Addressing Semantic Drift in Generative Question Answering with Auxiliary Extraction

Chenliang Li, Bin Bi, Ming Yan
Wei Wang, Songfang Huang
Alibaba Group
{lcl193798, b.bi, ym119608}@alibaba-inc.com
{hebian.ww, songfang.hsf}@alibaba-inc.com

Abstract

Recently, question answering (QA) based on machine reading comprehension has become popular. This work focuses on generative QA which aims to generate an abstractive answer to a given question instead of extracting an answer span from a provided passage. Generative QA often suffers from two critical problems: (1) summarizing content irrelevant to a given question, (2) drifting away from a correct answer during generation.

In this paper, we address these problems by a novel Rationale-Enriched Answer Generator (REAG), which incorporates an extractive mechanism into a generative model. Specifically, we add an extraction task on the encoder to obtain the rationale for an answer, which is the most relevant piece of text in an input document to a given question. Based on the extracted rationale and original input, the decoder is expected to generate an answer with high confidence. We jointly train REAG on the MS MARCO QA+NLG task and the experimental results show that REAG improves the quality and semantic accuracy of answers over baseline models.

1 Introduction

Question Answering (QA) has come a long way from answer sentence selection, relationship QA to machine reading comprehension (MRC). Recently, QA has become an essential problem in natural language understanding and a major milestone towards human-level machine intelligence. Current mainstream approaches (Chen et al., 2017; Wang et al., 2018; Yan et al., 2018) treat MRC as a process of extracting a consecutive piece of text from a document to a given question.

Despite the great success in extractive MRC (Wang et al., 2018; Chen et al., 2020), in real-world applications, correct answers may span different passages or even not be literally present in the passages. Directly extracting a consecutive answer span is often inadequate. Therefore, the ability of generating an abstractive answer is needed, which requires a QA model to summarize the main content in a paragraph that is relevant to a given question.

Question	does gameplay programmer need math skill
Passage	A good computer programmer is more of a problem solver and logical thinker than a math buff. Besides, the industry is peppered with many computer programmers who do not really know much about mathematics.
Gold	no, gameplay programmer does not need math skill.
PALM	yes, gameplay programmer is a math buff.
REAG	no, gameplay programmer does not need math skill.

Table 1: An example of the "semantic drift" issue in generative reading comprehension from the MARCO dataset (Nguyen et al., 2016). The text span of words in blue is the rationale extracted by REAG.

Answering questions in natural language can be beneficial to a variety of QA applications, and has led to the development of smart devices such as Siri, Cortana and Alexa. However, compared with answer extraction, answer generation for reading comprehension is more challenging, and has been less explored. A major challenge in generative reading comprehension comes from out-of-control generation of abstractive answers. Although much work has been done in neural language generation (NLG), e.g., KIGN(Li et al., 2018) for summarization, out-of-control generation remains an open question for generative QA which aims to produce correct and coherent answers. Specifically, we observed that generative models often generate answers semantically drifting away from the given

Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Short Papers), pages 942–947

passage and question, known as the "semantic drift" problem. As shown in Table 1, the baseline generative model PALM (Bi et al., 2020) generates an answer that has almost contrary semantics with the gold answer. In general, a generative model often suffers from two critical problems: (1) summarizing content irrelevant to a given question, and (2) drifting away from a correct answer during generation.

In this paper, we address these problems by a novel Rationale-Enriched Answer Generator (REAG), which incorporates an extractive mechanism into a generative model in order to leverage relevant information to a given question in the contextual passage. Specifically, we add an extraction task on the encoder to obtain the rationale for an answer, which is the most relevant piece of text in an input document to the given question. On one hand, the introduction of the supervised extraction task enables the encoder to learn the relevance between a question and a passage; On the other hand, the extracted rationale can be further used to guide the answer generation. Based on the extracted rationale and original input, the decoder is expected to summarize content relevant to a given question and generates an answer with high confidence. Finally, we jointly train REAG on the MS MARCO QA+NLG task based on the common bottom layers. The experimental results show that REAG improves the semantic accuracy of answers over the other state-of-the-art models.

2 Related Work

2.1 Machine Reading Comprehension

In recent years, machine reading comprehension has made great progress with the development of SQuAD (Rajpurkar et al., 2016) and MS MARCO (Nguyen et al., 2016). The current mainstream studies treat machine reading comprehension as answer span extraction from one passage (Rajpurkar et al., 2016, 2018) or multi-passages (Nguyen et al., 2016), which is usually done by predicting the start and end position of an answer. SLQA (Wang et al., 2018) improved answer quality with a hierarchical attention fusion network, which conducted attention and fusion horizontally and vertically across layers between a passage and a question. Recently, the BERT model Devlin et al. (2019) has proved effective for reading comprehension via unsupervised pre-training.

2.2 Generative Reading Comprehension

Bi et al. (2019) proposed a Knowledge-Enriched Answer Generator (KEAG) to compose a natural answer by exploiting and aggregating evidence from all four information sources available: question, passage, vocabulary and knowledge. Nishida et al. (2019a) proposed a multi-style generative model to generate an abstractive summary from the given question, passages and multi-style.

2.3 Reliable Text Generation

Compared with answer extraction, answer generation for reading comprehension is more challenging, and the major challenge in generative reading comprehension lies in out-of-control generation. Recently, some studies have been carried out on increasing the reliability of generation in the encoder-decoder framework(Liu et al., 2018; Li et al., 2018).

3 Rationale-Enriched Answer Generation

3.1 Rationale Span Extraction

In a generative reading comprehension task, every answer has its corresponding rationale, an extractive span in the passage, which can be derived by matching the passage text with the answer. The rationale can usually be located in a certain continuous area of the passage. We use continuous text span as the rationale to minimize the difficulty of the extraction task. Compared with the gold answer, the text span with the highest F1-score in passage is identified as the rationale for training supervision.

Based on the identified rationale, we introduce a rationale extraction task into the encoder. It enables the encoder to learn the relevance between the input question and the passage. Specifically, the encoder predicts whether each token of the passage should be included in the rationale. Every token in the rationale is labeled by 1 and the rest is labeled by 0.

Given input question Q and passage P, we first concatenate them together into an input sequence $X = \{x_1, x_2, ..., x_N\}$. Then we use a shared word embedding layer to project each of the vectors into d-dimensional vectors, and add to each the corresponding position embedding. The resulting vectors are then fed into the Transformer encoder to map the text into a sequence of encoder hidden states $\{h_1, h_2, ..., h_N\}$.

The encoder hidden states can be used to predict whether each token of the passage should be included in the rationale. Therefore, we add a fully connected layer with the sigmoid activation on top of the encoder, to compute the probability for each input word:

$$p_i^r = \text{sigmoid}(w_1 \cdot \text{relu}(W_2 h_i)) \quad (1)$$

where $h_i \in R^d$ is the output hidden state of the encoder for the i^{th} token.

This gives the probability p_i^r that the i^{th} token should be included in the rationale. We then calculate the averaged cross entropy, similar to(Ju et al., 2019), for the rationale extraction loss:

$$\mathcal{L}_{RE_j} = -\frac{1}{N}\sum_{i=1}^{N}(y_{ji}^r \log p_{ji}^r + (1-y_{ji}^r)\log(1-p_{ji}^r)), \quad (2)$$

where N is the number of input tokens. y_i^r is the rationale label for the i^{th} token, and $\mathcal{L}_{RE_j}$ represents the rationale loss for the j^{th} example in the training set.

3.2 Rationale-Enriched Answer Generation

This layer uses a stack of Transformer decoder blocks on top of the embeddings provided by the encoder's word embedding layer. The decoder is similar in structure to the encoder except that it includes a standard attention mechanism after each self-attention layer that attends to the output of the encoder. The rationale-aware hidden states output by the encoder are used for rationale extraction.

In calculating the decoder states s_t, an cross attention is introduced into the decoder to attend to the rationale-aware encoder hidden states. This results in the rationale-aware decoder hidden state s_t:

$$p(y_t|y_1, ..., y_{t-1}) = \text{softmax}(W^e(W^v s_t + b^v) + b^e) \quad (3)$$

During training, we minimize the negative log-likelihood of the answer word at each decoding time step. Let y_t^* denote the target word in the decoding time step t. The overall loss is then defined as:

$$\mathcal{L}_{GEN} = -\frac{1}{T}\sum_{t=1}^{T}\log p(y_t^*|y_1^*, ..., y_{t-1}^*, x, \theta) \quad (4)$$

where T denotes the length of a gold answer.

3.3 Joint Training and Prediction

The rationale extraction task and the answer generation task are designed to share the same embedding and the encoder. Therefore, we propose to train them together as multi-task learning. The joint objective function is formulated as follows:

$$\mathcal{L} = \mathcal{L}_{GEN} + \beta\mathcal{L}_{RE} \quad (5)$$

where β is a hyper-parameter that controls the weight of the rationale extraction task. During the training process, we use a linear decay schedule on the value of β, in order to rely more on the rationale extraction task for addressing the semantic drift problem at the early stage, following by more focus on the target generation task subsequently.

4 Experiments

4.1 Experiment Configuration

Dataset and Evaluation Metric. Given our objective of generating natural answers by document reading, the MARCO dataset [1] (Nguyen et al., 2016) released by Microsoft is a good fit for benchmarking REAG and other answer generation methods. We use the latest MARCO V2.1 dataset and focus on the "QA + Natural Language Generation" task in the evaluation. The data has been split into a training set (150k QA pairs), a dev set (12k QA pairs) and a test set (110k questions). Since true answers are not available in the test set and the task is retired now, we hold out the dev set for evaluation in our experiments, and test models for each question on its associated passages by concatenating them all together. Following Bi et al. (2019), we tune the hyper-parameters by cross-validation on the training set.

Implementation Details. Our REAG is based on PALM (Bi et al., 2020), an encoder-decoder generative language model pre-trained on a large corpus. It consists of a 12-layer encoder and 12-layer decoder with 768 embedding/hidden size, 3072 feed-forward filter size and 12 attention heads. REAG is trained with a dropout of 0.1 on all layers and attention weights. During training and testing, we truncate the text to 512 tokens and limit the length of the answer to 50 tokens. At test time, answers are generated using beam search with a beam size 5.

[1] https://microsoft.github.io/msmarco/

Model	ROUGE-L	BLEU-1
BIDAF+Seq2Seq[a]	34.15	29.68
S-Net[b]	42.71	36.19
S-Net+Seq2Seq[b]	46.83	39.74
gQA[c]	45.46	40.22
KEAG[d]	51.68	45.97
Masque[e]	69.77	65.56
PALM[f]	69.87	66.31
REAG	**70.98**	**69.12**

Table 2: Performance of generative reading comprehension in ROUGE-L and BLEU-1 on MARCO Q&A+NLG. All our ROUGE scores have a 95% confidence interval of at most ±0.25. [a](Seo et al., 2016);[b](Tan et al., 2017);[c](Mitra, 2017);[d](Bi et al., 2019);[e](Nishida et al., 2019b);[f](Bi et al., 2020).

Ablation	ROUGE-L	BLEU-1
REAG	**70.98**	**69.12**
✗ rationale-span extraction	69.87	66.31
✗ linear-decay joint training	70.45	68.28
✗ pre-training	69.54	68.12

Table 3: Ablation tests of REAG on the MARCO Q&A+NLG dataset.

4.2 Model Comparisons

Table 2 gives the comparison of other state-of-the-art QA models on the MARCO Q&A+NLG dataset in ROUGE-L and BLEU-1. From this table, we observe that generative QA models (e.g., REAG, PALM) are consistently superior to extractive models (e.g., BiDAF) in answer quality. Therefore, generative QA models establish a strong base architecture to be enhanced with the extra signals, which motivates this work. Among the generative models, REAG outperforms all the other state-of-the-art models with an improvement of over 2.8% BLEU-1 point and 1.1% ROUGE-L. Part of the results in the Table 2 are from (Bi et al., 2019), which re-running other researchers' code.

4.3 Ablation Study

We conduct ablation studies to assess the individual contribution of every component in REAG. Table 3 reports the results of full REAG and its ablations on the MS MARCO Q&A NLG dataset.

We evaluate how much rationale-span extraction contributes to generation quality by removing it from the REAG model. This ablation results in a drop from 70.98 to 69.87 on Rouge-L, demonstrating the role of the rationale-span extraction in REAG. In addition, we ablate the linear-decay joint-training which proves to be critical with over 0.5% drops on the metrics after the ablation. In order to exclude the influence of the pre-trained model, we ablate pre-training, retaining the rationale-span extraction. This ablation leads to a drop from 70.98 to 69.54 on Rouge-L, which demonstrates the power of REAG in generating high-quality answers without pre-training.

Method	Semantic Acc	ROUGE-L	BLEU-1
PALM	81.67	69.87	66.31
REAG	**84.33**	**70.31**	**68.59**

Table 4: Comparison of the semantic accuracy, ROUGE-L and BLEU-1 of REAG with those of PALM

	ROUGE-L	BLEU-1
Generated Answers	47.25	50.34
Gold Answers	38.14	43.12

Table 5: Agreement of generated/gold answers with extracted rationales for REAG

4.4 Quantitative Analysis on Semantic Drift

For generative reading comprehension, it is difficult to make the answer completely correct, because even if the semantics are correct, there may be some expression differences from the gold answer. Since neither ROUGE-L nor BLEU-1 can measure it, we conduct a human evaluation of the semantic accuracy. We randomly select 100 questions from the MARCO dev set, and manually evaluate whether the generated answers to these questions are semantically drifted. Table 4 reports the semantic accuracy of REAG and PALM obtained by human. Our REAG model surpasses PALM in generating correct answers without semantic drift. Although our REAG model improves over PALM by 1.1% in automatic evaluation metric ROUGE-L, it gives a 3.26% improvement in semantic accuracy. This shows the fact that in some cases automatic evaluation metrics, such as ROUGE-L and BLEU-1, do not reflect semantic accuracy.

In addition, we compute the agreement of generated/gold answers with extracted rationales for

Example 1	
Relevant Passage	Yes | No Thank you! Flu shots are not made for children under the age of 6 months. If you read the vaccine insert and studies regarding the flu shot and kids, you will see that flu shots don't even work for children under the age of 2.
Question	can a child get a flu vaccine under 6 months?
Gold Answer	No, a child under 6 months can't be given a flu vaccine.
PALM Answer	Yes, a child can get a flu vaccine under 6 months.
REAG Answer	No, a child cannot get a flu vaccine under 6 months.

Example 2	
Relevant Passage	Modesto, Stanislaus County Sales Tax Rate. Details. The sales tax in Modesto is 7.625%, which is about average for cities in Stanislaus County and lower than average for California (8%). Modesto is one of 21 cities in Stanislaus County with a distinct sales tax as listed by the California Board of Equalization. See all cities in Stanislaus County. Advertisement.
Question	what is the sales tax in california
Gold Answer	The sales tax in California is 8%.
PALM Answer	The sales tax in California is 7.625%
REAG Answer	The sales tax in California is 8%

Table 6: Examples of the output of REAG and PALM on the MARCO dataset. The text span of words in blue is the rationale extracted by REAG

REAG in ROUGE-L and BLEU-1. As shown in Table 5, the generated answers are strongly correlated with the rationales, demonstrating the effectiveness of leveraging the rationale signal. Also, the fact that the gold answers have a lower agreement with the rationales indicates that a generative model, as opposed to an extractive one, is needed for the MARCO Q&A+NLG task.

4.5 Case Study

Table 6 gives two examples to show the answers generated by the REAG model and the PALM model. In addition to the answers, we provide the rationales predicted by REAG's encoder to demonstrate the effectiveness of rationale extraction. In both examples, the rationale extraction module identifies the correct rationales, e.g., *Flu shots are not made for children under the age of 6 months.* and *California (8%)*.

In Example 1, PALM is confused by the noise "*Yes*" in the beginning of the passage, which leads to the contrary semantics of its generated answer. With the correctly extracted rationale, our REAG model generates an answer semantically consistent with the gold answer. In Example 2, PALM fails to identify a correct sales tax rate *8%* for California, so the response is incorrect and useless, even if it results in high ROUGE and BLEU scores against the gold answer. In contrast, based on the extracted rationale *California (8%)*, our REAG generates a semantically correct answer.

5 Conclusion and Future Work

This paper presents a novel model REAG that is designed to incorporate an extractive mechanism into a generative QA model. REAG introduces a new task on the encoder to extract rationales. Based on these rationales and original input, a rationale-enriched decoder is proposed to generate an answer with high confidence. The experimental results show that REAG significantly improves the quality and semantic accuracy of generated answers over state-of-the-art models.

References

Bin Bi, Chenliang Li, Chen Wu, Ming Yan, and Wei Wang. 2020. Palm: Pre-training an autoencoding&autoregressive language model for context-conditioned generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Bin Bi, Chen Wu, Ming Yan, Wei Wang, Jiangnan Xia, and Chenliang Li. 2019. Incorporating external knowledge into machine reading for generative 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)*, pages 2521–2530, Hong Kong, China. Association for Computational Linguistics.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading wikipedia to answer open-domain questions. *CoRR*, abs/1704.00051.

Kunlong Chen, Weidi Xu, Xingyi Cheng, Zou Xiaochuan, Yuyu Zhang, Le Song, Taifeng Wang, Yuan Qi, and Wei Chu. 2020. Question directed graph attention network for numerical reasoning over text. *arXiv preprint arXiv:2009.07448*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Ying Ju, Fubang Zhao, Shijie Chen, Bowen Zheng, Xuefeng Yang, and Yunfeng Liu. 2019. Technical report on Conversational Question Answering. *arXiv e-prints*, page arXiv:1909.10772.

Chenliang Li, Weiran Xu, Si Li, and Sheng Gao. 2018. Guiding generation for abstractive text summarization based on key information guide network. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 55–60, New Orleans, Louisiana. Association for Computational Linguistics.

Yizhu Liu, Zhiyi Luo, and Kenny Zhu. 2018. Controlling length in abstractive summarization using a convolutional neural network. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4110–4119, Brussels, Belgium. Association for Computational Linguistics.

Rajarshee Mitra. 2017. A Generative Approach to Question Answering. *arXiv e-prints*, page arXiv:1711.06238.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, and Rangan Majumder. 2016. Ms marco: A human generated machine reading comprehension dataset.

Kyosuke Nishida, Itsumi Saito, Kosuke Nishida, Kazutoshi Shinoda, Atsushi Otsuka, Hisako Asano, and Junji Tomita. 2019a. Multi-style generative reading comprehension. *CoRR*, abs/1901.02262.

Kyosuke Nishida, Itsumi Saito, Kosuke Nishida, Kazutoshi Shinoda, Atsushi Otsuka, Hisako Asano, and Junji Tomita. 2019b. Multi-style generative reading comprehension. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2273–2284, Florence, Italy. Association for Computational Linguistics.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Min Joon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2016. Bidirectional attention flow for machine comprehension. *CoRR*, abs/1611.01603.

Chuanqi Tan, Furu Wei, Nan Yang, Weifeng Lv, and Ming Zhou. 2017. S-net: From answer extraction to answer generation for machine reading comprehension. *CoRR*, abs/1706.04815.

Wei Wang, Ming Yan, and Chen Wu. 2018. Multi-granularity hierarchical attention fusion networks for reading comprehension and question answering. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1705–1714, Melbourne, Australia. Association for Computational Linguistics.

Ming Yan, Jiangnan Xia, Chen Wu, Bin Bi, Zhongzhou Zhao, Ji Zhang, Luo Si, Rui Wang, Wei Wang, and Haiqing Chen. 2018. A deep cascade model for multi-document reading comprehension. In *AAAI*.