Cloze-style Reading Comprehension

Qingchun Bai

2017-10-11

- [ACL2017] Attention-over-Attention Neural Networks for Reading comprehension
- [ACL2016] Text Understanding with the Attention Sum Reader Network
- [ACL 2017] Gated-Attention Readers for Text comprehension

Codes: https://github.com/bdhingra/ga-reader

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

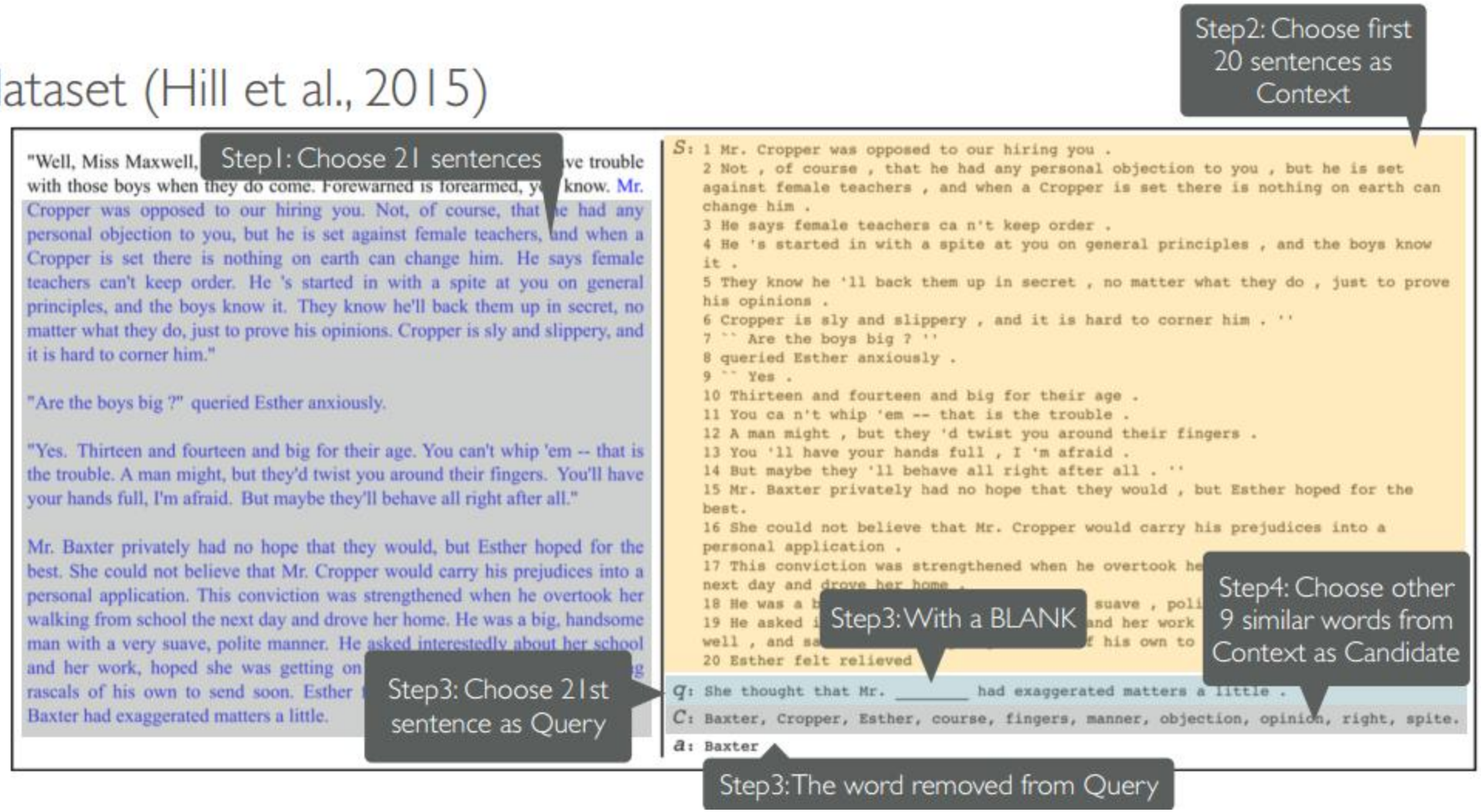
- Specifically, in cloze-style RC
  - Document: the same as the general RC
  - Query: a sentence with a blank
  - Candidate (optional): several candidates to fill in
  - Answer: a single word that exactly match the query (the answer word should appear in the document)

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)



Attention-over-Attention Neural Networks for Reading comprehension

AoA Reader

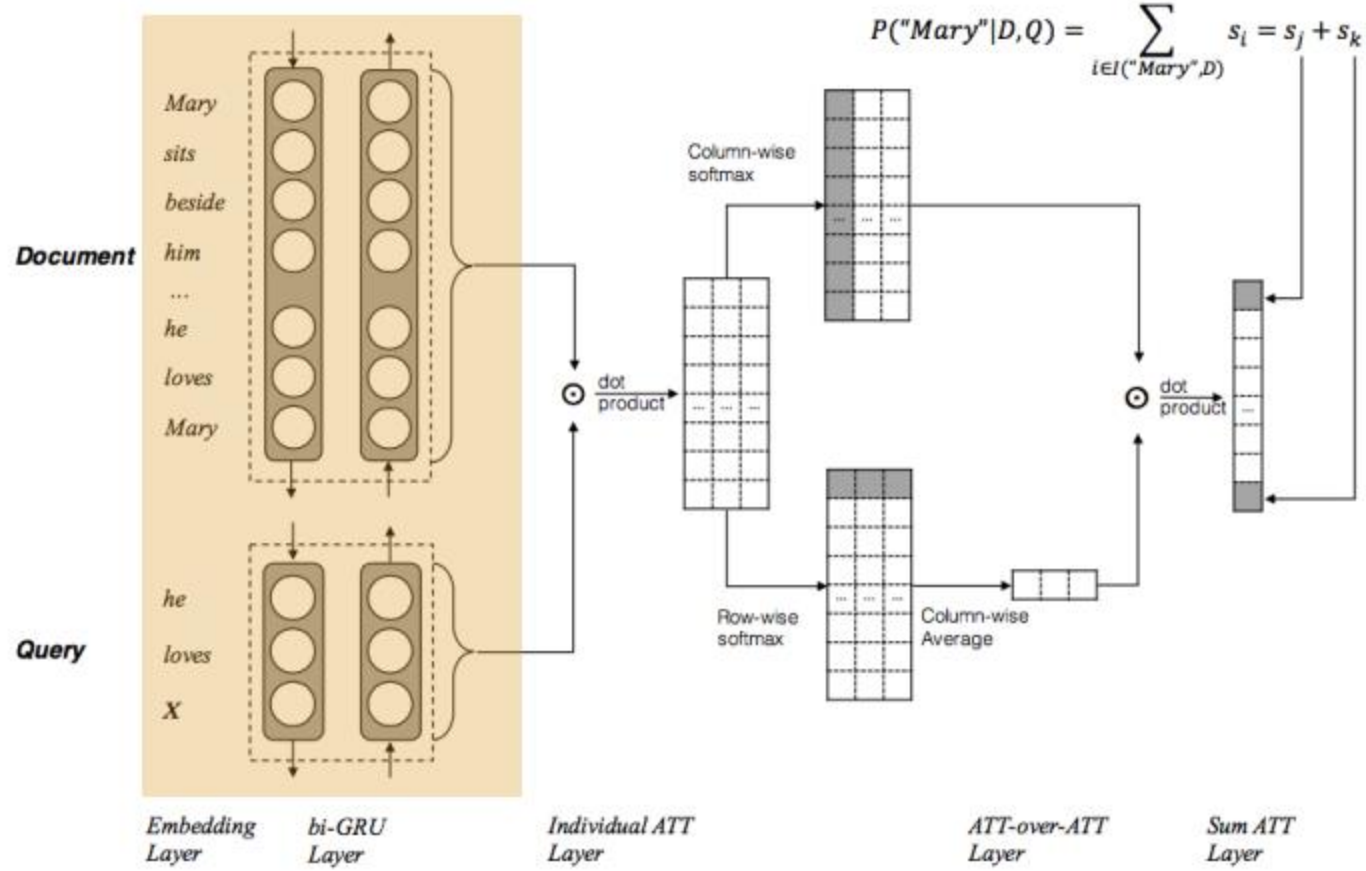
- **Contextual Embedding**
  - Transform document and query into contextual representations using GRU

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

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

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

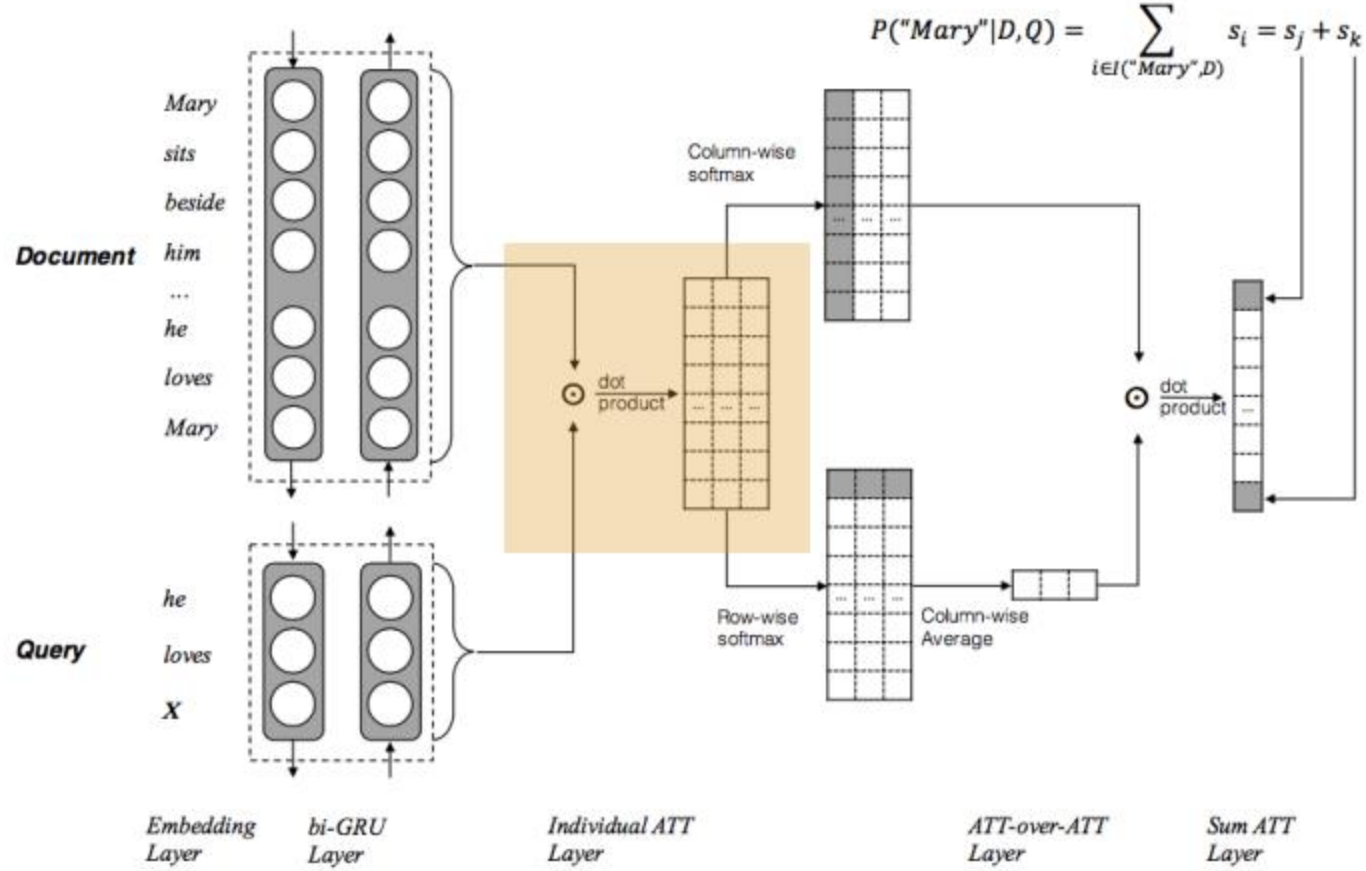
$$h_s(x) = [\overrightarrow{h_s(x)}; \overleftarrow{h_s(x)}] \quad (4)$$



AoA Reader

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

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

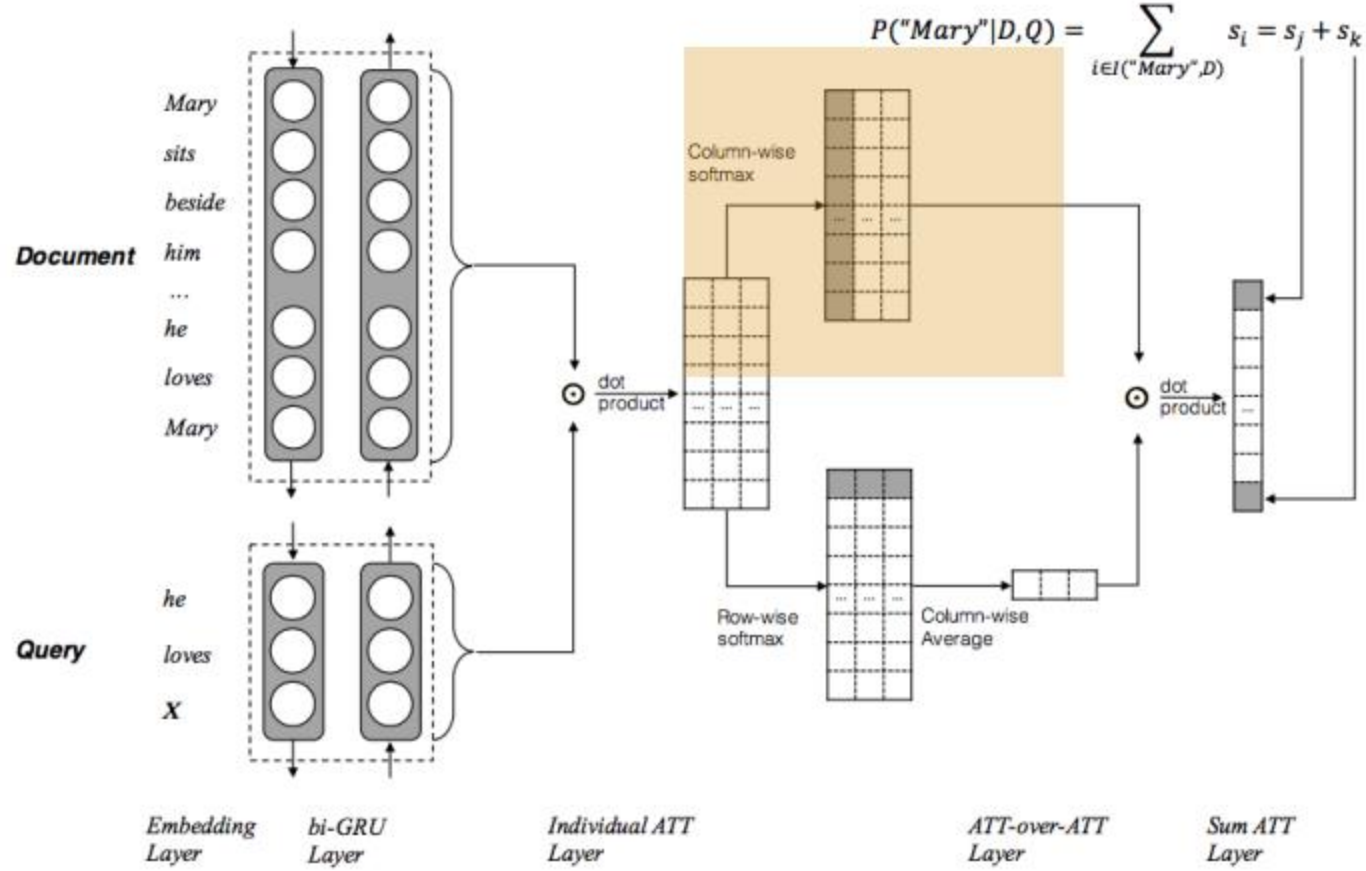


AoA Reader

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

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

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



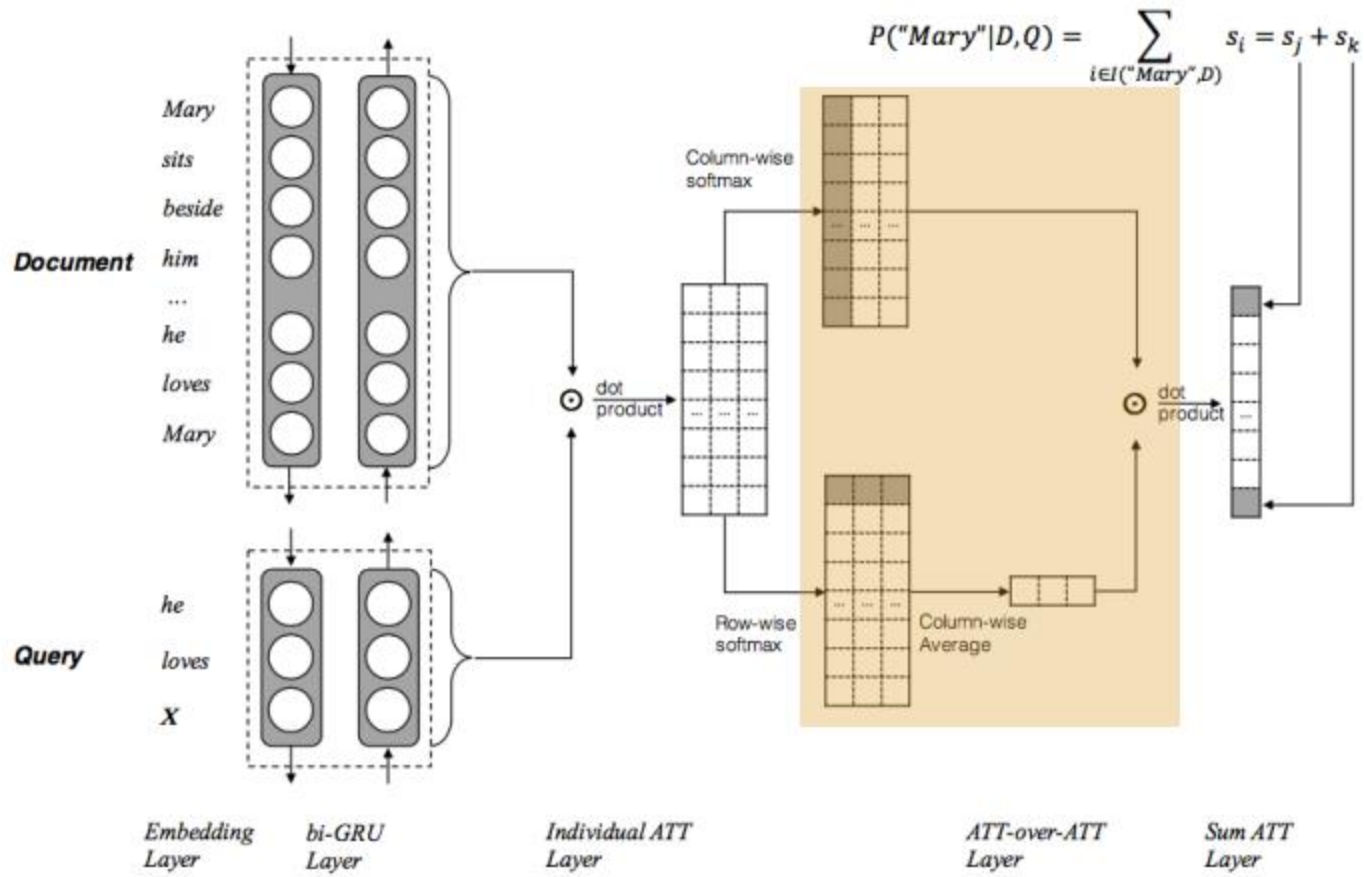
AoA Reader

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

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

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

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

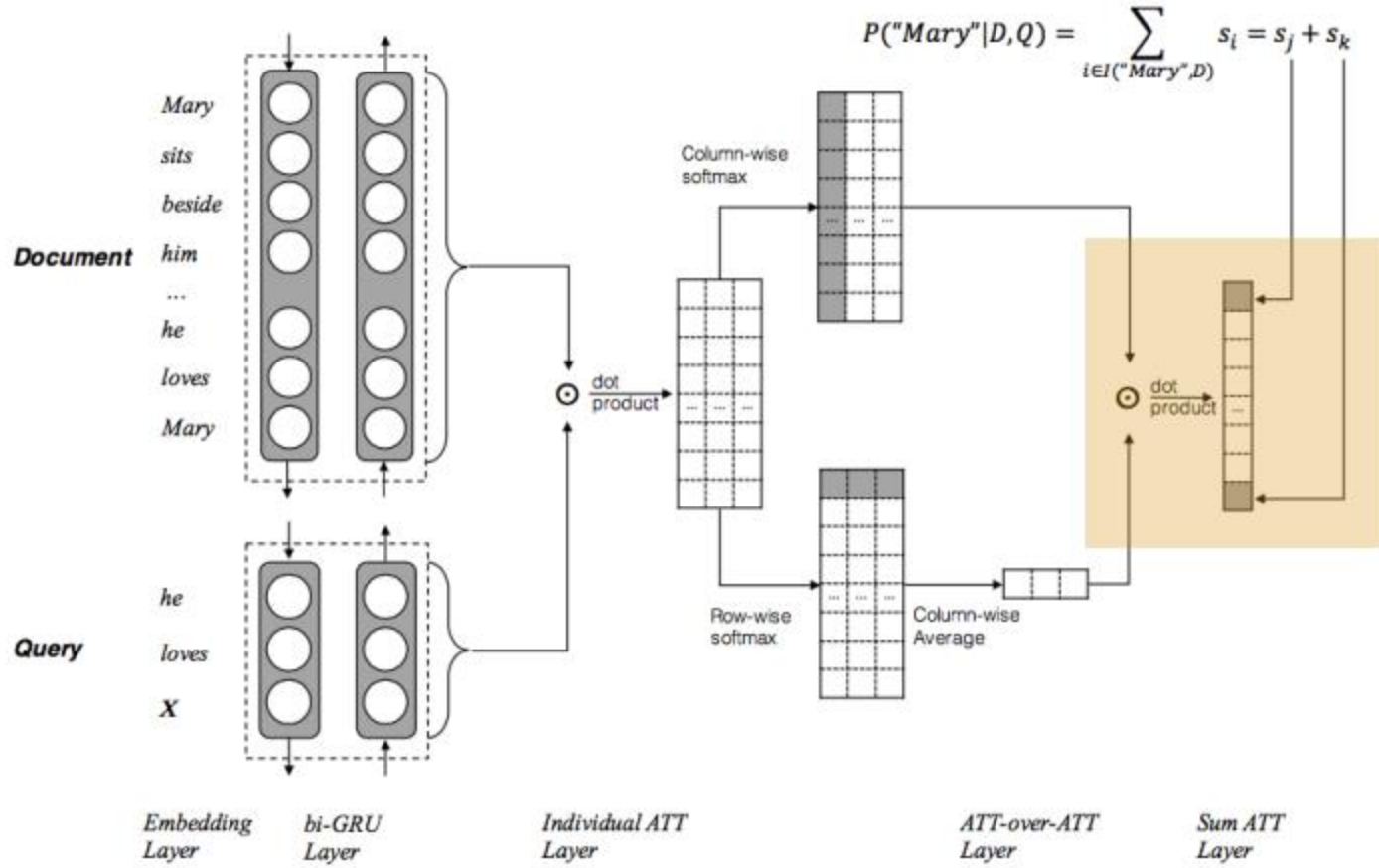


AoA Reader

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

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

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



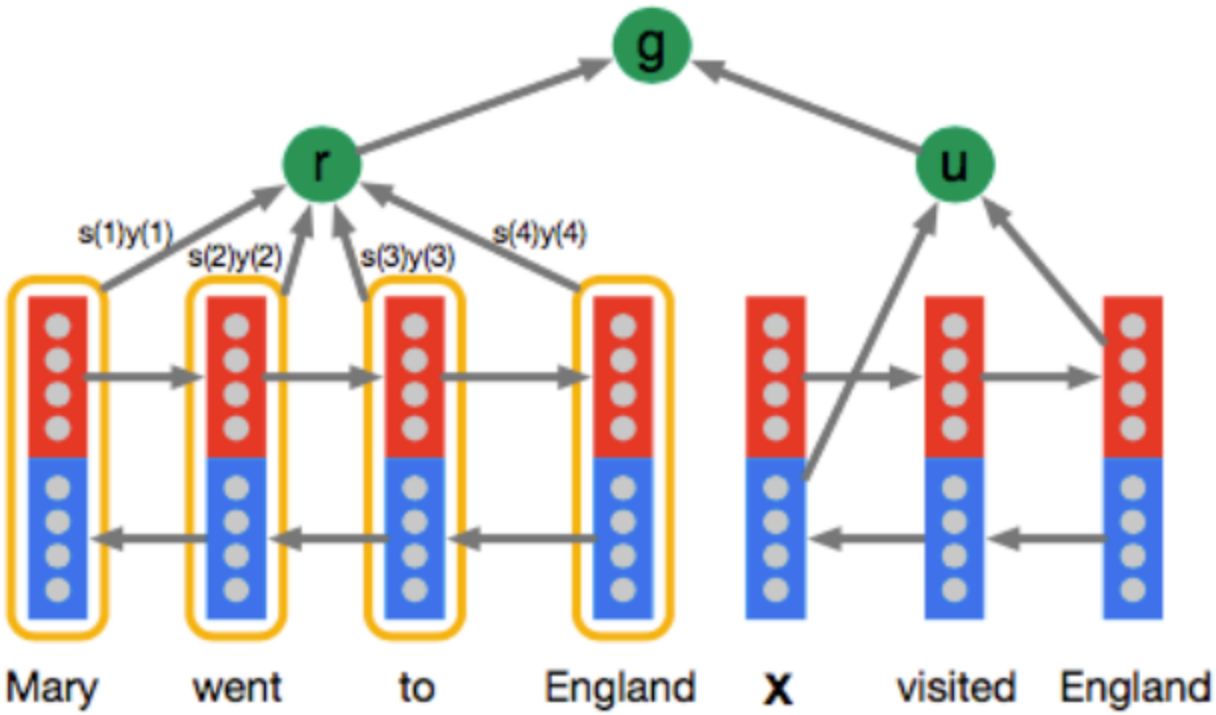
Experimental Results

- 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**

ATTENTIVE READER

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



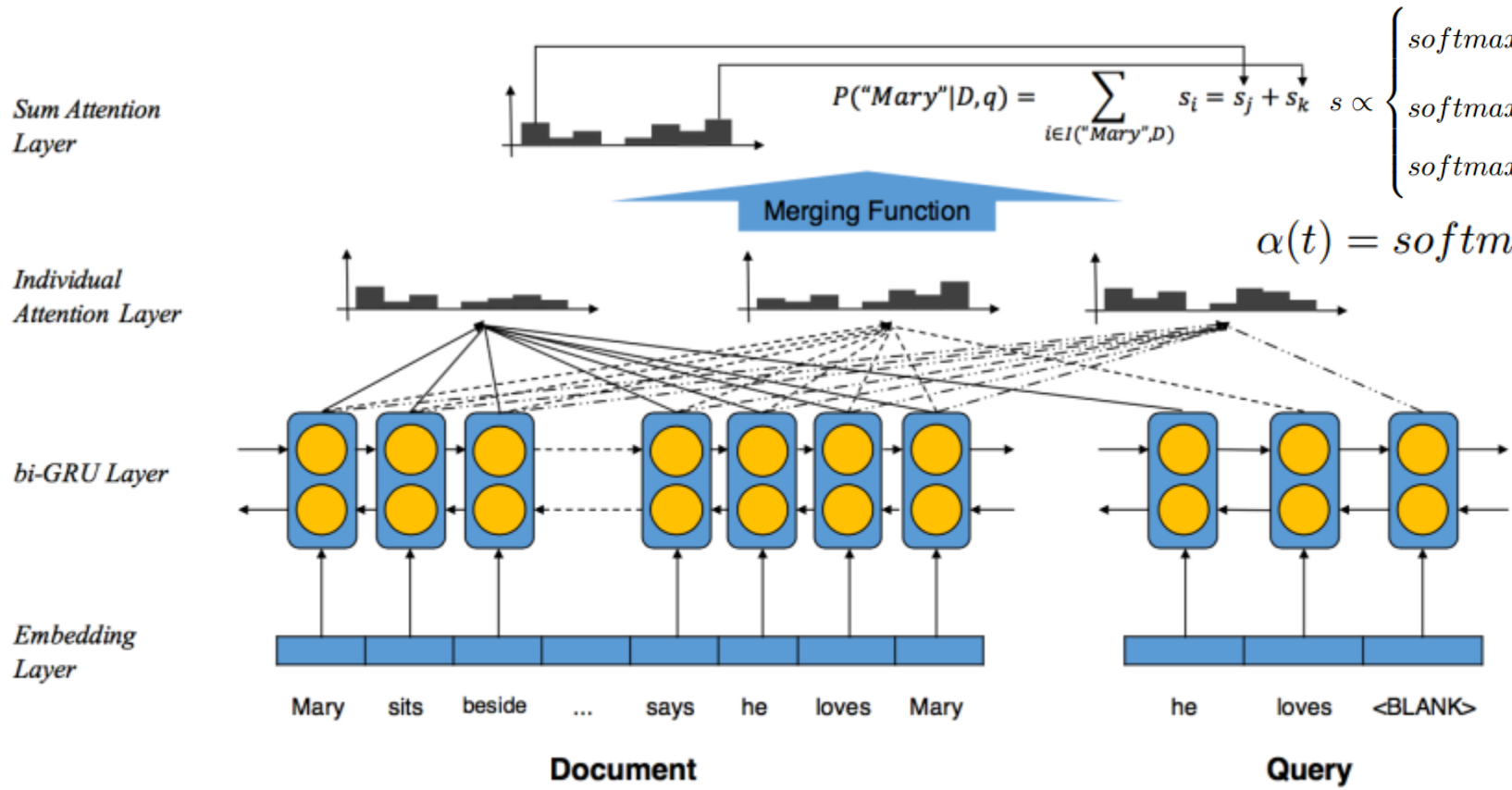
$$
\begin{aligned}
m(t) &= \tanh\left(W_{ym} y_d(t) + W_{um} u\right), \\
s(t) &\propto \exp\left(\mathrm{w}_{ms}^{\intercal} m(t)\right), \\
r &= y_d s,
\end{aligned}
$$

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

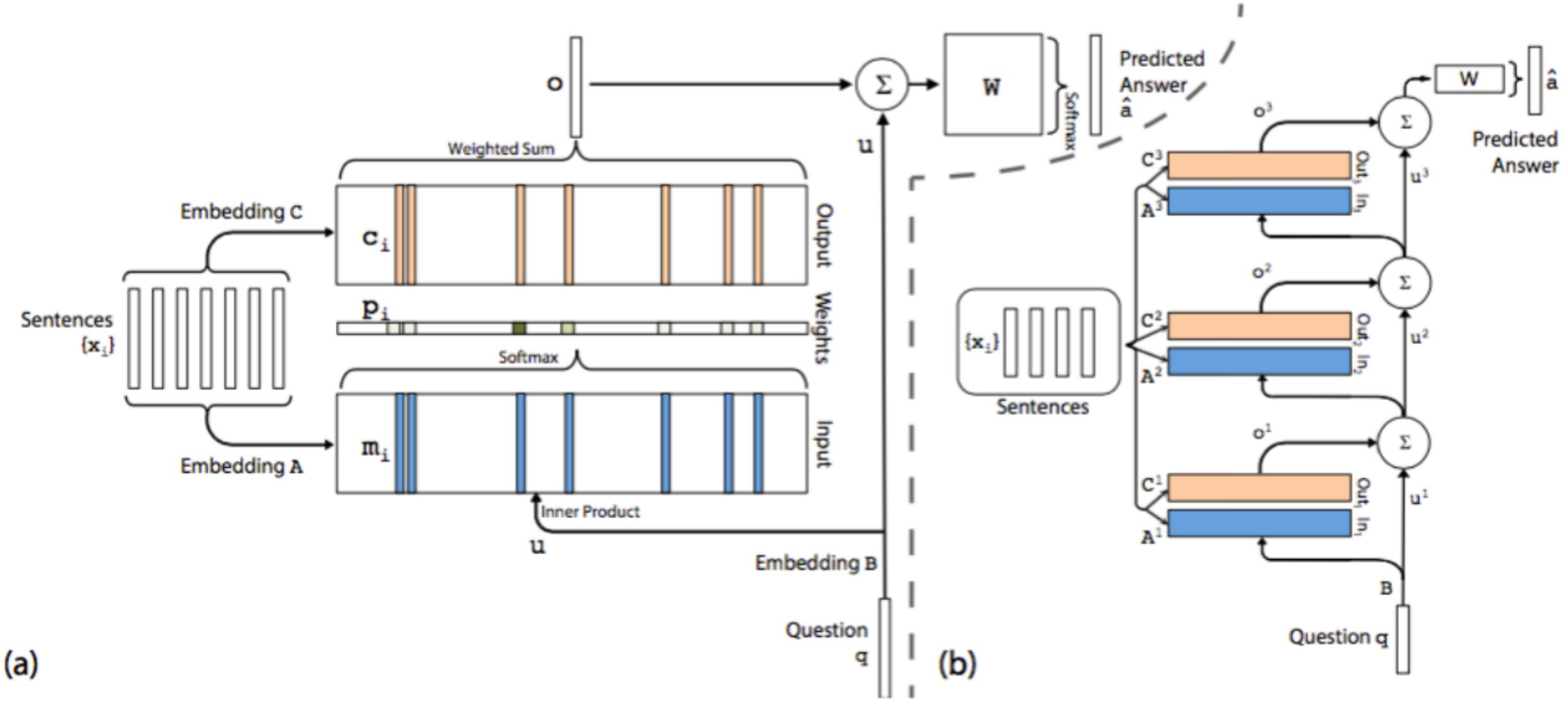
CONSENSUS ATTENTION READER

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

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



Memory Networks



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

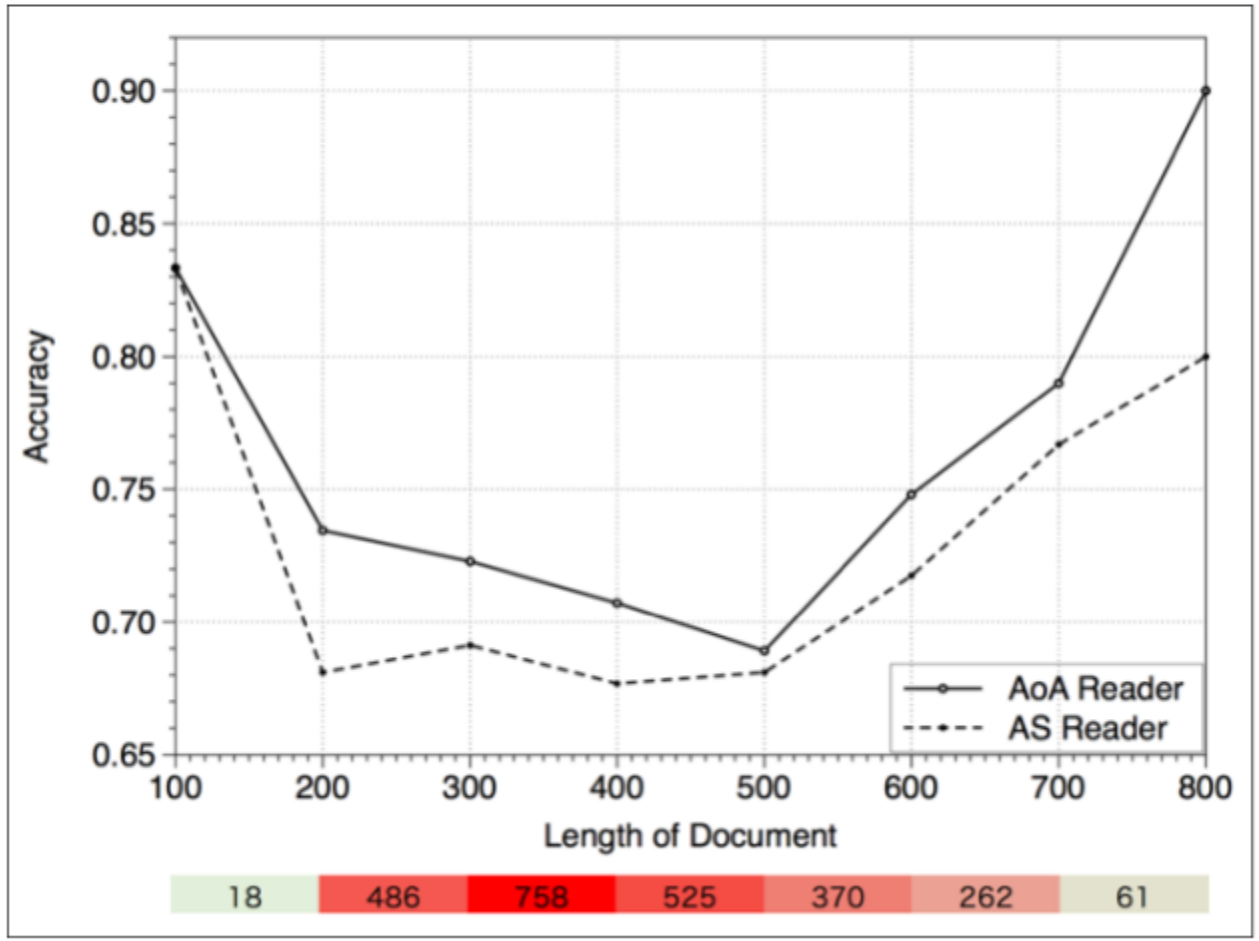
EXPERIMENTAL RESULTS

- **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

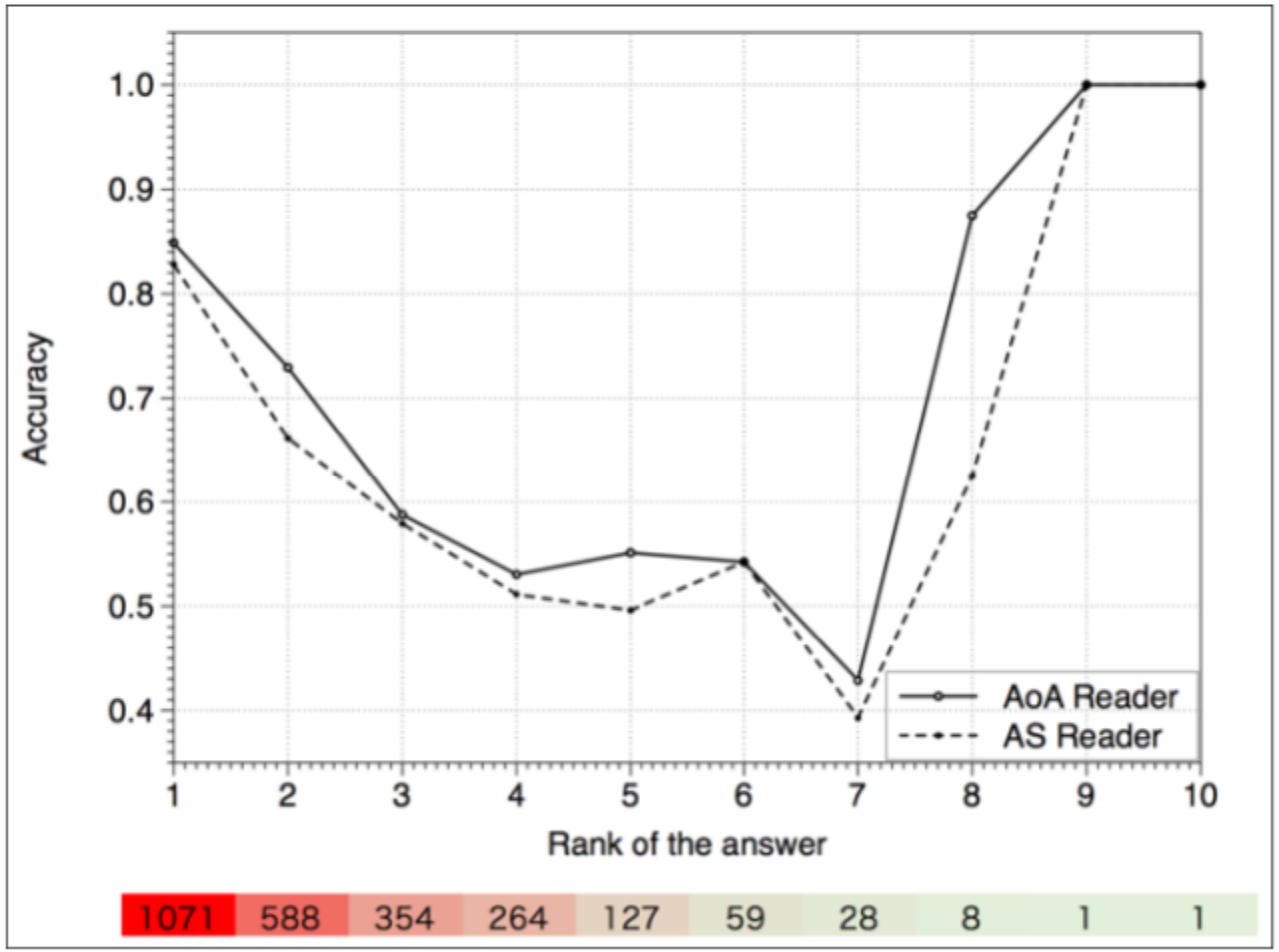
ANALYSIS

- 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

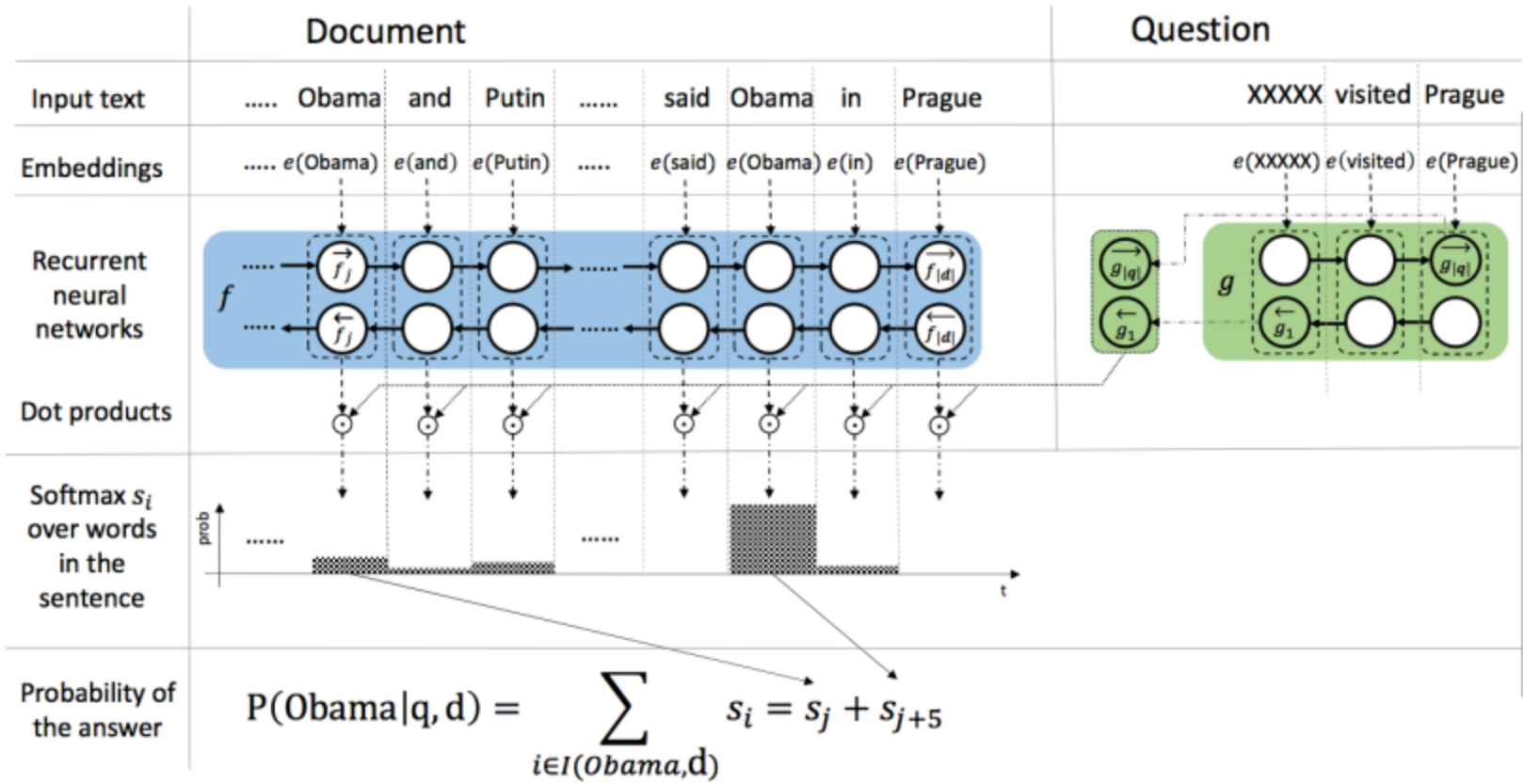
- Accuracy v.s. Frequency of answer
  - Most of the answers are the top frequent word among candidates
  - Tend to choose either high or low frequency word



Text Understanding with the Attention Sum Reader Network

ATTENTION SUM READER

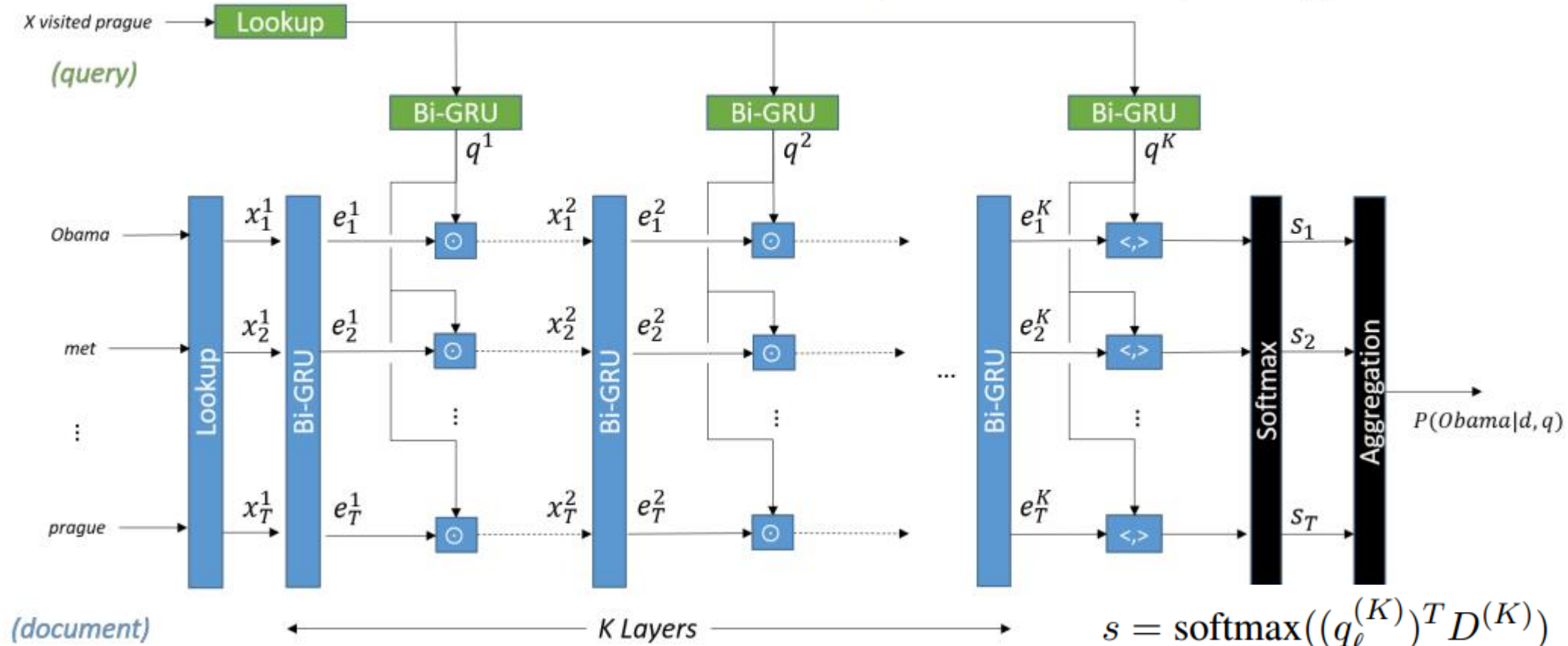
- Text Understanding with the Attention Sum Reader Network (Kadlec et al., 2016)



GATED-ATTENTION READER

GATED-ATTENTION READER

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



$$\alpha_i = \text{softmax}(Q^{\top} d_i)$$

$$\tilde{q}_i = Q\alpha_i$$

$$x_i = d_i \odot \tilde{q}_i$$

$$s = \text{softmax}((q_\ell^{(K)})^T D^{(K)})$$

$$\Pr(c|d,q) \propto \sum_{i \in \mathbb{I}(c,d)} s_i$$

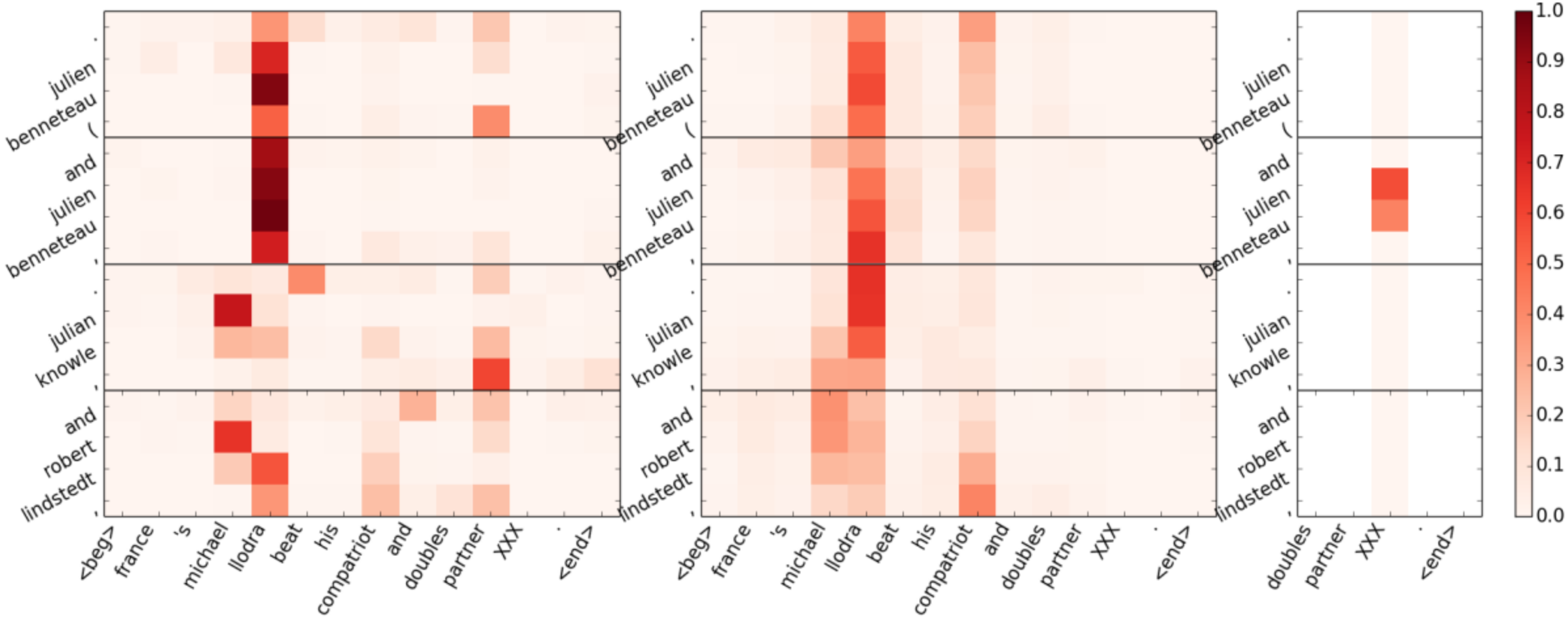
$$a^* = \text{argmax}_{c \in \mathcal{C}} \ \Pr(c|d,q).$$

Table 2: **Top:** Performance of different gating functions. **Bottom:** Effect of varying the number of hops K. Results on WDW without using the qe-comm feature and with fixed $L(w)$.

Gating Function	**Accuracy**	
	Val	Test
Sum	64.9	64.5
Concatenate	64.4	63.7
Multiply	**68.3**	**68.0**
K		
1 (AS) †	–	57
2	65.6	65.6
3	**68.3**	68.0
4	**68.3**	**68.2**

Model	CNN		Daily Mail		CBT-NE		CBT-CN	
	Val	Test	Val	Test	Val	Test	Val	Test
Humans (query) †	–	–	–	–	–	52.0	–	64.4
Humans (context + query) †	–	–	–	–	–	81.6	–	81.6
LSTMs (context + query) †	–	–	–	–	51.2	41.8	62.6	56.0
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	–	–	–	–
MemNets †	63.4	66.8	–	–	70.4	66.6	64.2	63.0
AS Reader †	68.6	69.5	75.0	73.9	73.8	68.6	68.8	63.4
DER Network †	71.3	72.9	–	–	–	–	–	–
Stanford AR (relabeling) †	73.8	73.6	77.6	76.6	–	–	–	–
Iterative Attentive Reader †	72.6	73.3	–	–	75.2	68.6	72.1	69.2
EpiReader †	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
ReasoNet †	72.9	74.7	77.6	76.6	–	–	–	–
NSE †	–	–	–	–	78.2	73.2	74.3	**71.9**
BiDAF †	76.3	76.9	80.3	79.6	–	–	–	–
MemNets (ensemble) †	66.2	69.4	–	–	–	–	–	–
AS Reader (ensemble) †	73.9	75.4	78.7	77.7	76.2	71.0	71.1	68.9
Stanford AR (relabeling,ensemble) †	77.2	77.6	80.2	79.2	–	–	–	–
Iterative Attentive Reader (ensemble) †	75.2	76.1	–	–	76.9	72.0	74.1	71.0
EpiReader (ensemble) †	–	–	–	–	76.6	71.8	73.6	70.6
AS Reader (+BookTest) † ‡	–	–	–	–	80.5	76.2	83.2	80.8
AS Reader (+BookTest,ensemble) † ‡	–	–	–	–	*82.3*	*78.4*	*85.7*	*83.7*
GA--	73.0	73.8	76.7	75.7	74.9	69.0	69.0	63.9
GA (update $L(w)$)	**77.9**	**77.9**	**81.5**	**80.9**	76.7	70.1	69.8	67.3
GA (fix $L(w)$)	77.9	77.8	80.4	79.6	77.2	71.4	71.6	68.0
GA (+feature, update $L(w)$)	77.3	76.9	80.7	80.0	77.2	73.3	73.0	69.8
GA (+feature, fix $L(w)$)	76.7	77.4	80.0	79.3	**78.5**	**74.9**	**74.4**	70.7

Layer-wise attention visualization of GA Reader trained on WDW-Strict. See text for details.



DOC: result sunday from the open 13 , a (euro) 512,750 ($ 697,400) atp world tour indoor hardcourt event at palais des sports (seedings in parentheses) : singl es final michael llodra , france , def . julien benneteau (8) , france , 6-3 , 6-4. doubles final michael llodra and julien benneteau , france (2) , def . j ulian knowle , austria and robert lindstedt , sweden (1) , 6-4 , 6-3 .

QRY: <beg> france 's michael llodra beat his compatriot and doubles partner XXX . <end>
ANS: julien benneteau

Thanks