Appendix for Paper 4498

1 Computational Cost

Since it is difficult to analyze the time complexity of deep learning models theoretically, we investigate the computational cost of baseline methods and the proposed KHAMA model on the SemEval-15 dataset, instead of analyzing the time complexity. All the models run on a single NVIDIA Tesla P100 GPU. At the training phase, KHAMA takes about 75 seconds per training epoch. Based on our empirical observation, most compared baselines take about 10-60 seconds per training epoch on an average . All models typically converge within less than 20 epochs by using the early stopping criterion. At the testing phase, our model takes about 1.32 milliseconds to predict the label for each instance, while the baselines take about 0.14-0.94 milliseconds on an average. The training time per epoch (seconds) and testing time (milliseconds) per instance on the SemEval-15 dataset are given in Table A.1.

Method	Training time per epoch(s)	Testing time per instance (ms)
JAIST	11.5	0.142
KeLP	17.2	0.266
BGMN	15.4	0.231
CNN	13.3	0.189
LSTM	22.6	0.369
Bi-LSTM-CRF	57.6	0.943
AP-LSTM	28.5	0.414
AI-CNN	47.4	0.817
KHAMA	75.6	1.323

Table A.1: Training time per epoch (seconds) and test time per instance (milliseconds) on SemEval-15 dataset.

2 About Multi-head Co-Attention Network

2.1 Multi-head attention

Multi-head attention produces 2-D attention weight matrix of size $b \times n$, which benefits capturing comprehensive information of the whole text from different perspectives. Specifically, multi-head attention projects the question Q^q and answer $\mu(Q^a)$ b times with different linear projections (see Eq.(13)). For each of these b projected versions of question and answer representations, we perform the single-head attention function in parallel, yielding n-dimensional attention vectors. These b vectors are concatenated to form the final attention matrix. More details of multi-head attention can be found in (Vaswani et al., 2017).

2.2 The Effect of Attention Head b

b is the number of heads in multi-head co-attention network. We investigate the effect of b by varying its value from 1 to 10 with step size 1. Note that when $b = 1$, attention matrix Σ reduces to a normal vector form, namely single-head attention. We report the experimental results on SemEval-2015 dataset in Figure 1. We can achieve best results when $b = 4$. As b increases from 1 to 10, the accuracy and F1 scores increase slightly till an optimal value (when $b = 4$), after which it decreases sharply.

3 Experimental Results with Mean and Variance

We summarize the **best results** in Tables A.2-A.3 for the two datasets. In addition, to evaluate the stability of our model, we run KHAMA ten times, and report the **mean** and the **standard deviation** in Tables A.2-A.3. Our model achieves statistically significantly better performance than the state-of-the-art competitors on the two datasets (t-test, p-value $<$ 0.05). For example, for the accuracy the proposed KHAMA method substantially and consistently outperforms other methods. As we know, it is difficult to boost 1 percent of accuracy for CQA.

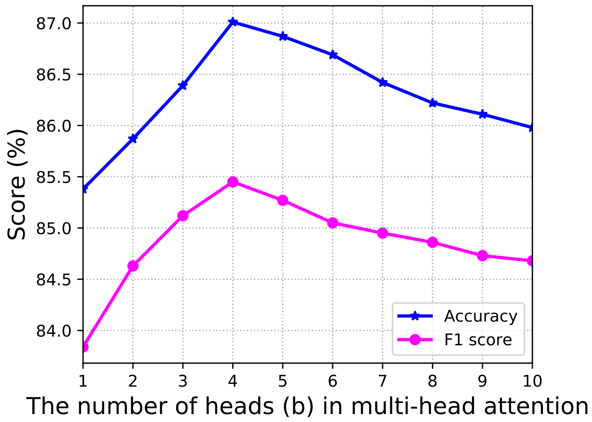


Figure 1: Experimental results of KHAMA on SemEval-2015 by varying the the numbers of heads (i.e., b) in multi-head co-attention network.

Method	Accuracy	F1 score
JAIST	79.10	78.96
KeLP	81.96	80.73
BGMN	81.24	80.22
CNN	77.33	76.92
LSTM	76.21	75.15
Bi-LSTM-attention	81.12	79.09
CNN-LSTM-CRF	82.15	81.33
AP-LSTM	79.45	79.06
AI-CNN	83.06	81.92
KHAMA (Ours)	**87.02** * (85.28 ±1.13)	**85.47*** (84.54±0.75)

Table A.2: Quantitative evaluation results on SemEval-2015. Numbers with * mean that improvement from our model is statistically significant over the baseline methods (t-test, p-value < 0.05).

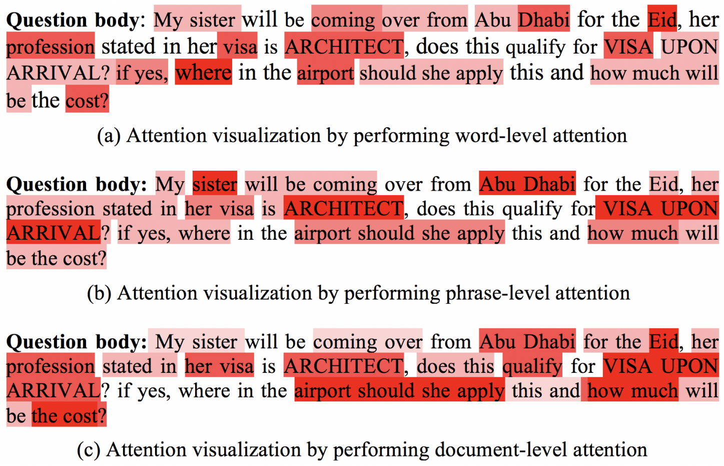
Method	Accuracy	F1 score	MAP
JAIST	73.78	68.04	87.24
Kelp	73.89	69.87	88.43
BGMN	74.75	75.39	87.68
CNN	73.22	72.14	86.21
LSTM	74.05	73.45	86.28
Bi-LSTM-attention	76.60	74.82	88.05
BGMN	74.75	75.39	87.68
CNN-LSTM-CRF	77.18	77.04	87.66
AP-LSTM	77.64	76.82	87.82
AI-CNN	78.24	77.75	88.33
KHAMA (Ours)	**82.34*** (81.41±0.70)	**81.13*** (80.25±0.94)	**90.76*** (89.21±0.92)

Table A.3: Quantitative evaluation results on SemEval-2017.

4 Intuition for the Proposed Architecture

4.1 Intuition for Using Knowledge-enhanced Hierarchical Attention Network

Hierarchical attention helps CQA model capture important information at different levels of granularity and acquire comprehensive information to identity suitable answers. As illustrated in Figure 2, we visualize the hierarchical attention weights for a question chosen from SemEval-15 dataset to analyze how the hierarchical attention mechanisms make effect. The color depth indicates the importance degree of the attention. The darker the color, the more important the word. First, we design a word-level mutual attention mechanism to establish the word-level correlations between the question and the commonsense knowledege from KB, and identify the important words such as "VISA" and "ARCHITECT". Second, we adopt the n-gram convolution operation to capture different local semantic units. For example, the n-gram patterns such as "Abu Dhabi", "UPON ARRIVAL" can be identified after performing phrase-level attention. Third, we use the document-level attention to select crucial context word chunks, such as "how much", to compose the knowledge-aware question representation.



Question body: My sister will be coming over from Abu Dhabi for the Eid, her profession stated in her visa is ARCHITECT, does this qualify for VISA UPON ARRIVAL? if yes, where in the airport should she apply this and how much will be the cost?

(a) Attention visualization by performing word-level attention

Question body: My sister will be coming over from Abu Dhabi for the Eid, her profession stated in her visa is ARCHITECT, does this qualify for VISA UPON ARRIVAL? if yes, where in the airport should she apply this and how much will be the cost?

(b) Attention visualization by performing phrase-level attention

Question body: My sister will be coming over from Abu Dhabi for the Eid, her profession stated in her visa is ARCHITECT, does this qualify for VISA UPON ARRIVAL? if yes, where in the airport should she apply this and how much will be the cost?

(c) Attention visualization by performing document-level attention

Figure 2: The knowledge-enhanced hierarchical attention weights of a question chosen from SemEval-2015.

4.2 The Intuition for Using External Knowledge

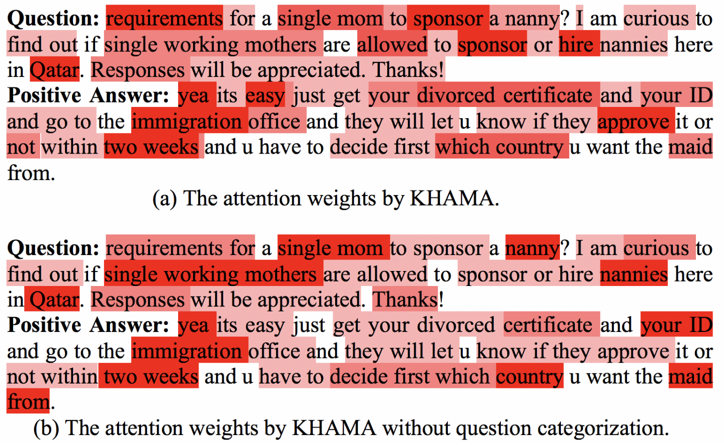
The large-scale real-world knowledge contained in the knowledge base plays a crucial role in answering natural language questions, but it is underutilized. Table A.4 lists an example question and its positive and negative answers. In the absence of the real-world background knowledge, negative answer may be scored higher than its positive counterpart, since the negative answer is more similar to the given question at the word level. Conversely, with the background knowledge, we can correctly identify the positive answer through the relevant facts contained in the KB such as (labor, related_to, work) and (permit, related_to, ticket of leave). However, even though recent advances in constructing large-scale knowledge bases have enabled QA systems to return an answer from a KB, the exploration of external knowledge from KBs is still a relatively new territory and under-explored.

Question	"How to go home without exit permit. My sponsor is not willing to give me exit permit in the fear of my not return. I am about 3 years at Qatar; and I want to go home on a vacation before Eid. Is there any way to go home without the exit permit from sponsor? I want to go home before Eid; no matter what does it cost; even cancellation of my visa. Thank you for your advice."
Positive answer	"You have 7 days... You left it a bit late to resign; no? If you are due to go home they have to give you a ticket. If not; then go to the Labour Department and report him."
Negative answer	"My sponsor is not willing to let me go in any way until all of my projects are finished. Can my embassy or the labor department help?"

Table A.4: Example of QA candidate pairs from SemEval-2017.

4.3 The Intuition for Using Category Classification

We list a QA pair from SemEval-2015 (in *Visas and Permits category*) that is correctly predicted by KHAMA but incorrectly predicted by KHAMA without the question categorization task (denoted as KHAMA w/o category), and visualize the attention scores of both models in Figure 3. We observe that KHAMA can assign higher scores to the important information in the question and answer such as "sponsor", "approve", "immigration" guiding by the category (i.e., *Visas and Permits*) of the input question. However, KHAMA w/o Category cannot recognize some important information about *Visas and Permits* (e.g., "sponsor", "approve") and gives wrong prediction since there are only a few overlapped information in word level between the question and the answer. Category classification can improve the quality of locating the salient information of a long question and enhance the question representation learning.



Question: requirements for a single mom to sponsor a nanny? I am curious to find out if single working mothers are allowed to sponsor or hire nannies here in Qatar. Responses will be appreciated. Thanks!

Positive Answer: yea its easy just get your divorced certificate and your ID and go to the immigration office and they will let u know if they approve it or not within two weeks and u have to decide first which country u want the maid from.

(a) The attention weights by KHAMA.

Question: requirements for a single mom to sponsor a nanny? I am curious to find out if single working mothers are allowed to sponsor or hire nannies here in Qatar. Responses will be appreciated. Thanks!

Positive Answer: yea its easy just get your divorced certificate and your ID and go to the immigration office and they will let u know if they approve it or not within two weeks and u have to decide first which country u want the maid from.

(b) The attention weights by KHAMA without question categorization.

Figure 3: The attention weights of an QA pair chosen from SemEval-2015, whose category label is *Visas and Permits*.

References

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*. pages 5998–6008.