QUASER: Question Answering with Scalable Extractive Rationalization

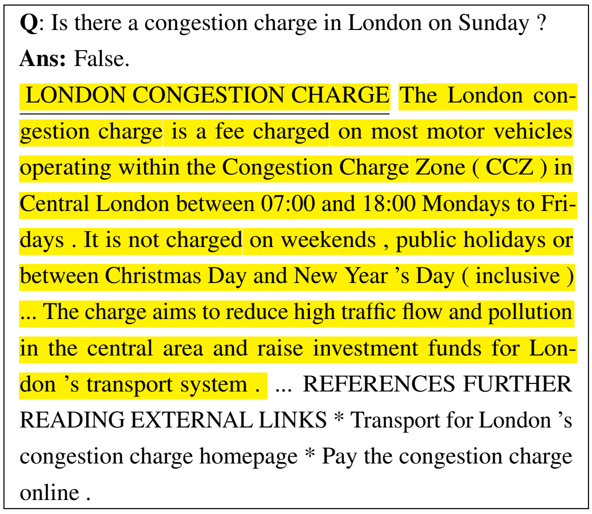
Anonymous ACL submission

Abstract

Designing NLP models that produce predictions by first extracting a set of relevant input sentences (i.e., rationales), is gaining importance as a means to improving model interpretability and to producing supporting evidence for users. Current unsupervised approaches are trained to extract rationales that maximize prediction accuracy, which is invariably obtained by exploiting spurious correlations in datasets, and leads to unconvincing rationales. In this paper, we introduce unsupervised generative models to extract dual-purpose rationales, which must not only be able to support a subsequent answer prediction, but also support a reproduction of the input query. We show that such models can produce more meaningful rationales, that are less influenced by dataset artifacts, and as a result, also achieve the state-of-the-art on rationale extraction metrics on four datasets from the ERASER benchmark, significantly improving upon previous unsupervised methods.

1 Introduction

While large pre-trained transformer models (Devlin et al., 2019; Raffel et al., 2019) have achieved state-of-the-art results on many question answering (QA) tasks, the process by which they generate their predictions is opaque. Therefore, to shed light on the prediction process and to increase user trust, training models to additionally present portions of the input i.e. rationales, as supporting evidence, has emerged as an effective solution (Lei et al., 2016; Yang et al., 2018). However, current approaches (Paranjape et al., 2020a) are trained to select rationales that optimize prediction accuracy, which is invariably achieved by exploiting dataset artifacts, and consequentially, results in unconvincing rationales. To alleviate these shortcomings, we introduce a generative approach to produce *dual-purpose rationales*, that are required to independently support a reproduction of the input query in addition to improving model prediction, thereby necessitating more meaningful rationales.



Q: Is there a congestion charge in London on Sunday ?

Ans: False.

LONDON CONGESTION CHARGE The London congestion charge is a fee charged on most motor vehicles operating within the Congestion Charge Zone (CCZ) in Central London between 07:00 and 18:00 Mondays to Fridays . It is not charged on weekends , public holidays or between Christmas Day and New Year 's Day (inclusive) ... The charge aims to reduce high traffic flow and pollution in the central area and raise investment funds for London 's transport system . ... REFERENCES FURTHER READING EXTERNAL LINKS * Transport for London 's congestion charge homepage * Pay the congestion charge online .

Figure 1: An example from the BoolQ data set in the ERASER benchmark (DeYoung et al., 2020), with human annotated rationales highlighted in yellow. Rationales predicted by a supervised BERT-based pipeline method (DeYoung et al., 2020) is shown underlined.

We focus on developing QA models that generate an answer based on a question and a (potentially long) passage, together with NL rationales. In this case, a rationale (or explanation) is defined as a minimal subset of passage sentences that is sufficient to answer the question. We present example questions, passages, answers from the BoolQ dataset (Clark et al., 2019) with human annotated rationales from the ERASER benchmark (DeYoung et al., 2020) in Figure 1. Supervised rationalization models for this task typically require large amount of expensive annotations, making unsupervised methods attractive. State-of-the-art unsupervised methods take a pipelined approach (Lehman et al., 2019; Paranjape et al., 2020a) where an extractor model classifies each sentence in the passage to be relevant or irrelevant to answering the question, while a separate model predicts the answer from the chosen relevant sentences. No parameters are shared between the two models to ensure faithfulness, i.e., the predicted answer relies only on the selected rationale sentences. Unsupervised methods also incorporate additional sparsity constraints (for example, adding sparsity inducing norms), to encourage the selection of a small number of sentences as rationales.

However, unsupervised pipelined approaches suffer from two main shortcomings. The first is that, their memory-intensive use of two separate models restrict them to making use of only base pre-trained models, making it difficult to scale to larger versions of pre-trained transformer models that significantly improve QA answer generation performance. Furthermore, they only use half the total capacity of the full (pipelined) neural network for answer prediction. Secondly, the sole objective for extracting rationales is answer prediction accuracy, which is invariably optimized by exploiting spurious correlations and dataset artifacts. As a result, the extracted rationales may explain dataset biases rather than present evidence for answering the question, resulting in unconvincing explanations.

Our contributions We propose a method for generating faithful explanations for query based tasks using a single model by adding a rationale selection module between the encoder and decoder of a Transformer model. We identify two key conceptual problems with existing rationalization schemes: reliance on spurious correlations, and lack of comprehensiveness constraints — a key metric for ensuring faithfulness of rationales. To address these, we propose a multi-task learning objective where we train our model jointly on a *forward objective* that predicts the answer given the question and passages, and a *backward objective* that predicts the question from the passage. We call our model trained using this forward-backward objective: `QUASER-FB`. We show that such joint training improves both answer accuracy and rationale selection performance while also improving faithfulness.

- Specifically, on four QA data sets in the ERASER (DeYoung et al., 2020) explainability benchmark, `QUASER-FB`, when initialized with T5-base (Raffel et al., 2019), achieves on an average 10.4% absolute improvement for answer generation and 7.8% absolute improvement for rationale selection over the previous unsupervised state of the art (Paranjape et al., 2020b). Lastly, our method achieves an average absolute improvement of 8.9% for answer generation over the *supervised* BERT-based pipeline model of DeYoung et al. (2020) on three out of four ERASER datasets.
- We show that augmenting our model with the question generation objective produces rationales that are 11.4% more comprehensive on two datasets in which comprehensiveness can be measured.
- Our method is scalable to large pre-trained transformer models (Vaswani et al., 2017) and achieves state-of-the-art performance while having roughly the same number of parameters as existing BERT-based supervised and unsupervised pipeline methods.

Finally, we show that the quality of rationales generated by our method are more correlated with answer accuracy than baselines, thereby making them more suitable to verify answer correctness (Lipton, 2018).

2 Preliminaries

	Answer EM		Rationale IOU F1	
Dataset	Qs. + Psg.	Psg. only	Qs. + Psg.	Psg. only
BoolQ	63	61	30	29
MultiRC	60	57	45	16

Table 1: Performance of BERT-based pipeline model of DeYoung et al. (2020) on two QA datasets under two settings: (a) Qs. + Psg.: where the model is trained to produce the answer given the question and passage, and (b) Psg. only: where the model has to generate the answer from the passage only.

In this section, we formally define the problem of faithful selective rationalization in question answering and describe some of the ways in which existing approaches can rely on spurious correlations to select rationales.

Question Answering tasks involve generating an answer Y given a question Q and a passage $X = (X_1, \ldots, X_n)$. Each sentence X_i, and the question Q, in-turn contain multiple tokens belonging to a vocabulary V, of size k. In this paper, we consider the setting where the answer belongs to a small finite set $\mathcal{Y}$ of size c. In the datasets that we evaluate our method on the answers are all binary ($c =$

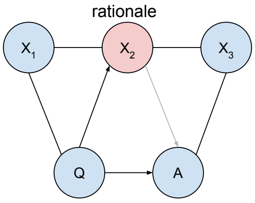


Figure 2: A generative model for QA where X_2 denotes the true rationales while X_1 and X_3 denote sentences that are correlated with the question and answer respectively.

2). A *faithful extractive rationale* is a subset $S \subseteq \{1, \ldots, n\}$ of sentences in the passage that is *used* by the model to generate the answer[1].

2.1 Faithful rationale selection

Yu et al. (2019) define three main desiderata for selecting faithful rationales: (a) **sufficiency:** the rationales should be sufficient to generate the answer Y, (b) **comprehensiveness**: all sentences which are useful for predicting the answer should be included in the rationales, and (c) **compactness**: the rationales should contain a small number of sentences. More formally, sufficiency entails selecting rationales S that maximize $I(Y, X_S \mid Q)$, where $I(\cdot, \cdot)$ denotes mutual information and X_S denotes the rationale. However, a model can trivially achieve sufficiency by choosing $S = \{1, \ldots, n\}$. Therefore, compactness ensures that rationales are succinct and interpretable. Lastly, Yu et al. (2019) define comprehensiveness as selecting rationales S such that $H(Y \mid X_{S^c}, Q) - H(Y \mid X_S, Q) \geq h$ for some constant h, where $H(\cdot)$ denotes Shannon entropy and S^c denotes the complement of S. The constant h can be interpreted as a *margin constraint* with a *large margin* implying *more comprehensive* rationale selection with X_{S^c} containing very little information about Y. This in turn encourages the model to select rationales based on robust features as opposed to relying on spurious correlations.

3 Motivation

We motivate the problems with current extractive rationalization schemes through a simplified probabilistic model. Fig. 2 shows a potential generative model for question answering, which is a causal partially directed acyclic graph (Pearl, 2009), with directed edges showing causal relationships. Given a question Q a function selects the relevant sentences from the passages which in this case is denoted by X_2. Then given the relevant sentences and the question another function produces the answer. There are additionally sentences that are only spuriously corrleated with the answer (X_3) and the question (X_1) — for instance, an overwhelming majority of the annotated rationales, and subsequently the correct answer, in BoolQ are at the very beginning of the passages. To verify the existence of sentences or tokens that are only correlated with the answer, we performed a quick experiment where we train the BERT-based pipeline model of DeYoung et al. (2020) to predict the answer from the passage only. Results of the experiments are shown in Table 1. Given access to the passage only, the model suffers a minimal ($\sim$ 2-3%) drop in answer generation performance while still achieving significant rationale selection performance [2]. Similar results have also been reported by Kaushik and Lipton (2018) on other benchmarks.

Spurious correlations The first problem with current approaches is that they do not preclude selection of sentences that are spuriously correlated with the answer. For instance, in Figure 2 given the question Q, X_3 is not independent of A, i.e. $X_3 \not\perp\!\!\!\perp A \mid Q$. Therefore, there is nothing preventing the model from selecting X_3 as a rationale for predicting A. Furthermore, stringent compactness constraints can result in the true rationale X_2 being excluded if the causal strength between the rationale and the answer is weak (figuratively denoted by a lighter colored edge). Since large pre-trained language models are known to store knowledge within their parameters (Roberts et al., 2020), they can predict the answer correctly just from the question. In the unsupervised setting where even the average size of rationales for a dataset is unknown, how to set the right compactness constraints is a challenging problem.

Lack of comprehensiveness constraints Second, existing approaches do not optimize for comprehensiveness. Selecting rationales that maximize answer accuracy (equivalently minimize $H(Y \mid X_S, Q)$) can result in reduced comprehensiveness of rationales. To see this observe that $H(Y \mid X_S, Q) \leq H(Y) = \log c$ which is fairly small to begin with and large neural networks

[1] We consider the setting where rationales are selected at the sentence level, although our method can also generate token level rationales.

[2] Note that MultiRC is a multiple choice QA task which has been converted to a binary QA task by appending the choice to the question and asking the model to predict if the choice is correct or wrong. Therefore, in the passage only experiments the model does not have access to the choice and only predicts True or False from the passage.

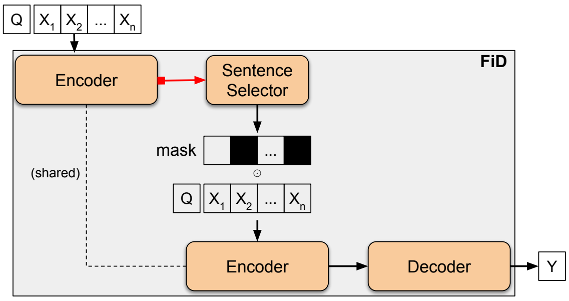


Figure 3: Model architecture. The two encoders represents the same encoder invoked twice. In the second invocation the sentence mask produced by the sentence selector is used for encoder side self-attention and encoder-decoder cross-attention to mask out sentences that are not part of the rationale. The red arrow denotes an unidirectional arrow along which gradients are not back-propagated during training.

are expressive enough to drive $H(Y \mid X_S, Q)$ to 0 during training by exploiting spurious correlations. Lewis and Fan (2019) intuitively refer to this as *loss saturation*. Therefore, the constraint $H(Y \mid X_{S^c}, Q) - H(Y \mid X_S, Q) \geq h$ is satisfied for a small margin h thereby resulting in reduced comprehensiveness.

Note that these two issues are complimentary. Incorporating explicit comprehensiveness constraints like those in (Yu et al., 2019) for answer generation is insufficient for excluding spuriously correlated sentences in rationales.

4 Method

The main idea behind our method is to select rationales that are simultaneously useful for generating both the question and the answer. Unlike (Yu et al., 2019), which requires an additional margin hyper-parameter h, we do not explicitly optimize for comprehensiveness but we demonstrate that augmenting our method with a question generation objective implicitly improves comprehensiveness. Like previous work, we represent rationales by a binary mask over sentences. We have a sentence selector m that takes as input a passage X (and optionally the question Q) and produces a binary mask $m(X) \in \{0,1\}^n$. Rationale selection is then denoted by $X \odot m(X)$ where $\odot$ denotes element-wise multiplication. For a question, passage, and answer triple (q, x, y), we learn a rationale selector $m(\cdot)$ by minimizing the following objective.

$$\begin{aligned} l(q,x,y) = &-p_Y(y \mid q, x \odot m(q,x)) \\ &- p_Q(q \mid x \odot m(x)) \\ &+ \lambda_1(\|m(x)\|_1 + \|m(x,q)\|_1) \quad (1) \end{aligned}$$

The above loss is averaged over all observed triples in the dataset to compute the training loss. We compute the likelihood of the answer $p_Y(y \mid q, m(q,x))$ and the likelihood of the observed question $p_Q(q \mid x \odot m(x))$ using the same sequence-to-sequence (seq2seq) model. In Eq. (1) λ_1 controls the compactness of the generated rationales. In our experiments we do not tune λ_1 and set it to a very small value. Note that two sets of masks $m(q,x)$ and $m(x)$ are different to allow for different sets of rationales for predicting the answer and question respectively. We do not add any (norm) constraints to the objective to encourage overlap between these two sets of rationales since that would introduce another hyperparameter. However, we observe that merely sharing the same sentence selector between the question and answer generation stages encourages sharing of rationales.

The above objective improves comprehensiveness and potentially robustness of the produced rationales due to the following reason. First, the objective encourages discovery of rationales that are jointly useful for generating the question and the answer making them less susceptible to be correlated with the answer alone. Next, the second term in (1) minimizes $H(Q \mid X \odot m(X))$. Since $H(Q)$ can be as large as $|Q| \log |V|$ where $|Q|$ is the number of tokens in Q, it can be difficult even for large pre-trained models to minimize $H(Q \mid X \odot m(X))$ from the knowledge encoded in their parameters or selecting a few sentences in $m(X)$ that are spuriously correlated with Q. Lastly, since the (parameters of) sentence selector $m(\cdot)$ is shared between question and answer generation stages, this encourages the answer generation mask $m(Q,X)$ and question generation mask $m(X)$ to be close to each other. Therefore, with the inclusion of more *causally* relevant sentences in $m(Q,X)$, the comprehensiveness constraint $H(Y \mid Q, X \odot \neg m(Q,X)) - H(Y \mid Q, X \odot m(Q,X)) + h$ is implicitly satisfied with a large margin h since $\neg m(Q,X)$ now contains very little information about Y.

Model Figure 3 shows our overall approach to generating faithful rationales for QA tasks. We modify the fusion-in-decoder (FiD) model of Izacard and Grave (2020) to generate rationales as follows. The FiD model has the standard Transformer architecture (Vaswani et al., 2017) consisting of an encoder and decoder. We first pass the inputs through the encoder and compute representations of the tokens in the inputs. A *sentence selector* then uses the token representations to mask out irrelevant sentences. Then we take the relevant sentences (rationales) and pass them through the encoder again to compute token representations that do not use sentences not in the rationales. These token representations are then passed through the decoder to compute the likelihood of the output. During training, we repeat this process twice to compute the likelihood of the answer given the question and passage as input and then compute the likelihood of the question given the passage. The final loss is the sum of the negative likelihoods of the question and answer as given in (1). Next, we describe each of the model components in detail.

Input representation In FiD, to effectively deal with long passages, each passage is broken down into multiple chunks or contexts and the question is concatenated with each chunk. Each context is then passed through the encoder of the transformer architecture to compute question-contextualized chunk representations. These representations are then concatenated and passed to the decoder which uses them to produce the answer. We modify this procedure by adding CLS tokens at the beginning and end of each sentence. For each sentence, the two CLS token embeddings are concatenated to compute the sentence representation which is used for sentence selection as described next.

Sentence selector Our main modification to FiD introduces a sentence selector that produces a binary mask over sentences from the sentence representations. Our sentence selector has the same architecture as that of (Paranjape et al., 2020b) which we describe here for completeness. Given sentence representations $v_i \in \mathbb{R}^{2d}$, for $1 \leq i \leq n$, which are obtained by concatenating the CLS token representations at the beginning and end of each sentence, the sentence selector computes the probability p_i of the i-th sentence being as rationale as follows:

$$u_i = \mathrm{dropout}(\mathrm{ReLU}(\mathbf{W}v_i))p_i \quad = \mathrm{sigmoid}(w^\top u_i)$$

The dropout parameter is set to 0.2 while $\mathbf{W} \in \mathbb{R}^{2d \times d}$ and $w \in \mathbb{R}^d$ are the parameters of the sentence selector. Since sampling a binary mask from the distribution $m_i \sim \mathrm{Bernoulli}(p_i)$ would break differentiablity of our model, we use the *Gumbel-sigmoid reparameterization trick* to sample a differentiable soft-mask $m_i \in (0, 1)$ as follows:

$$g \sim \mathrm{Gumbel}(0, 1), m_i = \mathrm{sigmoid}({}^{(\log p_i + g)}/{}_{\tau}),$$

where τ is a temperature parameter that we set to 0.7.

4.1 Comparison with pipeline models

Apart from differences in training objective, choice of base model (FiD), and pre-trained representations (T5), a key conceptual difference between our model and those of existing pipelined approaches is the use of single model with some shared parameters. While pipelined models have no shared parameters between the rationale extractor and the answer generator, the embedding layer is shared between the encoder and decoder in our model. This obviously has consequences for faithfulness. During inference, however, the encoder and decoder only rely on the rationales extracted by the sentence selector to generate the answer, in addition to the knowledge stored in their parameters. Note that even in pipelined models the answer generator can get exposed to information stored in different sentences during the course of training that are not part of eventual rationales, which the generator can use to answer questions during inference thereby affecting faithfulness. A key architectural choice that we make to improve faithfulness is not updating the encoder from the sentence selector during training. This also has the effect of improving the memory requirement of our model, since after the sentence masks are computed, the sentence representations can be discarded.

5 Experiments

Datasets We evaluate our method on four text classification tasks in the ERASER benchmark (DeYoung et al., 2020) which have been adapted as QA tasks: BoolQ (Clark et al., 2019), MultiRC (Khashabi et al., 2018), FEVER (Thorne et al., 2018), and Evidence Inference (Lehman et al., 2019). BoolQ and MultiRC are standard machine reading comprehension tasks involving boolean and multiple choice answers respectively. FEVER is a fact extraction and verification task adapted as a QA task in ERASER where the goal is to classify whether the given evidence (passage) supports or refutes the claim (question). Lastly, the Evidence

Inference dataset entails determining whether an *intervention* significantly increases, decreases, or has no effect, on an *outcome* with respect to a *comparator* of interest from clinical trial articles (passage). The intervention, outcome, and comparator triple are concatenated to form the query. We ignore datasets in the benchmark which have very short passages (between 1-2 sentences) like CoS-E and e-SNLI. We also do not consider the movie reviews dataset which is a sentiment classification task and has no query.

Baselines We compare our proposed methods (QUASER and QUASER-FB) against the information-bottleneck (IB) approach of Paranjape et al. (2020b) who report state-of-the-art unsupervised rationale extraction performance on the ERASER benchmark. Theirs is an unsupervised BERT based pipeline model with a sparsity inducing prior over masks. It is important to note that their method is not fully unsupervised as they use rationale metrics computed on the validation set for tuning conciseness of rationales and performing model selection. Whereas our method is fully unsupervised where we perform model selection purely based on answer generation performance and do not tune the sparsity controlling hyperparameter (λ_1) which we set to 0.01 for all our experiments as was done in (DeYoung et al., 2020). Furthermore, to deal with long passages in BoolQ and Evidence Inference which frequently exceed the maximum input length of 512 tokens for Transformer models, Paranjape et al. (2020b) use TF-IDF to extract a subset of the passage that has the highest overlap with the question, while we perform no such pre-processing. We also compare our method against the *supervised* BERT-based pipeline method (BERT-BERT) of DeYoung et al. (2020) which independently trains the rationale extractor to predict whether a sentence is a rationale or not on annotated gold rationales and then trains the classifier to predict the answer from the rationales. Lastly, we also report the performance of the baseline (full) that uses the entire passage to generate the output. The full baseline uses the same passage representation (i.e., number of contexts and maximum passage length) as our best performing model QUASER-FB. All methods have the same total number of parameters ($\approx$ 220M). We do not compare against the method of (Yu et al., 2019) since their use of three separate models in a three player game does not scale to using pre-trained models like BERT.

Metrics Following previous work, we use exact match for answer accuracy, and use intersection-over-union F1 score (IOU) and token F1 (TF1) score for evaluating rationale quality. IOU is computed by matching each predicted rationale with a gold rationale and computing the F1 score, where a match is considered positive if the overlap between the predicted and gold rationale exceeds a certain threshold. Like (Paranjape et al., 2020b) we use a threshold of 0.1. Token F1 score (TF1) simply computes the F1 score between the predicted and gold rationale at the token level and is not sensitive to the choice of the threshold. Since comprehensive rationales have been annotated for MultiRC and FEVER and rationale IOU recall directly measures comprehensiveness on these datasets (DeYoung et al., 2020), we will evaluate comprehensiveness using recall.

Training details As previously stated, FiD handles long passages by diving them into chunks of a certain maximum length. The number of chunks and the maximum length of chunks are hyperparameters in our model. We experiment with number of chunks in $\{4, 8, 10\}$ and maximum passage length of either 128 or 256 tokens. More details can be found in Appendix A.

6 Results

Table 2 shows the performance of our proposed methods vis-à-vis different baselines. The answer accuracy of our base model (QUASER) is uniformly better than the previous state-of-the-art unsupervised method (IB) of Paranjape et al. (2020b) across all four datasets, with QUASER achieving an average absolute improvement of 6.7% over IB. These gains come partly from using state-of-the-art base model (FiD) and pre-trained representations (T5-base). Even though our base model has almost the same number of parameters ($\approx$220M) as the BERT-based pipeline model of Paranjape et al. (2020b), we are able to use all the parameters for sentence selection and answer generation, whereas IB uses only half the parameters for answer generation. Using a larger pre-trained model can reduce faithfulness which we observe as an average drop of 2.2% rationale IOU performance of our base model across four datasets. However, it should be noted that IB is not a fully unsupervised method since they tune the sparsity hyper-parameter and

Method	BoolQ			MultiRC			FEVER			Evi. Inf.		
	Ans.	TF1	IOU	Ans.	TF1	IOU	Ans.	TF1	IOU	Ans.	TF1	IOU
full	73.8	36.0	34.0	80.6	29.0	28.0	93.2	26.7	27.4	69.9	3.0	2.5
BERT-BERT	61.6	14.4	28.2	63.1	44.3	46.0	87.7	81.2	83.5	69.8	47.6	53.5
IB	65.2	12.8	16.5	62.1	24.9	24.3	84.7	42.7	45.5	46.3	6.9	10.0
QUASER	69.9	2.9	3.3	76.8	39.8	41.2	88.2	37.6	40.2	50.3	2.9	2.6
QUASER-FB	70.2	34.4	34.6	78.1	41.4	42.9	90.8	39.0	42.1	60.9	3.6	3.1

Table 2: Answer accuracy and rational token and IOU F1 on four datasets in the ERASER benchmark. IB refers to the information-bottleneck approach of Paranjape et al. (2020b), BERT-BERT is the supervised BERT-based pipeline model of DeYoung et al. (2020), QUASER refers to our model trained to generate the answer only, while QUASER-FB denotes our model trained with multi-task objective of generating both the answer and the question.

perform model selection based on the development set rationale IOU, whereas we only use the answer accuracy for model selection.

Augmenting our base model (QUASER) with the question generation objective further improves answer accuracy uniformly across four datasets by 10.4% over IB while also improving rationale IOU by 6.6% on average across four datasets. Our final model QUASER-FB achieves significantly better rationale scores over IB on BoolQ and MultiRC, while almost achieving parity on FEVER. All methods perform poorly on the Evidence Inference dataset. Poor performance of our method on the Evidence Inference dataset is because of the extremely long passages in the dataset and our passage representation missing out most of the annotated rationales. The full passage representation (#contexts: 8 and maximum passage length: 256 tokens) input to our model has an rationale IOU recall of only 35.5, with QUASER-FB achieving a recall of 30.7. This can be partly addressed by using techniques in (Paranjape et al., 2020b) or using dense passage retrieval (Karpukhin et al., 2020) to find a smaller passage relevant to the question. Note that the rationale performance of QUASER-FB is similar to the "full" baseline on BoolQ which might indicate that our method is simply selecting all the sentences in the passage. The full baseline achieves (IOU) precision and recall of 25.0 and 62.2 respectively, while the corresponding numbers for QUASER-FB is 28.9 and 43.1, indicating that QUASER-FB is extracting compact rationales.

6.1 Analysis

To understand how augmenting answer generation with question generation improves faithful rationale extraction, which in turn improves answer accuracy, we dig further into rationale IOU metrics which are shown in Table 3. From the results we can conclude that question generation improves the recall (or comprehensiveness) of rationales not just for question generation but also for answer prediction. The recall of extracted rationales of QUASER-FB for answer generation is significantly better than those of QUASER while also improving or almost matching the precision of rationales of QUASER. Since answer accuracy also increases, we can also reasonably conclude that question generation improves both comprehensiveness and sufficiency and produces more robust rationales. Lastly, since recall directly measures comprehensiveness on the MultiRC and FEVER datasets (DeYoung et al., 2020), we can quantify the average improvement in comprehensiveness as 11.4%.

Figure 4 shows rationales predicted by our method and those of supervised BERT based pipeline model, and the unsupervised IB method. The examples qualitatively demonstrate how our method produces more comprehensive rationales.

Lastly, to test if the rationales generated by our method can be used by humans to gauge the correctness of the answer, we computed the Spearman's correlation between correctness (binary variable) and the IOU of the generated rationale. The correlation coefficient for QUASER-FB, BERT-BERT, and IB were 0.1, 0.05, and -0.02 respectively, thereby demonstrating that the rationales generated by our method were better suited for verifying answer correctness.

7 Related Work

Extractive rationalization (Lei et al., 2016) methods can either be supervised or unsupervised. Pruthi et al. (2020) propose weakly-supervised methods

Dataset	IB			QUASER			QUASER-FB					
	Answer generation			Answer generation			Answer generation			Question generation		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
BoolQ	15.7	63.5	25.2	8.4	2.1	3.3	28.9	43.1	34.6	22.2	62.1	32.7
MultiRC	20.1	30.7	24.3	27.9	78.3	41.2	28.8	83.7	42.9	16.4	95.9	28.0
FEVER	39.6	47.6	43.2	29.6	61.0	40.2	28.7	78.4	42.1	16.2	89.3	27.4
Evi. Inf.	5.1	11.3	7.0	1.5	8.2	2.6	1.6	30.7	3.1	1.1	34.8	2.2

Table 3: Rational IOU precision (P), recall (R), and F1 score (F1) for IB, QUASER, and QUASER-FB.

Q: do the white sox and cubs share a stadium ? **Ans:** False.

BERT-BERT: False. IB: False. QUASER-FB: False.

CUBS – WHITE SOX RIVALRY The Cubs – White Sox rivalry (also known as the Crosstown Classic ... geographical rivalry between the Chicago Cubs and the Chicago White Sox. **The Cubs are a member club of MLB 's National League (NL) Central division , and play their home games at Wrigley Field , located on Chicago 's North Side . The White Sox are a member club of MLB 's American League (AL) Central division , and play their home games at Guaranteed Rate Field , located on Chicago 's South Side .** *... The Chicago Transit Authority 's Red Line runs north ... stopping at Wrigley Field and Guaranteed Rate Field . ... In 1900 , Charles Comiskey moved ... In response , the team was renamed the " White Stockings " , which had been the original name of the Cubs from 1876 to 1889 .*

Q: is scottish law the same as english law ? **Ans:** False.

BERT-BERT: False. IB: False. QUASER-FB: False.

SCOTS LAW Scots law is the legal system of Scotland . **It is a hybrid or mixed legal system ...** ***Together with English law and Northern Irish law , it is one of the three legal systems of the United Kingdom .*** *... Although there was some indirect Roman law influence on Scots law , the direct influence of Roman law was slight up until around the 15th century . ... Legislation affecting Scotland may be passed by the Scottish Parliament , the United Kingdom Parliament , and the European Union . Some legislation passed by the pre-1707 Parliament of Scotland is still also valid .*

Figure 4: Examples from test set of the BoolQ data set where QUASER-FB achieves the best rationale IOU. Gold rationales are highlighted in yellow, while rationales predicted by QUASER-FB, BERT-BERT, and IB are shown in bold, underlined, and italicized words respectively.

for extracting supporting evidence. Among unsupervised methods, the closest to ours is the work of Paranjape et al. (2020b) who propose a simple sparse prior for the rationale extractor motivated by information bottleneck theory, and evaluate their approach on the ERASER benchmark (DeYoung et al., 2020). However, their method only optimizes for two (sufficiency and compactness) out of three desired characteristics of faithful rationales. Our method is also motivated by the work of Yu et al. (2019) who propose a general method for faithful rationalization for text classification problems by directly optimizing for comprehensiveness, in addition to sufficiency and compactness. However, their approach involves learning three different models within a cooperative game-theoretic framework and is not scalable while also requiring the need to tune comprehensiveness through a hyper-parameter.

Lastly, question generation has been previously considered by Lewis and Fan (2019) within the context of generative QA where they model the joint distribution of the question and answer. They show that question generation can increase robustness of the model on adversarial inputs.

8 Conclusion

We proposed a novel and scalable extractive rationalization method for QA tasks using a single Transformer model. By adding a question generation objective to our method, we showed that it is possible to extract rationales that rely on robust input features, thereby improving both faithfulness of the extracted rationales and answer accuracy.

While we showed that learning an evidence extractor by jointly predicting the question and the answer can improve rationale extraction, it is possible to combine the rationales extracted for the question and answer in more novel ways.

References

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. BoolQ: Exploring the surprising difficulty of natural yes/no questions. 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 2924–2936, Minneapolis, Minnesota. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. 2020. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4443–4458, Online. Association for Computational Linguistics.

Gautier Izacard and Edouard Grave. 2020. Leveraging passage retrieval with generative models for open domain question answering. *arXiv preprint arXiv:2007.01282*.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Divyansh Kaushik and Zachary C. Lipton. 2018. How much reading does reading comprehension require? a critical investigation of popular benchmarks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5010–5015, Brussels, Belgium. Association for Computational Linguistics.

Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. 2018. Looking beyond the surface: A challenge set for reading comprehension over multiple sentences. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 252–262, New Orleans, Louisiana. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Eric Lehman, Jay DeYoung, Regina Barzilay, and Byron C. Wallace. 2019. Inferring which medical treatments work from reports of clinical trials. 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 3705–3717, Minneapolis, Minnesota. Association for Computational Linguistics.

Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2016. Rationalizing neural predictions. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 107–117, Austin, Texas. Association for Computational Linguistics.

Mike Lewis and Angela Fan. 2019. Generative question answering: Learning to answer the whole question. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Zachary C Lipton. 2018. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57.

Bhargavi Paranjape, Mandar Joshi, John Thickstun, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020a. An information bottleneck approach for controlling conciseness in rationale extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1938–1952, Online. Association for Computational Linguistics.

Bhargavi Paranjape, Mandar Joshi, John Thickstun, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020b. An information bottleneck approach for controlling conciseness in rationale extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1938–1952, Online. Association for Computational Linguistics.

Judea Pearl. 2009. *Causality*. Cambridge university press.

Danish Pruthi, Bhuwan Dhingra, Graham Neubig, and Zachary C. Lipton. 2020. Weakly- and semi-supervised evidence extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3965–3970, Online. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How much knowledge can you pack into the parameters of a language model? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5418–5426, Online. Association for Computational Linguistics.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

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, Brussels, Belgium. Association for Computational Linguistics.

Mo Yu, Shiyu Chang, Yang Zhang, and Tommi Jaakkola. 2019. Rethinking cooperative rationalization: Introspective extraction and complement control. 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 4094–4103, Hong Kong, China. Association for Computational Linguistics.

A Training details.

We use the Adam optimizer (Kingma and Ba, 2015) with default parameters and learning rate of $1e^{-4}$ with linear decay to train all our models for a maximum of 20000 steps. We perform validation every 500 steps and select the model with the best validation set answer accuracy. All hyperparameters were tuned for answer accuracy on the validation set. We ran all our experiments on machines with 8 32GB GPUs.