Which Shortcut Solution Do Question Answering Models Prefer to Learn?

Kazutoshi Shinoda, Saku Sugawara, Akiko Aizawa

@shino__c kazutoshi.shinoda0516@gmail.com

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

AAAI-23@Washington, DC, USA

Summary

- We aim to verify our hypothesis: studying the learnability of shortcut solutions (i.e., how easy it is to learn a shortcut) in QA datasets is useful to construct training sets or to design data augmentation methods to avoid learning shortcuts.
- The primary finding is that **the more learnable a shortcut is, the less proportion of anti-shortcut training examples is required to avoid learning the shortcut**. Moreover, **data balancing alone is insufficient to avoid learning less learnable shortcuts**.

Background: Shortcut Learning of QA Models

In extractive and multiple-choice QA, existing studies have found models can learn several types of shortcut solutions. Various methods have been proposed to mitigate each shortcut solution independently. *However, these methods have not fully taken the characteristics of shortcuts into account.*
→*Hypothesis in Summary*

Examined Shortcut Solutions

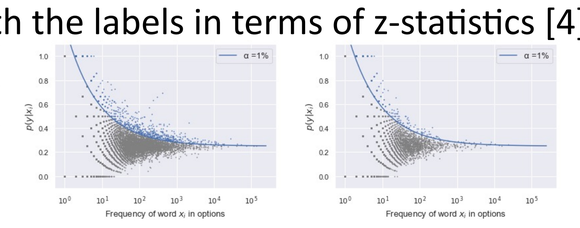
For extractive QA, we employed the following:

- Answer-Position [1]: Finding answers from the first sentence
- Word Matching [2]: Finding answers from the most similar sentence in context
- Type Matching [3]: Matching question and answer types

For multiple-choice QA, we adopted the two inspired by NLI.

- Word-label Correlation (Top-1): We identify the Top-1 word, which is the most highly correlated with the labels in terms of z-statistics [4].

RACE		ReClor	
w	z^*	w	z^*
and	23.6	a	6.7
above	20.7	result	5.3
may	20.7	an	5.1



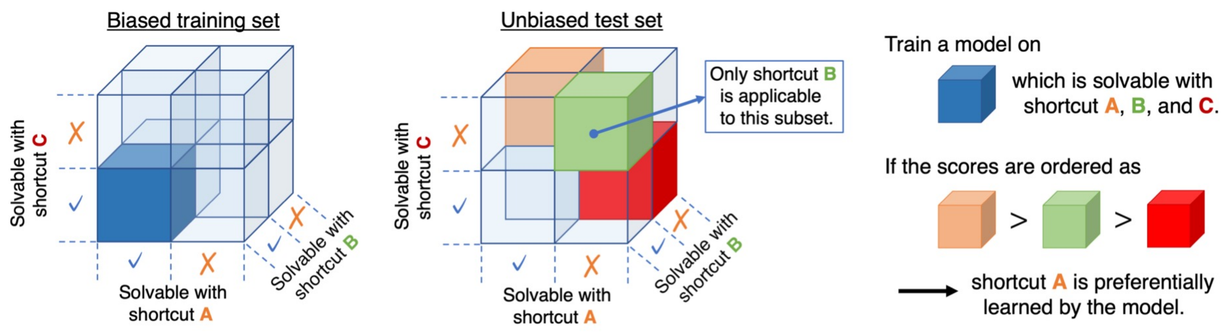
- Lexical Overlap: Choosing options that has the maximum lexical overlap with context+question.

For each shortcut solution k, we define a rule-based function to divide a dataset $\mathcal{D}$ into shortcut examples $\mathcal{D}_k$ and anti-shortcut examples $\overline{\mathcal{D}_k}$.

Experiments & Results

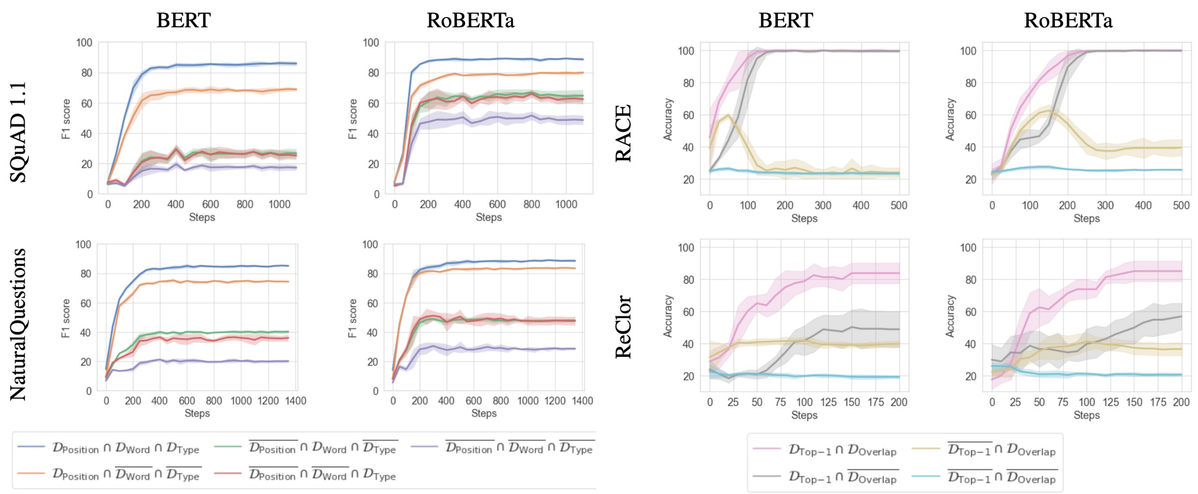
To verify our hypothesis, we first compare the learnabilities of the shortcuts with behavioral (**RQ1**), qualitative (**RQ2**), and quantitative (**RQ3**) analyses. Then, we study the requirements of anti-shortcut examples to avoid learning shortcuts (**RQ4**) and discuss the connection between the two.

RQ1: *When every shortcut is valid for answering every question in biased training sets, which shortcut do QA models prefer to learn?*



(Illustration of the behavioral analysis to answer RQ1.)

Results:

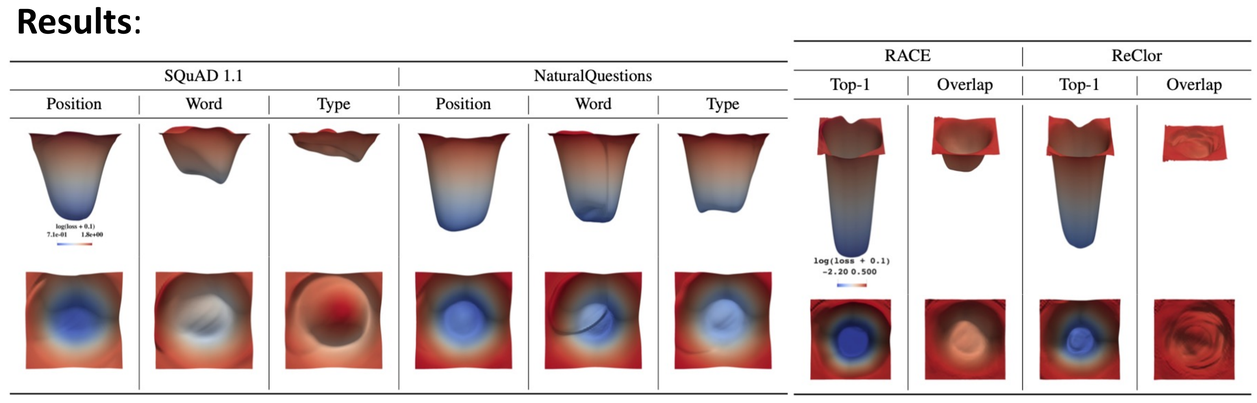


➡At the end of training,
Extractive QA: Position > Word ≒ Type
Multiple-choice QA: Top-1 > Overlap
In the early stage of training multiple-choice QA models, Overlap > Top-1, which may be due to the inductive bias of self attentions in transformers.

RQ2: *Why are certain shortcuts learned in preference to other shortcuts from the biased training sets?*

We train a model on  where only one shortcut is available and the others are not. We assume that the model learns the shortcut, which is verified experimentally. Then, we visualize the loss surface on the biased training set  . We repeat the same procedure for each shortcut.

Results:



➡ The preferred shortcuts (Position and Top-1) tend to lie in flatter and deeper loss surfaces in the parameter space.

RQ3: *How quantitatively different is the learnability for each shortcut?*

We propose **Rissanen Shortcut Analysis (RSA)** to quantitatively compare the learnabilities of shortcuts. In RSA, MDL on a biased dataset where only one of the shortcuts is available is approximated with the online code algorithm following [5].

➡ The availability of the preferred shortcuts (Position and Top-1) tends to make the task easier to learn.
➡ The orders of MDLs are roughly aligned with the previous experiments.

Results:

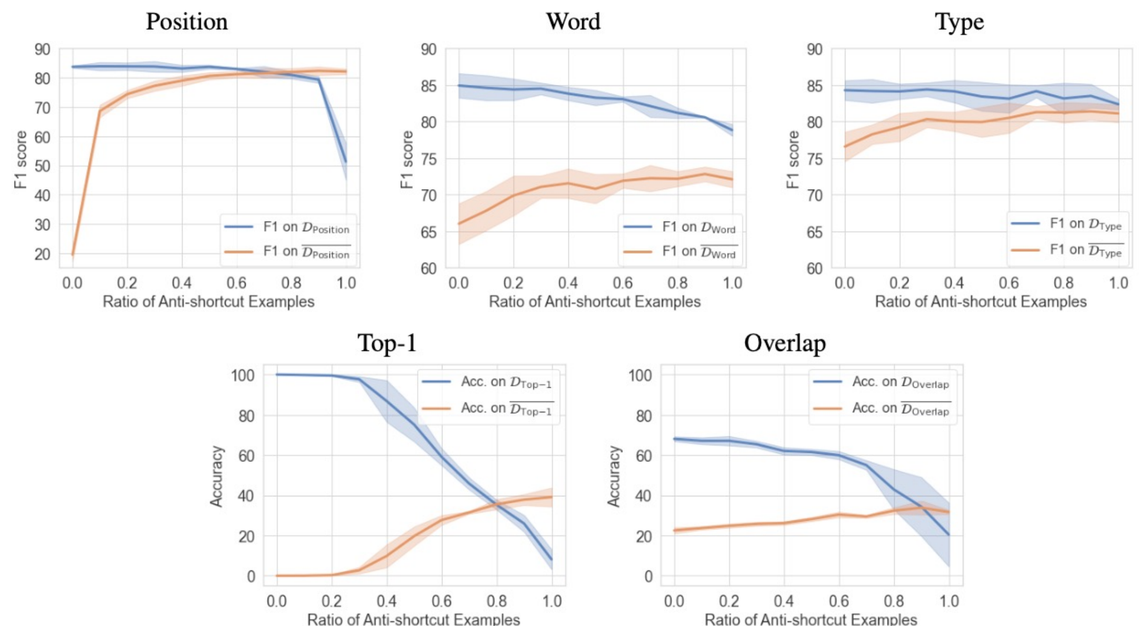
Shortcut	BERT	RoBERTa
SQuAD 1.1		
Position	4.65 ± 0.12	4.22 ± 0.23
Word	4.94 ± 0.24	3.73 ± 0.17
Type	5.75 ± 0.30	4.52 ± 0.06
NaturalQuestions		
Position	6.28 ± 0.15	5.37 ± 0.24
Word	12.24 ± 0.14	9.08 ± 0.20
Type	11.76 ± 0.55	8.83 ± 0.38
RACE		
Top-1	0.52 ± 0.34	0.41 ± 0.29
Overlap	4.16 ± 0.55	3.55 ± 0.10
ReClor		
Top-1	0.33 ± 0.07	0.28 ± 0.03
Overlap	0.55 ± 0.03	0.52 ± 0.02

(Minimum description lengths (kbits) on biased datasets such as or .)

RQ4: *What proportion of anti-shortcut examples in a training set is required to avoid learning a shortcut? Is it related to the learnability of shortcuts?*

We changed the proportion of anti-shortcut examples from 0 to 1 with the sizes of the training sets fixed. The scores on shortcut examples $\mathcal{D}_k$ and anti-shortcut examples $\overline{\mathcal{D}_k}$ are reported.

Results:



➡ The scores on $\mathcal{D}_k$ and $\overline{\mathcal{D}_k}$ are comparable when the proportion of anti-shortcut examples in training sets is 0.7, 0.8, and 0.9 for Position, Top-1, and Overlap shortcuts, resp.. Balancing $\mathcal{D}_k$ and $\overline{\mathcal{D}_k}$ in a training set could not mitigate the accuracy gap for Word and Type shortcuts completely.
➡ The requirements of the proportion of anti-shortcut examples are correlated with the learnabilies of the shortcuts studied in RQ1/2/3.

Discussion

- Our study suggests that **the learnability of shortcuts can be utilized to design new approaches for mitigating shortcut learning**.
- To avoid learning less learnable shortcuts, modifying loss functions or model architectures may be needed in addition to data balancing.

References

[1] Ko et al. 2020. Look at the First Sentence: Position Bias in Question Answering. In EMNLP.
[2] Sugawara et al. 2018. What Makes Reading Comprehension Questions Easier? In EMNLP.
[3] Weissenborn et al. 2017. Making Neural QA as Simple as Possible but not Simpler. In CoNLL.
[4] Gardner et al. 2021. Competency Problems: On Finding and Removing Artifacts in Language Data. In EMNLP.
[5] Perez et al. 2021. Rissanen Data Analysis: Examining Dataset Characteristics via Description Length. In ICML.

Links

Codes:

arXiv: