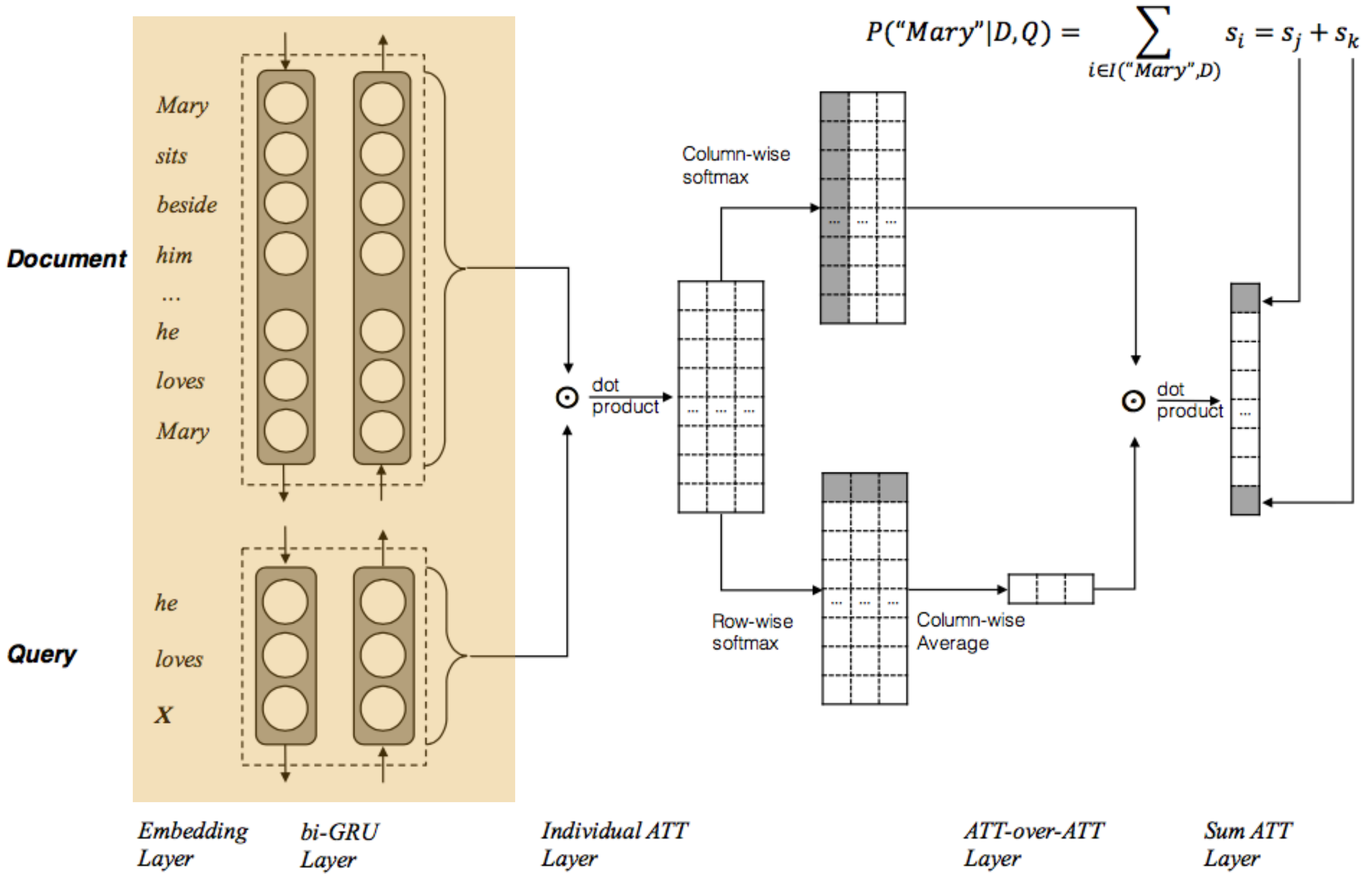
- **Contextual Embedding**
  - Transform document and query into contextual representations using GRU

$$e(x) = W_e \cdot x,\ where\ \ x \in \mathcal{D}, \mathcal{Q} \quad (1)$$

$$\overrightarrow{h_s(x)} = \overrightarrow{GRU}(e(x)) \quad (2)$$

$$\overleftarrow{h_s(x)} = \overleftarrow{GRU}(e(x)) \quad (3)$$

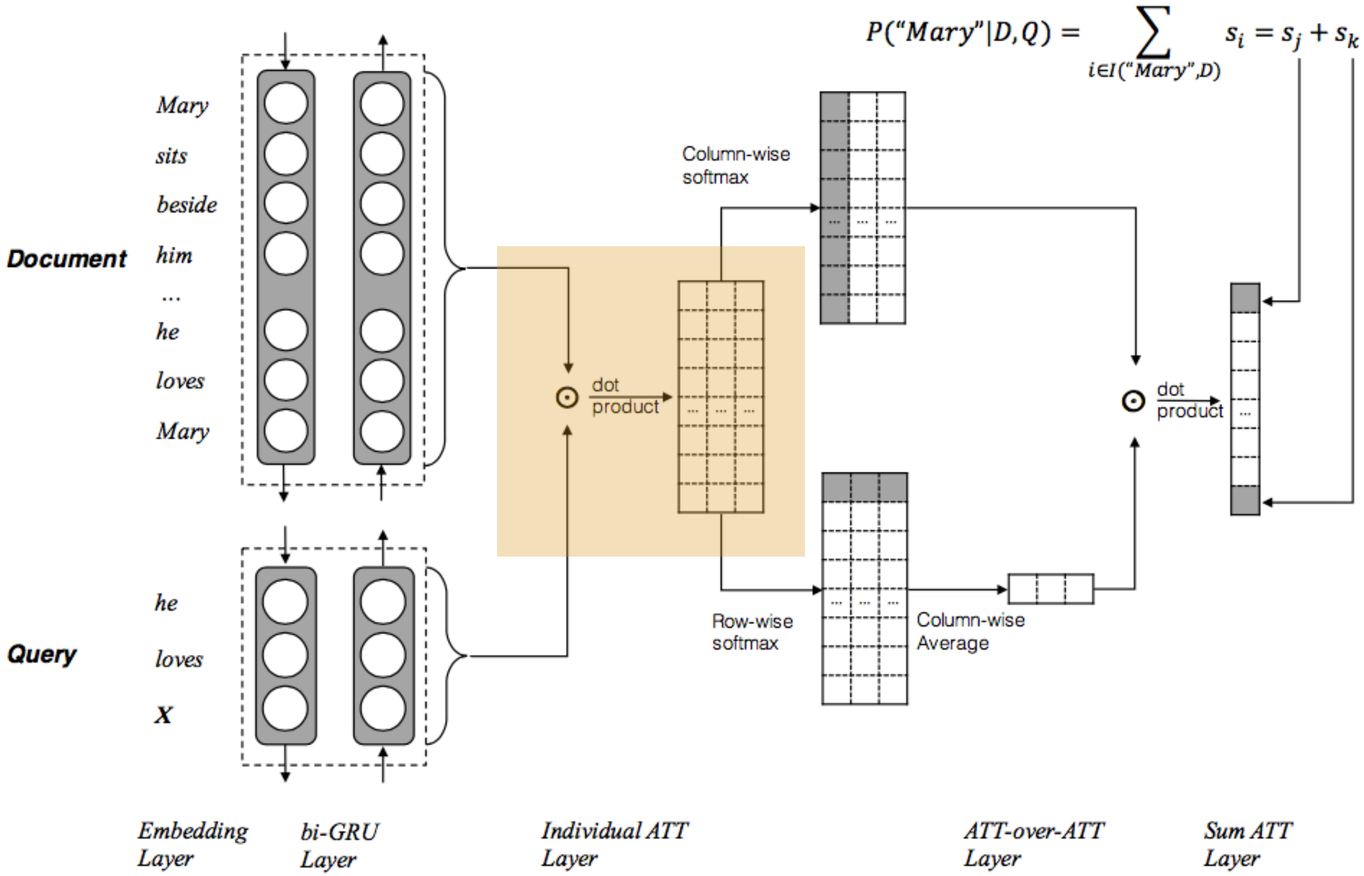
$$h_s(x) = [\overrightarrow{h_s(x)}; \overleftarrow{h_s(x)}] \quad (4)$$



AoA Reader

- **Pair-wise Matching Score**
  - Calculate 'similarity' between each document word and query word

$$M(i,j) = h_{doc}(i)^T \cdot h_{query}(j) \qquad (5)$$

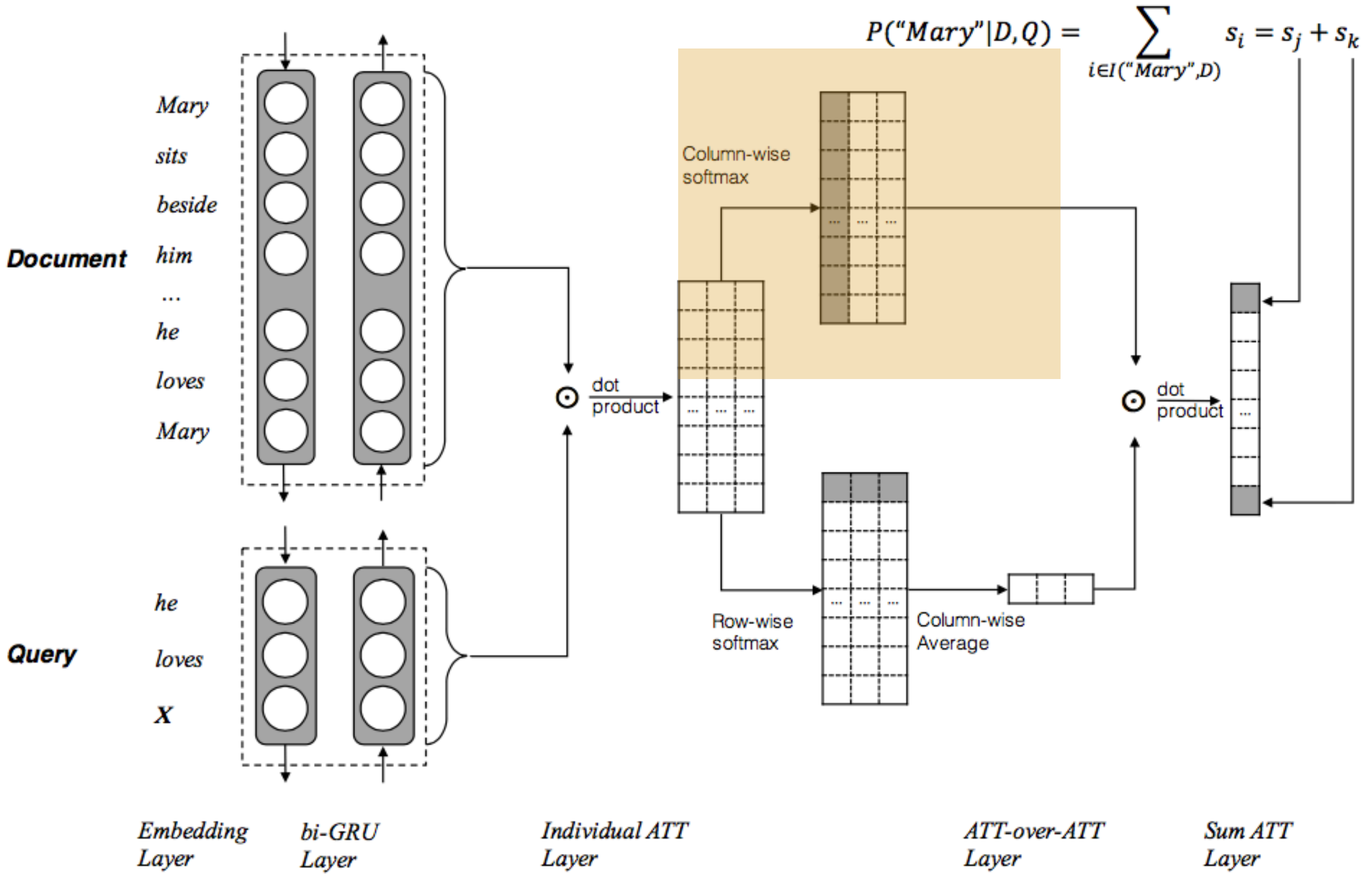


AoA Reader

- **Individual Attentions**
  - Calculate attention with respect to each query word

$$\alpha(t) = softmax(M(1,t), ..., M(|\mathcal{D}|,t)) \quad (6)$$

$$\alpha = [\alpha(1), \alpha(2), ..., \alpha(|\mathcal{Q}|)] \quad (7)$$



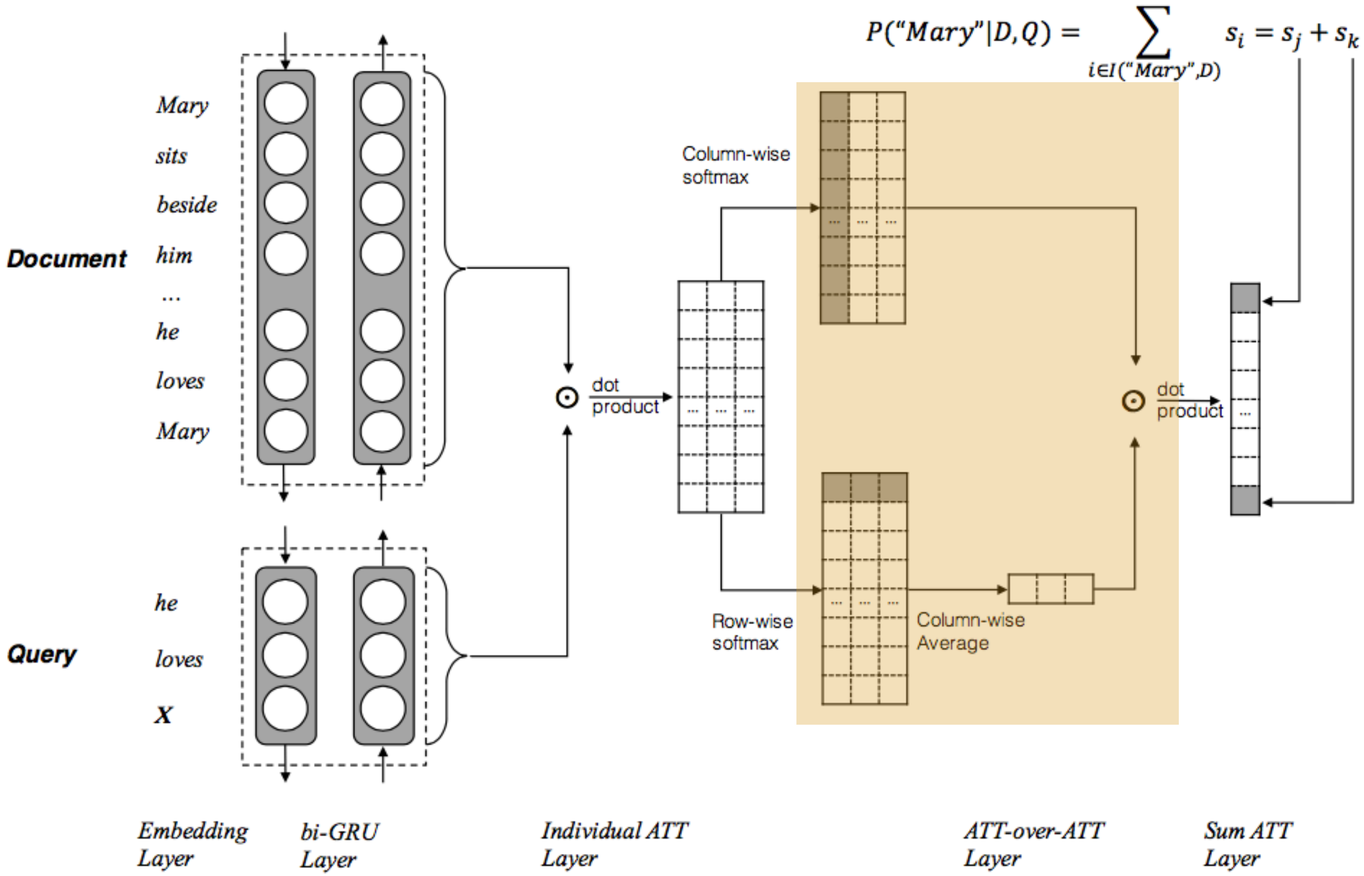
AoA Reader

- **Attention-over-Attention**
  - Dynamically assign weights to individual attentions

$$\beta(t) = softmax(M(t,1), ..., M(t,|\mathcal{Q}|)) \quad (8)$$

$$\beta = \frac{1}{n}\sum_{t=1}^{|\mathcal{D}|}\beta(t) \quad (9)$$

$$s = \alpha^T \beta \quad (10)$$

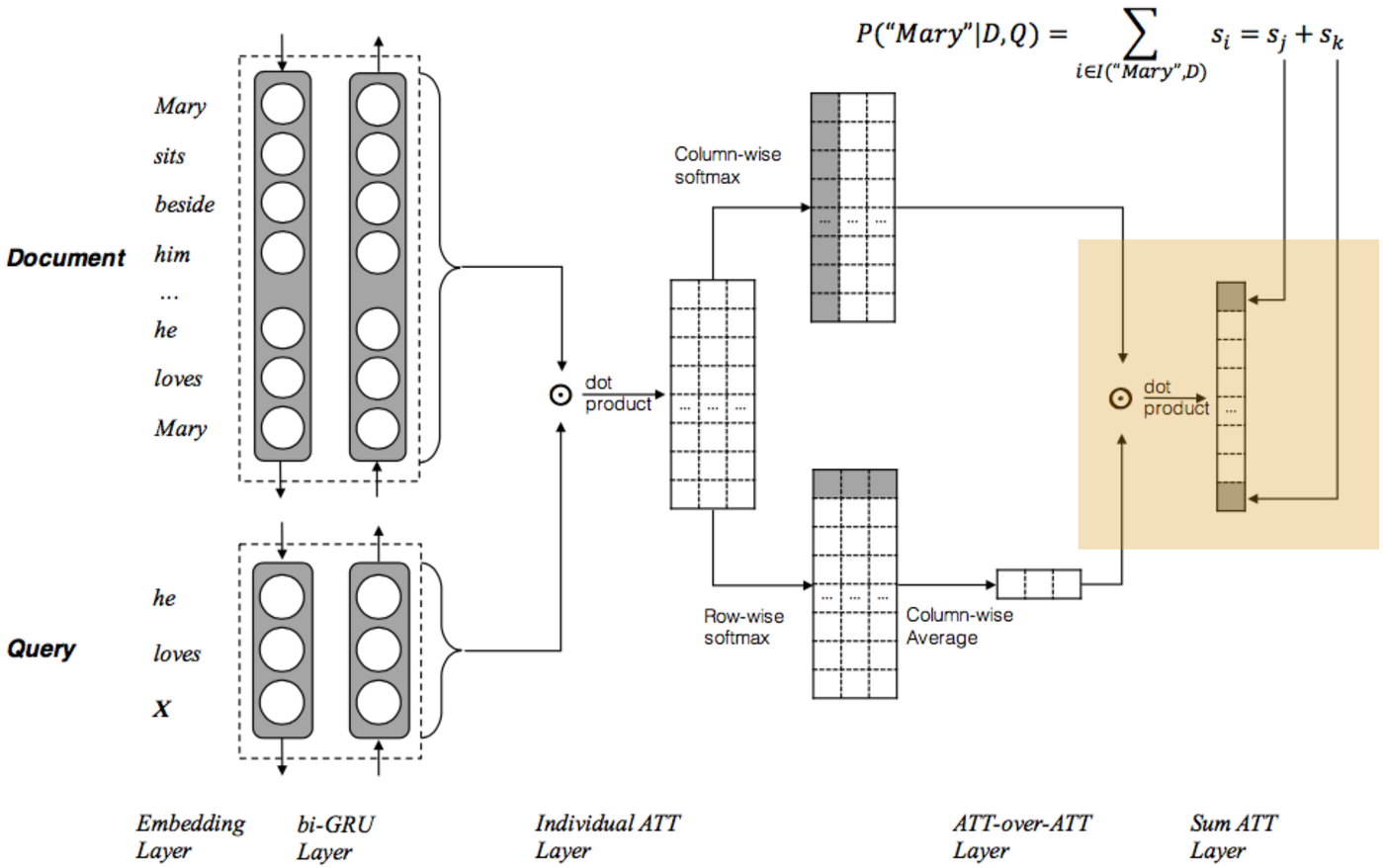


AoA Reader

- **Final Predictions**
  - Apply sum-attention mechanism (Kadlec et al., 2016) to get the final probability of the answer

$$P(w|\mathcal{D},\mathcal{Q}) = \sum_{i \in I(w,\mathcal{D})} s_i, \quad w \in V \qquad (11)$$

$$\mathcal{L} = \sum_i \log(p(x)) \quad , x \in \mathcal{A} \qquad (12)$$



EXPERIMENTS

- **Dataset**
  - CNN(Hermann et al., 2015) and CBT-NE/CN (Hill et al., 2015)
- **Parameters**
  - Embedding: uniform distribution [-0.05, 0.05] with l2-regularization, dropout 0.1
  - Hidden Layer: bi-GRU
  - Optimization: Adam(lr=0.001), gradient clipping 5, batch 32
- **Framework**: Theano + Keras

Experimental Results

- Single model performance

	CNN News		CBTest NE		CBTest CN	
	Valid	Test	Valid	Test	Valid	Test
Deep LSTM Reader (Hermann et al., 2015)	55.0	57.0	-	-	-	-
Attentive Reader (Hermann et al., 2015)	61.6	63.0	-	-	-	-
Human (context+query) (Hill et al., 2015)	-	-	-	*81.6*	-	*81.6*
MemNN (window + self-sup.) (Hill et al., 2015)	63.4	66.8	70.4	66.6	64.2	63.0
AS Reader (Kadlec et al., 2016)	68.6	69.5	73.8	68.6	68.8	63.4
CAS Reader (Cui et al., 2016)	68.2	70.0	74.2	69.2	68.2	65.7
Stanford AR (Chen et al., 2016)	72.4	72.4	-	-	-	-
GA Reader (Dhingra et al., 2016)	73.0	73.8	74.9	69.0	69.0	63.9
Iterative Attention (Sordoni et al., 2016)	72.6	73.3	75.2	68.6	72.1	69.2
EpiReader (Trischler et al., 2016)	**73.4**	74.0	75.3	69.7	71.5	67.4
AoA Reader	73.1	**74.4**	**77.8**	**72.0**	**72.2**	**69.4**

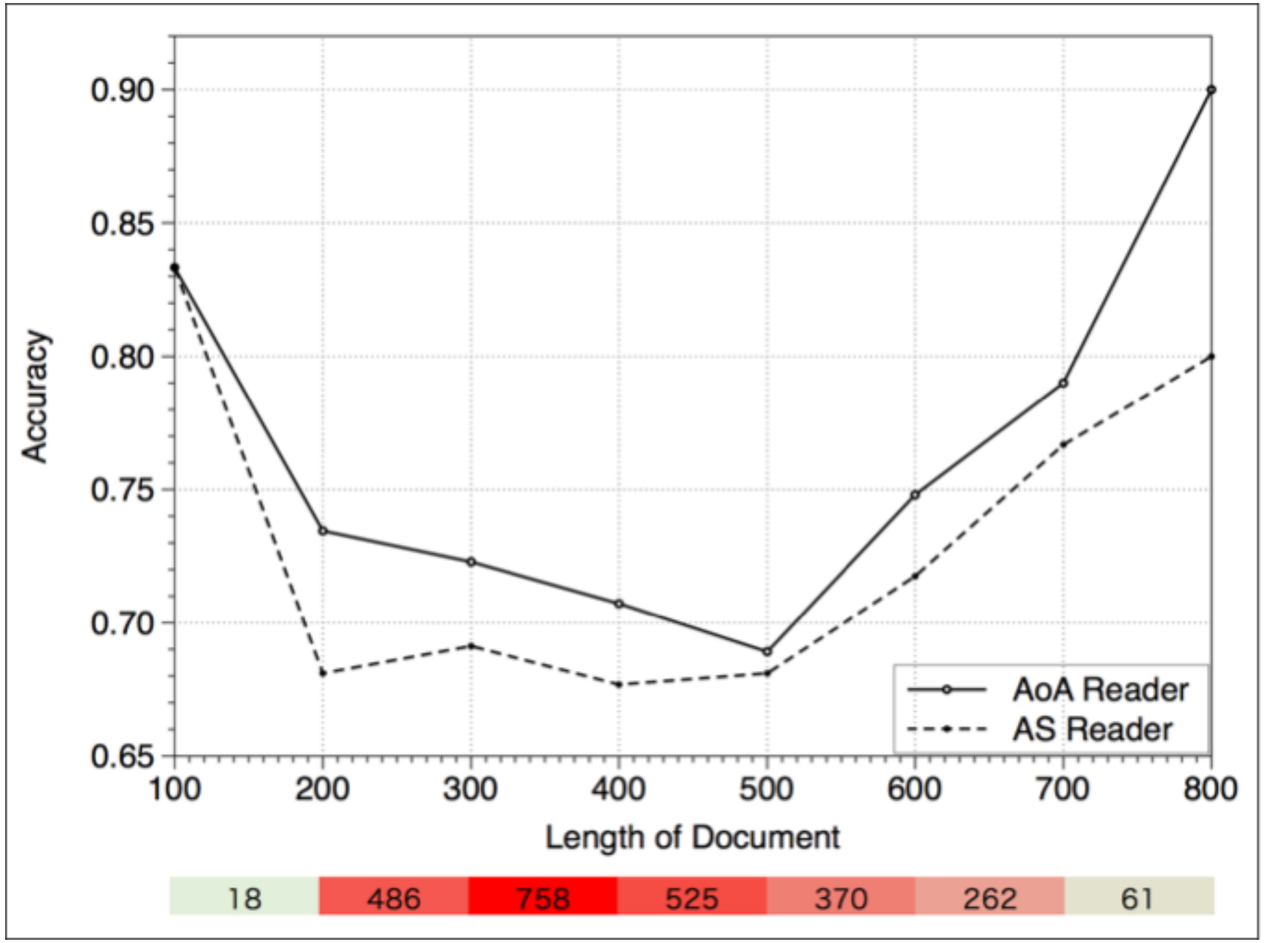
Experimental Results

- **Ensemble performance**
  - We use 4-model greedy ensemble approach

	CNN News		CBTest NE		CBTest CN	
	Valid	Test	Valid	Test	Valid	Test
MemNN (Ensemble)	66.2	69.4	-	-	-	-
AS Reader (Ensemble)	73.9	75.4	74.5	70.6	71.1	68.9
GA Reader (Ensemble)	76.4	77.4	75.5	71.9	72.1	69.4
EpiReader (Ensemble)	-	-	76.6	71.8	73.6	70.6
Iterative Attention (Ensemble)	74.5	75.7	76.9	72.0	74.1	**71.0**
AoA Reader (Ensemble)	-	-	**78.9**	**74.5**	**74.7**	70.8

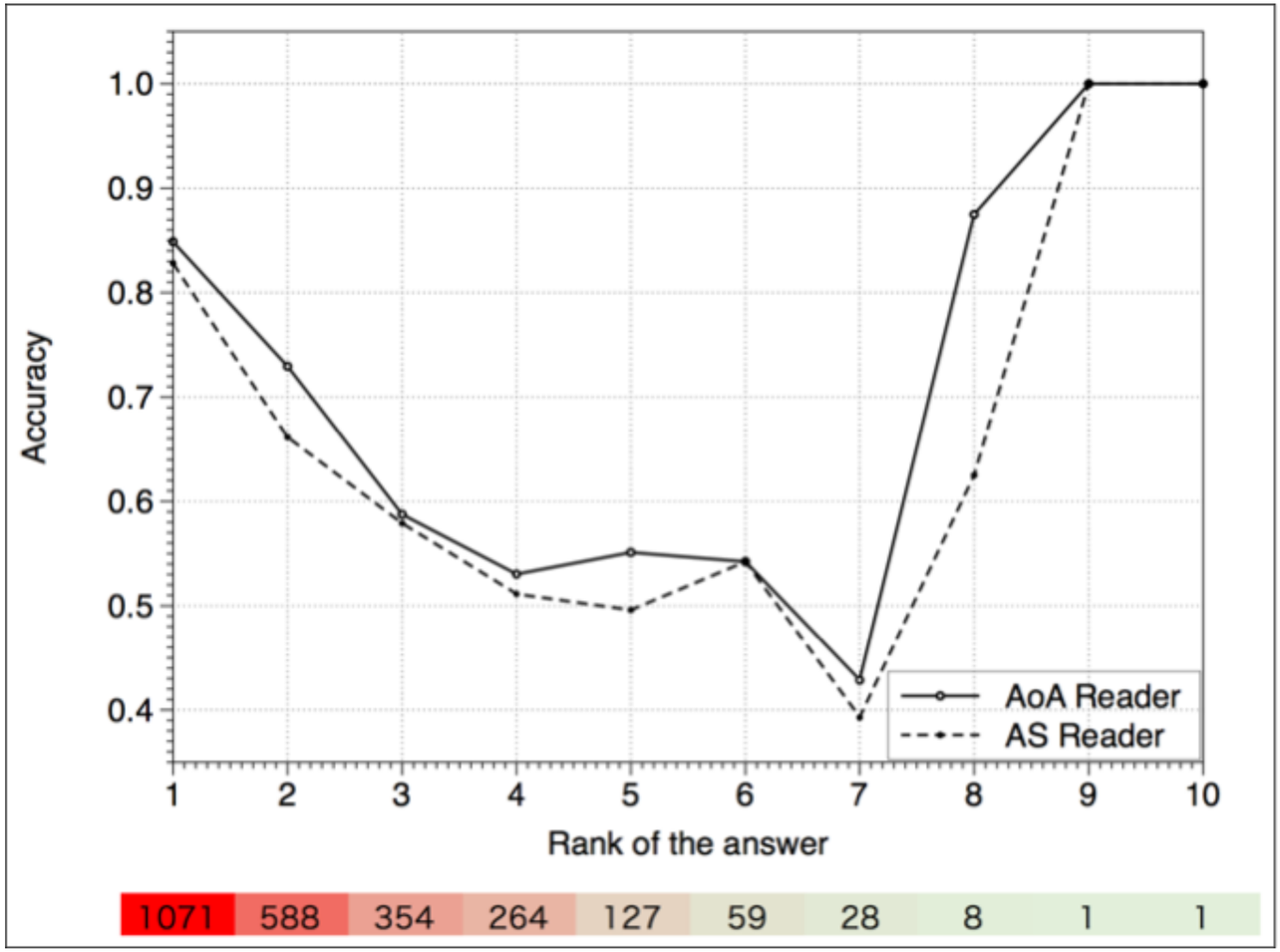
ANALYSIS

- Accuracy v.s. Length of Document
  - AoA Reader shows consistent improvements over AS Reader on different length of document
  - The improvements become larger when the length of document increases



ANALYSIS

- Accuracy v.s. Frequency of answer
  - Most of the answers are the top frequent word among candidates
  - Tend to choose either high or low frequency word

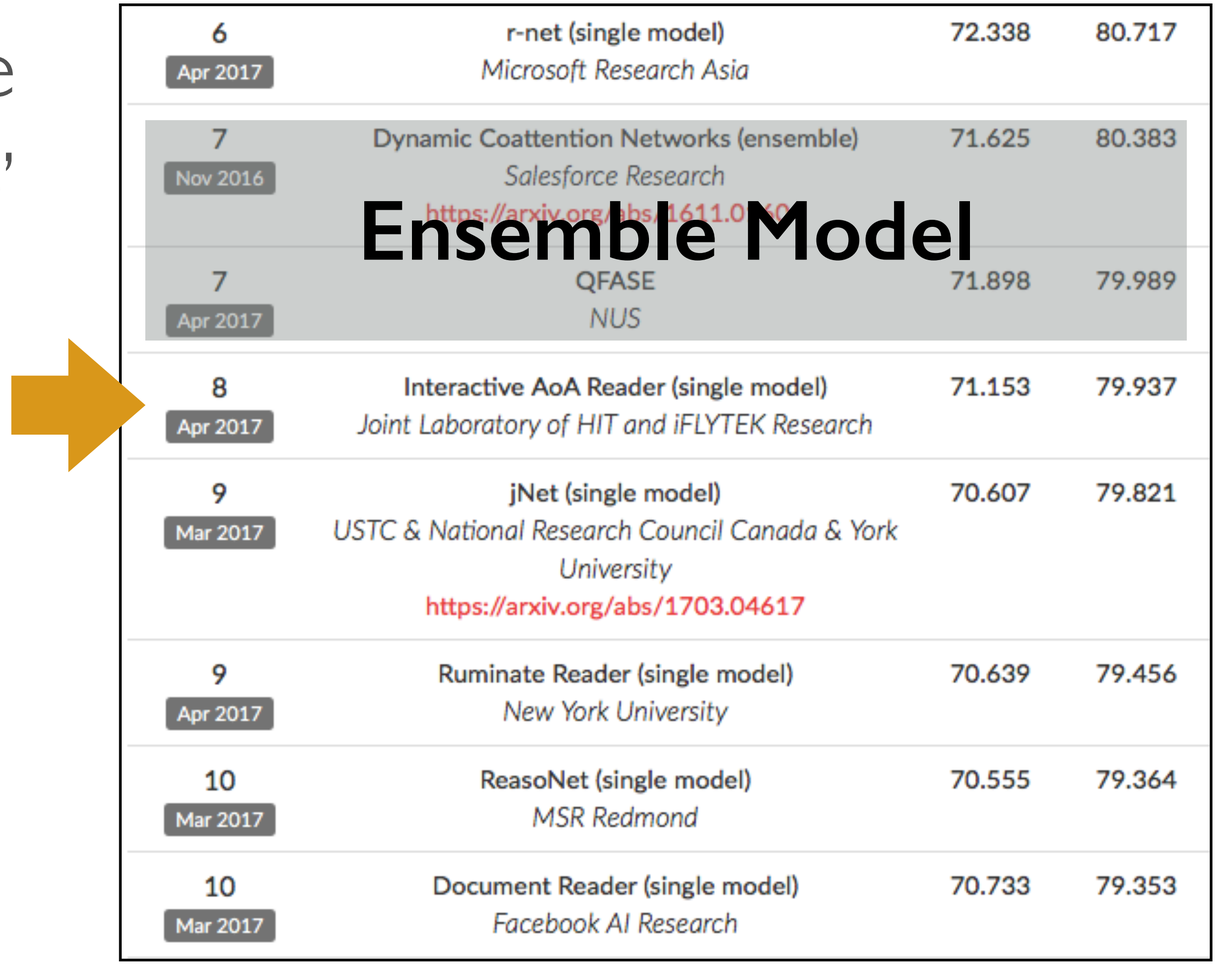


Conclusions

- Propose a novel mechanism called "**Attention-over-Attention**" to dynamically calculate weights between individual attentions
- Adopting both doc-to-query and query-to-doc attentions for final predictions
- Experimental results show significant improvements over various systems
- **Future Works**
  - Investigate more complex attention mechanism
  - Look into the problems that need comprehensive reasoning over several sentences

Interactive AoA Reader

- As a step further of our work, we upgrade our model as 'interactive'
- Shows good performance on Stanford SQuAD RC Task
- No.2 in single model ranking



6 Apr 2017	r-net (single model) *Microsoft Research Asia*	72.338	80.717
7 Nov 2016	Dynamic Coattention Networks (ensemble) *Salesforce Research* https://arxiv.org/abs/1611.01604	71.625	80.383
7 Apr 2017	QFASE *NUS*	71.898	79.989
8 Apr 2017	**Interactive AoA Reader (single model)** *Joint Laboratory of HIT and iFLYTEK Research*	71.153	79.937
9 Mar 2017	jNet (single model) *USTC & National Research Council Canada & York University* https://arxiv.org/abs/1703.04617	70.607	79.821
9 Apr 2017	Ruminate Reader (single model) *New York University*	70.639	79.456
10 Mar 2017	ReasoNet (single model) *MSR Redmond*	70.555	79.364
10 Mar 2017	Document Reader (single model) *Facebook AI Research*	70.733	79.353

CCL-CMRC2017

- The 1st Evaluation Workshop on Chinese Machine Reading Comprehension (CMRC2017)
- Hosted by CIPS, organized by Joint Laboratory of HIT and iFLYTEK (HFL)
- Co-located with CCL2017(2017.10.13 ~ 2017.10.15, Nanjing)
- Welcome to join us!

CCL阅读理解评测

事件	时间
预报名	2017年4月5日 ~ 2017年4月17日
正式报名	2017年4月18日 ~ 2017年4月25日
发布训练集和开发集	2017年5月3日
系统搭建及调整	2017年5月3日 ~ 2017年7月31日
提交系统验证开发集	2017年7月1日 ~ 2017年7月31日
提交系统验证测试集	2017年8月1日 ~ 2017年8月3日
撰写系统描述	2017年8月18日
召开CMRC2017大会	2017年10月13日

More Information

- **Paper download (through arXiv)**
  - https://arxiv.org/abs/1607.04423
- **General training tips & Leaderboard of Cloze-style RC (updates irregularly)**
  - https://github.com/ymcui/Eval-on-NN-of-RC
- **Personal Website (slides and new paper will be uploaded soon)**
  - http://ymcui.github.io

References

- Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473 .

- Danqi Chen, Jason Bolton, and Christopher D. Manning. 2016. A thorough examination of the cnn/daily mail reading comprehension task. In Association for Computational Linguistics (ACL).

- Colin Cherry and George Foster. 2012. Batch tuning strategies for statistical machine translation. In Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Lan- guage Technologies. Association for Computational Linguistics, Montre´al, Canada, pages 427–436. http://www.aclweb.org/anthology/N12-1047.

- Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics, pages 1724–1734. http://aclweb.org/anthology/D14-1179.

- Franc¸ois Chollet. 2015. Keras. https://github. com/fchollet/keras.

References

- Yiming Cui, Ting Liu, Zhipeng Chen, Shijin Wang, and Guoping Hu. 2016. Consensus attention-based neural networks for chinese reading comprehension. In Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers. Osaka, Japan, pages 1777–1786.

- Bhuwan Dhingra, Hanxiao Liu, William W Cohen, and Ruslan Salakhutdinov. 2016. Gated-attention readers for text comprehension. arXiv preprint arXiv:1606.01549 .

- Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In Advances in Neural Information Processing Systems. pages 1684–1692.

- Felix Hill, Antoine Bordes, Sumit Chopra, and Jason Weston. 2015. The goldilocks principle: Reading children's books with explicit memory representations. arXiv preprint arXiv:1511.02301 .

- Franz Josef Och. 1999. An efficient method for determining bilingual word classes. In Ninth Conference of the European Chapter of the Association for Computational Linguistics. http://aclweb.org/anthology/E99-1010.

References

- Rudolf Kadlec, Martin Schmid, Ondrej Bajgar, and Jan Kleindienst. 2016. Text understanding with the attention sum reader network. arXiv preprint arXiv:1603.01547 .
- Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 .
- Ting Liu, Yiming Cui, Qingyu Yin, Shijin Wang, Weinan Zhang, and Guoping Hu. 2016. Generating and exploiting large-scale pseudo training data for zero pronoun resolution. arXiv preprint arXiv:1606.01603 . In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (ACL-2017), to appear.
- Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. 2013. On the difficulty of training recurrent neural networks. ICML (3) 28:1310–1318.
- Andrew M Saxe, James L McClelland, and Surya Ganguli. 2013. Exact solutions to the nonlinear dynam- ics of learning in deep linear neural networks. arXiv preprint arXiv:1312.6120 .
- Alessandro Sordoni, Phillip Bachman, and Yoshua Bengio. 2016. Iterative alternating neural attention for machine reading. arXiv preprint arXiv:1606.02245

References

- Nitish Srivastava, Geoffrey E Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. Journal of Machine Learning Research 15(1):1929–1958.
- Andreas Stolcke. 2002. Srilm — an extensible lan- guage modeling toolkit. In Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP 2002). pages 901–904.
- Wilson L Taylor. 1953. Cloze procedure: a new tool for measuring readability. Journalism and Mass Communication Quarterly 30(4):415.
- Theano Development Team. 2016. Theano: A Python framework for fast computation of mathematical expressions. arXiv e-prints abs/1605.02688. http://arxiv.org/abs/1605.02688.
- Adam Trischler, Zheng Ye, Xingdi Yuan, and Kaheer Suleman. 2016. Natural language comprehension with the epireader. arXiv preprint arXiv:1606.02270 .
- Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. 2015. Pointer networks. In Advances in Neural Information Processing Systems. pages 2692–2700.

Thank you !

And time to revise camera-ready paper…

Contact: admin [at] ymcui [dot] com