# A Comprehensive Survey on Document-Level Information Extraction

**Hanwen Zheng**[1], **Sijia Wang**[1], **Lifu Huang**[1,2]
[1]Virginia Tech, [2]University of California, Davis
{zoez,sijiawang,lifuh}@vt.edu

## Abstract

Document-level information extraction (doc-IE) plays a pivotal role in the realm of natural language processing (NLP). This paper embarks on a comprehensive review and discussion of contemporary literature related to doc-IE. In addition, we conduct a thorough error analysis using state-of-the-art algorithms, shedding light on their limitations and remaining challenges for tackling the task of doc-IE. Our findings demonstrate that issues like entity coreference resolution and the lack of robust reasoning significantly hinder the effectiveness of document-level information extraction (doc-IE). Additionally, we uncover new challenges, including labeling noise and relation transitivity. The overarching objective of this survey paper is to provide valuable insights that can empower NLP researchers to further advance the performance of doc-IE.

## 1 Introduction

Natural language processing (NLP) triggers the present wave of artificial intelligence (Dosovitskiy et al., 2021; Liu et al., 2021; Zhang et al., 2021a; Zhang and Eskandarian, 2022). Information Extraction (IE) plays a vital role in all aspects of NLP by extracting structured information from unstructured texts (Lin et al., 2020; Wang et al., 2022). Document-level information extraction (**doc-IE**) has achieved significant progress, benefiting from the enormous data resources created by NLP researchers and the rapidly growing computational power resources (Yao et al., 2019; Xu et al., 2021b). However, several challenges persist within the realm of doc-IE research, such as entity coreference resolution, reasoning across long-span contexts, and lack of commonsense reasoning as shown in Figure 1. Furthermore, current doc-IE research predominantly focuses on restricted domains and languages (Zheng et al., 2019a; Yang et al., 2018; Tong et al., 2022; Li et al., 2021), which poses difficulties in fairly conducting model comparisons and hampers the generalizability of findings.

To gain a profound understanding of the current literature on doc-IE, we conduct a comprehensive survey of recent models and datasets for document-level relation extraction **(doc-RE)** and document-level event extraction **(doc-EE)**, focusing on those published in top NLP conferences such as ACL, EMNLP, and so on. These works span various languages and domains, providing a broad overview of advancements in the field. We also thoroughly analyze the errors of several state-of-the-art approaches and summarize several key remaining challenges and future research directions of doc-IE. The contributions of this survey paper include:

- To the best of our knowledge, we are the first to systematically review the literature on doc-IE, including both doc-EE and doc-RE.
- We conduct a thorough error analysis with the current state-of-the-art (SOTA) algorithms for doc-EE and doc-RE, and summarize several key remaining challenges that serve as a foundation for future advancements in doc-IE research and encourage researchers to further innovate and improve upon the various existing methodologies.

## 2 Task Definition

**Event Extraction** Event extraction (Grishman, 1997; Chinchor and Marsh, 1998; Ahn, 2006) is a task to identify and classify event triggers and relevant participants from natural language text. Formally, given a document consisting of a set of sentences where each sentence consists of a sequence of words, the objective of this task is to identify and extract the following components from a given document: **Event Mention**, which refers to the phrases or sentences denoting an event; **Event Trigger**, typically in the form of a verb that signals the occurrence of an event; **Event Type**, indicating the predefined type of event specified by the dataset, such as Conflict-Attack; **Argument Mention**, comprising entity mentions that provide additional details on the event, such as who, what, when, where, and how the event occurred; and finally, **Argument Role**, representing the role or type of argument associated with the entity.



*Proceedings of the Workshop on the Future of Event Detection (FuturED)*, pages 58–72

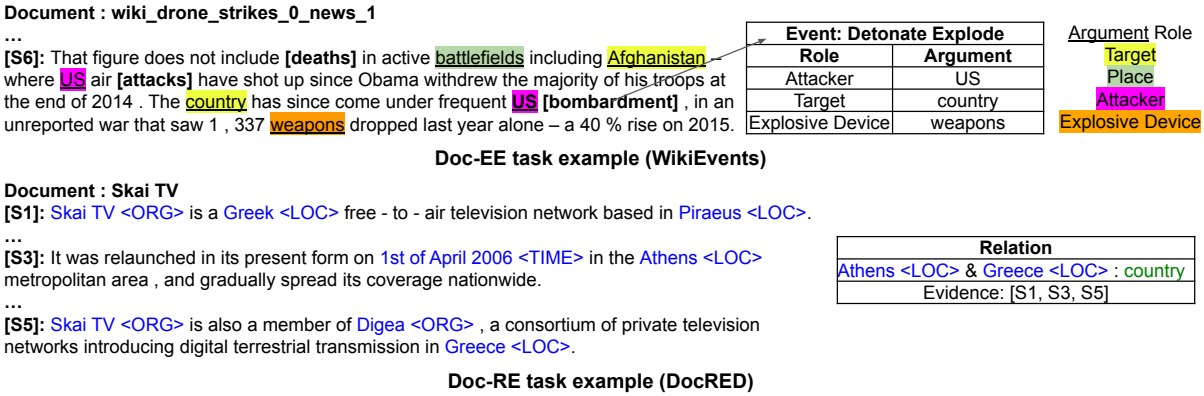


Figure 1: Examples of Document-Level Event Extraction (doc-EE) and Relation Extraction (doc-RE).

**Relation Extraction** Given a document $D$ with a set of sentences, we assume that $D$ also contains a set of entities $V = \{e_i\}_{i=1}^{N}$, which refer to units such as *People*, *Geographic Entity*, *Location*, *Organization*, *Date*, and *Number*. For each entity $e_i$, it might contain multiple entity mentions $e_i = \{m_j\}_{j=1}^{M}$, while each **Entity Mention** refers to a phrase within a text that identifies a specific entity. For instance, "NYC" and "the big apple" are both entity mentions for "New York City". The doc-RE task is to predict the relation types between an entity pair $(e_i, e_j)_{i,j \in \{1,\cdots,N\}, i \neq j}$, where $e_i$ stands for the subject and $e_j$ denotes the object. It is possible for an entity pair to have multiple relations, thereby rendering the task a multi-label classification problem. **Intra-sentence Relation** describes the relationship between entities within a single sentence, and the features within are often referred to as local features. **Inter-sentence Relation** refers to the relationship between entities across multiple sentences, and the features within are often referred to as global features.

## 3 Datasets

**Doc-EE Datasets** Existing doc-EE datasets are mainly collected from the news and financial domain. News is a large-scale accessible source of events like social emergencies and human life incidents, thus many datasets are created focusing on news events. Meanwhile, exploding volumes of digital financial documents, as a byproduct of continuous economic growth, have been created. Many datasets are created to help extract valuable structured information to detect financial risks or profitable opportunities. Statistics of the datasets for doc-EE are summarized in Table 1.

For news domain, **ACE-2005**[1] is a sentence-level event extraction (SEE) (Wang et al., 2022, 2023d) dataset but has been frequently used for evaluation in doc-EE. Unlike ACE-2005 which contains 5 groups of events covering *justice*, *life*, *business events*, etc, **MUC-4** (muc, 1992) focuses on one specific event type, *attack* events, containing 1,700 human-annotated news reports of terrorist attacks in Latin America collected by Federal Broadcast Information Services. MUC-4 includes six fine-grained incident types: *attack*, *kidnapping*, *bombing*, *arson*, *robbery*, and *forced work stoppage*, and four argument roles, including *individual perpetrator*, *organization perpetrator*, *physical target*, and *human target*. **Roles Across Multiple Sentences (RAMS)** (Ebner et al., 2020) is a crowd-sourced dataset with 9,124 event annotations on news articles from Reddit following the AIDA ontology[2]. **WikiEvents** (Li et al., 2021) follows the RAMS ontology containing 67 event types in a three-level hierarchy. Researchers used the BRAT interface for online annotation of event mentions (triggers and arguments) and event coreference separately. **CMNEE** (Zhu et al., 2024a) is a large-scale, open-source Chinese Military News Event Extraction dataset derived from the sentence-level military event detection dataset MNEE(Huang et al., 2022) and is manually annotated by human experts. **DocEE** (Tong et al., 2022) is the largest Doc-EE dataset to date. DocEE uses historical events and timeline events from Wikipedia as the candidate source to define 59 event types and 356 event argument roles. This dataset includes 27,485 document-level events and 180,528 event arguments that are manually labeled by humans.

For the financial domain, **ChFinAnn** (Zheng et al., 2019b) contains official disclosures such as

[1] https://catalog.ldc.upenn.edu/LDC2006T06

[2] https://aida.kmi.open.ac.uk/

| Dataset | # Docs | # Events | # Event types | # Roles | # Arguments | Ratio |
|---|---|---|---|---|---|---|
| ACE-2005[1] | 599 | 4,202 | 33 | 35 | 9,590 | - |
| MUC-4 (muc, 1992) | 1,700 | 1,514 | 4 | 5 | 2,641 | 13:2:2 |
| RAMS (Ebner et al., 2020) | 9,124 | 8,823 | 139 | 65 | 21,237 | 8:1:1 |
| WikiEvents (Li et al., 2021) | 246 | 3,951 | 67 | 59 | 5,536 | 10:1:1 |
| DocEE (Tong et al., 2022) | 27,485 | 27,485 | 59 | 356 | 180,520 | - |
| CMNEE (Zhu et al., 2024a) | 17,000 | 29,223 | 8 | 11 | 93,708 | 12:2:3 |
| ChFinAnn (Zheng et al., 2019b) | 32,040 | 47,824 | 5 | 35 | 289,871 | 8:1:1 |
| DCFEE (Yang et al., 2018) | 2,976 | 3,044 | 4 | 35 | - | 8:1:1 |
| DuEE-Fin (Zheng et al., 2019b) | 11,699 | 15,850 | 13 | 92 | 81,632 | 6:1:3 |

Table 1: Statistics of Doc-EE datasets. Ratio denotes training split ratio.

annual reports and earnings estimates, obtained from the Chinese Financial Announcement (CFA). The dataset has five event types: *Equity Freeze*, *Equity Repurchase*, *Equity Underweight*, *Equity Overweight* and *Equity Pledge*, with 35 different argument roles in total. In contrast to Doc-EE with one event in each document, 29.0% of the documents in ChFinAnn contain multiple events. **DCFEE** (Yang et al., 2018) comes from companies' official finance announcements and focuses on four event types: *Equity Freeze*, *Equity Pledge*, *Equity Repurchase*, and *Equity Overweight*. Data labeling was done through distant supervision. **DuEE-Fin** (Zheng et al., 2019b) is the largest human-labeled Chinese financial dataset. It is collected from real-world Chinese financial news and annotated with 13 event types. 29.2% of the documents contain multiple events and 16.8% of events consist of multiple arguments.

Several doc-RE datasets are from the biomedical domain. **Drug-gene-mutation (DGM)** (Jia et al., 2019) contains 4,606 PubMed articles, which are automatically labeled via distant supervision. DGM annotations include three entity types: *drugs*, *genes*, and *mutations*, and three relation types, including *drug-gene-mutation*, *drug-mutation*, and *gene-mutation relations*. **GDA** (Wu et al., 2019) gene-disease association corpus contains 30,192 titles and abstracts from PubMed articles that have been automatically labeled for *genes*, *diseases*, and *gene-disease associations* via distant supervision. **CDR** (Luan et al., 2018) is manually annotated for *chemicals*, *diseases*, and *chemical-induced disease (CID)* relations by domain experts. It contains the titles and abstracts of 1,500 PubMed articles and is split into training, validation, and test sets equally. **BioRED** (Luo et al., 2022) builds on previous biomedical datasets by including entity types such as gene/protein, disease, and chemical, along with gene-disease and chemical–chemical relations.

Additionally, doc-RE has been explored in other domains or languages. **DocRED** (Yao et al., 2019) is a human-annotated Doc-RE dataset, that includes 132,375 entities and 56,354 relational facts annotated on 5,053 Wikipedia documents. Doc-RED is generated by mapping Wikidata triples, originating from a comprehensive knowledge base closely intertwined with Wikipedia, onto complete English Wikipedia documents to get entity annotations. **RE-DocRED** (Tan et al., 2022b) refines 4,053 documents in the DocRED dataset targeting on resolving the problem of false negative samples. RE-DocRED increased the relation triples from 50,503 to 120,664 and decreased the *no_relation* samples by 3.1% by adding the missing relation triples back to the original DocRED. Moreover, **DocRED-FE** (Wang et al., 2023b) focus on fine-grained entity types; **DocRED-IE** (Bouziani et al., 2024) expands with five additional subtasks: *Mention Detection*, *Entity Typing*, *Entity Disambiguation*, *Coreference Resolution*, and their combinations, *Named Entity Recognition (NER) and Entity Linking* as in **DWIE** (Zaporojets et al., 2021). **KnowledgeNet** (Mesquita et al., 2019) offers links to a reference knowledge base (KB) for entity and relation annotations. **SciREX** (Jain et al., 2020) is a document-level relation extraction dataset that contains multiple IE tasks, such as Binary and N-ary relation classification. It consists of both automatic and human-annotated articles in the field of computer science. **HacRED** (Cheng et al., 2021) is a Chinese Doc-RE dataset collected from CN-DBpedia (Xu et al., 2017) that focuses on hard cases, such as long text and long distance between argument pairs, containing distractors or multiple homogeneous entity mentions. Statistics of the datasets for doc-RE are summarized in Table 2.

## 4 Methods

The fundamental challenge in doc-EE and doc-RE is to express document content in a concise and

| Dataset | Annotation | # Types | # Facts | % Inter-rel | # Train | # Dev | #Test |
|---|---|---|---|---|---|---|---|
| DGM (Jia et al., 2019) | Distant Supervision | 3 | - | 64.5% | 32,040 | - | - |
| CDR (Luan et al., 2018) | Human-annotated | 1 | - | 29.8% | 1,500 | 500 | 500 |
| GDA (Wu et al., 2019) | Distant Supervision | 1 | - | 15.6% | 30,192 | 5,839 | 1,000 |
| BioRED (Luo et al., 2022) | Combined | 2 | - | - | 4,178 | 1,162 | 1,163 |
| KnowledgeNet (Mesquita et al., 2019) | Human-annotated | 15 | 13,000 | - | - | - | - |
| DocRED (Yao et al., 2019) | Distant Supervision | 96 | 50,345 | 12.5% | 3,053 | 1,000 | 1,000 |
| Re-DocRED (Tan et al., 2022b) | Combined | 96 | 120,664 | 12.5% | 3,053 | 500 | 500 |
| DocRED-FE (Wang et al., 2023b) | Combined | 96 | 32,366 | - | 2,596 | 1,000 | - |
| DocRED-IE (Bouziani et al., 2024) | Automated | 96 | 37,486 | - | 3,008 | 300 | 700 |
| SciREX (Jain et al., 2020) | Human-annotated | 2 | - | 99.0% | 438 | 131 | 131 |
| HacRED (Cheng et al., 2021) | Combined | 26 | 65,225 | 25.4% | 9,231 | 1,500 | 1,500 |
| DWIE (Zaporojets et al., 2021) | Distant Supervision | 65 | 21,749 | - | 700 | - | 100 |

Table 2: Statistics of Doc-RE datasets.

| Task | Main Category | Sub Category | Approaches |
|---|---|---|---|
| Doc-EE | Multi-granularity-based | Sentence→ Paragraph→ Document | Yang et al. (2018), Huang and Jia (2021), Wang et al. (2023a) |
| | Graph-based | Heterogeneous graph | Zheng et al. (2019b), Xu et al. (2021d), Zhu et al. (2022), Xu et al. (2022), Zhang et al. (2024) |
| | Task-specific | Attention\Transformer | Yang et al. (2021), Liang et al. (2022), Liu et al. (2024) |
| | | Other Networks | Huang and Peng (2021) |
| | Generation-based | - | Li et al. (2021), Zeng et al. (2022), Huang et al. (2023) |
| | Memory-based | - | Du et al. (2022), Cui et al. (2022) |
| | LLM-based | - | Gatto et al. (2024), Zhou et al. (2024), Uddin et al. (2024) |
| Doc-RE | Multi-granularity-based | Sentence→ Paragraph→ Document | Tang et al. (2020) |
| | | Mention→ Entity | Jia et al. (2019) |
| | Graph-based | Heterogeneous graph | Quirk and Poon (2017), Peng et al. (2017), Song et al. (2018), Guo et al. (2019), Sahu et al. (2019), Christopoulou et al. (2019), Wang et al. (2020), Xu et al. (2021d), Zeng et al. (2020), Li et al. (2020), Zhang et al. (2020), Xu et al. (2023), Xu et al. (2021c), Zhu et al. (2024b), Mao et al. (2024) |
| | | Homogeneous graph | Nan et al. (2020) |
| | Task-specific | Attention\Transformer | Zhou et al. (2021), Tan et al. (2022a) |
| | | Other Networks | Xu et al. (2021a), Zhang et al. (2021b), Bouziani et al. (2024), Wang et al. (2023c), Ma et al. (2023) |
| | Evidence-based | Path reasoning | Huang et al. (2021) |
| | | Evidence retrieval | Xie et al. (2022), Xiao et al. (2022) |

Table 3: Typology of Doc-IE methods.

effective way such that key information is maintained. A typology of existing doc-EE and doc-RE approaches categorized by model design is shown in Table 3.

### 4.1 Doc-EE Approaches

**Multi-granularity-based Models** Multi-granularity-based designs employ two strategies: either addressing intermediate tasks using various models or utilizing the same model in a hierarchically ordered manner to independently tackle each subtask of information extraction, such as from sentence level to document level. The standard procedure involves concatenating features from each level to complete the IE tasks. **DCFEE** (Yang et al., 2018) first uses a sequence tagging model to automatically extract sentence-level events, and then proposes a key-event detection model based on a convolutional neural network (CNN) to predict document-level key event. **SCDEE** (Huang and Jia, 2021) uses graph attention network (GAT) to transform document-level features to event communities in order to detect event types at the sentence level. Wang et al. (2023a) collect sentence-level and document-level embeddings by various probing techniques to help probe event mentions in documents. Multi-granularity-based approaches improve the utilization of information across different granularities and the aggregation of global context, but they lose precision in co-reference resolution and capturing long-span dependencies.

**Graph-based Models** Graph-based models generally construct a graph with words, mentions, entities, arguments, or sentences as nodes and define different types of edges across the entire document, further predicting the relations by reasoning on the graph. **Doc2EDAG** (Zheng et al., 2019b) treats the

doc-EE task as an event table-filling task by generating an entity-based directed acyclic graph. It decides which entity node to expand until the graph is fully recovered. **GIT** (Xu et al., 2021d) propose a heterogeneous graph to extract corresponding arguments by expanding a constrained event type tree while tracking and storing records in global memory. **PTPCG** (Zhu et al., 2022) prune the complete graph by deciding whether entity pairs retain an edge based on semantic similarity between entities. **TSAR** (Xu et al., 2022) prune the Abstract Meaning Representation (AMR) graph with span information and surrounding events, and treat event argument extraction (EAE) as a link prediction task. However, while dependency graphs contain rich structural information, the pruning strategy may not always preserve relevant details. **GAM** (Zhang et al., 2024) builds a semantic mention graph capturing co-existence, co-reference, and co-type relations. Graph-based models enhance document representation by allowing the model to learn in an aggregated format, but they may struggle to identify the same event across multiple events and establish their relationships.

**Generation-based Models** **Bart-Gen** (Li et al., 2021) ask a PLM to fill in the blank in the Doc-EE templates. **EA2E** (Zeng et al., 2022) focuses on event-aware argument extraction by labeling arguments from nearby events in the document to enhance context and extracting event iteratively during generation. **S2C** (Huang et al., 2023) generates all possible arguments and predict the corresponding event arguments in a simple to complex order. A typical challenge that generation-based approaches face is in identifying precise spans.

**Memory-based Models** Du et al. (2022) stores gold-standard and previously generated events in memory, allowing the decoder to dynamically retrieve event knowledge and decode arguments based on event dependencies. **HRE** (Cui et al., 2022) mimics human reading with a two-stage process: rough reading detects event types, and elaborate reading extracts complete event records with arguments, updating memory with event type and argument information. Memory-based models require additional storage capacity, which can be challenging for large datasets, but they enable the model to retain event and argument dependencies effectively.

**LLM-based Models** LLM-based models leverage the extensive prior knowledge of large language models like LLAMA2 (Touvron et al., 2023) and GPT-4(OpenAI, 2024) for in-context learning. Gatto et al. (2024) investigates two data augmentation strategies for synthesizing document-level EAE samples and utilizes LLMs for slot-filling to address EAE tasks. Zhou et al. (2024) introduces the Link-of-Analogy Prompting technique, which guides LLMs in generating analogies to facilitate retrieval, mapping, and evaluation processes in a cross-event context. Uddin et al. (2024) provides several question-generation strategies such as prompting using GPT-4 to ask questions about the arguments of an event and inputs those questions to BART-based models for EAE. LLM-based models don't require additional training or fine-tuning, but their limitations lie in their computational demands and difficulty in tuning and optimizing prompts.

**Models with task-specific designs** Models with task-specific designs mostly rely on attention-based architectures or other NN-based (neural networks), which replicate complex interactions among arguments by implicitly capturing long-distance dependencies. **DE-PPN** (Yang et al., 2021) uses an encoder-decoder structure where the document encoder captures document-aware sentence and argument embeddings, while the decoder simultaneously decodes events, arguments, and roles. **ReDEE** (Liang et al., 2022) is the first to use entity relation information for doc-EE tasks, which utilizes SSAN (Xu et al., 2021a) to extract relation triples as input and calculates the attention between entities and candidate arguments to gain dependency. **DEED** (Huang and Peng, 2021) is an end-to-end model that utilizes Deep Value Networks (DVN), a structured prediction algorithm that effectively bridges the disparity between ground truth and prediction. This model directly incorporates event trigger prediction into DVN, thereby efficiently capturing cross-event dependencies for document-level event extraction. **DEEIA** (Liu et al., 2024) proposes a multi-event argument extraction method using a dependency-guided encoding module to enhance the correlation between prompts and contexts, and an event-specific information aggregation module to provide event-specific information for better contextual understanding. These task-oriented approaches effectively capture long-span dependencies but may overlook sentence-level information and often require long input lengths.

### 4.2 Doc-RE Approaches

**Multi-granularity-based Models** The first work on doc-RE using a multi-granularity method is by Jia et al. (2019), employing multiscale representation learning to aggregate mention representations and ensemble sub-relations. The **HIN** (Hierarchical Inference Network) (Tang et al., 2020) uses Bi-LSTMs at the token, sentence, and document levels to extract features as sequences and weighs the overall features with the attention mechanism to obtain both local and global information.

**Graph-based Models** **DISCREX** (Quirk and Poon, 2017) constructs a document graph with word-based nodes and edges representing intra- and inter-sentence level relations including dependency, adjacency, and discourse relations. Peng et al. (2017) contributes a Graph-LSTM model with a Bi-LSTM to encode the document graph to two directed acyclic graphs (DAG). Song et al. (2018) compares bidirectional graph LSTM with bidirectional DAG LSTM and concludes that the former, which retains the original graph structure, performs better. **AGGCNs** (Guo et al., 2019) proposes an end-to-end graph convolutional network (GCN) that encodes the entire graph using multi-head self-attention to learn edge weights and uses densely connected layers to extract global information. Sahu et al. (2019) designates words as individual nodes and establishes five types of edges to represent inter- and intra-sentence dependency. The model then uses an edge-oriented GCN to retain aggregated node representations.

**EoG** (Christopoulou et al., 2019) is a pioneering graph-based model. It uses entities as nodes and forms unique edge representations through the paths between nodes to better capture the paired relations. To predict relations between entity pairs, EoG makes iterative inferences on the path between the entities and aggregates every edge to a direct entity-entity edge. Many papers adapted from EoG can be divided into two main categories: homogeneous and heterogeneous graphs. **LSR** (Nan et al., 2020) uses graph structure as a latent variable to form a homogeneous graph. Unlike EoG which uses a human-constructed graph, LSR learns structured attention to refine the graph dynamically and constructs latent structures based on the previous refinement. For heterogeneous graphs, different types of edges are defined, representing unique features, functions, and even dual graphs. **GLRE** (Wang et al., 2020) utilizes a multi-layer relational GCN to learn global entity representations as queries in self-attention, while using sentence-level information as keys to learn local entity representations. **HeterGSAN** (Xu et al., 2021d) constructs a heterogeneous graph based on EoG and encodes it using a GAT. HeterGSAN improves the performance of relation classification by reconstructing a dependency-based path between each pair of entities. **POR** (Xu et al., 2023) builds upon HeterGSAN using a path-retrieving method on paired entities to extract path features through an LSTM.

Dual graphs are normally used to capture hierarchical information. **GAIN** (Zeng et al., 2020) utilizes a heterogeneous mention-level graph to model interactions between the document and all mentions. **GEDA** (Li et al., 2020) optimizes entity representations with two attention layers and a heterogeneous GCN layer. **DHG** (Zhang et al., 2020) propose a framework with two heterogeneous graphs: a structure modeling graph using words and sentences as nodes to better capture document structure information and a relation reasoning graph using mentions and entities as nodes to perform multi-hop relation reasoning. **DRN** (Xu et al., 2021c) passes encoded sentences and entities as a heterogeneous graph to a multi-layer GCN and meanwhile uses a self-attention mechanism to learn better contextual document-level representations.

**Models with task-specific designs** Models with task-specific designs focus on capturing contexts and entity information through tailored designs for document-level tasks, utilizing either adequate neural network structures or novel loss functions. **SSAN** (Xu et al., 2021a) integrates structural dependencies within and throughout the encoding stage of the network, not only enabling simultaneous context reasoning and structure reasoning but also efficiently modeling these dependencies in all network layers. **ATLOP** (Zhou et al., 2021) leverages pre-trained attention weights for localized context pooling and adopts an adaptive thresholding loss (ATL) to ensure that each entity maintains the same representation and balances the logits of positive and negative labels. **DocuNet** (Zhang et al., 2021b) divides model construction into three parts leveraging a u-shaped semantic segmentation network to refine entity feature extraction. **KD** (Tan et al., 2022a) calculates self-attention in the vertical and horizontal directions of a paired entity table as the axial attention to enhance entity pair representations. The authors propose an adaptive

focal loss (AFL) where the logits of entity relations are balanced with thresholds to address long-tailed classes.

**Path (Evidence)-based Models** Path-based models construct evidence paths and make relational decisions by reasoning on crucial information between entity pairs or sentences, instead of extracting features from the complete document. **THREE** (Huang et al., 2021) presents three kinds of paths to find the supporting sentences: consecutive paths, multi-hop paths, and default paths for entity pairs. **EIDER** (Xie et al., 2022) defines "evidence sentences", as a minimal number of sentences needed to predict the relations between certain pairs of entities in a document. **SAIS** (Xiao et al., 2022) utilizes two intermediary phases to obtain evidence information: pooled evidence retrieval, which distinguishes entity pairs with and without supporting sentences, and fine-grained evidence retrieval, which produces more interpretable evidence specific to each relation of an entity pair. Those approaches typically utilize supporting sentences to serve as evidence from existing datasets such as DocRED. The path-based approaches exhibit extraordinary performance because they align human perception and intuition in the doc-RE task, where we read through the whole document and evaluate sentences that are important for the task.

## 5 Discussion

To understand the limitations and remaining challenges of the current document-level IE approaches, we evaluate three state-of-the-art Doc-RE methods, **KD** (Tan et al., 2022a), **DRN** (Xu et al., 2021c), and **SAIS** (Xiao et al., 2022), on the DocRED and Re-DocRED datasets.Similarly, we also evaluate two state-of-the-art Doc-EE methods, graph-based model **TSAR** (Xu et al., 2022) and generative model **EA2E** (Zeng et al., 2022), on the WikiEvents dataset, and another two Doc-EE methods, graph-based model **PTPCG** (Zhu et al., 2022) and task-specific model **ReDEE** (Liang et al., 2022), on ChFinAnn dataset. For each work, we randomly select 50 errors and examine the cause of them. We finally conclude seven major types of errors for document-level information extraction. Figure 2, 3, 4 show the distribution of the seven types of errors on each dataset and Table 4 show several error examples.

**Entity coreference resolution** Document-level texts contain a large number of recognized entities along with coreferential words such as them, he, which, etc. Entity coreference resolution errors happen when the model fails to resolve all mentions in a document that refer to the same entity.

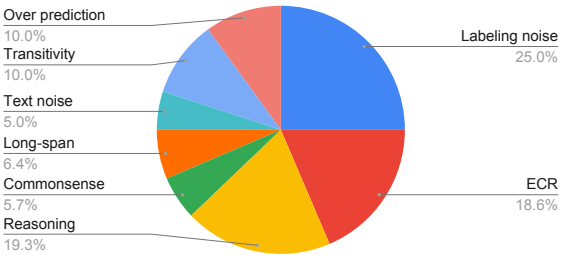

Figure 2: Doc-RE error distribution in DocRED and Re-DocRED

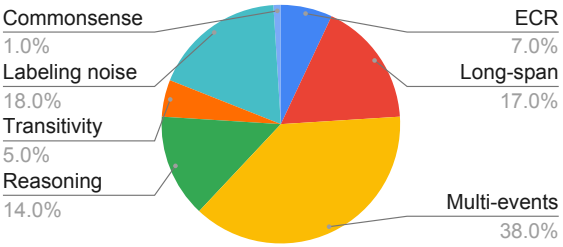

Figure 3: Doc-EE error distribution in ChFinAnn

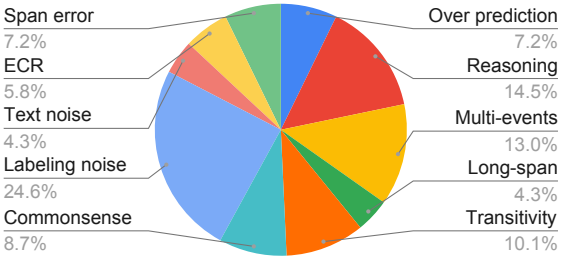

Figure 4: Doc-EE error distribution in WikiEvents

**Reasoning error** This type of error mainly relates to multi-hop logical reasoning. Document-level texts contain considerable amounts of information, so models may fail to give correct logical inferences based on the given information. Inferring from multi-hop information requires a model to have a high level of natural language understanding ability.

**Long-span** Document contains multiple sentences in a long span. This error happens when the model fails to capture the full context of a document or uses global information for inference.

**Commonsense knowledge** The error occurs when models fail to correctly extract relations or events or assume the wrong semantics due to a lack of commonsense and background knowledge, which humans are able to learn or understand instinctively. Many datasets are specific to some domains, and in the absence of relevant background and domain-specific knowledge, models may inac-

| Error Type | Text | GT | Pred |
|---|---|---|---|
| ECR | The game retains some common elements from previous **Zelda** <MISC> installments, such as the presence of **Gorons** <PER>, while introducing **Kinstones** <PER> and other new gameplay features. | The Legend of Zelda <MISC>, Gorons <PER> : characters | N/A |
| Multi-hop reasoning | **Parvathy** <PER> married film actor **Jayaram** <PER> who was her co-star in many films on... She has **two** <NUM> children, **Kalidas Jayaram** <PER> and **Malavika Jayaram** <PER>. | Jayaram <PER>, Kalidas Jayaram <PER>: child | N/A |
| Commonsense | **Olympic Gold** <MISC> is the official video game of the **XXV Olympic Summer Games** <MISC>, hosted by **Barcelona** <LOC>, **Spain** <LOC> in **1992** <TIME>. | XXV Olympic Summer Games <MISC> , Spain <LOC> : country | N/A |
| Over prediction | **The Link River** <LOC> is a short river connecting **Upper Klamath Lake** <LOC> to **Lake Ewauna** <LOC> in the city of **Klamath Falls** <LOC> in the **U.S.** <LOC> state of **Oregon** <LOC>. | N/A | Lake Ewauna <LOC>, Oregon <LOC> : located in the administrative territorial entity |
| Learned prior | **Ngoako Ramatlhodi** <PER>, a senior member of the **African National Congress** <ORG> , was **South Africa** <LOC> 's Minister | N/A | African National Congress <ORG>, South Africa <LOC> : country |
| Relation transitivity | At the **2007** <TIME> **European Indoor Athletics Championships** <MISC> he won a silver medal in the **4 x 400 metres** <NUM> relay , with teammates Ivan **Buzolin** <PER> , **Maksim Dyldin** <PER> and **Artem Sergeyenkov** <PER> | Artem Sergeyenkov <PER>, European Indoor Athletics Championships <MISC> : participant of | N/A |

Table 4: Examples of Doc-RE errors: the column of GT shows the ground truth event annotations while the column of Pred shows the predicted event mentions.

curately reason or misinterpret information.

**Relation transitivity error** Documents tend to have many entities appearing in the same sentence or across sentences. Relation transitivity errors occur when a model fails to correctly infer a relation between two entities based on their individual relations with a third entity. Additionally, not all relations are transitive, thus the model should correctly recognize when transitivity applies.

**Over prediction error** This error type refers to the spurious error (as we presented in Table 4) where there is no ground truth relation between two entities but the model predicts a relation, and can be caused by a number of reasons. For instance, when using large pre-trained language models to encode the documents, learned prior can cause models to make overconfident predictions.

In addition to shared error types with Doc-RE, we observe two more types of errors based on the WikiEvents and ChFinAnn datasets.

**Multi-events error** In Doc-EE tasks, documents contain multiple events that overlap or occur simultaneously, which requires the model to have sufficient training or advanced techniques to learn the inherent complexity of multi-event documents. In an event-trigger-annotated dataset such as WikiEvents, the model can fail at assigning arguments to the correct events or matching roles to arguments. In a trigger-not-annotated dataset like ChFinAnn, event detection errors may occur when models try to identify and differentiate distinct events within the document due to the complex contextual structure of each event, as shown in the example of Figure 5.

**Span errors** Models face span error types mainly associated with previous tasks like entity recognition or caused by the different linguistic features and complexities of datasets. For example, nominal mention recognition and argument span mismatch errors are common in many works, particularly in generative methods.

**Noisy data** This issue comprises natural language noises and labeling noises. Real-world documents contain noisy, unstructured, or poorly formatted content, causing difficulties in identifying entities and extracting relations. See further discussion in Section C of the Appendix.

## 6 Remaining Challenges

Current difficulties can be broadly categorized into three areas: information spread out, multiple mentions and multiple entity pairs throughout the entire document, some information must be deduced from several sentences or transferred by other relations in order to be discovered. The first two issues have been addressed by existing approaches using attention mechanisms and graph networks, though multiple-step reasoning is less widely focused. Existing methods rely on LLMs to learn syntactic features while neglecting the relation transitivity between entity pairs and the evidence trace of reasoning. Progressively, more methods try to use evidence sentences or evidence paths to infer complicated relations. Models continue to struggle with capturing commonsense and knowledge-based information as it is difficult to from the training data. Previous works have tried adaptive losses for balancing the positive and negative examples to allevi-


Figure 5: Multi-event error example in ChFinAnn: The colors in the sentence highlight the gold standard event annotations ( Event_0 , Event_1 , Event_2 ). The predicted event mentions and arguments are shown in the table. When predicting the arguments, e.g., *EquityHolder* role of Event_1 and Event_2, the model gets distracted by Event_0 and predicts *Zhu Jiang*. *December 28, 2018* and *200,000* are shared arguments of Event_1 and Event_2.

ate class imbalance problems. Existing works still struggle with long-tailed, ambiguous, and complicated classes, and have a hard time differentiating similar classes. Dataset-wise, creating annotated datasets for this task is time-consuming and expensive, which limits the amount of data available for training and evaluation. Domain-specific datasets differ from general datasets but are necessary for identifying relations that are specific to certain domains, understanding domain-specific terminology, and handling the high variability of language used in different domains.

There are several promising future directions. First, it is beneficial to incorporate entity coreference systems into doc-IE models, which we believe will play an important role in resolving ECR and multi-hop reasoning errors. Second, more investigations are needed to design a model with multi-hop reasoning capability. Finally, doc-EE and doc-RE can be supplementary tasks to each other. The information produced by these two tasks can provide a more complete picture of the information given in the document.

## 7 Conclusion

We conducted a thorough error analysis of current state-of-the-art algorithms, highlighting the limitations of existing approaches and identifying key challenges in document-level IE. Our analysis revealed that issues such as entity coreference resolution, insufficient reasoning capabilities, labeling noise, and relation transitivity significantly impact the performance of current models, providing insights for future research. Despite notable progress in the field, we conclude that persistent challenges within both datasets and models hinder the development of robust and generalizable solutions. Overcoming these obstacles will be essential for advancing document-level IE models in the future.

## Limitations

Due to the constraint that some state-of-the-art models had not released their code at the time we conducted the error analysis, we carefully selected iconic models featuring key designs and unique characteristics for evaluation. The current datasets include only Chinese and English data in the news, finance, biomedical, and Wikipedia domains; therefore, our analysis primarily focuses on studies using English and Chinese datasets within these domains. Nevertheless, we believe that our conclusions will generalize to other domains, languages, and future datasets. The limitations identified in this survey are expected to provide valuable insights and may reflect similar challenges in unexplored areas.

This survey focuses exclusively on text-only document-level information extraction (IE) due to the lack of research and datasets available for multimodal document-level IE. However, the challenges identified in this survey are expected to be critical and may serve as motivation for future research efforts in this area.

## Acknowledgements

This research is supported by the award No. 2238940 from the Faculty Early Career Development Program (CAREER) of the National Science Foundation (NSF). The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

## References


1992. *Fourth Message Uunderstanding Conference (MUC-4): Proceedings of a Conference Held in McLean, Virginia, June 16-18, 1992*.

David Ahn. 2006. The stages of event extraction. In *Proceedings of the Workshop on Annotating and Reasoning about Time and Events*, pages 1–8.

Nacime Bouziani, Shubhi Tyagi, Joseph Fisher, Jens Lehmann, and Andrea Pierleoni. 2024. Rexel: An end-to-end model for document-level relation extraction and entity linking. *Preprint*, arXiv:2404.12788.

Qiao Cheng, Juntao Liu, Xiaoye Qu, Jin Zhao, Jiaqing Liang, Zhefeng Wang, Baoxing Huai, Nicholas Jing Yuan, and Yanghua Xiao. 2021. Hacred: A large-scale relation extraction dataset toward hard cases in practical applications. In *FINDINGS*.

Nancy Chinchor and Elaine Marsh. 1998. Muc-7 information extraction task definition. In *Proceeding of the seventh message understanding conference (MUC-7), Appendices*, pages 359–367.

Fenia Christopoulou, Makoto Miwa, and Sophia Ananiadou. 2019. Connecting the Dots: Document-level Neural Relation Extraction with Edge-oriented Graphs. *arXiv preprint*. ArXiv:1909.00228 [cs].

Shiyao Cui, Xin Cong, Bowen Yu, Tingwen Liu, Yucheng Wang, and Jinqiao Shi. 2022. Document-Level Event Extraction via Human-Like Reading Process. *arXiv preprint*. ArXiv:2202.03092 [cs].

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Xinya Du and Claire Cardie. 2020. Event Extraction by Answering (Almost) Natural Questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 671–683, Online. Association for Computational Linguistics.

Xinya Du, Sha Li, and Heng Ji. 2022. Dynamic Global Memory for Document-level Argument Extraction. *arXiv preprint*. ArXiv:2209.08679 [cs].

Seth Ebner, Patrick Xia, Ryan Culkin, Kyle Rawlins, and Benjamin Van Durme. 2020. Multi-sentence argument linking. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Joseph Gatto, Parker Seegmiller, Omar Sharif, and Sarah M. Preum. 2024. Large language models for document-level event-argument data augmentation for challenging role types. *Preprint*, arXiv:2403.03304.

Ralph Grishman. 1997. Information extraction: Techniques and challenges. In *International summer school on information extraction*, pages 10–27. Springer.

Zhijiang Guo, Yan Zhang, and Wei Lu. 2019. Attention Guided Graph Convolutional Networks for Relation Extraction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 241–251, Florence, Italy. Association for Computational Linguistics.

Hongbin Huang, Jiao Sun, Hui Wei, Kaiming Xiao, Mao Wang, and Xuan Li. 2022. A dataset of domain events based on open-source military news.

Kung-Hsiang Huang and Nanyun Peng. 2021. Document-level Event Extraction with Efficient End-to-end Learning of Cross-event Dependencies. In *Proceedings of the Third Workshop on Narrative Understanding*, pages 36–47, Virtual. Association for Computational Linguistics.

Quzhe Huang, Yanxi Zhang, and Dongyan Zhao. 2023. From simple to complex: A progressive framework for document-level informative argument extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6129–6140, Singapore. Association for Computational Linguistics.

Quzhe Huang, Shengqi Zhu, Yansong Feng, Yuan Ye, Yuxuan Lai, and Dongyan Zhao. 2021. Three Sentences Are All You Need: Local Path Enhanced Document Relation Extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 998–1004, Online. Association for Computational Linguistics.

Yusheng Huang and Weijia Jia. 2021. Exploring Sentence Community for Document-Level Event Extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 340–351, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sarthak Jain, Madeleine van Zuylen, Hannaneh Hajishirzi, and Iz Beltagy. 2020. SciREX: A challenge dataset for document-level information extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7506–7516, Online. Association for Computational Linguistics.

Robin Jia, Cliff Wong, and Hoifung Poon. 2019. Document-Level N-ary Relation Extraction with Multiscale Representation Learning. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3693–3704, Minneapolis, Minnesota. Association for Computational Linguistics.

Kamran Kowsari, Kiana Jafari Meimandi, Mojtaba Heidarysafa, Sanjana Mendu, Laura Barnes, and Donald

Brown. 2019. Text classification algorithms: A survey. *Information*, 10(4):150.

Bo Li, Wei Ye, Zhonghao Sheng, Rui Xie, Xiangyu Xi, and Shikun Zhang. 2020. Graph Enhanced Dual Attention Network for Document-Level Relation Extraction. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1551–1560, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Sha Li, Heng Ji, and Jiawei Han. 2021. Document-Level Event Argument Extraction by Conditional Generation. *arXiv preprint*. ArXiv:2104.05919 [cs].

Yuan Liang, Zhuoxuan Jiang, Di Yin, and Bo Ren. 2022. RAAT: Relation-Augmented Attention Transformer for Relation Modeling in Document-Level Event Extraction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4985–4997, Seattle, United States. Association for Computational Linguistics.

Ying Lin, Heng Ji, Fei Huang, and Lingfei Wu. 2020. A joint neural model for information extraction with global features. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7999–8009, Online. Association for Computational Linguistics.

Wanlong Liu, Li Zhou, Dingyi Zeng, Yichen Xiao, Shaohuan Cheng, Chen Zhang, Grandee Lee, Malu Zhang, and Wenyu Chen. 2024. Beyond single-event extraction: Towards efficient document-level multi-event argument extraction. *Preprint*, arXiv:2405.01884.

Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9992–10002, Los Alamitos, CA, USA. IEEE Computer Society.

Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. 2018. Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3219–3232, Brussels, Belgium. Association for Computational Linguistics.

Ling Luo, Po-Ting Lai, Chih-Hsuan Wei, Cecilia N Arighi, and Zhiyong Lu. 2022. Biored: a rich biomedical relation extraction dataset. *Briefings in Bioinformatics*, 23(5):bbac282.

Youmi Ma, An Wang, and Naoaki Okazaki. 2023. DREEAM: Guiding Attention with Evidence for Improving Document-Level Relation Extraction. *arXiv preprint*. ArXiv:2302.08675 [cs] version: 1.

Yanxu Mao, Peipei Liu, and Tiehan Cui. 2024. Gega: Graph convolutional networks and evidence retrieval guided attention for enhanced document-level relation extraction. *Preprint*, arXiv:2407.21384.

Filipe Mesquita, Matteo Cannaviccio, Jordan Schmidek, Paramita Mirza, and Denilson Barbosa. 2019. KnowledgeNet: A benchmark dataset for knowledge base population. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 749–758, Hong Kong, China. Association for Computational Linguistics.

Guoshun Nan, Zhijiang Guo, Ivan Sekulić, and Wei Lu. 2020. Reasoning with Latent Structure Refinement for Document-Level Relation Extraction. *arXiv preprint*. ArXiv:2005.06312 [cs].

OpenAI. 2024. Chatgpt (gpt-4). `https://chat.openai.com`. Accessed: 2024-08-19.

Nanyun Peng, Hoifung Poon, Chris Quirk, Kristina Toutanova, and Wen-tau Yih. 2017. Cross-Sentence N-ary Relation Extraction with Graph LSTMs. *Transactions of the Association for Computational Linguistics*, 5:101–115. Place: Cambridge, MA Publisher: MIT Press.

Chris Quirk and Hoifung Poon. 2017. Distant Supervision for Relation Extraction beyond the Sentence Boundary. *arXiv preprint*. ArXiv:1609.04873 [cs].

Sunil Kumar Sahu, Fenia Christopoulou, Makoto Miwa, and Sophia Ananiadou. 2019. Inter-sentence Relation Extraction with Document-level Graph Convolutional Neural Network. *arXiv preprint*. ArXiv:1906.04684 [cs].

Linfeng Song, Yue Zhang, Zhiguo Wang, and Daniel Gildea. 2018. N-ary Relation Extraction using Graph-State LSTM. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2226–2235, Brussels, Belgium. Association for Computational Linguistics.

Qingyu Tan, Ruidan He, Lidong Bing, and Hwee Tou Ng. 2022a. Document-Level Relation Extraction with Adaptive Focal Loss and Knowledge Distillation. *arXiv preprint*. ArXiv:2203.10900 [cs].

Qingyu Tan, Lu Xu, Lidong Bing, Hwee Tou Ng, and Sharifah Mahani Aljunied. 2022b. Revisiting DocRED – Addressing the False Negative Problem in Relation Extraction. *arXiv preprint*. ArXiv:2205.12696 [cs].

Hengzhu Tang, Yanan Cao, Zhenyu Zhang, Jiangxia Cao, Fang Fang, Shi Wang, and Pengfei Yin. 2020. HIN: Hierarchical Inference Network for Document-Level Relation Extraction. In *Advances in Knowledge Discovery and Data Mining*, Lecture Notes in Computer Science, pages 197–209, Cham. Springer International Publishing.

MeiHan Tong, Bin Xu, Shuai Wang, Meihuan Han, Yixin Cao, Jiangqi Zhu, Siyu Chen, Lei Hou, and Juanzi Li. 2022. DocEE: A Large-Scale and Fine-grained Benchmark for Document-level Event Extraction. In *Proceedings of the 2022 Conference of*

*the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3970–3982, Seattle, United States. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Md Nayem Uddin, Enfa George, Eduardo Blanco, and Steven Corman. 2024. Generating uncontextualized and contextualized questions for document-level event argument extraction. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5612–5627, Mexico City, Mexico. Association for Computational Linguistics.

Barry Wang, Xinya Du, and Claire Cardie. 2023a. Probing representations for document-level event extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12675–12683, Singapore. Association for Computational Linguistics.

Difeng Wang, Wei Hu, Ermei Cao, and Weijian Sun. 2020. Global-to-Local Neural Networks for Document-Level Relation Extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3711–3721, Online. Association for Computational Linguistics.

Hongbo Wang, Weimin Xiong, Yifan Song, Dawei Zhu, Yu Xia, and Sujian Li. 2023b. Docred-fe: A document-level fine-grained entity and relation extraction dataset. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Jize Wang, Xinyi Le, Xiaodi Peng, and Cailian Chen. 2023c. Adaptive hinge balance loss for document-level relation extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3872–3878, Singapore. Association for Computational Linguistics.

Sijia Wang, Mo Yu, Shiyu Chang, Lichao Sun, and Lifu Huang. 2022. Query and extract: Refining event extraction as type-oriented binary decoding. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 169–182, Dublin, Ireland. Association for Computational Linguistics.

Sijia Wang, Mo Yu, and Lifu Huang. 2023d. The art of prompting: Event detection based on type specific prompts. In *ACL 2023*. Association for Computational Linguistics.

Ye Wu, Ruibang Luo, Henry C. M. Leung, Hing-Fung Ting, and Tak Wah Lam. 2019. Renet: A deep learning approach for extracting gene-disease associations from literature. In *RECOMB*.

Yuxin Xiao, Zecheng Zhang, Yuning Mao, Carl Yang, and Jiawei Han. 2022. SAIS: Supervising and Augmenting Intermediate Steps for Document-Level Relation Extraction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2395–2409, Seattle, United States. Association for Computational Linguistics.

Yiqing Xie, Jiaming Shen, Sha Li, Yuning Mao, and Jiawei Han. 2022. Eider: Empowering Document-level Relation Extraction with Efficient Evidence Extraction and Inference-stage Fusion. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 257–268, Dublin, Ireland. Association for Computational Linguistics.

Benfeng Xu, Quan Wang, Yajuan Lyu, Yong Zhu, and Zhendong Mao. 2021a. Entity Structure Within and Throughout: Modeling Mention Dependencies for Document-Level Relation Extraction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(16):14149–14157. Number: 16.

Bo Xu, Yong Xu, Jiaqing Liang, Chenhao Xie, Bin Liang, Wanyun Cui, and Yanghua Xiao. 2017. Cndbpedia: A never-ending chinese knowledge extraction system. In *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*.

Runxin Xu, Tianyu Liu, Lei Li, and Baobao Chang. 2021b. Document-level event extraction via heterogeneous graph-based interaction model with a tracker. *Preprint*, arXiv:2105.14924.

Runxin Xu, Peiyi Wang, Tianyu Liu, Shuang Zeng, Baobao Chang, and Zhifang Sui. 2022. A Two-Stream AMR-enhanced Model for Document-level Event Argument Extraction. *arXiv preprint*. ArXiv:2205.00241 [cs].

Wang Xu, Kehai Chen, and Tiejun Zhao. 2021c. Discriminative Reasoning for Document-level Relation Extraction. *arXiv preprint*. ArXiv:2106.01562 [cs].

Wang Xu, Kehai Chen, and Tiejun Zhao. 2021d. Document-Level Relation Extraction with Reconstruction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(16):14167–14175. Number: 16.

Wang Xu, Kehai Chen, and Tiejun Zhao. 2023. Document-Level Relation Extraction with Path Reasoning. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(4):1–14.

Hang Yang, Yubo Chen, Kang Liu, Yang Xiao, and Jun Zhao. 2018. DCFEE: A Document-level Chinese Financial Event Extraction System based on Automatically Labeled Training Data. In *Proceedings of ACL 2018, System Demonstrations*, pages 50–55, Melbourne, Australia. Association for Computational Linguistics.

Hang Yang, Dianbo Sui, Yubo Chen, Kang Liu, Jun Zhao, and Taifeng Wang. 2021. Document-level Event Extraction via Parallel Prediction Networks. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6298–6308, Online. Association for Computational Linguistics.

Yuan Yao, Deming Ye, Peng Li, Xu Han, Yankai Lin, Zhenghao Liu, Zhiyuan Liu, Lixin Huang, Jie Zhou, and Maosong Sun. 2019. DocRED: A Large-Scale Document-Level Relation Extraction Dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 764–777, Florence, Italy. Association for Computational Linguistics.

Klim Zaporojets, Johannes Deleu, Chris Develder, and Thomas Demeester. 2021. Dwie: An entity-centric dataset for multi-task document-level information extraction. *Information Processing & Management*, 58(4):102563.

Qi Zeng, Qiusi Zhan, and Heng Ji. 2022. EA2E: Improving Consistency with Event Awareness for Document-Level Argument Extraction. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2649–2655, Seattle, United States. Association for Computational Linguistics.

Shuang Zeng, Runxin Xu, Baobao Chang, and Lei Li. 2020. Double Graph Based Reasoning for Document-level Relation Extraction. *arXiv preprint*. ArXiv:2009.13752 [cs].

Ce Zhang and Azim Eskandarian. 2022. A quality index metric and method for online self-assessment of autonomous vehicles sensory perception. *ArXiv*, abs/2203.02588.

Ce Zhang, Azim Eskandarian, and Xuelai Du. 2021a. Attention-based neural network for driving environment complexity perception. *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*, pages 2781–2787.

Jian Zhang, Changlin Yang, Haiping Zhu, Qika Lin, Fangzhi Xu, and Jun Liu. 2024. A semantic mention graph augmented model for document-level event argument extraction. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1577–1587, Torino, Italia. ELRA and ICCL.

Ningyu Zhang, Xiang Chen, Xin Xie, Shumin Deng, Chuanqi Tan, Mosha Chen, Fei Huang, Luo Si, and Huajun Chen. 2021b. Document-level Relation Extraction as Semantic Segmentation. *arXiv preprint*. ArXiv:2106.03618 [cs] version: 2.

Zhenyu Zhang, Bowen Yu, Xiaobo Shu, Tingwen Liu, Hengzhu Tang, Wang Yubin, and Li Guo. 2020. Document-level Relation Extraction with Dual-tier Heterogeneous Graph. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1630–1641, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Shun Zheng, Wei Cao, Wei Xu, and Jiang Bian. 2019a. Doc2EDAG: An End-to-End Document-level Framework for Chinese Financial Event Extraction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 337–346, Hong Kong, China. Association for Computational Linguistics.

Shun Zheng, Wei Cao, Wei Xu, and Jiang Bian. 2019b. Doc2EDAG: An end-to-end document-level framework for chinese financial event extraction. In *EMNLP*.

Hanzhang Zhou, Junlang Qian, Zijian Feng, Hui Lu, Zixiao Zhu, and Kezhi Mao. 2024. Llms learn task heuristics from demonstrations: A heuristic-driven prompting strategy for document-level event argument extraction. *Preprint*, arXiv:2311.06555.

Wenxuan Zhou, Kevin Huang, Tengyu Ma, and Jing Huang. 2021. Document-Level Relation Extraction with Adaptive Thresholding and Localized Context Pooling. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(16):14612–14620. Number: 16.

Mengna Zhu, Zijie Xu, Kaisheng Zeng, Kaiming Xiao, Mao Wang, Wenjun Ke, and Hongbin Huang. 2024a. CMNEE:a large-scale document-level event extraction dataset based on open-source Chinese military news. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3367–3379, Torino, Italia. ELRA and ICCL.

Tong Zhu, Xiaoye Qu, Wenliang Chen, Zhefeng Wang, Baoxing Huai, Nicholas Jing Yuan, and Min Zhang. 2022. Efficient Document-level Event Extraction via Pseudo-Trigger-aware Pruned Complete Graph. *arXiv preprint*. ArXiv:2112.06013 [cs].

Xudong Zhu, Zhao Kang, and Bei Hui. 2024b. FCDS: Fusing constituency and dependency syntax into document-level relation extraction. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7141–7152, Torino, Italia. ELRA and ICCL.

## A Evaluation Metrics

In document-level information extraction (IE), the primary evaluation metrics are Precision (P), Recall (R), and Macro-F1 score (Kowsari et al., 2019). Additionally, for doc-RE, Ign F1 is also used as an evaluation metric (Yao et al., 2019) which refers to the F1 score that excludes relational facts shared by the training and dev/test sets. This metric is important for evaluating the generalizability of the model, as it disregards triples that have already been included in the annotated training dataset.

## B Performance of Existing Methods

Performance of Doc-RE Existing Methods are shown in Table 8, Table 5, and Table 6. Performance of Doc-EE Existing Methods are shown in Table 9 and Table 7.

| Model | F1 |
|---|---|
| SAIS$^{O}_{RE+CR+ET}$-SciBERT (Xiao et al., 2022) | 87.10 |
| DocuNet-SciBERT-base (Zhang et al., 2021b) | 85.30 |
| Eider(Rule)-SciBERT-base (Xie et al., 2022) | 84.54 |
| ATLOP-SciBERT-base (Zhou et al., 2021) | 83.90 |
| SSAN-SciBERT (Xu et al., 2021a) | 83.70 |

Table 5: Doc-RE GDA rank

| Model | F1 |
|---|---|
| SAIS$^{O}_{RE+CR+ET}$-SciBERT (Xiao et al., 2022) | 79.00 |
| DocuNet-SciBERT-base (Zhang et al., 2021b) | 76.30 |
| Eider(Rule)-SciBERT-base (Xie et al., 2022) | 70.63 |
| ATLOP-SciBERT-base (Zhou et al., 2021) | 69.40 |
| SSAN-SciBERT (Xu et al., 2021a) | 68.70 |

Table 6: Doc-RE CDR rank

| Model | F1 |
|---|---|
| ReDEE (Liang et al., 2022) | 81.90 |
| Git (Xu et al., 2021d) | 80.30 |
| PTPCG (Zhu et al., 2022) | 79.40 |
| SCDEE (Huang and Jia, 2021) | 78.90 |
| DE-PPN (Yang et al., 2021) | 77.90 |
| HRE (Cui et al., 2022) | 76.80 |
| Doc2EDAG (Zheng et al., 2019b) | 76.30 |

Table 7: Doc-EE ChFinAnn rank

## C Additional error analysis

**Noisy data** Natural language can be ambiguous or vague, leading to uncertainty in model inference. To overcome the limitations of the cost of creating annotated datasets, researchers commonly apply

| Model | F1 | Ign-F1 |
|---|---|---|
| KD-Rb-l (Tan et al., 2022a) | 67.28 | 65.24 |
| SSAN-RoBERTa-large+Adaptation (Xu et al., 2021a) | 65.92 | 63.78 |
| SAIS-RoBERTa-large (Xiao et al., 2022) | 65.11 | 63.44 |
| Eider-RoBERTa-large (Xie et al., 2022) | 64.79 | 62.85 |
| DocuNet-RoBERTa-large (Zhang et al., 2021b) | 64.55 | 62.40 |
| ATLOP-RoBERTa-large (Zhou et al., 2021) | 63.40 | 61.39 |

Table 8: Doc-RE DocRED rank

| Model | Arg Identification | | Arg Classification | |
|---|---|---|---|---|
| | Head F1 | Coref F1 | Head F1 | Coref F1 |
| TSAR$_{large}$ (Xu et al., 2022) | 76.62 | 75.52 | 69.70 | 68.79 |
| EA$^2$E (Zeng et al., 2022) | 74.62 | 75.77 | 68.61 | 69.70 |
| BART-Gen(Li et al., 2021) | 71.75 | 72.29 | 64.57 | 65.11 |
| OneIE(Li et al., 2021) | 61.88 | 63.63 | 57.61 | 59.17 |
| BERT-QA(Du and Cardie, 2020) | 61.05 | 64.59 | 56.16 | 59.36 |

Table 9: Doc-EE WikiEvent rank

automatic labeling strategies like distant supervision to generate large-scale training data. However, this leads to several minor problems due to noise and bias: nested entities (i.e., some entities can be embedded within other entities), false negative labels (i.e., entity pairs not known to be related but getting labeled as such in the dataset), and missing ground truth labels.

Note that Doc-EE errors vary between ChFinAnn and WikiEvents. There could be a number of factors behind the different Doc-EE error distribution between ChFinAnn and WikiEvents. One crucial factor is the diversity in underlying statistics between datasets due to their distinct domains and languages. Compared to the news dataset WikiEvents, the Chinese financial dataset ChFinAnn requires less commonsense comprehension. Each dataset contains unique linguistic features and complexities. WikiEvents has annotated trigger words, and arguments tend to be near the trigger words, whereas ChFinAnn can have events spread across the entire document and is more likely to interfere with other events. Therefore, long-span and multi-events are major error types in ChFinAnn. Moreover, various model designs and approaches usually aim to address specific challenges and optimize performance on the respective dataset.