DocRED: A Large-Scale Document-Level Relation Extraction Dataset

Yuan Yao†, Deming Ye†, Peng Li‡, Xu Han†, Yankai Lin†, Zhenghao Liu†, Zhiyuan Liu†, Lixin Huang†, Jie Zhou‡, Maosong Sun†

†Department of Computer Science and Technology, Tsinghua University, Beijing, China
‡Institute for Artificial Intelligence, Tsinghua University, Beijing, China
State Key Lab on Intelligent Technology and Systems, Tsinghua University, Beijing, China

Abstract

Multiple entities in a document generally exhibit complex inter-sentence relations, and cannot be well handled by existing relation extraction (RE) methods that typically focus on extracting intra-sentence relations for single entity pairs. In order to accelerate the research on document-level RE, we introduce DocRED, a new dataset constructed from Wikipedia and Wikidata with three features: (1) DocRED annotates both named entities and relations, and is the largest human-annotated dataset for document-level RE from plain text; (2) DocRED requires reading multiple sentences in a document to extract entities and infer their relations by synthesizing all information of the document; (3) along with the human-annotated data, we also offer large-scale distantly supervised data, which enables DocRED to be adopted for both supervised and weakly supervised scenarios. In order to verify the challenges of document-level RE, we implement recent state-of-the-art methods for RE and conduct a thorough evaluation of these methods on DocRED. Empirical results show that DocRED is challenging for existing RE methods, which indicates that document-level RE remains an open problem and requires further efforts. Based on the detailed analysis on the experiments, we discuss multiple promising directions for future research. We make DocRED and the code for our baselines publicly available at https://github.com/thunlp/DocRED.

1 Introduction

The task of relation extraction (RE) is to identify relational facts between entities from plain text, which plays an important role in large-scale knowledge graph construction. Most existing RE work focuses on sentence-level RE, i.e., extracting relational facts from a single sentence. In recent years, various neural models have been explored to encode relational patterns of entities for sentence-level RE, and achieve state-of-the-art performance (Socher et al., 2012; Zeng et al., 2014, 2015; dos Santos et al., 2015; Xiao and Liu, 2016; Cai et al., 2016; Lin et al., 2016; Wu et al., 2017; Qin et al., 2018; Han et al., 2018a).

Despite these successful efforts, sentence-level RE suffers from an inevitable restriction in practice: a large number of relational facts are expressed in multiple sentences. Taking Figure 1 as an example, multiple entities are mentioned in the document and exhibit complex interactions. In
order to identify the relational fact (*Riddarhuset, country, Sweden*), one has to first identify the fact that *Riddarhuset* is located in *Stockholm* from Sentence 4, then identify the facts *Stockholm is the capital of Sweden* and *Sweden is a country* from Sentence 1, and finally infer from these facts that the sovereign state of *Riddarhuset* is *Sweden*. The process requires reading and reasoning over multiple sentences in a document, which is intuitively beyond the reach of sentence-level RE methods. According to the statistics on our human-annotated corpus sampled from Wikipedia documents, at least 40.7% relational facts can only be extracted from multiple sentences, which is not negligible. Swampillai and Stevenson (2010) and Verga et al. (2018) have also reported similar observations. Therefore, it is necessary to move RE forward from sentence level to document level.

The research on document-level RE requires a large-scale annotated dataset for both training and evaluation. Currently, there are only a few datasets for document-level RE. Quirk and Poon (2017) and Peng et al. (2017) build two distantly supervised datasets without human annotation, which may make the evaluation less reliable. BC5CDR (Li et al., 2016) is a human-annotated document-level RE dataset consisting of 1,500 PubMed documents, which is in the specific domain of biomedicine considering only the “chemical-induced disease” relation, making it unsuitable for developing general-purpose methods for document-level RE. Levy et al. (2017) extract relational facts from documents by answering questions using reading comprehension methods, where the questions are converted from entity-relation pairs. As the dataset proposed in this work is tailored to the specific approach, it is also unsuitable for other potential approaches for document-level RE. In summary, existing datasets for document-level RE either only have a small number of manually-annotated relations and entities, or exhibit noisy annotations from distant supervision, or serve specific domains or approaches. In order to accelerate the research on document-level RE, we urgently need a large-scale, manually-annotated, and general-purpose document-level RE dataset.

In this paper, we present DocRED, a large-scale human-annotated document-level RE dataset constructed from Wikipedia and Wikidata (Erxleben et al., 2014; Vrandečić and Krötzsch, 2014). DocRED is constructed with the following three features: (1) DocRED contains 132,375 entities and 56,354 relational facts annotated on 5,053 Wikipedia documents, making it the largest human-annotated document-level RE dataset. (2) As at least 40.7% of the relational facts in DocRED can only be extracted from multiple sentences, DocRED requires reading multiple sentences in a document to recognize entities and inferring their relations by synthesizing all information of the document. This distinguishes DocRED from those sentence-level RE datasets. (3) We also provide large-scale distantly supervised data to support weakly supervised RE research.

To assess the challenges of DocRED, we implement recent state-of-the-art RE methods and conduct thorough experiments on DocRED under various settings. Experimental results show that the performance of existing methods declines significantly on DocRED, indicating the task document-level RE is more challenging than sentence-level RE and remains an open problem. Furthermore, detailed analysis on the results also reveals multiple promising directions worth pursuing.

## 2 Data Collection

Our ultimate goal is to construct a dataset for document-level RE from plain text, which requires necessary information including named entity mentions, entity coreferences, and relations of all entity pairs in the document. To facilitate more RE settings, we also provide supporting evidence information for relation instances. In the following sections, we first introduce the collection process of the human-annotated data, and then describe the process of creating the large-scale distantly supervised data.

### 2.1 Human-Annotated Data Collection

Our human-annotated data is collected in four stages: (1) Generating distantly supervised annotation for Wikipedia documents. (2) Annotating all named entity mentions in the documents and coreference information. (3) Linking named entity mentions to Wikidata items. (4) Labeling relations and corresponding supporting evidence.

Following ACE annotation process (Doddington et al., 2004), both Stage 2 and 4 require three iterative passes over the data: (1) Generating named entity using named entity recognition (NER) models, or relation recommendations us-
ing distant supervision and RE models. (2) Man-
ually correcting and supplementing recommenda-
tions. (3) Reviewing and further modifying the an-
notation results from the second pass for better ac-
curacy and consistency. To ensure the annotators
are well trained, a principled training procedure is
adopted and the annotators are required to pass test
tasks before annotating the dataset. And only care-
fully selected experienced annotators are qualified
for the third pass annotation.

To provide a strong alignment between text and
KBs, our dataset is constructed from the complete
English Wikipedia document collection and Wiki-
data\(^1\), which is a large-scale KB tightly integrated
with Wikipedia. We use the introductory sections
from Wikipedia documents as the corpus, as they
are usually high-quality and contain most of the
key information.

Stage 1: Distantly Supervised Annotation Gen-
eration. To select documents for human annota-
tion, we align Wikipedia documents with Wikidata
under the distant supervision assumption (Mintz
et al., 2009). Specifically, we first perform named
entity recognition using spaCy\(^2\). Then these
named entity mentions are linked to Wikidata
items, where named entity mentions with identi-
cal KB IDs are merged. Finally, relations between
each merged named entity pair in the document are
labeled by querying Wikidata. Documents con-
taining fewer than 128 words are discarded. To en-
courage reasoning, we further discard documents
containing fewer than 4 entities or fewer than 4 re-
lation instances, resulting in 107,050 documents
with distantly supervised labels, where we ran-
domly select 5,053 documents and the most fre-
quent 96 relations for human annotation.

Stage 2: Named Entity and Coreference An-
notation. Extracting relations from document re-
quires first recognizing named entity mentions and
identifying mentions referring to the same enti-
ties within the document. To provide high-quality
named entity mentions and coreference informa-
tion, we ask human annotators first to review, cor-
correct and supplement the named entity mention
recommendations generated in Stage 1, and then
merge those different mentions referring to the
same entities, which provides extra coreference in-
formation. The resulting intermediate corpus con-
tains a variety of named entity types including
person, location, organization, time, number and
names of miscellaneous entities that do not belong
to the aforementioned types.

Stage 3: Entity Linking. In this stage, we link
each named entity mention to multiple Wikidata
items to provide relation recommendations from
distant supervision for the next stage. To be spe-
cific, each named entity mention is associated with
a Wikidata item candidate set \(^3\) consisting of all
Wikidata items whose names or aliases literally
match it. We further extend the candidate set using
Wikidata items hyperlinked to the named entity
mention by the document authors, and recommen-
dations from an entity linking toolkit TagMe (Fer-
ragina and Scaiella, 2010). Specially, numbers and
time are semantically matched.

Stage 4: Relation and Supporting Evidence
Collection. The annotation of relation and sup-
porting evidence is based on the named entity
mentions and coreference information in Stage 2,
and faces two main challenges. The first challenge
comes from the large number of potential entity
pairs in the document. On the one hand, given
the quadratic number of potential entity pairs with
regard to entity number (19.5 entities on aver-
age) in a document, exhaustively labeling relations
between each entity pair would lead to intensive
workload. On the other hand, most entity pairs in
a document do not contain relations. The second
challenge lies in the large number of fine-grained
relation types in our dataset. Thus it is not feasible
for annotators to label relations from scratch.

We address the problem by providing human
annotators with recommendations from RE mod-
els, and distant supervision based on entity link-
ing (Stage 3). On average, we recommend 19.9
relation instances per document from entity link-
ing, and 7.8 from RE models for supplement. We
ask the annotators to review the recommendations,
remove the incorrect relation instances and sup-
plement omitted ones. We also ask the annotators
to further select all sentences that support the re-
served relation instances as supporting evidence.
Relations reserved must be reflected in the doc-
ument, without relying on external world knowl-
edge. Finally 57.2\% relation instances from entity
linking and 48.2\% from RE models are reserved.

---

\(^1\)We use the 2018-5-24 dump of English Wikipedia and
2018-3-20 dump of Wikidata.

\(^2\)https://spacy.io

\(^3\)To avoid losing relation recommendations due to predic-
tion errors in entity linking, we include multiple linking re-
results from different approaches in the candidate set.
In addition to the human-annotated data, we also collect large-scale distantly supervised data to promote weakly supervised RE scenarios. We remove the 5,053 human-annotated documents from the 106,926 documents, and use the rest 101,873 documents as the corpus of distantly supervised data. To ensure that the distantly supervised data and human-annotated data share the same entity distribution, named entity mentions are re-identified using Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019) that is fine-tuned on the human-annotated data. Then we merge the named entity mentions with identical KB IDs. Finally, relations between each merged entity pair are labeled via distant supervision.

3 Data Analysis

In this section, we analyze various aspects of DocRED to provide a deeper understanding of the dataset and the task of document-level RE.

Data Size. Table 1 shows statistics of DocRED and some representative RE datasets, including sentence-level RE datasets SemEval-2010 Task 8 (Hendrickx et al., 2010), ACE 2003-2004 (Doddington et al., 2004), TACRED (Zhang et al., 2017), FewRel (Han et al., 2018b) and document-level RE dataset BC5CDR (Li et al., 2016). We find that DocRED is larger than existing datasets in many aspects, including the number of documents, words, sentences, entities, especially in aspects of relation types, relation instances and relational facts. We hope the large-scale DocRED dataset could drive relation extraction from sentence level forward to document level.

Named Entity Types. DocRED covers a variety of entity types, including person (18.5%), location (30.9%), organization (14.4%), time (15.8%) and number (5.1%). It also covers a diverse set of miscellaneous entity names (15.2%) not belonging to the aforementioned types, such as events, artistic works and laws. Each entity is annotated with 1.34 mentions on average.

Relation Types. Our dataset includes 96 frequent relation types from Wikidata. A notable property of our dataset is that the relation types cover a broad range of categories, including relations relevant to science (33.3%), art (11.5%), time (8.3%), personal life (4.2%), etc., which means the relational facts are not constrained in any specific domain. In addition, the relation types are organized in a well-defined hierarchy and taxonomy, which could provide rich information for document-level RE systems.

Reasoning Types. We randomly sampled 300 documents from dev and test set, which contain 3,820 relation instances, and manually analyze the reasoning types required to extract these relations. Table 2 shows statistics of major reasoning types in our dataset. From the statistics on reasoning types, we have the following observations: (1) Most of the relation instances (61.1%) require reasoning to be identified, and only 38.9% relation instances can be extracted via simple pattern recognition, which indicates that reasoning is essential for document-level RE. (2) In relation instances with reasoning, a majority (26.6%) require logical reasoning, where the relations between two entities in question are indirectly established by a bridge entity. Logical reasoning requires RE systems to be capable of modeling interactions between multiple entities. (3) A notable number of relation instances (17.6%) need coreference reasoning, where coreference resolution must be performed first to identify target entities in a rich con-
Reasoning Types | % | Examples
--- | --- | ---
Pattern recognition | 38.9 | [1] *Me Musical Nephews* is a 1942 one-reel animated cartoon directed by Seymour Kneitel and animated by Tom Johnson and George Germanetti. [2] Jack Mercer and Jack Ward wrote the script. ...
Relation: publication_date | Supporting Evidence: 1
Logical reasoning | 26.6 | [1] “Nisei” is the ninth episode of the third season of the American science fiction television series *The X-Files*. ... [3] It was directed by David Nutter, and written by Chris Carter, Frank Spotnitz and Howard Gordon. ... [8] The show centers on FBI special agents *Fox Mulder* (David Duchovny) and Dana Scully (Gillian Anderson) who work on cases linked to the paranormal, called *X-Files*. ...
Relation: creator | Supporting Evidence: 1, 3, 8
Coreference reasoning | 17.6 | [1] *Dwight Tillery* is an American politician of the Democratic Party who is active in local politics of Cincinnati, Ohio. ... [3] He also holds a law degree from the University of Michigan Law School. [4] *Tillery* served as mayor of Cincinnati from 1991 to 1993.
Relation: educated_at | Supporting Evidence: 1, 3
Common-sense reasoning | 16.6 | [1] *William Busac* (1020-1076), son of William I, Count of Eu, and his wife Lesceline. ... [4] *William* appealed to King Henry I of France, who gave him in marriage *Adelaide*, the heiress of the county of Soissons. [5] *Adelaide* was daughter of Renaud I, Count of Soissons, and Grand Master of the Hotel de France. ... [7] *William* and *Adelaide* had four children: ...
Relation: spouse | Supporting Evidence: 4, 7

Table 2: Types of reasoning required for document-level RE on DocRED. The rest 0.3% requires other types of reasoning, such as temporal reasoning. The head, tail and relation are colored accordingly.

| Setting | # Doc. | # Rel. | # Inst. | # Fact |
| --- | --- | --- | --- | --- |
| Train W | 101,873 | 96 | 1,508,320 | 881,298 |
| S | 3,053 | 96 | 38,269 | 34,715 |
| Dev S,W | 1,000 | 96 | 12,332 | 11,790 |
| Test S,W | 1,000 | 96 | 12,842 | 12,101 |

Table 3: Statistics of data used for the two benchmark settings (Sec. 4): supervised setting (S) and weakly supervised setting (W).
and Gurevych, 2017; Christopoulou et al., 2018) are not efficient enough for document-level RE. Thus the efficiency of context-aware RE systems needs to be further improved to be applicable in document-level RE.

**Weakly Supervised Setting.** This setting is identical to the supervised setting except that the training set is replaced with the distantly supervised data (Sec. 2.2). In addition to the aforementioned two challenges, the inevitable wrong labeling problem accompanied with distantly supervised data is a major challenge for RE models under weakly supervised setting. Many efforts have been devoted to alleviating the wrong labeling problem in sentence-level RE (Riedel et al., 2010; Hoffmann et al., 2011; Surdeanu et al., 2012; Lin et al., 2016). However, noise in document-level distantly supervised data is significantly more than its counterpart in sentence-level. For example, for the recommended relation instances whose head and tail entities co-occur in the same sentence (i.e. intra-sentence relation instance) in Stage 4 of human-annotated data collection (Sec. 2.1), 41.4% are labeled as incorrect, while 61.8% inter-sentence relation instances are labeled as incorrect, indicating the wrong labeling problem is more challenging for weakly supervised document-level RE. Therefore, we believe offering distantly supervised data in DocRED will accelerate the development of distantly supervised methods for document-level RE. Moreover, it is also possible to jointly leverage distantly supervised data and human-annotated data to further improve the performance of RE systems.

**5 Experiments**

To assess the challenges of DocRED, we conduct comprehensive experiments to evaluate state-of-the-art RE systems on the dataset. Specifically, we conduct experiments under both supervised and weakly supervised benchmark settings. We also assess human performance and analyze the performance for different supporting evidence types. In addition, we conduct ablation study to investigate the contribution of different features. Through detailed analysis, we discuss several future directions for document-level RE.

**Models.** We adapt four state-of-the-art RE models to document-level RE scenario, including a CNN (Zeng et al., 2014) based model, an LSTM (Hochreiter and Schmidhuber, 1997) based model, a bidirectional LSTM (BiLSTM) (Cai et al., 2016) based model and the Context-Aware model (Sorokin and Gurevych, 2017) originally designed for leveraging contextual relations to improve intra-sentence RE. The first three models differ only at the encoder used for encoding the document and will be explained in detail in the rest of this section. We refer the readers to the original paper for the details of the Context-Aware model for space limitation.

The CNN/LSTM/BiLSTM based models first encode a document \( D = \{ w_i \}_{i=1}^n \) consisting of \( n \) words into a hidden state vector sequence \( \{ h_i \}_{i=1}^n \) with CNN/LSTM/BiLSTM as encoder, then compute the representations for entities, and finally predict relations for each entity pair.

For each word, the features fed to the encoder is the concatenation of its GloVe word embedding (Pennington et al., 2014), entity type embedding and coreference embedding. The entity type embedding is obtained by mapping the entity type (e.g., PER, LOC, ORG) assigned to the word into a vector using an embedding matrix. The entity type is assigned by human for the human-annotated data, and by a fine-tuned BERT model for the distantly supervised data. Named entity mentions corresponding to the same entity are assigned with the same entity id, which is determined by the order of its first appearance in the document. And the entity ids are mapped into vectors as the coreference embeddings.

For each named entity mention \( m_k \) ranging from the \( s \)-th word to the \( t \)-th word, we define its representation as \( m_k = \frac{1}{t-s+1} \sum_{i=s}^{t} h_i \). And the representation of an entity \( e_i \) with \( K \) mentions is computed as the average of the representations of these mentions: \( e_i = \frac{1}{K} \sum_{k} m_k \).

We treat relation prediction as a multi-label classification problem. Specially, for each entity pair \( (e_i, e_j) \), we first concatenate the entity representations with relative distance embeddings, and then use a bilinear function to compute the probability for each relation type:

\[
\hat{e}_i = [e_i; E(d_{ij})], \quad \hat{e}_j = [e_j; E(d_{ji})]
\]

\[
P(r|e_i, e_j) = \text{sigmoid}(\hat{e}_i^T W_r \hat{e}_j + b_r)
\]

where \([\cdot ; \cdot]\) denotes concatenation, \( d_{ij} \) and \( d_{ji} \) are the relative distances of the first mentions of the two entities in the document, \( E \) is an embedding matrix, \( r \) is a relation type, and \( W_r, b_r \) are relation type dependent trainable parameters.
| Model          | Dev Ign F1 | Dev Ign AUC | Dev F1 | Dev AUC | Test Ign F1 | Test Ign AUC | Test F1 | Test AUC |
|---------------|------------|-------------|--------|--------|------------|-------------|--------|--------|
| Supervised Setting | 37.99      | 31.47       | 43.45  | 39.41  | 36.44      | 30.44       | 42.33  | 38.98  |
| CNN           | 44.41      | 39.78       | 50.66  | 49.48  | 43.60      | 39.02       | 50.12  | 49.31  |
| LSTM          | 45.12      | 40.93       | 50.95  | 50.27  | 44.73      | 40.40       | 51.06  | 50.43  |
| BiLSTM        | 44.84      | 40.42       | 51.10  | 50.20  | 43.93      | 39.30       | 50.64  | 49.70  |
| Context-Aware |            |             |        |        |            |             |        |        |
| Weakly Supervised Setting | 26.35      | 14.18       | 42.75  | 38.01  | 25.40      | 13.46       | 42.02  | 36.86  |
| CNN           | 30.86      | 15.62       | 49.91  | 42.78  | 29.75      | 14.97       | 49.91  | 42.78  |
| LSTM          | 32.05      | 16.50       | 51.72  | 44.42  | 29.96      | 15.50       | 49.82  | 42.90  |
| BiLSTM        | 32.43      | 15.86       | 51.39  | 43.02  | 30.27      | 15.11       | 50.14  | 41.52  |
| Context-Aware |            |             |        |        |            |             |        |        |

Table 4: Performance of different RE models on DocRED (%).

**Evaluation Metrics.** Two widely used metrics F1 and AUC are used in our experiments. However, some relational facts present in both the training and dev/test sets, thus a model may memorize their relations during training and achieve a better performance on the dev/test set in an undesirable way, introducing evaluation bias. However, the overlap in relational facts between the training and dev/test sets is inevitable, since many common relational facts are likely to be shared in different documents. Therefore, we also report the F1 and AUC scores excluding those relational facts shared by the training and dev/test sets, denoted as Ign F1 and Ign AUC, respectively.

**Model Performance.** Table 4 shows the experimental results under the supervised and weakly supervised settings, from which we have the following observations: (1) Models trained with human-annotated data generally outperform their counterparts trained on distantly supervised data. This is because although large-scale distantly supervised data can be easily obtained via distant supervision, the wrong-labeling problem may harm the performance of RE systems, which makes weakly supervised setting a more difficult scenario. (2) An interesting exception is that LSTM, BiLSTM and Context-Aware trained on distantly supervised data achieve comparable F1 scores as those trained on human-annotated data but significantly lower scores on the other metrics, indicating that the overlap entity pairs between training and dev/test sets indeed cause evaluation biases. Therefore, reporting Ign F1 and Ign AUC is necessary. (3) Models leveraging rich contextual information generally achieve better performances. LSTM and BiLSTM outperform CNN, indicating the effectiveness of modeling long-dependency semantics in document-level RE. Context-Aware achieves competitive performance, however, it cannot significantly outperform other neural models. It indicates that it is beneficial to consider the association of multiple relations in document-level RE, whereas the current models are not capable of utilizing inter-relation information well.

**Human Performance.** To assess human performance on the task of document-level RE on DocRED, we randomly sample 100 documents from the test set and ask additional crowd-workers to identify relation instances and supporting evidence. Relation instances identified in the same way as Sec. 2.1 are recommended to the crowd-workers to assist them. The original annotation results collected in Sec. 2.1 are used as ground truth. We also propose another subtask of jointly identifying relation instances and supporting evidence, and also design a pipeline model. Table 5 shows the performance of RE model and human. Humans achieve competitive results on both the document-level RE task (RE) and the jointly identifying relation and supporting evidence task (RE+Sup), indicating both the ceiling performance on DocRED and the inter-annotator agreement are relatively high. In addition, the overall performance of RE models is significantly lower than human performance, which indicates document-level RE is a challenging task, and suggests ample opportunity for improvement.

| Method | RE P | RE R | RE F1 | RE+Sup P | RE+Sup R | RE+Sup F1 |
|--------|------|------|-------|----------|----------|-----------|
| Model  | 55.6 | 52.6 | 54.1  | 46.4     | 43.1     | 44.7      |
| Human  | 89.7 | 86.3 | 88.0  | 71.2     | 75.8     | 73.4      |

Table 5: Human performance (%).
Performance vs. Supporting Evidence Types.

Document-level RE requires synthesizing information from multiple supporting sentences. To investigate the difficulty of synthesizing information from different types of supporting evidence, we divide the 12,332 relation instances in development set into three disjoint subsets: (1) 6,115 relation instances with only one supporting sentence (denoted as single); (2) 1,062 relation instances with multiple supporting sentences and the entity pair co-occur in at least one supporting sentence (denoted as mix); (3) 4,668 relation instances with multiple supporting sentences and the entity pair do not co-occur in any supporting sentence, which means they can only be extracted from multiple supporting sentences (denoted as multiple). It should be noted that when a model predicts a wrong relation, we do not know which sentences have been used as supporting evidence, thus the predicted relation instance cannot be classified into the aforementioned subsets and computing precision is infeasible. Therefore, we only report recall of the RE model for each subset, which is 51.1% for single, 49.4% for mix, and 46.6% for multiple. This indicates that while multiple supporting sentences in mix may provide complementary information, it is challenging to effectively synthesize the rich global information. Moreover, the poor performance on multiple suggests that RE models still struggle in extracting inter-sentence relations.

Feature Ablations. We conduct feature ablation studies on the BiLSTM model to investigate the contribution of different features in document-level RE, including entity types, coreference information, and the relative distance between entities (Eq. 1). Table 6 shows that the aforementioned features all have a contribution to the performance. Specifically, entity types contribute most due to their constraint on viable relation types. Coreference information and the relative distance between entities are also important for synthesizing information from multiple named entity mentions. This indicates that it is important for RE systems to leverage rich information at document level.

Supporting Evidence Prediction. We propose a new task to predict the supporting evidence for relation instances. On the one hand, jointly predicting the evidence provides better explainability. On the other hand, identifying supporting evidence and reasoning relational facts from text are naturally dual tasks with potential mutual enhancement. We design two supporting evidence prediction methods: (1) Heuristic predictor. We implement a simple heuristic-based model that considers all sentences containing the head or tail entity as supporting evidence. (2) Neural predictor. We also design a neural supporting evidence predictor. Given an entity pair and a predicted relation, sentences are first transformed into input representations by the concatenation of word embeddings and position embeddings, and then fed into a BiLSTM encoder for contextual representations. Inspired by Yang et al. (2018), we concatenate the output of the BiLSTM at the first and last positions with a trainable relation embedding to obtain a sentence’s representation, which is used to predict whether the sentence is adopted as supporting evidence for the given relation instance. As Table 7 shows, the neural predictor significantly outperforms heuristic-based baseline in predicting supporting evidence, which indicates the potential of RE models in joint relation and supporting evidence prediction.

| Method         | Dev   | Test  |
|----------------|-------|-------|
| Heuristic predictor | 36.21 | 36.76 |
| Neural predictor            | **44.07** | **43.85** |

Table 7: Performance of joint relation and supporting evidence prediction in F1 measurement (%).

Discussion. We can conclude from the above experimental results and analysis that document-level RE is more challenging than sentence-level RE and intensive efforts are needed to close the gap between the performance of RE models and human. We believe the following research directions are worth following: (1) Exploring models explicitly considering reasoning; (2) Designing more expressive model architectures for collecting and synthesizing inter-sentence information; (3) Leveraging distantly supervised data to improve the performance of document-level RE.
6 Related Work

A variety of datasets have been constructed for RE in recent years, which have greatly promoted the development of RE systems. Hendrickx et al. (2010), Doddington et al. (2004) and Walker et al. (2006) build human-annotated RE datasets with relatively limited relation types and instances. Riedel et al. (2010) automatically construct RE dataset by aligning plain text to KB via distant supervision, which suffers from wrong labeling problem. Zhang et al. (2017) and Han et al. (2018b) further combine external recommendations with human annotation to build large-scale high-quality datasets. However, these RE datasets limit relations to single sentences.

As documents provide richer information than sentences, moving research from sentence level to document level is a popular trend for many areas, including document-level event extraction (Walker et al., 2006; Mitamura et al., 2015, 2017), fact extraction and verification (Thorne et al., 2018), reading comprehension (Nguyen et al., 2016; Joshi et al., 2017; Lai et al., 2017), sentiment classification (Pang and Lee, 2004; Prettenhofer and Stein, 2010), summarization (Nallapati et al., 2016) and machine translation (Zhang et al., 2018). Recently, some document-level RE datasets have also been constructed. However, these datasets are either constructed via distant supervision (Quirk and Poon, 2017; Peng et al., 2017) with inevitable wrong labeling problem, or limited in specific domain (Li et al., 2016; Peng et al., 2017). In contrast, DocRED is constructed by crowd-workers with rich information, and is not limited in any specific domain, which makes it suitable to train and evaluate general-purpose document-level RE systems.

7 Conclusion

To promote RE systems from sentence level to document level, we present DocRED, a large-scale document-level RE dataset that features the data size, the requirement for reading and reasoning over multiple sentences, and the distantly supervised data offered for facilitating the development of weakly supervised document-level RE. Experiments show that human performance is significantly higher than RE baseline models, which suggests ample opportunity for future improvement.

8 Acknowledgement

This work is supported by the National Key Research and Development Program of China (No. 2018YFB1004503), the National Natural Science Foundation of China (NSFC No. 61572273) and China Association for Science and Technology (2016QNRC001). This work is also supported by the Pattern Recognition Center, WeChat AI, Tencent Inc. Han is also supported by 2018 Tencent Rhino-Bird Elite Training Program.

References

Rui Cai, Xiaodong Zhang, and Houfeng Wang. 2016. Bidirectional recurrent convolutional neural network for relation classification. In Proceedings of ACL, pages 756–765.

Fenia Christopoulou, Makoto Miwa, and Sophia Ananiadou. 2018. A walk-based model on entity graphs for relation extraction. In Proceedings of ACL, pages 81–88.

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.

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.

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.

Paolo Ferragina and Ugo Scaiella. 2010. TAGME: on-the-fly annotation of short text fragments (by wikipedia entities). In Proceedings of CIKM, pages 1625–1628.

Xu Han, Pengfei Yu, Zhiyuan Liu, Maosong Sun, and Peng Li. 2018a. Hierarchical relation extraction with coarse-to-fine grained attention. In Proceedings of EMNLP, pages 2236–2245.

Xu Han, Hao Zhu, Pengfei Yu, Ziyun Wang, Yuan Yao, Zhiyuan Liu, and Maosong Sun. 2018b. 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, Zornitza Kozareva, Preslav Nakov, Dairmuid 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, the Workshop on Semantic Evaluation, pages 33–38.
Erik F Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: language-independent named entity recognition. In Proceedings of NAACL-HLT.

Patrick Verga, Emma Strubell, and Andrew McCallum. 2018. Simultaneously self-attending to all mentions for full-abstract biological relation extraction. In Proceedings of NAACL-HLT, pages 872–884.

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.

Yi Wu, David Bamman, and Stuart Russell. 2017. Adversarial training for relation extraction. In Proceedings of EMNLP, pages 1778–1783.

Minguang Xiao and Cong Liu. 2016. Semantic relation classification via hierarchical recurrent neural network with attention. In Proceedings of COLING 2016, pages 1254–1263.

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 Proceedings of EMNLP, pages 2369–2380.

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.

Jiacheng Zhang, Huanbo Luan, Maosong Sun, Feifei Zhai, Jingfang Xu, Min Zhang, and Yang Liu. 2018. Improving the transformer translation model with document-level context. In Proceedings of EMNLP, pages 533–542.

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 Appendices

A.1 Experimental Details

In this section, we provide more details of our experiments. To fairly compare the results of different models, we optimized all baselines using Adam, with learning rate of 0.001, $\beta_1 = 0.9$, $\beta_1 = 0.999$. The other experimental hyper-parameters used in our experiments are shown in Table 8. Additionally, due to the document-level distance between entities, distances are first divided into several bins $\{1, 2, \ldots, 2^k\}$, where each bin is associated with a trainable distance embedding.

| Hyper-parameter                        | Value |
|----------------------------------------|-------|
| Batch size                             | 40    |
| CNN hidden size                        | 200   |
| CNN window size                        | 3     |
| CNN dropout rate                       | 0.5   |
| LSTM hidden size                       | 128   |
| LSTM dropout rate                      | 0.2   |
| Word embedding dimension               | 100   |
| Entity type embedding dimension        | 20    |
| Coreference embedding dimension        | 20    |
| Distance embedding dimension           | 20    |

Table 8: Hyper-parameter settings.

A.2 Types of Named Entities

In this paper, we adapt the existing types of named entities used in Tjong Kim Sang and De Meulder (2003) to better serve DocRED. These types include “Person (PER)”, “Organization (ORG)”, “Location (LOC)”, “Time (TIME)”, “Number (NUM)”, and “other types (MISC)”. The types of named entities in DocRED and their covered contents are shown in Table 9.

| Types | Content                                                                 |
|-------|-------------------------------------------------------------------------|
| PER   | People, including fictional                                            |
| ORG   | Companies, universities, institutions, political or religious groups, etc.|
| LOC   | Geographically defined locations, including mountains, waters, etc.     |
|       | Politically defined locations, including countries, cities, states, streets, etc.|
| TIME  | Absolute or relative dates or periods                                   |
| NUM   | Percents, money, quantities                                             |
| MISC  | Products, including vehicles, weapons, etc.                             |
|       | Events, including elections, battles, sporting events, etc.             |
|       | Laws, cases, languages, etc                                              |

Table 9: Types of named entities in DocRED.

A.3 List of Relations

We provide the list of relations in DocRED, including Wikidata IDs, relation names and descriptions from Wikidata in Table 10 and 11.
| Wikidata ID | Name                                      | Description                                                                                     |
|------------|-------------------------------------------|-------------------------------------------------------------------------------------------------|
| P6         | head of government                        | head of the executive power of this town, city, municipality, state, country, or other governmental body |
| 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 |
| P27        | country of citizenship                    | the object is a country that recognizes the subject as its citizen                               |
| P30        | continent                                 | continent of which the subject is a part                                                        |
| P31        | instance of                               | that class of which this subject is a particular example and member. (Subject typically an individual member with Proper Name label.) Different from P279 (subclass of) |
| P35        | head of state                             | official with the highest formal authority in a country/state                                    |
| P36        | capital                                   | primary city of a country, state or other type of administrative territorial entity               |
| P37        | official language                         | language designated as official by this 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)                                   |
| P54        | member of sports team                     | sports teams or clubs that the subject currently represents or formerly represented              |
| P57        | director                                  | director(s) of this motion picture, TV-series, stageplay, video game or similar                  |
| P58        | screenwriter                              | author(s) of the screenplay or script for this work                                             |
| P69        | educated at                               | educational institution attended by the subject                                                 |
| P86        | composer                                  | person(s) who wrote the music; also use P676 for lyrician                                        |
| 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                                   |
| P118       | league                                    | league in which team or player plays or has played in                                            |
| P123       | publisher                                 | organization or person responsible for publishing books, periodicals, games or software         |
| P127       | owned by                                  | owner of the subject                                                                           |
| 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 |
| P136       | genre                                     | genre of a 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 |
| P140       | religion                                  | religion of a person, organization or religious building, or associated with this subject       |
| P150       | contains administrative territorial entity | list of direct subdivisions of an administrative territorial entity                                |
| P155       | follows                                   | immediately prior item in some series of which the subject is part. Use P1365 (replaces) if the preceding item was replaced, e.g. political offices, states and there is no identity between precedent and following geographic unit |
| P156       | followed by                               | immediately following item in some series of which the subject is part. Use P1366 (replaced by) if the item is replaced, e.g. political offices, states |
| P159       | headquarters location                     | specific location where an organization’s headquarters is or has been situated                   |
| 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                                  | producer(s) of this film or music work (film: not executive producers, associate producers, etc.) [use P72 to refer to the production company] |
| P166       | award received                            | award or recognition received by a person, organisation or creative work                         |
| P170       | creator                                   | maker of a creative work or other object (where no more specific property exists)                 |
| P171       | parent taxon                              | closest parent taxon of the taxon in question                                                   |
| P172       | ethnic group                              | subject’s ethnicity (consensus is that a VERY high standard of proof is needed for this field to be used. In general this means 1) the subject claims it him/herself, or 2) it is widely agreed on by scholars, or 3) is fictional and portrayed as such). |
| P175       | performer                                 | performer involved in the performance or the recording of a work                                |
| P176       | manufacturer                              | manufacturer or producer of this product                                                        |
| P178       | developer                                 | organisation or person that developed this item                                                  |
| P179       | series                                    | subject is part of a series, whose sum constitutes the object                                   |
| P190       | sister city                               | twin towns, sister cities, twinned municipalities and other localities that have a partnership or cooperative agreement, either legally or informally acknowledged by their governments |
| P194       | legislative body                          | legislative body governing this entity; political institution with elected representatives, such as a parliament/legislature or council |
| P205       | basin country                             | country that have drainage to/from or border the body of water                                  |

Table 10: Relation list (I), including Wikidata IDs, names and descriptions of relations in DocRED.
| Wikidata ID | Name | Description |
|------------|------|-------------|
| P206       | located in or next to body of water | branch to which this military unit, award, office, or person belongs, e.g. Royal Navy |
| P241       | military branch record label | brand and trademark associated with the marketing of subject music recordings and music videos |
| P264       | production company location | company that produced this film, audio or performing arts work |
| P276       | location subclass of | 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       | subsidiary 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) |
| P355       | part of | subsidiary of a company or organization, opposite of parent company (P749) |
| P361       | original language of work platform mouth of the watercourse | object of which the subject is a part. Inverse property of "has part" (P361). See also "has parts of the class" (P2670). |
| P364       | date of birth date of death | language in which a film or a performance work was originally created. Deprecated for written works; use P407 ("language of work or name") instead. |
| P400       | publication date publication date | platform for which a work has been developed or released / specific platform version of a software developed |
| P403       | creation date | the body of water to which the watercourse drains |
| P449       | original network member of | network(s) the radio or television show was originally aired on, including |
| P463       | chairperson country of origin residence date of birth date of death | 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). |
| P488       | country of origin | presiding member of an organization, group or body |
| P495       | has part date on which the subject was born | country of origin of the creative work or subject item |
| P527       | has part date on which the subject was born | part of this subject. Inverse property of "part of" (P361). |
| P551       | date on which the subject was born | the place where the person is, or has been, resident |
| P569       | date on which the subject was born | date on which the subject died |
| P570       | date on which the subject was born | date or point in time when the organization/subject was founded/created |
| P576       | date on which the subject was born | date or point in time on which an organisation was dissolved/disappeared or a building demolished; see also discontinued date (P2669) |
| P577       | date on which the subject was born | date or point in time a work is first published or released |
| P580       | publication date | indicates the time an item begins to exist or a statement starts being valid |
| P582       | start time | indicates the time an item ceases to exist or a statement stops being valid |
| P585       | end time | time and date something took place, existed or a statement was true |
| P607       | point in time | battles, wars or other military engagements in which the person or item participated |
| P674       | conflict | characters which appear in this item (like plays, operas, operettas, books, comics, films, TV series, video games) |
| P676       | characters | author of song lyrics; also use P86 for music composer |
| P706       | location where persons were active | located on the specified landform. Should not be used when the value is only political/administrative (provinces, states, countries, etc.). Use P131 for administrative entity. |
| P710       | creation date | person, group of people or organization (object) that actively takes/took part in the event (subject). Preferably qualify with "object has role" (P3831). Use P1923 for team participants. |
| P737       | creation date | this person, idea, etc. is informed by that other person, idea, etc., e.g. "Heidegger was influenced by Aristotle". |
| P740       | location where a group or organization was formed | location where a group or organization was formed |
| P749       | parent organization | parent organization of an organisation, opposite of subsidiaries (P355) |
| P800       | notable work | notable scientific, artistic or literary work, or other work of significance among subject's works |
| P807       | separated from | subject was founded or started by separating from identified object |
| P840       | creation date | the narrative of the work is set in this location |
| P937       | applies to jurisdictional feature | location where persons were active |
| P1001      | applies to jurisdictional feature | the item (an institution, law, public office ...) belongs to or has power over or applies to the value (a territorial jurisdiction: a country, state, municipality, ...) |
| P1056      | product or material | material or product produced by a government agency, business, industry, facility, or process |
| P1198      | unemployment rate | portion of a workforce population that is not employed |
| P1336      | territory claimed by | administrative divisions that claim control of a given area |
| P1344      | participant of replaces | event a person or an organization was a participant in, inverse of P710 or P1923 |
| P1365      | person or item replaced | 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 | person or item which replaces another. Use P156 (followed by) if the item is not replaced (e.g. books in a series), nor identical |
| P1376      | country of origin | country, state, department, canton or other administrative division of which the municipality is the governmental seat |
| P1412      | languages spoken, written or signed | language(s) that a person speaks or writes, including the native language(s) |
| P1441      | work in which this fictional entity | work in which this fictional entity (Q14897293) or historical person is present 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.) |
| P3373      | present in work | related to the work (P355) |

Table 11: Relation list (II), including Wikidata IDs, names and descriptions of relations in DocRED.