Implicit Event Argument Extraction With Argument-Argument Relational Knowledge

Kaiwen Wei, Xian Sun, *Senior Member, IEEE*, Zequn Zhang, Li Jin, Jingyuan Zhang, Jianwei Lv, and Zhi Guo

Abstract—As a challenging sub-task of event argument extraction, implicit event argument extraction seeks to identify document-level arguments that play direct or implicit roles in a given event. Prior work mainly focuses on capturing direct relations between arguments and the event trigger; however, the lack of reasoning ability imposes limitations to the extraction of implicit arguments. In this work, we propose an **A**rgument-argument **R**elation-enhanced **E**vent **A**rgument extraction (AREA) learning framework to tackle this issue through reasoning in event frame-level scope. The proposed method leverages related arguments of the expected one as clues, and utilizes such argument-argument dependencies to guide the reasoning process. To bridge the distribution gap between oracle knowledge used in the training phase and the imperfect related arguments in the test stage, we introduce a conventional knowledge distillation strategy to drive a final model that can work without extra inputs by mimicking the behaviour of a well-informed teacher model. In addition, considering that conventional knowledge distillation methods transfer knowledge individually, we integrate it with a novel relational knowledge distillation mechanism to explicitly capture the structural mutual argument-argument relation. Moreover, since the training process is not compatible with the real situation, a curriculum learning method is further introduced to make the training process smoother. Experimental results demonstrate that the learning framework obtains state-of-the-art performance on the RAMS and Wikievents datasets. Ablation study and further discussion also show it could handle long-range dependency and implicit argument problems effectively.

Index Terms—Implicit event argument extraction, argument-argument relation, knowledge distillation, curriculum learning

1 Introduction

How to extract structured knowledge from the massive amount of disordered data and make it available for subsequent tasks is an urgent problem. To this end, information extraction has gained increasing popularity. It aims to extract specific entity, event, or fact information from natural language text, and turn such information into a table-like form of arrangement. As a particular part of information extraction, event argument extraction, which seeks to identify arguments that play specific roles with respect to a given trigger [1], also benefits in various domains, including document summarization [2], [3], question answering [4], [5] knowledge base construction [6], [7], and event graph [8], [9], etc.

In this article, we investigate a more challenging sub-problem of event argument extraction, namely Implicit Event Argument Extraction (IEAE). Unlike traditional event argument extraction task that only processes a single sentence, arguments in IEAE could span multiple sentences. As shown in Fig. 1, given a trigger word *shooting* and its event type *conflict/attack/firearmattack*, an IEAE system aims to extract four corresponding arguments along with their roles in brackets: *mass murder* (*target*), *firearms* (*instrument*), *Andrey Shpagonov* (*attacker*), and *Tatarstan* (*place*).

Mainstream methods that extract event arguments mainly focus on learning pair-wise information between the given trigger and its arguments. [10], [11], [12], [13] cast argument extraction as a relation classification problem to extract pairs of trigger and candidate arguments. Research works [14], [15] utilize event trigger as the predicate and leverages the semantic role labelling model [16], [17] to identify arguments. Former state-of-the-art approaches [18], [19], [20] formulate event argument extraction as a Machine Reading Comprehension (MRC) problem through asking trigger and role-specific questions. Although these works achieve promising success in single sentence-level event argument extraction, current methods struggle in IEAE due to the following critical issues:

1. *Long-Range Dependency*. Since arguments could span multiple sentences, there exist long-range and cross-sentence dependencies between arguments and the given trigger, which are hard to be captured via existing methods.

- *Kaiwen Wei, Xian Sun, and Jianwei Lv are with the Aerospace Information Research Institute, Key Laboratory of Network Information System Technology (NIST), Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100190, China, and also with the School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing 100190, China. E-mail: {weikaiwen19, lvjianwei18}@mails.ucas.ac.cn, sunxian@aircas.ac.cn.*
- *Zequn Zhang, Li Jin, and Zhi Guo are with the Key Laboratory of Network Information System Technology (NIST), Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100190, China. E-mail: {zqzhang1, guozhi}@mail.ie.ac.cn, jinlimails@gmail.com.*
- *Jingyuan Zhang is with the alibaba Damo Academy, Beijing 100102, China. E-mail: weishi.zjy@alibaba-inc.com.*

Manuscript received 10 December 2021; revised 30 September 2022; accepted 30 October 2022. Date of publication 24 November 2022; date of current version 8 August 2023.
The work was supported in part by the Strategic Priority Research Program of the Chinese Academy of Sciences under Grant Y835120378, and in part by the National Natural Science Foundation of China under Grant 62206267.
(Corresponding author: Xian Sun.)
Recommended for acceptance by Y. Chen.
This article has supplementary downloadable material available at https://doi.org/10.1109/TKDE.2022.3218830, provided by the authors.
Digital Object Identifier no. 10.1109/TKDE.2022.3218830

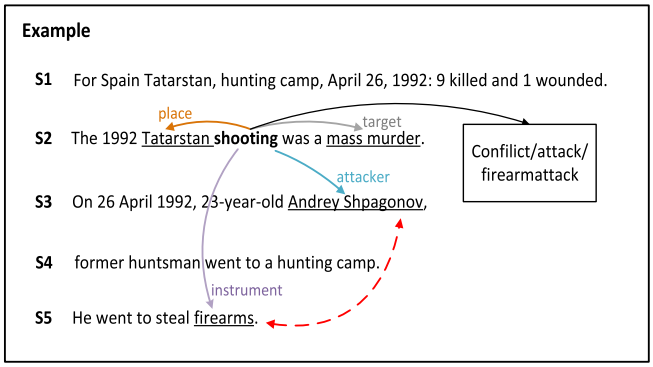

Fig. 1. An instance of implicit event argument extraction on RAMS. Solid lines link the event **trigger**, event type (in the solid box), arguments, and argument roles. The dashed line connects two implicitly related arguments that could be inferred from each other.

2. *Implicit Arguments*. Extracting implicit event arguments requires the ability to reason over event roles, and it is difficult for prior methods to learn these indirect relations.

We attribute these limitations to the fact that current methods are mainly designed to capture direct relations between arguments and the given event trigger. This pair-wise learning paradigm lacks the ability of effective reasoning. We also observe that in MRC-based event argument extraction methods, in addition to the trigger, the *related arguments*, which refer to arguments in the same event except for the required one, could provide some other information to perform reasoning. For example, as shown in Fig. 1, if we have already known *Andrey Shpagonov* plays the *attacker* role of a *firearmattack* event, intuitively, *firearms* could be the instrument of *attacker*. Implicit relations may lie between the two arguments *firearms* and *attacker*, helping identify *firearms* and its role *instrument*. In this manner, arguments corresponding to roles defined in the event frame-level scope could act as clues to perform reasoning. Such argument-argument dependencies could be utilized as relay nodes to capture long-range dependencies.

Nevertheless, the important issues of related arguments are under-exploited. [21] models event arguments as supervising attention information to promote trigger extraction. [1] proposes to learn the association of arguments, but this method works on golden-standard candidate spans, which is unavailable in real-world applications. Existing methods could also be extended by incorporating related arguments and their roles, e.g., concatenating information of related arguments into the original text and interacting with each other by multi-head and multi-layer attention operations in the pre-trained language model. However, since the model is trained with golden-standard arguments, arguments predicted to be imperfect might introduce noise and affect the performance in the test stage. To handle the train-test disparity caused by unavailable *oracle knowledge* [22] in the test stage, we propose a teacher-student framework, named conventional knowledge distillation (CKD), to transfer the knowledge from a well-informed teacher to a student trained without extra information.

Although introducing CKD narrows the gap between the training and testing phase, the relations between arguments are not explicitly captured since CKD only transfers the point-to-point knowledge from the teacher to the student.

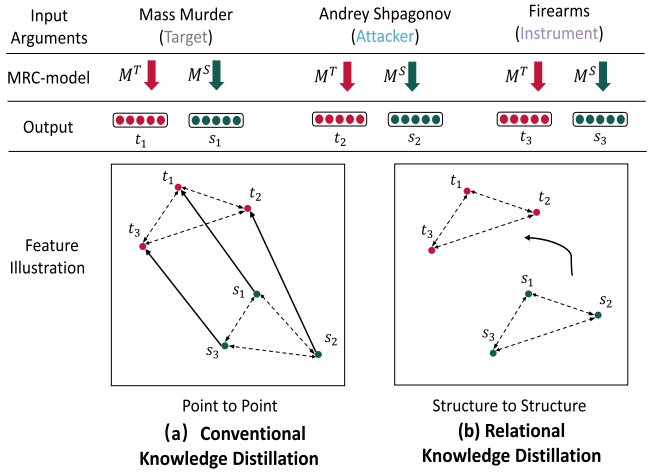

Fig. 2. Comparison of conventional knowledge distillation (CKD) and relational knowledge distillation (RKD) in IEAE, where M^T and M^S are encoding functions of the teacher model and the student model. t_1, t_2, t_3 and s_1, s_2, s_3 are encoded argument features, respectively. In the feature illustration field, solid arrows indicate distillation directions, and dashed arrows mean relations among arguments. CKD seeks to make the student have analogous output positions to the teacher, but RKD hopes the student has similar output topological structure as the teacher.

As shown in Fig. 2, CKD methods simply focus on point-to-point representations (e.g., sample s_1 to sample t_1), but neglect the pair-wise relations (e.g., relation $< s_1, s_2, s_3 >$ to relation $< t_1, t_2, t_3 >$). Typically, in feature space, samples with high similarity cohere together, while samples with low similarity separate from each other. The sample composition structure in the feature space reflects the mutual relation between each other. As a consequence, in CKD, the correlation between arguments of the student has many differences from the teacher', affecting the cohesiveness or separateness of samples. To incorporate such structure-to-structure relations between arguments, we propose a relational knowledge distillation (RKD) mechanism by encouraging the student to have similar output topological structure as the teacher.

Moreover, because the teacher model is trained with all percentages of oracle knowledge, it maximally captures the information of argument-argument correlation. While the input of student model does not contain any oracle knowledge in reality, making it challenging for the model to extract the expected argument by relying on the others. Intuitively, the training process of the student model is more complicated than that of the teacher. Based on this observation, inspired by the curriculum theory [23] that a machine learning model could be trained better by feeding data following the order from easiest to hardest, we introduce a curriculum learning strategy during training to promote the learning of the student model. This method views the proportion of given oracle knowledge as a criterion to evaluate the difficulty of the training process, and gradually increases the learning complexity of the student model to make it more compatible with the real situation, thus a better model is obtained.

In summary, our contributions are listed as follows:

- We introduce an **A**rgument-argument **R**elation enhanced **E**vent **A**rgument extraction (AREA) learning framework for implicit event argument extraction. Argument-argument relational knowledge is

incorporated with reasoning and captures long-range dependencies among different triggers and arguments.
- The proposed method incorporates the implicit relation between event arguments in MRC-model. Knowledge distillation and curriculum learning are both utilized to drive a model that does not require extra tools to produce reasoning clues, and could incorporate argument-argument knowledge effectively.
- A novel relational knowledge distillation framework is integrated with the conventional point-to-point methods, which explicitly captures the structural argument-argument relations. To the best of our knowledge, this is the first work to introduce relational knowledge distillation in the IEAE field.
- Our approach outperforms existing methods and achieves state-of-the-art performance on the RAMS and Wikievents datasets. Ablation study and further discussion also show our method could handle long-range dependency and implicit arguments problems effectively.[1]

2 Related Work

2.1 Event Argument Extraction

Event Argument Extraction (EAE) seeks to extract entities with specific roles in an event. Methods that learn direct relation between arguments and triggers have achieved substantial progress in this field. [10] and [11] first applied convolutional neural networks and recurrent neural network methods to EAE task. [12] utilized graph convolutional neural network based on dependency trees to encode syntax-level information. [21] proposed to build a supervised attention mechanism to force the model to focus more on entities than other parts. In recent years, pre-trained language models (e.g., BERT [24] and ELMo [25]) have made considerable improvements in many natural language processing (NLP) tasks since they incorporate universal language representations from a large amount of unlabeled data. [26] directly applied BERT representations for EAE task, and their model has achieved great performance without designing task-specific architectures or using external resources. Besides, there is a trend to formulate EAE as a Question Answering (QA) problem, and several MRC models report performing well [18], [20], [27]. These methods leverage manual-designed templates to ask role-specific questions, and then extract boundaries of the expected arguments. Nevertheless, most of existing studies are carried out within single sentence scope.

Implicit Event Argument Extraction (IEAE) is a less studied problem where arguments could span multiple sentences and appear in an implicit way. There are only a few works for IEAE. [14], [15] formulated IEAE as a semantic role labelling task and extracted arguments by classifying phrase pairs. These methods only explicitly consider direct relations between triggers and arguments. [28], [29], [30] also introduced generation-based methods, but the argument-argument relations are not explicitly considered. [1] took the relation between arguments into account; however, their method could only deal with argument linking task that identifies the role of a given argument span, which is not available in a realistic situation.

2.2 Knowledge Distillation

Knowledge distillation is first proposed by [31]. conventional knowledge distillation (CKD) typically transfers individual sample from the output of the teacher model to the student model's. [32], [33] leveraged knowledge distillation to generate a much more lightweight student model. [34] utilized feature extractor trained with labelled data as the teacher to teach a student trained with unlabeled data in open-domain. [35] combined knowledge distillation with adversarial training, using entity information as a supervision signal to enhance learning. In this work, we employ the knowledge distillation training strategy to handle the train-test disparity caused by unavailable oracle knowledge in the test stage through driving a student model to learn the behaviour of a well-informed teacher.

Despite achieving promising accomplishments, CKD methods typically only involve individual knowledge distillation [36], while ignoring the relation between samples. As mentioned in the introduction, mutual relations of samples (e.g., sentences or event arguments) also provide valuable information. To capture the structured mutual relation, in the domains of computer vision (CV), [37] transferred the relation knowledge of images from the teacher to the student model. [38] introduced a knowledge distillation method based on correlation congruence, where the distilled knowledge contains image-level information and the correlations between images. [39], [40] leveraged feature embedding or probabilistic distribution as criteria to model the association between images. In NLP fields, although models such as graph neural networks [41] or transformer-based models [42] can be utilized to fuse the features of two samples, the structural mutual knowledge is still hard to be captured. Especially for IEAE, the relation between arguments requires a more fine-grained understanding. To utilize the structural relationships between arguments, we integrate a relational knowledge distillation (RKD) mechanism into the conventional knowledge distillation framework in this work. To the best of our knowledge, there is no related work in the event argument extraction field.

2.3 Curriculum Learning

Curriculum learning is a learning strategy originally proposed by [23] that trains a neural network better through increasing data complexity of training data. The key to curriculum learning is how to measure the difficulty and how to schedule a difficulty reduction strategy. For difficulty measures, many studies evaluate the complexity of the training process from different angles: in CV domains, [43] took the number of objects in a picture as the measure of complexity. [44] defined the difficulty from the perspective of shape variability. Curriculum learning is also broadly adopted in many NLP domains, [45], [46], [47] utilized sentence length, parse tree depth, and word rarity to measure the hardness, respectively. As for training schedulers, they can be divided into discrete and continuous categories. The training difficulty rises in steps with the manually defined rules for discrete schedulers [23], [48]. And the training

[1]The code are available at https://github.com/wkw1259/AREA

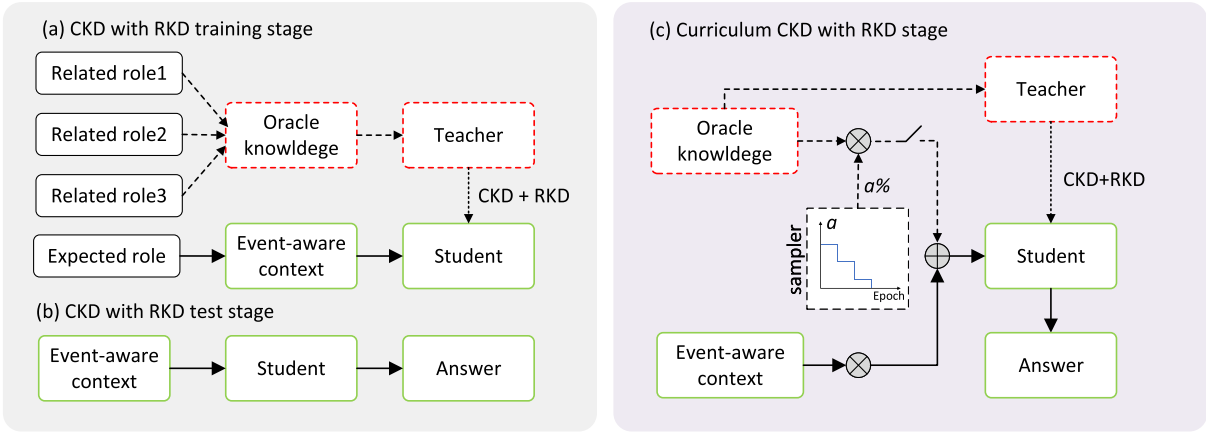

Fig. 3. The architecture of the AREA learning framework. Training and test stages are shown in (a) and (b), Respectively. (c) Shows the curriculum distillation strategy. Data flow of oracle knowledge in the training step is illustrated with dashed lines, and 'role' in the box is short for the argument role. Besides, 'CKD' and 'RKD' are short for 'conventional knowledge distillation' and 'relational knowledge distillation', respectively. 'Oracle knowledge' means the knowledge brought from related arguments, and 'event-aware context' denotes the input question without oracle knowledge.

complexity of continuous schedulers is smoothly rising [49], [50]. In this work, since data with rich related arguments is easier to be learned than those without extra inputs, we promote the training of our student model by increasing the learning complexity of the distillation process, which discretely decreases the proportion of given arguments.

3 Proposed Methodology

Overall, AREA framework consists of two training steps to drive a model that could utilize argument-argument knowledge within the event frame-level scope for IEAE. Details are shown in Fig. 3. The workflow of AREA is as follows: an MRC-based teacher model M^T is first trained with oracle knowledge composing of golden-standard related arguments to exploit argument-argument mutual relation and obtain the capacity to reason. Then a student model M^S that does not have access to this oracle information is driven with the guidance of M^T to be used in practice. A curriculum learning mechanism is further conducted to make the training process compatible with real situations.

The following sub-sections are organized as follows: First, we give the preliminary of the IEAE task and MRC-based model. Then a conventional knowledge distillation strategy is introduced to bridge the gap between the training and inference stage. After that, we illustrate the relational knowledge distillation framework to model argument-argument relations. Finally, we utilize curriculum learning to drive a better model by gradually increasing the training difficulty of the student model.

3.1 Preliminary

We formulate IEAE as a QA problem and leverage the MRC-based model to extract answer spans. The process of MRC-based QA is first asking a question about a specific argument type, then the MRC-based models yield the corresponding answer as the corresponding argument (i.e., return the words that meet the requirement of the question). Specifically, for each argument type, the provided information consists of a tuple $< q, c >$, where q and c refer to the question and context, respectively. In practice, the question q should contain information about a trigger, the event type, and the role of expected argument. We aim to extract a span s in the context that contains the answer to the question.

Formally, given the context $\boldsymbol{C} = \{w_i\}_{i=1}^{n}$ consisting of n words and a known event trigger with the corresponding event type, we seek to identify a set of argument tuples $\mathbb{A} = \{a_j\}_{j=1}^{m} = \{(a_{s_j}, a_{e_j}, role_j)\}_{j=1}^{m}$, where $\mathbb{A}$ is the argument set of the given context; a_j indicates the jth golden-standard argument; a_{s_j} and a_{e_j} are the start and end index of the jth argument, respectively; $role_j$ is the role of this argument.

In addition, we introduce some terminology to understand the proposed AREA framework more easily: (1) *Expected argument*: The event argument belonging to the specific argument type that MRC-based model is asked about. (2) *Related arguments*: The ground-truth event arguments in the same event except for the expected argument. (3) *Oracle knowledge*: The knowledge that is unavailable but assists in prediction (e.g., the future trading information in financial investment), which is first proposed from [22]. In IEAE, we refer to oracle knowledge as the knowledge from the related argument that assists in predicting the expected argument.

3.2 Question Generation

The key of MRC-based QA is to generate questions that contain information about text spans to be extracted. We leverage a template-based question generation strategy to acquire meaningful descriptions of the desired event argument in this work. Assuming there are m arguments in a sample (multi-sentence document), the question template is used to extract arguments with the role of Arg_Type is as follows:

$[Event\_Type]$ $[Arg\_Type]$ with external knowledge $[arg_1]$ as $[role_1]$ and $[arg_2]$ as $[role_2]$ ... and $[arg_{m-1}]$ as $[role_{m-1}]$ in $[Trigger]$.

where $[Trigger]$ and $[Event\_Type]$ should be filled in with event trigger and the corresponding event type, respectively; $[Arg\_Type]$ denotes the role of the expected argument; $[arg]$ and $[role]$ are related arguments and their role types in the same event. Elements in underlines contain oracle knowledge and are excluded during the test stage. It is worth noting that we exclude the information of the

expected argument, and only fill in the question template with the rest m-1 related arguments and their role types. As a result, the MRC-based model can only implicitly infer the expected argument based on the frame-level information between related arguments.

Take the sentences in Fig. 1 as an example, to extract the argument with role "instrument", the generated question for the teacher model should be: "Conflict/attack/firearmattack instrument with extra knowledge mass murder as target and Andrey Shpagonov as attacker and Tatarstan as place in shooting." The elements in underlines consist of related arguments containing the oracle knowledge, which should be excluded during testing. As a result, the event-aware context question at the test stage is: "Conflict/attack/firearmattack instrument in shooting."

3.3 MRC-Based Argument Extraction

We employ the pre-trained language model BERT [24] as the backbone of our MRC-based argument extraction model. The text input is formulated as

$$[CLS]\ question\ [SEP]\ context\ [SEP],$$

where $[CLS]$ and $[SEP]$ are special tokens defined in BERT; $question$ refers to the query generated with our template, and $context$ denotes the context words where arguments are extracted.

This input sequence is then converted into an embedding matrix E and used as the input of the MRC model. We leverage BERT to build semantic representation for each word in the context. For those words containing more than one token, only the first token of this word can be kept.

After the encoding stage, we utilize hidden states from the last BERT layer to represent each token

$$H = \mathrm{BERT}(E), \tag{1}$$

where $H \in \mathbb{R}^{n \times d}$, n is the sequence length, and d is the hidden layer dimension of BERT.

This encoding stage makes a deep fusion between the question and the context by interacting between multi-head and multi-layer attention. In order to explicitly inform the model of the location of the trigger word, we further introduce a position sequence to represent the location of the trigger. The positions of the sequence are set as 1 when the trigger appears, and the others are marked as 0. After that, we convert the position sequence to the positional embedding matrix E_p by looking it up in a randomly initialized embedding table. The concatenations of positional embedding and hidden states are then utilized to produce two probability vectors of the start and end positions

$$\begin{aligned} p_{start} &= \mathrm{softmax}(W_s(H \oplus E_p)/\tau) \\ p_{end} &= \mathrm{softmax}(W_e(H \oplus E_p)/\tau), \end{aligned} \tag{2}$$

where $\oplus$ is the operator of concatenation and τ is the parameter of softmax temperature.

We use cross-entropy between the prediction and golden labels as our training criterion to optimize our model. The following two losses are used for training the start and end index predictions

$$\begin{aligned} \mathcal{L}_{start} &= \mathrm{CE}(p_{start}, Y_{start}) \\ \mathcal{L}_{end} &= \mathrm{CE}(p_{end}, Y_{end}), \end{aligned} \tag{3}$$

where Y_{start} and Y_{end} are ground-truth labels for the index of desired span, respectively. For the situation where no answer exists in the context (missing role of the event), we point these two heads to the $[CLS]$ token. The overall loss of the basic MRC model is formulated as

$$\mathcal{L}_{CE} = \mathcal{L}_{start} + \mathcal{L}_{end}. \tag{4}$$

3.4 Teacher-Student Framework

Although oracle knowledge about related arguments in the same event could provide clues to assist reasoning in the training stage, this golden-standard information is not available for the test stage in practice. This train-test disparity may lead to a performance drop when noisy, or even unrelated arguments are used in the test stage.

To bridge this gap, we adopt the teacher-student framework (namely CKD) to drive a model that is capable of reasoning without the requirement of extra clues. Specifically, as shown in Fig. 3a, we first input question Q^{full} that contains all categories of oracle knowledge to obtain a well-trained teacher model M^T. Then M^T is utilized to generate hidden states H^T and the span distributions p^T_{start} and p^T_{end}. Likewise, a student model M^S, which does not utilize oracle information, produces hidden states H^S and index distributions p^S_{start} and p^S_{end}. The student M^S acquires knowledge from the teacher M^T through learning to have similar behaviours in both hidden vectors and prediction distributions

$$\begin{aligned} \mathcal{L}_{KL} &= (\mathrm{KL}(p^T_{start}, p^S_{start}) \\ &\quad + \mathrm{KL}(p^T_{end}, p^S_{end}))/2 \\ \mathcal{L}_{MSE} &= \mathrm{MSE}(H^T, H^S), \end{aligned} \tag{5}$$

where KL and MSE are short for KL-divergence loss and mean squared error loss, respectively. The teacher outputs p^T_{start} and p^T_{end} could be regarded as "soft labels", which is smoother than conventional "hard labels" that use zero-ones sequences to represent ground-truth labels.

Both the teacher M^T and the student M^S share the same architecture but with diverse parameters. The weights of M^T are fixed and we only optimize the parameters of the student model in the knowledge distillation stage.

The overall loss of the student M^S under our teacher-student framework combines the cross-entropy, KL-divergence and mean squared error loss. Formally, the total loss is formulated as

$$\mathcal{L}_{T,S} = \mathcal{L}_{CE} + \alpha\mathcal{L}_{KL} + \beta\mathcal{L}_{MSE}, \tag{6}$$

where α, β are weight coefficients for different components.

3.5 Relational Knowledge Distillation

In order to explicitly incorporate the structural argument-argument relations in a given context, we further propose a relational knowledge distillation (RKD) mechanism. It aims at transferring structural knowledge using mutual relations of different related arguments from the teacher to the student.

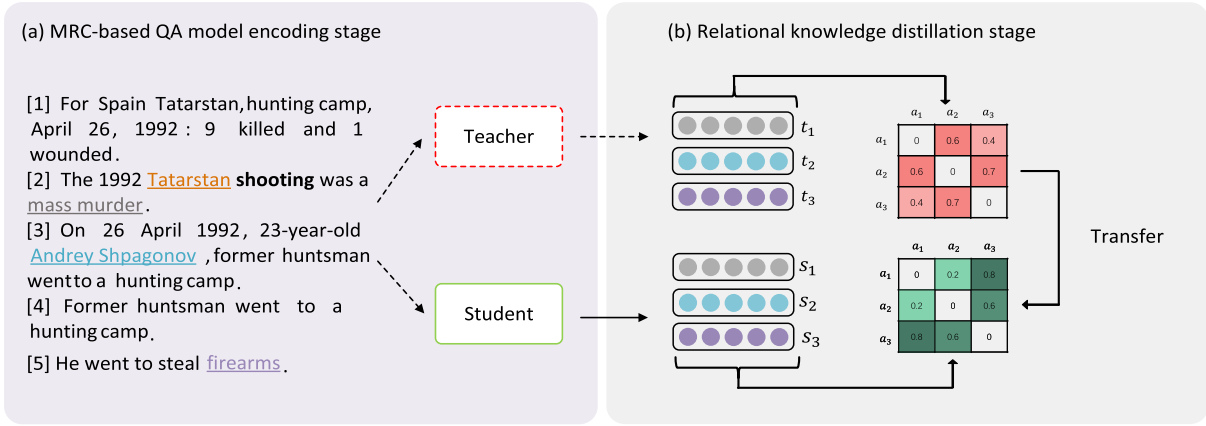

Fig. 4. The workflow of relational knowledge distillation: (a) input sentences to student and teacher models, where the expected argument is "Tatarstan"; (b) relational knowledge distillation transfers the argument-argument relational knowledge from teacher to student, where (t_1, t_2, t_3) and (s_1, s_2, s_3) are associated features of related arguments (a_1, a_2, a_3) encoded from teacher and student, respectively. Please note that we illustrate the naive MMD RKD as an example, where the elements in the matrices denote the relations between different related arguments. Best view in colour.

The workflow of RKD is illustrated in Fig. 4: First, the whole document containing multiple sentences is fed to the teacher and the student model, respectively. Assuming there are m arguments in the given context, except for the expected argument, we get the representations of the rest $m-1$ related arguments from the teacher $(t_1, t_2, \ldots, t_{m-1})$ and the student $(s_1, s_2, \ldots, s_{m-1})$ model according to the corresponding event argument position in the content. Second, take the teacher model as an example, we leverage a matrix V_t to represent the argument-argument relation, where each element v_t^{ij} represents the relation of the ith argument on the jth argument. Likewise, the student model could get the matrix V_s that has the same size as V_t.

During the RKD process, we calculate the element-wise loss at the corresponding positions of V_s and V_t. The goal of RKD is to minimize the argument relational distance between the teacher and the student. Formally, the overall objective of RKD could be expressed as

$$\mathcal{L}_{RKD} = \sum_{i,j} l(\varphi(s_i, s_j), \varphi(t_i, t_j)), \tag{7}$$

where φ is an optional function to calculate pair-wise argument relations, l is a loss function that penalizes the difference between the teacher and the student.

In this article, we present two options to calculate the argument-argument relations: (1) Feature-level RKD, which indicates calculating the correlations on the last hidden layer of BERT. (2) Logit-level RKD, which means calculating the mutual relations on the predicted start or end logits. Please note that in the following descriptions, for simplicity, we take the teacher model as an example. After calculating by the same potential relation function, the student model could also get the corresponding relational representations.

3.5.1 Feature-Level RKD

Since the feature space of BERT has high dimensionality, it is not easy to capture the complex correlations between event arguments. To overcome this problem, in this section, we propose to leverage kernel functions to calculate the relations between diverse event arguments in the feature space.

Formally, let a_i and a_j be two arguments of the given context. After being encoded by BERT, the hidden states of two arguments are represented as H_i and H_j. Since one argument may consist of more than one word, we conduct mean-pooling operations on the word sequences, and finally get the representations of the two arguments as t_i and t_j, respectively. Three kinds of kernel functions are provided to calculate feature-level relation:

1) Naive MMD, which is defined in [51] to measure the distance between two distributions with kernel functions. It reflects the distance between mean embeddings

$$\varphi(t_i, t_j) = \left| \frac{1}{d}\sum_{i=1}^{d} t_i - \frac{1}{d}\sum_{j=1}^{d} t_j \right|, \tag{8}$$

where d is the hidden layer dimension of the related argument representations.

2) Dot Production [52], which calculates the element-wise dot product of different arguments

$$\varphi(t_i, t_j) = t_i^{\top} \cdot t_j. \tag{9}$$

3) Gaussian RBF, which is a commonly used kernel function whose value depends only on the euclidean distance from the original space [38]

$$\varphi(t_i, t_j) = \exp\left(-\frac{\|t_i - t_j\|_2^2}{2\delta^2} \right). \tag{10}$$

Compared to Naive MMD and Dot Production, Gaussian RBF is more flexible and suitable for higher dimensional spaces. Based on Gaussian RBF, the kernel function could be approximated by the P-order Taylor series as

$$\begin{aligned} \varphi(t_i, t_j) &= \exp\left(-\varepsilon \left\| t_i - t_j \right\|^2\right) \\ &\approx \sum_{p=0}^{P} \exp(-2\varepsilon) \frac{(2\varepsilon)^p}{p!} \left(t_i \cdot t_j^{\top}\right)^p, \end{aligned} \tag{11}$$

where ε is a tunable hyperparameter.

3.5.2 Logit-Level RKD

We could also compute the relation function between diverse arguments at the predicted start or end logit level.

Formally, given triple-wise arguments a_i, a_j and a_k, we could get the representations of start logits $t_{s,i}$, $t_{s,j}$, $t_{s,k}$, and end logits $t_{e,i}$, $t_{e,j}$, $t_{e,k}$. Two kinds of knowledge distillation methods are introduced to compute the logit-level relations

1) Distance-wise calculation

$$\varphi(t_{\mathrm{s},i}, t_{\mathrm{s},j}) = \frac{1}{\mu_s}\|t_{\mathrm{s},i} - t_{\mathrm{s},j}\|_2$$
$$\varphi(t_{\mathrm{e},i}, t_{\mathrm{e},j}) = \frac{1}{\mu_e}\|t_{\mathrm{e},i} - t_{\mathrm{e},j}\|_2, \tag{12}$$

where μ_s and μ_e are scalars, which denote the averaged distance by calculating the 2-norm start/end logits of pair-wise arguments in all the related arguments of the sample, which is defined as

$$\mu_s = \sum_{t_{s,i},t_{s,j}} \left\|t_{s,i} - t_{s,j}\right\|_2, \mu_e = \sum_{t_{e,i},t_{e,j}} \left\|t_{e,i} - t_{e,j}\right\|_2. \tag{13}$$

The overall start and end logit relation could be formulated as

$$\varphi(t_i, t_j) = \varphi(t_{\mathrm{s},i}, t_{\mathrm{s},j}) + \varphi(t_{\mathrm{e},i}, t_{\mathrm{e},j}), \tag{14}$$

2) Angle-wise calculation

$$\varphi\big(t_{s,i}, t_{s,j}, t_{s,k}\big) = \cos\angle t_{s,i}t_{s,j}t_{s,k} = \big\langle e_s^{ij}, e_s^{kj}\big\rangle, \tag{15}$$

where

$$e_s^{ij} = \frac{t_{s,i} - t_{s,j}}{\left\|t_{s,i} - t_{s,j}\right\|_2}, e_s^{kj} = \frac{t_{s,k} - t_{s,j}}{\left\|t_{s,k} - t_{s,j}\right\|_2}. \tag{16}$$

Likewise, we could also get the end logit relation representation $\varphi(t_{e,i}, t_{e,j}, t_{e,k})$ based on Equation (15). And the final start and end logit relation of the triple could be calculated as

$$\varphi\big(t_i, t_j, t_k\big) = \varphi\big(t_{s,i}, t_{s,j}, t_{s,k}\big) + \varphi\big(t_{e,i}, t_{e,j}, t_{e,k}\big). \tag{17}$$

Extended from the RKD definition in Eq. (7), the angle-wise logit-level RKD loss among three related arguments could be formulated as

$$\mathcal{L}_{RKD} = \sum_{i,j,k} l(\varphi(s_i, s_j, s_k), \varphi(t_i, t_j, t_k)). \tag{18}$$

Between the two logit-level RKD methods, distance-wise distillation calculates the euclidean distance between arguments at the logit-level, which reflects the distances of pairs of arguments. Since angle-wise calculation provides a higher-order property than a distance, it may provide a more precise simulation of modelling the argument-argument relation.

Overall, we utilize Huber loss for both feature-level and logit-level RKD, which is defined as

$$l(x, y) = \begin{cases} \frac{1}{2}(x-y)^2 & \text{for } |x-y| \le 1 \\ |x-y| - \frac{1}{2}, & \text{otherwise} \end{cases}. \tag{19}$$

After getting the output of RKD, since the feature/logit-level RKD losses do not transfer point-to-point knowledge from the teacher, it is not adequate to use them alone when the individual output values themselves are crucial, e.g., classification layer for IEAE. As a result, we integrate RKD into CKD framework. Modified from Equation (6), the overall loss of the student M^S under our teacher-student framework is

$$\mathcal{L}_{T,S} = \mathcal{L}_{CE} + \alpha\mathcal{L}_{KL} + \beta\mathcal{L}_{MSE} + \gamma\mathcal{L}_{RKD}, \tag{20}$$

where γ is a new weight coefficients to control the proportion of RKD loss.

Note that oracle knowledge in the question template, marked with underlines, is not available in a realistic test situation. In addition, the positions of arguments in RKD are not provided in real-life test scenarios either. In this work, we only utilize oracle knowledge to guide our teacher model to capture argument-argument relational information in the training stage. As illustrated in Fig. 3b, for the test stage of our student model M^S, we discard these extra inputs and fill in slots with event-aware context, which only consists of the event trigger, event type, and the expected argument type. Furthermore, as oracle knowledge is included in the input of the teacher model, in the distillation process, we mask out the question part of the text input for both teacher and student models, and only distill the knowledge of the context part.

3.6 Curriculum Learning Strategy

In this subsection, a curriculum distillation strategy is further combined with CKD and RKD. We view the disparity between the training and test stage from the perspective of learning complexity. Clues in the form of related arguments and their roles are explicitly given to the teacher model to promote reasoning. While for the student model (the inference stage), there are no golden-standard clues, making it challenging for the model to extract the expected argument by relying on associated ones. Intuitively, the training process of the student model is more complicated than that of the teacher.

Algorithm 1. Curriculum Learning Strategy

Input: I^{All}, I, M^T, M^S
Output: p^S_{start}, p^S_{end}
for $q \leftarrow$*100 to 0* **do**
  // sample training question set
  $I^{Train} = Sample(I^{All}, I, q\%)$
  // cache teacher status
  $H^T, p^T_{start}, p^T_{end} = M^T(I^{All})$
  // get student status
  $H^S, p^S_{start}, P^S_{end} = M^S(I^{Train})$
  Apply knowledge distilling to M^S following Equation (20)
end
while *not converged* **do**
  Utilize I to train M^S following Equation (20) without RKD
end

Inspired by the curriculum theory that a machine learning model could be trained better by feeding data following the order from easiest to hardest, we introduce a curriculum learning strategy to promote the learning of the student model. We utilize the proportion of the given arguments to

measure the complexity of the learning task and data points in the IEAE task. As shown in Fig. 3c, at the beginning of the distillation stage, we utilize questions containing oracle knowledge with all related arguments to train the student as a warm-up procedure. Then we gradually reduce the proportion of the given arguments and finally transit to using no extra arguments as in a realistic situation. Note that all teacher models utilize oracle knowledge throughout the whole process of their training. And when there is no oracle knowledge in the input data, we stop utilizing RKD.

Details of the curriculum learning strategy are shown in Algorithm 1. I^{ALL} and I are two sets of training instances with all golden-standard arguments and without extra knowledge used to build questions, respectively. M^T is a well-informed teacher model trained with templates containing all percentages of oracle knowledge. M^S is the student model.

At each training step, first, we sample a batch of instances following Bernoulli distribution, and the probability of selecting an example from the I^{ALL} is $q\%$. Second, we cache the hidden state, start, and end distribution of the teacher model with I^{All} as input. Finally, we utilize all cached status from the teacher model to distill knowledge to student network. As the training stage progresses, the value of q gradually decreases from 100 to 0, leading to the learning difficulty of batches of data from easier to harder. Note that we evaluate the performance of M^S using data without extra arguments in questions. The early stop strategy is conducted to avoid over-fitting when the obtained F1 score on the development set no longer improves after several iterations.

4 Experiments

4.1 Experiment Setup

4.1.1 Dataset

We conduct experiments on the RAMS [14], which is annotated with 139 event types and 65 corresponding argument roles. Each instance consists of a 5-sentences context around the typed event trigger, and there are several typed arguments to be extracted. RAMS dataset consists of 7329, 924, and 871 instances in the training, development, and test set, respectively. To verify the robustness of our proposed model, we also introduce the Wikievents dataset [28], which provides 246 documents, annotated with 50 event types and 59 semantic roles.

4.1.2 Evaluation

An argument is considered correctly identified when the predicted offset fits the golden-standard span. If both the span and the role of an extracted argument are matched with golden-standard one, this argument is correctly classified. Precision (P), Recall (R), and F measure (F1) are adopted as valuation metrics. In the experiments, gold event type information is used in the type constrained decoding (TCD) setting [14].

4.2 Overall Performance

We evaluate the proposed framework against the following baseline models:

1) *BERT-CRF* [53] is a method that combines BERT with Condition Random Field (CRF) [54].
2) *Ebner's* [14] is a semantic role labelling-based method with greedy decoding.
3) *Zhang's* [15] is a two-step head-based model that first predicts head-words of an argument and then expands to the full span.
4) *RESIN* [55] is a method that utilizes multiple sources, multiple languages, multiple modalities data as external knowledge to promote IEAE.
5) *FEAE* [36] is a framework that first trains four teacher models with different proportions of oracle knowledge, and then transfers the knowledge of the teachers to the student model using curriculum knowledge distillation.
6) *BART-Gen* [28] is a method that formulates event argument extraction as a conditional generation task.
7) *DocMRC w/ In-Domain* [56] is a MRC-based model that utilizes in-domain data as augmentation.
8) *DocMRC w/ Impl. DA* [56] is a MRC-based model that utilizes task-related corpus to build a unified training framework.

Since IEAE is a newly proposed task, there are only a few existing works on it. To demonstrate the effectiveness of our method, we also adopt several strong methods from the event argument extraction task and report performances of these baselines and their variants:

9) *GTT* [29] is a generation-based model that "generates" a sequence of role-filler entities. It is conducted on template-filling tasks. For a fair comparison, we further incorporate the trigger word and position information by modifying the question template. For each event category, we enumerate all possible argument roles in a predefined order. More details are illustrated in the supplemental material Section 4, (*available online*).
10) *Student* is our base model that extracts arguments with MRC framework based on [18].
11) *Student_SUP* is a variant where argument information is explicitly modelled with supervising attention mechanism based on [21].
12) *Student_GCN* is a variant where graph nodes are built by named entities extracted from Stanford corenlp toolkit,[2] and adopts multi-hop graph convolutional network for reasoning based on [12].
13) *Student_MKD* is a multi-teacher knowledge distillation framework where four student models trained with various random seeds are used as teachers, and then distill to another student model.
14) *Student_DA* is a variant that utilizes questions with different proportions of oracle knowledge as the data augmentation strategy.
15) *Student_BAG* is a variant that ensembles five well-trained student models through a bagging paradigm.

2. http://stanfordnlp.github.io/CoreNLP/

TABLE 1
Overall Performance on the Test Set of RAMS Dataset and Baseline Methods

	Argument Identification			Argument Classification		
	P	R	F1	P	R	F1
BERT-CRF	-	-	-	39.9	40.7	40.3
Ebner's	-	-	-	**68.8**	14.3	23.7
Zhang's	47.93	35.07	40.50	-	-	-
BART-Gen	-	-	-	41.9	42.5	42.2
DocMRC w/ In-Domain♣	-	-	-	42.6	46.1	44.3
DocMRC w/ Impl. DA♣	-	-	-	43.4	48.3	45.7
RESIN♣	-	-	-	-	-	48.6
FEAE♠	60.87	47.70	53.49	53.17	42.76	47.40
GTT	-	-	-	38.80	36.54	37.63
Student	55.28	44.04	49.03	47.47	39.40	43.06
Student_SUP	57.63	44.49	50.21	51.82	40.29	45.33
Student_GCN	57.34	44.98	50.42	49.37	40.48	44.49
Student_MKD♠	56.87	44.88	50.17	49.44	39.30	43.79
Student_DA♠	61.23	42.07	49.87	54.06	36.73	43.74
Student_BAG♠	57.56	43.99	49.87	50.26	38.56	43.64
Teacher*	54.27	51.85	53.03	50.64	49.13	49.88
Teacher_R	54.61	37.62	44.55	32.29	32.87	32.57
Teacher_MT	55.73	40.33	46.80	48.72	34.80	40.60
FEAE - *multi*	58.61	46.78	52.03	51.06	42.27	46.25
AREA(ours)	**59.49**	**48.44**	**53.29**	52.29	**43.60**	**47.55**

** indicates ground-truth related arguments are used in the test stage. ♣ and ♠ mean leveraging external knowledge or data/model augmentation. **Bold** numbers denote the best results that are obtained without extra knowledge or data/model augmentation.*

16) *Teacher* is a variant with the same architecture as the student, and it is trained and tested with oracle knowledge.
17) *Teacher_R* has the same setting as the Teacher but tested with raw text.
18) *Teacher_MT* [57] is a variant where answering histories from previous turns are fused to the current question in a multi-turn manner.
19) *FEAE - multi* is a variant of FEAE where only one teacher model is adopted in the knowledge distillation stage.

The experimental results on RAMS are shown in Table 1, where * indicates ground-truth related arguments are used in the test stage. ♣ and ♠ mean leveraging external knowledge and data/model augmentation, respectively. We can conclude that: (1) MRC-based methods exceed those strong baselines that directly learn pair-wise relations between event targets and candidate arguments. We attribute these improvements to two folds: first, MRC models could capture relations between arguments from the prior knowledge contained in task descriptions, and they implicitly learn such information during the encoding stage through the QA framework. Second, these approaches explicitly learn the structural argument-argument relations during RKD. (2) Compared to RESIN using a massive amount of external knowledge, AREA achieves similar results with only utilizing golden-standard arguments as clues, and Teacher even outperforms RESIN. One crucial reason is that excessive use of external resources may introduce worthless out-of-domain information, which affects the training process of the model. (3) Without sophisticated fine-tuning on multiple well-trained teacher models, AREA yields better results than FEAE on argument classification task, and it is superior to FEAE - *multi*. Such discovery shows that the argument-argument relations contribute to IEAE, and RKD could effectively capture such information. But the effect of model ensemble is limited. Besides, although FEAE gets higher results on argument identification than AREA, it is achieved at the cost of a substantial increase in computational complexity and storage: as shown in Table 2, FEAE spends nearly 2.5 times as much as AREA training time and storage space. (4) Both AREA and DocMRC-related models are MRC-based models, but the latter leverages external corpus. AREA still achieves better experimental results, illustrating the advantage of the argument-argument knowledge. (5) GTT does not perform well on RAMS dataset. A possible reason is that compared to MUC-4 dataset [58], which only contains 5 event argument roles, RAMS dataset has 139 event types and 65 roles. Because each type of event has different argument types to be decoded, it could be more difficult for GTT to accurately extract the start/end token from the sentences in RAMS. (6) With the same architecture of feature extractor, Student_SUP, Student_GCN, Student_DA, and AREA surpass the Student, and the Teacher that utilizes oracle knowledge in both the training and test stage performs best. These results indicate the effectiveness of related arguments and verify our intuition that reasoning in the event frame-level scope contributes to IEAE. (7) The result gaps among Teacher, Teacher_R, and Teacher_MT clearly show that the train-test disparity could affect the inference procedure. Compared with Teacher_MT, AREA obtains a gain of 6.95

TABLE 2
Training Time and Model Storage Comparison Between FEAE and AREA

	FEAE	AREA
Time/(h)	28.15	10.98
Storage/(M)	2089	835

TABLE 3
Ablation Study on the Test Set of AREA

	F_i	F_c
Teacher*	53.03	49.88
AREA - *rkd - cl - ckd*	49.03	43.06
AREA - *rkd - cl*	50.35	44.75
AREA - *rkd*	52.03	46.25
AREA - *cl*	51.04	45.84
Ours	**52.29**	**47.55**

F_i and F_c mean F1 scores of argument identification and classification.

points in F1, indicating the effectiveness of our teach-student learning strategy. An explanation is that in Teacher_MT, incorrect answers in the previous turn may bring noise and seriously affect the results of subsequent answers. However, AREA is trained with golden-standard related arguments, thus could alleviate such problem of error accumulation. (8) Student_SUP that does not require extra NLP tools to build an explicit graph outperforms Student_GCN. Our method further obtains an improvement of 2.22 absolute points in the argument classification task. These results demonstrate that implicit reasoning and explicit capturing argument-argument relation are powerful solutions to extract the expected event arguments. Another reason is that building explicit reasoning graphs could not avoid introducing noises. (9) Student_MKD, Student_DA, and Student_BAG leverage model integration or data augmentation mechanism to improve performance on IEAE, but AREA still outperforms those methods. These findings reveal that incorporating relations among arguments is more effective than regular model integration mechanisms.

4.3 Ablation Study

We conduct an ablation study to investigate the effect of each component. There are three variations:

1) We remove relational knowledge distillation (*-rkd*).
2) We remove curriculum learning (*-cl*).
3) We remove conventional point-to-point knowledge distillation framework (*-ckd*).

The ablation study results are shown in Table 3. We can observe that: (1) Removing any component of our method leads to a substantial decline on both argument identification and classification, which indicates that each of the three compositions is effective. (2) Conventional knowledge distillation brings as large as 1.69 absolute points in F1 for argument classification. By mimicking the behaviour of a well-informed teacher, our method could effectively obtain the ability of reasoning in event frame-level scope, thus achieving better performances. (3) The curriculum strategy could promote the training process of our student model by gradually filling in the gap between train and test inputs. (4) Introducing relational knowledge distillation further improves the performance of the model since it provides more accurate guidance of structural argument-argument relation information, which is beneficial for IEAE.

4.4 Experiments on Wikievents

To verify the robustness of our proposed model, we also conduct experiments on the Wikievents dataset, which has the same task setting as the RAMS dataset. From the experiment results in Table 4, we observe that: (1) Even if strong baseline models leverage generation-based model, implicit data augmentation, or external corpus with annotator trained on in-domain data, AREA still achieves state-of-the-art results in argument classification on Wikievents corpus. This finding reveals the robustness of AREA on different datasets. (2) Since AREA and DocMRC-related methods are based on MRC models, the main differences are our proposed teacher-student framework and their data augmentation methods. We attribute the experiment promotion of AREA to the argument-argument knowledge, which could be effectively transferred from the teacher model to the student model by the proposed CKD and RKD mechanism. (3) From the ablation study experiments, removing each component results in some degree of performance degradation, which illustrates the effectiveness of each module in AREA.

TABLE 4
Overall Performance and Ablation Studies of AREA on Wikievents Dataset

	F1_i	F1_c
BART-Gen	-	41.7
DocMRC w/ In-Domain♣	-	42.1
DocMRC w/ Impl. DA♣	-	43.3
Teacher*	50.3	48.1
AREA	**48.2**	**44.5**
AREA-*rkd - cl - ckd*	43.6	40.8
AREA-*rkd - cl*	44.9	41.8
AREA-*rkd*	47.5	43.9
AREA-*cl*	46.4	42.6
AREA	**48.2**	**44.5**

** Indicates ground-truth related arguments are used in the test stage, and ♣ denotes leveraging external knowledge.*

4.5 Impact of Relational Knowledge Distillation

Experiments are conducted to investigate the impact of different relational knowledge distillation methods. The experimental results are shown in Table 5, and we can observe meaningful patterns: (1) Most methods including feature-level and logit-level RKD have positive effects on IEAE for both argument identification and classification. It indicates that the structured relationship between arguments can be captured effectively through an appropriate approach. (2) An unexpected observation is that the dot production method does not perform well. A possible reason is that in the experiment, naïve MMD and Gaussian

TABLE 5
Experiment Results of Different Relation Knowledge Distillation Mechanism

	RKD method	F_i	F_c
	None	52.03	46.25
Logit-level RKD	Distance-wise	52.23	46.29
	Angle-wise	**52.92**	47.08
Feature-level RKD	Naive MMD	52.64	47.07
	Dot Production	51.35	45.80
	Gaussian RBF	52.29	**47.55**

TABLE 6
Performance Breakdown by Argument-Trigger Distance d on RAMS Development Set

	$d=-2$ (7.6%)		$d=-1$ (10.4%)		$d=0$ (76.6%)		$d=1$ (3.5%)		$d=2$ (1.9%)	
	F1_i	F1_c	F1_i	F1_c	F1_i	F1_c	F1_i	F1_c	F1_i	F1_c
Zhang's	-	14.0	-	14.0	-	41.2	-	15.7	-	4.2
BART-Gen	-	17.7	-	16.8	-	44.8	-	16.6	-	9.0
DocMRC w/ Impl. DA♣	-	21.0	-	20.3	-	46.6	-	17.2	-	**12.2**
Teacher*	27.59	27.59	23.95	22.49	56.20	52.38	30.07	27.62	9.88	9.88
Student	3.77	3.77	14.49	13.77	51.75	44.00	20.48	17.78	5.79	2.89
FEAE	25.96	23.72	23.61	19.33	55.65	49.20	**26.10**	**25.00**	**7.65**	5.35
AREA	**26.28**	**24.17**	**24.03**	**21.01**	**56.48**	**49.29**	25.77	23.60	6.06	6.06

The percentages below the distance d represent the proportions of the data of that type to the total data volume. ♣ denotes leveraging external knowledge.

RBF utilize the euclidean distances between argument representation as the evaluation metrics, while the dot production method leverage the dot productions. Since the vectors involved in the dot product calculation are not normalized, the calculated values are easily affected by the "norm of vectors", and this unstable loss will cause the model not to converge well. As a result, the dot production approach can not capture the argument relation as well as other methods. (3) Similar to the experiment results in [38], Gaussian RBF achieves the best performance on argument classification compared to other RKD methods. This result indicates that Gaussian distribution could well fit the argument-argument relationship in higher dimensional spaces.

4.6 Performance on Argument Linking

We present the performances of AREA and baselines on the argument linking task in Table 7. For a fair comparison, we follow the argument linking experiment settings in [14]: for each event the model is given the (gold) trigger span and the (gold) spans of the arguments, and for each role the model finds the best argument(s) to fill it. To achieve this, we enumerate all possible combinations of argument spans and argument roles. The goal is to classify whether the argument span could fill the particular argument role. Specifically, we add the expected argument into the question and apply binary classification on the vector of $[CLS]$ token to decide whether the argument plays the given role in the event. Besides, we also incorporate the context of the given argument span to the question. The overall question template we used for argument linking is as follows:

TABLE 7
Performance on Argument Linking

	P	R	F1
Ebner's -TCD	62.8	74.9	68.3
Ebner's +TCD	78.1	69.2	73.3
Joint [1]	79.6	**80.2**	**79.9**
Teacher*	85.5	87.5	86.5
Student	66.4	77.3	71.5
FEAE	82.0	71.6	76.6
AREA	**84.2**	71.5	77.3

* *Indicates ground-truth related arguments are used in the test stage.*

$[Event\_Type]$ $[Arg\_Type]$ $[word]$ with external knowledge $[arg_1]$ as $[role_1]$ and $[arg_2]$ as $[role_2]$ ... and $[arg_n]$ as $[role_n]$ in $[Trigger]$.

where the additional part $[word]$ is the context of the argument according to the given argument span.

From the experiment, we find that AREA has 9 points improvement in F1 score compared to Ebner's -TCD. Compared with the FEAE model that leverages model augmentation methods, AREA also has a level of elevation. Results of this study indicate that argument-argument relational knowledge also contributes to improving the performance of argument linking.

4.7 Performance Breakdown by Distance

To test our method's ability to capture long-range dependencies, we list the performance breakdown on different sentence distances between arguments and the given trigger in Table 6. Compared to FEAE, AREA is more lightweight by replacing model ensemble with RKD and is superior in most scenarios. This finding demonstrates the effectiveness of RKD in capturing argument-argument relational knowledge. For local arguments (where $d=0$), AREA has a huge improvement when compared to Student, and it is competitive with Teacher. Those observations indicate that argument-argument relations are conducive to IEAE in straightforward scenarios.

As for $d=-1$ and $d=1$ situations, nearly all models achieve comparable experimental results. To explore the reasons, we find: (1) The number of samples at $d=1$ is 1.9 times higher than at $d=2$. Besides, the long-range dependence problem is less severe for $d=1$. As a result, $d=1$ achieves much better experimental results than $d=2$. (2) There are 40 categories to classify for $d=-1$, while there are only 22 classes for $d=1$. The former includes 91% of all the latter's category numbers. As a result, the simpler $d=1$ case yields comparable experimental results with $d=-1$.

Additionally, we observe that all models perform poor at $d=2$ than at $d=-2$. The reason could be stated as follows: (1) Insufficient data: the situation of $d=2$ only occupies 1.9% of the total data in the training dataset, while the situation of $d=-2$ accounts for 7.4%. As a result, compared to the situation of $d=0$, the situation of $d=\pm 2$ have a large degree of performance degradation. Especially for $d=2$, the extremely low-resource and long-range dependency scenario makes the model hard to be trained adequately. (2) Imbalanced category distribution between training and

TABLE 8
Case Study of AREA and Student on RAMS Test Set

Category	Example
Long-range dependency	E1: ... In the onslaughts [ISIS]$_{attacker}$ committed killings of whole families for their cooperation with Syrian Army troops, according to Reuters, with some of those killed being beheaded. The residents in the area of the **massacre** called {al - Bagilya}$_{place}$, had received Russian humanitarian aid earlier. ...
Implicit argument	E2: ... Hyatt dials the long telephone number, reaches " Martin ", and tells him that Litvinenko is gravely ill in hospital, the {victim}$_{target}$ of an apparent **poisoning** by two mysterious {Russians}$_{attacker}$. Facebook Twitter Pinterest Police investigate Litvinenko' [poisoning]$_{instrument}$ at the Millennium hotel in central London. Photograph : Alessia Pierdomenico / Reuters

*The **bold** text indicates the trigger word. Ground-truth related arguments are marked in blue with {curly braces} span indicator, while arguments correctly predicted by AREA are represented by the [square brackets] spans with red role types. For brevity, we do not draw the wrongly predicted results of Student model.*

development dataset: for instance, when $d = 2$, the data distribution of training dataset and development dataset is diverse, the 'origin' argument type occupies 6.25% of the training data, while in the development dataset, the percentage is only 2.4%. This problem makes the model difficult to make accurate predictions. However, since our AREA enables the model to reason within the event frame scope and capture the argument-argument relational knowledge, it is natural that our method could mitigate the performance degradation in long-range dependency situations. We sort all argument roles in the $d = \pm 2$ cases by the number of occurrences and find the top five categories are *place*, *recipient*, *instrument*, *participant*, and *attacker*, which cover more than 56% of the total number. Intuitively, there are strong semantic associations between the aforementioned roles and other roles defined in the frame scope.

4.8 Further Discussion

4.8.1 BERT Attention Analysis

The intuition of the attention weights experiment is that when predicting the expected argument, the more attention the related argument is paid to, the more argument-argument knowledge the model learns. To have a better understanding of how AREA improves the MRC model, we conduct two experiments to illustrate the reasoning process with attention weights of the BERT backbone. Specifically, we first conduct an experiment to statistically reveal the effectiveness of argument-argument knowledge. Second, we list some concrete examples to prove our intuition.

TABLE 9
Results on the Top-10 BERT Attention Heads

Related argument	Expected argument	Teacher	AREA
place	damager destroyer	1.53	1.19
recipient	territory or facility	1.36	1.08
destination	granter	1.09	1.07
origin	extraditer	1.14	1.06
participant	beneficiary	1.53	1.04
hidding place	vehicle	1.02	1.04
destination	retreater	1.49	1.04
destroyer	instrument	1.08	1.04
place	employee	1.16	1.04
recipient	surrenderer	1.03	1.03

These values are normalized over all instances with such related argument role pairs.

Following the experiment setting in[59], we extract the top 10 most substantial attention heads from all the 144 BERT-base heads pointing from expected argument to related argument. We enumerate and average those top 10 attention heads from 314 all possible argument role pairs on RAMS test set and find that Teacher and AREA have larger averaged values than Student with 295 and 240 argument pairs, respectively. The result indicates that our approach is able to well guide the BERT model to learn oracle information by modifying the corresponding attention weights and guiding the expected argument to focus more on the clues brought by related argument.

In addition, in Table 9, we show the 10 most notable samples where the listed numbers are averaged attention values. For simplicity, we normalizes those values by student attention outputs, i.e., each value of Teacher and AREA is divided by the corresponding output of Student. It should be noted that the averaged attention weights among different role pairs are numerically incomparable. But in a particular pair, both AREA and Teacher tend to have a larger value than that of Student. Those results indicate that Teacher model could well learn relations between arguments, and our curriculum knowledge distillation framework effectively transfers such knowledge from Teacher to Student. As a result, AREA learns reasoning by paying more attention to the related arguments. For example, in the first instance, intuitively, when looking for *damager destroyer*, arguments with the role of *place* could provide clues about where this event occurs. Accordingly, AREA concentrates more on *place* than Student, and has a higher attention value than Student.

4.8.2 Case Study

In this section, we further illustrate how AREA could alleviate long-range dependencies and implicit argument problems on the RAMS test set. As shown in Table 8, we give representative examples where Student model misses the correct answers, while AREA is able to correctly find them. In the table, ground-truth related arguments are marked in blue with curly braces span indicator, while arguments correctly predicted by AREA are represented by the square brackets spans with red role types.

For the scenario of long-range dependencies in E1, it is difficult to identify the argument of role *attacker* because there are too many words between the argument *ISIS* and the trigger *massacre*. It is challenging for conventional methods to get such long-range dependency problem by regular syntactic

relations. However, there is a strong implicit semantic relationship between *attacker* and *place*, i.e., *place* provides some properties to find where the event takes place, which is conducive to extract *attacker*. Moreover, AREA could better capture such oracle knowledge than Student model, thus AREA successfully finds *ISIS* and classifies it as *attacker*. For the implicit argument situations in E2, since there is no direct association between argument *poisoning* (argument type is *instrument*) and trigger word *poisoning*, Student model fails to identify argument *poisoning*. But argument-argument relations provide the priory that there is an implicit connection between argument role *target*, *attacker*, and *instrument* (intuitively, when in certain circumstances, attacker could use the instrument to attack target). Consequently, AREA successfully recalls argument *poisoning*.

4.8.3 Error Analysis

Following [15], [56], we conduct an error analysis, which samples out 100 error event frames from the development set of RAMS. Overall, there are 219 annotated arguments and 177 predicted ones. The errors could be divided into the following typical categories: (1) *Partial match*, which means there are intersections between predicted results and ground-truth labels. For example, the ground-truth is "French port city of Marseille" while our AREA predicts "Marseille". It accounts for 12.4%. We attribute this problem to the inconsistency of human annotation [14]. (2) *Co-reference*, which means there are co-reference problems in the annotations, such as "us", etc. It accounts for 11.3%. (3) *Multiple role arguments*, which means one event argument may simultaneously play different event argument roles. It accounts for 7.3%. For instance, "refugees" could be "recipient" and "beneficiary", which increases the difficulty of classification. (4) *Location mismatch*, which means there is a mismatched relationship between the ground-truth labels and the predicted results. For example, annotators mark "Russia" as "place", while AREA predicts "Moscow" as "place". It accounts for 3.4%.

5 Conclusion and Future Work

In this article, we exploit the effectiveness of argument-argument relations within the event frame scope for extracting document-level implicit event arguments. We propose an AREA learning framework to train an MRC-based QA model that could focus on argument-argument mutual relations to identify implicit arguments. Specifically, the proposed method leverages a teacher-student framework to avoid the requirement of extra clues and could conduct reasoning with the guidance in event frame-level scope. Besides, we propose a relational knowledge distillation mechanism to explicitly capture structural argument-argument relations, which has not been explored before, to the best of our knowledge. Moreover, to make the student model more compatible with the real situation, a curriculum learning method is further introduced. Experiments show that our method surpasses state-of-the-art baselines on the RAMS and Wikievents datasets, and could substantially alleviate long-range dependency and implicit argument problems in IEAE.

References

[1] Y. Chen, T. Chen, and B. Van Durme, "Joint modeling of arguments for event understanding," in *Proc. 1st Workshop Comput. Approaches Discourse*, 2020, pp. 96–101.

[2] X. Qian, M. Li, Y. Ren, and S. Jiang, "Social media based event summarization by user-text-image co-clustering," *Knowl. Based Syst.*, vol. 164, pp. 107–121, 2019.

[3] R. Pasunuru et al., "Data augmentation for abstractive query-focused multi-document summarization," in *Proc. 35th AAAI Conf. Artif. Intell.*, 2021, pp. 13666–13674.

[4] P. Verma, S. R. Marpally, and S. Srivastava, "Asking the right questions: Learning interpretable action models through query answering," in *Proc. 35th AAAI Conf. Artif. Intell.*, 2021, pp. 12024–12033.

[5] I. Abdelaziz, S. Ravishankar, P. Kapanipathi, S. Roukos, and A. G. Gray, "A semantic parsing and reasoning-based approach to knowledge base question answering," in *Proc. 35th AAAI Conf. Artif. Intell.*, 2021, pp. 15985–15987.

[6] A. Hürriyetoglu et al., "Cross-context news corpus for protest events related knowledge base construction," in *Proc. Autom. Knowl. Base Construction*, 2020, pp. 308–335.

[7] F. Xue, R. Hong, X. He, J. Wang, S. Qian, and C. Xu, "Knowledge-based topic model for multi-modal social event analysis," *IEEE Trans. Multimedia*, vol. 22, no. 8, pp. 2098–2110, Aug. 2020.

[8] R. Han, Q. Ning, and N. Peng, "Joint event and temporal relation extraction with shared representations and structured prediction," in *Proc. Conf. Empir. Methods Natural Lang. Process.*, 2019, pp. 434–444.

[9] T. Dasgupta, R. Saha, L. Dey, and A. Naskar, "Automatic extraction of causal relations from text using linguistically informed deep neural networks," in *Proc. 19th Annu. SIGdial Meeting Discourse Dialogue*, 2018, pp. 306–316.

[10] Y. Chen, L. Xu, K. Liu, D. Zeng, and J. Zhao, "Event extraction via dynamic multi-pooling convolutional neural networks," in *Proc. 53rd Annu. Meeting Assoc. Comput. Linguistics*, 2015, pp. 167–176.

[11] T. H. Nguyen, K. Cho, and R. Grishman, "Joint event extraction via recurrent neural networks," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics*, 2016, pp. 300–309.

[12] X. Liu, Z. Luo, and H. Huang, "Jointly multiple events extraction via attention-based graph information aggregation," in *Proc. Conf. Empir. Methods Natural Lang. Process.*, 2018, pp. 1247–1256.

[13] L. Sha, F. Qian, B. Chang, and Z. Sui, "Jointly extracting event triggers and arguments by dependency-bridge RNN and tensor-based argument interaction," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, Art. no. 726.

[14] S. Ebner, P. Xia, R. Culkin, K. Rawlins, and B. V. Durme, "Multi-sentence argument linking," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 8057–8077.

[15] Z. Zhang, X. Kong, Z. Liu, X. Ma, and E. Hovy, "A two-step approach for implicit event argument detection," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 7479–7485.

[16] M. Surdeanu, R. Johansson, A. Meyers, L. Màrquez, and J. Nivre, "The CoNLL 2008 shared task on joint parsing of syntactic and semantic dependencies," in *Proc. 12th Conf. Comput. Natural Lang. Learn.*, 2008, pp. 159–177.

[17] J. Hajic et al., "The CoNLL-2009 shared task: Syntactic and semantic dependencies in multiple languages," in *Proc. 13th Conf. Comput. Natural Lang. Learn.*, 2009, pp. 1–18.

[18] X. Du and C. Cardie, "Event extraction by answering (almost) natural questions," in *Proc. Conf. Empir. Methods Natural Lang. Process.*, 2020, pp. 671–683.

[19] F. Li et al., "Event extraction as multi-turn question answering," in *Proc. Conf. Empir. Methods Natural Lang. Process.*, 2020, pp. 829–838.

[20] Y. Zhang et al., "A question answering-based framework for one-step event argument extraction," *IEEE Access*, vol. 8, pp. 65420–65431, 2020.

[21] S. Liu, Y. Chen, K. Liu, and J. Zhao, "Exploiting argument information to improve event detection via supervised attention mechanisms," in *Proc. 55th Annu. Meeting Assoc. Comput. Linguistics*, 2017, pp. 1789–1798.

[22] Y. Fang et al., "Universal trading for order execution with oracle policy distillation," in *Proc. 35th AAAI Conf. Artif. Intell.*, 2021, pp. 107–115.

[23] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *Proc. 26th Annu. Int. Conf. Mach. Learn.*, 2009, pp. 41–48.

[24] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics*, 2019, pp. 4171–4186.

[25] M. E. Peters et al., "Deep contextualized word representations," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics*, 2018, pp. 2227–2237.
[26] S. Yang, D. Feng, L. Qiao, Z. Kan, and D. Li, "Exploring pre-trained language models for event extraction and generation," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 5284–5294.
[27] J. Liu, Y. Chen, K. Liu, W. Bi, and X. Liu, "Event extraction as machine reading comprehension," in *Proc. Conf. Empir. Methods Natural Lang. Process.*, 2020, pp. 1641–1651.
[28] S. Li, H. Ji, and J. Han, "Document-level event argument extraction by conditional generation," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics*, 2021, pp. 894–908.
[29] X. Du, A. M. Rush, and C. Cardie, "Template filling with generative transformers," in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2021, pp. 909–914.
[30] X. Du, A. M. Rush, and C. Cardie ., "GRIT: Generative role-filler transformers for document-level event entity extraction," in *Proc. Eur. Chapter Assoc. Comput. Linguistics*, 2021, pp. 634–644.
[31] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," 2015, *arXiv:1503.02531*.
[32] X. Jiao et al., "TinyBERT: Distilling BERT for natural language understanding," in *Proc. Conf. Empir. Methods Natural Lang. Process.*, 2020, pp. 4163–4174.
[33] R. Tang, Y. Lu, L. Liu, L. Mou, O. Vechtomova, and J. Lin, "Distilling task-specific knowledge from BERT into simple neural networks," 2019, *arXiv:1903.12136*.
[34] M. Tong et al., "Improving event detection via open-domain trigger knowledge," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 5887–5897.
[35] J. Liu, Y. Chen, and K. Liu, "Exploiting the ground-truth: An adversarial imitation based knowledge distillation approach for event detection," in *Proc. 33rd AAAI Conf. Artif. Intell.*, 2019, pp. 6754–6761.
[36] K. Wei, X. Sun, Z. Zhang, J. Zhang, Z. Guo, and L. Jin, "Trigger is not sufficient: Exploiting frame-aware knowledge for implicit event argument extraction," in *Proc. 59th Annu. Meeting Assoc. Comput. Linguistics*, 2021, pp. 4672–4682.
[37] W. Park, D. Kim, Y. Lu, and M. Cho, "Relational knowledge distillation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3962–3971.
[38] B. Peng et al., "Correlation congruence for knowledge distillation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 5006–5015.
[39] H. Chen, Y. Wang, C. Xu, C. Xu, and D. Tao, "Learning student networks via feature embedding," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 1, pp. 25–35, Jan. 2021.
[40] N. Passalis, M. Tzelepi, and A. Tefas, "Probabilistic knowledge transfer for lightweight deep representation learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 5, pp. 2030–2039, May 2021.
[41] M. Qu, T. Gao, L. A. C. Xhonneux, and J. Tang, "Few-shot relation extraction via Bayesian meta-learning on relation graphs," in *Proc. 37th Int. Conf. Mach. Learn.*, 2020, Art. no. 729.
[42] A. Vaswani et al., "Attention is all you need," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 6000–6010.
[43] Y. Wei et al., "STC: A simple to complex framework for weakly-supervised semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 11, pp. 2314–2320, Nov. 2017.
[44] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *Proc. 26th Int. Conf. Mach. Learn.*, 2009, pp. 41–48.
[45] E. A. Platanios, O. Stretcu, G. Neubig, B. Póczos, and T. M. Mitchell, "Competence-based curriculum learning for neural machine translation," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics*, 2019, pp. 1162–1172.
[46] Y. Tsvetkov, M. Faruqui, W. Ling, B. MacWhinney, and C. Dyer, "Learning the curriculum with Bayesian optimization for task-specific word representation learning," in *Proc. 54th Annu. Meeting Assoc. Comput. Linguistics*, 2016, pp. 130–139.
[47] T. Kocmi and O. Bojar, "Curriculum learning and minibatch bucketing in neural machine translation," in *Proc. Int. Conf. Recent Adv. Natural Lang. Process.*, 2017, pp. 379–386.
[48] V. I. Spitkovsky, H. Alshawi, and D. Jurafsky, "From baby steps to leapfrog: How "less is more "in unsupervised dependency parsing," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics Hum. Lang. Technol.*, 2010, pp. 751–759.
[49] G. Hacohen and D. Weinshall, "On the power of curriculum learning in training deep networks," in *Proc. 36th Int. Conf. Mach. Learn.*, 2019, pp. 2535–2544.
[50] G. Penha and C. Hauff, "Curriculum learning strategies for IR: An empirical study on conversation response ranking," 2019, *arXiv:1912.08555*.
[51] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. J. Smola, "A kernel method for the two-sample problem," 2008, *arXiv:0805.2368*.
[52] T. Lin, A. RoyChowdhury, and S. Maji, "Bilinear CNN models for fine-grained visual recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1449–1457.
[53] P. Shi and J. Lin, "Simple BERT models for relation extraction and semantic role labeling," 2019, *arXiv:1904.05255*.
[54] J. D. Lafferty, A. McCallum, and F. C. N. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proc. 18th Int. Conf. Mach. Learn.*, 2001, pp. 282–289.
[55] H. Wen et al., "RESIN: A dockerized schema-guided cross-document cross-lingual cross-media information extraction and event tracking system," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics*, 2021, pp. 133–143.
[56] J. Liu, Y. Chen, and J. Xu, "Machine reading comprehension as data augmentation: A case study on implicit event argument extraction," in *Proc. Conf. Empir. Methods Natural Lang. Process.*, 2021, pp. 2716–2725.
[57] Y. Chen, T. Chen, S. Ebner, A. S. White, and B. V. Durme, "Reading the manual: Event extraction as definition comprehension," in *Proc. Conf. Empir. Methods Natural Lang. Process.*, 2020, pp. 74–83.
[58] Fourth Message Uunderstanding Conference (MUC-4): Proceedings of a Conference Held in McLean, Virginia, June, 16–18, 1992, 1992. [Online]. Available: https://aclanthology.org/M92-1000
[59] K. Clark, U. Khandelwal, O. Levy, and C. D. Manning, "What does BERT look at? An analysis of bert's attention," in *Proc. 2019 ACL Workshop BlackboxNLP: Anal. Interpreting Neural Netw. NLP, BlackboxNLP@ACL 2019*, 2019, pp. 276–286.

Kaiwen Wei received the BSc degree from Chongqing University, Chongqing, China, in 2019. He is currently working toward the PhD degree with the Aerospace Information Innovation Institute, Chinese Academy of Sciences, China. His research interests include deep learning, natural language processing and information extraction.

Xian Sun (Senior Member, IEEE) received the BSc degree from the Beijing University of Aeronautics and Astronautics, Beijing, China, in 2004, and the MSc and PhD degrees from the Institute of Electronics, Chinese Academy of Sciences, Beijing, China, in 2009. He is currently a professor with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China. His research interests include machine learning, computer vision, geospatial data mining.

Zequn Zhang received the BSc degree from Peking University, Beijing, China, in 2012, and the PhD degree from Peking University, in 2017. He is currently a research assistant with the Aerospace Information Innovation Institute, Chinese Academy of Science, Beijing, China. His research interests include information fusion, knowledge graph and natural language processing.

Li Jin received the BS degree from Xidian University, Xi'an, China, in 2012, and the PhD degree from the Department of Computer Science and Technology, Tsinghua University, Beijing, China, in 2017. He is currently an associate professor with the Aerospace Information Research Institute, Chinese Academy of Sciences. His research interests include machine learning, knowledge graph and geographic information processing.

Jingyuan Zhang received the BSc degree from the Wuhan University of Technology, Wuhan, China, in 2016, and the PhD degree from Aerospace Information Research Institute, Chinese Academy of Sciences, China. He is the senior algorithm engineer with Alibaba Damo Academy now. His research interests include deep learning, natural language processing and information extraction.

Jianwei Lv received the bachelor's degree from the Minzu University of China, Beijing, China, in 2018. He is currently working toward the PhD degree with the Aerospace Information Research Institute, University of Chinese Academic of Sciences, Beijing, China. His research interests include natural language processing and event extraction.

Zhi Guo received the BSc degree from Tsinghua University, Beijing, China, in 1998, and the MSc and PhD degrees from the Institute of Electronics, Chinese Academy of Sciences, China, in 2003. He is a professor with the Aerospace Information Research Institute, Chinese Academy of Sciences, China. His research interests include computer vision, geospatial data mining, and remote sensing image understanding.