Multi-turn and Multi-Granularity Reader for Document-level Event Extraction

Anonymous ACL-IJCNLP submission

Abstract

Most existing event extraction (EE) works mainly focus on extracting events from one sentence. However, in real-world applications, event arguments of one event always scatter across sentences and multiple events co-exist frequently in one document. Thus these scenarios require document-level event extraction (DEE) which aims to extract events across sentences from a document. In this paper, we propose a new paradigm of DEE by formulating it as a machine reading comprehension (MRC) task (i.e., the extraction of event arguments is cast to identifying the answer span from the document). The MRC formalization comes with two advantages: firstly, the MRC-based method can provide end-to-end document-level modeling for DEE. Secondly, the question can provide semantic information of roles. Moreover, for addressing the unique challenges (arguments-scattering and multi-events) of DEE, we introduce an multi-turn and multi-granularity reader to aggregate the extracted argument information and capture hierarchical nature of a document. The empirical results demonstrate that our method achieves superior performance on the MUC-4 and the ChFinAnn datasets, increasing the state-of-the-art (SOTA) results to 58.14 (+3.72) and 77.5 (+1.3) respectively.

1 Introduction

Event extraction (EE) aims at extracting events from unstructured raw texts, which has received growing interest these years. As a fundamental and challenging task in natural language processing (NLP), EE can produce valuable structured information to facilitate many NLP applications such as knowledge base construction, question answering, language understanding and so on (Ji and Grishman, 2011; Berant et al., 2014). A great number of previous works (Chen et al., 2015; Nguyen et al., 2016; Yang et al., 2019; Wang et al., 2019; Li et al., 2020b; Du and Cardie, 2020b) focus on the sentence-level EE (SEE) which aims to detect events and extract arguments from one sentence. However, in real-world applications, many scenarios need document-level EE (DEE) which aims to extract events from a whole document.

DEE focus on a more challenging and more realistic setting: extracting events with their arguments from whole document. In contrast to SEE, DEE has two critical complications: **1) arguments-scattering**: arguments of one event scatter across multiple sentences in a document. For example, as shown in Figure 1, the arguments of *Event-1* are distributed in different sentences ($S_4 - S_6$) dispersedly and the extraction from an individual sentence will lead to incomplete results. It requires a view of a larger context to determine argument spans and capture long-distance dependencies among arguments across sentences. **2) multi-events**: there may be multiple events that co-occur in a document. As shown in Figure 1, there are two events *Event-1* and *Event-2* in a document with the same event type and there is no obvious textual boundary between the two events. This challenge requires DEE model to determine extracted argument belong to which event. To this end, previous works (Yang et al., 2018a; Zheng et al., 2019) formulated DEE as a two-step paradigm: from sentence-level candidate argument extraction to document-level event fusion. Although the above-mentioned works for DEE have achieved success, there are two problems: First, these methods for DEE are based on the sentence-level extraction, which lacks integrating document-level information for candidate argument extraction and the two-step paradigm will also cause error propagation. Second, the classification-based method for DEE is incapable of modeling the semantic information of event role labels explicitly.

In this paper, we propose a **M**ulti-turn and **M**ulti-

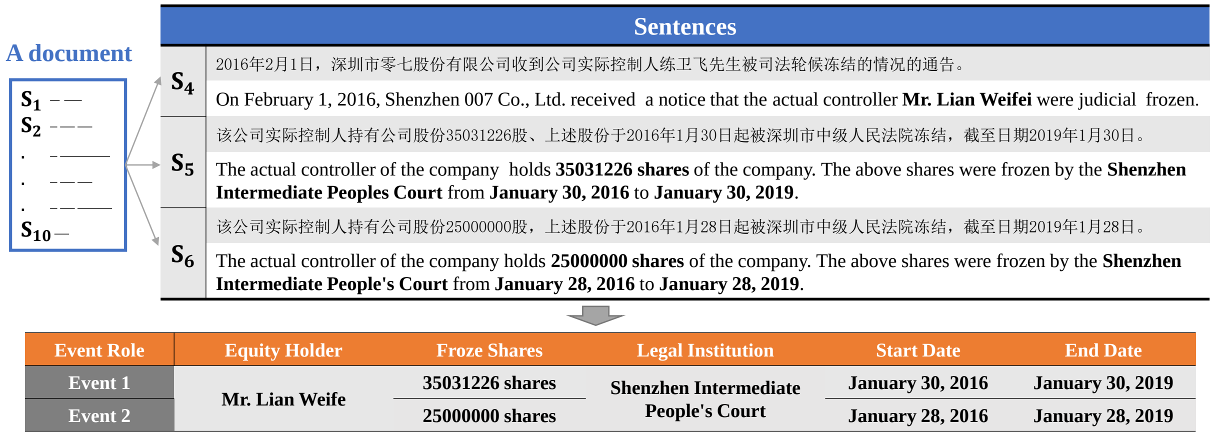

Figure 1: A sample of document-level event extraction with two *Equity Freeze* events whose arguments scatter across multiple sentences. In the document, only three sentences ($S_4 - S_6$) are shown, the lists *Event 1* and *Event 2* are annotated structured events and words in bold-faced are event arguments with specific roles.

granularity **R**eader (MMR) for DEE that can extract events from whole document directly without the stage of preliminary sentence-level extraction. Specifically, the MMR is based on a machine reading comprehension (MRC) formulation. The MRC-based formulation comes with two advantages for handling the task of DEE: 1) Modeling document-level information directly. the MRC-based framework can learn and inference event information in a document directly. 2) Semantic information of roles. Compared with the tagging-based method where categories are merely class index, the MRC-based model can provide external evidence for roles by encoding the role-specific query. For example, in the task of DEE, the event role type *Legal Institution* is treated as a one-hot vector in category classification method. But in the MRC formulation, the query(e.g., *"the legal institution that executes this freeze, usually institutions or courts."*) encourages the model to retrieve information about *institutions* or *courts* from long texts directly.

Despite the benefits of modeling the extraction task in the form of MRC-based paradigm, there are still some bottlenecks when apply the MRC-based paradigm to the task of DEE. The first one is how to capture long-distance dependencies between arguments effectively as they may scatter across sentences. The second one is how to model lengthy document while most of MRC methods are based on the Transformer (Vaswani et al., 2017) architecture which is limited to a fixed-length (e.g., 512) input. To address the challenges for DEE, we make the following improvements under the framework of MRC. Firstly, we introduce a multi-turn MRC form for DEE to better model the relationship between arguments explicitly. An event recorder is designed to encode the event histories (i.e., the extracted arguments for an event) that can guide the extraction of the corresponding arguments for the current event. Secondly, we introduce a multi-granularity reader for modeling the long texts and capturing hierarchical nature of a document. The transformer-based encoder is designed to dynamically learn the local context (e.g., sentence-level) and the global context(e.g., document-level).

In experiments, we evaluate our model on the widely used DEE datasets (MUC-4 and ChFinAnn) and the experimental results under the standard evaluation demonstrate the effectiveness of our proposed method. Specifically, our method achieves performance over current state-of-the-art (SOTA) models with 3.72, 1.3 improvements on the MUC-4 and ChFinAnn respectively. Additionally, we conduct experiments with few-shot settings and results prove that the MRC-based DEE method can be well transferred to new event types or event roles with a few samples.

In summary, our contributions are as follows:

- We formulate the document-level event extraction as a MRC paradigm that can introduce the semantic information of roles and model the document-level information directly.

- We propose a multi-turn and multi-granularity reader to model the dependencies between arguments and capture hierarchical nature of document.

- We conduct extensive experiments on both widely used DEE datasets (ChFinAnn and MUC-4). Results show that our model significantly outperforms the baseline models and also demonstrate promising results in addressing few-shot scenarios.

2 Related Work

2.1 Event Extraction

A great number of EE researches focus on the SEE and most of them are based on the expert-annotated benchmark ACE 2005 (Doddington et al., 2004) dataset. In recent years, as neural networks proved the effectiveness for NLP, many approaches (Chen et al., 2015; Nguyen et al., 2016; Yang et al., 2019; Chan et al., 2019; Björne and Salakoski, 2018; Yang et al., 2019) have been proposed to improve performance on this task by employing deep learning models.

As many real-world applications need DEE, there are two widely used datasets (MUC-4 and ChiFinAnn) for exploring it. The first one is the task of document-level event role filler extraction which is based on the classic MUC-4 dataset (MUC-4, 1992). This task aims to identify event role fillers with associated role types (i.e., Perpetrator Individual, Perpetrator Organization, Target, Victim and Weapon) from context. Recent works explores the local and additional context to extract the role fillers by manually designed linguistic features (Patwardhan and Riloff, 2009; Huang and Riloff, 2011, 2012) or neural-based contextual representation (Du and Cardie, 2020a; Du et al., 2020; Chen et al., 2020).

For exploring the real challenges (i.e., multi-events and arguments-scattering) for DEE, DCFEE (Yang et al., 2018a) proposed a pipeline method that contains a neural-based sequence tagging model for SEE and a key-event detection model with an arguments-completion strategy for DEE. Doc2EDAG (Zheng et al., 2019) proposed an event tables filling method with entity-based path expanding which achieves the state-of-art for DEE. Although these methods have achieved success for DEE, there are two key issues. First, these works were based on a two-stage process from sentence-level extraction to document-level fusion which lacks modeling document-level information. Second, these works ignored the explicit semantic information of roles. In this work, we formulate DEE as an MRC task that can model the lengthy document directly and capture the semantic information of roles.

2.2 Machine Reading Comprehension

In recent years, the MRC task has been widely investigated since the release of large-scale corpora (Rajpurkar et al., 2016; Joshi et al., 2017; Lai et al., 2017; Yang et al., 2018b). The main-stream MRC models extract text spans from passages given the questions and achieved good results (Seo et al., 2016; Wang and Jiang, 2016; Shen et al., 2017; Zheng et al., 2020; Devlin et al., 2019). Most of these MRC models tackle the text span extraction by predicting the starting and ending position of the answer based on two multi-class classifiers. They treat questions and documents as sequences and focus on building interaction between them, where the attention mechanism is most widely used. And pre-trained language model like BERT (Devlin et al., 2019) has proved to be extremely helpful for MRC tasks.

Recently, there have been explorations on formulating non-QA NLP tasks as machine reading comprehension. (Levy et al., 2017; Li et al., 2019) transform the extraction of entities and relations as a multi-turn QA formalization. (Li et al., 2020c) formalize the NER task as an MRC question answering task to address overlapping or nested entities. (Liu et al., 2020; Du and Cardie, 2020b; Li et al., 2020a) introduced an MRC paradigm for event extraction in an end-to-end manner. extraction Different from the work above, we focus on the DEE with more complicated scenarios (i.e., long context dependencies need to be captured) and unique challenges (i.e., multi-events and arguments-scattering). We show that the proposed multi-turn MRC model with an event recorder can solve these challenges well and achieve state-of-art results on DEE.

3 Methodology

Figure 2 illustrates our approach which is denoted by MMR (Multi-turn and Multi-granularity Reader) for DEE. Specifically, the extraction of events in a document is transformed into multi-turn MRC steps as follows: Firstly, a multi-granularity reader is designed to aggregate contextualized representations for tokens from multiple granularities (local and global). Then, based on the contextual representation of the document, we can get the event type by a linear classifier. Secondly, we construct questions for each event role based on the

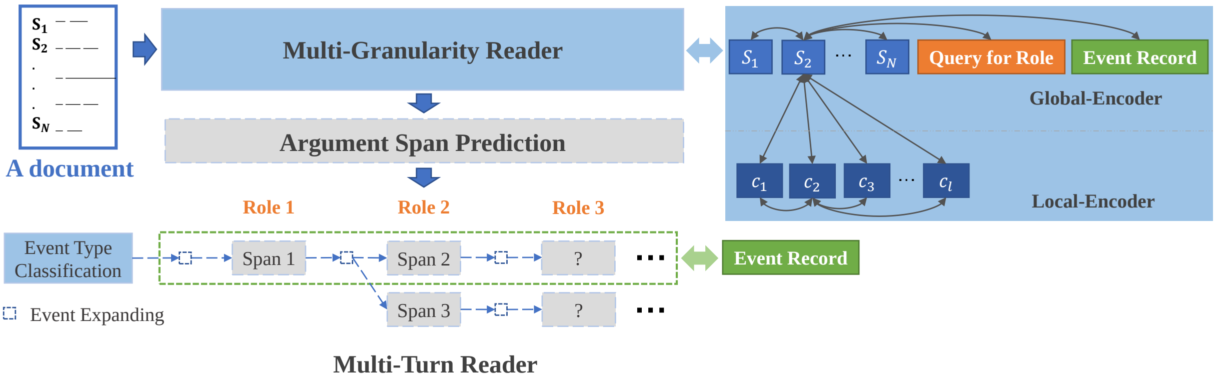

Figure 2: The overview of the proposed model MMR for DEE. Given a document with context, the extraction for each role is mapping to a multi-turn MRC task that answer the constructed question following a predefined event role order by a multi-granularity reader.

definition of role types from the predefined event schema given the predicted event type. Note that the event role query construction strategy cost very little manpower engineering than hand-designed templates. Finally, event arguments are extracted progressively by answering the question until all event roles are traversed. As there will be multiple events co-occur in a document, an event expanding operation is used for generating another extracted event. <u>And an event recorder is applied to conserve the historical event information which can guide the extraction of the remaining arguments for one event.</u>

3.1 Task Definition

Before introducing MMR for DEE in this section, we first describe the task formalization for it. Formally, we denote $\mathcal{T}$ and $\mathcal{R}$ as the predefined event types and role categories, respectively. Given an input document comprised of N_s sentences $\mathcal{D} = \{S_i\}_i^{N_s}$, the DEE task aims to extract one or more structured events $Y = \{y_i\}_i^k$ where each event y_i^t with event type t contains a series of roles $(r_i^1, r_i^2, \ldots, r_i^n)$ filled by arguments $(a_i^1, a_i^2, \ldots, a_i^n)$. k is the number of events contained in the document, n is the number of predefined roles for the event type t, $t \in \mathcal{T}$ and $r \in \mathcal{R}$.

3.2 Multi-granularity Reader

Previous works (Seo et al., 2016; Wang and Jiang, 2016; Shen et al., 2017; Zheng et al., 2020) have shown that the MRC framework can learn and inference in a document through the question-context pair. Most of them are based on the the Transformer architecture (Vaswani et al., 2017) with multi-layers self-attention mechanism to model long dependencies between tokens with limited sequence length (e.g., BERT (Devlin et al., 2019) allows fixed-length (e.g., 512) inputs, but the average token length in the task of DEE is 762). A straightforward solution for modeling lengthy document is sliding window, but this method sacrifices the possibility that the distant tokens "pay attention" to each other. To break the length limitation and model capture hierarchical nature of a document, we propose a multi-granularity reader to encode document-aware information for each token. The multi-granularity encoder mechanism is composed of three parts: local transform, global transformer and global-to-local attention.

Local Encoder. A local transformer is designed to capture the local contextual (sentence-level) representation for each token. Specifically, given a document $\mathcal{D} = \{S_i\}_i^{N_s}$ with N_s sentences, and each sentence S_i with a sequence of tokens $[c_{i,1}, c_{i,2}, \ldots, c_{i,l}]$, where l is the sentence length. Each sentence S_i is fed to the context encoder, which outputs the contextualized representations. In this paper, we adopt the Transformer (Vaswani et al., 2017) as a primary context encoder to get the local contextualized representation for each token in sentence S_i:

$$h_{i,1}, \ldots, h_{i,l} = \mathrm{Enc}_{\mathrm{Local}}(c_{i,1}, \ldots, c_{i,l}) \quad (1)$$

where $H_i \in \mathbb{R}^{d_h \times l}$, d_h denotes the hidden size. The local contextual representation $H_i \in \mathbb{R}^{d_h}$ of sentence S_i can be obtained by the max-pooling operation over the token sequence in sentence S_i. Similarly, the query embedding H_q can obtained by the same local-transformer over the tokens sequence in question q_m for role type m.

Global Encoder. To enable the awareness of document-level contexts and role-specific query for sentences, we employ a document-aware encoder to facilitate the interaction between all sentences and a role specific query. We employ the Transformer module, Transformer-global, as the encoder to get the document-aware embedding for sentences and facilitate the interaction between all sentences and query.

$$H_1, \ldots, H_{N_s} = \mathrm{Enc}_{\mathrm{Global}}(H_1, \ldots, H_{N_s}, H_q) \tag{2}$$

Global-to-Local Attention. To aggregate document-aware representation for each token, we construct a global-to-local attention to leverage the document-level context. Specifically, given the local contextual representation for i-th sentence and j-th token $h_{i,j}$ and document-aware global representation H_i for sentence i, where each token is calculated as follows:

$$z_{i,j} = \sum_{j=1}^{N_s} \mathrm{Softmax}(Q_h K_h^{\mathrm{T}}) V_h \tag{3}$$

where $Q_h = W_q h_{i,j}$, $K_h = k_q H_j$ and $V_h = V_q H_j$ are linear transformations. $z_{i,j}$ is the representation for each token in a document which incorporates a contextual representation of the global information. Then we sum the local representation $h_{i,j}$ and the global representation $z_{i,j}$ to get the fused representations for each token in the document.

3.3 Event Type Classification and Question Construction

Through the multi-granularity reader, the document-aware representation for each sentence in a document can be obtained. To predict event type in a document, we conduct a binary classification for each event type over the document representation that is calculated by operating the max-pooling over all sentence representations $\mathcal{H}_s$. Then, given the predicted event type, we construct queries for roles based on the definition of role types from the predefined event schema. Note that the event role query construction strategy cost very little manpower engineering than hand-designed templates. The extraction for each role type is mapping to an MRC sub-task: answer the corresponding questions following a manually defined event role order. In this way, argument spans can be extracted from the context following the order gradually, where each answer is either an argument or a special empty filler NA.

3.4 Argument Span Prediction

Considering that the document might have multiple arguments for a specific query, we apply a classification layer to the hidden representation for each token to predict the BIO boundary labels. For each token h_i, the probability of the candidate BIO label can be calculated as:

$$P(y_{label}|h_i) = \mathrm{softmax}(W \cdot h_i + b) \tag{4}$$

where $W \in \mathbb{R}^{d_h \times N_l}$ and $b \in \mathbb{R}^{d_h \times N_l}$. N_l is the size of BIO label set, and y_{label} denotes the predicted boundary label. Consequently, argument span can be extracted from the label sequence by identifying the boundaries given a document with role-specific query.

3.5 Event Expanding and Event Recording

As there may be multi-events co-occur in the document, to handle the challenge of multi-events effectively, we adopt a heuristics approach for event expanding. Specifically, for each MRC sub-task with a role-specific question, if there are multiple answers with different mentions, we recognize that a new event will be generated. To model the long-distance dependencies among arguments for DEE, We design an event recorder to encode the historical information for each event. With this design, each MRC sub-task can own unique event histories that can distinctly guide the extraction of later arguments. Specifically, the structured historical event information (i.e., extracted arguments with specific roles) is concatenated with role types to form a sequence and the local transformer is applied to encode the historical event record as the contextual representation H_E.

3.6 Training and Testing

During training, we calculate a cross-entropy loss for each argument prediction as follows:

$$\mathcal{L}_{ap} = \mathrm{CrossEntropy}(y_{label}, L_{label}) \tag{5}$$

Then, we sum the sum prediction loss for events prediction with preconditioned steps before multi-turn MRC as follows:

$$\mathcal{L}_{all} = \lambda_1 \sum_{i}^{N_r} \mathcal{L}_{ap} + \lambda_2 \mathcal{L}_{ec} \tag{6}$$

where $\mathcal{L}_{ec}$ are the cross-entropy loss function for event type classification and N_r is the number of role types for event type t. λ_1, λ_2 and λ_3 are hyper-parameters. At inference, given a document as the input, the events are extracted by answering the corresponding questions following a predefined role order.

4 Experiments and Analysis

4.1 Datasets and Evaluation Metrics

MUC-4. The MUC-4 dataset consists of 1,700 documents with a fixed set of event types (e.g., terrorist events) and associated role types (i.e., Perpetrator Individual, Perpetrator Organization, Target, Victim and Weapon). There are 1300 documents for training, 200 documents (TST1+TST2) for development, and 200 documents (TST3+TST4) for testing. As the task of role filler extraction aims to identify spans of text and there is no phase to determine extracted role fillers belong to which event (i.e., no multi-events evaluation). Following the prior work (Du and Cardie, 2020a), we adopt head noun phrase match and exact match accuracy to compare the extractions against gold role fillers for evaluation. Our results are reported as Precision (P), Recall (R), and F-measure (F-1) score for the macro average for all the event roles.

ChFinAnn. Doc2EDAG (Zheng et al., 2019) conducted a large-scale document-level event extraction dataset Chinese financial announcements (ChFinAnn) which contains 32,040 documents in total with five financial event types: *Equity Freeze* (EF), *Equity Repurchase* (ER), *Equity Underweight* (EU), *Equity Overweight* (EO) and *Equity Pledge* (EP). Following (Zheng et al., 2019), we leverage the ChFinAnn data to evaluate our proposed method with the same train, development, and test set. We evaluate our method as the same evaluation standard as Doc2EDAG (Zheng et al., 2019) as there may be multiple events in a document. Specifically, for each document, we pick one predicted event with one most similar ground-truth event without replacement to calculate Precision (P), Recall (R), and F-measure (F-1) for each event type. As an event type often includes multiple roles, micro-averaged role-level scores are calculated as the final document-level event extraction metric.

Implementation Details We adopt Transformer-base, which has 12 layers, 768 hidden units, and 12 attention heads, as our local encoder. For the global encoder, we set the number of transformer layers as 4. During training, we set λ_1=0.1 and λ_2=0.9 and employ the AdamW optimizer with the learning rate 2e-5 for training 50 epochs and pick the best parameters by the validation score on the development set. Besides, we denote the ascending order of the empty argument ratio as the event role order for multi-turn QA because more informative event histories can facilitate later argument extraction.

4.2 Results on MUC-4

For the MUC-4 dataset, event role fillers are extracted by answering the question for role types.

Baselines. **GLACIER** (Patwardhan and Riloff, 2009) used a sentential event recognizer to select sentences and then applied a plausible role-filler recognizer to extract role fillers as results. **TIER** (Huang and Riloff, 2011) extract role fillers from the secondary context which processes the extraction into three stages: classifying narrative document, recognizing event sentence, and noun phrase analysis. **Cohesion Extract** (Huang and Riloff, 2012) identifies candidate role fillers in the document and then refines the candidate set with cohesion sentence classifier. **MGR** (Du and Cardie, 2020a) propose a tagging-based model to dynamically incorporate paragraph- and sentence-level representations based on contextualized embeddings produced by the pre-trained language model.

Main Results. Table 1 gives the main results on the MUC-4 for head noun match and exact match. MMR achieves significant improvements overall baselines for the task of document-level role filler extraction. The performance improvement benefits from formulating DEE as a multi-turn MRC-based paradigm that can handle the challenge of arguments-scattering for DEE by capturing the long-distance dependencies between arguments explicitly. Compared with the SOTA method MGR which is a tagging based model, MMR improves 2.61, 3.93 F1 scores for head noun match and exact match respectively. It also proves that our MRC-based model MMR can provide external semantic information of roles and model the lengthy document directly, which benefits to the document-level event role filler extraction.

4.3 Results on ChFinAnn

For the ChFinAnn dataset, events and their arguments with specific role types are extracted by multi-turn question answers for different predicted event types (i.e., EF, ER, EU, EO, and EP).

Models	Head Noun Match			Exact Match		
	P	R	F1	P	R	F1
GLACIER (Patwardhan and Riloff, 2009)	47.80	57.20	52.08	-	-	-
TIER (Huang and Riloff, 2011)	50.80	61.40	55.60	-	-	-
Cohesion Extract (Huang and Riloff, 2012)	57.80	59.40	58.59	-	-	-
MGR (Du and Cardie, 2020a)	56.44	**62.77**	59.44	52.03	**56.81**	54.32
MMR	**62.34**	57.82	**60.45**	**58.76**	53.68	**56.10**

Table 1: Overall precision (P), recall (R) and F1 scores (F1) evaluated for document-level event role fillers extraction on the MUC-4 test set.

Models	EF			ER			EU			EO			EP			Avg.
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	F1
DCFEE-O	66.0	41.6	51.1	84.5	81.8	83.1	62.7	35.4	45.3	51.4	42.6	46.6	64.3	63.6	63.9	58.0
DCFEE-M	51.8	40.7	45.6	83.7	78.0	80.8	49.5	39.9	44.2	42.5	47.5	44.9	59.8	66.4	62.9	55.7
GreedyDec	79.5	46.8	58.9	83.3	74.9	78.9	68.7	40.8	51.2	69.7	40.6	51.3	85.7	48.7	62.1	60.5
Doc2EDAG	77.1	64.5	70.2	**91.3**	83.6	87.3	80.2	65.0	71.8	82.1	69.0	75.0	80.0	74.8	77.3	76.3
MMR-one	**81.2**	48.7	60.9	82.9	73.2	77.8	**81.2**	45.1	58.1	75.5	45.8	57.0	84.3	50.8	63.4	63.4
MMR	**78.4**	**65.5**	**71.3**	89.3	**88.1**	**88.7**	79.5	**66.4**	**72.4**	**83.5**	**71.4**	**76.9**	82.3	**74.1**	**78.0**	**77.4**

Table 2: Overall event-level precision (P), recall (R) and F1 scores (F-1) evaluated for document-level event extraction on the ChiFinAnn test set.

Models	EF		ER		EU		EO		EP		Avg.	
	S.	M.	S.	M.	S.	M.	S.	M.	S.	M.	S.	M.
DCFEE-O	56.0	46.5	86.7	54.1	48.5	41.2	47.7	45.2	68.4	61.1	61.5	49.6
DCFEE-M	48.4	43.1	83.8	53.4	48.1	39.6	47.1	42.0	67.0	60.0	58.9	47.7
GreedyDec	75.9	40.8	81.7	49.8	62.2	34.6	65.7	29.4	88.5	42.3	74.8	39.4
Doc2EDAG	80.0	61.3	89.4	68.4	77.4	64.6	79.4	69.5	85.5	72.5	82.3	67.3
MMR-one	79.2	52.1	88.2	49.3	70.3	49.4	74.2	44.7	87.4	45.2	79.8	48.1
MMR	**81.2**	**61.8**	**89.8**	**70.1**	**77.9**	**65.4**	**80.8**	**71.7**	**86.2**	**72.6**	**83.2**	**68.3**

Table 3: F1 scores for all event types and the averaged ones (Avg.) on single-event (S.) and multi-event (M.) sets evaluated on the ChiFinAnn dataset

Baselines. **DCFEE** (Yang et al., 2018a) proposed a tagging-based model for SEE and a key-event detection model with an arguments-completion strategy for DEE. In the comparisons, there are two versions of DCFEE: **DCFEE-O** only extract one event, **DCFEE-M** extract multiple events from one document. **Doc2EDAG** (Zheng et al., 2019), which aims to directly generate event tables based on the recognized entities to conduct table-filling in the document. There is a simple baseline of Doc2EDAG, named **GreedyDec**, which only fills one event table entry greedily. To be fair, we also introduce a simple baseline of MMR, MMR-one, that only predicts one event by extracting argument greedily given the specific role query.

Main Results. Table 2 shows the comparison between our model and baseline methods on the ChFinAnn for each event type. MMR improves 1.0, 1.3, 0.9, 2.0, 1.1, 0.9 F1-score over the SOTA method Doc2EDAG on the event type EF, ER, EU, EO and EP respectively. Compared with the SOTA method Doc2EDAG which based on a two-stage process from sentence-level extraction to document-level fusion, MMR can model the lengthy document directly and capture the relationship between arguments explicitly for addressing the challenges of arguments-scattering which is beneficial to DEE. Additionally, as the baseline of our proposed method, MMR-one can achieve the best performance compared with DCFEE-O and GreedyDec while all of them only predict one event for a document, which also proves the effectiveness of our proposed multi-turn and multi-granularity reader.

Results on Multi-Event. Table 3 shows F1 scores for different scenarios: single-event (i.e., documents contain just one event record) and multi-event (i.e., documents contain multi-events). MMR still maintains the highest extraction performance

Model	EF	ER	EU	EO	EP	Avg.
MMR	73.5	87.4	74.4	75.8	78.4	77.9
-GlobalEnc	-2.1	-3.4	-1.7	-2.6	-3.2	-2.6
-SemanticRole	-5.1	-3.8	-4.3	-4.7	-3.6	-4.3
-EventRecord	-9.2	-12.8	-13.1	-17.5	-14.3	-13.4

Table 4: F1-score of ablation studies on DE-PPN variants for each event type and the averaged (Avg.).

for all cases. As the multi-events is extremely challenging, MMR improves 2.3 averaged F1-score over the Doc2EDAG. Results proved the effectiveness of our event recorder which can guide the extraction of the later arguments for an event and address the challenge of multi-events on DEE.

4.4 Ablation Studies

In this section, to verify the effectiveness of each component of MMR, we conduct ablation studies on the next variants by evaluated on the ChFinAnn dataset: 1) *-GlobalEnc*: removing the Transformer-based global encoder, which can support the document-aware information for decoding. 2) *-SemanticRole*: replacing the event role specific question with initial embedding for each role type. 3) *-EventRecord*: removing the event recorder which can guide the extraction of the corresponding later arguments for the current event. The results are shown in Table 4 and we can observe that: 1) the global encoder is of prime importance that enhances the document-aware representations for tokens in a document and contributes +2.6 F1-score on average; 2) the used of event role specific question achieves a better performance compared with the learnable embeddings for roles, because the nature language question can provide more semantic information about the event role types. 3) the event recorder is a very important component for arguments extraction with +13.4 F1-score improvement which indicates that the event recorder can modeling the dependencies between arguments.

4.5 Effect of Different Decoder Layers

To investigate the importance of the multi-granularity reader, we explore the effect of different layers of the global encoder. Specifically, the number of decoder layers is set to 0,1,2,3 and 4, where 0 means removing this decoder. 1) The effect of different event decoder layers are shown in the left of Figure 3, and our method can achieve the best average F1-score when the number of layers is set to be 4. We conjecture that more layers of the global allow for better integrating document-aware information.

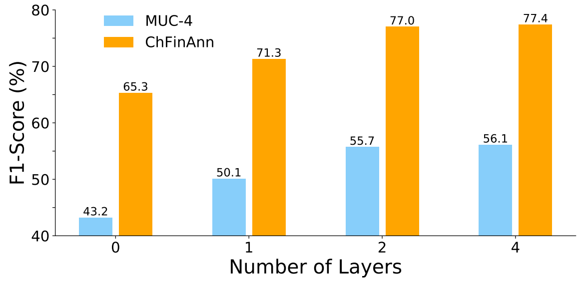

Figure 3: F1-score for performance differences global encoder layers.

Models	0%	25%	50%	100%
DCFEE	0	22.5	37.8	45.6
Doc2EDAG	0	47.3	66.4	71.1
MMR	**32.7**	**56.0**	**69.4**	**73.2**

Table 5: F1-score on different ratios of training data for a new event type.

4.6 Generalization Ability

To demonstrate the ability of transfer learning for new event type, we we conduct experiments on the ChFinAnn dataset where we keep one event type (EP) as the testing and the others (EF, ER, EU and EO) as the training set. Formally, we train our model using all training event type set to acquire prior knowledge. Then, we finetune the model using no or few samples of the test event type and evaluate the results in the remaining event type (the event type is preknown). Table 5 presents the results. We observe that MMR achieves a 32.7 F1-score without any data for event type EP, which benefits from the MRC-based formulation. Furthermore, with the samples increasing for fine-tuning, MMR can a better performance. Results illustrate the effectiveness of our model to handle new event types with only a few samples.

5 Conclusion

In this paper, we propose a multi-turn and multi-granularity reader for the task of DEE. Our model is based on a MRC formalization which comes with two key advantages: first, the question can provide semantic information for event roles. Second, the MRC-based method can model document-level information directly. Moreover, we introduce a multi-turn and multi-granularity reader to model the dependencies between arguments and capture hierarchical nature of document. The experimental results show that our proposed method obtains SOTA results on the MUC-4 and the ChiFinAnn datasets, which indicates the effectiveness and generalization of our method.

References

Jonathan Berant, Vivek Srikumar, Pei-Chun Chen, Abby Vander Linden, Brittany Harding, Brad Huang, Peter Clark, and Christopher D Manning. 2014. Modeling biological processes for reading comprehension. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1499–1510.

Jari Björne and Tapio Salakoski. 2018. Biomedical event extraction using convolutional neural networks and dependency parsing. In *Proceedings of the BioNLP 2018 workshop*, pages 98–108, Melbourne, Australia. Association for Computational Linguistics.

Yee Seng Chan, Joshua Fasching, Haoling Qiu, and Bonan Min. 2019. Rapid customization for event extraction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 31–36, Florence, Italy. Association for Computational Linguistics.

Pei Chen, Hang Yang, Kang Liu, Ruihong Huang, Yubo Chen, Taifeng Wang, and Jun Zhao. 2020. Reconstructing event regions for event extraction via graph attention networks. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 811–820, Suzhou, China. Association for Computational Linguistics.

Yubo Chen, Liheng Xu, Kang Liu, Daojian Zeng, and Jun Zhao. 2015. Event extraction via dynamic multi-pooling convolutional neural networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 167–176.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

George R Doddington, Alexis Mitchell, Mark A Przybocki, Lance A Ramshaw, Stephanie M Strassel, and Ralph M Weischedel. 2004. The automatic content extraction (ace) program-tasks, data, and evaluation. In *Lrec*, volume 2, page 1. Lisbon.

Xinya Du and Claire Cardie. 2020a. Document-level event role filler extraction using multi-granularity contextualized encoding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8010–8020, Online. Association for Computational Linguistics.

Xinya Du and Claire Cardie. 2020b. 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, Alexander Rush, and Claire Cardie. 2020. Document-level event-based extraction using generative template-filling transformers. *arXiv preprint arXiv:2008.09249*.

Ruihong Huang and Ellen Riloff. 2011. Peeling back the layers: detecting event role fillers in secondary contexts. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 1137–1147. Association for Computational Linguistics.

Ruihong Huang and Ellen Riloff. 2012. Bootstrapped training of event extraction classifiers. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 286–295. Association for Computational Linguistics.

Heng Ji and Ralph Grishman. 2011. Knowledge base population: Successful approaches and challenges. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pages 1148–1158.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. RACE: Large-scale ReAding comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 785–794, Copenhagen, Denmark. Association for Computational Linguistics.

Omer Levy, Minjoon Seo, Eunsol Choi, and Luke Zettlemoyer. 2017. Zero-shot relation extraction via reading comprehension. In *Proceedings of the 21st Conference on Computational Natural Language*

Learning (CoNLL 2017), pages 333–342, Vancouver, Canada. Association for Computational Linguistics.

Fayuan Li, Weihua Peng, Yuguang Chen, Quan Wang, Lu Pan, Yajuan Lyu, and Yong Zhu. 2020a. Event extraction as multi-turn question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 829–838, Online. Association for Computational Linguistics.

Manling Li, Qi Zeng, Ying Lin, Kyunghyun Cho, Heng Ji, Jonathan May, Nathanael Chambers, and Clare Voss. 2020b. Connecting the dots: Event graph schema induction with path language modeling. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 684–695, Online. Association for Computational Linguistics.

Xiaoya Li, Jingrong Feng, Yuxian Meng, Qinghong Han, Fei Wu, and Jiwei Li. 2020c. A unified MRC framework for named entity recognition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5849–5859, Online. Association for Computational Linguistics.

Xiaoya Li, Fan Yin, Zijun Sun, Xiayu Li, Arianna Yuan, Duo Chai, Mingxin Zhou, and Jiwei Li. 2019. Entity-relation extraction as multi-turn question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1340–1350, Florence, Italy. Association for Computational Linguistics.

Jian Liu, Yubo Chen, Kang Liu, Wei Bi, and Xiaojiang Liu. 2020. Event extraction as machine reading comprehension. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1641–1651, Online. Association for Computational Linguistics.

MUC-4. 1992. Fourth Message Uunderstanding Conference (MUC-4). In *In Proceedings of FOURTH MESSAGE UNDERSTANDING CONFERENCE (MUC4)*, McLean, Virginia.

Thien Huu Nguyen, Kyunghyun Cho, and Ralph Grishman. 2016. Joint event extraction via recurrent neural networks. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 300–309.

Siddharth Patwardhan and Ellen Riloff. 2009. A unified model of phrasal and sentential evidence for information extraction. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*, pages 151–160. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2016. Bidirectional attention flow for machine comprehension. *arXiv preprint arXiv:1611.01603*.

Yelong Shen, Po-Sen Huang, Jianfeng Gao, and Weizhu Chen. 2017. Reasonet: Learning to stop reading in machine comprehension. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1047–1055.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Shuohang Wang and Jing Jiang. 2016. Machine comprehension using match-lstm and answer pointer. *arXiv preprint arXiv:1608.07905*.

Xiaozhi Wang, Ziqi Wang, Xu Han, Zhiyuan Liu, Juanzi Li, Peng Li, Maosong Sun, Jie Zhou, and Xiang Ren. 2019. Hmeae: Hierarchical modular event argument 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 5781–5787.

Hang Yang, Yubo Chen, Kang Liu, Yang Xiao, and Jun Zhao. 2018a. 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.

Sen Yang, Dawei Feng, Linbo Qiao, Zhigang Kan, and Dongsheng Li. 2019. Exploring pre-trained language models for event extraction and generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5284–5294.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018b. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Bo Zheng, Haoyang Wen, Yaobo Liang, Nan Duan, Wanxiang Che, Daxin Jiang, Ming Zhou, and Ting Liu. 2020. Document modeling with graph attention networks for multi-grained machine reading comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6708–6718, Online. Association for Computational Linguistics.

Shun Zheng, Wei Cao, Wei Xu, and Jiang Bian. 2019. Doc2edag: An end-to-end document-level framework for chinese financial event extraction. *arXiv preprint arXiv:1904.07535*.