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

Kazutoshi Shinoda
shinoda@is.s.u-tokyo.jp

Saku Sugawara
saku@nii.ac.jp

Akiko Aizawa
aizawa@nii.ac.jp

東京大学 THE UNIVERSITY OF TOKYO

大学共同利用機関法人 情報・システム研究機構 国立情報学研究所 National Institute of Informatics

Our dataset is released at:

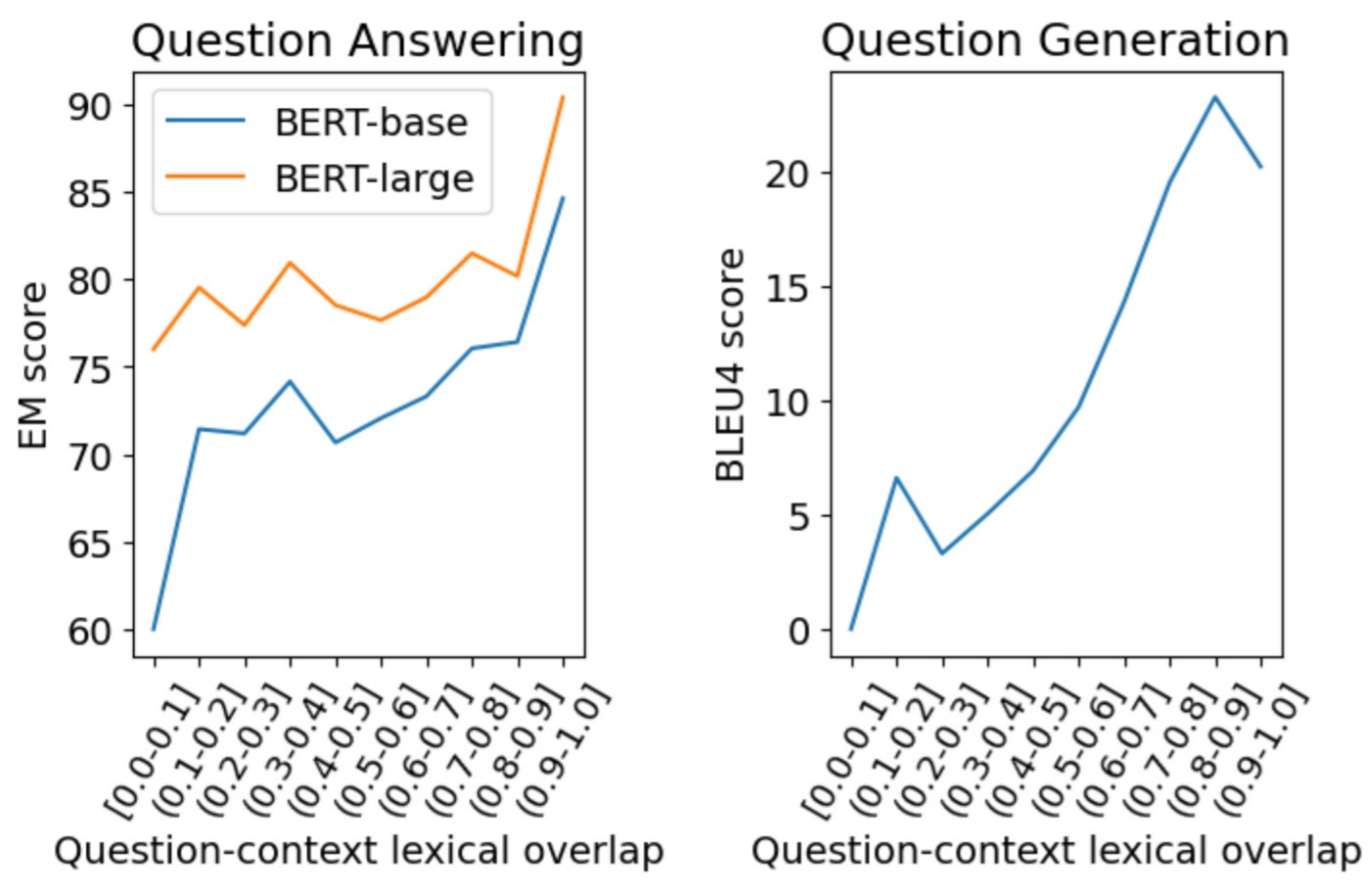
Summary

- Not only QA but also QG models are biased in terms of question–context lexical overlap; that is, **QG models tend to generate questions with high lexical overlap**.
- **Data augmentation using recent neural QG models does not debias QA models in terms of lexical overlap**; rather, it frequently degrade the QA performance on questions with lower overlap.
- **The proposed data augmentation method based on synonym replacement** for augmenting questions with low overlap **is simple yet effective to mitigate the degraded QA performance**.

QA & QG Models are Biased in Terms of Lexical Overlap.

In this study, question–context lexical overlap (QCLO) is defined as:

$$\mathrm{QCLO} = \frac{|\boldsymbol{Q} \cap \boldsymbol{C}|}{|\boldsymbol{Q}|}$$



The performance of both QA and QG models trained on SQuAD is degraded when QCLO is lower. The degraded QG performance on questions with low QCLO could exacerbate the degraded QA performance.

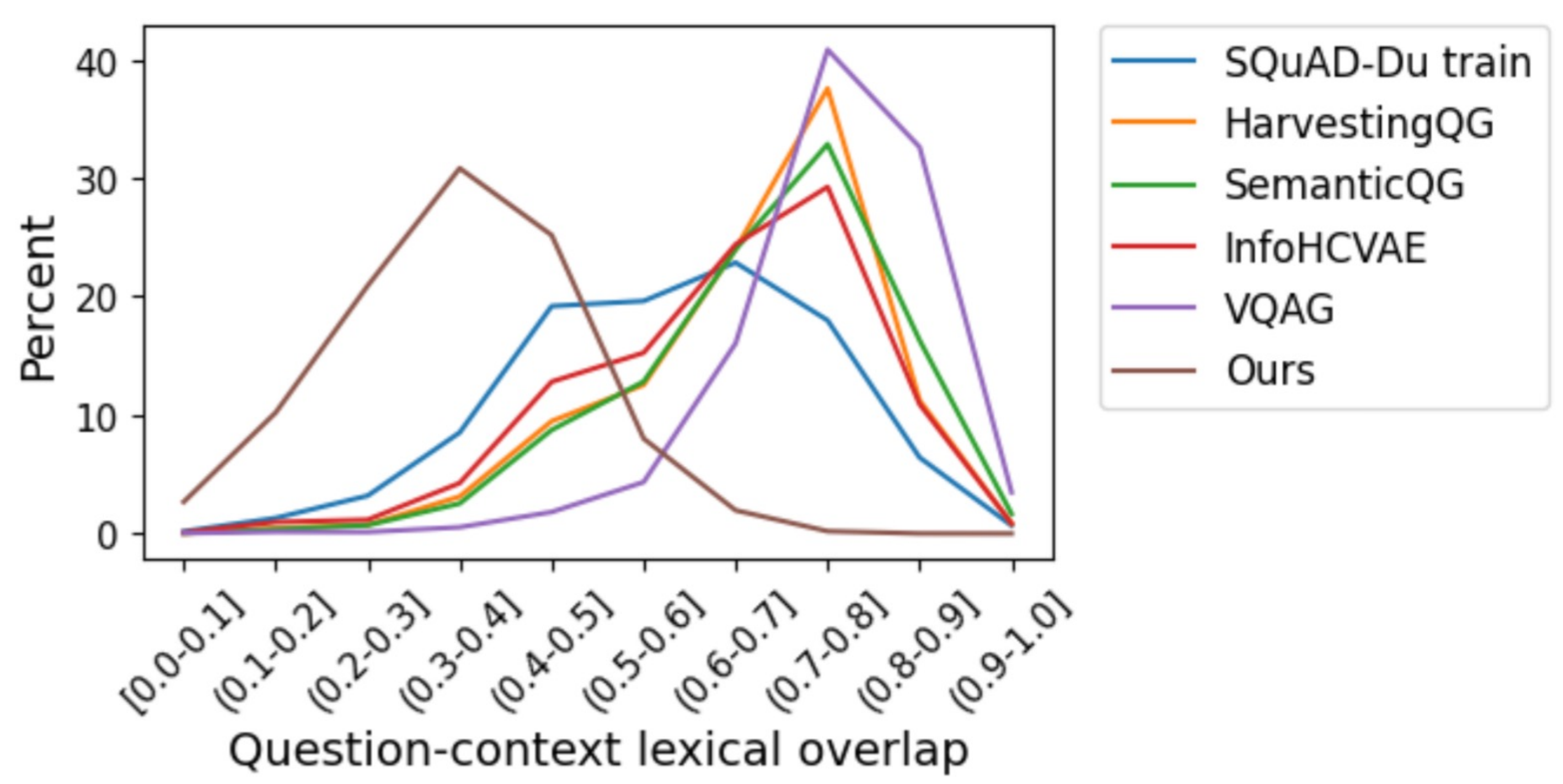
Synonym Replacement Can Reduce Lexical Overlap Bias.

We propose a simple method for specifically augmenting questions with low QCLO. The procedure is as follows:

(1) List all the overlapping words between question and context.

(2) Replace every word in the list with one of its synonyms from WordNet.

(3) If the QCLO decreases after (2), add the question to our dataset.



QCLO in the obtained dataset is substantially lower than SQuAD and other QG methods.

Experiment & Result:

Dataset:

- SQuAD-Du (train: 76k, dev: 11k, test: 12k)

Model: BERT-base & -large

Baseline QG methods:

HarvestngQG, SemanticQG, InfoHCVAE, VQAG

Evaluation dataset:

- Hard: subset of SQuAD-Du where QCLO ≤ 0.3
- Easy: subset of SQuAD-Du where QCLO > 0.3

		$\mathrm{SQuAD}^{\mathrm{Du}}_{\mathrm{test}}$ (EM/F1)		
Model	Train Source	Hard	Easy	ALL
base	$\mathrm{SQuAD}^{\mathrm{Du}}_{\mathrm{train}}$	70.88/81.99	73.22/84.75	73.06/84.57
	+ HarvestingQG	69.28/79.92	73.15/84.20	72.90/83.93
	+ SemanticQG	71.68/82.49	**74.39/85.59**	**74.21/85.39**
	+ InfoHCVAE	73.47/**83.91**	73.50/85.08	73.48/84.99
	+ VQAG	71.60/83.07	73.79/85.23	73.63/85.08
	+ Ours	**73.60**/83.49	73.08/84.41	73.11/84.34
large	$\mathrm{SQuAD}^{\mathrm{Du}}_{\mathrm{train}}$	77.93/87.84	**79.33/89.88**	**79.24/89.74**
	+ HarvestingQG	76.99/86.61	77.58/88.28	77.54/88.17
	+ SemanticQG	76.99/87.29	77.82/88.68	77.77/88.59
	+ InfoHCVAE	76.00/87.55	78.02/88.90	77.87/88.80
	+ VQAG	77.33/87.70	78.98/89.36	78.86/89.25
	+ Ours	**78.40/88.52**	77.94/89.00	77.96/88.97

While recent neural QG methods often degrade the scores on the Hard subset, ours consistently achieved the best EM score on the Hard subset.

Conclusion

Future research in QG for QA should exercise caution to prevent the amplification of dataset bias.