Why do you think that? Exploring Faithful Sentence–Level Rationales Without Supervision

Max Glockner and **Ivan Habernal** and **Iryna Gurevych**
Ubiquitous Knowledge Processing Lab (UKP-TUDA)
Department of Computer Science, Technische Universität Darmstadt
https://www.ukp.tu-darmstadt.de

Abstract

Evaluating the trustworthiness of a model's prediction is essential for differentiating between 'right for the right reasons' and 'right for the wrong reasons'. Identifying textual spans that determine the target label, known as faithful *rationales*, usually relies on pipeline approaches or reinforcement learning. However, such methods either require supervision and thus costly annotation of the rationales or employ non-differentiable models. We propose a differentiable training–framework to create models which output faithful rationales on a sentence level, by solely applying supervision on the target task. To achieve this, our model solves the task based on each rationale individually and learns to assign high scores to those which solved the task best. Our evaluation on three different datasets shows competitive results compared to a standard BERT blackbox while exceeding a pipeline counterpart's performance in two cases. We further exploit the transparent decision–making process of these models to prefer selecting the correct rationales by applying direct supervision, thereby boosting the performance on the rationale–level.[1]

1 Introduction

Large pre-trained language models, such as BERT (Devlin et al., 2018) or RoBERTa (Liu et al., 2019b) gain impressive results on a large variety of NLP tasks, including reasoning and inference (Rogers et al., 2020). Despite this success, research shows that their strong performance can rely, to some extent, on dataset–specific artifacts and not necessarily on the ability to solve the underlying task (Gururangan et al., 2018; Schuster et al., 2019; Gardner et al., 2020). Thus, these observations undermine the models' trustworthiness and impede their deployment in situations where 'blindly trusting' the model is deemed irresponsible (Sokol and Flach, 2020). Explainability has thus emerged as an increasingly popular field (Gilpin et al., 2018; Guidotti et al., 2018).

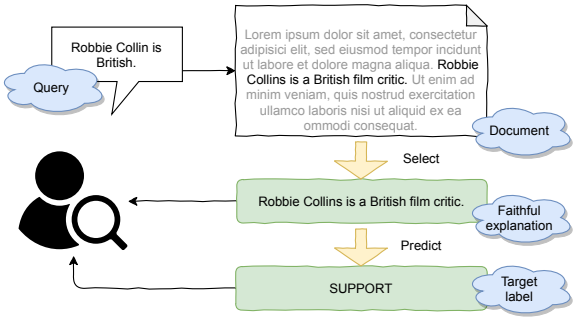

Figure 1: Example of the proposed rationale selecting process on one of the datasets (FEVER): Given a query and a document, our model selects the best rationale and predicts the label solely based on this selection.

We aim at faithful explanations – the identification of the actual reason for *the model's* prediction, which is essential for accountability, fairness, and credibility (Chakraborty et al., 2017; Wu and Mooney, 2019) to evaluate whether a model's prediction is based on the correct evidence. The recently published ERASER benchmark (DeYoung et al., 2020) provides multiple datasets with annotated *rationales*, i.e., parts of the input document, which are essential for correct predictions of the target variable (Zaidan et al., 2007). By contrast to post-hoc techniques to identify relevant input parts such as LIME (Ribeiro et al., 2016) or input reduction (Feng et al., 2018), we focus on models that are faithful by design, in which the selected rationale matches the full underlying evidence used for the prediction.

Existing strategies mostly rely on REINFORCE (Williams, 1992) style learning (Lei et al., 2016; Yu et al., 2019) or on

[1]Code available at https://github.com/UKPLab/emnlp2020-faithful-rationales

Findings of the Association for Computational Linguistics: EMNLP 2020, pages 1080–1095

training two disjoint models (Lehman et al., 2019; DeYoung et al., 2020), in the latter case depending on rationale supervision. This poses critical limitations as rationale annotations are costly to obtain and, in many cases, not available. Additionally, only when the model can select the "best" rationale from the full context we obtain an unbiased indicator for artifacts within a dataset that may influence models without rationale supervision.

In our proposed setup, we turn the hard selection into a differentiable problem by (a) decomposing each document into its residual sentences, and (b) similar to Clark and Gardner (2018) optimize the weighted loss based of each of these candidates. We show that this end–to–end trainable model (see Figure 1) can compete with a standard BERT on two reasoning tasks without rationale–supervision, and even slightly improve upon it, when supervised towards gold rationales. Our quantitative analysis shows how we can exploit these extracted rationales to identify the model's decision boundaries and annotation artifacts of a multi–hop reasoning dataset.

2 Related Work

Understanding the deep neural networks' decisions has gained increasing interest in the research community (DeYoung et al., 2020; Alishahi et al., 2019; Wallace et al., 2019; Jacovi and Goldberg, 2020). Several works are concerned with post–hoc techniques to explain decisions of blackbox models (Ribeiro et al., 2016; Feng et al., 2018; Camburu et al., 2019). Visualizing attention weights has been heavily used, but is known to be insufficient (Jain and Wallace, 2019; Serrano and Smith, 2019). Other works focus on making the models themselves more interpretable via neural module networks (Jiang and Bansal, 2019; Gupta et al., 2020), graph–based networks (Tu et al., 2019; Qiu et al., 2019), pipeline models (Lehman et al., 2019), or by generating textual explanations (Camburu et al., 2018; Rajani et al., 2019; Liu et al., 2019a). Rather than only producing this explanation as additional output, Latcinnik and Berant (2020) base the target prediction on this automatically created hypothesis.

Some approaches jointly use rationales to explain the predictions and boost performance without ensuring faithfulness (Zaidan et al., 2007; Melamud et al., 2019; Strout et al., 2019). Recent work use Gumbel Softmax (Maddison et al., 2016) to identify token–level rationales to avoid using REINFORCE (Bastings et al., 2019; Pfeiffer et al., 2019).

FEVER
Claim
Joan Crawford has had four marriages. (SUPPORTS)
Document
[...] Following a public appearance in 1974 , after which unflattering photographs were published , Crawford withdrew from public life and became increasingly reclusive until her death in 1977 . (R1) Crawford married four times . (R2) Her first three marriages ended in divorce ; the last ended with the death of husband Alfred Steele . Crawford 's relationships with her two older children , Christina and Christopher , were acrimonious . [...]

MultiRC
Question
What are we seeing when we see lightning ?
Answer
The discharge of electrons (TRUE)
Document
[...] Over time the differences increase . (R1) Eventually the electrons are discharged . This is what we see as lightning . You can watch an awesome slow - motion lightning strike below . [...]

Figure 2: While the example from FEVER provides two alternative single-sentence rationales (R1 and R2), the MultiRC example requires considering two sentences at once for a single rationale (R1).

Very recent work (Jain et al., 2020) aims similarly to us, to infer faithful rationales based on its impact on the target prediction without supervision, thereby relying on a dedicated explanation technique to identify rationales and an additional model for the prediction. This work is different in that we (a) rely on the same network weights for rationale selection *and* target prediction, and (b) provide quantitative analysis about the decision criteria of the models on the reasoning tasks.

3 Experimental Setup

3.1 Datasets

We conduct our experiments on three different datasets as provided by ERASER. Specifically, we use FEVER (Thorne et al., 2018), MultiRC (Khashabi et al., 2018), and Movies (Zaidan et al., 2007) as shown in Table 1. We limit ourselves to this sub-set of ERASER, as they require the identification of rationales from multi–sentence documents (as opposed to single sentences). Further, our approach must process the full sample, including the document, within the same minibatch. We do not consider datasets if their documents' size imposes memory issues with pre–trained language

models, as this would require external preprocessing, which is not controlled by the model.

	FEVER	MultiRC	Movies
# Samples	97,957	24,029	1,600
Rationales / Sample	1.0	1.5	8.7
Minimum reasoning–hops			
One	96,702	-	1,597
Two	1,133	17,345	-
Three	73	5,134	2
Four	27	1,547	-
Five+	22	3	-

Table 1: Properties of the datasets (train). In MultiRC rationales are annotated for each *question*. The numbers here reflect counts per (*question*, *answer*) tuple.

FEVER is a large fact–checking dataset based on Wikipedia. Given a claim and a relevant document, the model must either *support* or *refute* the claim[2]. In FEVER, multiple alternative rationales may exist, each of which can be used to refute or support a claim.

MultiRC is a multi–hop–reasoning multiple–choice dataset. It encompasses a variety of genres. Each question is annotated with a single rationale, which always consists of multiple sentences. For each question, an arbitrary number of correct answers exists. Examples for both datasets can be found in Figure 2.

Movies is a sentiment dataset of movie reviews. As opposed to the other two corpora, it (a) does not require reasoning between the document and an additional claim/question, and (b) contains rationale–annotations on a span–level. Though we are primarily interested in sentence–level reasoning tasks, we apply our method to this dataset and map its annotations to sentences.

3.2 Our Model

Task Overview We propose a model that (a) explains its decisions by outputting which input parts are used for the predictions as faithful rationales and (b) achieves performance comparable to a standard blackbox approach. Importantly, the model must be able to select rationales that are useful to solve the target task, without relying on additional supervision. We achieve this by first creating multiple smaller samples for each original sample — each associated with a potential rationale — and solving the task based on each sub–sample individually. Similar to Clark and Gardner (2018), each sub–sample is associated with a learned score. Our model utilizes this score to jointly predict the target and the rationale. Instead of learning these scores via direct supervision (Min et al., 2019), our approach can derive them solely based on how useful each rationale is for solving the target task.

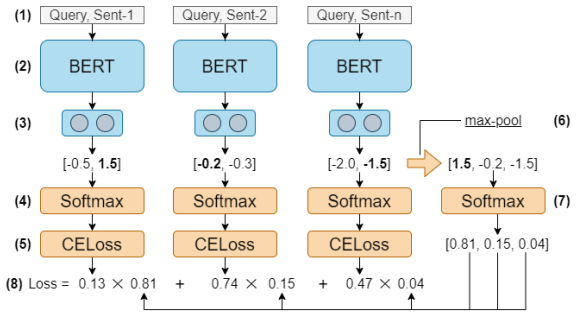

Figure 3: Model architecture. Each sample is split into its sentences (1), each individually encoded via BERT (2) followed by a linear layer (3). The loss for each input part is calculated separately (4,5). The score is computed via max–pooling (6), normalized (7) to compute the weighted loss (8). The input part with the highest score (6) is used for prediction.

Single–Sentence without Rationale Supervision Given a sample, the model must predict the label y based on a query q, i.e., the concatenation of the question and answer (MultiRC) or the claim (FEVER), and a document D. Instead of optimizing the objective given (q, D), we split D into segments and solve the overall task for each segment individually. We opt to split each document into sentences, as a trade-off between capturing enough semantic information within each segment while restricting each candidate's amount of information. Because some samples may be solved without any context (Schuster et al., 2019), we add a *query–only* part, which is associated with no sentence ($\emptyset$). Hence, for each (q_k, D_k) with D_k containing n_k sentences $s_{k,i}$, we create new input samples x_k^{new} with $|x_k^{new}| = n_k + 1$ as

$$x_k^{new} = \Big[(q_k, \emptyset), (q_k, s_{k,1}), (q_k, s_{k,2}), ..., (q_k, s_{k,n})\Big] \quad (1)$$

We use a standard model $\mathbf{m}$ to compute the logits z_k (without softmax) based on all (q_k, $s_{k,i}$) in x_k^{new} within the same minibatch. All experiments use BERT-base-uncased (Devlin et al., 2018) with a linear layer on top of the `[CLS]` token

$$z_k = \mathbf{m}(x_k^{new}); \quad z_k \in \mathbb{R}^{|x_k^{new}| \times t} \quad (2)$$

[2]Note that this task–setup and dataset from DeYoung et al. (2020) differs from the original FEVER (Thorne et al., 2018).

whereas t reflects the number of target labels. Based on z_k we compute $|x_k^{new}|$ losses l_k via softmax and cross–entropy based on each (q_k, $s_{k,i}$) individually. Likewise, $|x_k^{new}|$ different target predictions $\hat{y}_k$ are computed. Not all (q_k, $s_{k,i}$) contain the right information to properly solve the target task. Similar to Clark and Gardner (2018); Min et al. (2019) we rely on confidence scores to identify the best prediction, based on the most relevant rationale. To do so we must (a) compute scalar values $c_{k,i}$ as confidence scores for each (q_k, $s_{k,i}$), and (b) ensure that high scores $c_{k,i}$ are assigned to those input parts, that are most useful from *the model's* perspective. We compute c_k via row–wise max–pooling over z_k as it represents the value of the selected class:

$$c_k = \max(z_k); \quad c_k \in \mathbb{R}^{|x_k^{new}|} \tag{3}$$

The key idea is to multiply these c_k with the losses l_k to compute the overall loss, s.t. high losses will be associated with low confidence and vice-versa. Yet, we cannot merely multiply both terms, as this would allow the model to decrease the loss towards minus infinity only by assigning high negative values to all c_k without optimizing towards the actual label. To overcome this problem and obtain meaningful scores c_k solely based on how useful each rationale is for the target task, we normalize all c_k via softmax to obtain weights $w_{k,i}$ for each (q_k, $s_{k,i}$). As an overall objective, we minimize the weighted sum of losses using these weights:

$$w_{k,i} = \frac{e^{\frac{c_{k,i}}{\tau}}}{\sum_{j=1}^{|c_k|} e^{\frac{c_{k,j}}{\tau}}}; \quad \text{argmin}_\theta \left(\sum_{k=1}^{|x|} \sum_{i=1}^{|w_k|} w_{k,i} l_{k,i} \right) \tag{4}$$

The rationale behind this is threefold: A right prediction, i.e., a low loss $l_{k,i}$, is only possible for informative sentences *from the model's perspective*. First, by allowing the model to distribute the weights for the losses amongst all candidates, it can neglect non–informative sentences when learning to assign *low* values (to high losses). Second, by normalizing these scores, it cannot ignore all sentences, but must assign comparatively *higher* scores to at least one (q_k, $s_{k,i}$). Hence, to minimize the overall loss, high values must be assigned to the best suited (q_k, $s_{k,i}$), i.e., with the lowest (expected) loss. Finally, by deriving these scores directly from the predicted class, the same function for prediction and selection is used and optimized. The hyperparameter τ is the temperature of softmax, controlling the distribution of the softmax function. Higher values for τ result in a softer distribution, i.e., the loss is more evenly distributed amongst rationale candidates. Lower values result in a more hardened distribution, i.e., the model focuses quicker on one selected rationale. For both, prediction and training, all rationales are always considered. The process is visually exemplified in Figure 3 and, for the most part (steps 2–5), resembles a standard setup.

Prediction For predictions, we select the sentence with the highest confidence from all sentences as the rationale $\hat{r}$, and the prediction based on $\hat{r}$ as the target $\hat{y}$:

$$\hat{r} = \text{argmax}(w); \;\; \hat{y} = \text{argmax}(z_{\hat{r}}) \tag{5}$$

Though the rationale is *faithful* on a sentence–level, we note that it does not indicate whether *all* information of $\hat{r}$ is relevant to the model.

Rationale supervision We believe that rationales without supervision provide more trustworthy explanations. They are not affected by an additional objective and solely are selected if they are useful for the target task. Nevertheless, we experimentally show how rationale–supervision can be applied by jointly (Yin and Roth, 2018) supervising on the target and rationale. To compute the rationale–loss as an additional objective, we treat slightly adapted confidence values c_k^* as a multi–label problem via a sigmoid layer and binary cross–entropy loss.

$$c^*_{k,i} = \begin{cases} \max(z_{k,i}) & \text{if } x^{new}_{k,i} \text{ is not a gold–rationale.} \\ z_{k,i,y} & \text{if } x^{new}_{k,i} \text{ is a gold–rationale.} \end{cases} \tag{6}$$

This ensures that the correct class's confidence is increased even if the model (currently) predicts the wrong class.

Multiple Sentences Due to the memory consumption, encoding all (ordered) permutations of sentences up to a certain length through BERT is infeasible. To allow the model to select multiple sentences, for each permutation up to a length h, their representation is computed by max-pooling over the `[CLS]` token embeddings of its sentences. We experiment with up to two sentences.

4 Results

All experiments use AllenNLP (Gardner et al., 2018) and BERT-base-uncased (Devlin et al., 2018) as provided by Wolf et al. (2019). We manually

	Target		Rationale			Target & Rationale	
	F1a	**Acc.**	**P**	**R**	**F1**	**Acc. Full**	**Acc. Part**
FEVER							
Majority	33.2	49.6	-	-	-	-	-
BERT Blackbox	90.2 ±0.4	90.2 ±0.4	-	-	-	-	-
Pipeline $\mathbb{S}$ (DeYoung et al., 2020)	87.7	87.8	88.3	87.7	88.0	78.1	79.0
Single-Sentence Selecting $\mathbb{U}$	90.1 ±0.8	90.1 ±0.8	80.0 ±4.3	79.4 ±4.3	79.7 ±4.3	72.2 ±4.5	73.2 ±4.5
Single-Sentence Selecting $\mathbb{S}$	90.7 ±0.7	90.7 ±0.7	**92.3** ±0.1	**91.6** ±0.1	**91.9** ±0.1	**83.9** ±0.4	**84.9** ±0.4
Two-Sentence Selecting $\mathbb{U}$	90.6 ±0.2	90.6 ±0.2	84.0 ±0.9	83.5 ±1.0	83.8 ±0.9	76.5 ±1.0	77.7 ±1.0
Two-Sentence Selecting $\mathbb{S}$	**91.1** ±0.5	**91.1** ±0.5	91.7 ±0.5	91.1 ±0.5	91.4 ±0.5	**83.9** ±0.8	84.8 ±0.7
MultiRC							
Majority	36.3	57.2	-	-	-	-	-
BERT Blackbox	67.3 ±1.3	67.7 ±1.6	-	-	-	-	-
Pipeline $\mathbb{S}$ (DeYoung et al., 2020)	63.3	65.0	66.7	30.2	41.6	0.0	44.8
Single-Sentence Selecting $\mathbb{U}$	65.2 ±3.5	66.8 ±3.8	34.6 ±24.5	15.5 ±10.9	21.4 ±15.1	0.0 ±0.0	23.3 ±16.6
Single-Sentence Selecting $\mathbb{S}$	**67.4** ±0.4	**69.1** ±1.3	**74.3** ±1.1	33.5 ±0.5	46.1 ±0.6	0.0 ±0.0	54.0 ±0.9
Two-Sentence Selecting $\mathbb{U}$	66.7 ±2.7	67.7 ±3.0	44.4 ±11.0	19.9 ±5.0	27.5 ±6.9	0.1 ±0.0	31.2 ±7.4
Two-Sentence Selecting $\mathbb{S}$	65.5 ±3.6	67.7 ±1.5	65.8 ±0.2	**42.3** ±3.9	**51.4** ±2.8	**7.1** ±2.6	**55.7** ±1.2
Movies							
Majority	33.3	50.0	-	-	-	-	-
BERT Blackbox	**90.1** ±0.3	**90.1** ±0.3	-	-	-	-	-
Pipeline $\mathbb{S}$ (DeYoung et al., 2020)	86.0	86.0	**87.9**	60.5	71.7	40.7	**82.4**
Single-Sentence $\mathbb{U}$	53.3 ±14.1	60.6 ±7.4	50.1 ±13.1	34.0 ±8.5	40.4 ±10.1	18.4 ±7.3	37.4 ±13.8
Single-Sentence $\mathbb{S}$	85.6 ±3.6	85.8 ±3.5	86.9 ±2.5	**62.4** ±0.1	**72.6** ±0.9	**43.9** ±0.6	81.4 ±3.9

Table 2: Mean performance and standard deviation for all models. $\mathbb{U}$ represents models without supervision on the rationale, $\mathbb{S}$ indicates supervision is applied on the rationale. The first two columns measure the performance on the target task using macro–averaged F1 and accuracy. The next three columns specify **P**recision, **R**ecall and **F1** of the rationales on a sentence–level. The last two columns jointly show the performance based on a correct rationale *and* target. Majority is only computed for the target–task performance.

tune hyper-parameters for standard BERT baseline models and the sentence–selecting models, and show results in Table 2. We report results for the best configurations using three different seeds. We additionally report results of the BERT–to–BERT pipeline models from ERASER, which are based on the implementation of Lehman et al. (2019).

Metrics As opposed to DeYoung et al. (2020) we choose sentences as the lexical unit for rationales. We report precision, recall, and F1 for the rationales rather than token–level IOU, to avoid that the length of sentences impacts the metrics[3]. As we are interested to understand whether a model makes the right prediction for the right reasons, we focus on *sufficiency* of selected rationales rather than comprehensiveness: The claim of FEVER in Figure 2 shows two valid rationales. Only one of these is required to support the claim. To compute precision, recall, and F1 w.r.t. sufficiency, we, therefore, compute these metrics based on the single, most similar[4] gold–rationale when evaluating any of the models. We additionally report the joint accuracy of the target task and the rationale. Here we consider a prediction correct for the right reason, when it correctly predicts the target and *all* sentences of one gold–rationale (Acc. Full). A weaker measure (Acc. Part) only requires the intersection of the selected sentences and one gold–rationale to be non–empty. As multi–hop classification tasks tend to be easy to "trick" (Chen and Durrett, 2019), this joint evaluation with the underlying evidence gives a better impression of the performance on the task itself.

Observations The *Target* columns in Table 2 show that our models can compete with the standard BERT on both reasoning tasks FEVER and MultiRC. This is especially surprising for single–sentence models on the multi–hop reasoning task MultiRC. We find that the single–sentence model $\mathbb{U}$ is more sensitive towards seeds, yielding in a slightly lower overall performance and higher variance on MultiRC (see Appendix B). We believe this is because, given an unfortunate initialization, the model can focus on arbitrary features to quickly on this challenging dataset. Applying rationale supervision helps to stabilize this by improving the selected rationales rather than generally reaching higher target performances. The BERT–to–BERT pipeline makes its prediction based on the best *sin-*

[3]To simplify comparisons with future work, we report the original ERASER metrics in Appendix A.

[4]Determined by highest F1 on the sentence–level.

gle sentence and can only fairly be compared with the single–sentence selecting models. The unsupervised approach is far behind all other models on the movies dataset, which we partly attribute to the small training data combined with the much larger document size. Primarily, however, we find (see Section §5.1) that by design, our approach is unsuitable for this kind of data, which due to its discussing nature, contains evidence for both labels within the same document.

The closest measure for "right for the right reasons" is represented by Acc. Full. Yet, it can only measure whether the prediction is based on the correct rationale on a sentence level, whereas it may still solely rely on certain contained words. Assuming comprehensive rationale annotations[5], the opposite can be said, i.e., 92.9% of MultiRC are not classified correctly for the right reasons. Note that both, the single–sentence models and the BERT–to–BERT pipeline, are bound to reach an 0% for Acc. Full on MultiRC, since they can only select a single sentence as the rationale.

5 Analysis

Leveraging the information about the used rationales, we closer analyze decision criteria for FEVER and MultiRC, and why our method performed poorly on Movies. Further, except for the two–sentence $\mathbb{S}$ models on MultiRC, no other model selects two sentences as a rationale in more than 1.3%. We partly attribute this to the less–than–optimal aggregation via max–pooling. As these are only selected due to the additional supervision, not for the utility to solve the overall task[6], we focus on single–sentence models.

5.1 Poor Performance on Movies Dataset

Without rationale supervision, our approach by far lacks behind its counterparts. To better understand the reason for this performance gap, we analyze the underlying data and the predictions. We find that our models $\mathbb{U}$ reach an average recall of 0.93 and 0.32 for `NEG` and `POS` respectively on the dev set — despite the balanced training data. We emphasize that this is due to a very different nature of the data, compared to FEVER and MultiRC: Rather than all sentences within a document containing the same sentiment, they usually discuss pro and cons, and hence contain evidence for the gold label, as well as the opposite label. An extract of such a document can be seen in Figure 4 and two full examples in Appendix E. During prediction, even for humans, it is impossible to predict the correct overall sentiment based on isolated, out of context sentences of opposing stances. An additional problem arises during training in our setup: For the presented example, the model must either learn to either predict the label `NEG` even for sentences with clearly (only) positive indicators, or learn to reduce their confidence values c_k to mitigate their impact. Either way, this naturally compromises its ability to detect the opposite sentiment. This discussion–based nature of Movies significantly differs from MultiRC and Fever. In the latter case, each document only contains evidence for *or* against a claim, not both. In this case, the model must not learn contradicting patterns and only lower the confidence for irrelevant sentences, consistent with both labels. Both, the pipeline and the model $\mathbb{S}$, show that by guiding the model towards gold rationales, it can detect sentences for the overall movie sentiment. Without this guiding, however, our approach seems not suitable for such tasks.

(35) the scenes between nick and danny are very good, and i actually got a feel for their characters; a bond forms between them that holds parts of the film together.
(36) chow and wahlberg are both good actors; chow is a pro, and can do this kind of stuff in his sleep.
(37) wahlberg seems less at home in this atmosphere, but he's still fun to watch.
(38) i also liked the subplot involving danny 's father; brian cox's performance is powerful, and his character makes a compelling moral compass for danny.
(39) but the film ultimately fails, mostly at the hands of insane incoherence and overly - familiar action scenes.

Figure 4: An extract of a movie review with an overall **negative sentiment**. Sentence 35–38 in isolation contain positive sentiment, whereas sentence 39 shows strong negative sentiment. Only the underlined span in line 39 constitutes a gold rationale and represents the overall sentiment.

5.2 Learning curves

We investigate the impact of the amount of available training data for the three different models blackbox, model $\mathbb{S}$, and $\mathbb{U}$. To limit the data's impact, we create three random subsets of the training data of different sizes and report the average performance of each of the models on these subsets in Figure 5. All three models show similar trends across all training sizes for MultiRC. On FEVER,

[5]FEVER does not provide comprehensive rationale–annotations.

[6]We show supporting analysis for this in Appendix D.

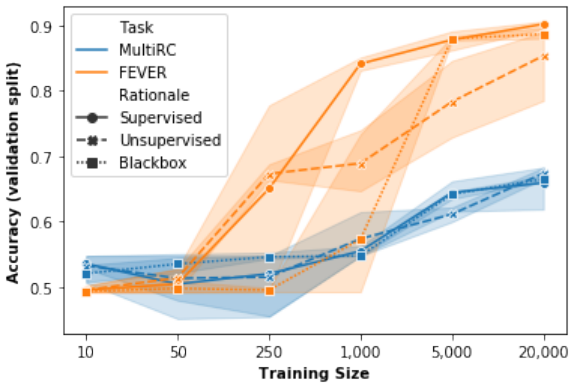

Figure 5: Accuracy (validation split) of BERT Blackbox and single–sentence models by training sizes.

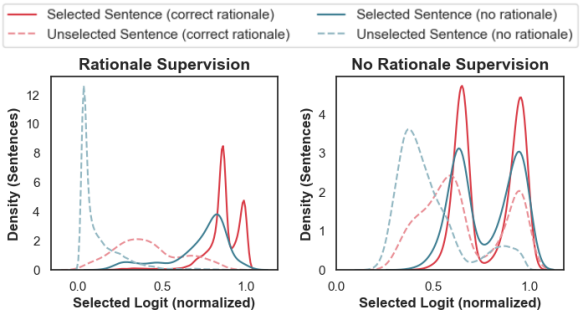

Figure 6: KDE plots of the single–sentence models ($\mathbb{S}$ left) and ($\mathbb{U}$ right) for FEVER, showing the relative frequency for each category individually based on globally normalized logits $z_{k,i}$ of the selected label.

the rationale–supervision offers an additional boost in scenarios with little data. Without rationale–supervision, it tends to require more data to reach its peak performance.

5.3 Model decisions on FEVER

Both (best) single–sentence models $\mathbb{U}$ and $\mathbb{S}$ perform very strong and predict the same label in 93.8% of all cases, from which they select the same rationale in 86%. We, therefore, focus on how supervision affects the model internally. Specifically, we exploit the fact that relevance and prediction are jointly encoded and optimized within the same logits $z_{k,i}$. In Figure 6 we compare these $z_{k,i}$ from a global perspective after normalizing them using min–max–normalization. Applying rationale–supervision leads to more decisive predictions, as the vast majority of unselected sentences scores close to the global minimum, whereas selected sentences have scores close to the maximum. Invalid selected rationales tend to be shifted slightly more towards the lower end than selected correct rationales. This looks very different for model $\mathbb{U}$. Most importantly, a non–trivial amount of unselected sentences reached scores very close to the global maximum.

	BERT	Single $\mathbb{U}$	Single $\mathbb{S}$
F1 (SUPPORT)	67.8 ±0.6	68.1 ±0.4	**71.1** ±1.9
F1 (REFUTE)	61.3 ±2.4	62.1 ±0.7	**64.4** ±2.4
F1a	64.5 ±1.5	65.1 ±0.2	**67.8** ±3.7

Table 3: Evaluation of BERT and single–sentence selecting models on the symmetric FEVER testset (Schuster et al., 2019) (717 samples)

Does it learn semantically better decision criteria with supervison? A possible reason why such high values occur for unchosen sentences is that the selected rationale is not substantial for a correct target prediction. Schuster et al. (2019) identify n–grams within claims that highly correlate with certain classes. By adding new evidence and claims for each of their selected claims they design a symmetric test-set, which cannot be solved using such artifacts. Intuitively, similar to Stacey et al. (2020), applying rationale–supervision (model $\mathbb{S}$) forces the model to learn — based on the rationale — high and low values for the same claim, i.e. containing the same artifacts. It should therefore be more sensitive for the context and not rely on claim–only features. We show the performance on this symmetric test set in Table 3. Despite a small improvement, it still lacks far behind the performance on FEVER. Even the model $\mathbb{U}$ rarely selects the claim–only as the rationale, suggesting that at least partially, additional context helps to solve the task properly. Yet, it shows that smaller lexical units than sentences as a rationale may be beneficial in such cases.

5.4 Model decisions on MultiRC

What is the impact of rationale supervision? The ceiling performance on the target task remains the same, even with rationale–supervision. We analyze the validity of the selected rationales on the validation split to shed light on (a) how the model can achieve a strong performance, and (b) how rationale supervision affects the model. For simplicity, we select the best performing single–sentence models and group the predictions by the gold and predicted target label in Table 4. The model $\mathbb{U}$ results show that evidence of positive samples is more likely to get selected. While the correctly predicted positive samples mostly rely on gold *evidence for* the answer, for correctly pre-

	T-T	T-F	F-T	F-F
$\mathbb{U}$ Rationale Prec.	79.4	62.3	45.9	36.2
$\mathbb{S}$ Rationale Prec.	86.5	78.2	52.4	80.3
Δ-Rationale Prec.	**+7.1**	**+15.9**	**+6.5**	**+44.1**

Table 4: Precision of the selected rationale by the best single–sentence models on MultiRC, grouped by the (Gold - Predicted) labels **T**rue and **F**alse.

dicted negative samples, the *absence of supporting evidence* seems sufficient, rather than explicit evidence against it. Note that none of these "evidence" is truly sufficient, as multiple sentences are technically required. To see whether this behavior is due to our training method or helpful for the underlying data, we re–evaluate the best performing BERT on the validation set and exclude all gold–rationales from the documents. The results show a recall of 28.4 (True) and 81.8 (False)[7], suggesting a similar behavior. Hence, the major benefit from rationale–supervision is to predict the label `False` based on the correct sentence, which is not required to solve the overall task. To limit this property of future datasets, we believe it is important to add unanswerable instances, as done for instance by Thorne et al. (2018) or Rajpurkar et al. (2018).

What kind of sentences are selected as a rationale?

We jointly look at the selected sentences with the target prediction of both models $\mathbb{U}$ and $\mathbb{S}$ and observe a high correlation with word overlap of the question and the answer. Figure 7 shows KDE plots of the selected sentences based on the percentage of non–stopwords[8] of the question and answer respectively, that are also contained within the selected sentence. We make multiple observations: Positive predictions mostly depend on a high overlap with the answer. The overlap with the question has a lower priority. Especially for the model $\mathbb{S}$, a clear decision boundary between rationales for both labels can be seen based on the lexical overlap. Interestingly, also Yadav et al. (2019), to a large part, rely on similar lexical features for their unsupervised detection of justification sentences on MultiRC. In line with the previous section, rationale supervision only has a limited impact on positive predictions. A significant difference is shown

[7]Compared to 54.5 (True) and 79.3 (False)

[8]We use spaCy to exclude punctuation and stopwords and seaborn (Waskom et al., 2017) with default parameters for plotting.

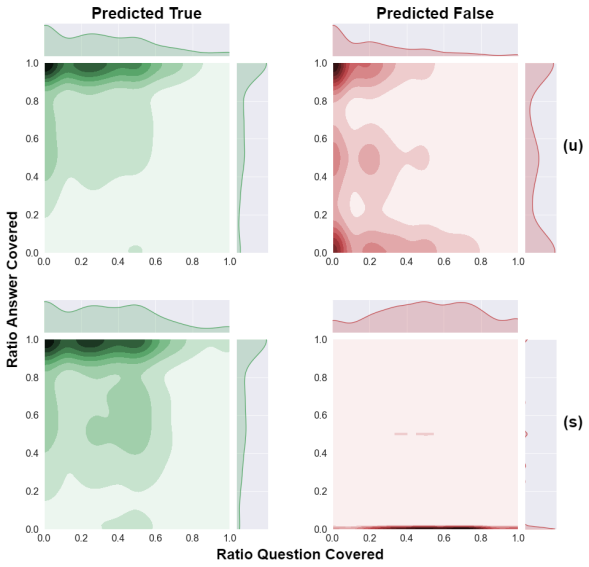

Figure 7: KDE plots for word overlaps between Question/Answer and the selected rationale of single–sentence models on MultiRC with (bottom) and without (top) rationale supervision..

for the negative predictions. Whereas model $\mathbb{U}$ tends to select rationales for both labels based on similar criteria, the selected rationales for samples predicted `False` by model $\mathbb{S}$ almost entirely have lexical overlaps with the question only. This intuitively makes sense, as the same rationales are valid for each question. Negative rationales should therefore be relevant for the question, not for the answer. We show some examples in Appendix C.

Are single sentences sufficient for MultiRC?

It has been shown that noisy detection of evidence can already improve the performance on MultiRC (Wang et al., 2019), yet this should not be possible via single sentences. To see whether BERT exploits such biases, we follow Gururangan et al. (2018) and identify samples within the test–set that are solvable using a single–hop only, i.e., these which the single–sentence $\mathbb{U}$ model classified correctly. To limit the impact of lucky guesses, we group samples by the number of these models that could solve them in Table 5. As pointed out in Section 4, one of our single–sentence models $\mathbb{U}$ on MultiRC performed poorly due to its seed sensitivity . To exclude impacts from this specific model and group the test–split by meaningful criteria, we retrain BERT blackbox and model $\mathbb{U}$ with a new random seed, reaching an F1a score of 66.3 and 67.6 respectively on the test set. We select the best three seeds of both model types for splitting the data (model $\mathbb{U}$) and evaluation (BERT blackbox).

	3/3	2/3	1/3	0/3
Size	2,314	1,114	779	641
Logistic Regression (F1)				
True	70.2	61.4	59.6	48.6
False	69.2	29.0	15.5	9.1
F1a	69.7	45.2	37.5	28.8
BERT Blackbox (F1)				
True	91.7±0.9	62.3±3.5	42.3±5.0	14.8±3.5
False	94.9±0.6	69.1±1.5	45.5±5.9	8.7±4.3
F1a	93.3±0.8	65.7±1.0	43.9±1.7	11.8±1.4
ΔF1a	**+25.3**±1.1	**–2.3**±0.8	**–24.1** ±1.6	**–56.2** ±1.4

Table 5: Average performance of BERT models based on subsets of the test–split that can be solved using a single sentence, compared with a lexical overlap logistic regression. ΔF1a measures the difference w.r.t. the performance on the full test set. Columns indicate how many single–sentence models $\mathbb{U}$ could solve each contained instance correctly.

Lexical Overlap Logistic Regression Additionally, we mimic our observations with the high lexical overlap using a simple logistic regression. We calculate a rationale score $r = w_q q_s + w_a a_s$ for each sentence s, whereas q_s and a_s represent the absolute/relative word overlap of the sentence with the question and answer respectively. For each sample, the sentence with the highest r is selected as a rationale (shorter sentences are preferred as a tie–breaker) and used to train a logistic regression (LR), breaking down the multi–hop reasoning task to two digits based on a single sentence. We run a grid–search with different values for w_q and w_a and select the model with the highest F1a score of 63.5 on the validation set (F1a score of 58.1 on the test set), using absolute word overlaps, $w_q = 0.4$ and $w_a = 1.0$.

Results The performances are shown in Table 5. BERT performs strongly on samples that can be solved using a single sentence while struggling with the same instances as model $\mathbb{U}$. Further, a simple logistic regression shows a similar trend. On the easiest (and largest) part it even exceeds the performance of the full test–set of any BERT model. The results suggest that high performance does not indicate successful multi–hop reasoning[9].

6 Discussion

Limitations From a technical perspective, a limitation is memory consumption, as the model must process all rationale candidates of the same instance within the same minibatch. Though single–sentence rationale can be processed, encoding all combinations of multiple sentences via BERT is problematic. Future work could investigate better sampling strategies or a greedy breadth search to reduce the number of candidates. Another limitation is the inability of coreference resolution between different sentences and the consideration of the context in general. Solving this is non–trivial, as we essentially buy faithfulness by explicitly omitting all other information than the selected sentence(s). While this does not seem crucial in the evaluated datasets, it poses potential dangers for malicious attacks, most importantly, when considering the permutations of multiple sentences. Therefore, we recommend to always show the identified evidence in context when using our approach in the real world.

Conclusion We proposed a conceptually simple approach to allow models to extract faithful rationales, which can compete with standard BERT on two reasoning tasks without supervision and even improve the overall performance, when supervising on the rationale. We showed that by outputting faithful rationales, it is possible to not only compare models based on the target performance alone, but also quantify how well even those correct predictions are based on the correct evidence. Our analysis showed that exploiting this knowledge about the selected rationales helps shed light on the models' the decision mechanism for debugging purposes and on the underlying data.

Acknowledgments

We thank the anonymous reviewers for their constructive feedback. This research work has been funded by the German Federal Ministry of Education and Research and the Hessen State Ministry for Higher Education, Research and the Arts within their joint support of the National Research Center for Applied Cybersecurity ATHENE, and the"Data Analytics for the Humanities" grant by the Hessian Ministry of Higher Education, Research, Science and the Arts. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan X Pascal GPU and the Titan Xp Pascal GPU used for this research.

References

Afra Alishahi, Grzegorz Chrupała, and Tal Linzen. 2019. Analyzing and interpreting neural networks

[9]This is not the official, hidden test–set of MultiRC.

for NLP: A report on the first BlackboxNLP workshop. *Natural Language Engineering*, 25(4):543–557.

Joost Bastings, Wilker Aziz, and Ivan Titov. 2019. Interpretable Neural Predictions with Differentiable Binary Variables". In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2963–2977, Florence, Italy. Association for Computational Linguistics.

Oana-Maria Camburu, Eleonora Giunchiglia, Jakob Foerster, Thomas Lukasiewicz, and Phil Blunsom. 2019. Can I Trust the Explainer? Verifying Post-hoc Explanatory Methods. In *NeurIPS 2019 Workshop on Safety and Robustness in Decision Making*.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-SNLI: Natural Language Inference with Natural Language Explanations. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 9539–9549. Curran Associates, Inc.

Supriyo Chakraborty, Richard Tomsett, Ramya Raghavendra, Daniel Harborne, Moustafa Alzantot, Federico Cerutti, Mani Srivastava, Alun Preece, Simon Julier, Raghuveer M Rao, et al. 2017. Interpretability of deep learning models: A survey of results. In *2017 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*, pages 1–6. IEEE.

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: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4026–4032, Minneapolis, Minnesota. Association for Computational Linguistics.

Christopher Clark and Matt Gardner. 2018. Simple and Effective Multi-Paragraph Reading Comprehension. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 845–855, Melbourne, Australia. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805*.

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.

Shi Feng, Eric Wallace, Alvin Grissom II, Mohit Iyyer, Pedro Rodriguez, and Jordan Boyd-Graber. 2018. Pathologies of Neural Models Make Interpretations Difficult. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3719–3728, Brussels, Belgium. Association for Computational Linguistics.

Matt Gardner, Yoav Artzi, Victoria Basmova, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, et al. 2020. Evaluating NLP Models via Contrast Sets. *arXiv preprint arXiv:2004.02709*.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. 2018. AllenNLP: A Deep Semantic Natural Language Processing Platform. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pages 1–6, Melbourne, Australia. Association for Computational Linguistics.

L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal. 2018. Explaining Explanations: An Overview of Interpretability of Machine Learning. In *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 80–89.

Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2018. A Survey of Methods for Explaining Black Box Models. *ACM Comput. Surv.*, 51(5).

Nitish Gupta, Kevin Lin, Dan Roth, Sameer Singh, and Matt Gardner. 2020. Neural Module Networks for Reasoning over Text. In *International Conference on Learning Representations (ICLR)*.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation Artifacts in Natural Language Inference Data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Alon Jacovi and Yoav Goldberg. 2020. Towards Faithfully Interpretable NLP Systems: How Should We Define and Evaluate Faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4198–4205, Online. Association for Computational Linguistics.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

Sarthak Jain, Sarah Wiegreffe, Yuval Pinter, and Byron C. Wallace. 2020. Learning to Faithfully Rationalize by Construction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4459–4473, Online. Association for Computational Linguistics.

Yichen Jiang and Mohit Bansal. 2019. Self-Assembling Modular Networks for Interpretable Multi-Hop Reasoning. 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 4473–4483, Hong Kong, China. 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.

Veronica Latcinnik and Jonathan Berant. 2020. Explaining Question Answering Models through Text Generation. *arXiv preprint arXiv:2004.05569*.

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.

Hui Liu, Qingyu Yin, and William Yang Wang. 2019a. Towards Explainable NLP: A Generative Explanation Framework for Text Classification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5570–5581, Florence, Italy. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*.

Chris J Maddison, Andriy Mnih, and Yee Whye Teh. 2016. The concrete distribution: A continuous relaxation of discrete random variables. *arXiv preprint arXiv:1611.00712*.

Oren Melamud, Mihaela Bornea, and Ken Barker. 2019. Combining Unsupervised Pre-training and Annotator Rationales to Improve Low-shot Text Classification. 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 3884–3893, Hong Kong, China. Association for Computational Linguistics.

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, Florence, Italy. Association for Computational Linguistics.

Jonas Pfeiffer, Aishwarya Kamath, Iryna Gurevych, and Sebastian Ruder. 2019. What do Deep Networks Like to Read? *arXiv preprint arXiv:1909.04547*.

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, Florence, Italy. Association for Computational Linguistics.

Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Explain Yourself! Leveraging Language Models for Commonsense Reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4932–4942, Florence, Italy. Association for Computational Linguistics.

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, Melbourne, Australia. Association for Computational Linguistics.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1135–1144, New York, NY, USA. Association for Computing Machinery.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A Primer in BERTology: What we know about how BERT works. *arXiv preprint arXiv:2002.12327*.

Tal Schuster, Darsh Shah, Yun Jie Serene Yeo, Daniel Roberto Filizzola Ortiz, Enrico Santus, and Regina Barzilay. 2019. Towards Debiasing Fact Verification 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 3410–3416, Hong Kong, China. Association for Computational Linguistics.

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, Florence, Italy. Association for Computational Linguistics.

Kacper Sokol and Peter Flach. 2020. Explainability fact sheets: A Framework for Systematic Assessment of Explainable Approaches. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 56–67, Barcelona, Spain. ACM.

Joe Stacey, Pasquale Minervini, Haim Dubossarsky, Sebastian Riedel, and Tim Rocktäschel. 2020. There is Strength in Numbers: Avoiding the Hypothesis-Only Bias in Natural Language Inference via Ensemble Adversarial Training. *arXiv preprint arXiv:2004.07790*.

Julia Strout, Ye Zhang, and Raymond Mooney. 2019. Do Human Rationales Improve Machine Explanations? In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 56–62, Florence, Italy. 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.

Ming Tu, Kevin Huang, Guangtao Wang, Jing Huang, Xiaodong He, and Bowen Zhou. 2019. Select, Answer and Explain: Interpretable Multi-hop Reading Comprehension over Multiple Documents. *arXiv preprint arXiv:1911.00484*.

Eric Wallace, Jens Tuyls, Junlin Wang, Sanjay Subramanian, Matt Gardner, and Sameer Singh. 2019. AllenNLP Interpret: A Framework for Explaining Predictions of NLP 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): System Demonstrations*, pages 7–12, Hong Kong, China. Association for Computational Linguistics.

Hai Wang, Dian Yu, Kai Sun, Jianshu Chen, Dong Yu, David McAllester, and Dan Roth. 2019. Evidence Sentence Extraction for Machine Reading Comprehension. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 696–707, Hong Kong, China. Association for Computational Linguistics.

Michael Waskom, Olga Botvinnik, Drew O'Kane, Paul Hobson, Saulius Lukauskas, David C Gemperline, Tom Augspurger, Yaroslav Halchenko, John B. Cole, Jordi Warmenhoven, Julian de Ruiter, Cameron Pye, Stephan Hoyer, Jake Vanderplas, Santi Villalba, Gero Kunter, Eric Quintero, Pete Bachant, Marcel Martin, Kyle Meyer, Alistair Miles, Yoav Ram, Tal Yarkoni, Mike Lee Williams, Constantine Evans, Clark Fitzgerald, Brian, Chris Fonnesbeck, Antony Lee, and Adel Qalieh. 2017. mwaskom/seaborn: v0.8.1 (september 2017).

Ronald J Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *ArXiv*, abs/1910.03771.

Jialin Wu and Raymond Mooney. 2019. Faithful Multimodal Explanation for Visual Question Answering. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 103–112, Florence, Italy. Association for Computational Linguistics.

Vikas Yadav, Steven Bethard, and Mihai Surdeanu. 2019. Quick and (not so) Dirty: Unsupervised Selection of Justification Sentences for Multi-hop Question Answering. 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 2578–2589, Hong Kong, China. Association for Computational Linguistics.

Wenpeng Yin and Dan Roth. 2018. TwoWingOS: A Two-Wing Optimization Strategy for Evidential Claim Verification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 105–114, 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 4085–4094, Hong Kong, China. Association for Computational Linguistics.

Omar Zaidan, Jason Eisner, and Christine Piatko. 2007. Using "Annotator Rationales" to Improve Machine Learning for Text Categorization. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 260–267, Rochester, New York. Association for Computational Linguistics.

A ERASER Metrics

	F1a	IOU F1	Token F1
FEVER			
Lei et al. $\mathbb{U}$	71.8	0.0	0.0
Lei et al. $\mathbb{S}$	71.9	21.8	23.4
DeYoung et al. $\mathbb{S}$	87.7	83.5	81.2
Single–Sentence $\mathbb{U}$	90.1 ±0.8	75.6 ±4.0	73.7 ±3.9
Single–Sentence $\mathbb{S}$	90.7 ±0.7	**87.3** ±0.1	**85.1** ±0.1
Two–Sentence $\mathbb{U}$	90.6 ±0.2	79.5 ±0.9	77.5 ±0.8
Two–Sentence $\mathbb{S}$	**91.1** ±0.5	86.8 ±0.5	84.6 ±0.5
MultiRC			
Lei et al. $\mathbb{U}$	64.8	0.0	0.0
Lei et al. $\mathbb{S}$	65.5	27.1	45.6
DeYoung et al. $\mathbb{S}$	63.3	41.6	41.2
Single–Sentence $\mathbb{U}$	65.2 ±3.5	21.4 ±15.1	20.9 ±14.8
Single–Sentence $\mathbb{S}$	**67.4** ±0.4	46.1 ±0.6	45.0 ±0.7
Two–Sentence $\mathbb{U}$	66.7 ±2.7	27.5 ±6.9	27.7 ±6.7
Two–Sentence $\mathbb{S}$	65.5 ±3.6	**51.4** ±2.8	**49.0** ±2.6
Movies			
Lei et al. $\mathbb{U}$	**92.0**	1.2	32.2
Lei et al. $\mathbb{S}$	91.4	**12.4**	**28.5**
DeYoung et al. $\mathbb{S}$	86.0	7.5	14.5
Single–Sentence $\mathbb{U}$	53.3 ±14.1	3.2 ±1.3	7.8 ±2.5
Single–Sentence $\mathbb{S}$	85.6 ±3.6	7.0 ±0.2	15.3 ±0.3

Table 6: Results on the original ERASER metrics together with their reported performance using the REINFORCE approach by Lei et al. (2016) and the BERT–to–BERT pipeline by DeYoung et al. (2020).

B MultiRC Sensitivity to Seeds

Model	F1a	Acc	Rat. P.	Acc. Part
Single–Sent–1 $\mathbb{U}$	69.4	69.9	53.1	37.1
Single–Sent–2 $\mathbb{U}$	60.9	61.3	0.0	0.0
Single–Sent–3 $\mathbb{U}$	65.5	69.0	50.7	32.8
Blackbox–1	67.9	69.0	-	-
Blackbox–2	68.6	68.4	-	-
Blackbox–3	65.4	65.5	-	-

Table 7: Performance of BERT blackbox models and Single–Sentence $\mathbb{U}$ models across different seeds on MultiRC.

C Examples for MultiRC with and without supervision

We show representative samples for both gold labels with the same and distinct target predictions. In cases where only one model is correct, `True` labelled samples are mostly classified correctly by $\mathbb{U}$ (82.7%), `False`-labelled samples by $\mathbb{S}$ (81.5%). We decide whether to show distinct or same rationales, depending on the majority of cases within each of these categories.

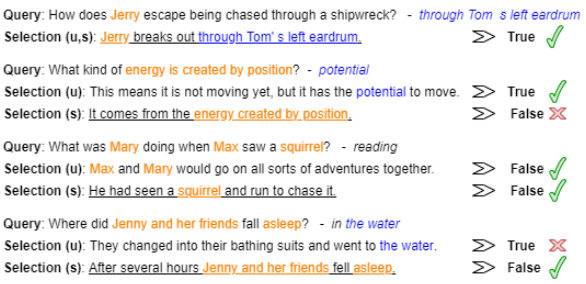

Figure 8: Examples from MultiRC with selected rationales and their prediction for the single-sentence model with $\mathbb{S}$ and without $\mathbb{U}$ rationale–supervision. Underlined sentences are part of the gold–rationale, word–overlaps are highlighted with colors.

D Two-Sentence Models on MultiRC with and without supervision

% Same Prediction	Prediction both sents False	Prediction both sents True
Sentence (Shared)	79.4%	96.8%
Sentence (New)	99.5%	51.3%

Table 8: Change of target prediction based on single sentences of model $\mathbb{S}$, when identifying two sentences as rationale. Columns indicate the classification based on the identified rationale. Rows show how many of these instances are still classified the same, when only using the same single sentence as rationale, as used by model $\mathbb{U}$ (**Shared**), or by the additional sentence, only selected with rationale–supervision (**New**).

On MultiRC, the two–sentence model $\mathbb{U}$ selects a single sentence as the rationale in 99.0%, whereas the model $\mathbb{S}$ selects two sentences on 51.4%. In 83.4% both models predict the same target $\hat{y}$. Based on these, we consider all instances, where model $\mathbb{S}$ selects the same sentence as model $\mathbb{U}$ plus one additional sentence as a rationale, to identify whether (i) both sentences are relevant, (ii) the shared sentence is relevant, or (iii) the additional sentence is relevant for model $\mathbb{S}$. Instead of looking at the prediction of the joint rationale of both

sentences of model $\mathbb{S}$, i.e., the selected rationale with the highest confidence score, we now look at the predictions of both selected sentences individually. Table 8 shows whether the prediction of model $\mathbb{S}$ remains stable for both predicted labels if only one of the sentences out of the two–sentence rationale is used. For `False` predictions, the additional sentence (only selected when supervised) has a major impact on the prediction and seems most relevant. This is in line with our observations in Section 5.4, namely that supervision affects the decision mechanism predicting this label. For the prediction of `True`, in almost all cases the same sentence as the one selected by model $\mathbb{U}$ yields in the same prediction. The additional sentence in isolation, however, changes the prediction to `False` in almost half of all cases. Though bound to our approach, these results suggest that rationale–supervision may yield in selecting rationales that are not required by the model to solve the target task, but rather the rationale–objective , thereby losing some of their faithfulness. This may be a relevant consideration when measuring faithfulness on a more fine–granular level.

E Movies Examples

Figure 9 shows an example of positive sentiment in which the model disregards sentences with clear positive stances and selects a sentence containing "scary" as the rationale. Figure 10 shows how the model correctly selects a sentence of positive stance but interprets this sentence as negative. Both examples show that sentences with opposing stances occur by discussing the plot and the movie in general.

(1) there ' s a thin line between satire and controversy , and mike nichols (the birdcage , wolf) has directed a sharp and very honest look at a us presidential election .
(2) based on the book written by " anonymous " (actually former " newsweek " writer joe klein) , john travolta plays governor jack stanton .
(3) but he does n ' t actually play stanton .
(4) he plays bill clinton ; just the same as emma thompson no doubt plays the first lady and billy bob thorton is the campaign manipulator james carville (although the credits will of course say otherwise) .
(5) the film is taken from the perspective of henry burton (adrian lester) , a morally correct and somewhat hesitant new advisor to stanton .
(6) he searches for justice and dignity in the ugliest possible situations , and whether it be keeping the history of his boss ' pants under wraps or contemplating digging up dirt on another politician , he approaches his work with a keen desire to skillfully serve his country and his fellow workers .
(7) richard jemmons (billy bob thorton) and daisy green (maura tierney) team up with henry as the would - be president ' s advisors , and hire lesbian veteran libby holden (kathy bates) as the campaign ' s eccentric " tougher than dirt " incriminator .
(8) together they face all sorts of sexual allegations , the irritatingly discourteous media and other witty politicians in the election race .
(9) in its satire and controversy , primary colors is a similar film to wag the dog : they both are not afraid to wipe their noses in the nitty - gritty and take a bold look at something that will never has honesty as a virtue .
(10) but whereas wag showed us how much affect a few people can have on the media , primary colors is much more concerned with fleshing out it ' s characters , letting us understand what they want and why , and making us truly appreciate the humanity and rectitude that they graciously represent .
(11) seeing john travolta play bill clinton
(12) *so confidently and justly is enough to make the film more than worth a look . and the rest of the cast also make*
(13) *superb performances - adrian lester sharply portrays the intellect of henry whilst kathy bates is perfect as the robust and energetic libby holden .*
(14) at occasions , you ca n ' t help but feel that these terrific characters are going to waste .
(15) there are long slabs of time where john travolta (unquestionably the most interesting to watch) is missed from the screen ; and since it is awkwardly structured as henry ' s story we are often forced to watch scenes that perhaps are not so necessary to the central plot - or even the point of the film .
(16) *having said that , make no mistake - primary colors is always enjoyable to watch*
(17) .
(18) but frequently we have to ask ourselves - exactly what are we watching ?
(19) most of the first half of its duration is a lightheaded look at melodramatic confrontations that seem so genuine we can not help but laugh , but the way primary colors chooses to finish tackles aspects that are very contrary , and almost unsuitable , to the rest of the film .
(20) *but as i mentioned before , there is a thin line between satire and controversy - and for the most part , primary colors delivers an entertaining indulgence of political matters combined with a far - from - overpowering look at winning the public ' s opinion .*
(21) although at occasions the film may jump around a little too freely , focus is never lost on how important and vulnerable the subject matter really is .
(22) thankfully , it is clear to make the distinction on what is entertaining movie cosmetics and what is a provocative documentation of something
(23) **so really it ' s scary .**

Figure 9: Example of Movies (dev) with gold label `POS` and predicted label `NEG`. *Italic* sentences are gold (sentence–level) rationales, **bold** is the selected rationale.

(1) buffalo ?
(2) 66 is a very rarely known movie that stars vincent gallo and christina ricci .
(3) gallo plays a very troubled man , who was sent to jail for gambling .
(4) once out of jail , he must visit his parents , who he told he was married .
(5) the truth is he is n ' t married .
(6) to try to impress them , he kidnaps a girl (christina ricci) from a tap dancing class to act as his wife .
(7) the film is very cheaply made , and it shows it throughout a lot of the movie , but you do n ' t need money to make a good film .
(8) buffalo ?
(9) 66 does n ' t always stay with the realistic concept , and sometimes goes through outrageous events .
(10) gallo ' s parents , played by angelica huston and ben gazarra , are two very strange individuals .
(11) the mother plays a football fanatic and the father plays a quiet man with odd habits .
(12) gallo and ricci arrive at his parent ' s house , and
(13) some extremely funny scenes take place within the house .
(14) *ricci ' s performance during the scene at gallo ' s parent ' s home are very well done .*
(15) *there is constantly humor involved in the interesting dinner table scenes .*
(16) *the way the movie was filmed in this particular part of the movie were interesting and creative .*
(17) *they seemed very mediocre , but they worked out just fine*
(18) .
(19) *gallo ' s character is developed very well .*
(20) the impression that he is very depressed and confused is very clear .
(21) *gallo gives a performance that makes you believe what the character is going through .*
(22) his character goes through many , many problems , just like many people in real life .
(23) this character seemed very realistic to me .
(24) ricci ' s character is funny and different .
(25) she does n ' t care much that she has been kidnaped , in fact , she falls in love the man who kidnaped her !
(26) ***ricci is a very wonderful actress and she is starting to get the recognition that she deserves***
(27) .
(28) buffalo ?
(29) 66 is n ' t all laughs though .
(30) many scenes are very dramatic and depressing .
(31) gallo ' s character was so realistic , he was extremely disturbing .
(32) some scenes are supposed to come off as funny , but they actually seemed sad and real to life .
(33) the film sometimes drags along , not giving much material .
(34) i really would have liked to see gallo ' s parents a lot more , and i would have liked to see the characters developed more .
(35) overall , buffalo ?
(36) 66 is n ' t as good as some people put it up to be .
(37) the bottom line -
(38) a few hysterical scenes save this film from sinking to the bottom .

Figure 10: Example of Movies (dev) with gold label `POS` and predicted label `NEG`. *Italic* sentences are gold (sentence–level) rationales, **bold** is the selected rationale.