A Context Paragraph Extraction

We gather a diverse collection of everyday situations from a corpus of personal narratives (Gordon and Swanson, 2009) from the ICWSM 2009 Spinn3r Blog Dataset (Burton et al., 2009). It contains over 1.6 million of non-spam weblog entries describing everyday personal events. For each personal story, we use spaCy[4] for sentence segmentation and tokenization. In order to get a short substory as context, we apply pre-trained BERT[5] (Devlin et al., 2018) model to predict a confidence score for each two consecutive sentences from the story, and then segment each story into multiple paragraphs so that each paragraph contains between 30 and 150 words. For each blog, we randomly sample one paragraph as context to create questions and answers.

B Additional Details on AMT Instructions

To make the questions more challenging for an AI system, we recommend the workers use less words from the paragraph for correct answers and make incorrect answers more appealing by using words from the paragraph as much as possible. We also encourage workers to provide all the candidate answers with similar length and style.

We restrict this task to the workers in English-speaking countries (United States, Canada, and United Kingdom) and with more than 5,000 HITs with at least a 99% acceptance rate. To ensure quality, we also create a qualification task.

C Implementation Details

For baseline methods, we use their released implementations from open source projects and retrain them on our dataset. All approaches follow the same pre-processing steps: segment each paragraph into multiple sentences, and tokenize each sentence as well as question and candidate answers with spaCy. For BERT-FT based approaches, we optimize the parameters with grid search: training epochs 10, learning rate $l \in \{$2e-5, 3e-5, 5e-5$\}$, gradient accumulation steps $g \in \{1, 4, 8\}$, training batch size $b \in \{2g, 3g, 4g, 5g\}$. We will make all the resources and implementations publicly available.

[4] https://spacy.io/

[5] Through the whole paper, BERT refers to the pre-trained BERT large uncased model from https://github.com/huggingface/pytorch-pretrained-BERT

D Impact of Training Data Size

To explore the impact of the amount of training data, we divide the whole training dataset into 10-fold and successively add another 10% into the training data. We use BERT-FT approach for comparison. Figure 8 shows the learning curve. We can see that, the performance goes up as we add more training data. However, we do not observe significant improvement when we further increase the training data after 15K questions.

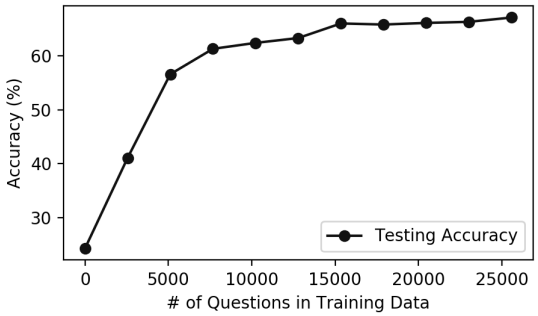

Figure 8: Performance on COSMOS with various amount of training data

E Details for Generative Evaluation

For generative evaluation, we base on the OpenAI pre-trained GPT2 transformer language model,[6] which has 117M parameters, and fine-tune it with all $\lfloor$*Paragraph*, *Question*, *Correct Answer*$\rceil$ in COSMOS QA training set with top-k sampling, where $k \in \{3, 10, 50, 100, 1000\}$. After fine-tuning, we use GPT2-FT to generate 10 candidate answers conditioned on each $\lfloor$*Paragraph*, *Question*$\rceil$ from development and test sets. Note that for all training, development and test sets, we omit the questions to which the correct answer in COSMOS QA is "None of the above".

For automatic evaluation, we compare each generated candidate answer against the original human authored correct choice in COSMOS QA, and average all metric scores with 10 sets of candidate answers. For AMT based human evaluation, we randomly sample 200 paragraphs and questions, and for each question we randomly sample 4 automatically generated answers from the outputs of GPT2 without fine-tuning and GPT2-FT. For each question set, we ask 3 workers to select all plausible correct answers from the 4 candidate choices or "None of the above". Workers are paid

[6] https://github.com/huggingface/pytorch-pretrained-BERT

$0.1 per question set. For each question, we consider an automatically generated answer as correct only if all 3 workers determined it as correct.