Robustifying Multi-hop Question Answering through Pseudo-Evidentiality Training

Kyungjae Lee[1] **Seung-won Hwang**[2*] **Sang-eun Han**[1] **Dohyeon Lee**[1]
[1]Yonsei University [2]Seoul National University

Abstract

This paper studies the bias problem of multi-hop question answering models, of answering correctly without correct reasoning. One way to robustify these models is by supervising to not only answer right, but also with right reasoning chains. An existing direction is to annotate reasoning chains to train models, requiring expensive additional annotations. In contrast, we propose a new approach to learn evidentiality, deciding whether the answer prediction is supported by correct evidences, without such annotations. Instead, we compare counterfactual changes in answer confidence with and without evidence sentences, to generate "pseudo-evidentiality" annotations. We validate our proposed model on an original set and challenge set in HotpotQA, showing that our method is accurate and robust in multi-hop reasoning.

1 Introduction

Multi-hop Question Answering (QA) is a task of answering complex questions by connecting information from several texts. Since the information is spread over multiple facts, this task requires to capture multiple relevant facts (which we refer as evidences) and infer an answer based on all these evidences.

However, previous works (Min et al., 2019; Chen and Durrett, 2019; Trivedi et al., 2020) observe "disconnected reasoning" in some correct answers. It happens when models can exploit specific types of artifacts (*e.g.*, entity type), to leverage them as **reasoning shortcuts** to guess the correct answer. For example, assume that a given question is: "which country got independence when World War II ended?" and a passage is: "Korea got independence in 1945". Although information ("World War II ended in 1945") is insufficient, QA models predict "Korea", simply because its answer type is country (or, using shortcut).

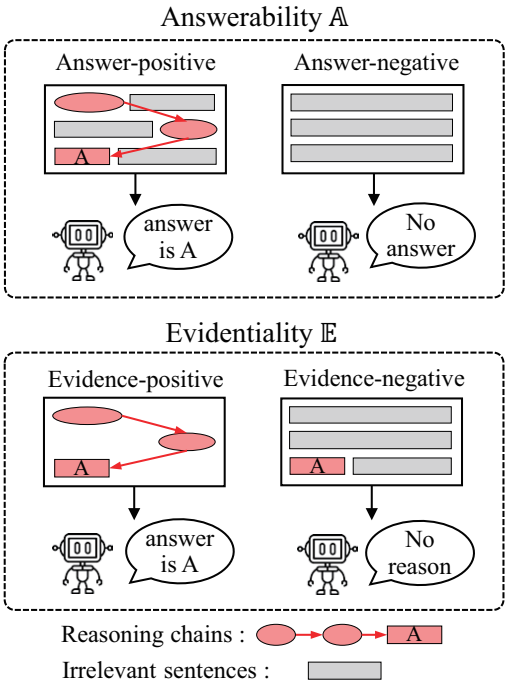


Figure 1: Overview of our proposed supervision: using Answerability and Evidentiality

To address the problem of reasoning shortcuts, we propose to supervise "**evidentiality**" – deciding whether a model answer is supported by correct evidences (see Figure 1). This is related to the problem that most of the early reader models for QA failed to predict whether questions are not answerable. Lack of answerability training led models to provide a wrong answer with high confidence, when they had to answer "unanswerable". Similarly, we aim to train for models to recognize whether their answer is "unsupported" by evidences, as well. In our work, along with the answerability, we train the QA model to identify the existence of evidences by using passages of two types: (1) **Evidence-positive** and (2) **Evidence-negative** set. While the former

*correspond to seungwonh@snu.ac.kr

Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, pages 6110–6119
August 1–6, 2021.

has both answer and evidence, the latter does not have evidence supporting the answer, such that we can detect models taking shortcuts.

Our first research question is: how do we **acquire** evidence-positive and negative examples for training without annotations? For evidence-positive set, the closest existing approach (Niu et al., 2020) is to consider attention scores, which can be considered as pseudo-annotation for evidence-positive set. In other word, sentence S with high attention scores, often used as an "interpretation" of whether S is causal for model prediction, can be selected to build evidence-positive set. However, follow-up works (Serrano and Smith, 2019; Jain and Wallace, 2019) argued that attention is limited as an explanation, because causality cannot be measured, without observing model behaviors in a counterfactual case of the same passage without S. In addition, sentence causality should be aggregated to measure group causality of multiple evidences for multi-hop reasoning. To annotate group causality as "pseudo-evidentiality", we propose *Interpreter* module, which removes and aggregates evidences into a group, to compare predictions in observational and counterfactual cases.

As a second research question, we ask how to **learn** from evidence-positive and evidence-negative set. To this end, we identify two objectives: (O1) QA model should not be overconfident in evidence-negative set, while (O2) confident in evidence-positive. A naive approach to pursue the former is to lower the model confidence on evidence-negative set via regularization. However, such regularization can cause violating (O2) due to correlation between confidence distributions for evidence-positive and negative set. Our solution is to selectively regularize, by purposedly training a biased model violating (O1), and decorrelate the target model from the biased model.

For experiments, we demonstrate the impact of our approach on HotpotQA dataset. Our empirical results show that our model can improve QA performance through pseudo-evidentiality, outperforming other baselines. In addition, our proposed approach can orthogonally combine with another SOTA model for additional performance gains.

2 Related Work

Since multi-hop reasoning tasks, such as HotpotQA, are released, many approaches for the task have been proposed. These approaches can be categorized by strategies used, such as graph-based networks (Qiu et al., 2019; Fang et al., 2020), external knowledge retrieval (Asai et al., 2019), and supporting fact selection (Nie et al., 2019; Groeneveld et al., 2020).

Our focus is to identify and alleviate reasoning shortcuts in multi-hop QA, without evidence annotations. Models taking shortcuts were widely observed from various tasks, such as object detection (Singh et al., 2020), NLI (Tu et al., 2020), and also for our target task of multi-hop QA (Min et al., 2019; Chen and Durrett, 2019; Trivedi et al., 2020), where models learn simple heuristic rules, answering correctly but without proper reasoning.

To mitigate the effect of shortcuts, adversarial examples (Jiang and Bansal, 2019) can be generated, or alternatively, models can be robustifed (Trivedi et al., 2020) with additional supervision for paragraph-level "sufficiency" – to identify whether a pair of two paragraphs are sufficient for right reasoning or not, which reduces shortcuts on a single paragraph. While the binary classification for paragraph-sufficiency is relatively easy (96.7 F1 in Trivedi et al. (2020)), our target of capturing a finer-grained sentence-evidentiality is more challenging. Existing QA model (Nie et al., 2019; Groeneveld et al., 2020) treats this as a supervised task, based on sentence-level human annotation. In contrast, ours requires no annotation and focuses on avoiding reasoning shortcuts using evidentiality, which was not the purpose of evidence selection in the existing model.

3 Proposed Approach

In this section, to prevent reasoning shortcuts, we introduce a new approach for data acquiring and learning. We describe this task (Section 3.1) and address two research questions, of generating labels for supervision (Section 3.2) and learning (Section 3.3), respectively.

3.1 Task Description

Our task definition follows *distractor* setting, between *distractor* and *full-wiki* in HotpotQA dataset (Yang et al., 2018), which consists of 112k questions requiring the understanding of corresponding passages to answer correctly. Each question has a candidate set of 10 paragraphs (of which two are positive paragraphs $\mathcal{P}^+$ and eight are negative $\mathcal{P}^-$), where the supporting facts for reasoning are scattered in two positive paragraphs. Then,

given a question $\mathcal{Q}$, the objective of this task is to aggregate relevant facts from the candidate set and estimate a consecutive answer span $\mathcal{A}$. For task evaluation, the estimated answer span is compared with the ground truth answer span in terms of F1 score at word-level.

3.2 Generating Examples for Training Answerability and Evidentiality

Answerability for Multi-hop Reasoning

For answerability training in single-hop QA, datasets such as SQuAD 2.0 (Rajpurkar et al., 2018) provide labels of answerability, so that models can be trained not to be overconfident on unanswerable text.

Similarly, we build triples of question $\mathcal{Q}$, answer $\mathcal{A}$, and passage $\mathcal{D}$, to be labeled for answerability. HotpotQA dataset pairs $\mathcal{Q}$ with 10 paragraphs, where evidences can be scattered to two paragraphs. Based on such characteristic, concatenating two positive paragraphs is guaranteed to be answerable/evidential and concatenating two negative paragraphs (with neither evidence nor answer) is guaranteed to be unanswerable. We define a set of answerable triplets $(\mathcal{Q}, \mathcal{A}, \mathcal{D})$ as **answer-positive** set $\mathbb{A}^+$, and an unanswerable set as **answer-negative** set $\mathbb{A}^-$. From the labels, we train a transformer-based model to classify the answerability (the detail will be discussed in the next section).

However, answerability cannot supervise whether the given passage has all of these relevant evidences for reasoning. This causes a lack of generalization ability, especially on examples with an answer but no evidence.

Evidentiality for Multi-hop Reasoning

While learning the answerability, we aim to capture the existence of reasoning chains in the given passage. To supervise the existence of evidences, we construct examples: **evidence-positive** and **evidence-negative** set, as shown in Figure 1.

Specifically, let E_* be the ground truth of evidences to infer $\mathcal{A}$, and $\mathcal{S}_*$ be a sentence containing an answer $\mathcal{A}$, corresponding to $\mathcal{Q}$. Given $\mathcal{Q}$ and $\mathcal{A}$, expected labels $\mathcal{V}_E$ of evidentiality, indicating whether the evidences for answering are sufficient in the passage, are as follow:

$$\begin{aligned} \mathcal{V}_E(\mathcal{Q}, \mathcal{A}, \mathcal{D}) \models True \quad &\Leftrightarrow \quad E_* = \mathcal{D},\ \mathcal{A} \subset \mathcal{D} \\ \mathcal{V}_E(\mathcal{Q}, \mathcal{A}, \mathcal{D}) \models False \quad &\Leftrightarrow \quad E_* \not\subset \mathcal{D},\ \mathcal{A} \subset \mathcal{D} \end{aligned} \tag{1}$$

We define a set of passages satisfying $\mathcal{V}_E \models True$ as **evidence-positive** set $\mathbb{E}^+$, and a set satisfying $\mathcal{V}_E \models False$ as **evidence-negative** set $\mathbb{E}^-$.

Since we do not use human-annotations, we aim to generate "pseudo-evidentiality" annotation. First, for **evidence-negative** set, we modify answer sentence $\mathcal{S}_*$ and unanswerable passages, and generate examples with the three following types:

- 1) Answer Sentence Only: we remove all sentences in answerable passage except $\mathcal{S}_*$, such that the input passage $\mathcal{D}$ becomes $\mathcal{S}_*$, which contains a correct answer but no other evidences. That is, $\mathcal{V}_E(\mathcal{Q}, \mathcal{A}, \mathcal{S}^*) \models False$.

- 2) Answer Sentence + Irrelevant Facts: we use irrelevant facts with answers as context, by concatenating $\mathcal{S}_*$ and unanswerable $\mathcal{D}$. That is, $\mathcal{V}_E(\mathcal{Q}, \mathcal{A}, (\mathcal{S}^*; \mathcal{D})) \models False$, where $\mathcal{D} \in \mathcal{P}^-$.

- 3) Partial Evidence + Irrelevant Facts: we use partially-relevant and irrelevant facts as context, by concatenating $\mathcal{D}_1 \in \mathcal{P}^+$ and $\mathcal{D}_2 \in \mathcal{P}^-$. That is, $\mathcal{V}_E(\mathcal{Q}, \mathcal{A}, (\mathcal{D}_1; \mathcal{D}_2)) \models False$.

These **evidence-negative** examples do not have all relevant evidences, thus if a model predicts the correct answer on such examples, it means that the model learned reasoning shortcuts.

Second, building an **evidence-positive** set is more challenging, because it is difficult to capture multiple relevant facts, with neither annotations E_* nor supervision. Our distinction is obtaining the above annotation from model itself, by interpreting the internal mechanism of models. On a trained model, we aim to find influential sentences in predicting correct answer $\mathcal{A}$, among sentences in an answerable passage. Then, we consider them as a pseudo evidence-positive set. Since such pseudo labels relies on the trained model which is not perfect, 100% recall of $\mathcal{V}_E(\mathcal{Q}, \mathcal{A}, \mathcal{D}) \models True$ in Eq. (1) is not guaranteed, though we observe 87% empirical recall (Table 1).

Section 1 discusses how interpretation, such as attention scores (Niu et al., 2020), can be pseudo-evidentiality. For QA tasks, an existing approach (Perez et al., 2019) uses answer confidence for finding pseudo-evidences, as we discuss below:

(A) Accumulative interpreter: to consider multiple sentences as evidences, the existing approach (Perez et al., 2019) iteratively inserts sentence $\mathcal{S}_i$ into set E^{t-1}, with a highest probability at t-th iter-

ation, as follows:

$$\Delta P_{\mathcal{S}_i} = P(\mathcal{A}|\mathcal{Q}, \mathcal{S}_i \cup E^{t-1}) - P(\mathcal{A}|\mathcal{Q}, E^{t-1})$$
$$\hat{E}^t = \underset{\mathcal{S}_i}{\text{argmax}}\, \Delta P_{\mathcal{S}_i}, \quad E^t = \hat{E}^t \cup E^{t-1} \tag{2}$$

where E^0 starts with the sentence $\mathcal{S}_*$ containing answer $\mathcal{A}$, which is minimal context for our task. This method can consider multiple sentences as evidence by inserting iteratively into a set, but cannot consider the effect of **erasing** sentences from reasoning chain.

(B) Our proposed *Interpreter*: to enhance the interpretability, we consider both **erasing** and **inserting** each sentence, in contrast to accumulative interpreter considering only the latter. Intuitively, erasing evidence would change the prediction significantly, if such evidence is causally salient, which we compute as follows:

$$\Delta P_{\mathcal{S}_i} = P(\mathcal{A}|\mathcal{Q}, \mathcal{D}) - P(\mathcal{A}|\mathcal{Q}, (\mathcal{D} \backslash \mathcal{S}_i)) \tag{3}$$

where $(\mathcal{D} \backslash \mathcal{S}_i)$ is a passage out of sentence $\mathcal{S}_i$. We hypothesize that breaking reasoning chain, by erasing $\mathcal{S}_i$, should significantly decrease $P(A|\cdot)$. In other words, $\mathcal{S}_i$ with higher $\Delta P_{\mathcal{S}_i}$ is salient. Combining the two saliency scores in Eq. (2),(3), our final saliency is as follows:

$$\Delta P_{\mathcal{S}_i} = P(\mathcal{A}|\mathcal{Q}, \mathcal{S}_i \cup E^{t-1}) - \cancel{P(\mathcal{A}|\mathcal{Q}, E^{t-1})} + \cancel{P(\mathcal{A}|\mathcal{Q}, \mathcal{D})} - P(\mathcal{A}|\mathcal{Q}, (\mathcal{D} \backslash (\mathcal{S}_i \cup E^{t-1}))) \tag{4}$$

where the constant values can be omitted in argmax. At each iteration, the sentence that maximize $\Delta P_{\mathcal{S}_i}$ is selected, as done in Eq. (2). This promotes selection that increases confidence $P(\mathcal{A}|\cdot)$ on important sentences, and decreases confidence on unimportant sentences. We stop the iterations if $\Delta P_{\mathcal{S}_i} < 0$ or $t = T$, then the final sentences in $E_{t=T}$ are a pseudo evidence-positive set $\mathbb{E}^+$. To reduce the search space, we empirically set $T = 5$[1].

Briefly, we obtain the labels of answerability and evidentiality, as follows:

- Answer-positive $\mathbb{A}^+$ and negative $\mathbb{A}^-$ set: the former has both answer and evidences, and the latter has neither.
- Evidence-positive $\mathbb{E}^+$ and negative $\mathbb{E}^-$ set: the former is expected to have all the evidences, and the latter has an answer with no evidence.

[1]Based on observations that 99% in HotpotQA require less than 6 evidence sentences for reasoning.

3.3 Learning Answerability & Evidentiality

In this section, our goal is to learn the above labels of answerability and evidentiality.

Supervising Answers and Answerability (Base)

As optimizing QA model is not our focus, we adopt the existing model in (Min et al., 2019). As the architecture of QA modal, we use a powerful transformer-based model – RoBERTa (Liu et al., 2019), where the input is `[CLS] question [SEP] passage [EOS]`. The output of the model is as follows:

$$h = \text{RoBERTa (Input)} \in \mathbb{R}^{n \times d}$$
$$O^s = f_1(h), \;\; O^e = f_2(h)$$
$$P^s = softmax(O^s), \;\; P^e = softmax(O^e) \tag{5}$$

where f_1 and f_2 are fully connected layers with the trainable parameters $\in \mathbb{R}^d$, P^s and P^e are the the probabilities of start and end positions, d is the output dimension of the encoder, n is the size of the input sequence.

For answerability, they build a classifier through the hidden state $h_{[0,:]}$ of `[CLS]` token that represents both $\mathcal{Q}$ and $\mathcal{D}$. As HotpotQA dataset covers both yes-or-no and span-extraction questions, which we follow the convention of (Asai et al., 2019) to support both as a multi-class classification problem of predicting the four probabilities:

$$P^{cls} = softmax(W_1 h_{[0,:]}) = [p_{span},\; p_{yes},\; p_{no},\; p_{none}] \tag{6}$$

where p_{span}, p_{yes}, p_{no}, and p_{none} denote the probabilities of the answer type being span, `yes`, `no`, and no answer, respectively, and $W_1 \in \mathbb{R}^{4 \times d}$ is the trainable parameters. For training answer span and its class, the loss function of example i is the sum of cross entropy losses (D_{CE}), as follows:

$$D_{CE}(P_i, \mathcal{A}_i) = -\Big(log(P^s_{s_i}) + log(P^e_{e_i})\Big)$$
$$D_{CE}(P^{cls}_i, C_i) = -log(P^{cls}_{c_i})$$
$$\mathcal{L}_A(i) = D_{CE}(P_i, \mathcal{A}_i) + D_{CE}(P^{cls}_i, C_i) \tag{7}$$

where s_i and e_i are the starting and ending position of answer $\mathcal{A}$, respectively, and c_i is the index of the actual class C_i in example i.

Supervising Evidentiality

As overviewed in Section 1, Base model is reported to take a shortcut, or a direct path between answer $\mathcal{A}$ and question $\mathcal{Q}$, neglecting implicit intermediate

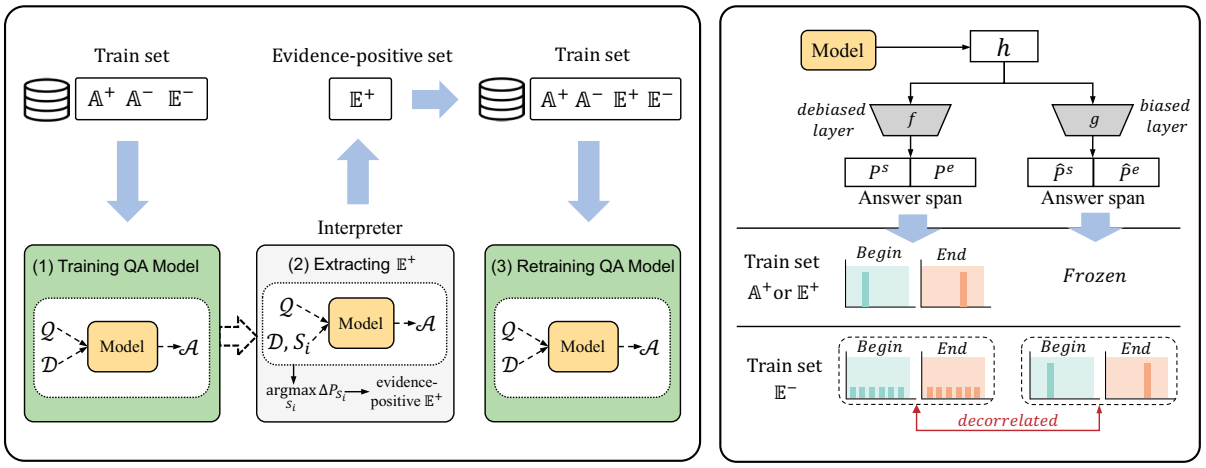


Figure 2: Learning of our proposed approach: (a) Training QA model for evidentiality, extracted by *Interpreter*. (b) Our QA predictor for learning decorrelated features on biased examples.

paths (evidences). Specifically, we present the two objectives for unbiased models:

- (O1): QA model should not be overconfident on passages with no evidences (*i.e.*, on $\mathbb{E}^-$).
- (O2): QA model should be confident on passages with both answer/evidences (*i.e.*, on $\mathbb{E}^+$)

For (O1), as a naive approach, one may consider a regularization term to avoid overconfidence on evidence-negative set $\mathbb{E}^-$. Overconfident answer distribution would be diverged from uniform distribution, such that Kullback–Leibler (KL) divergence $KL(p||q)$, where p and q are the answer probabilities and the uniform distribution, respectively, is high when overconfident:

$$\mathcal{R} = \sum_{i \in \mathbb{E}^-} D_{KL}(P(\mathcal{A}_i|\mathcal{Q}_i, \mathcal{D}_i)||P_{uniform}) \quad (8)$$

where $P_{uniform}$ indicates uniform distribution. This regularization term $\mathcal{R}$ forces the answer probabilities on $\mathbb{E}^-$ to be closer to the uniform one.

However, one reported risk (Utama et al., 2020; Grand and Belinkov, 2019) is that suppressing data with biases has a side-effect of lowering confidence on unbiased data (especially on in-distribution). Similarly, in our case, regularizing to keep the confidence low for $\mathbb{E}^-$, can cause lowering that for $\mathbb{E}^+$, due to their correlation. In other words, pursuing (O1) violates (O2), which we observe later in Figure 3. Our next goal is thus to decorrelate two distributions on $\mathbb{E}^+$ and $\mathbb{E}^-$ to satisfy both (O1) and (O2).

Figure 2(b) shows how we feed the hidden states h into two predictors. Predictor f is for learning the target distribution and predictor g is purposedly trained to be overconfident on evidence-negative set $\mathbb{E}^-$, where this biased answer distribution is denoted as $\hat{P}$. We regularize target distribution P to diverge from the biased distribution of $\hat{P}$.

Formally, the biased answer distributions $\hat{P}$ ($\hat{P}^s$ and $\hat{P}^e$) are as follows:

$$\begin{aligned} \hat{O}^s &= g_1(h), \quad \hat{O}^e = g_2(h) \\ \hat{P}^s &= softmax(\hat{O}^s), \quad \hat{P}^e = softmax(\hat{O}^e) \end{aligned} \quad (9)$$

where g_1 and g_2 are fully connected layers with the trainable parameters $\in \mathbb{R}^d$. Then, we optimize $\hat{P}$ to predict answer $\mathcal{A}$ on evidence-negative set $\mathbb{E}^-$, which makes layer g biased (taking shortcuts), and regularize f by maximizing KL divergence between P and fixed $\hat{P}$. The regularization term of example $i \in \mathbb{E}^-$ is as follows:

$$\hat{\mathcal{R}}(i) = D_{CE}(\hat{P}_i, \mathcal{A}_i) - \lambda D_{KL}(\hat{P}_i||P_i) \quad (10)$$

where λ is a hyper-parameter. This loss $\hat{\mathcal{R}}$ is optimized on only evidence-negative set $\mathbb{E}^-$.

Lastly, to pursue (O2), we train on $\mathbb{E}^+$, as done on $\mathbb{A}^+$. However, in initial steps of training, our *Interpreter* is not reliable, since the QA model is not trained enough yet. We thus train without $\mathbb{E}^+$ for the first K epochs, then extract $\mathbb{E}^+$ at K epoch and continue to train on all sets, as shown in Figure 2(a). In the final loss function, we apply different

losses as set $\mathbb{E}$ and $\mathbb{A}$:

$$\mathcal{L}_{total} = \sum_{i \in \mathbb{A}^{+,-}} \mathcal{L}_A(i) + \sum_{i \in \mathbb{E}^-} \hat{\mathcal{R}}(i) + \sum_{i \in \mathbb{E}^+} u(t-K) \cdot \mathcal{L}_A(i) \quad (11)$$

where the function u is a delayed step function (1 when epoch t is greater than K, 0 otherwise).

3.4 Passage Selection at Inference Time

For our multi-hop QA task, it requires to find answerable passages with both answer and evidence, from candidate passages. While we can access the ground-truth of answerability in training set, we need to identify the answerability of $(\mathcal{Q}, \mathcal{D})$ at inference time. For this, we consider two directions: (1) Paragraph Pair Selection, which is specific to HotpotQA, and (2) Supervised Evidence Selector trained on pseudo-labels.

For (1), we consider the data characteristic, mentioned in Section 3.1; we know one pair of paragraphs is answerable/evidential (when both paragraphs are positive, or $\mathcal{P}^+$). Thus, the goal is to identify the answerable pair of paragraphs, from all possible pairs $\mathcal{P}_{ij} = \{(p_i, p_j) : p_i \in \mathcal{P}, p_j \in \mathcal{P}\}$ (denoted as **paired-paragraph**). We can let the model select one pair with highest estimated answerability, $1 - p_{none}$ in Eq. (6), and predict answers on the paired passage, which is likely to be evidential.

For (2), some pipelined approaches (Nie et al., 2019; Groeneveld et al., 2020) design an evidence selector, extracting top k sentences from all candidate paragraphs. While they supervise the model using ground-truth of evidences, we assume there is no such annotation, thus train on pseudo-labels $\mathbb{E}^+$. We denote this setting as **selected-evidences**. For evidence selector, we follow an extracting method in (Beltagy et al., 2020), where the special token `[S]` is added at ending position of each sentence, and $h_{[S_i]}$ from BERT indicates i-th sentence embedding. Then, a binary classifier $f_{evi}(h_{[S_i]})$ is trained on the pseudo-labels, where f_{evi} is a fully connected layer. During training, the classifier identifies whether each sentence is evidence-positive (1) or negative (0). At inference time, we first select top 5 sentences[2] on paragraph candidates, and then insert the selected evidences into QA model for testing.

[2]Table 1 shows the precision and recall of top5 sentences.

Table 1: The precision and recall of pseudo evidences from *Interpreter*, compared to the ground truth (GT).

	# of sent	Prec	Recall
GT evidences	2.38	100.	100.
Answerable $\mathbb{A}^+$	6.45	36.94	100.
$\mathbb{E}^+$ (Train set)	3.64	61.13	86.64
$\mathbb{E}^+$ (Dev set)	5.00	46.12	90.35

While we discuss how to get the answerable passage above, we can use the passage setting for evaluation. To show the robustness of our model, we construct a challenge test set by excluding easy examples (*i.e.*, easy to take shortcuts). To detect such easy examples, we build a set of **single-paragraph** $\mathcal{P}_i$, that none of it is evidential in HotpotQA, as the dataset avoids having all evidences in a single paragraph, to discourage single-hop reasoning. If QA model predicts the correct answer on the (unevidential) single-paragraph, we remove such examples in HotpotQA, and define the remaining set as the challenge set.

4 Experiment

In this section, we formulate our research questions to guide our experiments and describe evaluation results corresponding to each question.

Research Questions To evaluate the effectiveness of our method, we address the following research questions:

- **RQ1**: How effective is our proposed method for a multi-hop QA task?
- **RQ2**: Does our *Interpreter* effectively extract pseudo-evidentiality annotations for training?
- **RQ3**: Does our method avoid reasoning shortcuts in unseen data?

Implementation Our implementation settings for QA model follow RoBERTa (Base version with 12 layers) (Liu et al., 2019). We use the Adam optimizer with a learning rate of 0.00005 and a batch-size of 8 on RTX titan. We extract the evidence-positive set after 3 epoch (K=3 in Eq. (11)) and retrain for 3 epochs. As a hyper-parameter, we search λ among $\{1, 0.1, 0.01\}$, and found the best value (λ=0.01), based on 5% hold-out set sampled from the training set.

Table 2: The comparison of the proposed models on the original set and challenge set.

	Model	Input at Inference	Question Answering (F1)	
			Original Set	Challenge Set
without external knowledge				
B-I:	Single-paragraph QA	Single-paragraph	68.65	0.0
B-II:	Single-paragraph QA	Paired-paragraph	62.01	30.07
O-I:	Our model	Single-paragraph	32.61	19.81
O-II:	Our model	Paired-paragraph	68.08	41.69
O-III:	Our model (full)	Selected-evidences	**70.21**	**44.57**
with external knowledge				
C-I:	Asai et al. (2019)	Retrieved-evidences	73.30	48.54
C-II:	Asai et al. (2019) + Ours	Retrieved-evidences	73.95	50.15

Table 3: The ablation study on our full model.

Model	QA (F1)	
	Original	Challenge
Our model (full)	**70.21**	**44.57**
(A) remove $\mathbb{E}^+$	68.51	40.78
(B) remove $\mathbb{E}^+$ & $\mathbb{E}^-$	66.42	40.75
(C) replace $\hat{\mathcal{R}}$ with $\mathcal{R}$	69.64	42.54

Metrics We report standard F1 score for HotpotQA, to evaluate the overall QA accuracy to find the correct answers. For evidence selection, we also report F1 score, Precision, and Recall to evaluate the sentence-level evidence retrieval accuracy.

4.1 RQ1: QA Effectiveness

Evaluation Set

- **Original Set**: We evaluate our proposed approach on multi-hop reasoning dataset, HotpotQA[3] (Yang et al., 2018). HotpotQA contains 112K examples of multi-hop questions and answers. For evaluation, we use the HotpotQA dev set (distractor setting) with 7405 examples.

- **Challenge Set**: To validate the robustness, we construct a challenge set where QA model on **single-paragraph** gets zero F1, while such model achieves 67 F1 in the original set. That is, we exclude instances with F1 > 0, where the QA model predicts an answer without right reasoning. The exclusion makes sure the baseline obtains zero F1 on the challenge set. The number of surviving examples in our challenge set is 1653 (21.5% of dev set).

[3]https://hotpotqa.github.io/

Baselines, Our models, and Competitors As a baseline, we follow the previous QA model (Min et al., 2019) trained on single-paragraphs. We test our model on single-paragraphs, paired-paragraphs and selected evidences settings discussed in Section 3.4. As a strong competitor, among released models for HotpotQA, we implement a state-of-the-art model (Asai et al., 2019)[4], using external knowledge and a graph-based retriever.

Main Results This section includes the results of our model for multi-hop reasoning. As shown in Table 2, our full model outperforms baselines on both original and challenge set.

We can further observe that **i)** when tested on single-paragraphs, where forced to take shortcuts, our model (O-I) is worse than the baseline (B-I), which indicates that B-I learned the shortcuts. In contrast, O-II outperforms B-II on paired-paragraphs where at least one passage candidate has all the evidences.

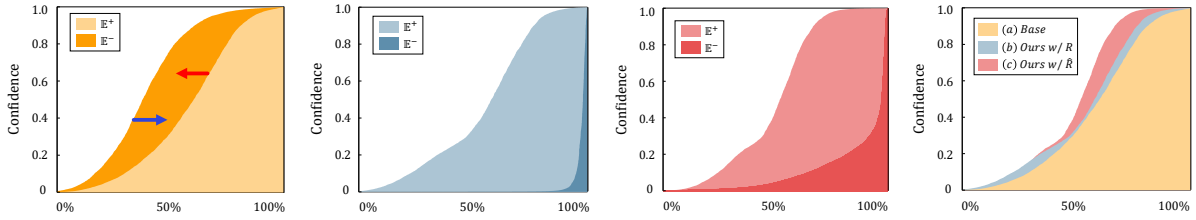
ii) When tested on evidences selected by our method (O-III), we can improve F1 scores on both original set and challenge set. This noise filtering effect of evidence selection, by eliminating irrelevant sentences, was consistently observed in a supervised setting (Nie et al., 2019; Groeneveld et al., 2020; Beltagy et al., 2020), which we could reproduce without annotation.

iii) Combining our method with SOTA (C-I) (Asai et al., 2019) leads to accuracy gains in both sets. C-I has distinctions of using external knowledge of reasoning paths, to outperform models without such advantages, but our method can contribute to complementary gains.

[4]Highest performing model in the leaderboard of HotpotQA with public code release

Table 4: The comparison of the proposed models for evidence selection

Model	Evidence Selection		
	F1	Precision	Recall
Retrieval-based AIR (Yadav et al., 2020)	66.16	**63.06**	69.57
Accumulative-based interpreter on our QA model	54.05	53.56	62.38
(a) *Interpreter* on Single-paragraph QA	56.76	57.50	63.71
(b) *Interpreter* on our QA model w/ $\mathcal{R}$	**70.30**	62.04	**87.10**
(c) *Interpreter* on our QA model (full)	69.35	61.09	86.59



(a) Single-paragraph QA (b) Ours w/ $\mathcal{R}$ (c) Ours w/ $\hat{\mathcal{R}}$ (full) (d) Three models on $\mathbb{E}^+$

Figure 3: **Confidence Analysis:** Confidence scores of three models in the ascending order, on $\mathbb{E}^+$ (light color) and $\mathbb{E}^-$ (dark colar). (a) Base model trained on single-paragraphs. (b) Our model with $\mathcal{R}$. (c) Our full model with $\hat{\mathcal{R}}$. (d) Comparison of three models on $\mathbb{E}^+$.

Ablation Study As shown in Table 3, we conduct an ablation study of O-III in Table 2. In (A), we remove $\mathbb{E}^+$ from *Interpreter*, in training time. On the QA model without $\mathbb{E}^+$, the performance decreased significantly, suggesting the importance of evidence-positive set. In (B), we remove evidentaility labels of both $\mathbb{E}^+$ and $\mathbb{E}^-$, and observed that the performance drop is larger compared to other variants. Through (A) and (B), we show that training our evidentiality labels can increase QA performance. In (C), we replace $\hat{\mathcal{R}}$ with $\mathcal{R}$, removing layer g to train biased features. On the replaced regularization, the performance also decreased, suggesting that training $\hat{\mathcal{R}}$ is effective for a multi-hop QA task.

4.2 RQ2: Evaluation of Pseudo-Evidentiality Annotation

In this section, we evaluate the effectiveness of our *Interpreter*, which generates evidences on training set, without supervision. We compare the pseudo evidences with human-annotation, by sentence-level. For evaluation, we measure sentence-level F1 score, Precision and Recall, following the evidence selection evaluation in (Yang et al., 2018).

As a baseline, we implement the retrieval-based model, AIR (Yadav et al., 2020), which is an unsupervised method as ours. As shown in Table 4, our *Interpreter* on our QA model outperforms the retrieval-based method, in terms of F1 and Recall, while the baseline (AIR) achieves the highest precision (63.06%). We argue recall, aiming at identifying all evidences, is much critical for multi-hop reasoning, for our goal of avoiding disconnected reasoning, as long as precision remains higher than precision of answerable $\mathbb{A}^+$ (36.94%), in Table 1.

As variants of our method, we test our *Interpreter* on various models. First, when comparing (a) and (c), our full model (c) outperforms the baseline (a) over all metrics. The baseline (a) trained on single-paragraphs got biased, thus the evidences generated by the biased model are less accurate. Second, the variant (b) trained by $\mathcal{R}$ outperforms (c) our full model. In Eq. (8), the loss term $\mathcal{R}$ does not train layer g for biased features, unlike $\hat{\mathcal{R}}$ in Eq. (10). This shows that learning g results in performance degradation for evidence selection, despite performance gain in QA.

4.3 RQ3: Generalization

In this section, to show that our model avoids reasoning shortcuts for unseen data, we analyze the confidence distribution of models on the evidence-positive and negative set. In dev set, we treat the ground truth of evidences as $\mathbb{E}^+$, and a single sentence containing answer as $\mathbb{E}^-$ (each has 7K $\mathcal{Q}$-$\mathcal{D}$ pairs). On these set, Figure 3 shows **confidence** $P(\mathcal{A}|\mathcal{Q},\mathcal{D})$ of three models; (a), (b), and (c) men-

tioned in Section 4.2. We sort the confidence scores in ascending order, where y-axis indicates the confidence and x-axis refers to the sorted index. Thus, the colored area indicates the dominance of confidence distribution. Ideally, for a debiased model, the area on evidence-positive set should be large, while that on evidence-negative should be small.

Desirably, in Figure 3(a), the area under the curve for $\mathbb{E}^-$ should decrease for pursuing (O1), moving along *blue* arrow, while that of $\mathbb{E}^+$ should increase for (O2), as *red* arrow shows. In Figure 3(b), our model with $\mathcal{R}$ follows *blue* arrow, with a smaller area under the curve for $\mathbb{E}^-$, while keeping that of $\mathbb{E}^+$ comparable to Figure 3(a). For the comparison, Figure 3(d) shows all curves on $\mathbb{E}^+$. In Figure 3(c), our full model follows both directions of *blue* and *red* arrows, which indicates that ours satisfied both (O1) and (O2).

5 Conclusion

In this paper, we propose a new approach to train multi-hop QA models, not to take reasoning shortcuts of guessing right answers without sufficient evidences. We do not require annotations and generate pseudo-evidentiality instead, by regularizing QA model from being overconfident when evidences are insufficient. Our experimental results show that our method outperforms baselines on HotpotQA and has the effectiveness to distinguish between evidence-positive and negative set.

Acknowledgements

This research was supported by IITP grant funded by the Korea government (MSIT) (No.2017-0-01779, XAI) and ITRC support program funded by the Korea government (MSIT) (IITP-2021-2020-0-01789).

References

Akari Asai, Kazuma Hashimoto, Hannaneh Hajishirzi, Richard Socher, and Caiming Xiong. 2019. Learning to retrieve reasoning paths over wikipedia graph for question answering. In *International Conference on Learning Representations*.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Jifan Chen and Greg Durrett. 2019. Understanding dataset design choices for multi-hop reasoning. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*.

Yuwei Fang, Siqi Sun, Zhe Gan, Rohit Pillai, Shuohang Wang, and Jingjing Liu. 2020. Hierarchical graph network for multi-hop question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8823–8838.

Gabriel Grand and Yonatan Belinkov. 2019. Adversarial regularization for visual question answering: Strengths, shortcomings, and side effects. In *Proceedings of the Second Workshop on Shortcomings in Vision and Language*, pages 1–13.

Dirk Groeneveld, Tushar Khot, Ashish Sabharwal, et al. 2020. A simple yet strong pipeline for hotpotqa. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8839–8845.

Sarthak Jain and Byron C Wallace. 2019. Attention is not explanation. 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 3543–3556.

Yichen Jiang and Mohit Bansal. 2019. Avoiding reasoning shortcuts: Adversarial evaluation, training, and model development for multi-hop qa. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2726–2736.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Sewon Min, Eric Wallace, Sameer Singh, Matt Gardner, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2019. Compositional questions do not necessitate multi-hop reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4249–4257.

Yixin Nie, Songhe Wang, and Mohit Bansal. 2019. Revealing the importance of semantic retrieval for machine reading at scale. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2553–2566.

Yilin Niu, Fangkai Jiao, Mantong Zhou, Ting Yao, Minlie Huang, et al. 2020. A self-training method for machine reading comprehension with soft evidence extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3916–3927.

Ethan Perez, Siddharth Karamcheti, Rob Fergus, Jason Weston, Douwe Kiela, and Kyunghyun Cho. 2019. Finding generalizable evidence by learning to convince q\&a models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2402–2411.

Lin Qiu, Yunxuan Xiao, Yanru Qu, Hao Zhou, Lei Li, Weinan Zhang, and Yong Yu. 2019. Dynamically fused graph network for multi-hop reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6140–6150.

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.

Sofia Serrano and Noah A Smith. 2019. Is attention interpretable? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2931–2951.

Krishna Kumar Singh, Dhruv Mahajan, Kristen Grauman, Yong Jae Lee, Matt Feiszli, and Deepti Ghadiyaram. 2020. Don't judge an object by its context: Learning to overcome contextual bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11070–11078.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2020. Is multihop qa in dire condition? measuring and reducing disconnected reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8846–8863.

Lifu Tu, Garima Lalwani, Spandana Gella, and He He. 2020. An empirical study on robustness to spurious correlations using pre-trained language models. *Transactions of the Association for Computational Linguistics*, 8:621–633.

Prasetya Ajie Utama, Nafise Sadat Moosavi, and Iryna Gurevych. 2020. Mind the trade-off: Debiasing nlu models without degrading the in-distribution performance. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8717–8729.

Vikas Yadav, Steven Bethard, and Mihai Surdeanu. 2020. Unsupervised alignment-based iterative evidence retrieval for multi-hop question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4514–4525.

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.