Look to the Right: Mitigating Relative Position Bias in Extractive Question Answering

Kazutoshi Shinoda, Saku Sugawara, Akiko Aizawa

@shino__c

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

BlackboxNLP at EMNLP2022, Abu Dhabi, UAE

Contributions

- We discover that an extractive QA model can exploit relative positional cues as shortcuts, which deteriorates the generalization to relative positions that are unseen during training.
- We devise a new biased model that works with an existing debiasing algorithm, thereby improving the robustness to unseen relative positions.

Background: Shortcut Cues in extractive QA

QA models have been shown to exploit lexical overlap between question and context [1], types of Q&A [2], and absolute answer positions [3], resulting in the degraded generalization to anti-biased examples. Understanding what shortcut cues QA models can exploit would be beneficial for building generalizable QA models in real-world applications.

Relative Position Bias in Extractive QA

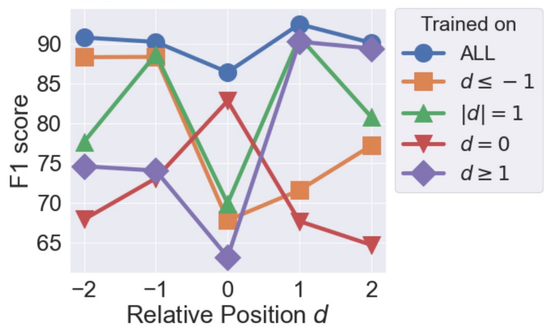
In this study, the relative position d of an answer is defined as the relative distance from the answer to the closest overlapping word.

Look to the right!

Context	... This changed in **1924** with formal requirements developed for graduate degrees, including offering Doctorate (PhD) degrees ...
Question	The granting of Doctorate degrees first occurred in what year at Notre Dame?
Relative Position	-1

(**Bold**: answer, ____: overlapping word)

Right Figure: Performance of BERT-base trained on full or subset of the SQuAD training set. We find that a QA model lacks robustness to unseen relative positions when the relative positions in the training set are biased.

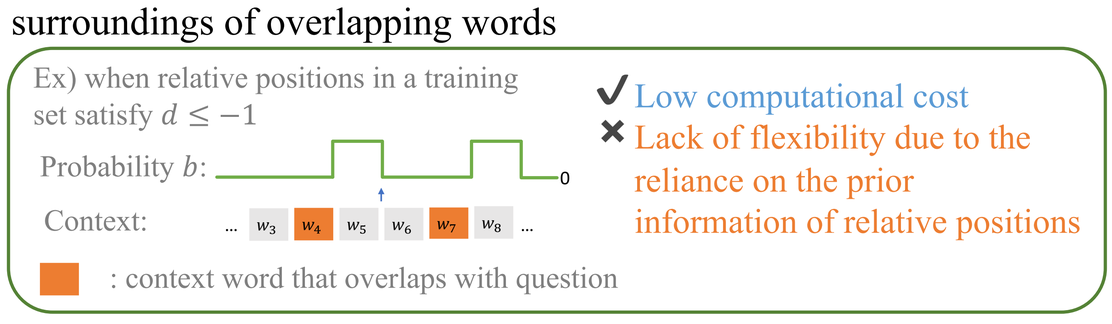


Method

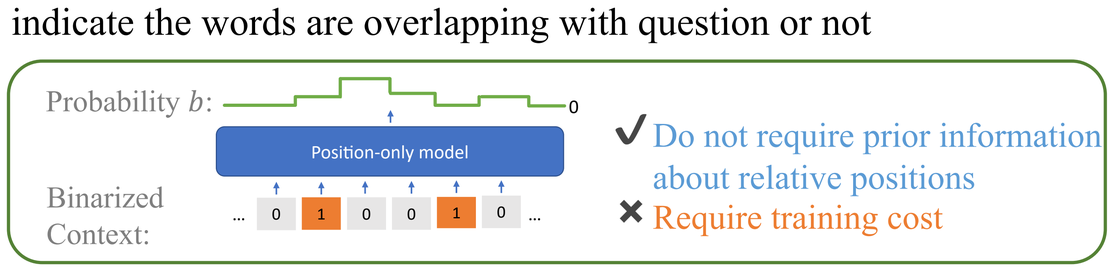
We compared two biased models and two debiasing algorithms for improving the robustness to unseen relative positions.

(1) **Training a biased model b**

- Answer Prior (AnsPrior): Heuristic that assigns probabilities to the surroundings of overlapping words



- Position-only model (PosOnly): Training BERT-base on binarized inputs where the context words are replaced with binary tokens that indicate the words are overlapping with question or not



(2) **Training a main model p by minimizing XE loss computed with $\hat{p}$**

- BiasProduct [4]
  $\hat{p} = \mathrm{softmax}(\log p + \log b)$ ← b is frozen during training
- LearnedMixin [4]
  $\hat{p} = \mathrm{softmax}(\log p + g(c,q) \log b)$
  where $g(c,q)$ is a learnable function that takes context c and question q as inputs and outputs a logit (≥ 0)

(3) **Using only the main model p for prediction at test time**

➡The training algorithms in (2) are expected to encourage the main model to learn solutions that do not rely solely on relative positions that the biased model uses. (Please refer to [4] for mathematical explanations.)

Experiments & Results

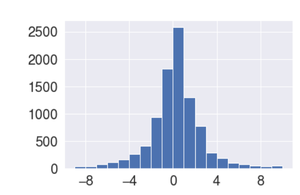
I. Generalization to Unseen Relative Positions

Training set: a subset of the SQuAD training set where d satisfy one of the predefined conditions, Evaluation set: the SQuAD development set for each range of d, Model: BERT-base.

Trained on	Model	Evaluated on $d \leq -3$	$d = -2$	$d = -1$	$d = 0$	$d = 1$	$d = 2$	$d \geq 3$
ALL	BERT-base	82.19	90.82	90.25	86.47	92.49	90.14	81.43
$d \leq -1$	BERT-base	78.17	88.34	88.38	67.82	71.62	77.22	69.54
$d \leq -1$	BiasProduct-AnsPrior	73.00	84.34	85.61	46.32	25.23	64.91	59.06
$d \leq -1$	LearnedMixin-AnsPrior	79.07	89.27	89.01	68.52	72.35	80.43	70.31
$d \leq -1$	BiasProduct-PosOnly	75.04	83.90	83.22	73.80	81.35	81.79	73.27
$d \leq -1$	LearnedMixin-PosOnly	77.00	86.72	86.25	74.26	82.66	82.81	75.94
$\|d\| = 1$	BERT-base	65.62	77.69	88.70	69.96	90.88	80.84	66.42
$\|d\| = 1$	BiasProduct-AnsPrior	60.44	75.07	56.44	49.32	52.37	72.85	57.98
$\|d\| = 1$	LearnedMixin-AnsPrior	73.42	83.39	88.70	74.24	90.47	85.51	73.52
$\|d\| = 1$	BiasProduct-PosOnly	72.41	80.59	84.01	73.34	87.61	83.11	72.09
$\|d\| = 1$	LearnedMixin-PosOnly	73.76	80.63	86.10	74.50	89.64	82.98	72.04
$d = 0$	BERT-base	60.75	67.94	73.11	82.85	67.72	64.74	52.88
$d = 0$	BiasProduct-AnsPrior	56.25	65.15	69.05	81.07	65.10	62.95	49.43
$d = 0$	LearnedMixin-AnsPrior	59.66	69.62	72.53	83.06	68.04	66.03	53.29
$d = 0$	BiasProduct-PosOnly	62.97	67.88	70.22	78.66	66.69	69.12	59.88
$d = 0$	LearnedMixin-PosOnly	65.09	70.47	72.51	81.32	68.29	68.47	59.54
$d \geq 1$	BERT-base	68.03	74.63	74.08	63.21	90.28	89.44	75.42
$d \geq 1$	BiasProduct-AnsPrior	58.63	63.13	29.08	39.22	88.53	88.34	72.29
$d \geq 1$	LearnedMixin-AnsPrior	70.71	77.22	76.82	66.67	90.87	89.75	76.31
$d \geq 1$	BiasProduct-PosOnly	68.54	78.13	78.58	70.72	85.17	81.59	72.90
$d \geq 1$	LearnedMixin-PosOnly	71.17	80.41	79.97	71.33	87.53	84.33	74.24

Gray cells: seen relative positions, white cells: unseen relative positions

Results:



- When using the full training set, the accuracy is lower when $|d|$ is larger. This may be due to the skewed distribution of d in the training set (➡).
- The BERT-base baseline maintains scores for seen relative positions compared to "ALL" but degrade the scores for unseen relative positions.
- LearnedMixin achieved higher scores than BiasProduct in most cases. Learning the degree to which the predictions of a biased model is effective.
- LearnedMixin-PosOnly achieved the best scores in white cells, but LearnedMixin-AnsPrior is better in gray cells. = *Trade-off*

II. Effect of Mitigating Relative Position Bias in Normal Settings

Training set: the full SQuAD training set, Evaluation set: two subsets of the SQuAD development set. One is $c \cap q \neq \phi$ where c and q have common words, and the other is $c \cap q = \phi$ where c and q have no words in common.

Trained on	Model	Evaluated on $c \cap q \neq \phi$	$c \cap q = \phi$
ALL	BERT-base	87.94	67.11
ALL	BiasProduct-PosOnly	84.83	59.88
ALL	LearnedMixin-PosOnly	87.37	80.44

Results: LeranedMixin-PosOnly increased 13 points on the subset $c \cap q = \phi$ while slightly degrading the score on the subset $c \cap q \neq \phi$. Our method may mitigate the reliance on overlapping words in normal settings.

Conclusion

- We showed that relative positions of answers can be shortcut cues for the extractive QA model, causing the performance degradation for unseen relative positions.
- The proposed LeranedMixin-PosOnly improved the robustness to unseen relative positions even when trained on intentionally filtered training sets. Moreover, when applied to the full training set, it improved the scores on examples without lexical overlap between question and context.
- Mitigating the trade-off between scores for seen and unseen relative positions is future work.

Citation

[1] Jia and Liang, EMNLP 2017
[2] Lewis and Fan, ICLR 2019
[3] Ko et al., EMNLP 2020
[4] Clark et al., EMNLP 2019