D-REX: Dialogue Relation Extraction with Explanations
Alon Albalak, Varun Embar, Yi-Lin Tuan, Lise Getoor, William Yang Wang
UC Santa Barbara, UC 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 data. We propose our model-agnostic framework, D-REX, a policy-guided semi-supervised algorithm that explains and ranks relations. 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 about 90% of the time, human annotators prefer D-REX’s explanations over a strong BERT-based joint relation extraction and explanation model. Finally, our evaluations on a dialogue relation extraction dataset show that our method is simple yet effective and achieves a state-of-the-art F1 score on relation extraction, improving upon existing methods by 13.5%. Our code, data and infrastructure are available online.1

1 Introduction
Traditional relation extraction approaches discover relations that exist between entities within a single sentence. Recently, several relation extraction approaches have been proposed, focusing on cross-sentence relation extraction, the task of extracting relations between entities that appear in separate sentences (Peng et al. 2017; Quirk and Poon 2017; Han and Wang 2020; Wang et al. 2019). To perform well, these approaches must understand the context surrounding relations and reason across multiple sentences. While cross-sentence relation extraction methods have been used to extract relations from long-form dialogues (Yu et al. 2020; Chen et al. 2020), the dialogue domain provides additional challenges not seen in the medical literature or wikipedia domains (Peng et al. 2017; Quirk and Poon 2017; Yao et al. 2019).

A crucial step towards performing cross-sentence relation extraction in multi-entity and multi-relation dialogues is to understand the context surrounding relations and entities (e.g., who said what, and to whom) which current state-of-the-art models lack. Figure 1 shows an example from the DialogRE dataset where the current state-of-the-art model (listed as 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 meant to be 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 relation extraction. They define trigger words as “the smallest span of contiguous text which clearly indicates the existence of the given relation”. In the context of relation extraction, these trigger words can be used as potential explanations.

Our work aims to extract explanations that clearly indicate a relation while also benefiting a relation extraction model by providing cross-sentence reasoning. Our proposed approach, D-REX, makes use of multiple learning signals to train an explanation extraction model. We utilize trigger words as a partial supervision signal and demonstrate that full trigger word annotation is unnecessary for learning meaningful explanations. In addition to partial supervision from labeled triggers, we propose multiple reward functions

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 predictions. The model initially classifies Speaker 2 and Chandler as girl/boyfriend. After predicting the explanation “your friend”, D-REX correctly re-ranks the relation as friends.

1https://github.com/alon-albalak/D-REX
used with a policy gradient, allowing the model to explore the explanation space and find explanations that benefit the re-ranking model.

In order to predict relation- and entity-specific explanations in D-Rex, we pose relation extraction as a relation re-ranking task with explanation extraction as an intermediate step. We show that this is not possible for a model trained to perform both tasks simultaneously. Figure 1 shows an example where D-Rex initially predicts an incorrect relation but is able to identify a meaningful explanation for the correct relation, which allows it to re-rank the relations accurately.

Our contributions are summarized as follows:

• We propose D-Rex, Dialogue Relation Extraction with eXplanations, a novel system trained by policy gradient and semi-supervision.
• D-Rex is model-agnostic and can be applied to any method of relation extraction and explanation extraction/generation.
• We show that D-Rex outperforms a strong baseline in explanation quality, with human annotators preferring D-Rex explanations over 90% of the time.
• We demonstrate that by conditioning on D-Rex extracted explanations, the current state-of-the-art model in relation extraction in dialogue can be improved by 13.5%.

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 \( E, R \) 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 \( s, o \in E \) are subject and object entities in the given dialogue and \( r \in R \) 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. Our work considers an explanation to be good if it strongly indicates that a relation holds, and for this purpose we consider triggers to be short explanations.

2.1 Relation Extraction

Given a dialogue \( d \), subject \( s \), and object \( o \), the goal of relation extraction is to predict the relation(s) that hold between \( s \) and \( o \). In addition to standard relation extraction, we also consider the problem of relation extraction with additional evidence in the form of a trigger or predicted explanation. Formally, this is the same as relation extraction, except that we are given an additional explanation, \( e_x \).

2.2 Explanation Extraction

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

3 Baseline Models

We first introduce approaches for relation extraction and explanation extraction based on state-of-the-art language models. We then propose a multitask approach that performs both tasks jointly. Our approaches use BERTbase (Devlin et al. 2019) and RoBERTa base (Liu et al. 2019) pre-trained models 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 \( e_x \), 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

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 \( W \in \mathbb{R}^{K \times H} \), 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 \( P_i = \text{sigmoid}(CW^T_i) \). We compute the standard cross-entropy loss for each relation as

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

where \( y_i \) denotes whether relation \( r_i \) holds.

For the standard relation extraction task, the input includes only the subject, object, and dialogue, which is replication of the current state-of-the-art, BERTs (Yu et al. 2020). For the relation extraction conditioned on explanation task, the input additionally includes the explanation concatenated to the front of the sequence.

3.2 Explanation Extraction

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

\[3\]Pre-trained models obtained from https://github.com/huggingface/transformers (Wolf et al. 2020)
the loss, we take the mean of the cross-entropy losses per token for the start position and add them to the mean of cross-entropy losses per token for the end position. Formally, let $|d|$ be the number of tokens in dialogue $d$, then

$$L_{EX} =\frac{1}{|d|}\sum_{i} \left( y_i^S \cdot \log(P^S_{Ti}) + (1 - y_i^S) \cdot \log(1 - P^S_{Ti}) \right) + \left( y_i^E \cdot \log(P^E_{Ti}) + (1 - y_i^E) \cdot \log(1 - P^E_{Ti}) \right)$$

(3)

where $y_i^S$ and $y_i^E$ 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][SEP][SEP][SEP][SEP], then we consider there to be no predicted explanation.

### 3.3 Joint Model

The joint relation extraction and explanation extraction model uses the standard input from §3.1. The joint model utilizes a BERT or RoBERTa backbone, and uses classification and span prediction layers identical to those in the relation extraction and explanation extraction models. Similarly, the loss is computed as the weighted sum of relation extraction and explanation extraction losses:

$$L_J = \alpha L_{RE} + (1 - \alpha) L_{EX}$$

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

One of the disadvantages of this joint prediction model is the challenge of including multiple explanations for multiple possible relations between a pair of entities. 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 simple method to simultaneously predict multiple relations and explanations, while determining which explanation belongs to which relation. 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. This idea is 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.

**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.). 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 denoted by $R(s, o, d)$.

**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, 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) with some adjustment to the formulation of the proposed 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 by $EX(r,s,o,d)$.

### Relation Re-Ranking Model (RR)

In our algorithm and discussions, we use $RR$ to denote the relation re-ranking model. In the D-REX training algorithm, $RR$ will be 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 $(e_x,s,o,d)$ and the output is a relation ranking, denoted as $RR(e_x,s,o,d)$.

#### 4.2 D-REX Algorithm

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 $e_x_i = EX(r_{pred_i}, s,o,d)$. Finally, for each predicted explanation, the re-ranking model computes $k$ probabilities for each relation type, using $(e_x_i, s,o,d)$ 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 both a policy gradient as well as supervision when available. For the policy gradient, 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 $L_{RE}^R(s,o,d)$ be the loss for $R$, given the subject, object, and dialogue: $s,o,d$. And let $L_{RE}^{RR}(e_x,s,o,d)$ be the loss of $RR$, given the explanation, subject, object, and dialogue: $e_x,s,o,d$. Then we define the relation re-ranking reward as:

$$R_{RR} = L_{RE}^R(s,o,d) - L_{RE}^{RR}(e_x,s,o,d)$$  \hspace{1cm} (4)

**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, Monroe, and Jurafsky 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 $e_x$ is the given explanation, let $d_{mask}(e_x)$ be the dialogue with $e_x$ replaced by mask tokens. Then, the leave-one-out reward is defined as:

$$R_{LOO} = L_{RE}(s,o,d_{mask}(e_x)) - L_{RE}(s,o,d)$$  \hspace{1cm} (5)

Because $L_{RE}$ 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 $e_x = d[i:j] = [T_i,T_{i+1},...,T_j]$ and the probabilities of $i,j$ being predicted are calculated as $P^a_i$ and $P^b_j$ according to equation [2].

---

**Algorithm 1:** The proposed training algorithm for D-REX

**Input:** Pre-trained ranking, explanation, and re-ranking models: $R$, $EX$, $RR$

**Data:** Dataset: $D$

```plaintext
for $(s,r,o,t,d)$ in $D$ do
    Compute ranking loss: $L_{RE}^R(s,o,d)$
    for $i$ in $r_{pred}$ do
        $e_x_i \leftarrow EX(r_{pred_i}, s,o,d)$
        Compute Re-ranking loss: $L_{RE}^{RR}(e_x_i,s,o,d)$
            // Equation 1
        Compute Re-Ranking Reward: $R_{RR}$
            // Equation 2
        Compute Leave-one-out Reward: $R_{LOO}$
            // Equation 3
    Compute policy gradient with rewards $R_{RR}$, $R_{LOO}$
        // Equation 4
    if $t$ not empty then
        Compute $L_{EX}$
            // Equation 3
    end
end
Update $EX$, $RR$ parameters with calculated losses
```

Because $R$ is stationary, $EX$ maximizes this function by minimizing $L_{RE}^{RR}$. Of course, $EX$ can only minimize $L_{RE}$ by interacting through its predicted explanations.

---
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

$$\mathcal{L}_{\text{EXP}} = -((P_{\pi}) + \log(P_{\pi}^c)) \ast (R_{\text{RR}} + \mathcal{L}_{\text{LOO}})$$  \hspace{1cm} (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, \(\mathcal{L}_{\text{EX}}\), as defined in equation $^5$.

Relation Extraction Re-ranking Model Optimization

We train RR with supervision. As discussed previously, this model takes as input a predicted explanation from the explanation model, in addition to the subject, object, and dialogue inputs. For RR we use a cross-entropy loss, \(\mathcal{L}_{\text{RR}}\), calculated in the same way as the ranking model.

5 Experimental Evaluation

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

5.1 Experimental settings

All models are implemented in PyTorch, 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 20 epochs, and D-REX models are further trained for at most 30 additional epochs with the best model determined based on relation extraction F1 scores computed on validation data. All experiments were repeated five times and we report the mean score along with standard deviation. D-REX models use a top-k of five and are initialized from the best performing models with the same backbone. For example, D-REXBERT uses two copies of \(R_{\text{BERT}}\) (Yu et al. 2020) to initialize the ranking and re-ranking models and \(E_{\text{BERT}}\) to initialize the explanation model. To train the joint model, we do not calculate \(\mathcal{L}_{\text{EX}}\) for relational triples which do not have a labeled trigger and we select \(\alpha\) from \{0.25, 0.5, 0.75\} and set \(\alpha\) to 0.5 based on validation performance.

DialogRE Dataset

We evaluate our models on the DialogRE English dataset which contains 36 relation types and 1,788 dialogues originating from the Friends TV show (Yu et al. 2020). For our experiments, we use the train/validation/test split V2-EN as defined by Yu et al. In the train/validation/test splits there are 6290/1992/1921 relational triples and 2446/830/780 labeled triggers. For D-REX, labeled triggers are only utilized during training, therefore models receive trigger supervision on 24% of the total relational triples.

---

Table 1: Relation extraction results. Yu et al. 2020 is equivalent to \(R_{\text{BERT}}\).

| Model                  | F1(\(\sigma\)) | MRR(\(\sigma\)) |
|------------------------|----------------|-----------------|
| Yu et al. 2020         | 59.2(1.9)      | 74.8(1.3)       |
| Joint \(R_{\text{BERT}}\) | 59.4(1.7)      | 74.0(0.9)       |
| D-REX \(R_{\text{BERT}}\) | 59.9(0.5)      | 75.4(0.1)       |
| \(R_{\text{RoBERTa}}\) | 64.2(1.6)      | 77.9(1.0)       |
| Joint \(R_{\text{RoBERTa}}\) | 65.2(0.3)     | 78.3(0.3)       |
| D-REX \(R_{\text{RoBERTa}}\) | 67.2(0.3)     | 79.4(0.3)       |

Table 2: Human evaluator preferences between different explanation extraction methods. NL and DL are samples where No Label exists, and where the predicted explanation differs from the Label, respectively. Results presented are percentages of preference.

| Method                | Win(%) | Tie(%) | Lose(%) |
|-----------------------|--------|--------|---------|
| Random (NL)           | 79.9   | 10.4   | 9.8     |
| Joint \(R_{\text{RoBERTa}}\) (NL) | 38.5   | 52.3   | 9.2     |
| Ground truth (DL)     | 12.1   | 44.3   | 43.7    |

Evaluation Metrics

Following Yu et al. we evaluate our models using F1 score. Additionally, we 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. We compute MRR as

$$MRR = \frac{\sum_{i=1}^{\mid\mathcal{E}\mid} \sum_{j=1}^{\mid\mathcal{R}\mid} 1}{\mid\mathcal{E}\mid}$$

where \(\mathcal{E}\) is the set of entity-pairs, \(\mathcal{R}\) is the set of relations, and rank\(_{ij}\) is the ranking of the jth ground truth relation for dialogue i.

To evaluate models in the explanation extraction task, we consider both automatic and manual evaluations. For automatic evaluations, we use a token-wise F1 score and exact match (EM); however, we only use this to measure baseline models as re-ranking models are encouraged to deviate from the labeled triggers.

5.2 Automatic Evaluation

In Table 1 we compare the previous state-of-the-art in relation extraction on dialogRE, \(R_{\text{BERT}}\) (Yu et al. 2020) with the methods presented in this paper. First, we see that even though D-REX is designed to introduce human-understandable explanations, it still has modest improvements over \(R_{\text{BERT}}\), which focuses on relation extraction, while Joint has no significant improvement. Finally, we see a five point absolute improvement in F1 from the baseline model when using RoBERTa, and an additional three point improvement when using D-REX. Note that there is less than a single point difference between \(R_{\text{BERT}}\) and D-REX \(R_{\text{BERT}}\), but
Comparative Analysis We employ crowd-workers for manual comparisons of D-REXRoBERTa against 3 baselines: randomly selected strings of 1-4 words, predictions from the joint model, and labeled triggers. For each comparison, 3 crowd-workers were given the full dialogue and a natural language statement corresponding to a relational triple, with explanations highlighted in the dialogue. The crowd-workers were then asked to specify which of the highlighted explanations was more indicative of the relation, or they could be equally indicative. 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 174 samples from D-REX with random strings and the joint model on NL, as well as 174 samples from DL. Table 2 shows the results. 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 new explanations. However, on DL, D-REX’s explanations were 3.6 times less likely to be preferred.

Table 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.

| Model   | F1  | Leave-one-out(↓) |
|---------|-----|------------------|
| D-REXRoBERTa (Full) | 67.2 | 83.9 |
| - reranking reward | 66.0 | 84.9 |
| - LOO reward | 67.1 | 85.4 |

Table 4: 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 in \[5.4\]. Smaller LOO is better.

5.3 Human Evaluation

To better understand how our model performs in extracting explanations and what challenges still exist, we perform two analyses manually: a comparative analysis and an absolute analysis. We consider two sets of data for evaluation: samples where No Labeled trigger exists (NL) and samples where the predicted explanation Differs from the Labeled trigger (DL).

Absolute Analysis To better understand the quality of D-REX’s explanations, we randomly sample 100 explanations from both NL and DL for a fine-grained analysis. We classify the explanations into 4 categories: indicative, incorrect entity-pair, incorrect relation, and not indicative. Table 3 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.

5.4 Ablation Study

To fairly assess the benefit of each proposed reward individually, we perform an ablation study on the reward functions. In order to study explanation quality, we introduce a new metric that we believe forms a fair automatic judgment of extracted explanation quality; the Leave-One-Out metric. For a relation extraction model R, an explanation extraction model EX, and a dataset D, LOO is calculated as

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

where F1_R(D) is the F1 score of R on D and D_{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.

For our experiments, we always calculate LOO using the baseline model, R_BERT. From the results in Table 4 we see that both reward functions are essential to the final results.
Compared with $R_{RoBERTa}$, D-REX$_{RoBERTa}$ 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.

### 6 Limitations and Future Work

This study focuses on relations and entities found in multi-party conversations, and while there are some similarities between the dialogue domain, medical literature, and Wikipedia (e.g., multi-entity, multi-relation), it is not clear whether the methods presented in this paper can transfer well 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. 2019) 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 tagger annotation is labor-intensive, and so 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 2017).

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, Yin, and Wang 2019, Lei, Barzilay, and Jaakkola 2016). There have even been methods developed which, similarly to our reranking, 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.

### 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.
References

Bronstein, O.; Dagan, I.; Li, Q.; Ji, H.; and Frank, A. 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), 372–376. Beijing, China: Association for Computational Linguistics.

Camburu, O.-M.; Rocktäschel, T.; Lukasiewicz, T.; and Blunsom, P. 2018. e-SNLI: Natural Language Inference with Natural Language Explanations. In Bengio, S.; Wallach, H.; Larochelle, H.; Grauman, K.; Cesa-Bianchi, N.; and Garnett, R., eds., Advances in Neural Information Processing Systems, volume 31. Curran Associates, Inc.

Chen, H.; Hong, P.; Han, W.; Majumder, N.; and Poria, S. 2020. Dialogue Relation Extraction with Document-level Heterogeneous Graph Attention Networks. ArXiv, abs/2009.05092.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 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), 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics.

Han, X.; and Wang, L. 2020. A Novel Document-Level Relation Extraction Method Based on BERT and Entity Information. IEEE Access, 8: 96912–96919.

Hancock, B.; Varma, P.; Wang, S.; Bringmann, M.; Liang, P.; and Ré, C. 2018. Training Classifiers with Natural Language Explanations. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 1884–1895. Melbourne, Australia: Association for Computational Linguistics.

Hendrickx, I.; Kim, S. N.; Kozaareva, Z.; Nakov, P.; Ó Séaghdha, D.; Padó, S.; Pennacchiotti, M.; Romano, L.; and Szpakowicz, S. 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, 33–38. Uppsala, Sweden: Association for Computational Linguistics.

Kumar, S.; and Talukdar, P. 2020. NILE: Natural Language Inference with Faithful Natural Language Explanations. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 8730–8742. Online: Association for Computational Linguistics.

Lei, T.; Barzilay, R.; and Jaakkola, T. 2016. Rationalizing Neural Predictions. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, 107–117. Austin, Texas: Association for Computational Linguistics.

Li, J.; Monroe, W.; and Jurafsky, D. 2016. Understanding Neural Networks through Representation Erasure. CoRR, abs/1612.08220.

Liu, H.; Yin, Q.; and Wang, W. Y. 2019. Towards Explainable NLP: A Generative Explanation Framework for Text Classification. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 5570–5581. Florence, Italy: Association for Computational Linguistics.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. CoRR, abs/1907.11692.

Loshchilov, I.; and Hutter, F. 2017. Fixing Weight Decay Regularization in Adam. CoRR, abs/1711.05101.

Ormándi, R.; Saleh, M.; Winter, E.; and Rao, V. 2021. WebRED: Effective Pretraining and Finetuning For Relation Extraction On The Web. CoRR, abs/2102.09681.

Peng, N.; Poon, H.; Quirk, C.; Toutanova, K.; and tau Yih, W. 2017. Cross-Sentence N-ary Relation Extraction with Graph LSTMs. Transactions of the Association for Computational Linguistics, 5: 101–115.

Quirk, C.; and Poon, H. 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, 1171–1182. Valencia, Spain: Association for Computational Linguistics.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; and Sutskever, I. 2019. Language Models are Unsupervised Multitask Learners.

Shahbazi, H.; Fern, X.; Ghaeini, R.; and Tadepalli, P. 2020. Relation Extraction with Explanation. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 6488–6494. Online: Association for Computational Linguistics.

Sutton, R. S.; and Barto, A. G. 2018. Reinforcement learning: An introduction. MIT press.

Wang, H.; Focke, C.; Sylvester, R.; Mishra, N.; and Wang, W. Y. 2019. Fine-tune Bert for DocRED with Two-step Process. CoRR, abs/1909.11898.

Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; Davison, J.; Shleifer, S.; von Platen, P.; Ma, C.; Jerneite, Y.; Plu, J.; Xu, C.; Scao, T. L.; Gugger, S.; Drame, M.; Lhoest, Q.; and Rush, A. M. 2020. Transformers: State-of-the-Art Natural Language Processing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 38–45. Online: Association for Computational Linguistics.

Yao, Y.; Ye, D.; Li, P.; Han, X.; Lin, Y.; Liu, Z.; Liu, Z.; Huang, L.; Zhou, J.; and Sun, M. 2019. DocRED: A Large-Scale Document-Level Relation Extraction Dataset. CoRR, abs/1906.06127.

Yu, D.; and Ji, H. 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), 44–53. Berlin, Germany: Association for Computational Linguistics.

Yu, D.; Sun, K.; Cardie, C.; and Yu, D. 2020. Dialogue-Based Relation Extraction. In Proceedings of the 58th Annual Meeting of the Association for Computational Lin-
guistics, 4927–4940. Online: Association for Computational Linguistics.

Yu, X.; Zhang, H.; Song, Y.; Song, Y.; and Zhang, C. 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), 5123–5132. Hong Kong, China: Association for Computational Linguistics.

Zahiri, S. M.; and Choi, J. D. 2017. Emotion Detection on TV Show Transcripts with Sequence-based Convolutional Neural Networks. CoRR, abs/1708.04299.

Zhang, Y.; Zhong, V.; Chen, D.; Angeli, G.; and Manning, C. D. 2017. Position-aware Attention and Supervised Data Improve Slot Filling. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 35–45. Copenhagen, Denmark: Association for Computational Linguistics.

Zhou, E.; and Choi, J. D. 2018. They Exist! Introducing Plural Mentions to Coreference Resolution and Entity Linking. In Proceedings of the 27th International Conference on Computational Linguistics, 24–34. Santa Fe, New Mexico, USA: Association for Computational Linguistics.

### Table 5: Trigger prediction results.

| model       | token F1(σ) | EM(σ) | LOO(σ) |
|-------------|-------------|-------|--------|
| EXBERT      | 62.1(3.1)   | 54.1(1.9) | 82.2(0.4) |
| JointBERT   | 43(1.3)     | 38.6(1.4) | 89.0(1.0) |
| D-REXBERT   | 50.5(1.1)   | 45.7(1.7) | 84.4(1.6) |
| EXRoBERTa   | 66.5(2.2)   | 58.4(2.0) | 82.0(0.4) |
| JointRoBERTa| 49(0.7)     | 47(0.7)   | 86.2(0.8) |
| D-REXRoBERTa| 57.2(2.1)   | 51.6(1.6) | 83.9(0.4) |

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 5 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. JointBERT scores nearly 20 points below its counterpart BERT model, while the JointRoBERTa 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 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 layout 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.
Dialogue 1

Which of the highlighted texts in the conversation below better indicate the following relation:

Speaker 2 and Speaker 1 are (or were) lovers.

Speaker 1: What did you just say?
Speaker 2: You roll another hard eight and we get married here tonight.
Speaker 1: Are you serious?!
Speaker 2: Yes! I love you! I've never loved anybody as much as I love you.
Speaker 1: I've never loved anybody as much as I love you.
Speaker 2: Okay, so if an eight comes up, we take it as a sign and we do it! What do you say?
Speaker 1: Okay!
Speaker 2: Okay! Come on! Let's go! All right!

- Yellow is a better indicator
- They are equal
- Orange is a better indicator

Figure 4