Machine Reading Comprehension as Data Augmentation: A Case Study on Implicit Event Argument Extraction

Jian Liu, Yufeng Chen, Jinan Xu
Beijing Jiaotong University, School of Computer and Information Technology, China
{jianliu, chenyf, jaxu}@bjtu.edu.cn

Abstract

Implicit event argument extraction (EAE) is a crucial document-level information extraction task that aims to identify event arguments beyond the sentence level. Despite many efforts for this task, the lack of enough training data has long impeded the study. In this paper, we take a new perspective to address the data sparsity issue faced by implicit EAE, by bridging the task with machine reading comprehension (MRC). Particularly, we devise two data augmentation regimes via MRC, including: 1) implicit knowledge transfer, which enables knowledge transfer from other tasks, by building a unified training framework in the MRC formulation, and 2) explicit data augmentation, which can explicitly generate new training examples, by treating MRC models as an annotator. The extensive experiments have justified the effectiveness of our approach — it not only obtains state-of-the-art performance on two benchmarks, but also demonstrates superior results in a data-low scenario.

1 Introduction

Textual event descriptions may span over multiple sentences. Implicit event argument extraction (EAE) (Ebner et al., 2020; Zhang et al., 2020), a crucial task for event information extraction, aims to identify event arguments *beyond the sentence level*. For example, in a document describing an AirstrikeMissileStrike event (shown in Figure 1), implicit EAE requires a model to recognize a global event argument "Syria", fulfilling the semantic role of PLACE. Note the argument is one-sentence away from the event trigger *bombarding*.

One key challenge faced by implicit EAE is *data sparsity* — owing to the complex interdependencies between triggers and arguments, it is expensive to label training data for the task. The existing datasets, which typically contain several dozens of documents, are too small to train a model for capturing regularities underlying how event arguments appear in texts (Li et al., 2021). For example, even the state-of-the-art model, trained on the full corpus of RAMS (Ebner et al., 2020), attains only 5% in F1 when an event argument is two-sentence away from the trigger (Zhang et al., 2020).

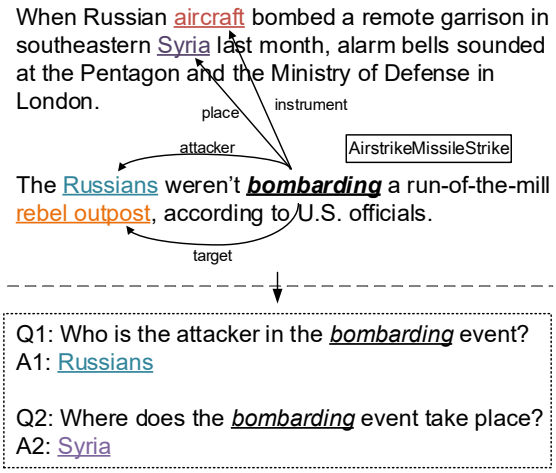

Figure 1: An example of implicit EAE (Above), and the illustration of framing implicit EAE as MRC (Bottom).

This paper attempts to provide a new perspective to address the data sparsity issue faced by implicit EAE. Motivated by previous works handling information extraction via machine reading comprehension (MRC) (Levy et al., 2017; Li et al., 2020; Du and Cardie, 2020b; Liu et al., 2020), we note implicit EAE may be more akin to MRC, as both of them are document-level tasks. For example, we may use a prompt question *Where does the bombarding event take place?* to extract the event argument "Syria", as shown in Figure 1 (Bottom). This formulation implies new ways to address implicit EAE, by leveraging resources in the domain of MRC to boost learning.

We devise two complementary data augmentation methods based on MRC for implicit EAE. The first one is *implicit knowledge transfer*, which aims to build a unified training framework, in the MRC formulation, to facilitate training multiple tasks

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 2716–2725
November 7–11, 2021.

together. It has two main advantages. First, by framing implicit EAE as MRC, we can directly use the sophisticated models in MRC to handle the task, which are shown to be excelled at capturing document-level clues (Devlin et al., 2019). Second, under this framework, we can leverage datasets in other tasks to boost learning. For example, we show it is possible to transfer knowledge from a wide range of tasks, including SQuAD question answering, FrameNet semantic role labeling, and ACE sentence-level event extraction.

Our second method performs data augmentation in a more explicit way, treating a pre-trained MRC model as an annotator to label new training instances. For example, we may use a question *Who is the attacker in the bombarding event?* to query external (unlabeled) documents, and regard documents with answers as new training examples annotated with an ATTACKER role. Compared with implicit knowledge transfer, explicit data augmentation can generate tangible training examples, which is shown to benefit a wide range of previous models (e.g., that based on sequence labeling (Shi and Lin, 2019)) for the task. Moreover, we show explicit data augmentation demonstrates better performance than implicit knowledge transfer does for addressing a zero-shot transfer scenario (§ 6.2).

The expensive experiments on two datasets, RAMS (Ebner et al., 2020) and WikiEvents (Li et al., 2021), have justified the effectiveness of our approach. Particularly, our method achieves substantially improvement over previous methods (+3% in F1 on the average). It also demonstrates promising results for capturing long-range dependencies. Moreover, equipped with the two data augmentation strategies, our approach can fit well with the data-low scenario. For example, on RAMS, with 1% of training data, our approach obtains over 30% in F1, yet previous methods only achieve an F1 score less than 10%.

Our contributions are summarized as follows:

- We study a new view to address the data sparsity issue faced by implicit EAE, by bridging it with MRC. Besides being the first work introducing the MRC formulation to implicit EAE, our work may encourage more studies investigating data augmentation via MRC.
- We propose two novel data augmentation regimes for implicit EAE via MRC — implicit knowledge transfer and explication data augmentation. Their application scopes are carefully explored with extensive evaluations.
- We set up state-of-the-art performance on two implicit EAE benchmarks. We have released our code at https://github.com/jianliu-ml/DocMRC to facilitate further exploration.

2 Related Work

2.1 Implicit Event Argument Extraction

Implicit EAE has long been studied under the MUC-4 paradigm (MUC, 1992), with a core subtask to extract all roll fillers of an event template (Grishman and Sundheim, 1996; Huang and Riloff, 2011; Du and Cardie, 2020a). This line of work is further extended by studies on implicit semantic role labeling (Ruppenhofer et al., 2009; Moor et al., 2013). Despite many advances, the datasets provided by the above evaluations are relatively small, which have long impeded the study on the task. Recently, Ebner et al. (2020) propose a new benchmark, annotating over 3,000 documents for implicit EAE, which has inspired many studies. Following the work, Zhang et al. (2020) devise a head-to-region approach, demonstrating very promising results; Gangal and Hovy (2020) investigate to what extent the pre-trained language model can benefit learning. Very recently, Li et al. (2021) investigate a generative perspective on the task, achieving state-of-the-art performance.

Nevertheless, the currently available datasets are still too small to train a learning based model to achieve decent performance. In this work, we propose a new perspective to address the data sparsity challenge, by bridging the task with machine reading comprehension (Hermann et al., 2015).

2.2 MRC for Information Extraction

Recently, there is a surge of work investigating addressing information extraction tasks using machine reading comprehension. To name a few, Levy et al. (2017); Li et al. (2019) cast relation extraction into question answering; Li et al. (2020) address named entity recognition via MRC; Du and Cardie (2020b); Liu et al. (2020) formulate sentence-level event extraction as MRC. But most works focus on problem reformulation, which rarely consider the data issue. By comparison, our work fills the gap by proposing a new perspective leveraging MRC for data augmentation, which is also the first

work extending MRC to implicit EAL. Additionally, we show our approach can also boost learning for sentence-level event extraction task (§ 6.5).

2.3 Data Augmentation for EAE

Due to the fine-trained annotation of events, data augmentation for event argument extraction is generally challenging. The existing methods are based on distantly supervision (Chen et al., 2017; Yang et al., 2018), leveraging external knowledge bases to generate new training data. However, such works rely on a great deal of expertise limited to domain/language. The work of Yang et al. (2019) combines entity substitution with pre-trained language models, achieving improved performance. But the newly generated examples may be a bit of rigid as the entities remain the same. Different from previous works, we study the possibilities of leveraging MRC for data augmentation. Our method on the one hand does not rely on complex domain knowledge and on the other hand can generate more diversified training data.

3 Implicit EAE in MRC Formulations

We formulate implicit EAE as follows: Assume a document $\mathcal{D}$ contains a set of events $\mathcal{E}$, each represented by an event trigger (e.g., *bombarding* in the previous example). The type of an event $e \in \mathcal{E}$ can determine a set of roles the arguments may take, denoted by $\mathcal{R}_e$. For each semantic role $r \in \mathcal{R}_e$, implicit EAE requires a model to find an event argument a, which is a textual span in $\mathcal{D}$, resulting in a (r, a) pair[1]. Different from previous methods addressing the task via sequence labeling (Shi and Lin, 2019) or span ranking (Ebner et al., 2020), we propose a new perspective based on MRC.

Query Generation. To frame implicit EAE as MRC, we transfer each semantic role r into a question q_r. We devise a template based method operating in three steps: 1) Role Categorization, in which we categorize r as person-based, general, or place-based one to select proper interrogative words (e.g., Who, What, and When). 2) Trigger Format Conversion, where we convert verb-based triggers into their noun formats, using WordNet (Miller, 1992). 3) Query Realization, in which we use the templates in Table 1 to realize the final question. Consider the previous example in Figure 1. Our method can yield two questions *Who is the attacker in the bombarding event?* and *Where does the bombarding event take place?* for the role of ATTACKER and PLACE respectively.

[1]We use $a = \epsilon$ to indicate the case where no event argument fulfills the role of r.

Role Type	Query Generation Template
Person-Based	Who is the $[]_{\text{role}}$ in the $[]_{\text{trigger}}$ event?
General	What is the $[]_{\text{role}}$ of the $[]_{\text{trigger}}$ event?
Place-Based	Where does the $[]_{\text{trigger}}$ event take place?

Table 1: Templates for query generation. $[]_{\text{role}}$ and $[]_{\text{trigger}}$ denote the name of a semantic role and the event trigger (in a format of noun) respectively.

Trigger-Aware Representation Learning. Given the document $\mathcal{D}$ and a question q_r, we build a BERT encoder (Devlin et al., 2019) to learn their joint representations. Particularly, we first construct an extended sequence S = [CLS] q_r [SEP] $\mathcal{D}$ [SEP][2] to concatenate q_r and $\mathcal{D}$. Then, considering multiple events may be contained by $\mathcal{D}$, to indicate which event is currently focused we also devise *trigger-aware embeddings*, modifying BERT's segmentation embeddings to indicate the location of an event trigger (the trigger's segmentation embeddings are sets as 1 instead of 0 as in conventional BERT). Finally, we use BERT encoder to encode S and take the output of its last hidden layer as the joint representation, which is denoted by $\boldsymbol{H}^{\mathcal{D}}_{q_r} \in \mathbb{R}^{N \times d}$, where N is the length of the extended sequence and d is the hidden dimension of BERT.

Argument Extraction as Answer Generation. Based on $\boldsymbol{H}^{\mathcal{D}}_{q_r}$, we compute two normalized vectors, containing the probabilities for the start and end positions of an event argument a over S:

$$\boldsymbol{p}_{start} = \text{softmax}(\boldsymbol{H}^{\mathcal{D}}_{q_r}\boldsymbol{w}_{start}) \quad (1)$$

$$\boldsymbol{p}_{end} = \text{softmax}(\boldsymbol{H}^{\mathcal{D}}_{q_r}\boldsymbol{w}_{end}) \quad (2)$$

where $\boldsymbol{w}_{start} \in \mathbb{R}^d$ and $\boldsymbol{w}_{end} \in \mathbb{R}^d$ are parameters to be learned. The predicted locations of a correspond to the positions having the largest values in $\boldsymbol{p}_{start}$ and $\boldsymbol{p}_{end}$. For the case $a = \epsilon$, i.e., no event argument corresponds to r, we assume the start/end position of a is 0. Namely, the leading token [CLS] in S is treated as a no-answer indicator.

Based on the above formulation, we next detail our two MRC-based data augmentation regimes.

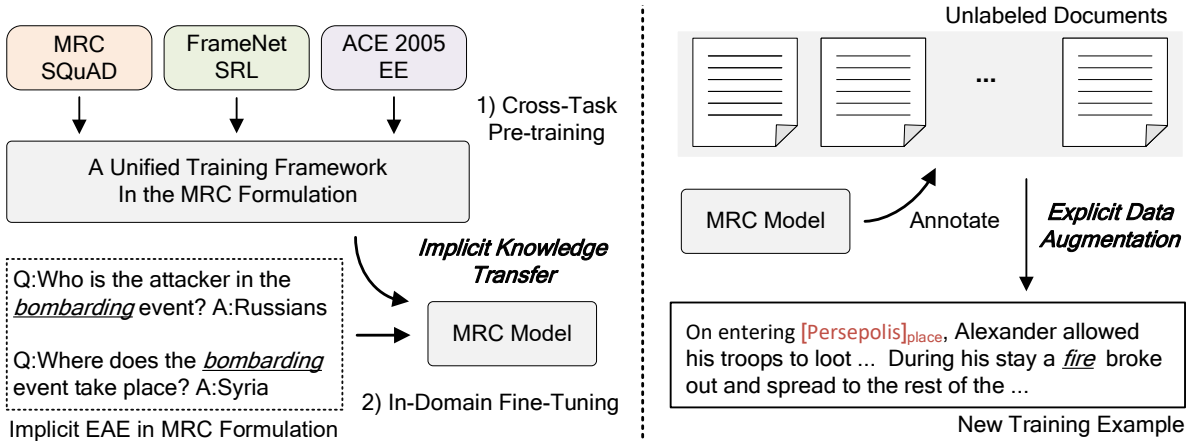

Figure 2: Illustration of two data augmentation regimes via MRC: implicit knowledge transfer (Left), which builds a unified training framework to connect related tasks, and explicit data augmentation (Right), which uses a pre-trained MRC model as annotator to label new training examples.

4 Data Augmentation via MRC

Based on the above proxy of implicit EAE and MRC, we devise two data augmentation regimes: implicit knowledge transfer (§ 4.1) and explicit data augmentation (§ 6.2).

4.1 Implicit Knowledge Transfer

As shown in Figure 2 (Left), implicit knowledge transfer aims to build a unified training framework, which therefore facilitates knowledge transfer from other tasks into implicit EAE. We adopt a pre-training followed by fine-tuning learning paradigm.

Cross-Task Pre-Training. After setting up a MRC model, we first pre-train it using the training data in other tasks. In addition to adopting MRC dataset, i.e., SQuAD 2.0 (Rajpurkar et al., 2018), we also use corpora in FrameNet semantic role labeling (SRL) (Atkins et al., 2003) and ACE event extraction (EE) for pre-training[3], by framing these tasks as a MRC problem in a similar way. The following pre-training objective is adopted:

$$\mathcal{J}_{cross} = -\sum_{\mathcal{T}} \sum_{(\hat{\mathcal{D}},\hat{q},\hat{a})} \log P(\hat{a}|\hat{\mathcal{D}},\hat{q}) \quad (3)$$

where $\mathcal{T}$ ranges over each task and $(\hat{\mathcal{D}},\hat{q},\hat{a})$ ranges over each training example in a MRC formulation. $P(\hat{a}|\hat{\mathcal{D}},\hat{q})$ denotes the likelihood of predicting $\hat{a}$ given $\hat{\mathcal{D}}$ and $\hat{q}$, which equals to $\boldsymbol{p}_{start}[a_{start}]$ + $\boldsymbol{p}_{end}[a_{end}]$, where a_{start} and a_{end} denote the golden start and end positions of $\hat{a}$ in $\hat{\mathcal{D}}$.

[2][CLS] and [SEP] are special tokens used in BERT.

[3]FrameNet SRL and ACE EE use a different event ontology from that of implicit EAE, and therefore it is generally hard to use their datasets for data augmentation.

In-Domain Fine-Tuning. After the pre-training stage converges, we fine-tune the model using in-domain data, with the following training objective:

$$\mathcal{J}_{in} = -\sum_{\mathcal{D}} \sum_{e} \sum_{(r,a)} \log P(a|\mathcal{D},e,r) \quad (4)$$

where $\mathcal{D}$ ranges over each document; e ranges over each event instance in $\mathcal{D}$; (r,a) indicates a role-argument pair. In this way, the knowledge learned from other tasks can be implicitly transferred into the implicit EAE task, which is shown to benefit learning largely in the data-low scenarios (§ 6.1).

4.2 Explicit Data Augmentation

One drawback of implicit knowledge transfer is that it cannot generate explicit training data, and therefore it only supports learning in a MRC formulation. We propose another data augmentation strategy, which can generate explicit examples and benefit models in any formulation for implicit EAE.

Automatic Data Annotation. As shown in Figure 2 (Right), the core idea of explicit data augmentation is to use the pre-trained MRC model as an annotator, to label new instances from unlabeled documents. Given a source document $\mathcal{D}'$, the following steps are conducted: 1) Identify all event triggers in $\mathcal{D}'$, using an event detector pre-trained on the in-domain data. 2) For each event trigger e', enumerate each semantic role r' determined by the event type, and convert r' as a question $q'_{r'}$. 3) Use the pre-trained MRC model to predict answer a' by using $q'_{r'}$ as a prompt. 4) If $a' \neq \epsilon$, construct a new training example $(\mathcal{D}', e', r', a')$. To enhance the annotation quality, we only consider answers whose likelihoods are above a threshold λ. Please

refer to § 5.1 for implementation and the statistics of the generated training examples.

Joint Training Strategy. The following objective is devised to combine the original training data with the automatically generated data for training:

$$\mathcal{J} = -\sum_{\mathcal{D}}\sum_{e}\sum_{(r,a)}\log P(a|\mathcal{D},e,r) \quad (5)$$

$$-\delta\sum_{\mathcal{D}'}\sum_{e'}\sum_{(r',a')}\log P(a'|\mathcal{D}',e',r') \quad (6)$$

where δ is a weight balancing their contributions. The overall process of explicit data augmentation can be seen as "eliciting" knowledge from a pre-trained MRC model, and as the training set is explicitly expanded, it has the potential to benefit any model (e.g., that based on sequence labeling (Shi and Lin, 2019) or span prediction (Ebner et al., 2020)) proposed for implicit EAE.

5 Experiments

5.1 Experimental Setup

Datasets. We conduct our experiments on two implicit EAE benchmarks RAMS (Ebner et al., 2020) and WikiEvents (Li et al., 2021). RAMS provides 3,993 paragraphs in total, annotated with 139 event types and 65 semantic roles; WikiEvents provides 246 documents, annotated with 50 event types and 59 semantic roles. Table 2 gives the detailed data statistics. For evaluation, we use Precision (P), Recall (R), and F1 score (F1) as evaluation metrics. Our experimental results are based on the Exact Match (EM) criterion: only when the predicted argument span matches exactly a gold one, do we count it a correct prediction.

Implementations. In our MRC model, we use a BERT-base-uncased encoder (Devlin et al., 2019), to keep consistent with previous studies (Ebner et al., 2020; Li et al., 2021). As for implicit knowledge transfer, in the pre-training stage, the MRC model achieves 83.5%, 72.1%, and 70.1% in F1 on SQuAD 2.0, FrameNet SRL, and ACE EE respectively, matching the state-of-the-art performance (Devlin et al., 2019; Shi and Lin, 2019; Liu et al., 2020); in the fine-tuning state, we tune parameters on the development set, and finally the batch size is set as 20, chosen from [1, 5, 10, 20, 30, 40]; the learning rate is set as 2e-5, chosen from [1e-5, 2e-5, ..., 1e-4]. As for explicit data augmentation, we adopt a BERT-based event detector (Yang et al., 2018), achieving 77.8% and 75.6% in F1 on RAMS and WikiEvents respectively. We select 500,000 unlabeled documents from the NYT portion of Gigaword[4] as source documents, and set the likelihood threshold λ to 1.95. Finally, our method generates 37,283 documents annotated with 46,632 events and 55,723 arguments for RAMS, and 7,491 documents annotated with 9,633 events and 13,499 arguments for WikiEvents. δ is set as 0.8 for joint training, chosen from [0.1, 0.2, ..., 1].

Dataset	Split	# Doc.	# Event	# Argument
RAMS	Train	3,194	7,329	17,026
	Dev	399	924	2,188
	Test	400	871	2,023
WikiE	Train	206	3,241	4,542
	Dev	20	345	428
	Test	20	365	566

Table 2: Data statistics of RAMS (Ebner et al., 2020) and WikiEvents (WikiE) (Li et al., 2021).

Baseline Models. The following state-of-the-art methods are treated as baselines for comparison:

- BERT-CRF (Shi and Lin, 2019), which combines BERT with Condition Random Field (Lafferty et al., 2001), achieving state-of-the-art performance on sentence-level SRL task.
- SpanSel (Ebner et al., 2020), a method based on span ranking (Lee et al., 2017), which enumerates each possible span in a document to identify the most likely event arguments.
- Head-Expand (Zhang et al., 2020), which extends SpanSel, by first identifying an argument's head, and then its region. It achieves state-of-the-art performance on RAMS.
- BART-Gen[5] (Li et al., 2021), a concurrent work to ours, adopts a generative perspective to address implicit EAE, based on the BART architecture (Lewis et al., 2020).

Our approach is denoted by DocMRC.

5.2 Experimental Result

Table 3 shows the performance of different models on RAMS and WikiEvents. Following Ebner et al.

[4] https://catalog.ldc.upenn.edu/LDC2003T05

[5] BART-Gen adopts BART-large architecture, which has much more parameters than BERT-base used in other methods (Ebner et al., 2020; Zhang et al., 2020). To make a fair comparison, we modify the configuration to BART-base.

Setting	Method	RAMS			WikiEvents			WikiEvents (CR)		
		P	R	F1	P	R	F1	P	R	F1
w/o Type Const.	BERT-CRF (Shi and Lin, 2019)	36.7	41.1	38.8	54.4	23.8	33.1	54.1	26.1	35.2
	SpanSel (Ebner et al., 2020)	38.0	38.4	38.2	56.2	26.2	35.7	55.6	27.6	36.9
	Head-Expand (Zhang et al., 2020)	-	-	40.1	55.4	25.4	34.8	56.7	27.7	37.2
	BART-Gen (Li et al., 2021)	20.7	30.3	24.6	14.2	7.8	10.1	12.6	10.9	11.7
	DocMRC w/ In-Domain	40.1	44.5	42.2	58.2	29.6	39.2	**58.9**	31.1	40.7
	DocMRC w/ Impl. DA	**41.2**	**45.2**	**43.1**	**58.5**	**30.5**	**40.1**	56.9	**32.3**	**41.2**
w/ Type Const.	BERT-CRF (Shi and Lin, 2019)	39.9	40.7	40.3	57.2	22.5	32.3	57.8	25.8	35.7
	SpanSel (Ebner et al., 2020)	38.2	43.6	40.7	57.8	29.2	38.8	57.8	32.9	41.9
	Head-Expand (Zhang et al., 2020)	-	-	41.8	57.8	30.6	40.0	58.1	33.1	42.2
	BART-Gen (Li et al., 2021)	41.9	42.5	42.2	60.0	32.0	41.7	61.2	33.1	43.0
	DocMRC w/ In-Domain	42.6	46.1	44.3	**61.7**	32.0	42.1	63.0	33.9	44.2
	DocMRC w/ Impl. DA	**43.4**	**48.3**	**45.7**	60.2	**33.7**	**43.3**	**64.2**	**36.2**	**46.3**

Table 3: Results on the RAMS and WikiEvents. "w/ Type Constraint" and "w/o Type Constraint" indicate whether gold event types are known or not. P, R, and F1 denote precision, recall, and F1 respectively. WikiEvents (CR) indicates the co-reference relation is considered into evaluation (Li et al., 2021).

Method	Trigger-Argument Distance d				
	$-2_{[4\%]}$	$-1_{[8\%]}$	$0_{[83\%]}$	$1_{[4\%]}$	$2_{[2\%]}$
BERT-CRF	14.0	14.0	41.2	15.7	4.2
SpanSel	15.0	12.2	44.1	12.6	6.6
Head-Expand	15.6	15.3	43.4	**17.8**	8.5
BART-Gen	17.7	16.8	44.8	16.6	9.0
DocMRC (IDA)	**21.0**	**20.3**	**46.6**	17.2	**12.2**

(a) Results on RAMS.

Method	Trigger-Argument Distance d				
	$-2_{[3\%]}$	$-1_{[6\%]}$	$0_{[88\%]}$	$1_{[2\%]}$	$2_{[1\%]}$
BERT-CRF	7.7	2.2	41.1	9.9	-
SpanSel	7.2	4.3	42.7	5.6	-
Head-Expand	10.0	4.1	43.2	6.6	-
BART-Gen	8.9	5.6	43.1	10.1	-
DocMRC (IDA)	**14.5**	**7.8**	**44.1**	**14.3**	**1.7**

(b) Results on WikiEvents (CR).

Table 4: Results on cases with different trigger-argument distances. The ratio of each case is given in the bracket. (IDA) denotes implicit knowledge transfer.

(2020), we adopt two experimental settings, where "w/ Type Constraint" and "w/o Type Constraint" indicate considering gold event types or not. In WikiEvents, the setting of taking co-reference into consideration is denoted by WikiEvents (CR). We denote our approaches with only in-domain training and with implicit knowledge transfer by "w/ In-Domain" and "w/ Impl. DA" respectively.

The experimental results have justified the effectiveness of our approach. Particularly, our approach with implicit knowledge transfer attains the best F1 on the two datasets with different settings, outperforming previous methods by over 3% on the average. Moreover, we note the model with only in-domain training can achieve the state-of-the-art performance, suggesting the effectiveness of problem re-formulation. Implicit knowledge transfer can further boost learning, particularly in Recall (+1.5% on the average). This implies that the knowledge transferred from other tasks enhances the generalization of the model. Additionally, we note a large performance drop of BART-Gen in the setting of "w/o Type Constraint", where event types are known. This suggests it is heavily dependent on correctly predicting the event types. By contrast, our approach doesn't rely on golden event types that much to extract event arguments.

Table 4 gives the performance of different models addressing cases with different trigger-argument distance[6]. The results suggest that our approach is excelled at capturing long-range dependencies. For example, on RAMS, in the case where the event argument is two-sentence ahead the trigger (d=-2), our full approach achieves 21.0% in F1, outperforming previous methods by 3.3%. Nevertheless, there are still many rooms for improvement.

6 Discussion

6.1 Impact of Implicit Knowledge Transfer

To better understand the impact of implicit knowledge transfer, we compare the performance of different models in a stimulated data-low scenario, where we vary the ratio of in-domain training examples for fine-tuning. This scenario also covers

[6]The results are based on the development set following Zhang et al. (2020). We adopt the setting of "w/ Type Constrain" with in-domain training to simplify discussion.

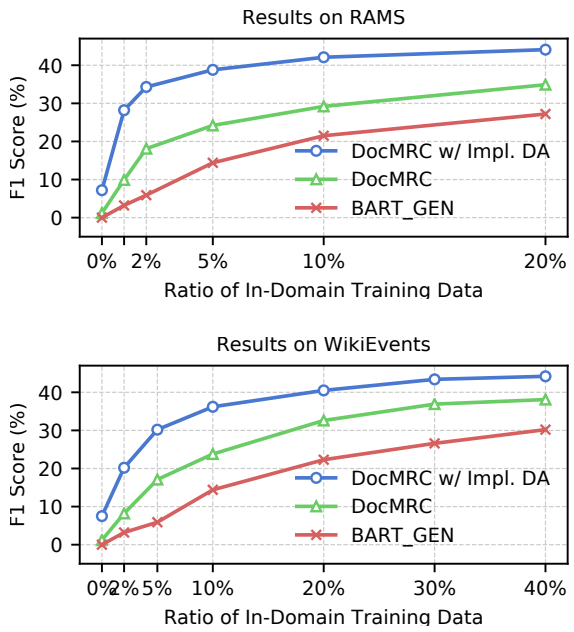

Figure 3: Results in the data-low scenario on RAMS and WikiEvents (CR). 0% denotes a zero-shot scenario where no in-domain data is used for fine-tuning.

Method	P	R	F1
DocMRC (In-Domain)	42.6	46.1	44.3
DocMRC w/ Impl. DA (SQuAD)	42.6	46.9	44.6
DocMRC w/ Impl. DA (FrameNet)	43.1	**47.2**	**45.1**
DocMRC w/ Impl. DA (ACE EE)	**43.3**	46.3	44.6
DocMRC w/ Impl. DA (ALL)	**43.4**	**48.3**	**45.7**

(a) Results on RAMS.

Method	P	R	F1
DocMRC (In-Domain)	63.0	33.9	44.2
DocMRC w/ Impl. DA (SQuAD)	**63.8**	34.8	**45.9**
DocMRC w/ Impl. DA (FrameNet)	63.0	**35.5**	45.4
DocMRC w/ Impl. DA (ACE EE)	63.2	35.1	45.1
DocMRC w/ Impl. DA (ALL)	**64.2**	**36.2**	**46.3**

(b) Results on WikiEvents (CR).

Table 5: Results of implicit knowledge transfer on RAMS and WikiEvents (CR).

Corpus	Method	P	R	F1
RAMS	DocMRC w/ Impl. DA	4.0	10.6	5.8
	DocMRC w/ Expl. DA	**10.4**	**11.7**	**11.0**
	BERT-CRF w/ Expl. DA	9.8	8.6	9.2
	Head-Expand w/ Expl. DA	9.3	9.5	9.4
WikiE	DocMRC w/ Impl. DA	7.9	6.1	6.9
	DocMRC w/ Expl. DA	**10.1**	**10.9**	**10.5**
	BERT-CRF w/ Expl. DA	8.6	9.2	8.9
	Head-Expand w/ Expl. DA	9.1	8.1	8.6

Table 6: Performance of explicit data augmentation for zero-shot evaluation on RAMS and WikiEvents.

a zero-shot transfer case with completely no in-domain training data. Figure 3 gives the results.

The results demonstrate the advantage of our implicit knowledge transfer method clearly. For example, with only 1% of in-domain data, on the RAMS corpus our method augmented with implicit knowledge transfer achieves about 30% in F1, while our method with only in-domain training and other methods achieve less than 10% in F1. Moreover, we note implicit knowledge transfer can even support zero-shot scenario — it achieves 5.8% and 6.9% in F1 on the two datasets without using any in-domain training data.

Table 5 shows the impact of using different tasks for pre-training. From the results, each task can boost learning, and their impacts are complementary. FrameNet and SQuAD lead to larger improvements than ACE, perhaps because they have large, diversified, and wide-coverage datasets.

6.2 Impact of Explicit Data augmentation

Explicit data augmentation, compared with implicit knowledge transfer, has an advantage to generate tangible training examples. We study its performance regarding 1) zero-shot transfer evaluation, and 2) boosting previous models for learning.

Table 6 gives the results of zero-shot evaluation, where we only use the automatically generated data to train a model. From the results, explicit data augmentation yields better performance than implicit knowledge transfer for addressing the zero-shot scenario. A plausible explanation for its effectiveness is that: by using a MRC model as an annotator, we can distill specific knowledge fitting to the event ontology out to boost learning. Moreover, we note the data generated by explicit data augmentation can also help other models, e.g., BERT-CRF and Head-Expand, to address the zero-shot scenario.

Table 7 gives the results of joint training, based on RAMS. From the results, explicit data augmentation improves the performance of different approach by 1.0% in F1 on the average, demonstrating its effectiveness. Nevertheless, we show explicit data augmentation underperforms implicit knowledge transfer in joint training (44.4% v.s. 45.7% in F1). This implies that implicit knowledge transfer may be more preferable than the explicit data generation strategy when we can obtain relatively abundant in-domain training data.

Method	F1	w/ EDA	ΔF1
BERT-CRF (Shi and Lin, 2019)	40.3	41.5	+1.2
SpanSel (Ebner et al., 2020)	40.7	41.5	+0.8
Head-Expand (Zhang et al., 2020)	41.8	42.9	+1.1
BART-Gen (Li et al., 2021)	42.2	43.1	+0.9
DocMRC	43.1	44.4	+1.3

Table 7: Impact of explicit data augmentation (w/ EDA) on RAMS. **ΔF1** denotes the performance gap.

Examples
(1) On entering [Persepolis]$_{\text{place}}$, Alexander allowed his troops to loot the city for several days. Alexander stayed for five months. During his stay a *fire*$_{\text{fireexplosion}}$ broke out and spread to the rest of the city.
(2) With the assistance of the god Hermes, Hector's father [Priam]$_{\text{participant}}$ goes to [Achilles' tent]$_{\text{place}}$ to plead with Achilles for the return of Hector's body so that he can be buried. Achilles relents and promises a truce for the duration of the funeral, lasting 9 days with a *burial*$_{\text{funeralvigil}}$ on the 10th.
(3) Lincoln rarely raised objections; but in an 1859 case, where he defended a cousin, [Peachy Harrison]$_{\text{defendant}}$, who was *accused*$_{\text{chargeindict}}$ of [killing a man]$_{\text{crime}}$, he angrily protested the judge's decision to exclude evidence favorable to his client.

Table 8: Generated examples by explicit data augmentation on RAME. The identified event triggers are shown in *underline*, and the identified event arguments are shown in the brackets, with their roles in subscripts.

6.3 Case Study

Table 8 gives three examples generated by our explicit data augmentation method. From the results, our approach does identify global event arguments. For example, in (1), our approach finds out that "Persepolis", which is two-sentence away from the event trigger *fire*, is an event argument fulfilling the role of PLACE. The above examples can server as perfect training data to boost learning. Nevertheless, we note a skewed distribution of the automatically labeled data. Particularly, the following roles have the most instances: INSTRUMENT: 7456 (20.0%), TARGET: 6365 (17.0%), RECIPIENT: 6108 (16.3%), and PLACE: 4169 (11.2%). But roles such as INSPECTOR, DAMAGER, JAILER, and DETAINEE have less than 10 instances. Our approach fails to identify instances of EXTRADITER and TERRITORYORFACILITY. The reason of the skewed distribution is that our approach relies on a question answering formulation to label examples, however, it is difficult to design proper questions for some rules (e.g., TERRITORYORFACILITY).

Method	P	R	F1
QAEE (Du and Cardie, 2020b)	56.9	49.8	53.1
DocMRC	56.9	50.1	53.4
DocMRC w/ Impl. DA	**57.2**	**53.8**	**55.5**
BERT-CRF	53.2	49.2	51.1
BERT-CRF w/ Exp. DA	**54.1**	**53.1**	**52.8**

Table 9: Results on ACE 2005 (sentence-level) EAE.

6.4 Error Analysis

Following Zhang et al. (2020), we conduct an error analysis, by sampling out 100 error cases from the development set of RAMS. We identify four typical errors: 1) **Partial Match**, which accounts for 16%. For example, the golden annotation of an ATTACKER in "the Palestine solidarity", but our approach predicts "Palestine solidarity". This issue is partially derived from the inconsistency of human annotation (Ebner et al., 2020). 2) **Spurious Semantic**, which accounts for 8%. For example, our approach incorrectly predicts that "Japan" fulfills a PLACE role in "... Japan had accepted the terms ...", owing to not fully understanding the sentence semantic. 3) **Commonsense**, which accounts for 3%. For example, our approach fails to predict that "computer network" fulfills the role of GIVER in *acquires* given a text: "... into the computer network. Someone *acquires* the information ...". How to master commonsense for reasoning is still an open challenge in implicit EAE. 4) **Co-reference**, which accounts for 4%. Different from RAMS, the dataset of WikiEvents has noted this issue and considered co-reference into evaluation, which improves about 2-point in F1 according to Table 3.

6.5 Impact on Sentence-Level EAE

Table 9 gives the result of our approach on the ACE sentence-level EAE task. We compare our method with QAEE (Du and Cardie, 2020b), which adopts a fine-grained query generation strategy (we directly use the trigger prediction result of QAEE to ensure comparability). The results have justified the effectiveness of our approach. Particularity, with implicit knowledge transfer, our approach outperforms QAEE by 2.4% in F1. Additionally, we show explicit data augmentation can also benefit learning — it leads to +1.7% in F1 for the model based on sequence labeling (Shi and Lin, 2019).

7 Conclusion

In this paper we take a new view to handle the data sparsity challenge faced by implicit EAE. Two data augmentation regimes based on MRC are devised, which can implicitly transfer knowledge from related tasks, or generate new training data explicitly, to boost learning. The extensive experiments have justified the effectiveness of our approach. In the future, we would design better question generation method and apply our method to other tasks.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No.6210021758). This work is also supported by Fundamental Research Funds for the Central Universities (No. 2021RC234), and the National Key R&D Program of China (2019YFB1405200).

References

1992. *MUC4 '92: Proceedings of the 4th Conference on Message Understanding*. Association for Computational Linguistics, USA.

Sue Atkins, Michael Rundell, and Hiroaki Sato. 2003. The contribution of framenet to practical lexicography. *International Journal of Lexicography*, 16.3:333–357.

Yubo Chen, Shulin Liu, Xiang Zhang, Kang Liu, and Jun Zhao. 2017. Automatically labeled data generation for large scale event extraction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 409–419, Vancouver, Canada. Association for Computational Linguistics.

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

Xinya Du and Claire Cardie. 2020a. Document-level event role filler extraction using multi-granularity contextualized encoding. In *Association for Computational Linguistics (ACL)*.

Xinya Du and Claire Cardie. 2020b. Event extraction by answering (almost) natural questions.

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*, pages 8057–8077, Online. Association for Computational Linguistics.

Varun Gangal and Eduard Hovy. 2020. Bertering rams: What and how much does bert already know about event arguments? – a study on the rams dataset.

Ralph Grishman and Beth Sundheim. 1996. Message understanding conference-6: A brief history. In *Proceedings of the 16th Conference on Computational Linguistics - Volume 1*, COLING '96, page 466–471, USA. Association for Computational Linguistics.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 1693–1701. Curran Associates, Inc.

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

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, page 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. End-to-end neural coreference resolution. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 188–197, Copenhagen, Denmark. Association for Computational Linguistics.

Omer Levy, Minjoon Seo, Eunsol Choi, and Luke Zettlemoyer. 2017. Zero-shot relation extraction via reading comprehension. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 333–342, Vancouver, Canada. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Sha Li, Heng Ji, and Jiawei Han. 2021. Document-level event argument extraction by conditional generation.

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

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

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

George A. Miller. 1992. WordNet: A lexical database for English. In *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992*.

Tatjana Moor, Michael Roth, and Anette Frank. 2013. Predicate-specific annotations for implicit role binding: Corpus annotation, data analysis and evaluation experiments. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013) – Short Papers*, pages 369–375, Potsdam, Germany. Association for Computational Linguistics.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

Josef Ruppenhofer, Caroline Sporleder, Roser Morante, Collin Baker, and Martha Palmer. 2009. SemEval-2010 task 10: Linking events and their participants in discourse. In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions (SEW-2009)*, pages 106–111, Boulder, Colorado. Association for Computational Linguistics.

Peng Shi and Jimmy Lin. 2019. Simple BERT models for relation extraction and semantic role labeling. *CoRR*, abs/1904.05255.

Hang Yang, Yubo Chen, Kang Liu, Yang Xiao, and Jun Zhao. 2018. DCFEE: A document-level Chinese financial event extraction system based on automatically labeled training data. In *Proceedings of ACL 2018, System Demonstrations*, pages 50–55, Melbourne, Australia. Association for Computational Linguistics.

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

Zhisong Zhang, Xiang Kong, Zhengzhong Liu, Xuezhe Ma, and Eduard Hovy. 2020. 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, Online. Association for Computational Linguistics.