Can Question Generation Debias Question Answering Models? A Case Study on Question–Context Lexical Overlap

Kazutoshi Shinoda[1,2] **Saku Sugawara**[2] **Akiko Aizawa**[1,2]
[1]The University of Tokyo
[2]National Institute of Informatics
shinoda@is.s.u-tokyo.ac.jp
{saku,aizawa}@nii.ac.jp

Abstract

Question answering (QA) models for reading comprehension have been demonstrated to exploit unintended dataset biases such as question–context lexical overlap. This hinders QA models from generalizing to under-represented samples such as questions with low lexical overlap. Question generation (QG), a method for augmenting QA datasets, can be a solution for such performance degradation if QG can properly debias QA datasets. However, we discover that recent neural QG models are biased towards generating questions with high lexical overlap, which can amplify the dataset bias. Moreover, our analysis reveals that data augmentation with these QG models frequently impairs the performance on questions with low lexical overlap, while improving that on questions with high lexical overlap. To address this problem, we use a synonym replacement-based approach to augment questions with low lexical overlap. We demonstrate that the proposed data augmentation approach is simple yet effective to mitigate the degradation problem with only 70k synthetic examples.[1]

1 Introduction

Question answering (QA) for machine reading comprehension is a central task in natural language understanding, which requires a model to answer questions given textual contexts. Pretrained language models have been successfully applied to QA and achieve scores higher than those of humans on benchmark datasets such as SQuAD (Rajpurkar et al., 2016). However, QA models have been demonstrated to exploit unintended dataset biases instead of the intended solutions, and lack robustness to challenge test sets whose distributions are different from those of training sets (Jia and Liang, 2017; Sugawara et al., 2018; Gan and Ng, 2019; Ribeiro et al., 2019), which could be a serious problem in real-world applications.

Question generation (QG) has also been extensively studied to augment QA datasets (Du et al., 2017; Du and Cardie, 2018). It is demonstrated that QG can improve not only the in-domain generalization but also the out-of-distribution generalization capability of QA models (Zhang and Bansal, 2019; Lee et al., 2020; Shinoda et al., 2021). In other areas, data augmentation techniques have been successfully used to reduce dataset biases and increase the performance of machine learning models on under-represented samples in vision (McLaughlin et al., 2015; Wong et al., 2016) and language (Zhao et al., 2018; Zhou and Bansal, 2020). Thus, we assume that QG is useful to debias QA models and improve its robustness by augmenting QA datasets. However, it has not been fully studied whether existing QG models can contribute to debiasing QA models (i.e., improve the robustness of QA models to under-represented questions).

In this study, we focus on question–context lexical overlap, inspired by the findings presented in Sugawara et al. (2018). Their work revealed that questions having low lexical overlap with context tend to require reasoning skills rather than superficial word matching, and existing QA models are not robust to these questions (Table 1). To see if data augmentation with recent neural QG models can improve the robustness to those questions, we analyze the performance of BERT (Devlin et al., 2019) trained on SQuAD v1.1 (Rajpurkar et al., 2016) augmented with them. Our analysis reveals that data augmentation with neural QG models frequently sacrifices the QA performance of the BERT-base model on questions with low lexical overlap, while improving that on questions with high lexical overlap. We conjecture that this is because neural QG models frequently generate questions with high lexical overlap as indicated in Table

[1]Our data is publicly available at https://github.com/KazutoshiShinoda/Synonym-Replacement.

Proceedings of the 3rd Workshop on Machine Reading for Question Answering , pages 63–72

C					
Besides earning a reputation as a respected entertainment device, the iPod has also been accepted as a business device. Government departments, major institutions and international organisations have turned to the iPod line as a delivery mechanism for business communication and training, such as the Royal and Western Infirmaries in Glasgow, Scotland, where iPods are used to train new staff.					
Q	A	$\frac{\lvert Q \cap C \rvert}{\lvert Q \rvert}$	$C, Q \rightarrow A'$	$C, A \rightarrow Q'$	$\frac{\lvert Q' \cap C \rvert}{\lvert Q' \rvert}$
Where is Royal and Western Infirmaries located?	Glasgow, Scotland	5/8 = 0.62	Glasgow, Scotland (✓)	Where is the Royal and Western Infirmaries located? (✓)	6/9 = 0.67
Aside from recreational use, in what other arena have iPods found use?	business	4/14 = 0.29	entertainment (✗)	The iPod has been accepted as what kind of device? (✗)	7/11 = 0.64

Table 1: Examples of ground-truth question–answer pairs and predictions of question answering (BERT-base (Devlin et al., 2019)) and generation (SemanticQG (Zhang and Bansal, 2019)) models. C: context, Q: question, A: answer, $\frac{\lvert Q \cap C \rvert}{\lvert Q \rvert}$: question–context lexical overlap, A': predicted answer, Q': generated question. Overlapping words in the questions are underlined.

1. This behavior can be interpreted as a consequence of the recent QG models pursuing higher average BLEU scores on SQuAD, which inherently contains reference questions with high lexical overlap, by copying many words from contexts to generate questions. By doing so, QG models can amplify the lexical overlap bias in the original dataset.

To address the performance degradation, we use a simple data augmentation approach using synonym replacement to generate questions with low question–context lexical overlap. We found that the proposed approach not only debiases the dataset but also improves the QA performance on questions with low lexical overlap with only 70k synthetic examples, whereas conventional neural QG approaches use more than one million synthetic examples.

In summary, our contributions are as follows:

- We found that not only QA but also QG models are biased in terms of question–context lexical overlap; that is, QG models fail to generate questions with low lexical overlap (§2).
- We discovered that data augmentation using recent neural QG models does not contribute to debias QA datasets; rather, it frequently degrades the QA performance on questions with low lexical overlap, while improving that on questions with high lexical overlap (§4).
- We demonstrated that the proposed simple data augmentation approach using synonym replacement (§3) for augmenting questions with low lexical overlap is effective to improve QA performance on questions with low lexical overlap with only 70k synthetic examples (§4), while preserving or slightly hurting the overall accuracy.

2 Revisiting the QA and QG Performance in Terms of Question–Context Lexical Overlap

In this paper, we denote question–context lexical overlap as QCLO. We define QCLO as the ratio of the overlapping words between question Q and context C to the total number of words in question.[2] Precisely, QCLO is calculated as

$$\text{QCLO} = \frac{|Q \cap C|}{|Q|}. \quad (1)$$

The second example in Table 1 indicates a question with lower QCLO is neither answered nor generated correctly by neural models. To investigate this phenomenon, we first analyze the QA and QG performance in terms of QCLO.

Experimental setups For QA, we use the fine-tuned BERT-base and -large models (Devlin et al., 2019). For QG, we use SemanticQG (Zhang and Bansal, 2019).[3] For the dataset, we use the SQuAD-Du dataset; the train, dev, and test split of SQuAD v1.1 (Rajpurkar et al., 2016) proposed by Du et al. (2017), which we denote as $\text{SQuAD}^{\text{Du}}_{\text{train}}$, $\text{SQuAD}^{\text{Du}}_{\text{dev}}$, $\text{SQuAD}^{\text{Du}}_{\text{test}}$, respectively. This split

[2] When computing lexical overlap, we do not exclude stop words because even overlapping stop words are important cues to determine the correct answer.

[3] We used the ELMo+QPP&QAP (Zhang and Bansal, 2019) model for QG.

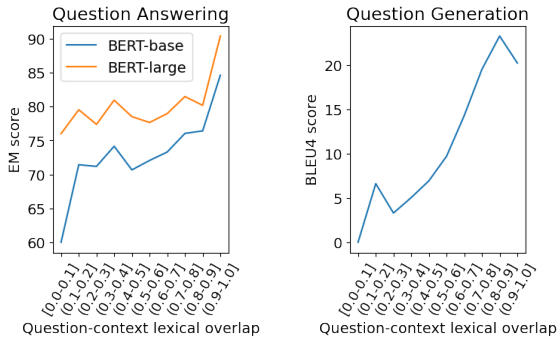

Figure 1: Exact match (EM) score of BERT models and BLEU-4 score of SemanticQG (Zhang and Bansal, 2019) on the test set of SQuAD-Du for each range of Question–Context Lexical Overlap (QCLO). See Eq. 1 for the definition of QCLO. Both the QA and QG models degrade the scores on questions with low QCLO.

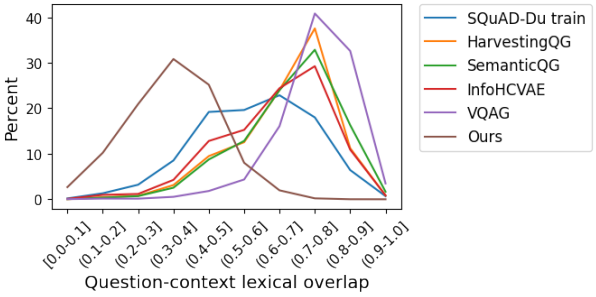

Figure 2: The percentages of questions in the datasets, SQuAD-Du (Du et al., 2017), HarvestingQG (Du and Cardie, 2018), SemanticQG (Zhang and Bansal, 2019), InfoHCVAE (Lee et al., 2020), VQAG (Shinoda et al., 2021), and ours (§3), for each range of QCLO. While neural question generation models are biased towards generating questions with high QCLO, ours can generate questions with low QCLO.

is commonly used in the QG literature (Du and Cardie, 2018; Zhang and Bansal, 2019) because the original test set is not released. The numbers of question, answer and context triples in $\text{SQuAD}^{\text{Du}}_{\text{train}}$, $\text{SQuAD}^{\text{Du}}_{\text{dev}}$, and $\text{SQuAD}^{\text{Du}}_{\text{test}}$, are 76k, 11k, and 12k, respectively.

Results We show the result in Figure 1. This indicates that the performance of the BERT models on the questions with lower QCLO is degraded compared to the questions with higher QCLO. For QG, the BLEU-4 score (Papineni et al., 2002) is highly correlated with QCLO, which means that the model fails to generate questions with low QCLO accurately.

We also show the distributions in terms of QCLO of questions generated by recent neural QG models (HarvestingQG (Du and Cardie, 2018), SemanticQG (Zhang and Bansal, 2019), InfoHCVAE (Lee et al., 2020), and VQAG (Shinoda et al., 2021)) in Figure 2. This indicates that all the QG models are biased towards generating questions with higher QCLO than $\text{SQuAD}^{\text{Du}}_{\text{train}}$, which is used to train those QG models.

Based on the result, we suspect that when neural QG is used to augment a QA dataset, the degraded QG performance on questions with low QCLO could exacerbate the degraded QA performance. Our experiments in §4 show that this is often true. We hypothesize that this is caused by the strong tendency of neural QG models to generate questions with high QCLO as shown in Figure 2.

3 Method

We assume that if we augment questions with low QCLO unlike existing neural QG approaches, the robustness of QA models to questions with low QCLO can be improved. In this section, we describe the proposed method for generating questions with low QCLO. We extend the idea of synonym replacement used in (Wei and Zou, 2019) to reduce the lexical overlap. The proposed method is as follows:

1. List all the overlapping words between question and context.
2. Replace every word in the listed words other than predefined stop words with one of its synonyms chosen randomly from WordNet (Miller, 1995), and obtain a synthetic question.
3. If the lexical overlap decreases after synonym replacement, add the synthetic question to our dataset; if not, discard the question.

After repeating this procedure once for every ground-truth question in the training set, we obtain 70k synthetic questions with significantly lower lexical overlap, as indicated in Figure 2 (ours). For example, *What is heresy mainly at odds with?* is converted into *What is heterodoxy mainly at odds with?*, and *How many documents remain classified?* is converted into *How many text file remain classified?*. Because *heterodoxy*, *text*, and *file* do not appear in the contexts, the lexical overlap is reduced in each example.

It is worth mentioning a couple of limitations of our method. First, synonym replacement may slightly change the meaning of questions depending on the context. Second, our approach relies on the assumption that annotated questions are available, which makes it impossible to apply to unlabeled passages.

4 Experiments

To determine the effect of data augmentation on improving the QA model robustness to questions with low QCLO, we conducted experiments with several QG approaches.

Dataset We used the SQuAD-Du dataset as in §2. Considering the QCLO statistics of SQuAD displayed in Figure 2, we split $\text{SQuAD}^{\text{Du}}_{\text{dev}}$ and $\text{SQuAD}^{\text{Du}}_{\text{test}}$ into **Easy** and **Hard** subsets that contain questions with QCLO greater than 0.3, and the others, respectively. Our Easy and Hard subsets offered concise, yet sufficient, evaluation in terms of QCLO.

Baselines We adopted the following four baselines that use neural QG models for data augmentation.

- **HarvestingQG** (Du and Cardie, 2018) generates question–answer pairs from 10,000 top-ranking Wikipedia articles with neural answer extraction and question generation.[4] The size is 1.2 million.
- **SemanticQG** (Zhang and Bansal, 2019) is a QG model that uses reinforcement learning to generate semantically valid questions. Following this work, we generated questions using the publicly available model[5] from the same context–answer pairs as HarvestingQG. The size is 1.2 million.
- **InfoHCVAE** (Lee et al., 2020) is a question–answer pair generation model based on conditional variational autoencoder with mutual information maximization. We trained this model on $\text{SQuAD}^{\text{Du}}_{\text{train}}$, and then generated 50 questions and answers from each context in $\text{SQuAD}^{\text{Du}}_{\text{train}}$. The size is 824k.
- **VQAG** (Shinoda et al., 2021) is a question–answer pair generation model based on conditional variational autoencoder with explicit KL control. We used the publicly available dataset.[6] The size is 432k.

The distributions of the lexical overlap of these datasets are presented in Figure 2. We indicate that these methods are more biased towards high lexical overlap than $\text{SQuAD}^{\text{Du}}_{\text{train}}$, which was used as the training set for these QG models.

Experimental Setups As in our previous experiment (§2), we used BERT-base and -large models, whose total number of parameters are 110M and 340M, respectively. Dhingra et al. (2018) proposed to pretrain a QA model using synthetic data composed of cloze-style questions and then fine-tune it on the ground-truth data. We adopted the pretrain-and-fine-tune approach for the neural QG approaches, which generated over 1.2 million questions. However, as discussed by Zhang and Bansal (2019), we observed that when the size of the synthetic data was small or similar to the ground-truth data, a performance gain could not be obtained by the pretrain-and-fine-tune approach. Thus, for the proposed approach, which generated 70k questions, we fine-tuned QA models on the ground-truth data randomly mixed with the generated data. We used the Hugging Face's implementation of BERT (Wolf et al., 2019). We use the Adam (Kingma and Ba, 2014) optimizer with epsilon set to 1e-8. The batch size was 32 for all the settings. In both the pretraining and fine-tuning procedure, the learning rate decreased linearly from 3e-5 to zero. We train the QA models for one epoch for pretraining with synthetic data and two epochs for fine-tuning with $\text{SQuAD}^{\text{Du}}_{\text{train}}$.

Results The results of the data augmentation are displayed in Table 2. In all the settings, the proposed approach achieved the best EM score on the Hard subset. Notably, the proposed method significantly improved the performance by **2.72 (EM) / 1.50 (F1)** points using BERT-base on the Hard subset in the test set, while maintaining the overall scores compared to the no data augmentation baseline. This improvement indicates that the proposed approach for debiasing the dataset in terms of QCLO is helpful for addressing the performance

[4] https://github.com/xinyadu/harvestingQA

[5] https://github.com/ZhangShiyue/QGforQA

[6] https://github.com/KazutoshiShinoda/VQAG

		$\text{SQuAD}^{\text{Du}}_{\text{dev}}$ (EM/F1)			$\text{SQuAD}^{\text{Du}}_{\text{test}}$ (EM/F1)		
Model	Train Source	Hard	Easy	ALL	Hard	Easy	ALL
base	$\text{SQuAD}^{\text{Du}}_{\text{train}}$	72.31/81.11	80.74/88.39	80.35/88.05	70.88/81.99	73.22/84.75	73.06/84.57
	+ HarvestingQG	70.25/78.27	80.06/87.62	79.60/87.19	69.28/79.92	73.15/84.20	72.90/83.93
	+ SemanticQG	70.45/80.25	81.70/89.08	81.17/88.67	71.68/82.49	**74.39/85.59**	**74.21/85.39**
	+ InfoHCVAE	72.05/80.66	81.79/**89.35**	81.34/**88.95**	73.47/**83.91**	73.50/85.08	73.48/84.99
	+ VQAG	73.29/82.04	**81.88**/88.93	**81.48**/88.62	71.60/83.07	73.79/85.23	73.63/85.08
	+ Ours	**73.50/82.81**	80.34/87.81	80.02/87.58	**73.60**/83.49	73.08/84.41	73.11/84.34
large	$\text{SQuAD}^{\text{Du}}_{\text{train}}$	78.72/87.71	**87.06/93.23**	**86.67/92.98**	77.93/87.84	**79.33/89.88**	**79.24/89.74**
	+ HarvestingQG	79.13/86.92	85.55/92.12	85.26/91.88	76.99/86.61	77.58/88.28	77.54/88.17
	+ SemanticQG	79.96/87.73	85.90/92.57	85.62/92.35	76.99/87.29	77.82/88.68	77.77/88.59
	+ InfoHCVAE	77.85/86.44	85.25/92.15	84.91/91.89	76.00/87.55	78.02/88.90	77.87/88.80
	+ VQAG	79.50/87.55	86.68/93.01	86.35/92.76	77.33/87.70	78.98/89.36	78.86/89.25
	+ Ours	**81.37/88.33**	86.49/92.78	86.25/92.57	**78.40/88.52**	77.94/89.00	77.96/88.97

Table 2: QA performance with data augmentation. EM/F1 scores on the Hard (where QCLO ≤ 0.3) and Easy (where QCLO > 0.3) subsets, and the whole set of $\text{SQuAD}^{\text{Du}}_{\text{dev}}$ and $\text{SQuAD}^{\text{Du}}_{\text{test}}$ are reported.

degradation. However, the proposed approach degraded the scores on the Easy subsets when using BERT-large. Addressing the trade-off between the scores in the Hard and Easy subsets using BERT-large is future work.

When using BERT-base, the neural QG baselines except for HarvestingQG improved the scores on the Easy subset; however, the baselines except for InfoHCVAE often degraded the scores on the Hard subset. This could be due to the tendency to generate questions with high QCLO (Figure 2).

When using BERT-large, the QG approaches often fail to improve the scores in both the Hard and Easy subsets. Generating useful examples for a larger model is more challenging than for a smaller one according to these results. Utilizing pretrained language models for QG may be useful given the fact that only RNNs are used in all the baseline QG methods in our experiments.

HarvesingQG was not effective in almost all the settings. Comparing its scores with those of SemanticQG, which used the same context–answer pairs as HarvestingQG, some feature of generated questions other than lexical overlap appeared to be critical in improving the QA scores on the Easy subset, because the distributions of QCLO of two synthetic datasets were similar to each other (see Figure 2).

For further boosting the overall average score, we can make an ensemble prediction using the best performing models in the Easy and Hard subsets, although improving the overall scores is not the main focus in this paper. The performance gains were positive but not very significant in our case. We leave utilizing the ensemble prediction to address the performance trade-off to future work.

5 Qualitative Analysis

To demonstrate the effect of the baseline QG models and proposed method qualitatively, we present examples in both the Hard (QCLO ≤ 0.3) and Easy (QCLO > 0.3) subsets in Table 3. The first two examples show that only the QA model trained with the proposed method could correctly answer the questions. Answering the questions in these examples required a knowledge of synonyms, such as "recreational" vs. "entertainment," "besides" vs. "aside from," "employees" vs. "workers," and "kill oneself" vs. "commit suicide." These examples imply that the proposed data augmentation method based on synonym replacement enabled the QA model to acquire knowledge regarding synonyms. This kind of reasoning beyond superficial word matching is indispensable for QA systems to achieve human-level language understanding.

The third example in Table 3 displays an example where data augmentation using the neural QG models made the original prediction incorrect. This example implies that current QG models may harm the robustness of QA models to questions with low QCLO. As Geirhos et al. (2020) discussed, if QG models just amplify the dataset bias, QA models could learn dataset-specific solutions (i.e., shortcuts) and fail to generalize to challenge test sets.

In contrast, the fourth and fifth examples in Table 3 display examples in the Easy subset where

Besides earning a reputation as a respected entertainment *(Original, InfoHCVAE)* device, the iPod has also been accepted as a **business** *(Ours)* device. Government departments, major institutions and international organisations have turned to the iPod line as a delivery mechanism for business communication and training, such as the Royal and Western Infirmaries *(HarvestingQG, SemanticQG, VQAG)* in Glasgow, Scotland, where iPods are used to train new staff. — Aside from recreational use, in what other arena have iPods found use? (QCLO: 0.29)
In **2010** *(Ours)*, a number of workers committed suicide at a Foxconn operations in China. Apple, HP, and others stated that they were investigating the situation. Foxconn guards have been videotaped beating employees. Another employee killed himself in 2009 *(Original, HarvestingQG, SemanticQG, VQAG)* when an Apple prototype went missing, and claimed in messages to friends, that he had been beaten and interrogated. — In what year did Chinese Foxconn emplyees* kill themselves? (*: annotator's typo) (QCLO: 0.2)
The BBC began its own regular television programming from the basement of Broadcasting House, London, on 22 August 1932 *(HarvestingQG, SemanticQG)*. The studio moved to larger quarters in 16 Portland Place, London, in **February 1934** *(Original, Ours)*, and continued broadcasting the 30-line images, carried by telephone line to the medium wave transmitter at Brookmans Park, until 11 September 1935, by which time advances in all-electronic television systems made the electromechanical broadcasts obsolete. — When did the BBC first change studios? (QCLO: 0.25)
Peyton Manning *(VQAG)* became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39 *(Original, Ours)*. The past record was held by **John Elway** *(HarvestingQG, SemanticQG, InfoHCVAE)*, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver's Executive Vice President of Football Operations and General Manager. — Prior to Manning, who was the oldest quarterback to play in a Super Bowl? (QCLO: 0.88)
Despite being relatively unaffected by the embargo *(Original, HarvestingQG, VQAG, Ours)*, the UK nonetheless faced an oil crisis of its own - **a series of strikes by coal miners and railroad workers** *(SemanticQG, InfoHCVAE)* over the winter of 1973–74 became a major factor in the change of government. Heath asked the British to heat only one room in their houses over the winter. The UK, Germany, Italy, Switzerland and Norway banned flying, driving and boating on Sundays. Sweden rationed gasoline and heating oil. The Netherlands imposed prison sentences for those who used more than their ration of electricity. — What caused UK to have an oil crisis in its own country? (QCLO: 0.62)

Table 3: Illustrative predictions on $\text{SQuAD}^{\text{Du}}_{\text{dev}}$ and $\text{SQuAD}^{\text{Du}}_{\text{test}}$ by a BERT-base model trained on $\text{SQuAD}^{\text{Du}}_{\text{train}}$ (Original), +HarvestingQG, +SemanticQG, +InfoHCVAE, +VQAG, and +Ours. The ground truth answers are in **bold**. The incorrectly predicted answers are written in red. The QA models that predict them are written in *italics*. The overlapping words in the questions are underlined. Question–context lexical overlap (QCLO) is given in parentheses.

data augmentation with neural QG models is beneficial, while the original and proposed models fail to answer them correctly. These examples require multiple-sentence reasoning, i.e., one has to read and understand multiple sentences to answer these questions. This observation implies that some under-represented features (e.g., multiple-sentence reasoning (Rajpurkar et al., 2016)) exist even in the Easy subset, and the existing neural QG models might amplify such features (possibly by copying many words from multiple sentences to formulate questions) and make it easy to capture them. Investigating what kind of features are learned by using data augmentation with neural QG models in more detail is future work.

6 Related Work

The Robustness of QA models Pretrained language models such as BERT (Devlin et al., 2019) have surpassed the human score on the SQuAD leaderboard.[7] However, such powerful QA models have been shown to exhibit the lack of robustness. A QA model that is trained on SQuAD is not robust to paraphrased questions (Gan and Ng, 2019), implications derived from SQuAD (Ribeiro et al., 2019), questions with low lexical overlap (Sugawara et al., 2018), and other QA datasets (Yogatama et al., 2019; Talmor and Berant, 2019; Sen and Saffari, 2020). Ko et al. (2020) showed that extractive QA model can suffer from positional bias and fail to generalize to different answer positions.

The lack of robustness demonstrated in these studies can be explained by shortcut learning of deep neural networks (Geirhos et al., 2020). A high score on an in-distribution test set can be achieved by just exploiting unintended dataset biases (Levesque, 2014). Therefore, evaluating QA models only on an in-distribution test set is not enough to evaluate the robustness of the QA models.

Question Generation for Question Answering QG has been studied extensively in order to augment QA datasets and boost the QA performance, which has been evaluated primarily on SQuAD (Du et al., 2017; Zhou et al., 2018; Yang et al., 2017; Zhang and Bansal, 2019). Question answer pair generation, which consists of answer candidate extraction and QG, has been also received attention because question-worthy answers for the input of QG are not freely available (Du and Cardie, 2018; Lee et al., 2020; Shinoda et al., 2021). The de facto standard of QG models is to utilize a copy mechanism (Gu et al., 2016; Gulcehre et al., 2016). The tendency of QG models to copy words from textual contexts as indicated in Figure 2 is partially due to this copy mechanism. While the existing QG works have increased the BLEU scores on SQuAD[8] and successfully generated fluent questions in terms of human scores, the bias regarding lexical overlap in QG has not received sufficient attention.

Data Augmentation and Dataset Bias Data augmentation has been widely used in other domains to reduce dataset biases such as the background bias in person re-identification (McLaughlin et al., 2015), the gender bias in coreference resolution (Zhao et al., 2018), and the lexical bias in natural language inference (Zhou and Bansal, 2020). These works repeated training examples or added synthetic data to increase under-represented samples and reduce the imbalance in a training set. Our proposed approach has the same motivation as these works.

On the other hand, data augmentation can unintentionally introduce or amplify dataset bias. Back-translation (Sennrich et al., 2016), which is the common data augmentation approach for machine translation, can introduce the translationese bias. That is, machine translation systems trained with back-translation, compared to ones without back-translation, can enhance the BLEU scores when the input is translationese (i.e., human-translated texts) but harm the BLEU scores when the input is naturally occurring texts (Edunov et al., 2020; Marie et al., 2020). This phenomenon is analogous to the observation in our work, where we demonstrated that SQuAD QG models are biased towards generating questions with high QCLO, and this tendency can harm the QA performance on questions with low QCLO while improving that on questions with high QCLO.

7 Conclusion

We demonstrated that not only QA models but also QG models are biased in terms of the question–context lexical overlap. To determine the influence of the bias, we analyzed the QA performance with data augmentation using the recent QG models. We demonstrated that they frequently degraded the QA

[7] https://rajpurkar.github.io/SQuAD-explorer/

[8] http://aqleaderboard.tomhosking.co.uk/squad

performance on questions with low lexical overlap, while improving that on questions with high lexical overlap when using BERT-base. To address this problem, we designed a simple approach using synonym replacement to debias a QA dataset. We demonstrated that the proposed approach improved the QA performance on questions with low lexical overlap while maintaining or slightly degrading the overall scores with only 70k synthetic examples.

Our results suggest that future research in QG for data augmentation should exercise caution to prevent the amplification of dataset bias in terms of lexical overlap. In addition, what features are learned by data augmentation with neural QG models is worth to be explored in more detail to clarify what is improved and what is not improved by QG. It is also worth investigating whether our findings still hold in other QA datasets where annotated questions have lower lexical overlap than those in SQuAD.

Acknowledgements

We would like to thank the anonymous reviewers for their detailed and valuable comments. This work was supported by NEDO SIP-2 "Big-data and AI-enabled Cyberspace Technologies," and JSPS KAKENHI Grant Numbers 21H03502, 20K23335.

References

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.

Bhuwan Dhingra, Danish Danish, and Dheeraj Rajagopal. 2018. Simple and effective semi-supervised question answering. 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 582–587, New Orleans, Louisiana. Association for Computational Linguistics.

Xinya Du and Claire Cardie. 2018. Harvesting paragraph-level question-answer pairs from Wikipedia. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1907–1917, Melbourne, Australia. Association for Computational Linguistics.

Xinya Du, Junru Shao, and Claire Cardie. 2017. Learning to ask: Neural question generation for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1342–1352, Vancouver, Canada. Association for Computational Linguistics.

Sergey Edunov, Myle Ott, Marc'Aurelio Ranzato, and Michael Auli. 2020. On the evaluation of machine translation systems trained with back-translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2836–2846, Online. Association for Computational Linguistics.

Wee Chung Gan and Hwee Tou Ng. 2019. Improving the robustness of question answering systems to question paraphrasing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6065–6075, Florence, Italy. Association for Computational Linguistics.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673.

Jiatao Gu, Zhengdong Lu, Hang Li, and Victor O.K. Li. 2016. Incorporating copying mechanism in sequence-to-sequence learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1631–1640, Berlin, Germany. Association for Computational Linguistics.

Caglar Gulcehre, Sungjin Ahn, Ramesh Nallapati, Bowen Zhou, and Yoshua Bengio. 2016. Pointing the unknown words. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 140–149, Berlin, Germany. Association for Computational Linguistics.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031, Copenhagen, Denmark. Association for Computational Linguistics.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Miyoung Ko, Jinhyuk Lee, Hyunjae Kim, Gangwoo Kim, and Jaewoo Kang. 2020. Look at the first sentence: Position bias in question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1109–1121, Online. Association for Computational Linguistics.

Dong Bok Lee, Seanie Lee, Woo Tae Jeong, Donghwan Kim, and Sung Ju Hwang. 2020. Generating diverse and consistent QA pairs from contexts with information-maximizing hierarchical conditional VAEs. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 208–224, Online. Association for Computational Linguistics.

Hector J. Levesque. 2014. On our best behaviour. *Artif. Intell.*, 212(1):27–35.

Benjamin Marie, Raphael Rubino, and Atsushi Fujita. 2020. Tagged back-translation revisited: Why does it really work? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5990–5997, Online. Association for Computational Linguistics.

Niall McLaughlin, Jesus M. Del Rincon, and Paul Miller. 2015. Data-augmentation for reducing dataset bias in person re-identification. In *2015 12th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–6, Karlsruhe, Germany. IEEE.

George A. Miller. 1995. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. 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.

Marco Tulio Ribeiro, Carlos Guestrin, and Sameer Singh. 2019. Are red roses red? evaluating consistency of question-answering models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6174–6184, Florence, Italy. Association for Computational Linguistics.

Priyanka Sen and Amir Saffari. 2020. What do models learn from question answering datasets? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2429–2438, Online. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Kazutoshi Shinoda, Saku Sugawara, and Akiko Aizawa. 2021. Improving the robustness of QA models to challenge sets with variational question-answer pair generation. In *Proceedings of the ACL-IJCNLP 2021 Student Research Workshop*, pages 197–214, Online. Association for Computational Linguistics.

Saku Sugawara, Kentaro Inui, Satoshi Sekine, and Akiko Aizawa. 2018. What makes reading comprehension questions easier? In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4208–4219, Brussels, Belgium. Association for Computational Linguistics.

Alon Talmor and Jonathan Berant. 2019. MultiQA: An empirical investigation of generalization and transfer in reading comprehension. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4911–4921, Florence, Italy. Association for Computational Linguistics.

Jason Wei and Kai Zou. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. 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 6382–6388, Hong Kong, China. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

S. C. Wong, A. Gatt, V. Stamatescu, and M. D. McDonnell. 2016. Understanding data augmentation for classification: When to warp? In *2016 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, pages 1–6, Gold Coast, QLD, Australia. IEEE.

Zhilin Yang, Junjie Hu, Ruslan Salakhutdinov, and William Cohen. 2017. Semi-supervised QA with generative domain-adaptive nets. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1040–1050, Vancouver, Canada. Association for Computational Linguistics.

Dani Yogatama, Cyprien de Masson d'Autume, Jerome Connor, Tomas Kocisky, Mike Chrzanowski, Lingpeng Kong, Angeliki Lazaridou, Wang Ling, Lei Yu, Chris Dyer, et al. 2019. Learning and evaluating general linguistic intelligence. *arXiv preprint arXiv:1901.11373*.

Shiyue Zhang and Mohit Bansal. 2019. 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), pages 2495–2509, Hong Kong, China. Association for Computational Linguistics.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. 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 15–20, New Orleans, Louisiana. Association for Computational Linguistics.

Qingyu Zhou, Nan Yang, Furu Wei, Chuanqi Tan, Hangbo Bao, and Ming Zhou. 2018. Neural question generation from text: A preliminary study. In *Natural Language Processing and Chinese Computing*, pages 662–671, Cham. Springer International Publishing.

Xiang Zhou and Mohit Bansal. 2020. Towards robustifying NLI models against lexical dataset biases. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8759–8771, Online. Association for Computational Linguistics.