D-REX: Dialogue Relation Extraction with Explanations

Alon Albalak[1], Varun Embar[2], Yi-lin Tuan[1], Lise Getoor[2], William Yang Wang[1]
[1]University of California, Santa Barbara [2]University of California, Santa Cruz
{alon_albalak, ytuan}@ucsb.edu
{vembar, getoor}@ucsc.edu
william@cs.ucsb.edu

Abstract

Existing research studies on cross-sentence relation extraction in long-form multi-party conversations aim to improve relation extraction without considering the explainability of such methods. This work addresses that gap by focusing on extracting explanations that indicate that a relation exists while using only partially labeled explanations. We propose our model-agnostic framework, D-REX, a policy-guided semi-supervised algorithm that optimizes for explanation quality and relation extraction simultaneously. We frame relation extraction as a re-ranking task and include relation- and entity-specific explanations as an intermediate step of the inference process. We find that human annotators are 4.2 times more likely to prefer D-REX's explanations over a joint relation extraction and explanation model. Finally, our evaluations show that D-REX is simple yet effective and improves relation extraction performance of strong baseline models by 1.2-4.7%.[1]

1 Introduction

Traditional relation extraction (RE) approaches discover relations that exist between entities within a single sentence. Recently, several approaches have been proposed which focus on cross-sentence RE, the task of extracting relations between entities that appear in separate sentences (Peng et al., 2017; Quirk and Poon, 2017; Han and Wang, 2020; Yao et al., 2019) as well as cross-sentence RE in dialogues (Yu et al., 2020; Chen et al., 2020; Xue et al., 2021; Qiu et al., 2021; Lee and Choi, 2021).

A crucial step towards performing cross-sentence RE in multi-entity and multi-relation dialogues is to understand the context surrounding relations and entities (e.g., who said what, and to whom). Figure 1 shows an example from the DialogRE dataset where a simple BERT-based model

Speaker 1: Could you please get the key off the back of the door for me.

Speaker 2: Oh yeah! Yeah!

Speaker 1: You tell your friend Chandler that we're definitely broken up this time.

Speaker 2: Okay!

Subject	Object	Initial Predicted Relation	D-REX Predicted Explanation	D-REX Predicted Relation
Speaker 2	Chandler	girl/boyfriend	your friend	friends

Figure 1: A sample dialogue between 2 speakers with actual D-REX predictions. The model initially classifies Speaker 2 and chandler, incorrectly, as girl/boyfriend. After predicting the explanation "your friend", D-REX correctly re-ranks the relation as friends.

(Initial Predicted Relation in Figure 1) gets confused by multiple entities and relations existing in the same dialogue (Yu et al., 2020). The model predicts the "girl/boyfriend" relation between Speaker 2 and Chandler, however, it is clear from the context that the "girl/boyfriend" relation is referring to a different pair of entities: Speaker 1 and Chandler.

One approach to encourage a model to learn the context surrounding a relation is by requiring the model to generate an explanation along with the relation (Camburu et al., 2018). In addition to the DialogRE dataset, Yu et al. (2020) introduces manually annotated *trigger words* which they show play a critical role in dialogue-based RE. They define trigger words as "the smallest span of contiguous text which clearly indicates the existence of the given relation". In the context of RE, these trigger words can be used as potential explanations.

Our work aims to extract explanations that clearly indicate a relation while also benefiting an RE model by providing cross-sentence reasoning. Our proposed approach, D-REX, makes use of multiple learning signals to train an explanation extraction model. First, D-REX utilizes trigger words as a partial supervision signal. Additionally, we pro-

[1]Code and data publicly available at https://github.com/alon-albalak/D-REX

pose multiple reward functions used with a policy gradient, allowing the model to explore the explanation space and find explanations that benefit the re-ranking model. Including these reward functions allows D-REX to learn meaningful explanations on data with less than 40% supervised triggers.

In order to predict relation- and entity-specific explanations in D-REX, we pose RE as a relation re-ranking task with explanation extraction as an intermediate step and show that this is not possible for a model trained to perform both tasks jointly.

Our contributions are summarized as follows:

- We propose D-REX, **D**ialogue **R**elation **E**xtraction with e**X**planations, a novel system trained by policy gradient and semi-supervision.
- We show that D-REX outperforms a strong baseline in explanation quality, with human evaluators preferring D-REX explanations over 90% of the time.
- We demonstrate that by conditioning on D-REX extracted explanations, relation extraction models can improve by 1.2-4.7%.

2 Problem Formulation

We follow the problem formulation of Yu et al.: let $d = (s_1 : u_1, s_2 : u_2, \ldots, s_n : u_n)$ be a dialogue where s_i and u_i denote the speaker ID and the utterance from the i^{th} turn, respectively. Let $\mathcal{E}, \mathcal{R}$ be the set of all entities in the dialogue and the set of all possible relations between entities, respectively. Each dialogue is associated with m relational triples <s, r, o> where $s, o \in \mathcal{E}$ are subject and object entities in the given dialogue and $r \in \mathcal{R}$ is a relation held between the s and o. Each relational triple may or may not be associated with a trigger t. It is important to note that there is no restriction on the number of relations held between an entity pair; however, there is at most one trigger associated with a relational triple. In this work, we consider an explanation to be of high quality if it strongly indicates that a relation holds, and for this purpose we consider triggers to be short explanations, though not always optimal in quality.

2.1 Relation Extraction (RE)

Given a dialogue d, subject s, and object o, the goal of RE is to predict the relation(s) that hold between s and o. We also consider RE with additional evidence in the form of a trigger or predicted explanation. Formally, this is the same as relation extraction with an additional explanation, ex.

2.2 Explanation Extraction (EE)

We formulate EE as a span prediction problem. Given a dialogue d consisting of n tokens T_1 through T_n, and a relational triple <s, r, o>, the goal of EE is to predict start and end positions, i, j in the dialogue, such that the explanation $ex = [T_i, T_{i+1}, \ldots, T_j]$ indicates that r holds between s and o.

3 Baseline Models

We first introduce approaches for RE and EE based on state-of-the-art language models. We then propose a multitask approach that performs both tasks jointly. Our approaches use $\text{BERT}_{\text{base}}$ (Devlin et al., 2019) and $\text{RoBERTa}_{\text{base}}$ (Liu et al., 2019b) pre-trained models[2], and follow their respective fine-tuning protocols.

For all models, we maintain a single input format, which follows from Yu et al.. Formally, for a dialogue d, subject s, object o, relation r, and explanation ex, the input sequence to all models is [CLS]$\{r/ex$[SEP]$\}s$[SEP]o[SEP]d, where $\{r/ex$[SEP]$\}$ denotes that the relation or explanation may be included depending on the task setting. For RoBERTa models, we use the <s> and </s> tokens rather than [CLS] and [SEP], respectively.

3.1 Relation Extraction (RE)

We follow the fine-tuning protocols of Devlin et al. and Liu et al. for BERT and RoBERTa classification models by using the output corresponding to the first token $C \in \mathbb{R}^H$ ([CLS] and <s>, respectively) as a latent representation of the entire input and train a classification matrix $W \in \mathbb{R}^{K \text{x} H}$, where K is the number of relation types and H is the dimension of the output representations from the language model. For each relation r_i, the probability of r_i holding between s and o in d is calculated as $P_i = \text{sigmoid}(CW_i^T)$. We compute the standard cross-entropy loss for each relation as

$$\mathcal{L}_{RE} = -\frac{1}{\text{K}} \sum_{i=1}^{\text{K}} y_i \cdot \log(P_i) + (1 - y_i) \cdot \log(1 - P_i) \quad (1)$$

where y_i denotes whether relation i holds.

[2]Pre-trained models obtained from https://github.com/huggingface/transformers (Wolf et al., 2020)

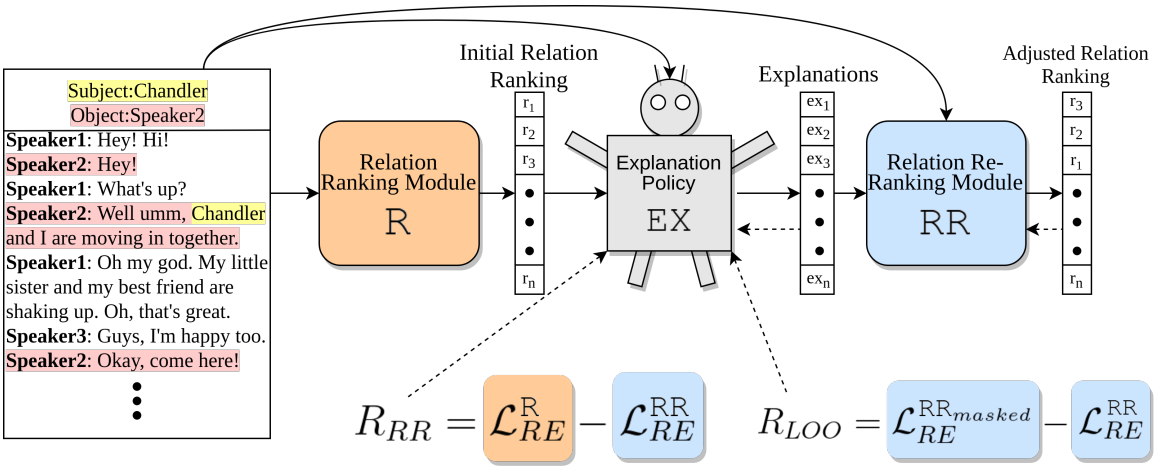

Figure 2: Overview of the D-REX system. The relation ***R***anking module ranks relations conditioned only on the subject, object, and the dialogue. The ***EX***planation policy extracts supporting evidence for the ranked relations by conditioning on individual relations in addition to the original input. The relation ***R***e***R***anking module conditions its rankings on supporting evidence from the explanation policy. In this hypothetical example, we see that relation 3 was originally ranked number 3 but had strong supporting evidence and was re-ranked in the number 1 spot. Solid lines represent model inputs/outputs, and dotted lines represent learning signals. Reward functions, $\mathcal{R}_{RR}$ and $\mathcal{R}_{LOO}$, are detailed in equations 4 and 5, respectively.

3.2 Explanation Extraction (EE)

For EE, we use the input described above, with a natural language phrasing of a relation appended to the beginning of the sequence. For example, if r is "per:positive_impression", then we concatenate "person positive impression" to the beginning.

We follow the fine-tuning protocol of Devlin et al. for span prediction. We introduce start and end vectors, $S, E \in \mathbb{R}^H$. If $T_i \in \mathbb{R}^H$ is the final hidden representation of token i, then we compute the probability of token i being the start of the predicted explanation as a dot product with the start vector, followed by a softmax over all words in the dialogue:

$$P^S_{T_i} = \frac{exp(S \cdot T_i)}{\sum_j exp(S \cdot T_j)} \quad (2)$$

To predict the end token, we use the same formula and replace the start vector S with the end vector E. To compute the loss, we take the mean of the cross-entropy losses per token for the start and end vectors. Formally, let $|d|$ be the number of tokens in dialogue d, then

$$\mathcal{L}_{EX} = -\frac{1}{|d|}\sum_{i}^{|d|} \big(y^S_i \cdot log(P^S_{T_i}) + (1 - y^S_i) \cdot \log(1 - P^S_{T_i})\big) + \big(y^E_i \cdot log(P^E_{T_i}) + (1 - y^E_i) \cdot \log(1 - P^E_{T_i})\big) \quad (3)$$

where y^S_i and y^E_i are the start and end labels. Because we want explanations extracted only from the dialogue, if the start or end token with largest log-likelihood occurs within the first l tokens, where l is the length of [CLS]r[SEP]s[SEP]o[SEP], then we consider there to be no predicted explanation.

3.3 Joint Relation and Explanation Model

The joint RE and EE model uses the standard input from §3. It utilizes a BERT or RoBERTa backbone, and has classification and span prediction layers identical to those in the RE and EE models. Similarly, the loss is computed as the weighted sum of RE and EE losses:

$$\mathcal{L}_{\mathcal{J}} = \alpha\mathcal{L}_{RE} + (1 - \alpha)\mathcal{L}_{EX}$$

where α is an adjustable weight. In practice, we find that $\alpha = 0.5$ works best.

Flaw of the joint model The disadvantage of the joint model is this: supposing that an entity pair has 2 relations, each explanation should be paired with a single relation. However, by making predictions jointly, there is no guaranteed mapping from predicted explanations to predicted relations. One method of solving this issue is to predict relations and explanations in separate steps. It is possible to first predict relations and then condition the explanation prediction on each individual relation and conversely. This idea forms the basis for D-REX.

4 D-REX

In this section, we introduce the D-REX system. We begin by introducing the models which make up the system. Next, we present the training and inference algorithms. Finally, we discuss the optimization objectives for each model in the system.

4.1 Models

The D-REX framework requires three components: an initial relation ranking model, an explanation model, and a relation re-ranking model, shown in Figure 2.

Initial Ranking Model (*R*) In our algorithm and discussions, we use *R* to denote the initial ranking model. There are no restrictions on *R*, it can be any algorithm which ranks relations (e.g., deep neural network, rule-based, etc.) such as (Yu et al., 2020; Lee and Choi, 2021). However, if *R* needs to be trained, it must be done prior to D-REX training; D-REX will not make any updates to *R*.

In our evaluations, we use the relation extraction model described in §3.1. The input to this model is (s,o,d) and the output is a ranking, $R(s, o, d)$.

Explanation Extraction Model (*EX*) In our algorithm and discussions, we use *EX* to denote the explanation model. In this paper we limit our experiments to extractive explanation methods, as opposed to generative explanation methods, however this is not a limitation of D-REX. The only limitation on the explanation model is that we require it to produce human-interpretable explanations. Thus, it is also possible to use generative models such as GPT-2 (Radford et al., 2019) or graph-based methods such as (Yu and Ji, 2016; Xue et al., 2021) with adjustments to the formulation of the reward functions.

In our evaluations, we use the model as described in §3.2. The input to *EX* is (r,s,o,d) and the output is an extracted phrase from d, denoted as $EX(r, s, o, d)$.

Relation Re-Ranking Model (*RR*) In our algorithm and discussions, we let *RR* denote the relation re-ranking model. In the D-REX training algorithm, *RR* is updated through gradient-based optimization methods, and must be able to condition its ranking on explanations produced by *EX*. In our experiments, we use the same model architecture as *R* and include an explanation as additional input to the model. The input to *RR* is (ex,s,o,d) and the output is a relation ranking, denoted as $RR(ex, s, o, d)$.

Algorithm 1: The proposed training algorithm for D-REX

Input : Pre-trained ranking, explanation, and re-ranking models: *R*, *EX*, *RR*
k: for number of relations to re-rank
Data: Dataset: $\mathcal{D}$
for (s, r, o, t, d) *in* $\mathcal{D}$ **do**
  Compute ranking loss: $\mathcal{L}^{R}_{RE}(s, o, d)$
  $r_{pred} \leftarrow R(\text{s,o,d})_{1:k}$
  for i *in* r_{pred} **do**
    $ex_i \leftarrow EX(r_{pred_i}, s, o, d)$
    Compute Re-ranking loss: $\mathcal{L}^{RR}_{RE}(ex_i, s, o, d)$; // Equation 1
    Compute Re-Ranking Reward: $\mathcal{R}_{RR}$; // Equation 4
    Compute Leave-one-out Reward: $\mathcal{R}_{LOO}$; // Equation 5
    Compute policy gradient with rewards R_{RR}, R_{LOO} ; // Equation 6
  end
  if t *not empty* **then**
    Compute $\mathcal{L}_{EX}$; // Equation 3
  end
  Update *EX*,*RR*parameters with calculated losses
end

4.2 D-REX Algorithm

The outline of this algorithm is shown in pseudocode in Algorithm 1.

Assuming that we have ranking, explanation, and re-ranking models *R*, *EX*, *RR*, then given a single datum (s, r, o, t, d), comprised of a subject, relation, object, trigger(may be empty), and dialogue, the D-REX algorithm operates as follows: The ranking model takes as input (s, o, d) and computes the probability of each relation from the predefined relation types. Next, we take the top-k ranked relations, $r_{pred} = R(s, o, d)_{1:k}$, and compute explanations. For $i = 1, ..., k$, explanations are computed as $ex_i = EX(r_{pred_i}, s, o, d)$. Finally, for each predicted explanation, the re-ranking model computes k probabilities for each relation type, using (ex_i, s, o, d) as the input to *RR*. The final probabilities for each relation type are computed as the mean across all k+1 predictions from *R* and *RR*.

4.3 Model optimization

We propose multiple optimization objectives to train an *EX* model that extracts explanations meaningful to humans and beneficial to the relation extraction performance while ensuring that *RR* maintains high-quality predictions.

Explanation Model Optimization We train *EX* with supervision on labeled samples, and a policy gradient for both labeled and unlabeled samples, allowing for semi-supervision. For the policy gradi-

ent, we introduce two reward functions: a relation re-ranking reward and a leave-one-out reward.

Re-ranking Reward The purpose of the re-ranking reward is to ensure that *EX* predicts explanations which benefit *RR*. Formally, let $\mathcal{L}^{R}_{RE}(s, o, d)$ be the loss for *R*, given the subject, object, and dialogue: s, o, d. And let $\mathcal{L}^{RR}_{RE}(ex, s, o, d)$ be the loss of *RR*, given the explanation, subject, object, and dialogue: ex, s, o, d. Then we define the relation re-ranking reward as:

$$\mathcal{R}_{RR} = \mathcal{L}^{R}_{RE}(s, o, d) - \mathcal{L}^{RR}_{RE}(ex, s, o, d) \quad (4)$$

Because *R* is stationary, *EX* maximizes this function by minimizing $\mathcal{L}^{RR}_{RE}$. Of course, *EX* can only minimize $\mathcal{L}^{RR}_{RE}$ through its predicted explanations.

Leave-one-out Reward The purpose of the leave-one-out reward is to direct *EX* in finding phrases which are essential to correctly classifying the relation between an entity-pair. This reward function is inspired by previous works which make use of the leave-one-out idea for various explanation purposes (Shahbazi et al., 2020; Li et al., 2016). We can calculate the leave-one-out reward using either *R* or *RR*, and it is calculated by finding the difference between the standard relation extraction loss and the loss when an explanation has been masked. Formally, if d is the original dialogue and ex is the given explanation, let $d_{mask}(ex)$ be the dialogue with ex replaced by mask tokens. Then, the leave-one-out reward is defined as:

$$\mathcal{R}_{LOO} = \mathcal{L}_{RE}(s, o, d_{mask}(ex)) - \mathcal{L}_{RE}(s, o, d) \quad (5)$$

Because $\mathcal{L}_{RE}$ is calculated using the same model for both the masked and unmasked loss, *EX* maximizes this reward function by maximizing the masked loss. Of course, the only interaction that *EX* has with the masked loss is through the explanation it predicts.

Policy Gradient We view *EX* as an agent whose action space is the set of all continuous spans from the dialogue. In this view, the agent interacts with the environment by selecting two tokens, a start and end token and receives feedback in the form of the previously discussed reward functions. Let i, j be the start and end indices that the explanation model selects and T_i be the i^{th} token, then $ex = d[i:j] = [T_i, T_{i+1}, \ldots, T_j]$ and the probabilities of i, j being predicted are calculated as $P^{S}_{T_i}$ and $P^{E}_{T_j}$ according to equation 2.

For both reward functions, we use a policy gradient (Sutton and Barto, 2018) to update the weights of the explanation model and calculate the loss as

$$\mathcal{L}_{EX_{PG}} = -(log(P^{S}_{T_i}) + log(P^{E}_{T_j})) * (R_{RR} + R_{LOO}) \quad (6)$$

Additionally, while training *EX* in the D-REX algorithm, we make use of supervision when available. In the case where supervision exists, we calculate an additional loss, $\mathcal{L}_{EX}$, as defined in equation 3.

Relation Extraction Re-ranking Model Optimization While training D-REX we train *RR* with labeled relations as supervision and use a cross-entropy loss, $\mathcal{L}^{RR}_{RE}$, calculated in the same way as *R* in Equation 1.

5 Experimental Evaluation

In this section, we present an evaluation of D-REX in comparison with baselines methods on the relation extraction and explanation extraction tasks.

5.1 Experimental settings

For our experiments, we re-implement the BERT_{S} model from (Yu et al., 2020) as well as a new version which replaces BERT with RoBERTa. In our paper, we refer to these models as R_{BERT} and R_{RoBERTa}. All models are implemented in PyTorch[3] and Transformers(Wolf et al., 2020), trained using the AdamW optimizer (Loshchilov and Hutter, 2018). All experiments were repeated five times and we report mean scores along with standard deviations. D-REX models use a top-k of five and are initialized from the best performing models with the same backbone. For example, $\text{D-REX}_{\text{BERT}}$ uses two copies of R_{BERT} (Yu et al., 2020) to initialize the ranking and re-ranking models and EX_{BERT} to initialize the explanation model. When training *Joint*, we do not calculate $\mathcal{L}_{EX}$ for relational triples without a labeled trigger. The full details of our training settings are provided in Appendix B.

DialogRE Dataset We evaluate our models on the DialogRE English V2 dataset[4] which contains dialogues from the Friends TV show (Yu et al., 2020), details of which are in Table 1. D-REX models are trained with trigger supervision on less than 40% of the training data, and make no use of dev or test set triggers. The learning signal for the remaining triples comes entirely from our rewards through a policy gradient.

[3] https://pytorch.org/

[4] Dataset collected from https://dataset.org/dialogre/ for research purposes only

DialogRE V2			
Dialogues	Relations	Relational Triples (train/dev/test)	Triggers (train/dev/test)
1788	36	6290/1992/1921	2446/830/780

Table 1: **Dataset details** for DialogRE. With only 2446 labeled triggers in the training set, D-REX models learn using only a policy gradient and no direct supervision on the remaining 3844 triples.

Model	F1(σ)	MRR(σ)
R_{BERT}	59.2(1.9)	74.8(1.3)
$Joint_{\text{BERT}}$	59.4(1.7)	74.0(0.9)
D-REX$_{\text{BERT}}$	**59.9**(0.5)	**75.4**(0.1)
R_{RoBERTa}	64.2(1.6)	77.9(1.0)
$Joint_{\text{RoBERta}}$	65.2(0.3)	78.3(0.3)
D-REX$_{\text{RoBERTa}}$	**67.2**(0.3)	**79.4**(0.3)
*GDPNet	60.2(1.0)	-
*TUCORE-GCN$_{\text{BERT}}$	65.5(0.4)	-
†SocAoG	**69.1**(0.5)	-

Table 2: **Relation extraction results on DialogRE V2**. *R* models are described in Section 3.1, *Joint* models in 3.3, and D-REX models in 4. R_{BERT} is a replication of BERT$_{\text{S}}$ from Yu et al. (2020). "*" denotes results taken from Lee and Choi (2021) and "†" from Qiu et al. (2021)

Evaluation Metrics We adopt separate evaluations for relation and explanation extraction.

First, for relation extraction, we evaluate our models using F1 score, following Yu et al. (2020), and additionally calculate the mean reciprocal rank (MRR), which provides further insight into a model's performance. For example, MRR is able to differentiate between a ground truth relation ranked 2nd or 10th, while the F1 score does not. In the dialogRE dataset, multiple relations may hold between a single pair of entities, so we use a variation of MRR which considers all ground truth relations, rather than just the highest-ranked ground truth relation.

For explanation extraction, we focus mainly on manual evaluations, but also propose the Leave-One-Out metric, introduced in section 5.4 for an ablation study.

5.2 Relation Extraction (RE) Evaluation

In Table 2, we compare the baseline RE model R_{BERT} with the methods presented in this paper. We also compare with three other methods which use similarly sized language models, but additionally utilize graph neural networks (GNN): GDPNet(Xue et al., 2021), TUCORE-GCN$_{\text{BERT}}$(Lee and Choi, 2021), and SocAoG(Qiu et al., 2021).

First, we see that even though D-REX is designed to introduce human-understandable explanations, it still has modest improvements over R_{BERT}, which focuses on RE, while *Joint* has no significant improvement. Next, we see a five point absolute improvement in F1 from the baseline model when using RoBERTa. The trend from BERT to RoBERTa is similar to results found by Lee and Choi (2021), where changing from a BERT$_{\text{base}}$ model to RoBERTa$_{\text{Large}}$(not shown here) improved their model performance significantly. Additionally, we see a 3 point improvement from *R* to D-REX when using RoBERTa (compared to 0.7 for BERT), which we believe is due to the better performing ranking model, which allows for D-REX to rely more on the input explanations. Finally, we see that by using GNNs, and task-specific dialogue representations, all three GNN-based methods can improve over the general BERT-based methods.

5.3 Explanation Extraction (EE) Evaluation

Automatic Evaluation Although the aim of this paper is not trigger prediction, for completeness and reproducibility, we include results on the test set of triggers in Appendix A.

Human Evaluation To better understand how our model performs in extracting explanations and what challenges still exist, we perform two analyses; a comparative and an absolute analysis. We consider two sets of data for evaluation: samples for the DialogRE test set where **N**o **L**abeled trigger exists (***NL***) and samples where the predicted explanation **D**iffers from the **L**abeled trigger (***DL***).

5.3.1 Comparative Analysis

In Table 3, we show the results for pairwise comparisons of explanations predicted by D-REX$_{\text{RoBERTa}}$ against 3 baselines: random strings of 1-4 words, predictions from $Joint_{\text{RoBERTa}}$, and labeled triggers. For each comparison, we employ 3 crowd-workers[5], who were given the full dialogue, a natural language statement corresponding to a relational triple, and the two proposed explanations highlighted in the dialogue[6]. The crowd-workers were asked to specify which of the highlighted explanations was most indicative of the relation, or

[5]Amazon Mechanical Turk workers were paid $0.35 per HIT, where a HIT includes 3 comparisons. We estimate an average HIT completion time of ~1.5 minutes, averaging ~$14 per hour. We only accept workers from AUS, CA, and USA.

[6]Example HIT included in Appendix 4

D-REX$_{\text{RoBERTa}}$ vs.	Win(%)	Tie(%)	Lose(%)
Random (*NL*)	79.9	10.4	9.8
Joint$_{\text{RoBERTa}}$ (*NL*)	38.5	52.3	9.2
Ground truth (*DL*)	12.1	44.3	43.7

Table 3: **Human evaluator preferences on explanation extraction methods**. *NL* and *DL* are samples where **N**o **L**abeled trigger exists, and where the predicted explanation **D**iffers from the **L**abel, respectively. Results presented are percentages of preference.

	Not Indicative	Incorrect Entity Pair	Incorrect Relation	Indicative
NL	29	19	18	34
DL	19	13	7	61

Table 4: **Explanation error analysis** on 100 samples where **N**o **L**abeled trigger exists (*NL*) and 100 where the predicted explanation **D**iffers from the **L**abel (*DL*).

they could be equal. For each comparison we use a majority vote, and if there was a three-way tie we consider the explanations to be equal. We compare D-REX with random strings and the joint model on 174 samples from ***NL***, as well as 174 samples from ***DL***.

In Table 3 we see that for ***NL***, D-REX produces explanations which were 4.2 times more likely to be outright preferred by crowd-workers than the joint model, suggesting that our reward functions properly guided the explanation policy to learn meaningful explanations on unlabeled data. Surprisingly, we found that on over 12% of samples with labeled triggers, evaluators outright preferred D-REX explanations over the ground truth trigger, suggesting that D-REX indeed finds some explanations which are better than the ground truth trigger.

In Appendix 5.5, we include 2 examples comparing explanations from D-REX and *Joint*.

5.3.2 Absolute Analysis

To better understand the quality of D-REX's explanations, we randomly sample 100 from both ***NL*** and ***DL*** for a fine-grained analysis. We classify the explanations into 4 categories: not indicative, incorrect entity-pair, incorrect relation, and indicative. "Indicative" and "Not indicative" have the obvious meanings, "Incorrect entity-pair" means that an explanation actually explains the correct relation, but between the incorrect entity-pair, and "Incorrect relation" means that the explanation indicates a relation different from the desired relation.

Table 4 shows the results. Interestingly, we see in the ***NL*** set, that errors were equally likely to come from either an explanation indicating the relation for an incorrect entity-pair as for the incorrect relation altogether. This is in contrast to the ***DL*** set, where D-REX was nearly half as likely to predict an explanation for an incorrect relation as it was for an incorrect entity-pair.

Additionally, in our fine-grained analysis, we also considered whether a relational triple was identifiable from the context alone and found that nearly 20% of the 200 samples had ambiguities which could not be resolved without outside knowledge. This suggests that there is likely a maximum achievable relation extraction score on the DialogRE dataset under the current setting.

Model	F1	Leave-one-out(↓)
D-REX$_{\text{RoBERTa}}$ (Full)	67.2	83.9
- reranking reward	66.0	84.9
- LOO reward	67.1	85.4

Table 5: **Ablation study** on reward functions. Leave-One-Out metric (LOO) measures how salient a predicted explanation is in determining a relation and is further defined and motivated in §5.4. Smaller LOO is better.

5.4 Ablation Study

To assess the benefit of each proposed reward individually, we perform an ablation study on the reward functions. In order to study explanation quality automatically, we introduce a new metric for explanation quality; the Leave-One-Out metric.

The Leave-One-Out (LOO) metric has a theoretical basis in the works of Li et al. (2016) and Ribeiro et al. (2016), where Li et al. (2016) use word erasure to determine a "word importance score". Here we define LOO formally. For a relation extraction model R, an explanation extraction model EX, and a dataset $\mathcal{D}$, LOO is calculated as

$$LOO(R, EX, \mathcal{D}) = \frac{\text{F1}_R(\mathcal{D}_{\text{MASK}}(EX))}{\text{F1}_R(\mathcal{D})}$$

where $\text{F1}_R(\mathcal{D})$ is the F1 score of R on $\mathcal{D}$ and $\mathcal{D}_{\text{MASK}}(EX)$ is the dataset where explanations predicted by EX are replaced by mask tokens. The LOO metric calculates how essential the predicted explanations are to the ability of the relation extraction model.

To show that LOO is an appropriate measure of explanation quality, we compute the Pearson correlation coefficient between token F1 score and LOO scores for models on labeled triggers, found in Table 6. With 6 models trained on 5 random seeds each, we have 30 data points and a correlation

Dialogue	Subject Object Relation
... Speaker 1: Oh, I'm just so exhausted from dragging around this huge ***engagement ring***! ... Speaker 7: Hey, I'm sorry. I should have given you guys my black book when I got married! Although it wasn't so much a book as a...napkin. With Janice's phone number on it. ...	Janice Speaker 7 girl/boy-friend
Speaker 1: Sir? Speaker 2: What's ***in*** it? Speaker 1: Goat cheese, water chestnuts and panchetta. ... Speaker 3: Joey, it's been three days, okay. You're just a little homesick, okay. Would you just try to relax. Just try to enjoy yourself. Speaker 2: You're different here too. You're mean in England. ...	England Speaker 3 visited_by

Figure 3: Two examples comparing predicted explanations from D-REX (underlined) and *Joint* (***bold***).

coefficient of -87.4 with $p = 2.4 * 10^{-8}$. Because we calculate the coefficient with respect to human-annotated triggers, this suggests that a low LOO correlates with explanations that humans would determine as indicative of the given relation.

For our experiments, we always calculate LOO using the baseline model, R_{BERT}. From the results in Table 5, we see that both reward functions benefit the final results. Compared with R_{RoBERTa}, D-REX$_{\text{RoBERTa}}$ gains 3 F1 points, but without the reranking reward, the model only gains 1.8 F1 score or 60% of the total possible improvement. This performance loss demonstrates that the reranking reward is critical to attaining the best score in relation extraction. Similarly, without the leave-one-out reward, the model's explanation quality, measured in LOO, is 1.5 points, or nearly 10% worse, demonstrating that the leave-one-out reward is beneficial in guiding the model to salient explanations.

5.5 Explanation Samples

Figure 3 shows two samples comparing explanations from D-REX and *Joint*. In both examples, even though there was no labelled trigger, each model was able to predict an explanation which correlates with the relation. Specifically, "engagement ring" and "got married" are related to the girl/boyfriend relation, and "in" and "mean in" can be associated with the visited_by relation. However, the bottom example shows that *Joint* did not consider the context surrounding it's explanation. The conversation is about food, and the visited_by relation is not relevant. On the other hand, D-REX finds the phrase "you're mean in", where "you're" refers to speaker3, and "in" refers to "England". This is clearly an explanation which indicates the correct relation between the correct entities.

5.6 Reduced Labels

All previous results use 100% of labeled triggers in the DialogRE dataset, which covers 40% of all relational triples. To test how few labeled triggers *EX* requires in order to learn meaningful explanations we ran a small scale experiment (1 random seed) using labeled triggers from only 5, 10, and 20% of relational triples. However, in the small tests we ran, we found that at 20% labeled triggers the *EX* model mostly predicts no explanations. Furthermore, at 10% and fewer labeled triggers, the model converges to the trivial solution in the explanation space which is to never predict any tokens.

We believe that this issue is due, in part, to two challenges: the search space over all possible start/end tokens is too large, and the policy gradient has a high variance. Although these results may seem discouraging, we believe this challenge can be overcome in the future by using algorithms which reduce variance in the policy gradient and by initializing *EX* with a model pre-trained in span extraction.

6 Limitations and Future Work

Firstly, this study focuses on learning explanations as well as relations in dialogue and DialogRE is the only currently available dataset with annotations for both tasks. A limitation of this study is the small scale at which we were able to test the methods. A future direction would be to learn explanations on a

different RE dataset and use the pre-trained model in D-REX, however it would be non-trivial for a model to transfer explanations learned on a wildly different domain. Additionally, it is theoretically possible to train D-REX with no labeled triggers at all, however, we were unsuccessful and in Section 5.6 we discuss these and additional negative results.

This study focuses on relations and entities found in multi-party conversations, and while there are similarities between the dialogue domain, medical literature, and wikipedia (e.g., multi-entity, multi-relation), it is not clear whether the methods from this paper can transfer to other such domains. We plan to investigate how well the proposed methods transfer to relations and entities in other domains such as news and web text (Zhang et al., 2017) and for other types of semantic relations as in Hendrickx et al. (2010) or Yao et al. (2019).

We acknowledge that this study is English-focused, and it is not clear that these methods can transfer to languages in other families such as afro-asiatic or sino-tibetan. Additionally, we think that it would be very interesting to see how these methods perform on languages with very different linguistic features; for example, languages with inflection such as Finnish. We leave non-English and multilingual variations of these methods to future work.

In this work, we do not focus on improving state-of-the-art trigger prediction. However, we recognize that trigger annotation is labor-intensive, and a possible use of D-REX would be to use predicted labels as a form of weak supervision for a system whose goal *is* to improve on trigger prediction.

7 Related Work

Recently, there have been numerous information extraction tasks proposed which involve dialogues, including character identification (Zhou and Choi, 2018), visual coreference resolution (Yu et al., 2019), emotion detection (Zahiri and Choi, 2018).

New settings for relation extraction have also been proposed, such as web text (Ormándi et al., 2021) and, in many ways similar to dialogue, document text (Yao et al., 2019). There have also been methods developed to include explanations in similar natural language understanding tasks (Camburu et al., 2018; Kumar and Talukdar, 2020; Liu et al., 2019a; Lei et al., 2016). There have even been methods developed which, similarly to our re-ranking, make use of an explanation as additional information (Hancock et al., 2018).

The work by Shahbazi et al. is aligned with our study. They also focus on relation extraction with explanations; however, their method is based on distant supervision from bags of sentences containing an entity-pair. Due to the cross-sentence nature of relations in dialogue, their method is not applicable here, although we draw inspiration from their work. They explain their model by considering the salience of a sentence to their model's prediction, similarly to our leave-one-out reward.

Also relevant to our work is that by Bronstein et al.. Their work focuses on the task of semi-supervised event trigger labeling, which is very similar to our semi-supervised prediction of relation explanations. In their work, they use only a small seed set of triggers and use a similarity-based classifier to label triggers for unseen event types.

Finally, there have been multiple recent works in dialogue RE which perform quite well by using graph neural networks (Xue et al., 2021; Qiu et al., 2021; Lee and Choi, 2021). However, they focus only on RE and not on explaining the relations.

8 Conclusion

In this work, we demonstrated that not only is it possible to extract relation explanations from multi-party dialogues, but these explanations can in turn be used to improve a relation extraction model. We formulated purpose-driven reward functions for training the explanation model and demonstrated their importance in learning high quality explanations. Our proposed approach, D-REX, is powered by a very simple reformulation of the traditional relation extraction task into a re-ranking task.

9 Ethical and Social Considerations

The methods proposed in this work on their own are not nefarious, however, proposed explanations should not be blindly accepted as fact. For the same reasons that language models may have ethical and social risks, so may our algorithm which is built on top of such language models. While we test only on TV show dialogues, were this technology to be put to use in non-scripted, real life conversations, there would need to be very thorough analysis of any ethical risks that the proposed explanations may have.

References

Ofer Bronstein, Ido Dagan, Qi Li, Heng Ji, and Anette Frank. 2015. Seed-based event trigger labeling: How far can event descriptions get us? In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 372–376, Beijing, China. Association for Computational Linguistics.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-snli: Natural language inference with natural language explanations. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

H. Chen, Pengfei Hong, Wei Han, Navonil Majumder, and Soujanya Poria. 2020. Dialogue relation extraction with document-level heterogeneous graph attention networks. *ArXiv*, abs/2009.05092.

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.

Xiaoyu Han and Lei Wang. 2020. A novel document-level relation extraction method based on bert and entity information. *IEEE Access*, 8:96912–96919.

Braden Hancock, Paroma Varma, Stephanie Wang, Martin Bringmann, Percy Liang, and Christopher Ré. 2018. Training classifiers with natural language explanations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1884–1895, Melbourne, Australia. Association for Computational Linguistics.

Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. 2010. SemEval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 33–38, Uppsala, Sweden. Association for Computational Linguistics.

Sawan Kumar and Partha Talukdar. 2020. NILE : Natural language inference with faithful natural language explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8730–8742, Online. Association for Computational Linguistics.

Bongseok Lee and Yong Suk Choi. 2021. Graph based network with contextualized representations of turns in dialogue. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 443–455, Online and Punta Cana, Dominican Republic. 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.

Jiwei Li, Will Monroe, and Dan Jurafsky. 2016. Understanding neural networks through representation erasure. *CoRR*, abs/1612.08220.

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. *CoRR*, abs/1907.11692.

Ilya Loshchilov and Frank Hutter. 2017. Fixing weight decay regularization in adam. *CoRR*, abs/1711.05101.

Ilya Loshchilov and Frank Hutter. 2018. Fixing weight decay regularization in adam.

Róbert Ormándi, Mohammad Saleh, Erin Winter, and Vinay Rao. 2021. Webred: Effective pretraining and finetuning for relation extraction on the web. *CoRR*, abs/2102.09681.

Nanyun Peng, Hoifung Poon, Chris Quirk, Kristina Toutanova, and Wen tau Yih. 2017. Cross-sentence n-ary relation extraction with graph lstms. *Transactions of the Association for Computational Linguistics*, 5:101–115.

Liang Qiu, Yuan Liang, Yizhou Zhao, Pan Lu, Baolin Peng, Zhou Yu, Ying Nian Wu, and Song-Chun Zhu. 2021. Socaog: Incremental graph parsing for social relation inference in dialogues. In *ACL/IJCNLP*.

Chris Quirk and Hoifung Poon. 2017. Distant supervision for relation extraction beyond the sentence boundary. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1171–1182, Valencia, Spain. Association for Computational Linguistics.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

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.

Hamed Shahbazi, Xiaoli Fern, Reza Ghaeini, and Prasad Tadepalli. 2020. Relation extraction with explanation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6488–6494, Online. Association for Computational Linguistics.

Richard S Sutton and Andrew G Barto. 2018. *Reinforcement learning: An introduction*. MIT press.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Fuzhao Xue, Aixin Sun, Hao Zhang, and Eng Siong Chng. 2021. Gdpnet: Refining latent multi-view graph for relation extraction. In *AAAI*.

Yuan Yao, Deming Ye, Peng Li, Xu Han, Yankai Lin, Zhenghao Liu, Zhiyuan Liu, Lixin Huang, Jie Zhou, and Maosong Sun. 2019. Docred: A large-scale document-level relation extraction dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 764–777.

Dian Yu and Heng Ji. 2016. Unsupervised person slot filling based on graph mining. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 44–53, Berlin, Germany. Association for Computational Linguistics.

Dian Yu, Kai Sun, Claire Cardie, and Dong Yu. 2020. Dialogue-based relation extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4927–4940, Online. Association for Computational Linguistics.

Xintong Yu, Hongming Zhang, Yangqiu Song, Yan Song, and Changshui Zhang. 2019. What you see is what you get: Visual pronoun coreference resolution in dialogues. 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 5123–5132, Hong Kong, China. Association for Computational Linguistics.

Sayyed M Zahiri and Jinho D Choi. 2018. Emotion detection on tv show transcripts with sequence-based convolutional neural networks. In *Workshops at the thirty-second aaai conference on artificial intelligence*.

Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D. Manning. 2017. Position-aware attention and supervised data improve slot filling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 35–45, Copenhagen, Denmark. Association for Computational Linguistics.

Ethan Zhou and Jinho D. Choi. 2018. They exist! introducing plural mentions to coreference resolution and entity linking. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 24–34, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

model	token F1(σ)	EM(σ)	LOO(σ)
EX_{BERT}	62.1(3.1)	54.1(1.9)	82.2(0.4)
$Joint_{\text{BERT}}$	43(1.3)	38.6(1.4)	89.0(1.0)
D-REX$_{\text{BERT}}$	50.5(1.1)	45.7(1.7)	84.4(1.6)
EX_{RoBERTa}	66.5(2.2)	58.4(2.0)	82.2(0.4)
$Joint_{\text{RoBERTa}}$	49(0.7)	47(0.7)	86.2(0.8)
D-REX$_{\text{RoBERTa}}$	57.2(2.1)	51.6(1.6)	83.9(0.4))

Table 6: **Trigger prediction results**. Leave-One-Out metric (LOO) measures how salient a predicted explanation is in determining a relation and is further defined in §5.4. Smaller LOO is better.

A Trigger prediction

In Table 6, we compare our methods for supervised explanation extraction with D-REX. Interestingly, we find that the joint model achieves the lowest F1 score for both the BERT and RoBERTa models. $Joint_{\text{BERT}}$ scores nearly 20 points below its counterpart BERT model, while the $Joint_{\text{RoBERTa}}$ model cuts that difference to just over 15 points below its RoBERTa counterpart. On the other hand, D-REX maintains a token F1 score within 10 points of its counterpart even though it has been trained to generalize beyond the labeled triggers.

B Hyperparameters

All models are trained using the AdamW optimizer (Loshchilov and Hutter, 2017) with a learning rate of $3e$-5 and batch sizes of 30. To determine the best learning rate, *R* and *EX* models were trained using learning rates in {$3e$-6, $1e$-5, $3e$-5, $1e$-4}. The best learning rate, $3e$-5, was determined by performance on a held out validation dataset. Baseline models (*R*, *EX*, and *Joint*) are trained for at most 30 epochs and we use validation-based early stopping to determine which model to test. D-REX models are trained for at most 30 additional epochs with the best model determined based on relation extraction F1 scores computed on validation data. We found the best validation result to always occur within the first 30 epochs. All experiments were repeated five times and we report the mean score along with standard deviation. To train the joint model, we do not calculate $\mathcal{L}_{EX}$ for relational triples which do not have a labeled trigger and we select α from {0.25,0.5,0.75} and set α to 0.5 based on validation performance.

C Crowd-Worker Sample

In Figure 4, we show a sample HIT that was provided to crowd-workers. Each crowd-worker was shown three examples. The layour is as follows: the top always asks the worker to decide which of the highlighted texts is a better indication of the relation. Next, a natural language interpretation of the relational triple is given, in this case, "Speaker 2 and Speaker 1 are (or were) lovers". Then, we show the entire dialogue along with highlighted spans of text for each explanation. Finally, at the bottom, we always provide the user with three choices: yellow is better, equal, or orange is better, where the user is only allowed to select one option.

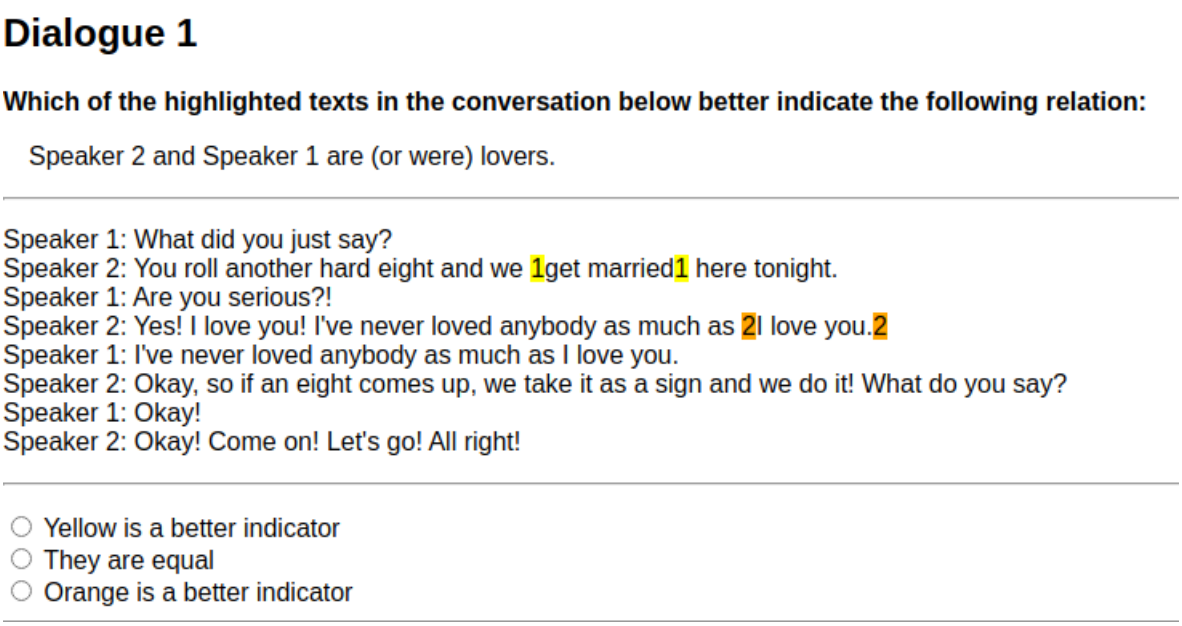

Figure 4: A sample HIT that was presented to crowd-workers for the comparative study of explanations.