CodRED: A Cross-Document Relation Extraction Dataset for Acquiring Knowledge in the Wild

Yuan Yao[1*], **Jiaju Du**[1*], **Yankai Lin**[2], **Peng Li**[2], **Zhiyuan Liu**[1†], **Jie Zhou**[2], **Maosong Sun**[1]
[1]Department of Computer Science and Technology
Institute for Artificial Intelligence, Tsinghua University, Beijing, China
Beijing National Research Center for Information Science and Technology, China
[2]Pattern Recognition Center, WeChat AI, Tencent Inc.
yuan-yao18@mails.tsinghua.edu.cn, i@dujiaju.me

Abstract

Existing relation extraction (RE) methods typically focus on extracting relational facts between entity pairs within single sentences or documents. However, a large quantity of relational facts in knowledge bases can only be inferred across documents in practice. In this work, we present the problem of cross-document RE, making an initial step towards knowledge acquisition in the wild. To facilitate the research, we construct the first human-annotated cross-document RE dataset CodRED. Compared to existing RE datasets, CodRED presents two key challenges: Given two entities, (1) it requires finding the relevant documents that can provide clues for identifying their relations; (2) it requires reasoning over multiple documents to extract the relational facts. We conduct comprehensive experiments to show that CodRED is challenging to existing RE methods including strong BERT-based models. We make CodRED and the code for our baselines publicly available at https://github.com/thunlp/CodRED.

1 Introduction

Relation extraction (RE), which aims to extract relations between entities from plain text, serves as an essential resource in populating knowledge bases (KBs) from large-scale corpora automatically. Existing RE systems typically focus on either sentence-level RE (Socher et al., 2012; Zeng et al., 2014, 2015; Lin et al., 2016; Qin et al., 2018) or document-level RE (Li et al., 2016; Peng et al., 2017; Quirk and Poon, 2017; Yao et al., 2019), and have achieved promising results on several public benchmarks. However, these works can only extract relational facts from single sentences or documents containing both two target entities, which inevitably limits the coverage of knowledge acquisition. According to our statistics on Wikipedia documents, for over 57.6% of the relational facts in Wikidata (Erxleben et al., 2014; Vrandečić and Krötzsch, 2014), the head and tail entities do not co-occur in a single document. This inspires that it is crucial to break through the limitations of document boundaries to acquire knowledge in the wild.

In this work, we make an initial step in this direction, presenting the problem of *cross-document RE (cross-doc RE)*, which requires a RE system to infer the relation between two entities by retrieving and reasoning over multiple documents. Compared to conventional sentence/document-level RE, cross-doc RE presents new challenges in two levels of granularity: (1) at the coarse-grained level, given an entity pair, RE systems are required to find multiple informative documents for each entity, instead of restricted to the sentence/document containing both entities; (2) on the fine-grained level, RE systems are required to perform both intra- and cross-document reasoning in multiple documents and then predict the relations by aggregating information. The challenges come from not only the non-trivial nature of each phase, but also the intrinsic inter-dependence among the phases.

Fig. 1 shows an example for cross-doc RE, in which *Amun-her-khepeshef* and *Merneptah* do not co-appear in a single document. To identify their relation, we need to first retrieve the relevant documents for each entity and then recognize two reasoning text paths in these documents. The first reasoning text path (the documents titled "Nefertari" and "Memeptah") shows that both target entities are the son of *Ramesses II*, and the second one indicates that they also share a common sister *Meritamen*. The information of these two reasoning text paths is complementary to each other and suggests the relation between *Amun-her-khepeshef* and *Merneptah* is `sibling`.

Although several datasets have been proposed for investigating cross-document reasoning (Over and Yen, 2004; Yang et al., 2018), there is still

[*] indicates equal contribution
[†] Corresponding author: Z.Liu (liuzy@tsinghua.edu.cn)

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 4452–4472
November 7–11, 2021.

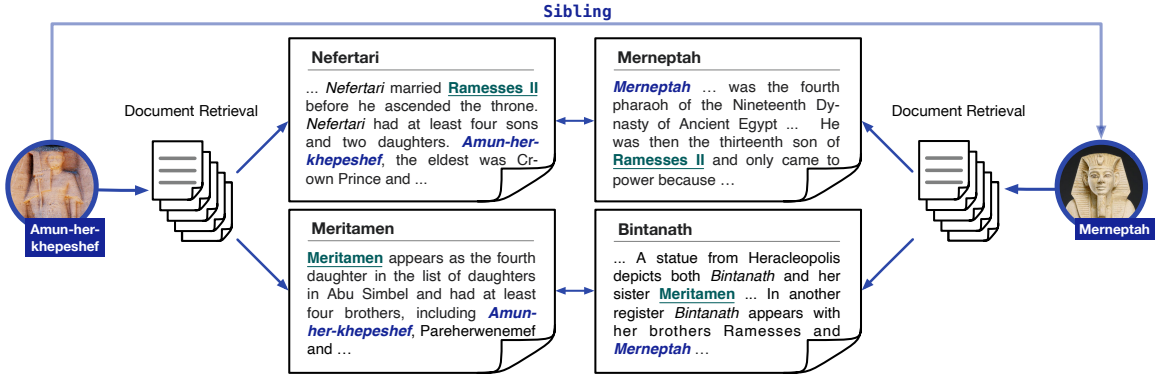

Figure 1: An example from CodRED. Two of the text paths between the target entity pair are shown. Each text path consists of two documents, which are connected by bridging entities. We only show evidence sentences in text paths for brevity. The ***target entity pair***, bridging entity *within* and **across** documents are highlighted accordingly.

no dataset designed for cross-doc RE. To facilitate the research, we construct the first human-annotated **Cr**oss-**d**ocument **R**elation **E**xtraction **D**ataset named as **CodRED**, aiming to test the RE systems' ability of knowledge acquisition in the wild. CodRED has the following features: (1) it requires natural language understanding in different granularity, including coarse-grained document retrieval, as well as fine-grained cross-document multi-hop reasoning; (2) it contains $30,504$ relational facts associated with $210,812$ reasoning text paths, as well as enjoys a broad range of balanced relations, and long documents in diverse topics; (3) it provides strong supervision about the reasoning text paths for predicting the relation, to help guide RE systems to perform meaningful and interpretable reasoning; (4) it contains adversarially-created hard NA instances to avoid RE models to predict relations by inferring from entity names instead of text information (Peng et al., 2020).

To assess CodRED, we propose two representative solutions based on the strong BERT-based RE architecture, including (1) a pipeline model that first extracts a relational graph for each document, and then reasons over these graphs to extract the relation; and (2) an end-to-end model that jointly considers text across different documents in text paths to predict the relation. We conduct comprehensive experiments under both closed and open settings on CodRED. Experimental results show that CodRED is very challenging to the strong BERT-based solutions, indicating ample room for further research.

2 Data Collection

In this section, we introduce the data collection process of the cross-doc RE dataset. Given an entity pair (h, t), cross-doc RE consists of two stages: (1) **Document Retrieval**, which finds multiple relevant documents of given entity pairs h and t from a large-scale corpus D, which could provide clues for identifying their relationship; (2) **Cross-Document Reasoning**, which reasons over the retrieved documents to predict the relation.

To focus on the problem of cross-doc RE, we only annotate the relational facts where the composing entities do not co-occur in a single document. As illustrated in Fig. 1, a relational fact can be better inferred through multiple complementary reasoning text paths (i.e., two documents that contain the head and tail entity respectively, and are connected by bridging entities) in the wild. Hence, we want to construct a cross-doc RE dataset in which each instance contains a relational fact with multiple reasoning text paths as well as strong supervision about supporting evidence.

However, it is infeasible for human annotators to label multiple reasoning text paths over documents for a relational fact from scratch. We thus carefully design a principled data collection pipeline for cross-doc RE. Specifically, we construct CodRED from the English Wikipedia and Wikidata through three stages: (1) Generating distantly supervised annotations from Wikipedia documents, which serves as relation recommendations for further human annotations; (2) Annotating relations and the corresponding supporting evidence by multiple independent crowdworkers; (3) Generating adversarial hard NA instances (i.e., entity pairs and text paths that do not express positive relations) to alleviate the reasoning shortcuts in RE. Here we introduce main procedure of data collection, and we refer readers to the appendix for more details.

2.1 Distantly Supervised Annotation Generation

To select relational facts and their relevant reasoning text paths for human annotations, we align Wikipedia articles with Wikidata under the distant supervision assumption (Mintz et al., 2009). To ensure the quality and encourage the diversity of the corpus, we select the articles in the Wikipedia popular page list[1] as the document candidates, which cover various topics in open domain. Specially, we first recognize named entity mentions in the documents by a BERT-based (Devlin et al., 2019) named entity recognition system. Then we link these named entity mentions to Wikidata, and merge the entity mentions with same IDs in Wikidata. Finally, we align each named entity pair from two different documents with their relations in Wikidata to form one reasoning text path of this entity pair.

However, different from previous works that adopt sentence-level (Riedel et al., 2010; Han et al., 2018) or document-level (Yao et al., 2019) distant supervision, we find that directly performing distant supervision for entities across documents will lead to substantial noise (i.e., over 95% raw relation labels from distant supervision are not expressed in the given text paths, according to manual verification on distantly supervised samples). To address the problem, we introduce additional requirements that there exists at least one relational reasoning chain between the target entity pair in two documents. Here, the reasoning chain is defined as a relational path between the entity pair (h, t), which is bridged by another entity e appearing in both documents, such that e has relation with h and t in Wikidata respectively. The reasoning chain can be formally denoted as $h \xrightarrow{r_i} e \xrightarrow{r_j} t$[2]. For example, in Fig. 1, *Amun-her-khepeshef* and *Merneptah* are linked by a reasoning chain consisting of two relational facts: *Amun-her-khepeshef* $\xrightarrow{\texttt{father}}$ *Ramesses II* $\xrightarrow{\texttt{child}}$ *Merneptah*. To alleviate the noise in reasoning chains, we ask experts to manually filter out frequent reasoning chains that cannot induce the target relations. We observe that this constraint can substantially alleviate the wrong-labeling problem, with less than 45% noise in the improved distantly supervised annotations.

In addition, we further sub-sample the annotations of frequent relations for two reasons: (1) to balance the relation distribution; (2) to prevent the strong correlation between the relations and documents (i.e., we make sure that the co-occurrence of any relation and document is fewer than 20), inspired by Welbl et al. (2018).

2.2 Human Annotation Generation

After obtaining distantly supervised relation annotations, we ask human annotators to label them to remove the noisy annotations in distant supervision. To ensure the dataset quality, we provide principled guidelines as well as training to the annotators, and utilize a test task to examine if the annotators understand our annotation principle. We also conduct regular quality inspections for each annotator, and update the feedback in the individual reports.

During the annotation, human annotators are asked to label (1) **text paths**, i.e., whether a relational fact can be supported by the given text path without external knowledge, and (2) **evidence sentences**, i.e., selecting a set of evidence sentences (if any) from the reasoning text path that can fully support the relational fact. Each reasoning text path is annotated independently by at least two annotators, and will be further annotated by a third annotator if there are disagreements in whether the relational fact can be supported.

After human annotation, each entity pair is associated with multiple reasoning text paths, which are labeled with either positive relations, or `NA` indicating no relation. The final relations between an entity pair are aggregated from all paths in between, by the union of the positive relations in each path. The final relation will be `NA` if there is no positive path in between. We discard the text paths if the relations can be extracted from one document, i.e., there are evidence annotations in only one document.

2.3 Adversarial NA Instance Generation

We find obvious reasoning shortcuts in our and most existing RE datasets (Peng et al., 2020), i.e., there are obvious correlations between some relations and entity names. This makes RE models could easily infer the relations from the entity names without performing complex reasoning in text, which may over-estimate their performance. To overcome this problem, we employ a novel adversarial `NA` instance generation strategy at entity-level, which requires RE models to pay more attention to understanding text. Moreover, we also add

[1] https://en.wikipedia.org/wiki/Wikipedia:Lists_of_popular_pages_by_WikiProject

[2] r_i and r_j can be relations or inverse relations in Wikidata.

Set	#Fact		#Path	
	Pos.	NA	Pos.	NA
Train	2,733	16,668	8,623	120,925
Dev	1,010	4,558	2,558	38,182
Test	1,012	4,523	2,505	38,019

Table 1: Statistics of data split. (#Fact: the number of relational facts; #Path: the number of reasoning text paths; Pos.: Positive.)

challenging path-level NA instances to test RE models' ability in reasoning in the presence of noise (i.e., there are NA text paths between entity pairs), which is important in real-world applications.

Entity-Level Adversarial NA Instance. We select challenging adversarial NA entity pairs, i.e., entity pairs that do not have relations in Wikidata but are assigned with high confidence of positive relations by RE models. Specifically, we first train a series of RE classifiers (i.e., CNN, LSTM and BERT, etc.) that extract the relations based on entity names.[3] Then for each positive entity pair (h, t), we generate an NA entity pair by replacing h or t with the top entity ranked by the confidence of the ensemble models. We generate $23,069$ adversarial NA entity pairs in total, reducing the percentage of positive instances to 15.6% in the dataset.[4]

Path-Level Adversarial NA Instance. To test the model ability of cross-document reasoning in the presence of noise in closed setting (see Sec. 3), we generate NA reasoning text paths for both human-annotated and adversarial NA entity pairs. Given an entity pair, we enumerate all possible reasoning text paths consisting of two documents that contain head and tail entities respectively, and share at least one common entity. To select hard NA paths, we choose the reasoning text paths that have the most shared entities between the composing documents.

3 Post-Processing and Benchmarks

We first introduce the data split process, including the split of positive and NA entity pairs. (1) **Positive entity pair split.** We split the positive entity pairs into training, development and test set, such that there is no overlap in entity names under the same positive relations, to prevent the correlation between relations and entity names. (2) **NA entity pair split.** Adversarially-created NA entity pairs (see Sec. 2.3) are randomly split into the three sets. Human-annotated NA entity pairs (see Sec. 2.2) are only put into training set to avoid the situation that there are unlabeled positive paths between the entity pair in open domain corpora, which could lead to false negative in evaluation in open setting (see following sections). Table 1 shows the statistics.

Since CodRED requires natural language understanding in different granularity, we design two benchmark settings to fully evaluate each required capability including (1) document retrieval, and (2) cross-document reasoning.

Closed Setting. In this setting, we test model capabilities in cross-document reasoning. Given an entity pair, RE models need to extract relations based on the given positive text paths and NA text paths. The first challenge comes from intra- and cross-document multi-hop reasoning in each text path. RE models need to first resolve complex interactions between entities within long documents, which may require logical, coreference and commonsense reasoning (Yao et al., 2019). Then RE models have to overcome the semantic gap between documents, and perform cross-document multi-hop reasoning through multiple potential bridging entities (4.7 on average) to establish the relation in each reasoning text path. The second challenge is that RE models need to synthesize all information in multiple text paths to obtain the final relation.

Open Setting. This setting fully tests the ability of RE in the wild. Given a target entity pair, models need to first retrieve relevant documents for the entity pair from full English Wikipedia corpus ($5,882,234$ documents in total, $3,646$ reasoning text path candidates for each entity pair on average), then perform cross-document reasoning with the retrieved documents to predict the relation. Compared with natural language queries in open domain question answering (Chen et al., 2017), the sparse query information in entity pairs presents unique challenges to document retrieval ability. The second challenge comes from both the quadratic number of potential paths (efficiency), and the fine-grained influence of document retrieval on the extraction of relations (effectiveness).

4 Data Analysis

In this section, we present data analysis of CodRED, including data statistics, required abilities in our dataset, and cross-document relation instances.

Data Statistics. CodRED enjoys diversity in open

[3] Here we use entity names to predict the relations, since we find it can effectively eliminate the reasoning shortcuts in our experiments, and also has better computation efficiency.

[4] The percentage reflects the sparsity of positive relations in real-world RE scenarios (Zhang et al., 2017).

Dataset	DR	CDR	IDR	ISR
TACRED				✓
FewRel				✓
KnowledgeNet				✓
BC5CDR			✓	✓
DialogRE			✓	✓
DocRED			✓	✓
CodRED	✓	✓	✓	✓

Table 2: Abilities required in different RE datasets. (DR: document retrieval, CDR: cross-document reasoning, IDR: intra-document reasoning, ISR: intra-sentence reasoning.)

domain in two aspects: relations and documents. (1) *Relations.* CodRED covers 276 relation types in different domains, including science (24.6%), work (21.3%) and art (8.7%), etc. Besides, CodRED contains 4,755 positive relational facts and 13,686 positive reasoning text paths, along with 25,749 NA relational facts and 197,126 NA reasoning text paths. CodRED exhibits balanced relation distribution, where the most frequent positive relation accounts for less than 4.5%. (2) *Documents.* The documents cover a variety of topics, including geography (28.7%), entertainment (19.6%), and society (8.5%), etc. The average length of documents is 2,416 words, presenting challenges for modeling long text in both efficiency and effectiveness. We refer readers to the appendix for more details.

Required Abilities. We compare required abilities of CodRED with existing RE datasets in Table 2, including (1) sentence-level RE datasets TACRED (Zhang et al., 2017), FewRel (Han et al., 2018) and KnowledgeNet (Mesquita et al., 2019), and (2) document-level RE datasets BC5CDR (Li et al., 2016), DocRED (Yao et al., 2019) and DialogRE (Yu et al., 2020). Compared with existing RE datasets that mainly focus on extracting relations from local contexts, i.e., single sentences or documents, CodRED presents unique challenges in document retrieval and cross-document reasoning.

Intra- and Cross-Document Reasoning. Cross-doc RE requires both intra- and cross-document multi-hop reasoning. For intra-document reasoning, we randomly sample 500 positive reasoning text paths and annotate the number of hops needed within the documents. 1.3 hops are required within documents on average, indicating that there are 2.6 hops in each path on average. For cross-document reasoning, a crucial challenge comes from multiple potential bridging entities between documents (4.7 on average). Each reasoning text path is labeled with 4.8 supporting sentences on average, accounting for 2.7% sentences in each path. This means that models need to select correct and meaningful sentences and bridging entities for cross-document reasoning from rich context and severe distractions.

5 Baselines

In this section, we design baseline models to assess the challenge of CodRED. In the closed setting, we design two representative baselines that perform cross-document reasoning based on strong architectures, including: (1) a *pipeline model* that first extracts a relational graph (i.e., graph containing entities and their relations) for each document, and then reasons over these graphs to extract the relation; and (2) an *end-to-end model* that jointly considers text across different documents in text paths to predict the relation. In the open setting, we first retrieve relevant documents and connect them into text paths, and then perform cross-document reasoning to predict the relation. We refer readers to the appendix for implementation details.

5.1 Document Retrieval

In the open setting, given an entity pair (h, t) and a document set D (i.e., full Wikipedia corpus), we first find relevant documents to extract their relation. Due to the large number of possible documents containing h and t respectively, we explore several strategies to retrieve the relevant documents and connect them into text paths. Specifically, we enumerate all possible text paths between the target entity pairs (i.e., two documents that contain h and t respectively with shared entities) as candidates. We first present a random baseline, where the candidate paths are randomly sampled. We also experiment with several heuristic retrieval strategies, where text paths are ranked by the heuristic scores. Specifically, the score of a text path (d_h, d_t) is given by: (1) *entity count*: multiplication of the occurrence number of h in d_h and the occurrence number of t in d_t, (2) *shared entity*: number of shared entities that appears in both d_h and d_t, or (3) *TF-IDF*: TF-IDF similarity (Manning et al., 2008) between the two documents. After ranking, we select top K paths with highest scores $\{(d_h^i, d_t^i)\}_{i=0}^{K}$.

5.2 Cross-Document Reasoning

Given the text paths between an entity pair, we present two baselines that perform cross-document reasoning for cross-doc RE, including a pipeline model and an end-to-end model.

5.2.1 Pipeline Model

We build a pipeline model that decomposes cross-document reasoning into three phases as follows:

1. **Intra-Document Relational Graph Extraction.** We predict the relations between the entities within each document containing head or tail entities using a BERT-based document-level RE model, resulting in a relational graph for each document.

2. **Cross-Document Relation Reasoning.** For each possible bridging entity e (i.e., any entity shared by two relational graphs), we predict the relation between the target entity pair based on the entity types of h, t and e, and relation r_i between (h, e), as well as the relation r_j between (e, t). Note that the prediction is only based on the relational graphs without considering text. Specifically, we feed the concatenation of the embeddings of r_i and r_j and embeddings of the types of h, t and e (e.g., person, organization and location) into a fully connected layer to obtain the relation distribution.

3. **Relation Aggregation.** We finally obtain the relation between the target entity pair by aggregating relation scores from all bridging entities. For each relation, the aggregated score is obtained by the max relation score from all possible bridging entities in all text paths.

5.2.2 End-to-end Model

Despite their simplicity, pipeline models usually suffer from error propagation. We also design an end-to-end model that jointly considers text across documents in text paths to predict the relation.

Specifically, given a text path, we adopt BERT as the text encoder. Since intra- and cross-document text understanding are both important components in cross-doc RE, we introduce two relation prediction tasks, including: (1) *Intra-document relation prediction*, where the model is asked to predict intra-document relations labeled by distant supervision as in Yao et al. (2019). (2) *Cross-document relation prediction*, where the model needs to predict cross-document relations labeled in CodRED.

Specifically, in cross-document relation prediction, documents are first concatenated and then tokenized. Then we add entity markers to mark the positions of head/tail/bridging entity mentions. The tokens are fed into BERT to obtain the text path representation $\mathbf{p}_i$. After that, to select meaningful paths in the presence of noise, following previous works on distantly supervised RE, we synthesize all informative paths by selective attention mechanism (Lin et al., 2016) and obtain the aggregated representation $\mathbf{x}$. The aggregated entity pair representation $\mathbf{x}$ is then fed into a fully connected layer followed by a softmax layer to obtain the distribution of the relation between the entity pair.

Besides the entity-level supervision, we also incorporate path-level supervision using an auxiliary classification task, where models need to predict the relation expressed in each path based on $\mathbf{p}_i$.

6 Experiments

In this section, we assess the challenges of CodRED in both closed and open benchmark settings.

6.1 Evaluation Metrics

In closed setting, following previous works (Zeng et al., 2015; Lin et al., 2016), we evaluate our model using aggregate precision-recall curves, and report the area under curve (AUC), the maximum F1 on the curve and Precision@K (P@K). In open setting, we first retrieve relevant documents (top 16 paths) from full Wikipedia corpus, and then use the models trained in the closed setting to infer the relation between the entity pair. We report the mean average precision (MAP), Recall@K (R@K) and mean reciprocal rank (MRR) to show the performance of document retrieval.

6.2 Overall Results

We report experimental results in both settings in Table 3, where document retrieval in open setting is based on the best performing entity count strategy. From the results we observe that: (1) The overall performance in the two benchmark settings is unsatisfactory for both baseline models, demonstrating the challenge of cross-doc RE. (2) The end-to-end model consistently outperforms the pipeline model by a large margin in both settings. This indicates that the pipeline model, i.e., simple adaptation of existing document-level RE approaches, cannot well handle cross-doc RE. The results show the necessity of developing RE models that jointly model text across different documents tailored for cross-doc RE. (3) The performance of models in open setting is significantly lower than their counterparts in closed setting. Document retrieval results in Table 4 also indicate that simple heuristic retrieval strategies cannot well serve cross-doc RE. In summary, the results show that CodRED is challenging to existing RE models, where retrieving relevant documents in open domain and reasoning over multiple documents present their unique challenges.

Setting	Model	Dev AUC	Dev F1	Dev P@500	Dev P@1000	Test AUC	Test F1	Test P@500	Test P@1000
Closed	Pipeline	17.45	30.54	30.60	26.70	18.94	32.29	32.00	28.70
	End-to-end	**47.94**	**51.26**	**62.80**	**51.00**	**47.46**	**51.02**	**65.00**	**51.20**
Open	Pipeline	14.07	26.45	27.00	19.90	16.26	28.70	30.00	24.50
	End-to-end	**40.86**	**47.23**	**59.00**	**46.30**	**39.05**	**45.06**	**57.80**	**45.10**

Table 3: Main results in two benchmark settings.

Strategy	MAP	MRR	R@16	R@100
Random	3.48	3.86	9.22	22.02
Shared Entity	7.61	8.25	19.74	42.14
TF-IDF	8.32	9.05	20.02	34.37
Entity Count	**19.83**	**23.87**	**36.73**	**59.66**

Table 4: Document retrieval results on the dev set.

Setting	Model	Sup.	AUC	F1	P@500
Closed	Pipeline	✓	16.65	30.54	26.20
			17.45	**30.85**	**30.60**
	End-to-end		45.30	48.55	60.80
		✓	**47.94**	**51.26**	**62.80**
Open	Pipeline	✓	13.52	26.45	22.40
			14.07	**26.51**	**27.00**
	End-to-end		**41.89**	46.19	57.40
		✓	40.86	**47.23**	**59.00**

Table 5: Ablation results on path-level supervision in two benchmark settings. Sup.: path-level supervision.

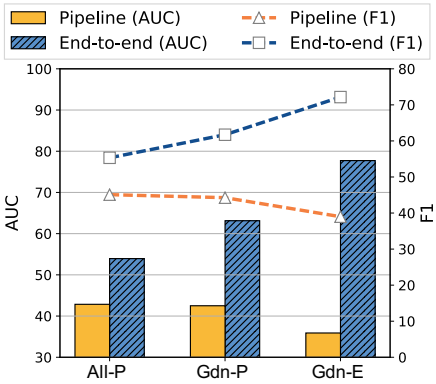

Figure 2: Experimental results when golden text paths or evidence sentences are given. All-P: all text paths, Gdn-P: golden text paths, Gdn-E: golden evidence.

6.3 Analysis

To provide better understanding of cross-doc RE and CodRED, we conduct comprehensive experiments and analysis. Unless otherwise specified, all the experiments and analysis are conducted in the closed setting on the development set.

Path-level Supervision. To investigate the effect of path-level annotations (i.e., annotation indicating whether the path expresses positive or NA relations) for cross-doc RE, we remove path-level supervision and report the results in Table 5, from which we observe that: (1) The model performance degrades in the closed setting when path-level supervision is removed. It indicates path-level supervision could effectively help to filter out the noise within multiple reasoning text paths. (2) In the open setting, the advantage of models supervised by path-level annotations shrinks. We hypothesize the reason is that the retrieved text paths in the open setting exhibit different distributions from the training set, making it difficult for the models to find the positive paths.

To verify the aforementioned hypothesis, we further evaluate the performance of relation classification given golden positive text paths and evidence sentences in the closed setting. Specifically, we remove the NA entity pairs, since they do not have golden positive text paths or evidence. Given an entity pair, we compare the performance of models that during evaluation are provided with (1) all text paths in between, (2) golden positive text paths, (3) golden evidence sentences. Results in Fig. 2 show that: (1) The performance of the end-to-end model improves significantly when golden positive text paths and evidence are given. This shows the importance and challenge of information selection from rich context for cross-doc RE. (2) The performance of the pipeline model degrades, since the number of reasoning chains in golden paths/evidence is very limited, which leads to over-fitting. (3) Extracting relations is challenging in CodRED even if golden evidence sentences are given, since models need to perform reasoning across multiple sentences and also overcome the semantic gap between different documents. In summary, the results indicate ample room for improvement in both selecting relevant information and reasoning over complex context.

Intra- v.s. Cross-Document Supervision. To investigate the importance of intra- and cross-document text understanding to cross-doc RE,

Model	AUC	F1	P@500	P@1000
End-to-end	**47.94**	**51.26**	**62.80**	**51.00**
w/o ID-Sup	15.67	26.56	33.20	26.50
w/o CD-Sup	10.14	17.21	19.00	15.40

Table 6: Ablation study on intra-document supervision (ID-Sup) and cross-document supervision (CD-Sup).

we ablate the corresponding supervision (see Sec. 5.2.2) and report the results in Table 6.[5] We observe dramatic drops in performance when removing either intra- or cross-document supervision. This shows that cross-doc RE requires deep text understanding both within and across documents.

Entity Names v.s. Context. Previous works have shown that RE systems tend to exploit shallow clues in existing datasets, i.e., predict relations based on entity names, instead of inferring from contexts (Peng et al., 2020). To investigate the contribution of each information source in CodRED, we ablate each information source and report the results in Table 7: (1) *Entity Only.* Models are given only names of the entity pair to predict their relation. (2) *Context Only.* The mentions of head and tail entities in documents are replaced by special mask tokens. Experimental results show that models struggle to predict relations only from entity names, and masking entity names does not dramatically hurt the performance. This indicates that there are no obvious correlations between relations and entity names in CodRED, due to the existence of adversarial `NA` entity pairs (see Sec. 2.3). In summary, although entity names can provide useful information in many cases, CodRED encourages RE models to infer relations by reasoning in rich context, instead of relying on shallow correlation between relations and entity names. In this sense, CodRED provides a more reasonable benchmark for knowledge acquisition systems.

7 Related Work

A variety of RE datasets have been constructed to promote the development of RE systems in recent years, which can be categorized in two main categories: (1) **Sentence-level RE datasets** focus on extracting relations on sentence-level, where the composing entities of a relational fact must co-appear in single sentences, with the relations labeled by human annotators (Doddington et al., 2004; Walker et al., 2006; Hendrickx et al., 2010; Han et al., 2018; Mesquita et al., 2019) or distant supervision (Riedel et al., 2010; Zhang et al., 2017; Elsahar et al., 2018). (2) **Cross-sentence RE datasets** focus on extracting cross-sentence relations from documents (Li et al., 2016; Peng et al., 2017; Quirk and Poon, 2017; Yao et al., 2019) or dialogues (Yu et al., 2020). Notably, NIST TAC SM-KBP 2019 Track[6] aims to extract and link document-level KBs from different languages and modalities. However, these datasets are still limited at sentence-level or document-level without considering cross-document reasoning, which restricts the coverage of knowledge acquisition. Hence, we extend RE to cross-document level, and construct a large-scale human-annotated dataset CodRED to facilitate further research.

Model	Ent.	Ctx.	AUC	F1	P@500
Entity Only	✓		10.46	21.19	21.70
Pipeline		✓	12.72	25.46	25.40
	✓	✓	17.45	30.54	30.60
End-to-end		✓	41.76	47.33	58.60
	✓	✓	**47.94**	**51.26**	**62.80**

Table 7: Ablation results on entity names (Ent.) and context (Ctx.).

Cross-document natural language understanding has received increasing interest in recent years. Several datasets have been constructed including cross-document question answering (Yang et al., 2018; Welbl et al., 2018) and cross-document summarization (Over and Yen, 2004; Owczarzak and Dang, 2011; Fabbri et al., 2019). In comparison with existing datasets, our dataset is tailored for the task of RE with fine-grained path and evidence annotations, and investigates the more open and challenging scenario of knowledge acquisition.

8 Conclusion

In this work, we study the problem of cross-doc RE. To facilitate the research for the problem, we present the first human-annotated dataset CodRED, and propose two representative solutions. Experimental results show that CodRED is challenging for strong RE models, indicating ample room for improvement. In this work, we focus on acquiring knowledge from text paths consisting of two documents. In the future, we plan to further explore longer text paths to better facilitate knowledge acquisition in the wild.

[5] The results do not include the pipeline model, since both supervisions are necessary for the model to infer the relation.

[6] https://tac.nist.gov/2019/SM-KBP/index.html

9 Ethical Considerations

In this section, we discuss the main ethical considerations of CodRED dataset: (1) Intellectual property protection. CodRED is constructed from Wikipedia and Wikidata, of which permissions are granted to copy, distribute and modify the contents under the terms of the Creative Commons Attribution-ShareAlike 3.0 Unported License and Creative Commons CC0 License respectively. (2) Privacy. The data collection procedure is designed for factual knowledge acquisition, and does not involve privacy issues. (3) Compensation. During relation annotation, the salary for annotating each relation instance is determined by the average time of annotation and local labor compensation standard. (4) Data characteristics. We refer readers to the appendix and data description file for more detailed characteristics of the dataset. (5) Potential problems. While principled measures are taken to ensure the quality of the dataset, there might still be potential problems with the dataset quality, which may lead to incorrect predictions in knowledge acquisition applications. However, moderate noise is common in large-scale modern KBs, even for human contributed contents, which should not cause serious issues.

10 Acknowledgement

This work is jointly funded by the Natural Science Foundation of China (NSFC) and the German Research Foundation (DFG) in Project Crossmodal Learning, NSFC 62061136001 / DFG TRR-169.

References

Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. 2019. Matching the blanks: Distributional similarity for relation learning. In *Proceedings of ACL*, pages 2895–2905.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to answer open-domain questions. In *Proceedings of ACL*, pages 1870–1879.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL: HLT*, pages 4171–4186.

George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, and Ralph Weischedel. 2004. The automatic content extraction (ACE) program-tasks, data, and evaluation. In *Proceedings of LREC*, pages 837–840.

Hady Elsahar, Pavlos Vougiouklis, Arslen Remaci, Christophe Gravier, Jonathon Hare, Frederique Laforest, and Elena Simperl. 2018. T-REx: A large scale alignment of natural language with knowledge base triples. In *Proceedings of LREC*.

Fredo Erxleben, Michael Günther, Markus Krötzsch, Julian Mendez, and Denny Vrandecic. 2014. Introducing wikidata to the linked data web. In *Proceedings of ISWC*, pages 50–65.

Alexander Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir Radev. 2019. Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model. In *Proceedings of ACL*, pages 1074–1084.

Xu Han, Hao Zhu, Pengfei Yu, Ziyun Wang, Yuan Yao, Zhiyuan Liu, and Maosong Sun. 2018. FewRel: A large-scale supervised few-shot relation classification dataset with state-of-the-art evaluation. In *Proceedings of EMNLP*, pages 4803–4809.

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 ACL*, pages 33–38.

Jiao Li, Yueping Sun, Robin J. Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J. Mattingly, Thomas C. Wiegers, and Zhiyong Lu. 2016. BioCreative V CDR task corpus: a resource for chemical disease relation extraction. *Database*, pages 1–10.

Yankai Lin, Shiqi Shen, Zhiyuan Liu, Huanbo Luan, and Maosong Sun. 2016. Neural relation extraction with selective attention over instances. In *Proceedings of ACL*, pages 2124–2133.

Christopher D Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to information retrieval*. Cambridge university press.

Filipe Mesquita, Matteo Cannaviccio, Jordan Schmidek, Paramita Mirza, and Denilson Barbosa. 2019. KnowledgeNet: A benchmark dataset for knowledge base population. In *Proceedings of EMNLP-IJCNLP*, pages 749–758.

Mike Mintz, Steven Bills, Rion Snow, and Daniel Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of ACL*, pages 1003–1011.

Paul Over and James Yen. 2004. An introduction to duc-2004. In *Proceedings of DUC*.

Karolina Owczarzak and Hoa Trang Dang. 2011. Overview of the tac 2011 summarization track: Guided task and aesop task. In *Proceedings of TAC*.

Hao Peng, Tianyu Gao, Xu Han, Yankai Lin, Peng Li, Zhiyuan Liu, Maosong Sun, and Jie Zhou. 2020. Learning from context or names? an empirical study on neural relation extraction. In *Proceedings of EMNLP*, pages 3661–3672.

Nanyun Peng, Hoifung Poon, Chris Quirk, Kristina Toutanova, and Wen-tau Yih. 2017. Cross-sentence n-ary relation extraction with graph LSTMs. *TACL*, 5:101–115.

Pengda Qin, Weiran Xu, and William Yang Wang. 2018. Robust distant supervision relation extraction via deep reinforcement learning. In *Proceedings of ACL*, pages 2137–2147.

Chris Quirk and Hoifung Poon. 2017. Distant supervision for relation extraction beyond the sentence boundary. In *Proceedings of EACL*, pages 1171–1182.

Sebastian Riedel, Limin Yao, and Andrew McCallum. 2010. Modeling relations and their mentions without labeled text. In *Proceedings of ECML-PKDD*, pages 148–163.

Richard Socher, Brody Huval, Christopher D. Manning, and Andrew Y. Ng. 2012. Semantic compositionality through recursive matrix-vector spaces. In *Proceedings of EMNLP-CoNLL*, pages 1201–1211.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*.

Christopher Walker, Stephanie Strassel, Julie Medero, and Kazuaki Maeda. 2006. ACE 2005 multilingual training corpus. *Linguistic Data Consortium, Philadelphia*, 57.

Johannes Welbl, Pontus Stenetorp, and Sebastian Riedel. 2018. Constructing datasets for multi-hop reading comprehension across documents. *TACL*, 6:287–302.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Processing of EMNLP*, pages 2369–2380.

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 ACL*, pages 764–777.

Dian Yu, Kai Sun, Claire Cardie, and Dong Yu. 2020. Dialogue-based relation extraction. In *Proceedings of ACL*, pages 4927–4940.

Daojian Zeng, Kang Liu, Yubo Chen, and Jun Zhao. 2015. Distant supervision for relation extraction via piecewise convolutional neural networks. In *Proceedings of EMNLP*, pages 1753–1762.

Daojian Zeng, Kang Liu, Siwei Lai, Guangyou Zhou, and Jun Zhao. 2014. Relation classification via convolutional deep neural network. In *Proceedings of COLING*, pages 2335–2344.

Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Manning Christopher D. 2017. Position-aware attention and supervised data improve slot filling. In *Proceedings of EMNLP*, pages 35–45.

A Data Collection Details

Named Entity Annotation. To generate distantly supervised relation annotations, we first annotate named entities in the documents by a named entity recognition system. We fine-tune a $\text{BERT}_{\text{LARGE}}$ (Devlin et al., 2019) model on DocRED (Yao et al., 2019), which achieves 0.91 F1 score on the DocRED validation set. Second, we link each entity mention to Wikidata by matching the mention to the name and aliases of the entities. We link the mention to the most frequent entity in Wikidata with the same name or aliases (if any). After linking the entity mentions to Wikidata, we merge the entity mentions in a document that are linked to the same entities to provide extra coreference information. Finally, each entity is associated with a set of documents that contain the entity.

Noisy Reasoning Chain Filtering. In distantly supervised annotation generation, we introduce requirements that there exists at least one reasoning chain between the labeled entity pair in the text path. To alleviate the noise in reasoning chains, we ask experts to filter out frequent noisy reasoning chains that cannot induce the target relations. Denote h, t, b as head, tail and bridging entities respectively. Generally, noisy reasoning chains can be categorized into two types as follows:

Type I. The relation between h and t is different from the relation induced from the reasoning chain. For example, relation `place of death` is different from the relation induced from $h \xrightarrow{\texttt{employer}} b \xrightarrow{\texttt{located in}} t$. Type I accounts for 37.8% noisy reasoning chains.

Type II. There is large uncertainty in inducing the relation from the reasoning chain. For example, relation `place of birth` cannot be induced from the relation $h \xrightarrow{\texttt{country of citizenship}} b \xrightarrow{\texttt{capital}} t$. Type II accounts for 62.2% noisy reasoning chains.

Human Annotation. The annotators mainly consist of undergraduate students, and receive principled training for 4 weeks on average to fully pass the test task and regular inspections. During annotation, in addition to the relational fact, we highlight the mentions of target entities and bridging entities, and provide possible reasoning chains to assist human annotation. The salary for each relation instance is determined by the average time of annotation and local labor compensation standard. We refer readers to data description in data supplement for the user interface of our annotation platform.

Adversarial Negative Instance Generation. To alleviate the obvious correlations between relations and entity names, we employ an adversarial negative instance generation strategy. Specifically, we select entity pairs that do not have relations in Wikidata but are assigned with high confidence of positive relations by RE models.

Given a positive entity pair (h, t), we generate negative entity pairs by replacing one of the entities. We first train several neural models that predicts the relation between an entity pair from entity names, including a BERT-based model, a CNN-based model, an LSTM-based model, a bilinear model, and a bag-of-words model. Specifically, the BERT-based model, CNN-based model, and LSTM-based model take the concatenation of entity names as input to predict the relation score. The bilinear model predicts the relation as follows:

$$s_r = \text{sigmoid}(\mathbf{h}\mathbf{M}_r\mathbf{t} + b_r), \tag{1}$$

where s_r is the score of relation r, $\mathbf{M}_r$ and b_r are learnable parameters. $\mathbf{h}$ and $\mathbf{t}$ are entity name embeddings obtained from BERT as follows:

$$\mathbf{h} = \text{BERT}(\text{h}) \tag{2}$$

$$\mathbf{t} = \text{BERT}(\text{t}), \tag{3}$$

where h and t are the name of the entity pair.

After that, for each positive entity pair (h, t), we generate a negative entity pair by replacing one of the entities. We first select top 100 entities as candidates using the bilinear model due to its efficiency, then select the top entities ranked by the ensemble models as negative entities.

Data Split. In positive entity pair split, we aim to split the positive entity pairs into training, development and test set, such that there is no overlap in entity names under the same positive relations, to prevent the correlation between relations and entity names. Specifically, for each positive relation, the corresponding relational facts are represented as an undirected graph, where nodes correspond to entity pairs, and there is an edge between two nodes if the entity pairs share a common entity. Then we randomly split the connected entity pairs in the graph into training, development and test set, and ensure that relations in development and test set appeared in the training set.

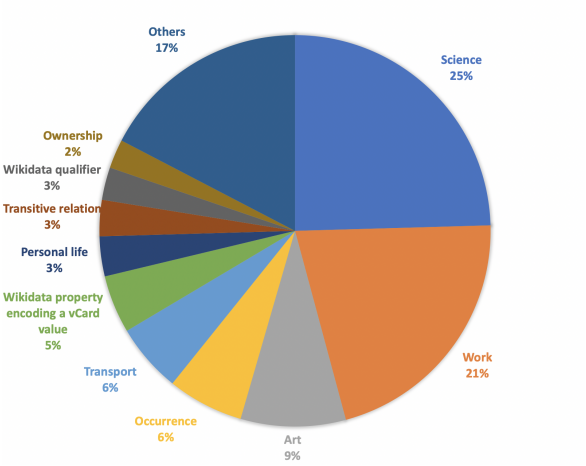

Figure 3: Relation domain distribution.

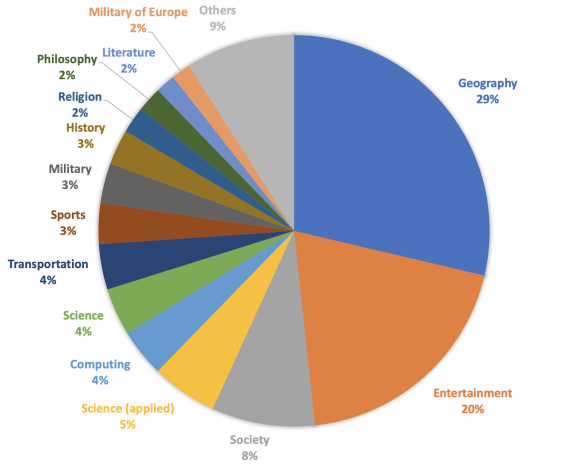

Figure 4: Document domain distribution.

B Data Distribution

We provide distribution of relations and documents. We show relation domain distribution in Fig. 3, and the document domain distribution in Fig. 4. We can see that CodRED enjoys diversity in both relations and documents. We also compare the length of documents in CodRED and existing document-level RE datasets in Table 8. We observe that documents in CodRED are much longer than those in existing document-level RE datasets, presenting new challenges to RE systems. We refer readers to data description in data supplement for relation documentation and reasoning chain distribution.

Dataset	Words/Doc.
BC5CDR (Li et al., 2016)	118
DocRED (Yao et al., 2019)	198
DialogRE (Yu et al., 2020)	226
CodRED	2,416

Table 8: Comparison of average document length between CodRED and document-level RE datasets.

C Baseline Implementation Details

We provide implementation details of the two proposed baseline methods, including the pipeline model and the end-to-end model. For both baselines, we adopt the $\text{BERT}_{\text{BASE}}$ (110M) implementation by Wolf et al. (2019).

C.1 Pipeline Model.

Intra-document Relational Graph Extraction. This phase aims to predict the relations between the entities within each document containing head or tail entities. Documents are first tokenized into word pieces (Wu et al., 2016). To extract the relation between two entities in a document, we mark the position of entity mentions. Specifically, inspired by Baldini Soares et al. (2019), we adopt special tokens as entity markers and insert them to the start and end of all mentions of an entity. Four special tokens (i.e., $\{\texttt{[UNUSED}i\texttt{]}\}_{i=0}^{3}$ from BERT vocabulary) are used to mark the start and end of two entities in a document.

After marking the entity mentions, we select relevant text in documents to encode. Since the documents in CodRED are typically very long, and the document length usually exceeds the 512 maximum input length of BERT, we extract text snippets surrounding the two entities in the document. Specifically: (1) if the distance between the nearest mentions of two entities is less than 512, we use the 512 tokens centered on the nearest mentions; (2) otherwise we extract 255 tokens centered on the first mention of each entity, and concatenate them to obtain the input tokens. A snippet will be shifted accordingly if the span encounters the document boundaries. A `[CLS]` token is put at the beginning, and a `[SEP]` token is concatenated at the end of the input tokens. Then we feed the tokens into BERT and take the `[CLS]` embedding in the last layer as the entity pair representation.

Finally, the entity pair representation is fed into a fully connected layer followed by a softmax layer to obtain the relation distribution. The target relations are labeled by distant supervision as in Yao et al. (2019). The intra-document relational graph extraction model achieves 53.75 F1 score on the validation set of DocRED (Yao et al., 2019).

Hyperparameters. The hyperparameters are selected by grid search based on AUC metric on the

validation set. The learning rate is 3e-5, selected from {2e-5, 3e-5, 5e-5}. The batch size is 32, selected from $\{16, 32, 64\}$. In cross-document relation reasoning phase, the dimension of entity type embedding and relation embedding is 256, selected from $\{128, 256\}$. We train our intra-document relational graph extraction model on 4 GeForce RTX 2080Ti GPUs for 2 epochs, which takes about 12 hours. The cross-document relation reasoning model is trained on a GeForce RTX 2080Ti GPUs for 20 epochs, which takes about 0.5 hours.

C.2 End-to-end Model

We provide details about the end-to-end model, including intra-document relation prediction and cross-document relation prediction. For intra-document relation prediction, we adopt the same approach in the pipeline model. Here we introduce details of cross-document relation prediction, including text path encoding and path aggregation.

Text Path Encoding. Given a text path (d_h^i, d_t^i) of an entity pair (h, t), we first encode it into representation. The text path encoding largely follows the implementation of the encoder of intra-document relational graph extraction model in the pipeline method. Documents are first concatenated and tokenized. Then entity markers are inserted to the start and end of all mentions of head, tail and bridging entities. We adopt unused tokens $\{\texttt{[UNUSEDi]}\}_{i=0}^{83}$ from BERT vocabulary as entity markers. $\{\texttt{[UNUSEDi]}\}_{i=0}^{3}$ are used to mark the start and end of head and tail entities. Bridging entities are marked by $\{\texttt{[UNUSEDi]}\}_{i=4}^{83}$ according to their occurrence order in the document containing the head entity. Next we select relevant text snippets surrounding the head/tail entity in the document in a similar approach to pipeline model. Finally we feed the tokens into BERT and take the `[CLS]` embedding in the last layer as the text path representation $\mathbf{p}_i$.

Path Aggregation. Given the representations of paths $\{\mathbf{p}_i\}_{i=0}^{K}$ between the entity pair, to select meaningful paths in the presence of noise, we adopt selective attention (Lin et al., 2016) to obtain the aggregated entity pair representation $\mathbf{x}$ as follows:

$$\mathbf{x} = \sum_{i \in K} \alpha_i \mathbf{p}_i, \tag{4}$$

where α_i is the weight of path p_i, and is defined as:

$$\alpha_i = \frac{\exp(e_i)}{\sum_{k \in K} \exp(e_k)}, \tag{5}$$

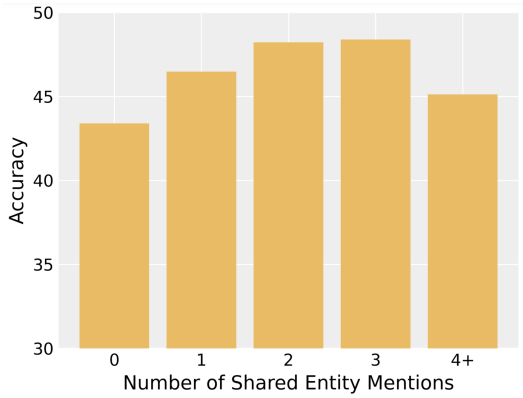

Figure 5: Accuracies of relation classification on text paths with different numbers of shared entity mentions on the development set.

where e_i is the attention score of the path p_i, which indicates how well the path p_i and the query relation r matches. Informative paths are expected to have higher attention scores. The attention score e_i is given by:

$$e_i = \mathbf{p}_i \mathbf{r}, \tag{6}$$

where $\mathbf{r}$ is the query embedding of relation r.

Hyperparameters. The hyperparameters are selected by grid search according to the AUC metric on the validation set. The learning rate is 3e-5, selected from {2e-5, 3e-5, 5e-5}. The batch size is selected 16, selected from $\{8, 16, 32\}$. We also choose the weight decay value 0.01. We train our model on 4 GeForce RTX 2080Ti GPUs for 10 epochs, which takes about 5 hours.

D Further Analysis

Performance w.r.t. Bridging Entities. Cross-doc RE requires cross-document reasoning via bridging entities. To investigate the challenge of cross-document reasoning with respect to possible bridging entity mentions, we report the model performance on text paths with different numbers of shared entity mentions between the two documents. Specifically, each text path (including positive and negative ones) is treated as an independent instance. We train a relation classification model that consists of a text path encoder and a relation predictor. Then we divide 2,558 positive text paths in the development set into subsets according to the number of shared entity mentions in the snippet, and report the accuracy of positive relation classification on each subset in Fig. 5. We observe that with the increase of shared entity mentions, the performance first improves slightly and then drops significantly.

We hypothesize the reason is that (1) when there are few shared entity mentions, increased shared entity mention number indicates smaller semantic gaps and more alternative reasoning chains between the documents; (2) when there is a number of shared entity mentions, further increment in the number will lead to complex context and severe distractions, making the reasoning process more challenging.

E Human Annotation.

During annotation, in addition to the relational fact, we highlight the mentions of target entities and bridging entities, and provide possible reasoning chains to assist human annotation. Fig. 6 shows the user interface of our annotation platform.

F Data Distribution

We provide the list of relations and their descriptions in Wikidata in Table 9, 10, 11, 12, 13 and 14.

Wikidata ID	Name	Description
P16	highway system	system (or specific country specific road type) of which the highway is a part
P17	country	sovereign state of this item; don't use on humans
P19	place of birth	most specific known (e.g. city instead of country, or hospital instead of city) birth location of a person, animal or fictional character
P20	place of death	most specific known (e.g. city instead of country, or hospital instead of city) death location of a person, animal or fictional character
P22	father	male parent of the subject. For stepfather, use "stepparent" (P3448)
P25	mother	female parent of the subject. For stepmother, use "stepparent" (P3448)
P26	spouse	the subject has the object as their spouse (husband, wife, partner, etc.). Use "unmarried partner" (P451) for non-married companions
P30	continent	continent of which the subject is a part
P36	capital	primary city of a country, state or other type of administrative territorial entity
P38	currency	currency used by item
P39	position held	subject currently or formerly holds the object position or public office
P40	child	subject has the object in their family as their offspring son or daughter (independently of their age)
P50	author	main creator(s) of a written work (use on works, not humans)
P53	family	family, including dynasty and nobility houses. Not family name (use P734 for family name).
P54	member of sports team	sports teams or clubs that the subject currently represents or formerly represented
P57	director	director(s) of film, TV-series, stageplay, video game or similar
P58	screenwriter	person(s) who wrote the script for subject item
P59	constellation	the area of the celestial sphere of which the subject is a part (from a scientific standpoint, not an astrological one)
P61	discoverer or inventor	the entity who discovered, first described, invented, or developed this discovery or invention
P69	educated at	educational institution attended by subject
P85	anthem	subject's official anthem
P86	composer	person(s) who wrote the music [for lyricist, use "lyrics by" (P676)]
P101	field of work	specialization of a person or organization; see P106 for the occupation
P102	member of political party	the political party of which this politician is or has been a member
P108	employer	person or organization for which the subject works or worked
P112	founded by	founder or co-founder of this organization, religion or place
P113	airline hub	airport that serves as a hub for an airline
P114	airline alliance	alliance the airline belongs to
P115	home venue	home stadium or venue of a sports team or applicable performing arts organization
P118	league	league in which team or player plays or has played in
P119	place of burial	location of grave, resting place, place of ash-scattering, etc, (e.g. town/city or cemetery) for a person or animal. There may be several places: e.g. re-burials, cenotaphs, parts of body buried separately.
P121	item operated	equipment, installation or service operated by the subject
P122	basic form of government	subject's government
P123	publisher	organization or person responsible for publishing books, periodicals, games or software
P126	maintained by	person or organization in charge of keeping the subject (for instance an infrastructure) in functioning order
P127	owned by	owner of the subject
P129	physically interacts with	physical entity that the subject interacts with
P131	located in the administrative territorial entity	the item is located on the territory of the following administrative entity. Use P276 (location) for specifying the location of non-administrative places and for items about events
P135	movement	literary, artistic, scientific or philosophical movement associated with this person or work
P136	genre	creative work's genre or an artist's field of work (P101). Use main subject (P921) to relate creative works to their topic
P137	operator	person or organization that operates the equipment, facility, or service; use country for diplomatic missions
P138	named after	entity or event that inspired the subject's name, or namesake (in at least one language)
P140	religion	religion of a person, organization or religious building, or associated with this subject
P144	based on	the work(s) used as the basis for subject item
P149	architectural style	architectural style of a structure
P150	contains administrative territorial entity	(list of) direct subdivisions of an administrative territorial entity
P155	follows	immediately prior item in a series of which the subject is a part [if the subject has replaced the preceding item, e.g. political offices, use "replaces" (P1365)]

Table 9: Relation list of CodRED, including Wikidata IDs, names and descriptions of relations.

Wikidata ID	Name	Description
P156	followed by	immediately following item in a series of which the subject is a part [if the subject has been replaced, e.g. political offices, use "replaced by" (P1366)]
P159	headquarters location	specific location where an organization's headquarters is or has been situated. Inverse property of "occupant" (P466).
P161	cast member	actor performing live for a camera or audience [use "character role" (P453) and/or "name of the character role" (P4633) as qualifiers] [use "voice actor" (P725) for voice-only role]
P162	producer	person(s) who produced the film, musical work, theatrical production, etc. (for film, this does not include executive producers, associate producers, etc.) [for production company, use P272, video games - use P178]
P169	chief executive officer	highest-ranking corporate officer appointed as the CEO within an organization
P170	creator	maker of this creative work or other object (where no more specific property exists)
P171	parent taxon	closest parent taxon of the taxon in question
P175	performer	performer involved in the performance or the recording of a musical work
P176	manufacturer	manufacturer or producer of this product
P177	crosses	obstacle (body of water, road, ...) which this bridge crosses over or this tunnel goes under
P178	developer	organisation or person that developed the item
P179	series	subject is part of a series, the sum of which constitutes the object
P180	depicts	depicted entity (see also P921: main subject)
P184	doctoral advisor	person who supervised the doctorate or PhD thesis of the subject
P190	twinned administrative body	twin towns, sister cities, twinned municipalities and other localities that have a partnership or cooperative agreement, either legally or informally acknowledged by their governments
P193	main building contractor	the main organization responsible for construction of this structure or building
P197	adjacent station	the stations next to this station, sharing the same line(s)
P199	business division	divisions of this organization
P205	basin country	country that have drainage to/from or border the body of water
P206	located in or next to body of water	sea, lake or river
P241	military branch	branch to which this military unit, award, office, or person belongs, e.g. Royal Navy
P264	record label	brand and trademark associated with the marketing of subject music recordings and music videos
P272	production company	company that produced this film, audio or performing arts work
P276	location	location of the item, physical object or event is within. In case of an administrative entity use P131. In case of a distinct terrain feature use P706.
P279	subclass of	all instances of these items are instances of those items; this item is a class (subset) of that item. Not to be confused with P31 (instance of)
P282	writing system	alphabet, character set or other system of writing used by a language, supported by a typeface
P286	head coach	on-field manager or head coach of a sports club (not to be confused with a general manager P505, which is not a coaching position) or person
P287	designed by	person(s) that designed the item
P291	place of publication	geographical place of publication of the edition (use 1st edition when referring to works)
P306	operating system	operating system (OS) on which a software works or the OS installed on hardware
P355	subsidiary	subsidiary of a company or organization, opposite of parent organization (P749)
P360	is a list of	common element between all listed items
P361	part of	object of which the subject is a part (it's not useful to link objects which are themselves parts of other objects already listed as parts of the subject). Inverse property of "has part" (P527, see also "has parts of the class" (P2670)).
P366	use	main use of the subject (includes current and former usage)
P375	space launch vehicle	type of rocket or other vehicle for launching subject payload into outer space
P397	parent astronomical body	major astronomical body the item belongs to
P398	child astronomical body	minor body that belongs to the item
P400	platform	platform for which a work was developed or released, or the specific platform version of a software product
P403	mouth of the watercourse	the body of water to which the watercourse drains
P404	game mode	a video game's available playing mode(s)
P408	software engine	software engine employed by the subject item
P411	canonization status	stage in the process of attaining sainthood per the subject's religious organization
P414	stock exchange	exchange on which this company is traded
P421	located in time zone	time zone for this item
P425	field of this occupation	activity corresponding to this occupation (use only for occupations - for people use Property:P101, for companies use P452)
P437	distribution	method (or type) of distribution for the subject
P449	original network	network(s) the radio or television show was originally aired on, including

Table 10: Relation list of CodRED, including Wikidata IDs, names and descriptions of relations.

Wikidata ID	Name	Description
P451	partner	someone in a relationship without being married. Use "spouse" for married couples.
P452	industry	industry of company or organization
P460	said to be the same as	this item is said to be the same as that item, but the statement is disputed
P461	opposite of	item that is the opposite of this item
P462	color	color of subject
P463	member of	organization or club to which the subject belongs. Do not use for membership in ethnic or social groups, nor for holding a position such as a member of parliament (use P39 for that).
P479	input method	input method or device used to interact with a software product
P483	recorded at	studio or location where a musical composition was recorded
P485	archives at	the institution holding the subject's archives
P488	chairperson	presiding member of an organization, group or body
P489	currency symbol description	item with description of currency symbol
P495	country of origin	country of origin of this item (creative work, food, phrase, product, etc.)
P504	home port	home port of the vessel (if different from "ship registry"): For civilian ships, the primary port from which the ship operates. Port of registry P532 should be listed in "Ship registry". For warships, this will be the ship's assigned naval base
P509	cause of death	underlying or immediate cause of death. Underlying cause (e.g. car accident, stomach cancer) preferred. Use 'manner of death' (P1196) for broadest category, e.g. natural causes, accident, homicide, suicide
P511	honorific prefix	word or expression used before a name, in addressing or referring to a person
P512	academic degree	academic degree that the person holds
P516	powerplant	equipment or engine used to power the subject
P520	armament	equippable weapon item for the subject
P521	scheduled service destination	airport or station connected by regular direct service to the subject; for the destination of a trip see P1444
P523	temporal range start	the start of a process or appearance of a life form relative to the geologic time scale
P527	has part	part of this subject. Inverse property of "part of" (P361). See also "has parts of the class" (P2670)
P546	docking port	intended docking port for a spacecraft
P551	residence	the place where the person is or has been, resident
P553	website account on	a website that the person or organization has an account on (use with P554) Note: only used with reliable source or if the person or organization disclosed it.
P559	terminus	the feature (intersecting road, train station, etc.) at the end of a linear feature
P598	commander of	for persons who are notable as commanding officers, the units they commanded
P607	conflict	battles, wars or other military engagements in which the person or item participated
P608	exhibition history	exhibitions where the item is or was displayed
P610	highest point	point with highest elevation in a region, on a path, of a race
P611	religious order	order of monks or nuns to which an individual or religious house belongs
P629	edition or translation of	is an edition or translation of this entity
P658	tracklist	songs contained in this item
P664	organizer	person or institution organizing an event
P674	characters	characters which appear in this item (like plays, operas, operettas, books, comics, films, TV series, video games)
P676	lyrics by	author of song lyrics; also use P86 for music composer
P703	found in taxon	the taxon in which the item can be found
P706	located on terrain feature	located on the specified landform. Should not be used when the value is only political/administrative (P131) or a mountain range (P4552).
P707	satellite bus	general model on which multiple-production satellite spacecraft is based
P710	participant	person, group of people or organization (object) that actively takes/took part in an event or process (subject). Preferably qualify with "object has role" (P3831). Use P1923 for team participants.
P725	voice actor	performer of a spoken role in a creative work such as animation, video game, radio drama, or dubbing over [use "character role" (P453) as qualifier] [use "cast member" (P161) for live acting]
P737	influenced by	this person, idea, etc. is informed by that other person, idea, etc., e.g. "Heidegger was influenced by Aristotle".
P739	ammunition	cartridge or other ammunition used by the subject weapon
P740	location of formation	location where a group or organization was formed
P747	has edition	link to an edition of this item
P749	parent organization	parent organization of an organisation, opposite of subsidiaries (P355)
P750	distributor	distributor of a creative work; distributor for a record label
P751	introduced feature	feature introduced by this version of a product item

Table 11: Relation list of CodRED, including Wikidata IDs, names and descriptions of relations.

Wikidata ID	Name	Description
P767	contributor(s) to the creative work	person or organization that contributed to a subject: co-creator of a creative work
P769	significant drug interaction	clinically significant interaction between two pharmacologically active substances (i.e., drugs and/or active metabolites) where concomitant intake can lead to altered effectiveness or adverse drug events.
P790	approved by	item is approved by other item(s) [qualifier: statement is approved by other item(s)]
P793	significant event	significant or notable events associated with the subject
P800	notable work	notable scientific, artistic or literary work, or other work of significance among subject's works
P832	public holiday	official public holiday that occurs in this place in its honor, usually a non-working day
P840	narrative location	the narrative of the work is set in this location
P852	ESRB rating	North American video game content rating - appropriate values are on property's talk page
P859	sponsor	organization or individual that sponsors this item
P880	CPU	central processing unit found within the subject item
P915	filming location	actual place where this scene/film was shot. For the setting, use "narrative location" (P840)
P921	main subject	primary topic of a work (see also P180: depicts)
P924	medical treatment	treatment that might be used to heal the medical condition
P931	place served by transport hub	territorial entity or entities served by this transport hub (airport, train station, etc.)
P937	work location	location where persons were active
P941	inspired by	work, human, place or event which inspired this creative work or fictional entity
P944	Code of nomenclature	the Code that governs the scientific name of this taxon
P945	allegiance	the country (or other power) that the person, or organization, served
P974	tributary	stream or river that flows into this main stem (or parent) river
P1001	applies to jurisdiction	the item (an institution, law, public office ...) or statement belongs to or has power over or applies to the value (a territorial jurisdiction: a country, state, municipality, ...)
P1027	conferred by	person or organization who awards a prize to or bestows an honor upon a recipient
P1037	director/manager	person who manages any kind of group
P1038	relative	family member (qualify with "type of kinship", P1039; for direct family member please use specific property)
P1050	medical condition	any state relevant to the health of an organism, including diseases and positive conditions
P1056	product or material produced	material or product produced by a government agency, business, industry, facility, or process
P1066	student of	person who has taught this person
P1071	location of final assembly	place where the item was made; location of final assembly
P1072	readable file format	file format a program can open and read
P1073	writable file format	file format a program can create and/or write to
P1079	launch contractor	organization contracted to launch the rocket
P1080	from fictional universe	subject's fictional entity is in the object narrative. See also P1441 and P1445
P1142	political ideology	political ideology of this organization or person
P1158	location of landing	location where the craft landed
P1192	connecting service	service stopping at a station
P1303	instrument	musical instrument that a person plays
P1308	officeholder	person who holds an office
P1327	professional or sports partner	person a professional or athlete works with
P1336	territory claimed by	administrative divisions that claim control of a given area
P1343	described by source	dictionary, encyclopaedia, etc. where this item is described
P1344	participant of	event a person or an organization was/is a participant in, inverse of P710 or P1923
P1346	winner	winner of an event or award - do not use for wars or battles
P1365	replaces	person or item replaced. Use P1398 (structure replaces) for structures. Use P155 (follows) if the previous item was not replaced or if predecessor and successor are identical.
P1366	replaced by	other person or item which continues the item by replacing it in its role. Use P156 (followed by) if the item is not replaced (e.g. books in a series), nor identical, but adds to the series without dropping the role of this item in that series
P1387	political alignment	political position within the political spectrum
P1389	product certification	certification for a product, qualify with P1001 ("applies to jurisdiction") if needed
P1399	convicted of	crime a person was convicted of
P1408	licensed to broadcast to	place that a radio/TV station is licensed/required to broadcast to

Table 12: Relation list of CodRED, including Wikidata IDs, names and descriptions of relations.

Wikidata ID	Name	Description
P1411	nominated for	award nomination received by a person, organisation or creative work (inspired from "award received" (Property:P166))
P1414	GUI toolkit or framework	framework or toolkit a program uses to display the graphical user interface
P1416	affiliation	organization that a person or organization is affiliated with
P1427	start point	starting place of this journey, flight, voyage, trek, migration etc.
P1431	executive producer	executive producer of a movie or TV show
P1433	published in	larger work that a given work was published in, like a book, journal or music album
P1434	takes place in fictional universe	the subject is a work describing a fictional universe, i.e. whose plot occurs in this universe.
P1435	heritage designation	heritage designation of a cultural or natural site
P1441	present in work	work in which this fictional entity (Q14897293) or historical person is present (use P2860 for works citing other works and P361/P1433 for works being part of / published in other works)
P1444	destination point	intended destination for this route (journey, flight, sailing, exploration, migration, etc.)
P1445	fictional universe described in	to link a fictional universe with a work that describes it: <universe> "described in the work:" <work>
P1454	legal form	legal form of an organization
P1532	country for sport	country a person or a team represents when playing a sport
P1535	used by	item or concept that makes use of the subject (use sub-properties when appropriate)
P1557	manifestation of	embodiment of a given concept
P1582	natural product of taxon	links a natural product with its source (animal, plant, fungal, algal, etc.)
P1622	driving side	side of the road that vehicles drive on in a given jurisdiction
P1716	brand	brand of a product
P1830	owner of	entities owned by the subject
P1876	vessel	vessel involved in this mission, voyage or event
P1877	after a work by	artist whose work strongly inspired/ was copied in this item
P1891	signatory	person, country, or organization that has signed an official document (use P50 for author)
P1923	participating team	Like 'Participant' (P710) but for teams. For an event like a cycle race or a football match you can use this property to list the teams and P710 to list the individuals (with 'member of sports team' (P54)' as a qualifier for the individuals)
P1990	species kept	taxa, preferably species, present at a zoo, botanical garden, collection, or other institution. NOT specific animals, not for any geographic location
P1995	health specialty	main specialty that diagnoses, prevent human illness, injury and other physical and mental impairments
P2079	fabrication method	method, process or technique used to grow, cook, weave, build, assemble, manufacture the item
P2094	competition class	official classification by a regulating body under which the subject (events, teams, participants, or equipment) qualifies for inclusion
P2175	medical condition treated	disease that this pharmaceutical drug, procedure, or therapy is used to treat
P2176	drug used for treatment	drug, procedure, or therapy that can be used to treat a medical condition
P2283	uses	item or concept used by the subject or in the operation
P2321	general classification of race participants	classification of race participants
P2341	indigenous to	area or ethnic group that a language, folk dance, cooking style, food or other cultural expression is found (or was originally found)
P2348	time period	time period (historic period or era, sports season, theatre season, legislative period etc.) in which the subject occurred
P2360	intended public	this work, product, object or event is intended for, or has been designed to that person or group of people, animals, plants, etc
P2389	organisation directed from the office or person	
P2408	set in period	historical, contemporary or future period the work is set in
P2416	sports discipline competed in	discipline an athlete competed in within a sport
P2499	league level above	the league above this sports league
P2500	league level below	the league below this sports league
P2522	victory	competition or event won by the subject
P2541	operating area	area this organisation operates in, serves or has responsibility for
P2546	sidekick of	close companion of a fictional character
P2564	Köppen climate classification	indicates the characteristic climate of a place
P2579	studied by	subject is studied by this science or domain
P2670	has parts of the class	the subject instance has parts of the object class (the subject is usually not a class)
P2743	this zoological name is coordinate with	links coordinate zoological names
P2789	connects with	item with which the item is physically connected

Table 13: Relation list of CodRED, including Wikidata IDs, names and descriptions of relations.

Wikidata ID	Name	Description
P2852	emergency phone number	telephone number to contact the emergency services
P2853	electrical plug type	standard plug type for mains electricity in a country
P2860	cites	citation from one creative work to another
P2868	subject has role	role or generic identity of subject (the item that the statement is on) in a certain context. For acting roles, use P453 ("character role"). For roles of the object/value of statements, use P3831 ("object has role").
P2925	domain of saint or deity	domain(s) which this saint or deity controls or protects
P2935	connector	connectors which the device has/supports
P2962	title of chess player	title awarded by a chess federation to chess players for achievement
P3018	located in protected area	the protected area a place or geographical feature belongs to
P3033	package management system	package management system used to publish the software
P3075	official religion	official religion in this administrative entity
P3091	mount	creature ridden by the subject, for instance a horse
P3095	practiced by	type of agents that study this subject or work in this profession
P3137	parent peak	parent is the peak whose territory this peak resides in, based on the contour of the lowest col
P3320	board member	member(s) of the board for the organization
P3342	significant person	person linked to the item in any possible way
P3373	sibling	the subject has the object as their sibling (brother, sister, etc.). Use "relative" (P1038) for siblings-in-law (brother-in-law, sister-in-law, etc.) and step-siblings (step-brothers, step-sisters, etc.)
P3448	stepparent	subject has the object as their stepparent
P3494	points classification	
P3966	programming paradigm	programming paradigm in which a programming language is classified
P4000	has fruit type	morphology of the fruit of this taxon, as defined in botany
P4044	therapeutic area	disease area in which a medical intervention is applied
P4132	linguistic typology	classification of languages according to their linguistic trait (as opposed to historical families like romance languages)
P4387	update method	method used by an app/OS to receive updates or self-update
P4446	reward program	reward program associated with the item
P4552	mountain range	range or subrange to which the geographical item belongs
P4614	drainage basin	area where precipitation collects and drains off into a common outlet, such as into a river, bay, or other body of water
P4743	animal breed	subject item belongs to a specific group of domestic animals, generally given by association
P4791	commanded by	commander of a military unit/army/security service, operation, etc.
P5025	gens	a group of families from Ancient Rome who shared the same nomen
P5096	member of the crew of	person who has been a member of a crew associated with the vessel or spacecraft. For spacecraft, inverse of crew member (P1029), backup or reserve team or crew (P3015)
P5658	railway traffic side	indicates for a country or a railway line whether rail traffic usually runs on the left or right hand side
P5826	majority opinion by	judicial opinion agreed to by more than half of the members of a court
P5869	model item	defines which item is a best practice example of modelling a subject, which is described by the value of this property, usage instructions at Wikidata:Model item
P5995	kit supplier	official supplier of sports goods to a given club or a national sports team
P6216	copyright status	copyright status for intellectual creations like works of art, publications, software, etc.
P6275	copyright representative	person or organisation who represents the copyright for this person or work of art
P6379	has works in the collection	collection that have works of this artist
P6885	historical region	geographic area which at some point in time had a cultural, ethnic, linguistic or political basis, regardless of present-day borders
P6942	animator	person creating animated sequences out of still images
P7047	enemy of	opponent character or group of this character or group
P7153	significant place	significant or notable places associated with the subject

Table 14: Relation list of CodRED, including Wikidata IDs, names and descriptions of relations.

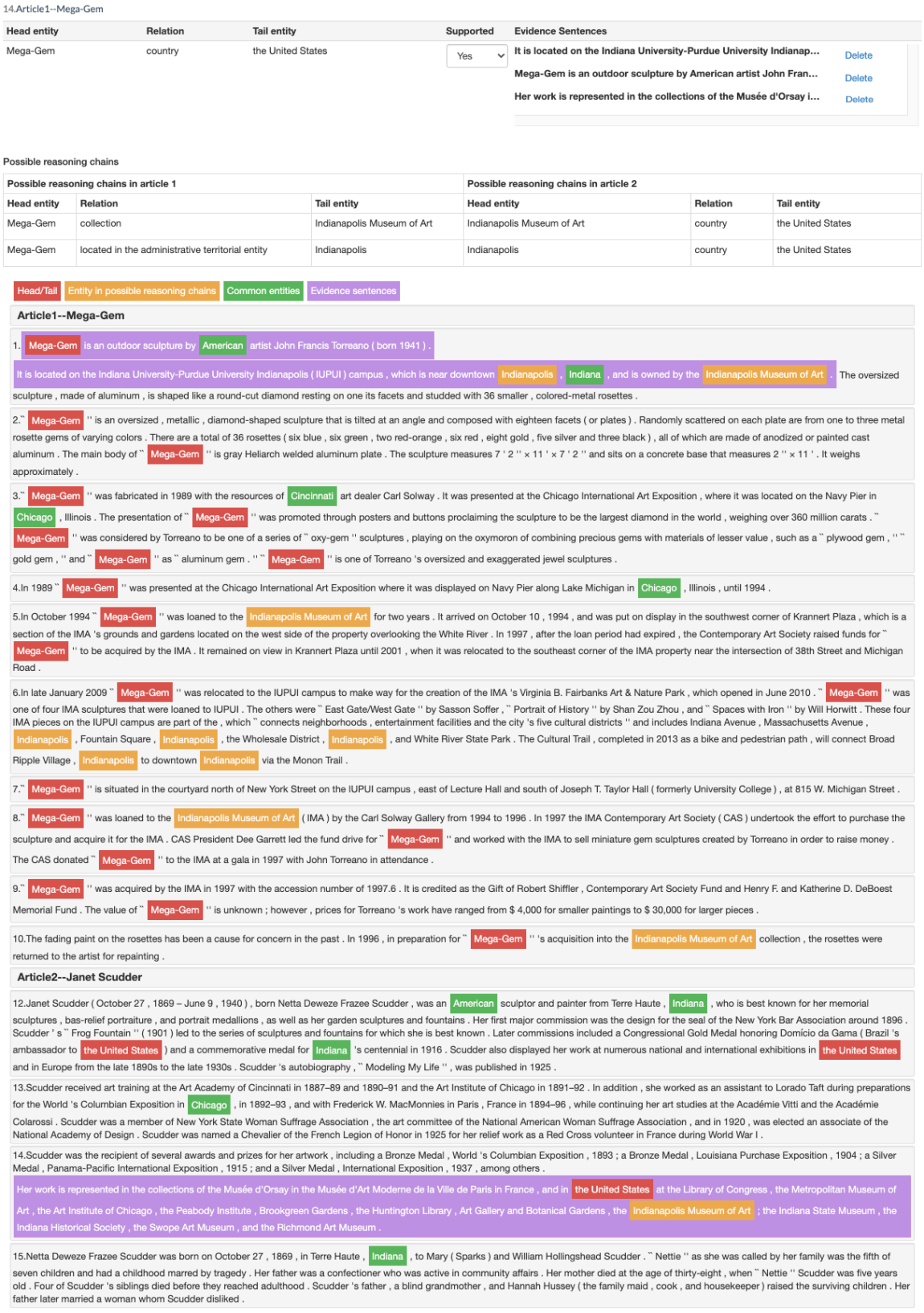

Figure 6: The annotation platform. Annotators are provided with relational fact and text paths. We also highlight the mentions of target entities and bridging entities, and provide possible reasoning chains to assist annotation.