Which Shortcut Solution Do Question Answering Models Prefer to Learn?

AAAI-23 @ Washington, DC, USA

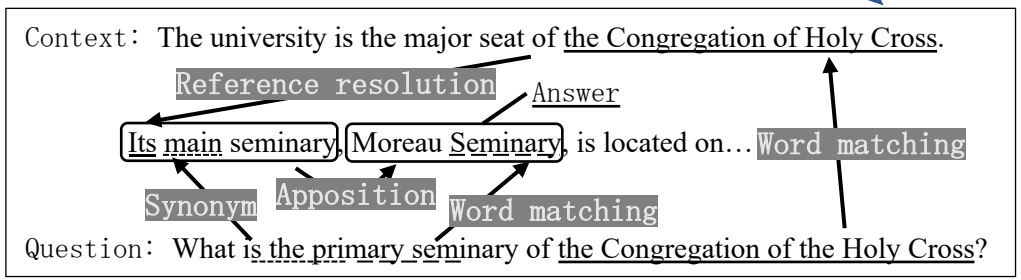
Kazutoshi Shinoda[1,2] Saku Sugawara[2] Akiko Aizawa[1,2]

[1] The University of Tokyo

[2] National Institute of Informatics

東京大学
THE UNIVERSITY OF TOKYO

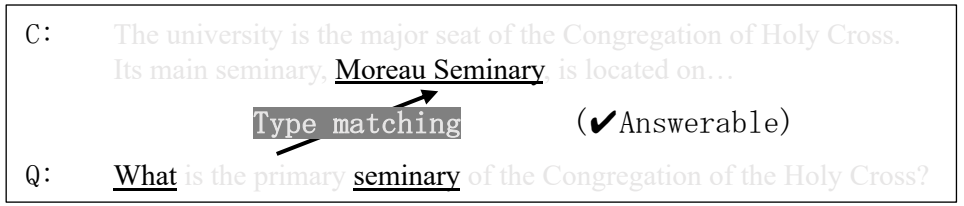
Background: Shortcut Learning of QA Models



Human-like intended solution

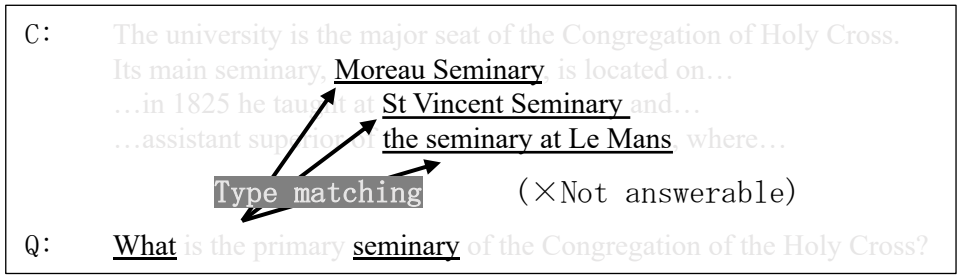
QA Model

Prefer to learn



Shortcut solution (Type matching)

✔Generalize



✖Do not generalize

An out-of-distribution example where a shortcut shortcut is not valid

Background: Shortcut Learning of QA Models

- In extractive and multiple-choice QA, existing studies found several types of shortcut solutions, which causes the lack of out-of-distribution generalization.

- Various methods (data augmentation, loss function, etc.) have been proposed to mitigate shortcut learning of QA models for each shortcut independently.

- **However, these mitigation methods have not fully taken the characteristics of shortcuts into account**.

Research Goal

- 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 new training sets or to design data augmentation methods **to avoid learning shortcuts**.

- 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 data balancing to avoid learning shortcuts (**RQ4**) and discuss the connection between learnability and data balancing.

Examined Shortcut Solutions

- Extractive QA:
  - Answer-**Position** (Ko et al. 2020)
    - Finding answers from the first sentences
  - **Word** Matching (Sugawara et al. 2018)
    - Finding answers from the most similar sentences in contexts
  - **Type** Matching (Weissenborn et al. 2017)
    - Matching types of questions and answers

Examined Shortcut Solutions

- Multiple-choice QA:
  - 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 (Gardner et al. 2021).

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

For the RACE dataset,

Top-1 shortcut = predicting that an option is correct if it contains “and”.

Examined Shortcut Solutions

- Multiple-choice QA:
  - 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 (Gardner et al. 2021).
  - 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** $\mathcal{D}_k$, where the shortcut is available, and **anti-shortcut examples** $\overline{\mathcal{D}_k}$, where it is not.

Research Questions

- RQ1: *When every shortcut is valid for answering every question in biased training sets, which shortcut do QA models prefer to learn?*
- RQ2: *Why are certain shortcuts learned in preference to other shortcuts from the biased training sets?*
- RQ3: *How quantitatively different is the learnability for each shortcut?*
- 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?*

Research Questions

- **RQ1**: ***When every shortcut is valid for answering every question in biased training sets, which shortcut do QA models prefer to learn?***
- RQ2: *Why are certain shortcuts learned in preference to other shortcuts from the biased training sets?*
- RQ3: *How quantitatively different is the learnability for each shortcut?*
- 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?*

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

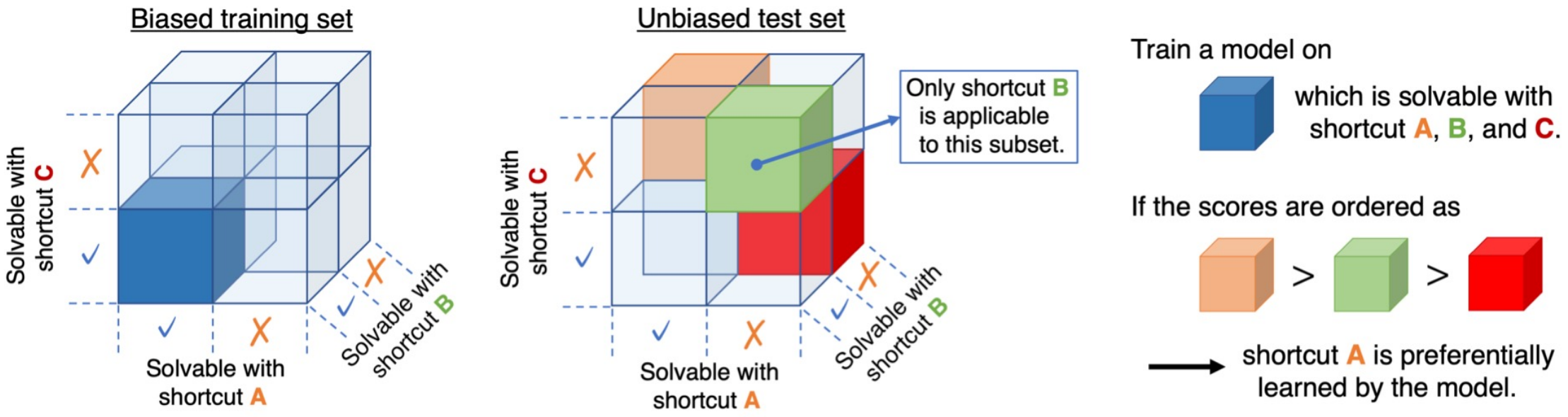
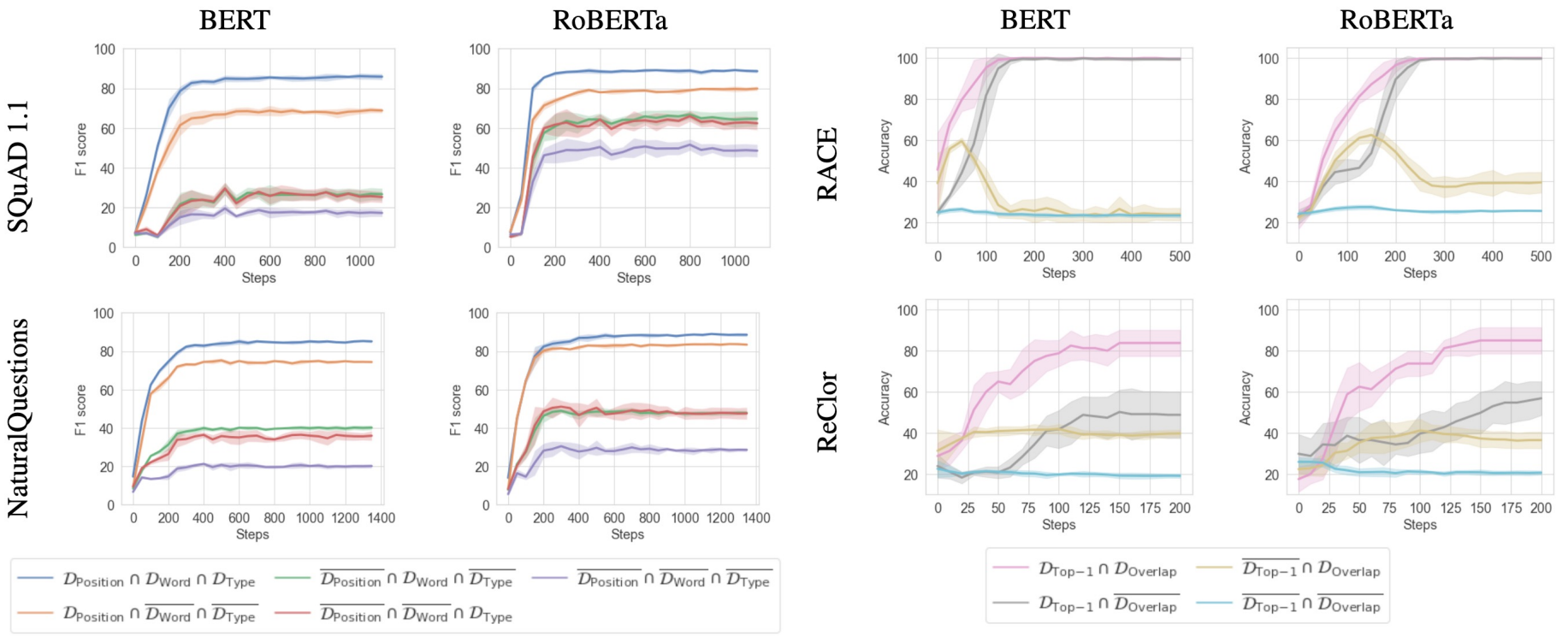


Illustration of the behavioral analysis to answer RQ1.

Results:

Extractive QA

Multiple-choice QA



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

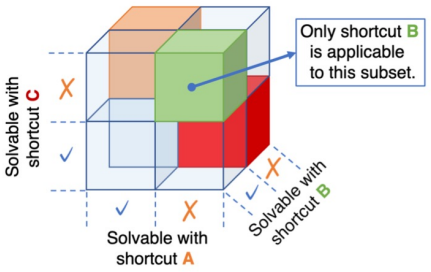
Research Questions

- RQ1: *When every shortcut is valid for answering every question in biased training sets, which shortcut do QA models prefer to learn?*
- **RQ2: *Why are certain shortcuts learned in preference to other shortcuts from the biased training sets?***
- RQ3: *How quantitatively different is the learnability for each shortcut?*
- 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?*

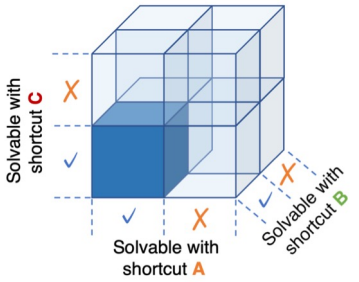
RQ2: *Why are certain shortcuts learned in preference to other shortcuts from the biased training sets?*

- (1) We train a model on a subset where only one shortcut is available and the others are not. We assume that the model learns the available shortcut, which is verified experimentally.
- (2) Then, we visualize the loss surface on the biased training set, which was used as the training set in RQ1. We repeat the same procedure for each shortcut.

(1) Train a model on one of the subsets:



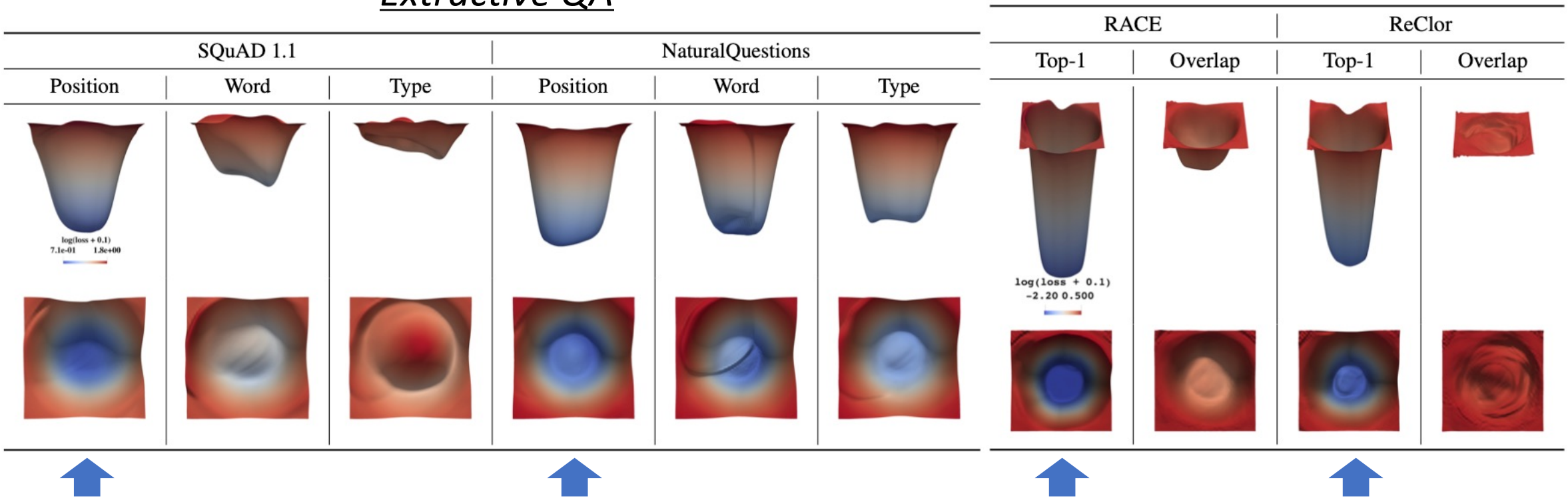
(2) Visualize the loss surface on the subset:



Results:

Extractive QA

Multiple-choice QA



- The preferred shortcuts (**Position** and **Top-1**) tend to lie in flatter and deeper loss surfaces in the parameter space.
- The orders of the flatness and depth of the loss surfaces are roughly correlated with the preferential order of learning shortcuts in RQ1.

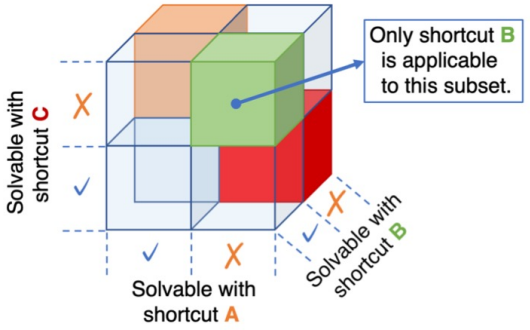
Research Questions

- RQ1: *When every shortcut is valid for answering every question in biased training sets, which shortcut do QA models prefer to learn?*
- RQ2: *Why are certain shortcuts learned in preference to other shortcuts from the biased training sets?*
- **RQ3: *How quantitatively different is the learnability for each shortcut?***
- 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?*

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

- We propose **Rissanen Shortcut Analysis (RSA)** to quantitatively compare the learnabilities of shortcuts. RSA approximates the minimum description length (MDL) on a biased dataset where only one of the shortcuts is available. MDLs are approximated with the online code algorithm following Perez et al. (2021).
- RSA can measure how easy it is to learn the task when one shortcut is available and the others are not.

RSA approximates the MDL of one of the subsets:



Results:

- 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.

	Shortcut	BERT	RoBERTa
Extractive QA	*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
Multiple-choice QA	*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 where only one shortcut is available.

Research Questions

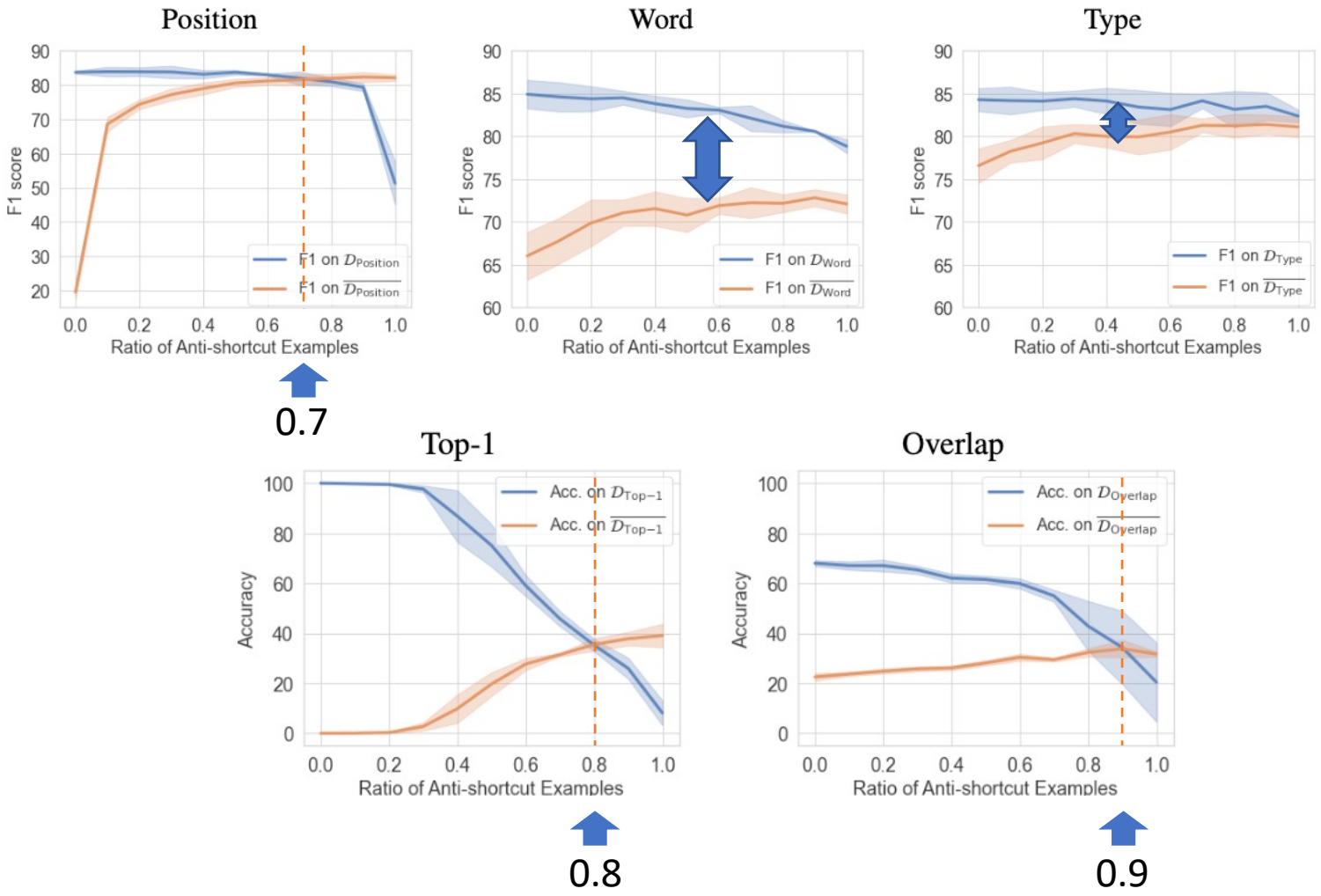
- RQ1: *When every shortcut is valid for answering every question in biased training sets, which shortcut do QA models prefer to learn?*
- RQ2: *Why are certain shortcuts learned in preference to other shortcuts from the biased training sets?*
- RQ3: *How quantitatively different is the learnability for each shortcut?*
- **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?***

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.

Conclusion & Discussion

- We experimentally showed that the learnabilities of shortcut solutions and the requirements of data balancing are roughly correlated.
- Our study suggests that **the learnability of shortcuts can be utilized to design new approaches for mitigating shortcut learning**.
  - For example, to avoid learning less learnable shortcuts, modifying loss functions or model architectures may be needed in addition to data balancing.
  - Similarly, to avoid learning more learnable shortcuts, only data balancing may be sufficient.

References

Ko et al. 2020. Look at the First Sentence: Position Bias in Question Answering. In EMNLP.

Sugawara et al. 2018. What Makes Reading Comprehension Questions Easier? In EMNLP.

Weissenborn et al. 2017. Making Neural QA as Simple as Possible but not Simpler. In CoNLL.

Gardner et al. 2021. Competency Problems: On Finding and Removing Artifacts in Language Data. In EMNLP.

Perez et al. 2021. Rissanen Data Analysis: Examining Dataset Characteristics via Description Length. In ICML.

Thank you!

Codes

arXiv

If you have any questions, please contact me via

@shino__c

kazutoshi.shinoda0516@gmail.com