中国科学院自动化研究所 Institute of Automation, Chinese Academy of Sciences

模式识别国家重点实验室 National Laboratory of Pattern Recognition

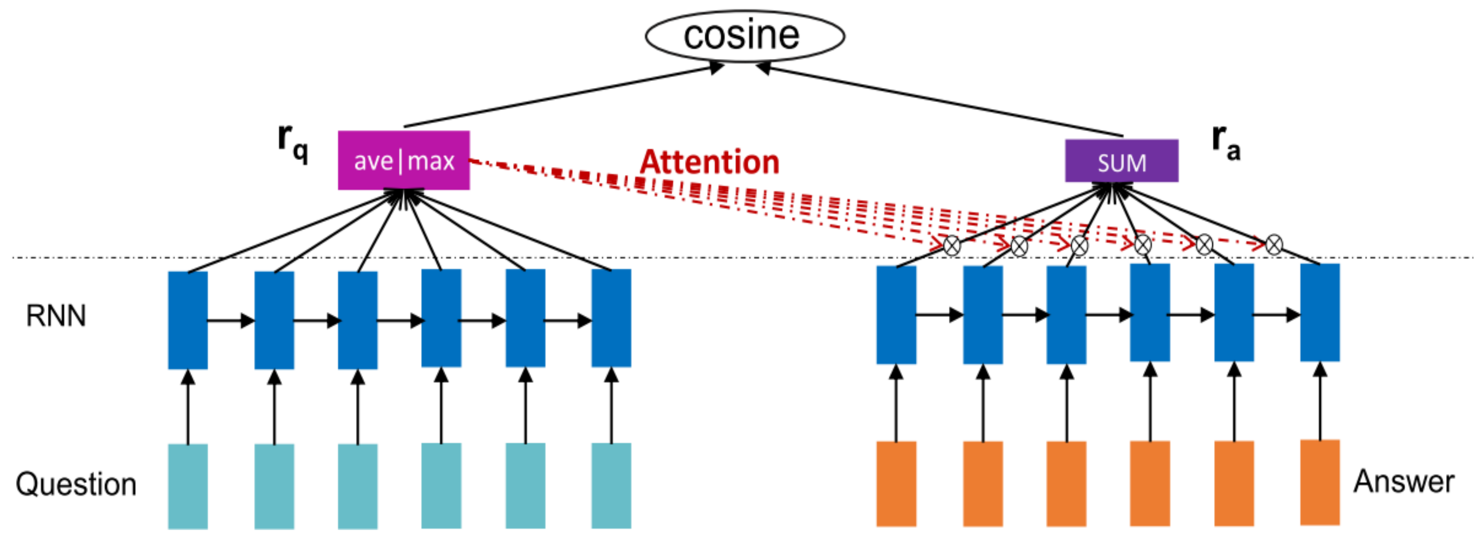
Inner Attention based Recurrent Neural Networks for Answer Selection

Bingning Wang, Kang Liu, Jun Zhao

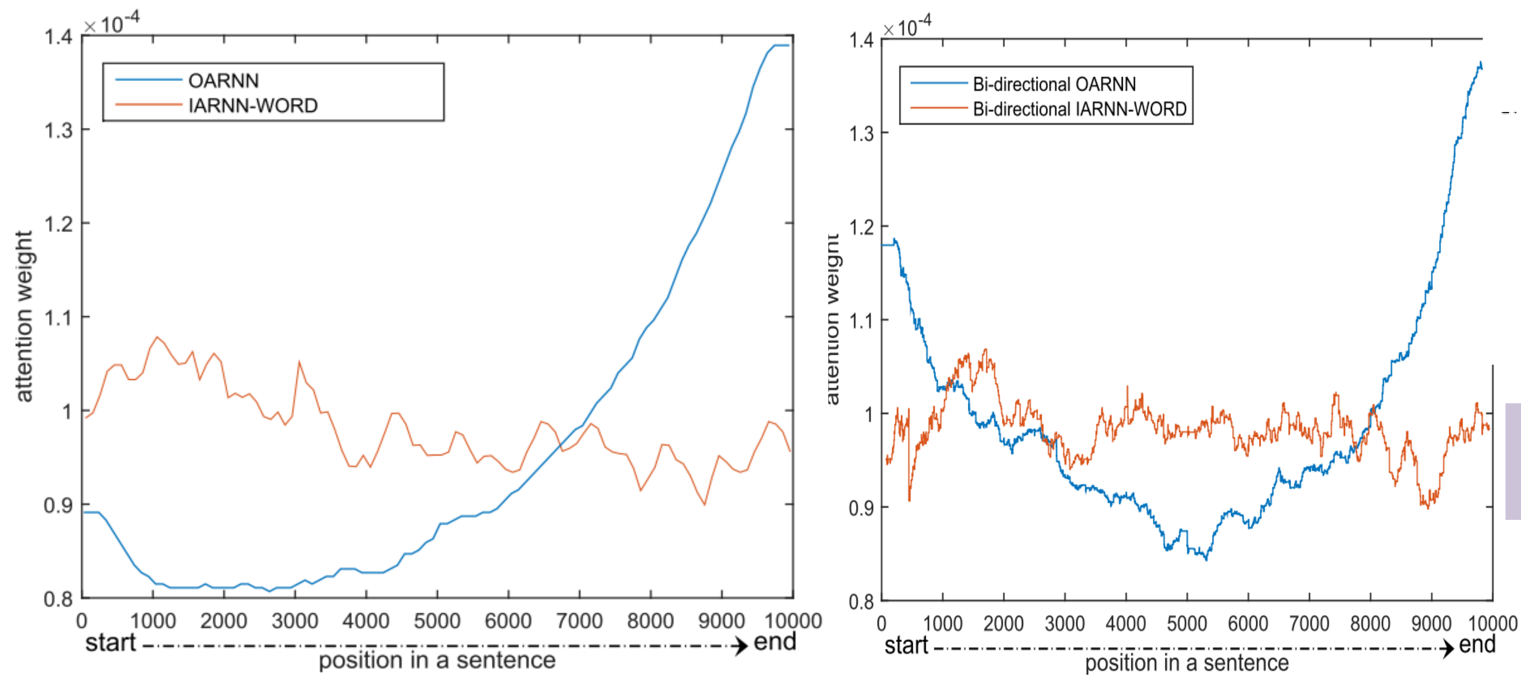
National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences

Background

In traditional attention based RNN models,the attention is added to the hidden states, but in RNN the hidden states near the end of the sentence are expected to capture more information, so it is bound to get more information from the resource.



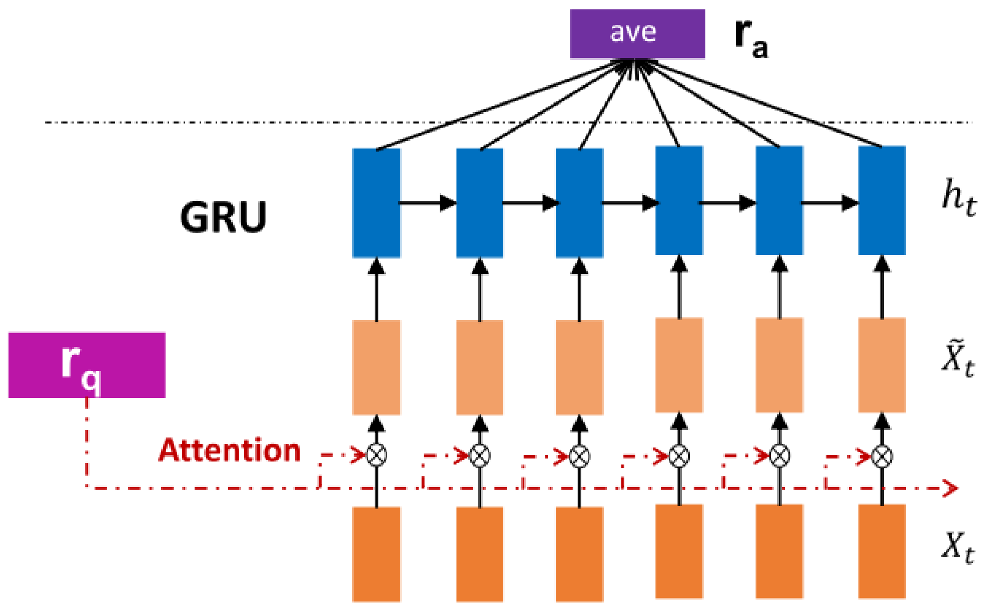
The attention may biased toward the later coming words in a sentence, which is illustrated in the following picture.



Methods

In order to solve the attention bias problem, we proposed three inner attention based RNN models that add attention before recurrent representation.

Model1:IARNN-WORD

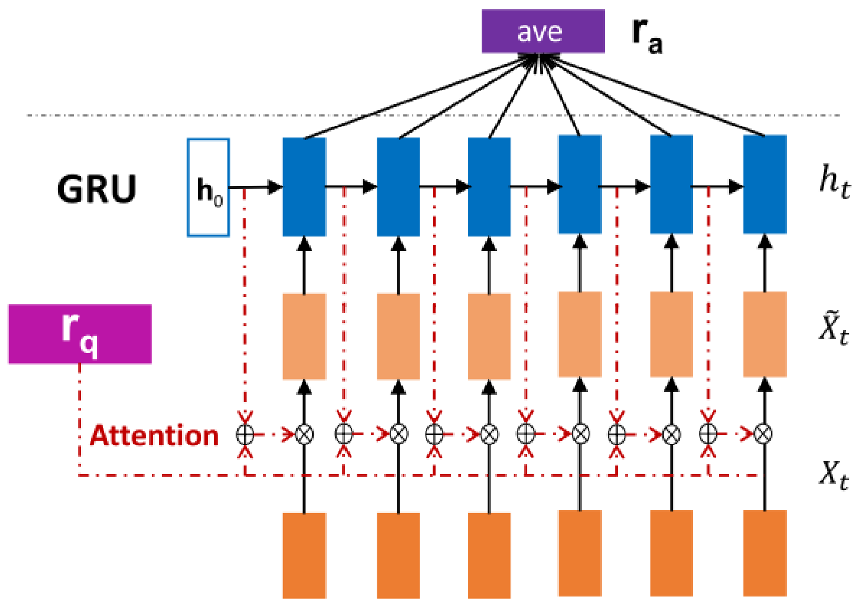


$$\alpha_t = \sigma(\mathbf{r}_q^T \mathbf{M}_{qi} \mathbf{x}_t)$$

$$\tilde{\mathbf{x}}_t = \alpha_t * \mathbf{x}_t$$

Instead of adding attention information to the hidden layers of RNN (GRU), we directly add this information to the original word embedding.

Model2:IARNN-CONTEXT



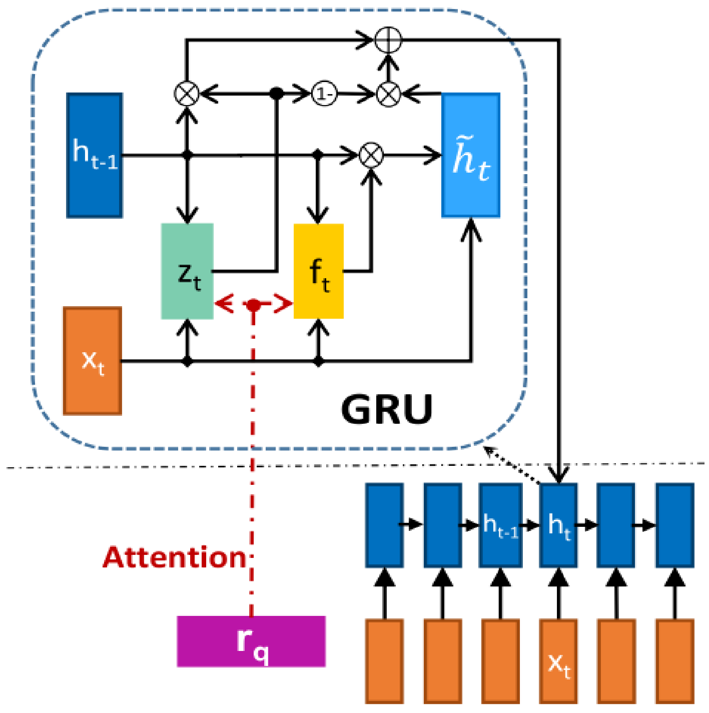
The IARNN-WORD did not take the context information into account, but the context for a word is important for determining its meaning and thus attention weights

$$\mathbf{w}_C(t) = \mathbf{M}_{hc}\mathbf{h}_{t-1} + \mathbf{M}_{qc}\mathbf{r}_q$$

$$\alpha_C^t = \sigma(\mathbf{w}_C^T(t)\mathbf{x}_t)$$

$$\tilde{\mathbf{x}}_t = \alpha_C^t * \mathbf{x}_t$$

Model3:IARNN-GATE



Directly embed the attention information into the recurrent activation unit, which take the attention information into recurrent process in a more generalized way.

$$\mathbf{z}_t = \sigma(\mathbf{W}_{xz}\mathbf{x}_t + \mathbf{W}_{hz}\mathbf{h}_{t-1} + \mathbf{M}_{qz}\mathbf{r}_q)$$

$$\mathbf{f}_t = \sigma(\mathbf{W}_{xf}\mathbf{x}_t + \mathbf{W}_{hf}\mathbf{h}_{t-1} + \mathbf{M}_{qf}\mathbf{r}_q)$$

$$\tilde{\mathbf{h}}_t = tanh(\mathbf{W}_{xh}\mathbf{x}_t + \mathbf{W}_{hh}(\mathbf{f}_t \odot \mathbf{h}_{t-1}))$$

$$\mathbf{h}_t = (1 - \mathbf{z}_t) \odot \mathbf{h}_{t-1} + \mathbf{z}_t \odot \tilde{\mathbf{h}}_t$$

IARNN-OCCAM

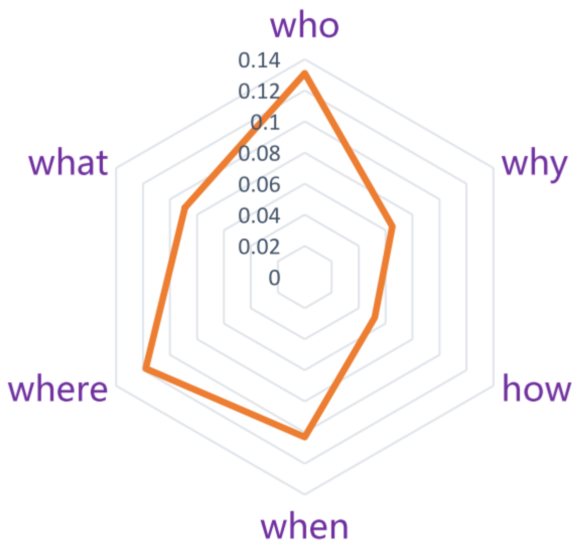
Occam's Razor: Among the whole words set, we choose those with fewest number that can represent the sentence.

$$n_p^i = \max\{\mathbf{w}_{qp}^T \mathbf{r}_q^i, \lambda_q\}$$

$$J_i^* = J_i + n_p^i \sum_{t=1}^{m} \alpha_t^i$$

for the speciPc question representation **r**, we use a vector wqp to project it into scalar value **n** and then we add it into the original objective J

Experiment



Occam regulation

System	MAP	MRR
(Wang and Nyberg, 2015) †	0.7134	0.7913
(Wang and Ittycheriah, 2015) †	0.7460	0.8200
(Santos et al., 2016) †	**0.7530**	**0.8511**
GRU	0.6487	0.6991
OARNN	0.6887	0.7491
IARNN-word	0.7098	0.7757
IARNN-Occam(word)	0.7162	0.7916
IARNN-context	0.7232	0.8069
IARNN-Occam(context)	0.7272	0.8191
IARNN-Gate	<u>0.7369</u>	<u>0.8208</u>

Trec-QA

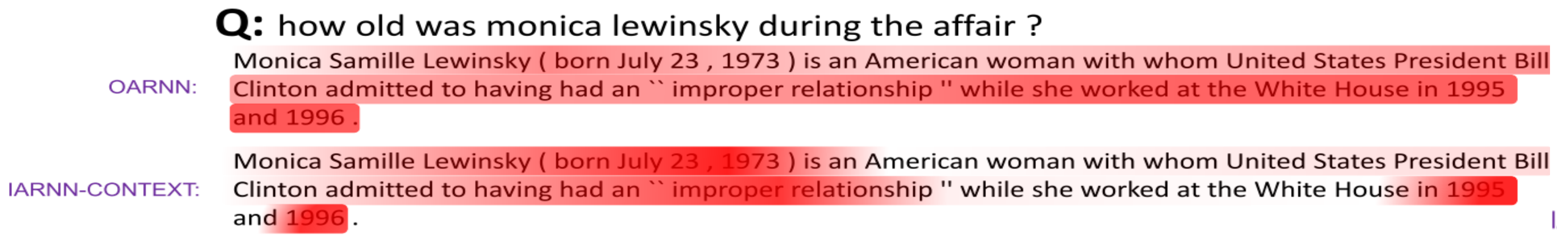
System	Dev	Test1	Test2
(Feng et al., 2015)	65.4	65.3	61.0
(Santos et al., 2016)	66.8	67.8	60.3
GRU	59.4	53.2	58.1
OARNN	65.4	66.1	60.2
IARNN-word	67.2125	67.0651	61.5896
IARNN-Occam(word)	69.9130	69.5923	63.7317
IARNN-context	67.1025	66.7211	63.0656
IARNN-Occam(context)	69.1125	68.8651	**65.1396**
IARNN-Gate	**69.9812**	**70.1128**	62.7965

Insurance-QA

System	MAP	MRR
(Yang et al., 2015)	0.652	0.6652
(Yin et al., 2015)	0.6921	0.7108
(Santos et al., 2016)	0.6886	0.6957
GRU	0.6581	0.6691
OARNN	0.6881	0.701
IARNN-word	0.7098	0.7234
IARNN-Occam(word)	0.7121	0.7318
IARNN-context	0.7182	0.7339
IARNN-Occam(context)	**0.7341**	**0.7418**
IARNN-Gate	0.7258	0.7394

Wiki-QA

Visualization

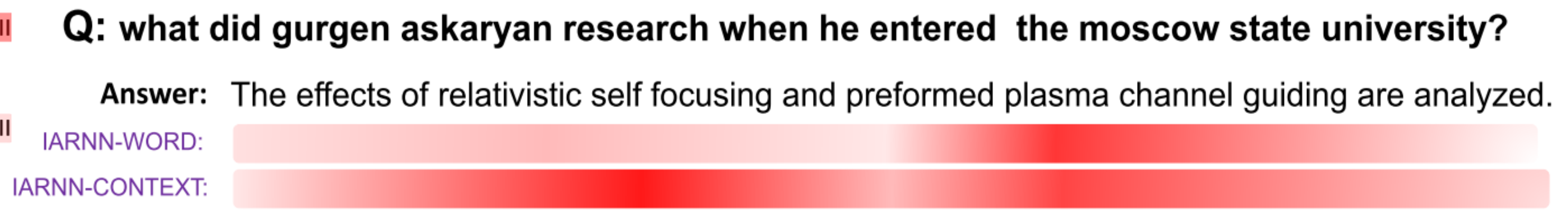


Q: how old was monica lewinsky during the affair ?

OARNN: Monica Samille Lewinsky (born July 23 , 1973) is an American woman with whom United States President Bill Clinton admitted to having had an `` improper relationship '' while she worked at the White House in 1995 and 1996 .

IARNN-CONTEXT: Monica Samille Lewinsky (born July 23 , 1973) is an American woman with whom United States President Bill Clinton admitted to having had an `` improper relationship '' while she worked at the White House in 1995 and 1996 .

An example demonstrates the advantage of IARNN in capturing the informed part of a sentence compared with OARNN.



Q: what did gurgen askaryan research when he entered the moscow state university?

Answer: The effects of relativistic self focusing and preformed plasma channel guiding are analyzed.

An example illustrates the IARNN-CONTEXT could attend the consecutive words in a sentence.