ISSN 1330-3651 (Print), ISSN 1848-6339 (Online)
https://doi.org/10.17559/TV-20250123002290
Original scientific paper

Anaphor-Aware Document-level Entity and Relation Joint Extraction with Curriculum Learning

Shunheng QI, Jiang ZHONG*, Kaiwen WEI*, Rongzhen LI, Hong YIN

Abstract: Document-level entity relation joint extraction aims at identifying semantic relationships between different entities within data, while assuming the entities are uncalibrated. Existing methods typically focus on enhancing the interaction between Coreference resolution (*COREF*) and Relation Extraction (*RE*) without adequately leveraging the anaphors present in the text which can provide clues for *COREF* and *RE*. Another challenge in this task is that the determination of the relationship between two entities requires logical reasoning through other entities. To alleviate these above challenges, we propose Anaphor-aware Entity and Relation Joint Extractor (AERJE), in which explicit anaphoric information is utilized in the two-stage model. We also adopt curriculum learning approach during the training process to bridge the gap between the model and external tool. Additionally, to enhance the model's logical reasoning ability at the mention level, we introduce a mention-level pairwise aggregation mechanism to allow the model to concentrate on information specific to mentions and anaphors pairs. Extensive experiments on the DocRED and Re-DocRED datasets demonstrate AERJE outperforms many former state-of-the-art methods. Our code is available now at https://github.com/PurRigiN/AERJE.

Keywords: curriculum learning; document-level entity and relation joint extraction; graph convolutional network

1 INTRODUCTION

Document-level RE aims at identifying semantic relationships between different entities within a text and representing them as triples. This task is of significant importance to various domains such as knowledge graph construction [1], information retrieval [2-4], recommendation [5, 6] and question-answering systems [7], [8]. Compared to sentence-level RE, document-level RE poses a greater challenge. For instance, in DocRED dataset, at least 40% of entity relationship facts can only be obtained by integrating information from multiple sentences [9].

Document-level entity relation joint extraction is to bridge the gap between theory and practical application where entities are usually not annotated. To complete the entity and relation joint extraction, one category of methods exhibits higher efficiency in which the task is transformed into a sequence-to-sequence task [10, 11]. Although this type of method is efficient, its accuracy is not competitive. Alternative approaches [12-14] are pipeline-based methods, in which named entity recognition, coreference resolution, and relation extraction are naturally performed step by step. Early method JEREX [12] introduces multi-instance learning to sequentially complete mention recognition, mention clustering, and relation classification. Subsequent study [13] argues that considering *COREF* and *RE* tasks separately overlooks the potential dependencies between two tasks, and then proposes Graph Compatibility to introduce explicit interaction between *COREF* and *RE*. Zhang et al. [14] believe that the aforementioned methods only consider unidirectional information flow and propose TAG for more comprehensive interaction.

However, the aforementioned methods do not adequately take into account the coreference information from anaphors, which is beneficial to gaining interactive information across sentences. Although some methods introduce coreference information to enhance the performance of relation extraction, some of which do so only in non-joint extraction RE tasks [15, 16], or fail to fully utilize the coreference information [15].

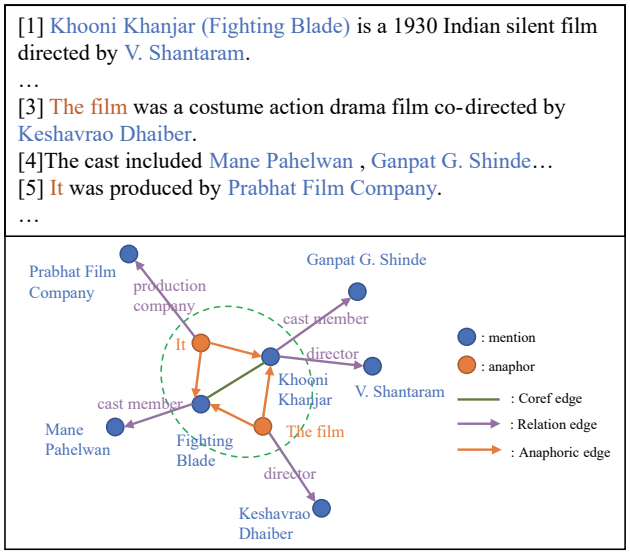


Figure 1 An example from Re-DocRED dataset. The blue and orange dots within the green dashed circle indicate that they belong to the same entity

Example from Re-DocRED dataset presented in Fig. 1 includes the original sentence and the relational triples depicted in graphical form. It is evident that anaphoric information is essential in determining relationships. Specifically, anaphor "The film" in sentence 3 and anaphor "It" in sentence 5 both refer back to the mention "Khooni Khanjar" and its alias "Fighting Blade". In sentence 3, it is indicated that "The film" has co-directors including "Keshavrao Dhaiber". From sentence 5, it can be inferred that "production company" of "It" is "Prabhat Film Company". By integrating sentences 1, 3 and 5, we can deduce the triples (Khooni Khanjar, production company, Prabhat Film Company) and (Khooni Khanjar, director, Keshavrao Dhaiber) through the two anaphors "It" and "The film". Additionally, it is necessary to discern the implicit triples (Khooni Khanjar, cast member, Ganpat G. Shinde) and (Khooni Khanjar, cast member, Mane Pahelwan) across sentences, which also represents a challenge in document-level RE.

To capture the coreference feature implied between mentions and anaphors for relation extraction, we propose

a novel Anaphor-aware Entity and Relation Joint Extractor (AERJE), in which supernumerary and explicit coreference information is incorporated in a two-stage entity and relation joint extraction model. We first let AERJE predict preliminary coreference scores and relation scores to construct coarse-level coreference and relation graphs. Subsequently, we integrate the anaphoric information generated by external tools to form the new heterogeneous coreference graph and relation graph. To mitigate the gap between AERJE and the external tool, we adopt curriculum learning into Relational Graph Convolutional Network (R-GCN). In document-level relation extraction, the relationship between two entities often requires logical inference through other entities to some extent. To enhance the model's capability in this regard, we propose mention-level pair-wise aggregation mechanism in classification to enable model adaptively focus on information specific to the current mentions or anaphors pair. Extensive experiments show AERJE achieves state-of-the-art results on DocRED and Re-DocRED datasets.

Our contributions are summarized as follows:

1. In order to adequately leverage the anaphoric information, we explicitly introduce external anaphoric information to construct new heterogeneous graphs for information propagation and adopting curriculum learning to smooth the incompatibility between model and external tool.
2. To enhance the model's ability to logically reason upon multiple mentions, we propose mention-level pair-wise aggregation mechanism, enabling the model to adaptively focus on information specific to the current mention or anaphor pair.
3. Extensive experiments demonstrate that proposed method is effective and the performance of AERJE achieves new state-of-the-art results on DocRED and Re-DocRED datasets.

2 RELATED WORK

In this section, two domains related to the work presented in this paper will be introduced, namely document-level entity and relation joint extraction, and curriculum learning.

2.1 Document-Level Entity Relation Joint Extraction

Different from normal document-level RE, entities are often not pre-annotated in practical applications, necessitating the completion of both entity recognition and relation extraction tasks. There are two main approaches to this challenge. The first are pipeline-style methods, and the second kind of methods transform the tasks into a seq2seq task. In the context of the former, JEREX [12] sequentially accomplishes mention recognition, mention clustering, and relation classification, and incorporates multi-instance learning into the model. Xu and Choi et al. [13] utilizes the same representation layer and the proposed Graph Compatibility to introduce explicit interaction between coreference resolution and relation extraction. The aforementioned methods suffer from cascading errors and exposure bias; hence some approaches transform the joint extraction task into a seq2seq task. Zeng et al. [17] introduced a copy mechanism in seq2seq model to address the issue of entity overlap in sentence-level relation extraction. CopyMTL [11] incorporates multi-task learning to resolve the limitation of previous seq2seq models, which could only identify single-token entities. Giorgi et al. [10] proposed a novel relation-to-sequence transformation method, extending the seq2seq approach from the sentence-level to document-level. But these methods do not take fully consideration of the coreference information brought by anaphors.

2.2 Curriculum Learning

Inspired by the principles underlying human cognitive processes, which involve learning simple concepts before gradually transitioning to more complex ones, curriculum learning [18] has been suggested as a training approach that introduces machine learning models to simpler samples initially, gradually advancing to more complex ones. Previous research [19-21] has indicated that curriculum learning can guide models toward more favorable parameter spaces and enhance generalization capabilities. Motivated by these findings, scholars have applied curriculum learning in a broad spectrum of domains, including computer vision [22-25], graph classification [26], event extraction [27], node classification [28], etc. For instance, in computer vision, Mousavi et al. [24] employ the entropy-alpha target decomposition method to assess the complexity level of each PolSAR image patch before feeding it into the neural network. Furthermore, curriculum learning has also been introduced in Natural Language Processing (NLP). Wei et al. [27, 29] introduces curriculum learning to smooth the incompatibility between training and real-world scenarios in their knowledge distillation strategy. For graph classification, CurGraph determines the difficulty scores of graphs by analyzing the intra-class and inter-class distributions of their embeddings and exposes a Graph Neural Network (GNN) to easy graphs before gradually moving on to hard ones [26]. CLNode [28] applies curriculum learning to node classification by using a selective training strategy that prioritizes nodes based on their quality, thereby mitigating the effects of mislabeled nodes.

3 PROBLEM DEFINITION

Considering a document $D = \{S_1, S_2, \ldots, S_l\}$ that consists of l sentences, our objective is to perform end-to-end identification of entities and to ascertain the relationships between these entities. In joint extraction, the entity mentions are unspecified. Therefore, before relation extraction, it is necessary to conduct mention detection and coreference resolution subtasks.

Mention extraction necessitates the identification of references with specific meanings as mentions $M = \{m_i\}_{i=1}^{\mathcal{M}}$ within document D, excluding anaphors. $\mathcal{M}$ denotes the number of mentions presented in one document.

Coreference resolution involves categorizing the extracted mentions $M = \{m_i\}_{i=1}^{\mathcal{M}}$ into clusters, thus representing them as a single entity, denoted by

$e = \{m_i\}_{i=1}^{N_e}$, which consists of one or more mentions m_i. $\xi_D = \{e_i\}_{i=1}^{E}$ represents the number of entities that appear in one document.

Relation extraction, which is performed on the identified entity clusters, involves predicting a subset of relationships $f(e_h, e_t) \subseteq R\mathcal{R}$ between pairs of entities $(e_h, e_t)_{h,\, t=1,\, \ldots,E;\, h \neq t}$, where the relationship set $\mathcal{R}$ is predefined.

4 METHODOLOGY

Fig. 2 shows the architecture of AERJE. To leverage clues provided by anaphors, the open-source tool NeuralCoref (https://github.com/huggingface/neuralcoref) is introduced. Firstly, after mention extraction and anaphor extraction, AERJE predicts preliminary coreference and relation scores to construct coarse-level coreference and relation graph. Subsequently, we integrate the anaphoric information generated by external tools to form the new heterogeneous coreference graph and relation graph. Combining the syntactic graph, we employ R-GCN on these three graphs, propagating coreference and relation information to enhance the interaction between *COREF* and RE tasks. Finally, we utilize the new mention embeddings to conduct fine-grained coreference resolution and relation extraction predictions. In the training stage, curriculum learning is adopted to bridge the gap between AERJE and NeuralCoref. In classification, mention-level pair-wise aggregation mechanism enables model to adaptively focus on information specific to the current mention or anaphor pair. The following sections will provide a detailed introduction to each component of the model.

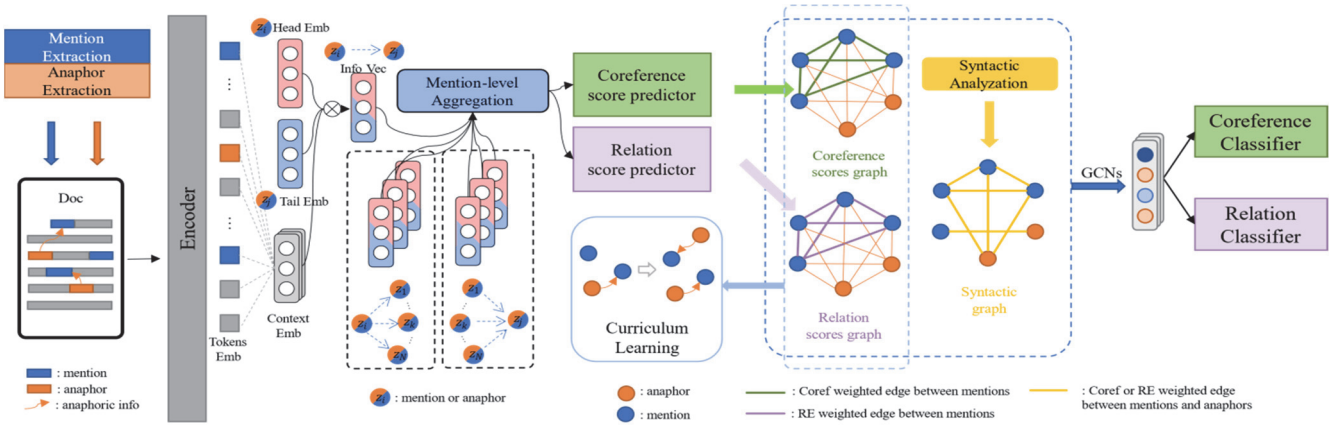


Figure 2 The architecture of our AERJE. AERJE firstly predicts coarse-level coreference and relation scores to construct graphs. Then it performs coreference and relation classification using new embeddings of mention or anaphor after graph convolution propagation

4.1 Mention Extraction and Anaphor Extraction

Before relation extraction is performed, we use BIO-based approach to extract the entity mentions. Building on the work of Devlin et al. [30], a pre-trained language model (*PLM*) is employed to convert the words into vector representation:

$$H = [h_1, h_2, \ldots, h_L]^{\top} = PLM([x_1, x_2, \ldots, x_L]) \quad (1)$$

where L means the number of tokens in a document. Furthermore, a simple classifier is employed to predict whether it is beginning word, inside word, or other word for each token. The model parameters are optimized by minimizing the cross-entropy loss function.

In order to leverage the anaphoric information within a text, we utilize the NeuralCoref open-source tool based on Spacy to identify pronouns and their corresponding anaphoric relationships with confidence scores. However, this tool identifies nouns as single tokens, whereas mentions can consist of multiple tokens. Naturally, an anaphor refers back to a mention if the noun token referred to by the anaphor is included within that mention. In such cases, an anaphor can be considered a special kind of mention, denoted by the symbol a. Consequently, the representation of an entity becomes $e = \{m_i\}_{i=1}^{N_e^m} \cup \{a_j\}_{j=1}^{N_e^a}$.

From the preceding discussion [13], refining the learning of relationships between mentions can enhance accuracy. Therefore, we need to integrate the anaphoric information extracted by the tool to construct labels at the mention level. Specifically, for entities e_s and e_o, if there exist relational facts $r_{s,o}$ between them, then for any mention or anaphor z_i in e_s, z_j in e_o, z_i and z_j express relation $r_{s,o}$, that is, $\forall z_i \in e_s, z_j \in e_o, r_{s,o} = f(z_i, z_j)$.

4.2 Curriculum Learning

Due to limitations in the external tool NeuralCoref itself, combined with the fact that model and NeuralCoref employ different datasets, the use of external tools for extracting mentions or anaphors pairs will inevitably introduce noise. Therefore, this study introduces curriculum learning to mitigate the impact of such noise, bridging the gap between AERJE and NeuralCoref.

Fig. 3 shows curriculum learning in AERJE. In previous studies, curriculum learning primarily encompasses two components: a difficulty measurer and a training scheduler [31]. For the former, the confidence scores predicted by NeuralCoref between mention-anaphor

pairs naturally align with the inherent difficulty of learning. Consequently, we employ such scores as a metric for gauging the difficulty in curriculum learning.

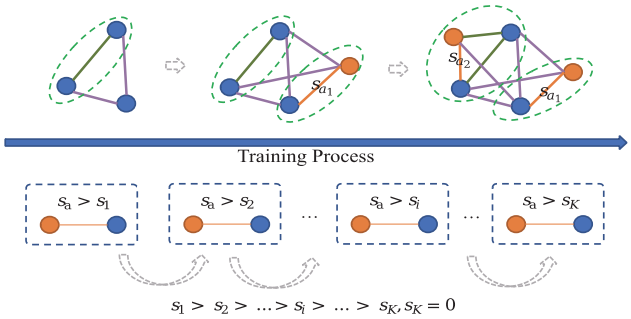


Figure 3 Curriculum learning process of gradually importing anaphors to construct label during training. The blue and orange dots within a green dashed circle indicate that they belong to the same entity

We extract such scores from NeuralCoref and arrange them in descending order of confidence, noted as (m, a, s_a), s_{ai} is the confidence scores between m and a_i. For the training scheduler, we then partition mention-anaphor pairs into K segments:

$$\left\{\left\{(m,a,s_a)\middle|s_a\right\rangle s_1\right\},\ \left\{(m,a,s_a)\middle|s_a\right\rangle s_2\right\},\ \ldots,\ \left\{(m,a,s_a)\middle|s_a\right\rangle s_K\right\}\Bigg\} \tag{2}$$

where $s_1 > s_2 > \ldots, > s_K, s_K = 0$. Similarly, we divide the training epochs into the same K phases. During the i th phase of training, we introduce anaphors involved in the set $\left\{(m,a,s_a)\middle|s_a\right\rangle s_i\right\}$ to construct coreference and relation graph which serve as supervisory signal to train R-GCN. Clearly, at the beginning stages of training, the value of s_i is relatively high, indicating that the model needs to learn simpler anaphoric information. As the training progresses, s_i decreases signifying that the anaphoric information the model needs to learn is becoming more complex.

4.3 Mention-Level Pair-Wise Aggregation Mechanism

To enable the model to capture context-specific information between mention or anaphor pairs during the inference phase, we adopt the localized context pooling method [32], utilizing a *PLM* to obtain the embeddings for each token as well as its multi-head attention to other tokens from last transformer layer. Then using $A_k^i, A_k^j \in R^L$ to represent the attention which z_i and z_j pay to other tokens in the k th attention head, we can derive the context vector $c^{(i,j)}$ specific to z_i, z_j:

$$\begin{aligned}&(H, A) = PLM\left([x_1, x_2, \ldots, x_L]\right),\\ &q^{(i,j)} = \sum_{k=1}^{H} A_k^i \cdot A_k^j,\ c^{(i,j)} = H^\top q^{(i,j)} \Big/ 1^\top q^{(i,j)}\end{aligned} \tag{3}$$

in which $A \in R^{H\times L\times L}$ is multi-head attention between tokens. $q^{(i,j)}$ reveals the composite attention from z_i, z_j pair towards the other tokens.

Subsequently, by integrating the context pooling vectors, we can obtain a new vector representation for the mention or anaphor n_i, n_j, which are then integrated to yield an info-vector used for classification:

$$\begin{aligned}&p^{(i,j)} = W_p vec\left(n_i n_j^\top\right),\\ &n_i = \tan h\left(W_h z_i + W_{c_h} c^{(i,j)}\right),\\ &n_j = \tan h\left(W_t z_j + W_{c_t} c^{(i,j)}\right)\end{aligned} \tag{4}$$

where $W_h, W_t, W_{c_h}, W_{c_t} \in \mathcal{R}^{d\times d}$ are learnable model parameters. $W_p \in \mathcal{R}^{d^2}$ are model parameters and $vec(\cdot)$ is the operation flattening matrix to vector.

In document-level RE, the relationship between two entities often requires some degree of logical inference through other entities. This is also the case at the mention level. For instance, when predicting the relationship between mentions or anaphors z_i, z_j naturally, ascertaining the relationships between z_i and other mentions, as well as between those other mentions and z_j, would provide the model with additional information, thereby enhancing the accuracy of relationship judgment from z_i to z_j. To bolster the model's capabilities in this regard, we propose mention-level pair-wise aggregation mechanism, enabling the model to adaptively focus on relevant mentions or anaphors during prediction. Prior to predicting coreference scores or relation scores, we utilize this mechanism to augment the mention-level info-vector representation as $\hat{p}^{(i,j)}$:

$$\begin{aligned}\hat{p}^{(i,j)} = p^{(i,j)} + \text{softmax}\left(\left(q^{(i,j)}\right)^\top K^{(i)}\right)V^{(i)} +\\ +\text{softmax}\left(\left(q^{(i,j)}\right)^\top K^{(j)}\right)V^{(j)}\end{aligned} \tag{5}$$

$$\begin{aligned}&q^{(i,j)} = W_Q p^{(i,j)},\ K^{(j)} = W_K P^{(j)},\ K^{(i)} = W_K P^{(i)},\\ &V^{(i)} = W_V P^{(i)},\ V^{(j)} = W_V P^{(j)}\end{aligned}$$

where W_Q, W_K, W_V are learnable weight matrices, and $P^{(i)} = \left[p^{(i,1)}, \ldots, p^{(i,N_e)}\right]^\top$ means stacking info-vectors representing relationships from z_i to other mentions or anaphors. $P^{(j)}$ is the same. Utilizing the enhanced info-vectors, we predict co-reference scores or relation scores using a two-layer linear classification network:

$$s^{(i,j)} = \text{sigmod}\left(W_2 ReLU\left(W_1 \hat{p}^{(i,j)}\right) + W_3\left(z_i \oplus z_j\right) + b\right) \tag{6}$$

where $W_1 \in \mathcal{R}^{d\times d}$, $W_2 \in \mathcal{R}^{1\times d}$, $w_3 \in \mathcal{R}^{1\times 2d}$, $b \in \mathcal{R}^d$ are learnable parameters. It should be noted that coreference and relation scores prediction share the same unaugmented vectors n_i, n_j but the subsequent fusion enhancement is independent of the two. Ultimately, we are able to obtain

the co-reference scores $S_c^{(i,j)}$ and relation scores $S_r^{(i,j)}$ for pairs of mentions or anaphors within an article.

4.4 Latent Graph Construction

Based on the predictions from the previous modules, we naturally construct two directed weighted graphs at the mention-level for coreference scores and relation scores. After normalization, we can obtain the coreference graph and the relation graph where nodes represent mentions or anaphors, and the edges are normalized score values: $G_c = \text{Softmax}(S_c)$, $G_r = \text{Softmax}(S_r)$.

Syntactic information in document plays a crucial role in relation identification. Incorporating syntactic dependency trees helps analyzing syntactic information and leading to improved representations. However, this approach is typically employed in sentence-level RE, whereas document-level RE focuses more on inter-sentential information. Therefore, in this work, similar to the baseline, we establish bidirectional links between mentions or anaphors that co-occur within the same sentence to construct a syntactic graph. The contextual information of each mention or anaphor can be conveyed through such edges.

4.5 Propagating with R-GCN

To enhance the interaction between coreference resolution and relation classification tasks, while leveraging inter-sentential syntactic information, we employ relational graph convolutional neural networks for information propagation. Similar to the baseline, we perform attention graph convolution operations on the coreference graph, relational graph, and syntactic graph, respectively, ultimately obtaining new mention-level node representations denoted as $\widehat{z_i}$:

$$z_i^{(l+1)} = \tan h\left(\sum_{t\in\{c,r,s\}}\sum_{j=1}^{N_e} g_t^{(i,j)} W_t^l z_j^l + b_t^l\right) \tag{7}$$

where t represents the type of graph, $g_t^{(i,j)}$ means the weighted value from mention or anaphor z_i to z_j in G_t. W_t^l, b_t^l are learnable parameters in l th layer. The initial node embedding z_i^0 is z_i.

4.6 Training and Inference

We employ a multi-task learning approach, utilizing a joint loss function to train the various modules. For the coarse-level coreference score prediction, relation score prediction, and fine-grained coreference score prediction, we optimize using binary cross-entropy loss, denoted as $\mathcal{L}_{cc}$, $\mathcal{L}_{cr}$, $\mathcal{L}_{fc}$ respectively.

For the fine-grained relation classification, we employ the adaptive thresholding loss [32] for optimization, denoted as $\mathcal{L}_{fr}$:

$$\begin{aligned}\mathcal{L}_{fr} = &-\sum_{r\in\mathcal{P}_\mathcal{T}} \log\left(\frac{\exp(\text{log}it_r)}{\sum_{r'\in\mathcal{P}_\mathcal{T}\cup\{TH\}}\exp(\text{log}it_{r'})}\right) - \\ &-\log\left(\frac{\exp(\text{log}it_{TH})}{\sum_{r'\in\mathcal{N}_\mathcal{T}\cup\{TH\}}\exp(\text{log}it_{r'})}\right)\end{aligned} \tag{8}$$

where positive classes $\mathcal{P}_\mathcal{T} \subseteq \mathcal{R}$ represent the relations that exist between z_i, z_j and the negative classes $\mathcal{N}_\mathcal{T} = \mathcal{R} - \mathcal{P}_\mathcal{T}$. The complete optimization loss is as follows, where α is used to adjust the losses between two stages:

$$L = \mathcal{L}_{cc} + \mathcal{L}_{cr} + \alpha\left(\mathcal{L}_{fc} + \mathcal{L}_{fr}\right) \tag{9}$$

To conduct fine-grained coreference score prediction and relation classification, we employ the same module mentioned above, taking the new mention-level embeddings obtained after convolution as input, to derive the classification vector $\hat{p}_{fine}^{(i,j)}$ at this stage. Predict the coreference score and relation score using the following formula:

$$s_{fine}^{(i,j)} = \text{sigmod}\left(U_2\text{ReLU}\left(U_1\hat{p}_{fine}^{(i,j)}\right) + U_3\left(\widehat{z_i}\oplus\widehat{z_j}\right) + b\right) \tag{10}$$

where $U_1 \in \mathcal{R}^{d\times d}$, $U_2 \in \mathcal{R}^{n\times d}$, $U_3 \in \mathcal{R}^{n\times 2d}$, $b \in \mathcal{R}^d$. When predict coreference score, n = 1, while $n = |\mathcal{R}+1|$ when predict relation score. Similar to TAG, we introduce learnable dummy class TH to serve as a threshold for addressing the multi-label relation classification task [32]. And the same decoding strategy is used to acquire entity clusters and entity-level relations.

5 EXPERIMENTS

5.1 Datasets

DocRED [9] dataset is a large-scale document-level relation extraction dataset including 96 relation types, 132375 entities and 63427 relational facts and is constructed from Wikipedia and Wikidata. The authors have also released more than 100000 articles based on distantly supervised data, enabling research on weakly supervised relation extraction. Re-DocRED [33] is a revised version of the DocRED dataset, aiming at addressing the false negative issue present in the original dataset, as well as correcting certain logical inconsistencies and coreference errors. The dataset statistics for Re-DocRED indicate that on average, there are approximately 34.7 triples per article.

5.2 Evaluation Metrics

In addressing entity and relation joint extraction task, it is necessary to employ metrics to separately assess the three sub-tasks: mention extraction, coreference resolution, and relation extraction. Drawing from previous research [14] , for the mention extraction task, we utilize the $F1$ metric. For the task of coreference resolution, the averaged

$F1$ score of *MUC*, B^3, and $CEAF_{\phi_4}$ is employed. In the context of relation extraction, the entity-level $F1$ score is utilized, where a prediction is deemed correct only when both the entities and their associated mentions are correctly identified. Additionally, Ign $F1$ score is used, which disregard relationships that are already present in the training and validation sets as well as in the training and testing sets, mitigating evaluation bias that may arise from the model's potential memory of certain relationship facts.

5.3 Baselines

KB-IE, proposed by Verlinden et al. [34], is to incorporate external knowledge bases (Wikidata & Wikipedia) to assist in entity and relation joint extraction but with LSTM in use. Eberts and Ulges [12] propose JEREX, which sequentially accomplishes mention identification, entity clustering, and relation classification, and incorporates multi-instance learning into the model. Giorgi et al. [10] propose seq2rel, a novel transformation method from relations to sequences, extending the seq2seq approach from sentence level to document level, while also addressing the issue of discontinuous entity mentions. Xu and Choi [13] propose JointM + GPGC to enhance the interaction between coreference resolution and relation extraction tasks by introducing "Graph Compatibility" into their model. Furthermore, pipeline is standard pipeline method baseline and JointM only shares the same text encoder. Zhang et al.argue that previous methods only consider unidirectional interactions between tasks and propose TAG, aiming to enhance the bidirectional interactions between *COREF* and RE tasks [14].

5.4 Hyperparameters Settings

In the experiments, we trained the model using the AdamW optimizer. We used a learning rate of 1e−4 for the *PLM* and 3e−5 for the rest of the model. The model was implemented using the PyTorch framework, with the *PLM* invocation facilitated by the transformers library from HuggingFace (https://huggingface.co/). We maintained the hidden size consistent with the *PLM*, and due to machine limitations, the batch size was set to 4. Observing the convergence rate, we set the training epoch to 50, using an A6000 GPU. All experimental results were conducted with 3 random seeds and are reported as the average values.

5.5 Experimental Results and Analysis

5.5.5 Overall Comparison

Tab. 1 presents the performance evaluation results of AERJE compared to other selected baseline methods on the DocRED dataset. The results indicate that AERJE achieved the best performance compared to the other selected baseline models, demonstrating the effectiveness of our approach for joint entity relation extraction.

TAG employs three distinct text encoders to experiment with varying effects on the DocRED dataset. To validate the efficacy of our approach, we similarly conducted experiments utilizing these three encoders. The enhancements in the final RE $F1$ and Ign-$F1$ results demonstrate that our method outperforms state-of-the-art (SOTA) model. Although the use of roberta-base and roberta-large as encoders did not yield significant improvements in *COREF* $F1$, it should not be overlooked that our reproduced *ME* is slightly lower than SOTA, indicating the effectiveness of our method in the *COREF* task.

Table 1 Overall performance on DocRED dataset. Some of previous methods only conducted experiments on test split, while we report results on both dev/test set, respectively. Note that JEREX and seq2rel utilize different partitions of the DocRED dataset and their results are for reference only

Method	Encoder	*ME*	*COREF*	*RE*-F1	*RE*-Ign *F*1
KB-IE	LSTM	-	83.6	25.7	-
JEREX	BERT-base	92.99*	82.79*	40.38*	-
seq2rel	BERT-base	-	-	38.2*	-
Pipeline	SpanBERT-base	92.56	84.09	38.29	35.88
Joint	SpanBERT-base	93.34	84.79	38.94	36.64
JointM+GPGC	SpanBERT-base	93.35	84.96	40.62	38.28
TAG	BERT-base	92.89/93.56	84.75/85.07	40.65/41.87	38.27/39.82
AERJE	BERT-base	**92.90**/**93.61**	**85.15**/**85.37**	**41.95 ± 0.12**/**42.14**	**39.80 ± 0.11**/**39.88**
TAG	RoBERTa-base	92.95/**93.63**	85.67/86.03	42.28/43.16	40.28/41.13
AERJE	RoBERTa-base	**92.99**/93.26	**85.74**/**86.21**	**42.60 ± 0.08**/**43.64**	**40.50 ± 0.07**/**41.59**
TAG	RoBERTa-large	**93.32**/**93.84**	85.87/86.37	43.21/44.97	41.22/42.88
AERJE	RoBERTa-large	93.19/93.42	**85.95**/**86.40**	**43.64 ± 0.10**/**45.30**	**41.37 ± 0.08**/**43.12**

Table 2 Overall performance of AERJE on Re-DocRED dataset

Method	Encoder	*ME*	*COREF*	*RE*-*F*1	*RE*-Ign *F*1
TAG	RoBERTa-base	**93.42**/**92.91**	86.49/85.61	49.34/49.38	48.21/48.47
AERJE	RoBERTa-base	93.19/92.67	**86.50**/**85.65**	**50.37**/**50.07**	**49.28**/**49.10**

In our comparison with the SOTA model (TAG), it can be observed that the performance on the Re-DocRED dataset is more significant compared to the DocRED dataset. This improvement is likely due to the fact that Tan et al. [33] corrected some of coreference errors on the DocRED dataset, while our method supplements with more specific coreference information using external tools and incorporates it into the model. Those improvements are anticipated.

5.5.2 Ablation Studies

To investigate the contribution of various modules within our model to the performance of joint entity-relation extraction, we designed three variants of the original model and conducted experiments on both the DocRED and Re-DocRED datasets. The three variants are as Tab. 3.

Table 3 Ablation studies of AERJE on DocRED and Re-DocRED dataset

dataset	variants	*COREF*	*RE-F*1	*RE*-Ign *F*1
DocRED	AERJE-RoBERTa-base	**85.74/86.21**	**42.60/43.64**	**40.50/41.59**
	w/o Anaphor	85.71/86.05	42.15/43.29	40.06/41.25
	w/o Curriculum Learning	85.67/86.06	42.28/43.30	40.25/41.18
	w/o Pair-wise Mec	85.73/86.05	41.85/43.35	39.90/41.29
Re-DocRED	AERJE-RoBERTa-base	**86.50/85.65**	**50.37/50.07**	**49.28/49.10**
	w/o Anaphor	86.32/85.44	49.54/49.7	48.49/48.77
	w/o Curriculum Learning	86.43/85.33	49.80/49.54	48.80/48.59
	w/o Pair-wise Mec	86.29/85.33	49.58/49.12	48.55/48.18

w/o Anaphor. The module that introduces anaphoric information is removed, including curriculum learning; the model solely utilizes mention-level Pair-wise Aggregation Mechanism to enhance the focus on relevant mentions.

w/o Curriculum Learning. The model incorporates anaphoric information but does not perform curriculum learning, while also employing mention-level pair-wise aggregation mechanism.

w/o Pair-wise Mec. The model introduces anaphoric information and undergoes noise suppression processing, but does not utilize mention-level Pair-wise Aggregation Mechanism.

Without the incorporation of anaphoric information, there is a decline in the performance metrics for both *COREF* and *RE* tasks on both two datasets. Particularly on the Re-DocRED dataset, the *RE-F*1 score decreased by 0.83 on the validation set, and the *RE*-Ign *F*1 score decreased by 0.79. This decrease indicates that explicitly introducing specific anaphoric information can facilitate the model in more accurately judging the relationships between mentions. The decline in the *COREF F*1 metric similarly suggests that explicit anaphoric information can enhance coreference resolution.

Subsequently, we introduced anaphoric information without curriculum learning. Ideally, the model's performance should slightly decrease, but not to a level lower than the first model variant. The results in the table confirm this hypothesis. Specifically, on the validation split of the Re-DocRED dataset, the *RE-F*1 and *RE*-Ign *F*1 scores decreased by 0.57 and 0.48, respectively.

Finally, we retain the introduction of anaphoric information and perform curriculum learning, eliminating mention-level pair-wise aggregation mechanism. It can be observed that the model's performance declines on both datasets, indicating that adaptively focusing on other pairs is helpful when the model judges the relationships between two mentions or anaphors.

5.5.3 Effectiveness of Curriculum Learning

To verify and explore the effectiveness of curriculum learning, we conducted experiments on Re-DocRED, comparing the performance of *COREF* with different hyperparameter K as shown in Tab. 4. The experiments reveals that the model performs well when $K = 4$. When K increases from 3 to 4, the model's performance in both *COREF* and RE tasks improves, indicating that our introduced method can effectively suppress noise brought by external tools. However, when K is set to 5 and 6, the performance on the *COREF* task remained the same as when $K = 4$, but there is a decline in performance on the RE task. This may because *COREF* and RE have different acceptance capabilities for samples in progressive learning, and are thus affected differently by the hyperparameter K. This further indicates that the hyperparameter K in joint entity relation extraction can only serve as a modest refining parameter, and continuing to increase it may harm the model's performance.

Table 4 *F*1 scores of AERJE on Re-DocRED dataset on both dev/test split with different hyperparameter *K*

K	3	4	5	6
*COREF-F*1	86.43/85.33	**86.50/85.65**	86.50/85.60	86.35/85.58
*RE-F*1	49.88/49.60	**50.37/50.07**	49.90/49.77	49.78/49.55
RE-Ign *F*1	48.92/48.62	**49.28/49.10**	49.07/48.77	48.98/48.51

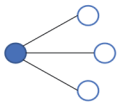
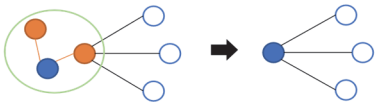
Document Text	Model	Schematic Diagram	Predicted Relation
[0] Ashwathy Kurup , better known by her stage name Parvathy, is an Indian film actress and classical dancer… … [4] Her notable works include Amrutham Gamaya, Thoovanathumbikal, Ponmuttayidunna Tharavu…	TAG		(Ashwathy, NA, Amrutham Gamaya) (Ashwathy, NA, Thoovanathumbikal) (Ashwathy, NA, Ponmuttayidunna Thravu)
	Our		(Ashwathy, cast member, Amrutham Gamaya) (Ashwathy, cast member, Thoovanathumbikal) (Ashwathy, cast member, Ponmuttayidunna Tharavu)
[0] Michael Imperioli (born March 26 , 1966) is an American actor , writer and director best known for… … [5] He wrote and directed his first feature film , The Hungry Ghosts , in 2008 .	TAG		(Michael Imperioli , NA, The Hungry Ghosts)
	Our		(Michael Imperioli , screenwriter, The Hungry Ghosts) (Michael Imperioli , director, The Hungry Ghosts)

Figure 4 Case Study from our AERJE and baseline model. Two cases are chosen from DocRED and Re-DocRED respectively. Schematic diagram is intended to aid intuitive understanding

For new datasets, to better extract anaphoric information, it is recommended to perform additional annotations and then train on the NeuralCoref. Additionally, we suggest conducting a statistical analysis of the new dataset (e.g., calculating data size, class distribution, and feature variance) and using this information to initialize the curriculum learning phase segmentation strategy. For instance, when the dataset exhibits severe class imbalance, the learning phase for simple samples can be extended to mitigate model bias.

5.5.4 Case Study

To demonstrate the effectiveness of our module, we conducted case study comparisons on DocRED and Re-DocRED datasets with the baseline and included a Schematic Diagram to intuitively display the relationships between mentions, anaphors, and entities. As shown in the first example, the baseline model failed to predict the relationship between "Ashwathy Kurup" and the other three works. Our model, however, by one of the introduced anaphors "Her" in sentence 4, explicitly connected "Ashwathy Kurup" with the three works, ultimately inferring the triples (Ashwathy, cast member, Amrutham Gamaya), (Ashwathy, cast member, Thoovanathumbikal), (Ashwathy, cast member, Ponmuttayidunna Tharavu). Similarly, in the second example, although mentions are far apart in the text, through the use of "He" in sentence 5, the model ultimately deduced that "Michael Imperioli" has both "screenwriter" and "director" relationships with the movie "The Hungry Ghosts".

6 CONCLUSION

To adequately leverage the anaphoric information in the text and enhance the model's logical reasoning capabilities in document-level entity and relation joint extraction, we introduce AERJE, an anaphor-aware two-stage graph neural network with curriculum learning and mention-level pair-wise aggregation mechanism. Specially, AERJE leverages the explicit anaphoric information between anaphors and mentions provided by external tool, and the mention-level pair-wise aggregation mechanism guides model to focus on the information specific to current mention and anaphor pair. The use of curriculum learning helps bridge the gap between AERJE and the external tool. Extensive experiments demonstrate that AERJE outperforms baseline models.

7 REFERENCES

[1] Yin, H., Zhong, J., Wang, C., Li, R., & Li, X. (2023). GS-InGAT: An interaction graph attention network with global semantic for knowledge graph completion. *Expert Systems with Applications*, *228*, 120380. https://doi.org/10.1016/j.eswa.2023.120380

[2] Siciliani, L., Taccardi, V., Basile, P., Di Ciano, M., & Lops, P. (2023). AI-based decision support system for public procurement. *Information Systems*, *119*, 102284. https://doi.org/10.1016/j.is.2023.102284

[3] Wei, K., Yang, Y., Jin, L., Sun, X., Zhang, Z., Zhang, J., Li, X., Zhang, L., Liu, J., & Zhi, G. (2023). Guide the Many-to-One Assignment: Open Information Extraction via IoU-aware Optimal Transport. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, *1: Long Papers*, 4971-4984. https://doi.org/10.18653/v1/2023.acl-long.272

[4] Tang, J., Xu, D., Cai, Q., Li, S., & Rezaeipanah, A. (2024). Towards a semi-supervised ensemble clustering framework with flexible weighting mechanism and constraints information. *Engineering Applications of Artificial Intelligence*, *136*, 108976. https://doi.org/10.1016/j.engappai.2024.108976

[5] Li, Z., Liu, F., Wei, Y., Cheng, Z., Nie, L., & Kankanhalli, M. (2024). Attribute-driven Disentangled Representation Learning for Multimodal Recommendation. *Proceedings of the 32nd ACM International Conference on Multimedia*, 9660-9669. https://doi.org/10.1145/3664647.3681148

[6] Elahi, E., Anwar, S., Al-kfairy, M., Rodrigues, J. J. P. C., Ngueilbaye, A., Halim, Z., & Waqas, M. (2025). Graph attention-based neural collaborative filtering for item-specific recommendation system using knowledge graph. *Expert Systems with Applications*, *266*, 126133. https://doi.org/10.1016/j.eswa.2024.126133

[7] Wei, K., Zhang, J., Zhang, H., Zhang, F., Zhang, D., Jin, L., & Yu, Y. (2024). *Chain-of-Specificity: An Iteratively Refining Method for Eliciting Knowledge from Large Language Models*.

[8] Li, Z., Guo, Y., Wang, K., Wei, Y., Nie, L., & Kankanhalli, M. (2023). Joint Answering and Explanation for Visual Commonsense Reasoning. *IEEE Transactions on Image Processing*, *32*, 3836-3846. https://doi.org/10.1109/TIP.2023.3286259

[9] Yao, Y., Ye, D., Li, P., Han, X., Lin, Y., Liu, Z., Liu, Z., Huang, L., Zhou, J., & Sun, M. (2019). *DocRED: A Large-Scale Document-Level Relation Extraction Dataset*. https://doi.org/10.18653/v1/P19-1074

[10] Giorgi, J., Bader, G., & Wang, B. (2022). A sequence-to-sequence approach for document-level relation extraction. *Proceedings of the 21st Workshop on Biomedical Language Processing*, 10-25. https://doi.org/10.18653/v1/2022.bionlp-1.2

[11] Zeng, D., Zhang, H., & Liu, Q. (2020). CopyMTL: Copy Mechanism for Joint Extraction of Entities and Relations with Multi-Task Learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, *34*(05). https://doi.org/10.1609/aaai.v34i05.6495

[12] Eberts, M. & Ulges, A. (2021). An End-to-end Model for Entity-level Relation Extraction using Multi-instance Learning. In P. Merlo, J. Tiedemann, & R. Tsarfaty (Eds.), *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 3650-3660. https://doi.org/10.18653/v1/2021.eacl-main.319

[13] Xu, L. & Choi, J. D. (2022). *Modeling Task Interactions in Document-Level Joint Entity and Relation Extraction*. https://doi.org/10.18653/v1/2022.naacl-main.395

[14] Zhang, R., Li, Y., & Zou, L. (2023). A Novel Table-to-Graph Generation Approach for Document-Level Joint Entity and Relation Extraction. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, *1: Long Papers)*, 10853-10865. https://doi.org/10.18653/v1/2023.acl-long.607

[15] Lu, C., Zhang, R., Sun, K., Kim, J., Zhang, C., & Mao, Y. (2023). *Anaphor Assisted Document-Level Relation Extraction*. https://doi.org/10.18653/v1/2023.emnlp-main.955

[16] Xue, Z., Zhong, J., Dai, Q., & Li, R. (2022). CorefDRE: Coref-Aware Document-Level Relation Extraction. *Knowledge Science, Engineering and Management*, 116-128. https://doi.org/10.1007/978-3-031-10989-8_10

[17] Zeng, X., Zeng, D., He, S., Liu, K., & Zhao, J. (2018). Extracting Relational Facts by an End-to-End Neural Model with Copy Mechanism. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, *1: Long Papers*, 506-514. https://doi.org/10.18653/v1/P18-1047

[18] Bengio, Y., Louradour, J., Collobert, R., & Weston, J. (2009). Curriculum learning. *Proceedings of the 26th Annual International Conference on Machine Learning*, 41-48. https://doi.org/10.1145/1553374.1553380

[19] Weinshall, D., Cohen, G., & Amir, D. (2018). Curriculum Learning by Transfer Learning: Theory and Experiments with Deep Networks. *Proceedings of the 35th International Conference on Machine Learning*, 5238-5246.

[20] Weinshall, D. & Amir, D. (2020). Theory of Curriculum Learning, with Convex Loss Functions. *Journal of Machine Learning Research*, *21*(222), 1-19.

[21] Wang, Q., Gu, Y., Yang, M., & Wang, C. (2021). Multi-attribute smooth graph convolutional network for multispectral points classification. *Science China Technological Sciences*, *64*(11), 2509-2522. https://doi.org/10.1007/s11431-020-1871-8

[22] Tang, J., Gong, Z., Tao, B., & Yin, Z. (2024). Advancing generalizations of multi-scale GAN via adversarial perturbation augmentations. *Knowledge-Based Systems*, *284*, 111260. https://doi.org/10.1016/j.knosys.2023.111260

[23] Zhang, B., Wang, Y., Hou, W., Wu, H., Wang, J., Okumura, M., & Shinozaki, T. (2021). FlexMatch: Boosting Semi-Supervised Learning with Curriculum Pseudo Labeling. *Advances in Neural Information Processing Systems*, *34*, 18408-18419.

[24] Mousavi, H., Imani, M., & Ghassemian, H. (2022). Deep Curriculum Learning for PolSAR Image Classification. *2022 International Conference on Machine Vision and Image Processing (MVIP)*, 1-5. https://doi.org/10.1109/MVIP53647.2022.9738781

[25] Wei, K., Du, R., Jin, L., Liu, J., Yin, J., Zhang, L., Liu, J., Liu, N., Zhang, J., & Guo, Z. (2024). Video Event Extraction with Multi-View Interaction Knowledge Distillation. *Proceedings of the AAAI Conference on Artificial Intelligence*, *38*(17), Article 17. https://doi.org/10.1609/aaai.v38i17.29891

[26] Wang, Y., Wang, W., Liang, Y., Cai, Y., & Hooi, B. (2021). CurGraph: Curriculum Learning for Graph Classification. *Proceedings of the Web Conference 202*1, 1238-1248. https://doi.org/10.1145/3442381.3450025

[27] Wei, K., Sun, X., Zhang, Z., Jin, L., Zhang, J., Lv, J., & Guo, Z. (2023). Implicit Event Argument Extraction With Argument-Argument Relational Knowledge. *IEEE Transactions on Knowledge and Data Engineering*, *35*(9), 8865-8879. https://doi.org/10.1109/TKDE.2022.3218830

[28] Wei, X., Gong, X., Zhan, Y., Du, B., Luo, Y., & Hu, W. (2023). CLNode: Curriculum Learning for Node Classification. *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, 670-678. https://doi.org/10.1145/3539597.3570385

[29] Wei, K., Sun, X., Zhang, Z., Zhang, J., Zhi, G., & Jin, L. (2021). Trigger is Not Sufficient: Exploiting Frame-aware Knowledge for Implicit Event Argument Extraction. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, *1*: *Long Papers,* 4672-4682. https://doi.org/10.18653/v1/2021.acl-long.360

[30] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *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)*, *1*, 4171-4186.

[31] Wang, X., Chen, Y., & Zhu, W. (2022). A Survey on Curriculum Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9), 4555-4576. https://doi.org/10.1109/TPAMI.2021.3069908

[32] Zhou, W., Huang, K., Ma, T., & Huang, J. (2021). Document-Level Relation Extraction with Adaptive Thresholding and Localized Context Pooling. *Proceedings of the AAAI Conference on Artificial Intelligence*, *35*(16), Article 16. https://doi.org/10.1609/aaai.v35i16.17717

[33] Tan, Q., Xu, L., Bing, L., Ng, H. T., & Aljunied, S. M. (2022). Revisiting DocRED - Addressing the False Negative Problem in Relation Extraction. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 8472- 8487. https://doi.org/10.18653/v1/2022.emnlp-main.580

[34] Verlinden, S., Zaporojets, K., Deleu, J., Demeester, T., & Develder, C. (2021). Injecting Knowledge Base Information into End-to-End Joint Entity and Relation Extraction and Coreference Resolution. *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 1952-1957. https://doi.org/10.18653/v1/2021.findings-acl.171

Contact information:

Shunheng QI
College of Computer Science, Chongqing University,
Chongqing 401331, China
E-mail: qishunheng@stu.cqu.edu.cn

Jiang ZHONG
(Corresponding author)
College of Computer Science, Chongqing University,
Chongqing 401331, China
E-mail: zhongjiang@cqu.edu.cn

Kaiwen WEI
(Corresponding author)
College of Computer Science, Chongqing University,
Chongqing 401331, China
E-mail: weikaiwen@cqu.edu.cn

Rongzhen LI
College of Computer Science, Chongqing University,
Chongqing 401331, China
E-mail: lirongzhen@cqu.edu.cn

Hong YIN
College of Computer Science, Chongqing University,
Chongqing 401331, China
E-mail: yinhong@cqu.edu.cn