Handling Anomalies of Synthetic Questions in Unsupervised Question Answering

Giwon Hong* **Junmo Kang*** **Doyeon Lim*** **Sung-Hyon Myaeng**
School of Computing, KAIST
Daejeon, Republic of Korea
{gch02518, junmo.kang, dylim, myaeng}@kaist.ac.kr

Abstract

Advances in Question Answering (QA) research require additional datasets for new domains, languages, and types of questions, as well as for performance increases. Human creation of a QA dataset like SQuAD, however, is expensive. As an alternative, an unsupervised QA approach has been proposed so that QA training data can be generated automatically. However, the performance of unsupervised QA is much lower than that of supervised QA models. We identify two anomalies in the automatically generated questions and propose how they can be mitigated. We show our approach helps improve unsupervised QA significantly across a number of QA tasks.

1 Introduction

Machine Reading Question Answering (MRQA) is a popular NLP task that attempts to answer questions from associated texts. A popular challenge is SQuAD (Rajpurkar et al., 2016), which has 100K ⟨*context, answer, question*⟩ triples created by humans, and other challenging tasks and datasets have been created. However, it is costly to create a dataset for a new QA task or expand an existing one.

Context	Generated Question
... Montecarlo Automobile began **manufacturing its first street legal GT car in 1989 ...**	How much does **manufacturing its street legal GT car in 1989 ... ?**
... it is hot in the summer and freezing in the **winter** ...	**When is it hot in the summer and freezing?**

Table 1: A copy-type question **(green)** and an unanswerable question **(red)**.

As an attempt to relieve the burden of human efforts, Lewis et al. (2019) proposed an *unsupervised question answering (UQA)* where the *answers* and *questions* are generated in an unsupervised way. The core of the approach is to create a cloze question from a context and then apply a translator to convert it into a natural question. This pioneering work provides a reasonable baseline for UQA but leaves a large room for improvement on question generation.

We have identified two anomalies of the unsupervised question generation (UQG) model, in addition to the question-type issues (Kang et al., 2019; Lim et al., 2020). First, a large number of the tokens in the generated questions are copied from the answer-containing sentences in the contexts, producing *copy-type* questions. As a result, a QA model trained with a large number of these "easy" questions is prone to fail when confronted with questions not directly derivable from the contexts. Second, the UQG model sometimes creates *unanswerable* questions that are semantically too far away from the context, which only acts as noise when training a QA model. An example of the two anomaly types is shown in Table 1.

The main thrust of this paper is to demonstrate that the anomalies of UQG can be overcome by means of relatively simple techniques, which produce higher quality questions that in turn help achieving better QA performance. The main goal of this study is to examine the potential of the individual methods. Note

*Equal contribution.

Proceedings of the 28th International Conference on Computational Linguistics, pages 3441–3448
Barcelona, Spain (Online), December 8-13, 2020

that a more complex method of modifying the QG process itself has been proposed concurrently (Kang et al., 2020). The key contributions of this paper are: 1) identification of the two problems associated with UQG: excessive copying of the context in question generation and unanswerable questions, 2) methods to mitigate the problems, which together generate higher quality questions, and 3) demonstration of the validity of the raised issues and efficacy and generalizability of the proposed methods.

2 Limitations of Unsupervised Question Generation

The analyses in this section probe the nature of the UQG problems and set the stage for developing our methods that alleviate the problems and help to improve the performance on various QA tasks.

2.1 Excessive Copying from Contexts

In the UQA work by Lewis et al. (2019), they generate questions by first creating cloze questions and then converting them into natural questions. Cloze questions are generated by selecting named entities or noun phrases from the context as the answers and converted into natural questions by utilizing Unsupervised Neural Machine Translator (UNMT) (Lample et al., 2018).

However, this approach would make the generated natural questions highly overlapped with the cloze questions and therefore with the contexts. This occurs because UMNT is basically an autoencoder trained to make the inputs and the outputs identical, leading itself to copy from cloze questions as much as possible. To measure the degree to which the copying problem exists, we conduct an experiment comparing the UQG dataset with the SQuAD 1.1 (Rajpurkar et al., 2016) dataset. We compute BLEU-4 scores (Papineni et al., 2002) between a question and the corresponding answer-containing sentence, assuming that the higher the BLEU score, the more copying. We observe a big difference between the two datasets, 23.35 for UQG and 3.02 for SQuAD, which proves that the UQG dataset contains an unreasonable number of copy-type questions, compared to the human-generated dataset. The excessive copying problem would hamper QA models from eliciting answers if the context does not include enough question words.

2.2 Unanswerable Questions

Since the UQG model inherits the limitations of a seq-to-seq model, a generated question may contain a word incoherent with the context. Despite the fact that the randomness makes the trained model robust and generalizable, it can make the questions deviate so much from the semantics of the context that they are unanswerable even by a human.

In order to verify the existence of this problem in the UQG dataset, we train a QA model[1] on SQuAD 2.0 (Rajpurkar et al., 2018), which contains no-answer questions, so that it learns how to discern unanswerable questions. When tested on the UQG dataset, the trained model identifies 41.8% of the questions as unanswerable. Since the model's ability in identifying no-answer cases is 84.25 in F1 when tested with the SQuAD 2.0 dev set, we can estimate that more than 30% (41.8×84.25) of the UQG instances are unanswerable. Note that SQuAD 2.0 is not used in any other experiments in order to make our approach completely "unsupervised".

3 Methods

Given the analysis result, we propose two approaches to refine UQA questions with no supervision. The proposed methods[2] are: 1) Paraphrasing questions, and 2) Trimming undesirable questions.

3.1 Paraphrasing

Inspired by previous work on the use of paraphrasing for QA (Duboue and Chu-Carroll, 2006; Gan and Ng, 2019), we attempt to mitigate the problem of excessive copying problem by diversifying the questions through paraphrasing. We opt for using back-translation because it is naturally compatible with the UQA framework. We leverage a publicly available translation system that can handle multiple

[1]The QA model is based on BERT-Large (Devlin et al., 2019).

[2]An overall pipeline procedure can be found in Algorithm 1 in Appendix A.1.

	Portuguese	Italian	French	Spanish	German	Russian	Arabic	Hindi	Chinese
BLEU	41.40	40.99	35.36	34.99	37.02	**28.80**	**25.32**	**21.13**	**15.59**

Table 2: BLEU-4 scores between original questions and paraphrased questions using different languages.

language pairs. An original UQG question Q_s in English (source) is translated into Q_t in one of the target languages, which is then back-translated into English to produce Q'_s.

We start with a hypothesis that copy-type questions can be maximally avoided by choosing a target language that is most widely different from English. To verify the hypothesis, we employ a BLEU-4 score (Papineni et al., 2002) between an original question and a paraphrased question. We sample 10K questions from the UQA dataset and paraphrase them using each of the nine different languages as the target for back-translation. We then calculate the average BLEU-4 score for each target language (Table 2). The higher the score, the more diversified questions we generate. As expected, there is a division between the European languages and others like Hindi, Arabic, and Chinese, with Russian in the middle. We assume the computed BLUE scores correlate with the desired diversification.

Instead of sticking to a single target language, we can also use multiple languages to apply different styles of diversification for English sentences. To make use of the linguistic differences between English and the target languages, we devise a more flexible method of using the inverse of the BLEU scores. We calculate the probability that a particular target language t should be selected, as in Equation (1), and use it as a distribution on which a particular language is sampled for a given question instance.

$$P(t) = \frac{exp(\frac{1}{BLEU(t,en)}/T)}{\sum_i exp(\frac{1}{BLEU(i,en)}/T)} \tag{1}$$

where T is a hyperparameter that controls the smoothness of the probability and is set to 1.2.[3]

3.2 Trimming Based on Confidence Scores

In an unsupervised setting, selecting the most useful subset for training is an important issue because it is hard to guarantee the quality of the questions. In addition, the previous paraphrasing scheme (Section 3.1) might produce low-quality questions for some instances. Avoiding a supervised method for quality enhancement, we propose to trim the questions that are likely to be either unanswerable or of copy-type after paraphrasing. To this end, we compute the confidence scores for individual questions using a BERT QA model trained on the 100K UQG instances to which paraphrasing has been applied.

A confidence score measures the degree to which the predicted answer A* matches the actual answer A given the triple $\langle C, A, Q\rangle$. It reflects how comfortably the question ends up predicting the answer. A high score indicates the question is easily answerable whereas a low score signals it is hardly answerable. More specifically, we obtain the hidden states H of the input tokens from the BERT-QA based scorer given the input $\langle Q, C\rangle$ and then take the logits for the start and end of the answer span:

$$H = BERT_{scorer}(Q, C),\ H \in \mathbb{R}^{|C|\times h} \tag{2}$$

$$L = H \cdot W + b,\ W \in \mathbb{R}^{h\times 2},\ b \in \mathbb{R}^{|C|\times 2} \tag{3}$$

where h is the size of hidden layer and W is a parameter to be learned. A confidence score is calculated by summing the logits for the *start* and the *end* tokens of the predicted answer within the context.

$$score = L_{[start,1]} + L_{[end,2]} \tag{4}$$

We hypothesize that questions with excessively high or low confidence scores are copy-type or unanswerable questions, respectively. To trim these low-quality questions, we first sort all the triple instances

[3]The number of instances sampled for each language by this inverse BLEU-based method is shown in Appendix A.2

Model	SQuAD 1.1	NewsQA	TriviaQA	SearchQA	HotpotQA	Natural Questions
Lewis et al. (2019)	54.30	26.72	23.32	32.94	25.15	20.38
+ Paraphrasing	57.06	28.24	31.95	33.91	28.41	29.31
+ Trimming	**58.51**	**28.54**	**32.22**	**35.05**	**29.49**	**31.15**

Table 3: F1 scores of the baseline and the proposed approach on various QA task datasets.

D_c in ascending order of their confidence scores. We then discard the top and bottom k instances in terms of their confidence scores, generating D_t, from which a subset of training instances are randomly sampled to create the final train set T_{final}. In the experiments to be described in Section 4, the threshold k and the number of instances in T_{final} are set to 300K and 100K, respectively.

4 Experiments

4.1 Experimental Settings

We use the dataset provided by Lewis et al. (2019) as our initial UQA dataset for our methods, as well as for training the baseline model. The baseline model is implemented and evaluated in our settings, instead of using the results provided by Lewis et al. (2019) for a fair comparison with our model in the in-domain MRQA datasets (Fisch et al., 2019). We use 100K training instances for all the experiments except for the investigation on the use of different target languages for the paraphrasing method (Table 4), where we use 10K instances. For evaluation of the proposed approaches, we use SQuAD v1.1 (Rajpurkar et al., 2016) (Table 4 and 5) and the in-domain MRQA shared task datasets (Table 3). The latter dataset includes not only SQuAD but also many other QA datasets such as NewsQA (Trischler et al., 2017), TriviaQA (Joshi et al., 2017), SearchQA (Dunn et al., 2017), HotpotQA (Yang et al., 2018), and NaturalQuestions (Kwiatkowski et al., 2019).

The QA model used for most of the experiments is based on the BERT-large (Devlin et al., 2019) model with the only exception being the investigation on the use of different target languages for back-translation, for which we use BERT-base model (Table 4). Note that our proposed methods are model-agnostic, meaning that it can be applied to any QA models. For the implementation, we follow the settings in Devlin et al. (2019). We conduct the experiments five times under the same setting to obtain reliable results. For the paraphrasing method, we use Google Translate [4] which supports the largest number of languages for translation.

4.2 Experimental Results

In Table 3, we compare the proposed methods to the baseline of the original UQA questions (Lewis et al., 2019) across multiple QA tasks including SQuAD 1.1 and those in the "MRQA in-Domain dev sets" (Fisch et al., 2019). By applying the paraphrasing method (the second row), we obtain significant improvements across all the tasks by a margin ranging from 0.98% to 8.93% in F1, with an average of 4.51%. After applying the trimming method (third row), the performance improves further, ranging from 1.82% to 10.77% in F1, with an average of 5.35%. We obtain a larger improvement for more "difficult" tasks; HotpotQA (Yang et al., 2018) is a multi-hop task, and Natural Questions (Kwiatkowski et al., 2019) and TriviaQA (Joshi et al., 2017) questions sometimes require linking multiple sentences to find an answer.

Table 4 shows a comparison among different paraphrasing methods. While the performance changes are not huge, it proves that the proposed method of sampling from the inverse BLEU scores is effective, compared to using a particular language chosen for the lowest and highest BLEU scores. The overall trend is that creating more diversified questions results in performance improvement.

Finally, we show that the trimming method works as intended (Table 5). The questions are divided into three parts, bottom (0-300K), middle (300-1600K), and top (1600-1900K), according to the confidence

[4]We use Google Translate API from https://pypi.org/project/googletrans/

Paraphrasing Method	EM	F1
Lewis et al. (2019)	38.6	47.8
Randomly sampled language	42.28	52.71
Portuguese (Highest in BLEU)	41.55	51.82
Chinese (Lowest in BLEU)	42.50	53.02
Sampled with Inverse-BLEU	**43.14**	**53.61**

Table 4: Effect of different paraphrase methods on SQuAD 1.1 dev set with BERT-base QA.

Sampling	BLEU	EM	F1
Random (All)	7.76	47.34	57.10
Bottom part only	**7.19**	45.05	55.33
Top part only	**8.41**	36.77	44.86
All w/o Top	7.43	48.18	57.86
All w/o Bottom	8.05	48.58	57.73
Middle part only	7.68	**49.17**	**58.54**

Table 5: Performance with confidence score intervals, evaluated on the SQuAD 1.1 dev set.

scores in ascending order. Note that the division threshold is set empirically, just to see the effect of a reasonable division. The BLEU score is measured between a question and a sentence that contains an answer. The result implies that the questions in the top part are the closest to the answer sentences, giving a negative impact by a big margin, whereas those in the bottom are least similar, giving a small help when they are trimmed. We also show that the BLEU score negatively correlates with the confidence score, which re-validates our trimming approach. In addition, the BLEU score of "Random" (to which our paraphrasing method has been applied) is significantly reduced compared to the UQA BLEU score (23.35) of Section 2.1, which can also support our paraphrasing approach.

5 Conclusion

We have identified two anomalies of questions from Unsupervised Question Answering (UQA) (Lewis et al., 2019): excessive copying of the context and unanswerable questions. We proposed methods for the two: machine-translation-based paraphrasing of questions with the novel idea of using inverse BLEU scores and the confidence-based pruning that trims low-quality questions. We show that the proposed approach can improve UQA performance across diverse QA tasks, providing new insights on how to improve it.

Acknowledgements

This work was supported by Institute for Information & communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (No. 2013-0-00131, Development of Knowledge Evolutionary WiseQA Platform Technology for Human Knowledge Augmented Services). Also, the authors would like to thank Haritz Puerto San Roman for his insighful discussions and comments on the experiments and the draft.

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.

Pablo Duboue and Jennifer Chu-Carroll. 2006. Answering the question you wish they had asked: The impact of paraphrasing for question answering. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 33–36.

Matthew Dunn, Levent Sagun, Mike Higgins, V Ugur Guney, Volkan Cirik, and Kyunghyun Cho. 2017. Searchqa: A new q&a dataset augmented with context from a search engine. *arXiv preprint arXiv:1704.05179*.

Adam Fisch, Alon Talmor, Robin Jia, Minjoon Seo, Eunsol Choi, and Danqi Chen. 2019. MRQA 2019 shared task: Evaluating generalization in reading comprehension. In *Proceedings of 2nd Machine Reading for Reading Comprehension (MRQA) Workshop at EMNLP*.

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.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611.

Junmo Kang, Haritz Puerto San Roman, and Sung-Hyon Myaeng. 2019. Let me know what to ask: Interrogative-word-aware question generation. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 163–171, Hong Kong, China, November. Association for Computational Linguistics.

Junmo Kang, Giwon Hong, Haritz Puerto San Roman, and Sung-Hyon Myaeng. 2020. Regularization of distinct strategies for unsupervised question generation. In *Proceedings of the Findings of 2020 Conference on Empirical Methods in Natural Language Processing*. To appear.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Phrase-based & neural unsupervised machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5039–5049, Brussels, Belgium, October-November. Association for Computational Linguistics.

Patrick Lewis, Ludovic Denoyer, and Sebastian Riedel. 2019. Unsupervised question answering by cloze translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4896–4910, Florence, Italy, July. Association for Computational Linguistics.

Doyeon Lim, Haritz Puerto San Roman, and Sung-Hyon Myaeng. 2020. Analysis of the semantic answer types to understand the limitations of mrqa models. *Journal of KIISE : Software and Applications*, 47(3):298–309.

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 on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

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

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.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2017. Newsqa: A machine comprehension dataset. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 191–200.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380.

A Appendix

A.1 Pipeline Procedure

Algorithm 1 Overall Pipeline Procedure

Input: the original 2M dataset D_o (Lewis et al., 2019)
Output: refined QA train set T_{final} (100K)

1: $D_p \leftarrow paraphrasing(D_o)$ ▷ Apply our paraphrasing method to D_o
2: $T_p, D'_p \leftarrow split(D_p)$ ▷ Sample 100K training set T_p uniformly out of paraphraased dataset D_p and obtain the rest set $D'_p = D_p - T_p$
3: $\theta_{scorer} \leftarrow train(T_p)$ ▷ Train a QA model (scorer) with 100K train set T_p
4: $D_c \leftarrow scoring(\theta_{scorer}, D'_p)$ ▷ Inference confidence scores for D'_p (1,900K) using the trained QA model θ_{scorer}
5: $D_t \leftarrow trimming(D_c)$ ▷ Apply our trimming method to filter out anomalies and obtain filtered dataset D_t (1,300K)
6: $T_{final} \leftarrow sample(D_t)$ ▷ Uniformly sample train set T_{final} (100K) out of D_t

A.2 Number of Instances Sampled for Each Language

As explained in Section 3, we propose the probabilistic method of using inverse BLEU scores to choose a language in applying back-translation for each instance. For a total of 100K instances (which is the final QA dataset), the number of instances sampled for each language is shown in Table 6.

	pt	it	fr	es	de	hi	ru	ar	cn	Total
# of instances	2,135	2,135	3,027	3,055	2,688	14,542	5,121	7,698	59,599	100,000

Table 6: The number of instances sampled for each language

A.3 Case Study

We evaluated the quality of the questions by taking a random sample from the instances with an extreme confidence score. Table 7 shows four questions with the top and the bottom confidence scores for answer spans, according to the weakly trained QA model. The questions belonging to the top score region have many tokens copied from the context while those in the bottom score region contain irrelevant words that differ significantly from the meaning of the context. These cases indicate that trimming the instances in the top and bottom regions would result in a superior dataset. We hypothesize that a low confidence score is due to the noise of the questions whereas a high confidence score is due to a high word overlap with the context (copy-type), both of which are likely to hurt the QA performance.

Class	Context	Generated Question
Top Confidence Score Region	Los Angeles **shot poorly in the early going**, and the Celtics jumped out to a first-quarter 24–12 lead which was cut to 59–56 at halftime. ...	Where you **shoot poorly in the early going** ?
	... The semi-**operatic innovations of Henry Purcell did not lead to a native operatic tradition, but** George Frideric Handel found **important royal patrons and enthusiastic public support in England. ...**	Who knew that the half-**operatic innovations of Henry Purcell did not lead to a native operatic tradition, but** it was an **important royal patrons and enthusiastic public support in England** ?
Bottom Confidence Score Region	In the earliest days of the BSA, some commissioners were paid by local benefactors and supporters to administer and "grow Scouting" on a daily basis. ...	**When and how not to, "grow the search" you handle every day ?**
	The minor district (king amphoe) was created on 18 July 1977, when the four tambon Po, Dot, Siao and Nong Ma were split off from Uthumphon Phisai District. ...	**Who were the four manufacturers, Po, and Nong Ma PRH Phisqa District split ?**

Table 7: An example showing the context and the generated question with a high and a low confidence score, respectively. The answer span is underlined in the contexts.