Trigger is Not Sufficient: Exploiting Frame-aware Knowledge for Implicit Event Argument Extraction

Kaiwen Wei[1,2], **Sun Xian**[1,2], **Zequn Zhang**[*1,2],
Jingyuan Zhang[1,2], **Zhi Guo**[1,2] **and Li Jin**[1,2]
[1]Key Laboratory of Network Information System Technology, Aerospace
Information Research Institute, Chinese Academy of Sciences
[2]School of Electronic, Electrical and Communication Engineering, University
of Chinese Academy of Sciences
weikaiwen19@mails.ucas.ac.cn, zqzhang1@mail.ie.ac.cn

Abstract

Implicit Event Argument Extraction seeks to identify arguments that play direct or implicit roles in a given event. However, most prior works focus on capturing direct relations between arguments and the event trigger. The lack of reasoning ability brings many challenges to the extraction of implicit arguments. In this work, we present a Frame-aware Event Argument Extraction (FEAE) 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 to guide the reasoning process. To bridge the gap between oracle knowledge used in the training phase and the imperfect related arguments in the test stage, we further introduce a curriculum knowledge distillation strategy to drive a final model that could operate without extra inputs through mimicking the behavior of a well-informed teacher model. Experimental results demonstrate FEAE obtains new state-of-the-art performance on the RAMS dataset.

1 Introduction

In this work, we investigate the problem of Implicit Event Argument Extraction (IEAE) (Ebner et al., 2020), which seeks to identify arguments that play specific roles respect to a given trigger (Chen et al., 2020). Unlike previous event argument extraction task that only processes a single sentence, arguments in IEAE could span multiple sentences. As shown in Figure 1, given a *conflict/attack/firearmattack* event triggered by the word *shooting*, an IEAE system is required to extract four corresponding arguments with their roles in brackets: *mass murder* (*target*), *firearms* (*instrument*), *Andrey Shpagonov* (*attacker*), and *Tatarstan* (*place*).

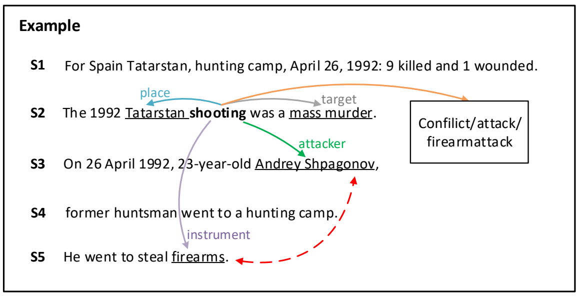
Example
S1 For Spain Tatarstan, hunting camp, April 26, 1992: 9 killed and 1 wounded.
place
target
S2 The 1992 Tatarstan shooting was a mass murder.
Confilict/attack/
firearmattack
attacker
S3 On 26 April 1992, 23-year-old Andrey Shpagonov,
S4 former huntsman went to a hunting camp.
instrument
S5 He went to steal firearms.

Figure 1: Instance of implicit event argument extraction on RAMS. Solid lines link the event **trigger**, event type, arguments, and argument roles. The dashed line connects two implicitly related arguments that could be inferred from each other.

Mainstream methods to extract event arguments focus on learning pair-wise information between arguments and the given trigger. Chen et al. (2015a); Nguyen et al. (2016a); Liu et al. (2018); Sha et al. (2018) cast argument extraction as a relation classification problem to extract pairs of trigger and candidate arguments. Ebner et al. (2020); Zhang et al. (2020b) utilize event trigger as the predicate and leverage semantic role labeling model (Surdeanu et al., 2008; Hajic et al., 2009) to identify arguments. Former state-of-the-art approaches (Du and Cardie, 2020; Li et al., 2020; Zhang et al., 2020a) formulate event argument extraction as a Machine Reading Comprehension (MRC) problem through asking trigger and role-specific questions. Despite the success of these works in single sentence 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 is hard to be captured through existing methods.

2.Implicit Arguments: Extracting implicit event arguments requires the ability to reason over event

[*]Corresponding Author.

*Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics
and the 11th International Joint Conference on Natural Language Processing*, pages 4672–4682
August 1–6, 2021. ©2021 Association for Computational Linguistics

roles, and it is difficult for prior methods to learn these indirect relations.

We attribute these limitations to that current works 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. Instead of only using trigger information, we observe that in MRC-based event argument extraction methods, the related arguments, which refer to arguments (also their roles) in the same event except for the required one, could provide information to perform reasoning. For example, as shown in Figure 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, helping identifying *firearms*. In this manner, arguments corresponding to roles defined in the event frame-level scope could act as clues to perform reasoning and be utilized as relay nodes to capture long-range dependencies.

Nevertheless, the importance of related arguments is under-exploited. Liu et al. (2017) model event arguments as supervising attention information to promote trigger extraction. Chen et al. (2020) propose to learn the association of arguments, but their method works on golden-standard candidate spans, which is unavailable in real-world applications. Existing methods could also be extended to incorporate related arguments and their roles by taking such information as inputs. However, since the model is trained with golden-standard arguments, predicted imperfect arguments might introduce noise and affect the performance in the test stage.

In this work, we introduce a Frame-aware Event Argument Extraction (FEAE) learning framework for IEAE. We extend the MRC-based method to allow reasoning in event frame-level scope by exploiting related arguments and their roles as clues to capture the argument-argument dependencies. This method could learn to extract implicit arguments of an event trigger and handle the long-range dependency problem. To bridge the gap between the unavailable oracle knowledge (Fang et al., 2021) and the imperfect test inputs, we introduce a teacher-student framework that drives a final model that could operate without extra inputs through mimicking the behavior of well-informed teachers. Inspired by the curriculum theory (Bengio et al., 2009), we further introduce a curriculum distillation strategy that gradually increases the learning complexity of the student model to make it more compatible with the real situation, thus driving a better model. In summary, our contributions in this work are as follows:

1) We introduce a Frame-aware Event Argument Extraction framework to train models for implicit event argument extraction. Event frame-level knowledge is incorporated to reason and capture long-range dependencies among triggers and arguments.

2) The proposed model learns to incorporate frame-level knowledge implicitly. Knowledge distillation and curriculum learning are utilized to drive a model that does not require extra tools to produce reasoning clues, and could incorporate frame-level knowledge implicitly.

3) Our approach outperforms existing methods significantly. We achieve new state-of-the-art performance on the RAMS dataset.

2 Related Work

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 significant progress in this field (Chen et al., 2015b; Nguyen et al., 2016b; Zhang et al., 2019; Liu et al., 2018). Recently, there is a trend to formulate EAE as a Question Answering (QA) problem, and several MRC models report performing well (Zhang et al., 2020a; Du and Cardie, 2020; Liu et al., 2020). These methods leverage role-specific questions to extract boundaries of the expected arguments. Implicit Event Argument Extraction (IEAE) is a less studied problem where arguments could span multiple sentences and appear in an implicit way. There have been only a few works for IEAE. Ebner et al. (2020); Zhang et al. (2020b) formulate IEAE as a semantic role labeling task and extract arguments by classifying phrase pairs. These methods only explicitly consider direct relations between triggers and arguments. Chen et al. (2020) also consider the relation among arguments, 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.

Knowledge Distillation is proposed to guide a student model to imitate a well-trained teacher model. It is first proposed by Hinton et al. (2015) and has been widely used in the natural language processing (NLP) field (Ruder and Plank, 2018; Gong et al., 2018; Lee et al., 2018; Jiao et al., 2020). 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 behavior of a well-informed teacher.

Curriculum Learning is a learning strategy firstly proposed by Bengio et al. (2009) that trains a neural network better through increasing data complexity of training data. It is broadly adopted in many NLP domains (Platanios et al., 2019; Huang and Du, 2019; Xu et al., 2020). 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 gradually increasing the learning complexity of the distillation process by decreasing the proportion of given arguments.

3 Method

Our FEAE framework consists of two training steps to drive a model that could utilize frame-level knowledge for IEAE, and details are shown in Figure 2. For single teacher situations, firstly we train an MRC-based teacher model M^T with oracle knowledge composing of golden-standard relevant arguments to exploit frame-aware information 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. Our framework can also be extended to multi-teacher circumstances.

In the following sub-sections, we will give the formulation of our task and our MRC-based model. After that, we will illustrate the curriculum knowledge distillation strategy to bridge the gap between the training and inference stage.

3.1 Task Formulation

We formulate IEAE as a QA problem and leverage the MRC-based model to extract answer spans. 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 the expected argument. We aim to extract a span s in the context that contains the answer to the question.

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

3.2 Frame-aware 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 about the desired event argument in this work. The question template we used to extract arguments with the role of Arg_Type is as follows:

$[Event\_Type]$ $[Arg\_Type]$ $\underline{\text{with } [arg_1] \text{ as } [role_1] \text{ and } [arg_2] \text{ as } [role_2] \ldots \text{ and } [arg_n] \text{ as } [role_n]}$ 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. The MRC-based model could be explicitly aware of the frame-level information by filling in this template, thus making better predictions.

3.3 MRC-based Argument Extraction

We employ the pre-trained language model BERT (Devlin et al., 2019) 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 inputs of the MRC model. We leverage BERT to build semantic representation for each word in the context. After the encoding stage, we utilize hidden states from the last BERT layer to represent each token:

$$H = \text{BERT}(E) \qquad (1)$$

This encoding stage makes a deep fusion between the question and the context by interactions

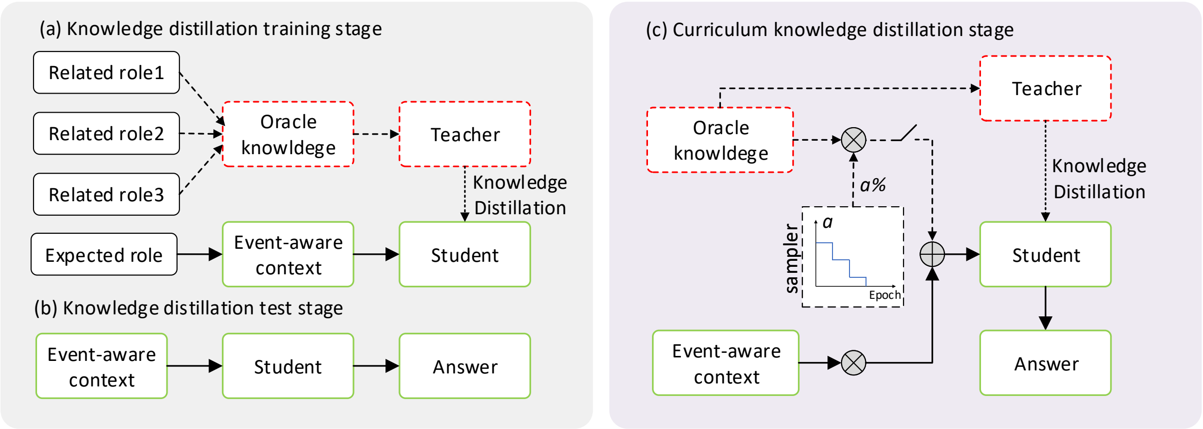

Figure 2: Architecture of the FEAE learning framework. Training and test stages are shown in (a) and (b), respectively. (c) shows the curriculum distillation strategy. Data flow of oracle knowledge in the training step is illustrated with dashed lines and 'role' in the box is short for the argument role.

between multi-head and multi-layer attention. In order to explicitly inform the model with the location of trigger word, we further introduce positional embedding to reflect the relevant distances between words and the specific trigger. The concatenations of positional embedding and hidden states are then utilized to produce two probability vectors of the start and end positions:

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

where E_p is the positional embedding matrix; $\oplus$ 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 start and end index predictions:

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

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 to drive a model that is capable of reasoning without the requirement of extra clues. Specifically, as shown in Figure 2 (a), we first input frame-aware 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 M^S distills knowledge from M^T through learning to have similar behavior in both hidden vectors and prediction distributions:

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

where KL and MSE are short for KL-divergence loss and mean squared error loss, respectively.

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 during the knowledge distillation stage. The overall loss of M^S under our teacher-student framework is formulated as:

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

where α and β are two weight coefficients.

Note that oracle knowledge in the question template, marked with underlines, is not available in a realistic test situation. In this work, we only utilize them to guide our teacher model to capture frame-aware information in the training stage. As illustrated in Figure 2(b), 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. Besides, as oracle knowledge is included in the input of the teacher model, during the distillation process we mask out the question part of the text input in both teacher and student models, and only distill the knowledge of context part.

This teacher-student framework could be further extended to a multi-teacher manner which enables a student model to capture knowledge from multiple perspectives. A teacher model could learn to focus on several patterns to apply reasoning by providing different combinations of related arguments. We drive four teachers trained with diverse templates to capture different categories of oracle knowledge among roles, which are represented with ALL, $ALL-1$, $ALL-2$, and $NONE$, respectively. These templates utilize arguments of different proportions. Take the example of the knowledge distillation training stage in Figure 2 (a), there is one expected argument to be extracted and three related arguments. ALL indicates we fill in the input template with all related arguments. $ALL-1$ denotes that we randomly enumerate the possibilities of two out of the three other arguments and leave one slot unfilled. Questions for $ALL-2$ and $NONE$ are generated in the same method where two or all slots remain unfilled.

For the multi-teacher situation, we distill knowledge into the student model from the four teachers mentioned above simultaneously. The overall multi-teacher distillation loss is formulated as:

$$\mathcal{L} = \sum_{k} \omega_k \mathcal{L}_{T_k,S} \tag{7}$$

where ω_k and $\mathcal{L}_{T_k,S}$ are the weighting factor and the loss function calculated with the k-th teacher model using equation 6, respectively.

3.5 Curriculum Distillation

In this subsection, we view the disparity between the training and test stage from the perspective of learning complexity and introduce our curriculum distillation strategy. Clues in the form of related arguments and their roles are explicitly given for 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 harder than that of the teacher.

Algorithm 1 Curriculum distillation strategy

```
Input: $I^{All}$, $I$, $\{M^{T_k}\}_{k=1}^{4}$, $M^S$
Output: $p_{start}^S$, $p_{end}^S$
for $a$←100 to 0 do
    // build training question set
    $I^{Train} = Sample(I^{All}, I, a\%)$
    for $k$←1 to 4 do
        // cache teacher status
        $H^{T_k}, p_{start}^{T_k}, p_{end}^{T_k} = M^{T_k}(I^{All})$
    end
    // get student status
    $H^S, p_{start}^S, P_{end}^S = M^S(I^{Train})$
    Apply knowledge distilling to $M^S$ following equation 7
end
while not coverage do
    Utilize $I$ to train $M^S$ following equation 7
end
```

Inspired by the curriculum theory that a machine learning model could be trained better by feeding data following the easier to harder order, we introduce a curriculum distillation strategy to promote the learning of student model. We utilize the proportion of given arguments to measure the complexity of the learning task and data points in IEAE task. As in Figure 2 (c), 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 given arguments and finally transit to using no extra arguments as in a realistic situation. Note that all teacher models utilize oracle knowledge as they are trained throughout the whole process.

Details of the curriculum distillation strategy are shown in Algorithm 1. I^{ALL} and I are two sets of training instances with all golden-standard arguments and no extra knowledge are used to build questions, respectively. $\{M^{T_k}\}_{k=1}^{4}$ are four well-informed teacher models trained with diverse templates that capture different categories of oracle knowledge. M^S is the student model. For each training step, firstly, we sample a batch of instances following Bernoulli distribution and the probability of selecting an example from the I^{ALL} is a%.

	Argument Identification			Argument Classification		
	P	R	F1	P	R	F1
Ebner's	-	-	-	**68.8**	14.3	23.7
Zhang's	47.93	35.07	40.50	-	-	-
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	60.87	**47.70**	**53.49**	53.17	**42.76**	**47.40**

Table 1: Overall performance on the test set of RAMS dataset (%) and baseline methods. * indicates ground-truth related arguments are used in the test stage. **Bold** numbers denote the best results that can be obtained without extra knowledge.

Secondly, we cache the hidden state, start and end distribution of the four teachers with I^{All} as input. Finally, we utilize all cached status from teacher models to simultaneously distill knowledge to student network. As the training stage progresses, the value of a 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. We apply the early stop strategy 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

Dataset. We conduct experiments on the RAMS[1] dataset, 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.

Evaluation and Hyperparameters. 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, then this argument is correctly classified. Precision (P), Recall (R), and F measure (F1) are adopted as valuation metrics. Besides, gold event type information is used in the type constrained decoding (TCD) setting.

In experiments, we adopt BERT-base, which has 12 layers, 768 hidden units, and 12 attention heads in every layer, as our MRC model. The batch size is set to 4 and the max sequence length is 512. We set the dimension of the trigger position embedding to 76 and the epoch is set to 7. We train the models with an Adam weight decay optimizer with an initial learning rate of 3e-5. The warming up portion for learning rate is 10%. Temperature τ is set to 1. And we set α as 0.5, β as 2e-3 to balance cross-entropy, KL-divergence, and MSE loss. The proportionality factor a in every epoch is set to 100, 70, 40, 30, 20, 10, 0. And the weighting factors $\{\omega_k\}_{k=1}^{4}$ from ALL, $ALL-1$, $ALL-2$, and $NONE$ are configured as 0.35, 0.25, 0.25, 0.15, respectively.

4.2 Overall Performance

Baselines. (1) **Ebner's** (Ebner et al., 2020) is a semantic role labeling-based method with greedy decoding. (2) **Zhang's** (Zhang et al., 2020b) is a two-step head-based model that first predicts headwords of an argument and then expands to the full span. Since IEAE is a newly proposed task, there are only a few existing works. To demonstrate the effectiveness of our method, we also adopt several strong methods from the EAE task and report performances of these baselines and their variants. (3) **Student** is our base model that extracts arguments with MRC framework based on Du and Cardie (2020). (4) **Student-SUP** is the variant where argument information is explicitly modeled with supervising attention mechanism based on Liu et al. (2017). (5) **Student-GCN** is the variant where graph nodes are built by named entities ex-

[1] https://nlp.jhu.edu/rams/

	F_i	F_c
Teacher*	53.03	49.88
FEAE - *multi - cl - kd*	49.03	43.06
FEAE - *multi - cl*	50.35	44.75
FEAE - *multi*	52.03	46.25
FEAE - *cl*	51.26	45.82
FEAE	**53.49**	**47.40**

Table 2: Ablation study on the test set of FEAE. F_i and F_c mean F1 scores of argument identification and classification.

	NONE	ALL-2	ALL-1	ALL	FEAE
F1_c	45.11	45.23	45.98	46.25	**47.40**

Table 3: Argument classification study with different proportions of arguments. ALL, ALL-1, ALL-2, and NONE denote models trained with various templates.

	P	R	F1
Ebner's -TCD	62.8	**74.9**	68.3
Ebner's +TCD	78.1	69.2	73.3
Teacher*	85.5	87.5	86.5
Student	66.4	77.3	71.5
FEAE	**82.0**	71.6	**76.6**

Table 4: Performance on argument linking.

tracted from Stanford corenlp toolkit[2], and adopts multi-hop graph convolutional network for reasoning based on Liu et al. (2018). (6) **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. (7) **Student-DA** is the variant that utilizes questions with different proportions of oracle knowledge as the data augmentation strategy. (8) **Student-BAG** is the variant that ensembles 5 well-trained student models through a bagging paradigm. (9) **Teacher** is the variant with the same architecture as the student, and it is trained and tested with oracle knowledge. (10) **Teacher-R** has the same setting as the Teacher but tested with raw text. (11) **Teacher-MT** is the variant where answering histories from previous turns are fused to the current question in a multi-turn manner.

From experimental results shown in Table 1, we can conclude that: (1) MRC-based methods exceed those directly learn pair-wise relations among event targets and candidate arguments, leading to strong baselines for IEAE. We attribute these improvements to that MRC models could capture relations among arguments implicitly during the encoding stage through the QA framework. These methods also benefit from the prior knowledge contained in task descriptions. (2) With the same architecture, Student-SUP, Student-GCN, Student-DA, and FEAE 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. (3) 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, our FEAE obtains a gain of 6.80 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, FEAE is trained with golden-standard related arguments, thus could alleviate such error accumulation problem. (4) Student-SUP that does not require extra NLP tools to build an explicit graph outperforms Student-GCN. Our method further obtains an improvement of 2.07 absolute points in the argument classification task. These results demonstrate that implicit reasoning is a powerful way to capture the interrelation between arguments. Another reason is that building explicit reasoning graphs could not avoid introducing noises. (5) The improvements of Student-MKD, Student-DA, and Student-BAG are marginal, illustrating that the improvement in our method is mainly from the architecture of knowledge distillation rather than introducing additional factors. (6) The proposed FEAE outperforms strong baselines and achieves new state-of-the-art results for both argument identification and argument classification. Without using extra inputs, our approach achieves results similar to the one with oracle knowledge. The performance gain clearly indicates that our FEAE could capture frame-aware information effectively.

Ablation Study. To investigate the effect of each component, we conduct an ablation study by removing multi-teacher (*-multi*), curriculum learning (*-cl*), and knowledge distillation framework (*-kd*). We train the model with oracle knowledge containing all related arguments when eliminating multi-teacher(*-multi*), results are shown in Table 2. We can observe that: (1) Knowledge distillation brings as large as 1.69 absolute points in F1 for argument classification. By mimicking the behavior of a well-informed teacher, our method could effectively ob-

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

	d =-2		d =-1		d =0		d =1		d =2	
	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
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**

Table 5: Performance breakdown by argument-trigger distance d on RAMS development set.

Category	Example
Long-range dependency	E1: **Genocide** will never remain in the past . By recognizing the genocide , it will force the $\{$Turkish$\}_{killer}$ government to take a brave step and look into its own history ... from the Turkish and $\{$Armenian$\}_{victim}$ embassies were present in the German parliament while the vote was taking place ...
Implicit argument	E2:. . . Critics of $\{$Putin$\}_{granter}$'s land grab plan say it will only increase the amount of $\{$Chinese workers$\}_{transporter}$ **immigrating** in masses across the border to work on newly - developed destination . Countered one Chinese businessman : " I think the Russians need to understand that if they do n't allow [Chinese investment]$_{destination}$ or Japanese ...

Table 6: Case study on RAMS test set. The **bold** text indicates the trigger word. Ground-truth relevant arguments are marked in blue with {curly braces} span indicator, while arguments correctly predicted by FEAE are represented by the [square brackets] spans with red role types.

tain the ability of reasoning in event frame-level scope, thus achieving better performances. (2) The curriculum strategy could promote the training process of our student model by gradually filling in the gap between train and test inputs. (3) Introducing multiple teachers could provide more accurate guidance from different views and enhance the knowledge distillation framework.

Impact of Frame-aware Knowledge. To get a better understanding of the impact of frame-aware knowledge, we show results with different teacher settings in Table 3, where we adopt a single-teacher curriculum knowledge distillation strategy in experiment. The main difference between these variants is the percentage of oracle knowledge utilized to train teachers, as shown in section 3.4. We find that with the increase of the percentage of ground-truth related argument (the completeness in event frame-level scope), the student could achieve better performance, verifying our assumption that frame-aware knowledge could provide essential information for IEAE. FEAE achieves the best results and shows the importance of capturing multi-view guidances.

Performance on Argument Linking. We present the performances of FEAE and baselines on the argument linking task in Table 4, where ground-truth argument spans are provided and these models are required to identify the role of each span. For our MRC variants, 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. We find that FEAE has an 8.3 points improvement in F1 score compared to Ebner's -TCD, and our FEAE also surpasses baselines. Results of this study indicate that frame-aware knowledge also contributes to improving the performance of argument linking.

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 5. Similar to Zhang et al. (2020b), we observe that all models have a performance drop for the non-local arguments (where $d = \pm 2$ or $d = \pm 1$). Compared with Student, FEAE achieves a gain of more than 4 times by summing the results in the condition of $d = \pm 2$, and the F1 score even increases by 6 times when $d = -2$. To explore the reasons, 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 covers 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. Since our FEAE enables the model to reason with frame-level knowledge, it is natural that our method could mitigate the performance degradation in long-range dependency situations.

4.3 Further Discussion

4.3.1 BERT Attention Analysis

To have a better understanding of how FEAE improves the MRC model, we conduct an experiment

Related argument	Expected argument	FEAE
damager destroyer	place	1.21
beneficiary	participant	1.14
origin	extraditer	1.13
giver	artifact money	1.12
retreater	destination	1.11

Table 7: Results on the Top 10 BERT attention heads. These values are averaged over all instances with such relevant argument role pairs.

to illustrate the reasoning process with attention weights of the BERT backbone. Following Clark et al. (2019), we extract the top 10 most significant 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 FEAE have larger averaged values than Student with 295 and 269 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 guide expected argument to focus more on the clues brought by related argument. In addition, we list the 5 most notable samples where the values are normalized by student averaged values in Table 7. It should be noted that the averaged attention weights among different role-pairs are numerically incomparable. But in a particular pair, FEAE tends to have a larger value than that of the student model, indicating that FEAE learns to reason by paying more attention to the relevant arguments. For example, in the first instance, intuitively, when looking for *place*, arguments with the role of *damager destroyer* could provide clues.

4.3.2 Case Study

In this section, we further illustrate how FEAE could alleviate long-range dependencies and implicit argument problems. As shown in Table 6, we give representative examples where student model misses the correct answers, while FEAE is able to correctly find them. For the scenario of long-range dependencies in E1, it is difficult to identify the argument of role *victim* because there are too many words between the argument *Armenian* and the trigger *Genocide*. However, there is a strong implicit semantic relationship between *killer* and *victim*. FEAE could better capture such oracle knowledge than student model, thus FEAE successfully find and classify *Armenian* as *victim*. For the implicit argument situations in E2, since there is no direct association between argument *Russian farms* and trigger word *immigrating*, student model falls to identify *Russian farms*. But frame-aware knowledge provides the priory that there is an implicit connection between argument role *transporter* and *passenger*. Consequently, FEAE successfully recalls argument *Russian farms*.

5 Conclusion and Future Work

In this paper, we exploit frame-aware knowledge for extracting implicit event arguments. Specifically, we introduce a curriculum knowledge distillation strategy, FEAE, to train an MRC model that could focus on frame-aware information to identify implicit arguments. The proposed method leverages a teacher-student framework to avoid the requirement of extra clues and could perform reasoning with the guidance in event frame-level scope. Experiments show that our method surpasses strong state-of-the-art baselines in RAMS, and could scientifically alleviate long-range dependency and implicit argument problems.

References

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48.

Yubo Chen, Liheng Xu, Kang Liu, Daojian Zeng, and Jun Zhao. 2015a. Event extraction via dynamic multi-pooling convolutional neural networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers*, pages 167–176. The Association for Computer Linguistics.

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

Yunmo Chen, Tongfei Chen, and Benjamin Van Durme. 2020. Joint modeling of arguments for event understanding. In *Proceedings of the First Workshop on Computational Approaches to Discourse*, pages 96–101.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does BERT look at? an analysis of bert's attention. *CoRR*, abs/1906.04341.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Xinya Du and Claire Cardie. 2020. Event extraction by answering (almost) natural questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 671–683. Association for Computational Linguistics.

Seth Ebner, Patrick Xia, Ryan Culkin, Kyle Rawlins, and Benjamin Van Durme. 2020. Multi-sentence argument linking. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 8057–8077. Association for Computational Linguistics.

Yuchen Fang, Kan Ren, Weiqing Liu, Dong Zhou, Weinan Zhang, Jiang Bian, Yong Yu, and Tie-Yan Liu. 2021. Universal trading for order execution with oracle policy distillation. In *AAAI*.

Chen Gong, Xiaojun Chang, Meng Fang, and Jian Yang. 2018. Teaching semi-supervised classifier via generalized distillation. In *IJCAI*, pages 2156–2162.

Jan Hajic, Massimiliano Ciaramita, Richard Johansson, Daisuke Kawahara, Maria Antònia Martí, Lluís Màrquez, Adam Meyers, Joakim Nivre, Sebastian Padó, Jan Stepánek, Pavel Stranák, Mihai Surdeanu, Nianwen Xue, and Yi Zhang. 2009. The conll-2009 shared task: Syntactic and semantic dependencies in multiple languages. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning: Shared Task, CoNLL 2009, Boulder, Colorado, USA, June 4, 2009*, pages 1–18. ACL.

Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the knowledge in a neural network. *CoRR*, abs/1503.02531.

Yuyun Huang and Jinhua Du. 2019. Self-attention enhanced cnns and collaborative curriculum learning for distantly supervised relation extraction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 389–398. Association for Computational Linguistics.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2020. Tinybert: Distilling BERT for natural language understanding. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings, EMNLP 2020, Online Event, 16-20 November 2020*, pages 4163–4174. Association for Computational Linguistics.

Seung Hyun Lee, Dae Ha Kim, and Byung Cheol Song. 2018. Self-supervised knowledge distillation using singular value decomposition. In *European Conference on Computer Vision*, pages 339–354. Springer.

Fayuan Li, Weihua Peng, Yuguang Chen, Quan Wang, Lu Pan, Yajuan Lyu, and Yong Zhu. 2020. Event extraction as multi-turn question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings, EMNLP 2020, Online Event, 16-20 November 2020*, pages 829–838. Association for Computational Linguistics.

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

Shulin Liu, Yubo Chen, Kang Liu, and Jun Zhao. 2017. Exploiting argument information to improve event detection via supervised attention mechanisms. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1789–1798.

Xiao Liu, Zhunchen Luo, and Heyan Huang. 2018. Jointly multiple events extraction via attention-based graph information aggregation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 1247–1256. Association for Computational Linguistics.

Thien Huu Nguyen, Kyunghyun Cho, and Ralph Grishman. 2016a. Joint event extraction via recurrent neural networks. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 300–309. The Association for Computational Linguistics.

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

Emmanouil Antonios Platanios, Otilia Stretcu, Graham Neubig, Barnabás Póczos, and Tom M. Mitchell. 2019. Competence-based curriculum learning for neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019,* *Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 1162–1172. Association for Computational Linguistics.

Sebastian Ruder and Barbara Plank. 2018. Strong baselines for neural semi-supervised learning under domain shift. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 1044–1054. Association for Computational Linguistics.

Lei Sha, Feng Qian, Baobao Chang, and Zhifang Sui. 2018. Jointly extracting event triggers and arguments by dependency-bridge RNN and tensor-based argument interaction. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 5916–5923. AAAI Press.

Mihai Surdeanu, Richard Johansson, Adam Meyers, Lluís Màrquez, and Joakim Nivre. 2008. The conll 2008 shared task on joint parsing of syntactic and semantic dependencies. In *Proceedings of the Twelfth Conference on Computational Natural Language Learning, CoNLL 2008, Manchester, UK, August 16-17, 2008*, pages 159–177. ACL.

Benfeng Xu, Licheng Zhang, Zhendong Mao, Quan Wang, Hongtao Xie, and Yongdong Zhang. 2020. Curriculum learning for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 6095–6104. Association for Computational Linguistics.

Tongtao Zhang, Heng Ji, and Avirup Sil. 2019. Joint entity and event extraction with generative adversarial imitation learning. *Data Intelligence*, 1(2):99–120.

Yunyan Zhang, Guangluan Xu, Yang Wang, Daoyu Lin, Feng Li, Chenglong Wu, Jingyuan Zhang, and Tinglei Huang. 2020a. A question answering-based framework for one-step event argument extraction. *IEEE Access*, 8:65420–65431.

Zhisong Zhang, Xiang Kong, Zhengzhong Liu, Xuezhe Ma, and Eduard Hovy. 2020b. A two-step approach for implicit event argument detection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7479–7485.