Transfer Learning to Generate Multiple Sentence Question with Leveraging Difference between Datasets

Kimihiro Hasegawa[†] **Takaaki Matsumoto**[††] **Ryoichi Takashima**[†]
Tetsuya Takiguchi[†] **Yasuo Ariki**[†] **Teruko Mitamura**[††]
[†]Kobe University
[††]Carnegie Mellon University
kimihiroh@stu.kobe-u.ac.jp, {tmatsumo, teruko}@andrew.cmu.edu
rtakashima@port.kobe-u.ac.jp, {takigu, ariki}@kobe-u.ac.jp

1 Introduction

Recently, question generation (QG) is getting more attention as the dual task of question answering (QA), improving the performance of QA models. The common setting of the text-based QG task is to output a question by taking a context and an answer as input. In this work, we focus on a semantic aspect of generated questions, specifically reasoning, aiming to generate multi-sentence questions (MSQs): questions that require reasoning over multiple sentences (Khashabi et al., 2018), which is more difficult to answer than single-sentence questions (SSQs).

In Figure 1, we show the example of MSQ and SSQ. When one answers to the MSQ, one needs to look at the first and second sentences to figure out "Imelda Staunton" is "the English stage and screen actress born in 1956", who was starred in "The Awakening", a 2011 British horror film." Then, one can answer the question by finding the director of "The Awakening." Contrary to the MSQ which contains several reasoning steps, the SSQ in Figure 1 can be answered by finding the name, looking at only the first sentence.

Due to the current success of neural sequence-to-sequence approaches, the performances of QG models have been significantly improved on traditional metrics, such as BLEU or ROUGE (Du and Cardie, 2018; Dong et al., 2019). However, considering an application of QG to measure the reading comprehension ability of humans, semantic quality of generated questions is an important issue. One of the semantic-quality problems which existing models fall into is a semantic drift problem, i.e., the semantics of the model-generated question drifts away from the given context and answer. (Zhang and Bansal, 2019)

Furthermore, in order to work on reasoning quality of QG, we do not have a parallel dataset that consists of both MSQ and SSQ for the same context and answer. Therefore, in this paper, we propose a transfer-learning QG approach and a method to create a parallel dataset to address the semantic quality, especially reasoning. Then, on the created parallel dataset, we fine-tune a transformer-based pre-trained language model and train a binary classifier to identify whether an input question belongs to MSQs or not, simultaneously. Filtering questions generated by the fine-tuned model based on the classifier's score, we can finally obtain MSQs with higher probability.

context: (1)The Awakening is a 2011 British horror film directed and co-written by Nick Murphy, starring Rebecca Hall, Dominic West, Isaac Hempstead-Wright and Imelda Staunton. (2)Imelda Mary Philomena Bernadette Staunton, CBE (born 9 January 1956) is an English stage and screen actress. (3)After training at the Royal Academy of Dramatic Art, ...
answer: Nick Murphy

MSQ: Who directed the 2011 British horror film starring the English stage and screen actress born in 1956?
SSQ: Who wrote the script for 'Awakening'?

Figure 1: An example of MSQ and SSQ.

According to our evaluation, the trained classifier successfully filters MSQs among generated questions, resulting in a higher ratio of MSQs than that in the generated questions before filtering. Moreover, the selected questions demonstrate higher scores on the automatic evaluation metrics.

We describe the architecture of our model and training steps more specifically in Section 3. Then, we explain the detailed settings of our experiment in Section 4, followed by the experiment's result and discussion in Section 5.

2 Related Works

Du and Cardie (2018) proposed the attention-based sequence-to-sequence neural model with leveraging information from the entire paragraph, which has become a baseline model on the neural QG task. Zhang and Bansal (2019) addresses the semantic drift in QG by introducing semantic-enhanced rewards. Dong et al. (2019) has achieved the current state-of-the-art performance by a large-scale language model pre-training strategy, which is

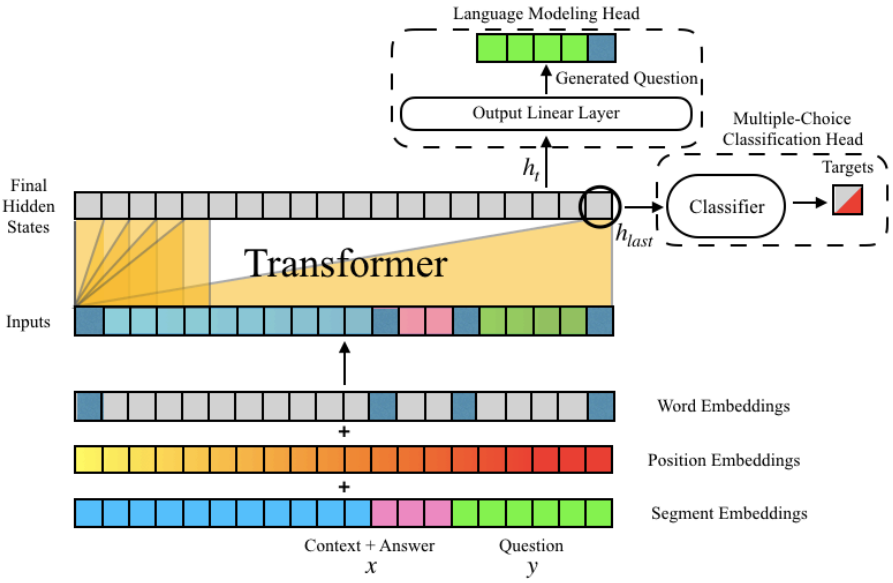

Figure 2: The LM+BC architecture and input embeddings.

similar to our work in terms of the model architecture and fine-tuning method. Despite the improvement of the performance w.r.t automatic machine evaluation, it remains a problem that questions generated by these models do not have sufficient semantic quality. In this work, we focus on the QG task, especially to improve reasoning quality of generated questions.

3 Transfer Learning Approach

To generate MSQs, in this work, we use the Generative Pre-trained Transformer, GPT-2, by Radford et al. (2019) that is a multi-layer Transformer decoder. GPT-2 based language model (LM) takes as input the concatenation of word embeddings, position embeddings, and segment embeddings. Word embeddings are constructed from context, answer, and question in this order with delimiters between each group of tokens, which are tokenized by byte pair encoding. Position embeddings and segment embeddings are added to indicate the position of each token and the group where each token belongs to. We follow the work of Radford et al. (2019) for pre-training GPT-2. Taking the embeddings as input, the LM is fine-tuned by teacher forcing to optimize the language modeling loss L_{lm}, where x indicates the embeddings of context and answer. The final hidden state of the self-attention model h_t, is fed into an output linear layer f_{out}, where the activation is a softmax function over the vocabulary to obtained next token probabilities, $p(y_t|...)$. Taking the ground truth y_t, as labels, a negative log-likelihood loss is computed from the probabilities as follows:

$$p(y_t|y_1, ..., y_{t-1}, x) = \text{softmax}(f_{out}(h_t))$$

$$L_{lm} = -\sum_{t=1}^{n} \log p(y_t|y_1, ..., y_{t-1}, x)$$

According to our pilot experiment[1], LM fine-tuned on an MSQ dataset (MSQ-LM) does not generate MSQs in the same ratio as the original dataset. More specifically, while ground truth questions contain MSQs for 93%, MSQ-LM generates MSQs for 50%, SSQs for 10%, and invalid questions for 40%. In short, the fine-tuned model is more likely to generate SSQs or invalid questions that are ungrammatical or need more information to answer. In order to solve this problem, we also use a classifier aiming to identify whether the input question belongs to MSQs or not. This binary classifier (BC) is fine-tuned on the next question classification loss L_{bc}. The last token's final hidden state h_{last}, is fed into a linear layer f_{bc}, where the activation is a sigmoid function σ, to obtain the probability of MSQ p_{msq}. Then, taking the gold question as a label, a negative log-likelihood loss is computed as follows:

$$p_{msq}(y_1, ..., y_n) = \sigma(f_{bc}(h_{last}))$$

$$L_{bc} = -y * \log p_{msq} - (1 - y) * \log(1 - p_{msq})$$

Namely, Our model, GPT-2-based language model with a classifier (LM+BC) shown in Figure 2, is fine-tuned to optimize a combination of these two losses. We follow the implementation of a dialogue-model by Wolf et al. (2019).

[1]The first author evaluated 30 randomly selected questions for each.

In order to train the classifier with a parallel dataset consisting of pairs of an MSQ and an SSQ for the same context and answer, we created a parallel dataset by leveraging a fine-tuned model on another dataset. Hence, the classifier learns the only difference between MSQs and SSQs. In our proposed approach, our model is fine-tuned by the following steps:

Step 1: Fine-tuning language model on SSQs We fine-tuned the LM on an SSQ dataset so that it generates questions with the characteristics of SSQs. We call this fine-tuned language model SSQ-LM.

Step 2: Parallel dataset creation Taking a pair of context and answer in an MSQ dataset as input, the SSQ-LM generates an SSQ for the corresponding pair so that each pair of context and answer has both an MSQ and an SSQ.

Step 3: Fine-tuning language model and classifier on MSQs and SSQs The LM+BC is fine-tuned on the created dataset, specifically, the language modeling loss is calculated only from MSQs, and the classifier is trained on both MSQs and SSQs. We call this fine-tuned language model with the trained classifier MSQ-LM+BC.

4 Experimental Setup

4.1 Datasets

SQuAD dataset mostly consists of SSQs based on Wikipedia articles (Rajpurkar et al., 2016). Since the test data is not released publicly for its integrity, we randomly split development set into a development set and test set, resulting in 87503 pairs for the training set, 8492 pairs for the development, and 9253 pairs for the test set.

As an MSQ dataset, we adopt the HotpotQA dataset (Yang et al., 2018), whose questions are created to require reasoning over multiple supporting documents to answer. Among the three levels of questions in the dataset, easy, medium, and hard, we use only medium and hard questions, because easy-level questions are likely to be SSQs from its collecting process. To remove noises, we filtered out questions whose length is over 30 words, keeping 92 % of the original dataset. Then, we split the filtered training data into training, development, and test set, resulting in 57000 pairs for the training set, 7125 pairs for the development set, and 7125 pairs for the test set.

4.2 Settings

In our experiment, we use pre-trained GPT-2 small (12-layer decoder-only transformer with 768 dimensional states and 12 attention heads, which is the smallest model out of four with 117M parameters) open-sourced by HuggingFace[2]. For fine-tuning, we use a single GPU, RTX 2070, limiting the number of tokens to 600 maximum, by teacher forcing algorithm. When decoding, we employ top-p (nucleus) filtering: sampling from next token distributions after filtered this distribution to keep only the highest probability tokens whoso cumulative probability exceeds the threshold, 0.9.

Table 1: Automatic evaluation results of SSQ generation on the SQuAD dataset.

	BLEU-4	METEOR	ROUGE-L
SSQ-LM	14.39	19.62	39.30
SemQG (2019)	18.37	22.65	46.68
UNILM (2019)	22.88	24.94	51.80

Table 2: Automatic evaluation results of MSQ generation on the HotpotQA dataset.

	BLEU-4	METEOR	ROUGE-L
MSQ-LM	20.77	23.14	40.76
MSQ-LM+BC	20.97	23.17	40.86
+ scores >7.0	**25.40**	**26.54**	**45.75**

5 Results

5.1 Quantitative Analysis

First, we evaluated the generation ability of our proposed model on the test set of the SSQ dataset. Table 1 shows the automatic evaluation results that indicate the similarity between the generated questions and the reference questions. Although other models perform better than our model on the metrics, we choose to use GPT-2 based model because of the following two reasons. (1) we put the main focus on the semantic quality of generated questions, especially reasoning. (2) our model is easy and reasonable to fine-tune with a relatively small amount of computation cost, compared to UNILM, which requires 2-4 32GB-GPUs. Next, from Table 2, we can see that the performances of MSQ-LM and MSQ-LM+BC are similar, which means the classifier added to our model does not affect significantly on the performance of the language model. We also illustrate the probability computed by the classifier, in Figure 3. According to the distribution, the classifier seems to have captured the feature difference between MSQs and SSQs by training. To investigate more, we looked at the score by the classifier before a sigmoid function, which is shown in Figure 4. Contrary to Figure 3, the scores in Figure 4 have variety, follow-

[2]https://github.com/huggingface/transformers

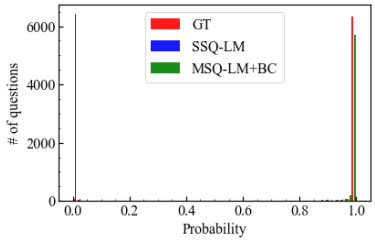

Figure 3: The probability histogram of the classifier.

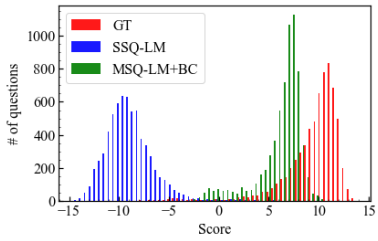

Figure 4: The output score histogram of the classifier.

ing a normal distribution. Based on the assumption that the scores represent the degree of MSQ (the higher the score is, the more probable the question is MSQ), we extracted questions whose score is more than 7.0 (around 30 % of test set) and conducted the same automatic evaluation, which is shown in "+ scores > 7.0" in tables. The results indicate the filtering by the classification score improves the scores on the evaluation metrics significantly.

5.2 Qualitative Analysis

We further sample 100 examples from the test set and conduct a human evaluation of checking whether the questions are MSQ or not. As for the sampling, "+scores>7.0" are sampled from the filtered examples, which accounts for about 30% in the entire test set, while others are sampled from the entire test set. Then, we take the average of the evaluation results done by three non-native subjects, which is shown in Table 3. From the table, we can observe the increase in the ratio of MSQ after the filtering. Although this does not prove the classifier's ability to understand the reasoning in text, it can be said that, to some extent, the classifier captures the characteristics of MSQs from the differences between two question types.

6 Conclusion

In this paper, we propose a transfer learning approach to generate multiple sentence questions, MSQs, with a simple classifier to capture the features of MSQs. In order to fine-tune our language model and train the classifier, we leverage two different Wikipedia-based QA datasets.

Table 3: Human evaluation results of MSQ generation on the HotpotQA dataset."GT" represents ground truth.

	MSQ	Not MSQ
GT	80.0%	20.0%
SSQ-LM	33.3%	66.7%
MSQ-LM+BC	62.7%	37.3%
+ scores >7.0	**78.7%**	21.3%

According to our machine and human evaluation, our approach can generate MSQs with higher probability than only the language model, after filtering based on the classification score.

References

Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. Unified language model pre-training for natural language understanding and generation. In *33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*, 2019.

Xinya Du and Claire Cardie. Harvesting paragraph-level question-answer pairs from wikipedia. In *Association for Computational Linguistics (ACL)*, 2018.

Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. Looking beyond the surface:a challenge set for reading comprehension over multiple sentences. In *Proceedings of North American Chapter of the Association for Computational Linguistics (NAACL)*, 2018.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016.

Thomas Wolf, Victor Sanh, Julien Chaumond, and Clement Delangue. Transfertransfo: A transfer learning approach for neural network based conversational agents. *CoRR*, abs/1901.08149, 2019. URL http://arxiv.org/abs/1901.08149.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2018.

Shiyue Zhang and Mohit Bansal. Addressing semantic drift in question generation for semi-supervised question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019.