Few-Shot Document-Level Event Argument Extraction

Xianjun Yang **Yujie Lu** **Linda Petzold**
{xianjunyang,petzold}@ucsb.edu
yujielu10@gmail.com
Department of Computer Science
University of California, Santa Barbara

Abstract

Event argument extraction (EAE) has been well studied at the sentence level but under-explored at the document level. In this paper, we study to capture event arguments that actually spread across sentences in documents. Prior works usually assume full access to rich document supervision, ignoring the fact that the available argument annotation is usually limited. To fill this gap, we present **FewDocAE**, a **Few**-Shot **Doc**ument-Level Event **A**rgument **E**xtraction benchmark, based on the existing document-level event extraction dataset. We first define the new problem and reconstruct the corpus by a novel N-Way-D-Doc sampling instead of the traditional N-Way-K-Shot strategy. Then we adjust the current document-level neural models into the few-shot setting to provide baseline results under in- and cross-domain settings. Since the argument extraction depends on the context from multiple sentences and the learning process is limited to very few examples, we find this novel task to be very challenging with substantively low performance. Considering FewDocAE is closely related to practical use under low-resource regimes, we hope this benchmark encourages more research in this direction. Our data and codes will be available online[1].

1 Introduction

Event argument extraction (EAE), a sub-task of event extraction, is a fundamental task for many downstream NLP applications in the IE community. For example, events and arguments play an important role in the knowledge base population from unstructured data (Ge et al., 2018; Li et al., 2021). And the real world public affairs management relies on recognizing the event arguments from daily news and social media (Yuan et al., 2018; Ritter et al., 2012). Although tremendous progress has been made under the supervised setting, current neural models typically rely on large-scale human-annotated data, which is not reliable considering the huge amounts of novel events and arguments emerging in many fields every day. Few-shot learning (FSL) (Fei-Fei et al., 2006) is proposed to tackle such limitations to make machine learning models more applicable given limited annotated examples and has been used a lot in the IE area (Han et al., 2018; Ding et al., 2021; Lai et al., 2021).

Previous research (Yang et al., 2019; Tong et al., 2020) mainly focuses on sentence-level event extraction, such as the popular ACE2005 (Doddington et al., 2004) dataset. In recent years, researchers have begun to realize that the complete event and arguments actually spread in a full document or paragraph (Li et al., 2021; Ebner et al., 2020; Li, 2022). And the focus starts to turn into document-level event extraction motivated by the newly proposed large-scale document-level corpora, namely the WikiEvents (Li et al., 2021), RAMS (Ebner et al., 2020) and the recent largest corpus DocEE (Li, 2022). Following these datasets, many novel methods for solving such new challenges brought by the longer context have also been investigated and witness significant progress (Du and Cardie, 2020; Li et al., 2021; Xu et al., 2022).

However, the traditional supervised learning methods heavily rely on large-scale annotated training data, but we are witnessing new events every day due to the rapid emergence of new affairs. Thus it is not durable to greedily make large collections of newly appeared events for real-life applications. Therefore, more attention has been paid to few-shot event extraction. But the current research has only considered the few-shot EAE on a single sentence (Lai et al., 2021, 2020; Deng et al., 2020), ignoring the big gap between realistic scenarios. Therefore, we aim to pave a new way for few-shot EAE at the document level towards urgent data scarcity problems on the complete documents.

The recently released large-scale document-level

[1] https://github.com/Xianjun-Yang/FewDocAE

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 8029–8046

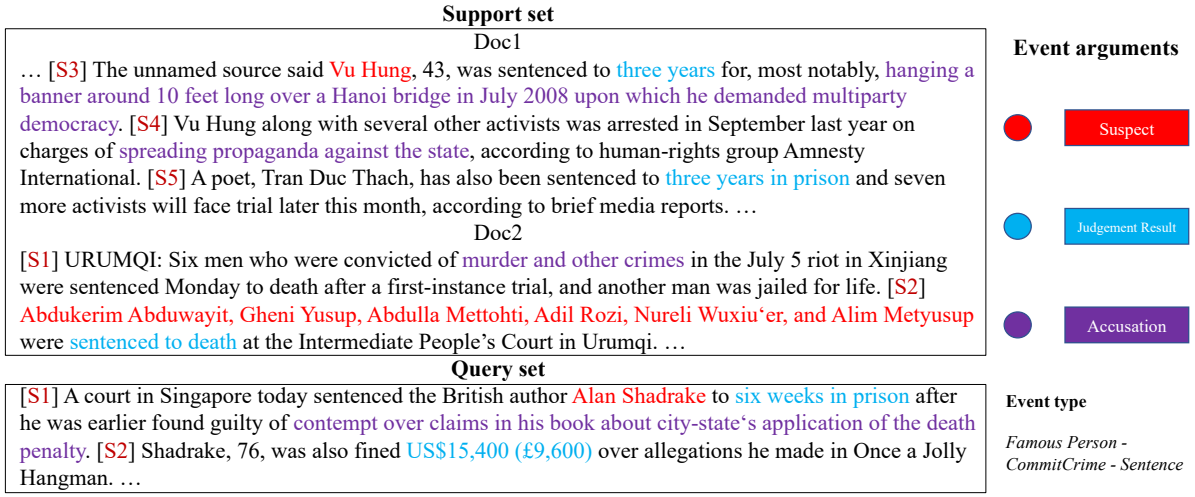

Figure 1: An example of a 3-Way-2-Doc episode consisting of a support and query set. Given a support set of 2 documents with 3 argument types, the goal is to extract all event arguments in the query document. During testing, the argument types are disjoint from type set in the training set. But the documents still share the same main event type.

EE datasets include RAMS (Ebner et al., 2020) and DocEE (Tong et al., 2022), and their statistics is shown in Table 1. RAMS and DocEE contain 139 and 59 event types, 65 and 158 arguments types, with a total collection of 9,124 and 21,450 documents, respectively. While other datasets such as WikiEvents (Li et al., 2021) and MUC-4 (Grishman and Sundheim, 1996) only contain extremely limited types and documents, thus not suitable for our settings. Therefore, we formulate our FewDocAE based on the largest DocEE dataset. Different from FSL for single sentences by traditional N-way-K-Shot sampling, a novel N-Way-D-Doc sampling strategy is proposed for our document-level task, as can be seen from the example in Fig. 1. Besides, previous few-learning problems often fall into the pitfall of robust and fair evaluation, and we follow the FLEX (Bragg et al., 2021) (Few-shot Language Evaluation across(X) many transfer types Principles) to design our settings to avoid such weaknesses. Moreover, prototypical networks (ProtoNet) (Snell et al., 2017) have been proven to be very powerful for solving few-shot problems by representing each category as a prototype in both Vision (Pan et al., 2019; Dong and Xing, 2018) and NLP (Sun et al., 2019; Gao et al., 2019) domains. We combine ProtoNet and a pre-trained language encoder for establishing baselines on our new task and provide a comprehensive analysis.

The key contributions of this work include:

- We are the first to introduce few-shot document-level event argument extraction, greatly extending supervised document-level EAE to few-shot scenarios.
- We reconstruct a realistic FewDocAE dataset along with a new few-shot sampling algorithm, N-Way-D-Doc sampling.
- We conduct comprehensive experiments to provide benchmark results and find the tasks are extremely challenging and worth further investigation.

2 Related Work

2.1 Document-Level Information Extraction

In the IE community, previous work mainly focuses on sentence-level tasks. For example, the commonly-used relation extraction benchmark TACRED (Zhang et al., 2017), the ACE05 [2] and KBP2017 [3] event extraction datasets, the CoNLL-2003 (Sang and De Meulder, 2003) and OntoNotes 5.0 (Pradhan et al., 2013) named entity recognition corpora, all focus on single sentence-level semantics. Since information extraction often involves document-level reasoning, recently there have been great efforts to contribute document-level benchmarks. For instance, the DocRED (Yao et al., 2019) is the largest dataset to extend relation extraction to the document level. The recent RAMS (Ebner et al., 2020) and DocEE (Tong et al., 2022) corpora focus on multi-sentence event extraction. Although

[2] https://catalog.ldc.upenn.edu/LDC2006T06
[3] https://tac.nist.gov/2017/KBP/

tremendous progress has been made in the information extraction area on sentence-level tasks, the emerging document-level datasets raise new challenges. The difficulty mainly comes from the long context semantic representation brought by multiple sentences and the extremely unbalanced label distribution.

2.2 Document-Level Event Argument Extraction

Event extraction can be classified into trigger-word and no-trigger word based extraction, including event detection and event arguments extraction. Many approaches and datasets (Petroni et al., 2018; Hürriyetoğlu et al., 2021; Giorgi et al., 2021; Zavarella et al., 2022) across diverse domains have been proposed for document-level argument extraction to go beyond single-sentence inference. For example, (Ebner et al., 2020) build RAMS to include cross-sentence argument annotations but still limits the arguments around the event in a 5-sentence window. An end-to-end generative transformer (Du et al., 2021) regards argument extraction as a template-filling task. Besides, (Li et al., 2021) formulates the task as conditional generation following event templates, and contributes to the WIKIEVENTS dataset, which consists of only 246 documents with less than one-fourth of annotated cross-sentence arguments. Very recently, (Du et al., 2022) introduce a new global neural generation-based framework by constructing a document memory store to record the contextual event information for improving capability. The largest document-level event extraction dataset is the DocEE (Tong et al., 2022), which consists of $27,000+$ events, $180,000+$ arguments over $27,485$ Wikipedia articles. In this paper, we use the DocEE as base set for our task.

2.3 Few-Shot Argument Extraction

Few-shot learning for information extraction is proposed to tackle such circumstances when only limited instances are annotated. There have been growing interests under few-shot settings for named entity recognition (Ding et al., 2021; Das et al., 2022), and relation extraction (Han et al., 2018; Popovic and Färber, 2022) under single-sentence and document-level scenarios. There has also been research (Deng et al., 2020; Lai et al., 2021; Feng et al., 2020; Lai et al., 2020) for few-shot event extraction within single-sentence. However, to the best of our knowledge, there is no research about document-level few-shot event argument extraction until the submission date.

To fill this gap, this work focuses on few-shot learning for document-level argument extraction. Instead of building a new dataset from scratch, we aim at leveraging the existing supervised dataset for reconstructing the instances by a novel N-Way-D-Doc sampling strategy, inspired by similar work (Sabo et al., 2021; Popovic and Färber, 2022).

3 Task Formulation

3.1 Argument Extraction Definition

The event argument extraction usually depends on the first-stage detected events, but we assume gold event labels to reduce the task complexity in our work. This is opposite to joint extraction where the task is to jointly extract all events and their associated arguments all at once (Sha et al., 2018; Yang and Mitchell, 2016). Since the DocEE dataset (Tong et al., 2022) follows the main event extraction (Hamborg et al., 2018) setting where no trigger words exist and the article title t and the article a itself together determine the event type, we follow their setting and assume the event type e is given, then aim at extracting all related arguments with types R_e.

Formally, given a document $D = \{w_1, ..., w_{|D|}\}$ and its corresponding event type e, where $|D|$ is the total number of words, the event argument extraction aims to detect the boundaries and types for all possible continuous spans $\{w_{start}, w_{end}\}$ in the document D according to event argument types R_e.

3.2 Document-Level Few-Shot Argument Extraction

Following previous work about sentence-level few-shot event detection (Deng et al., 2020), we define the document-level few-shot argument extraction as the following. Given the event instance e, its associated argument types set R_e, the support set S and the query set Q, the few-shot task T is defined as:

$$S = \{..., R_i^s, ...\}, R_i^s = (D_i^s, \{..., (b_i^s, t_i^s), ...)\}$$

$$Q = \{..., R_i^q, ...\}, R_i^q = (D_i^q, \{..., (b_i^q, t_i^q), ...\})$$

$$T = \{S, Q\}$$

where (b_i, t_i) represents the i-th event argument boundaries and type in document D_i in the support

Data Set	# Docs.	# ET.	# AT.	# Tok/Doc	# Sents/Doc	# ArgInst.	#ArgScat.
DocEE	21,450	59	358	678	30.71	109,395	10.2
RAMS	9,124	139	65	105	3.79	21,237	4.8

Table 1: A comparison of DocEE (Li, 2022) and RAMS (Ebner et al., 2020). Docs.: document, ET.: event type, AT.: event argument type, Tok/Doc: tokens per document, Sents/Doc.: sentences per document, ArgInst.: event arguments, ArgScat.: the number of sentences in which event arguments of the same event are scattered.

Types	Train		Dev		Test	
	#ET.	#AT.	#ET.	#AT.	#ET.	#AT.
In domain (small)	30	193	14	87	15	68
In domain (base)	49	303	10	51	10	50
Cross domain	49	302	10	53	10	54

Table 2: The statistics of chosen event and argument types in our three domain split settings.

Avg. Args	In domain (small)		In domain (base)		Cross domain	
	micro	macro	micro	macro	micro	macro
3w1d	4.41	4.23	4.40	4.58	4.47	4.35
3w2d	3.20	3.22	3.57	3.11	3.52	3.03
6w2d	6.56	6.35	7.16	6.28	6.56	6.35

Table 3: The average number of arguments in different settings in the training episodes. The micro average is calculated on average across all episodes on D-doc. The macro average is calculated on average across all argument types on D-doc.

s and query q set. $R_i^{s/q}$ is the set of all the annotated arguments in D_i and S/Q is the combinations of $R_i^{s/q}$ from different documents. Following the episode training for few-shot learning, a task T is one episode aiming to predict all the instances in Q given S. The few-shot learning is usually formalized as an N-Way-K-Shot problem, which means that there are N possible argument types and K supporting instances for each argument type for every task T. Note that the argument types set $R_{\textbf{train}}$ and $R_{\textbf{test}}$ are disjoint.

In practice, given all the support documents in S, we want to extract all the arguments for documents in Q. Since it is not guaranteed each support instance D_i^s contains only one argument, the K in the traditional N-Way-K-Shot setting is no longer guaranteed to be an integer. Previous research (Yang and Katiyar, 2020) tries to use greedy sampling to guarantee the strict K shots requirements for sentence-level few-shot NER task, but this is not applicable due to the sparse density of arguments in the document as also been observed by (Ding et al., 2021). And they loose the K shots requirement to $K{\sim}2K$ shots. However, this is still not realistic under a document-level setting since the arguments spread become even sparser and $K{\sim}2K$ shots are still not guaranteed. It is notable that (Popovic and Färber, 2022) tackle this problem for few-shot document-level relation extraction by D-Doc setting where both N and K are variables between documents and individual episodes. We argue that variable N is not suitable for deploying applications and will make models complicated, so we design a novel N-Way-D-Doc sampling, where N and D are both fixed, resulting in variable K shots.

3.3 Domain Split

To make our task closer to realistic applications, we consider three settings for investigating the difficulty of tasks and model performance. In the *In-domain* circumstance, we manually choose event and event argument types from the same coarse-grained label sets to ensure that they share the same domain knowledge. To explore the meta-learning ability with varying amounts of training bases, we further set *In domain (small)* and *In domain (base)* with a small and medium numbers of event types contained in the training set. The authors (Tong et al., 2022) provide *Cross domain* scenario, where the training and test labels are entirely disjoint, sharing no mutual domain information. We follow their splits for our domain adaptation task.

4 Constructing Few-Shot Episodes

In this section, we talk about the details regarding the dataset conversion and how we obtain episodes for training. For our FewDocAE task, the key components for constructing a realistic few-shot dataset are made of two parts: avoiding data leakage and constructing effective support and query pairs. Generally, to avoid leaking new event arguments information in the training phase we replace all other labels of arguments contained in the validation/test sets with `O`. In this way, it is guaranteed that they have disjoint argument types sets, and this is also close to the realistic scenarios. To provide each episode with a support and query set, we greedy sample instances to make sure they satisfy our N-Way-D-Doc choice. Also, they support and query

Algorithm 1 N-Way-D-Doc Sampling for Few-Shot Episode Generation

Require:
Dataset $\mathcal{X}$, Label set $\mathcal{Y}$, N, D
Ensure:
$\mathcal{S} \leftarrow \varnothing$; // Init the support set
CountN $\leftarrow 0$; //Init the count of entity types;
CountD $\leftarrow 0$; //Init the count of documents;
while CountN $\neq N$ or CountD $\neq D$ **do**
  Randomly sample $(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{X}, \mathcal{Y}$;
  Compute CountN and CountD;
  if CountN $> N$ or CountD $> D$ **then**
    Continue;
  else
    $\mathcal{S} = \mathcal{S} \bigcup (\boldsymbol{x}, \boldsymbol{y})$;
    Update CountN, CountD ;
  end if
end while

instances should always come from different documents.

4.1 Choosing Datasets

We initially consider the two largest document-level event extraction datasets for our tasks: RAMS (Ebner et al., 2020) is annotated in a 5-sentence window around each event trigger and contains $9,124$ annotated events from news based on an ontology of 139 event types and 65 roles. In addition, DocEE includes $21,450$ document-level events with $109,395$ arguments, making it the largest document-level event extraction dataset. Since we aim at building a document-level benchmark, we finally exclude RAMS for its 5-sentence limits and narrow argument types. Eventually, we choose DocEE for our FewDocAE task from their original release [4]. Besides, in order to make sure there are enough examples in each episode for support and query, we exclude all events and argument types that have lower than 2 annotated examples in the train/validation/test set.

4.2 Determining Event Arguments Types

In the released DocEE corpus, there are 31 hard news event types and 28 soft news event types with their corresponding arguments. Their original paper follows no-trigger words design (Nguyen et al., 2016; Zheng et al., 2019) and assumes one main event per document. For arguments, there are 358 event argument types belonging to those 59 event types. The argument annotation is done on the whole document resulting in some documents could contain up to 40 sentences and $7k$ words.

Since different events could still contain the same argument types, we could mask the overlapped arguments in two ways. The first one masks all arguments in the training set with `O` if they also appear in the val and test sets as used in (Ding et al., 2021), while in the second strategy, we can mask all the arguments in the val and test sets if they are shared by the training set. The intuition is that the former one has the risk that the model is trained with `O` types, but is forced to predict their true labels which are not `O` during testing. We believe the latter setting reduces the difficulty by making sure that the new arguments appearing during the prediction stage are not labeled with `O` during training, and use this strategy for all our experiments.

For the **In domain (small)** setting, we manually choose 14 and 15 event types for constructing the validation and test episodes, while leaving all other 32 event types for training purposes. This results in disjoint event sets for train/val/test, but they share some domain knowledge. Since the number of event types used for training is only about two times that of testing, we believe this setting is more challenging. The intuition is that for few-shot learning, we aim to get a good feature extractor during training using massive base data so that those learned features could benefit the model by predicting novel instances. We use this setting for exploring the few-shot learning limitation when the training base is small. For the **In domain (base)** setting, we use a larger training base with 49 event types and adopt the remaining 10 event types for validation and testing. Again, our choice of event splitting guarantees they share in domain event types. In contrast to previous situation, we now aim at investigating the full capability of few-shot learning given enough training base.

For **Cross domain**, we follow the original event splits in DocEE where the authors choose the natural disasters events as the target domain, including `Floods`, `Droughts`, `Earthquakes`, `Insect Disaster`, `Famine`, `Tsunamis`, `Mudslides`, `Hurricanes`, `Fire`, and `Volcano Eruption`, and leave the remaining 49 event types as source domains.

Besides, there are six arguments, including `Date`, `Causes`, `Areas affected`, `Location`,

[4] https://github.com/tongmeihan1995/DocEE

Casualties, and Losses that occur frequently in all the splits. To fully leverage the capability of meta-learning during the training phase, we leave those arguments in the training set only and mask them as O in the val and test sets. The full statistics of our domain split result is shown in Table 2.

4.3 Sampling Strategy

The traditional N-Way-K-Shot sampling for few-shot learning fails when applied to our document-level settings. The reason is that one document with the sparse spread of event arguments could contain one or many arguments. Thus greedy sampling K shots instance can not be guaranteed. Soft sampling methods like N-Way-$K{\sim}2K$ Shots in (Ding et al., 2021) still do not work for our long documents since $K{\sim}2K$ Shots are still hard to be satisfied. Note that (Popovic and Färber, 2022) adopts D-doc sampling for document-level relation extraction. However, their approach can not guarantee a fixed number of N classes. For example, their 1-Doc results in 2.18-Way-2.36-Shot instances on average. We argue that variable N is not ideal for our task.

We introduce a new sampling strategy, N-Way-D-Doc sampling as shown in Algorithm 1, to guarantee a fixed number of N classes within D documents. In our approach, we first pick N event argument types, then randomly sample D documents and keep all the arguments within N types in the D documents unchanged while discarding other argument types. This process is not finished until there are exactly N classes within D documents. Besides, few-shot learning is notoriously famous for lacking challenging-yet-realistic testing setups and failing to employ careful experimental design (Bragg et al., 2021). To make our benchmark more robust, we sample episodes by making sure they evenly come from combinations of different documents, events, and argument types.

4.4 Sampling Results

We sample around $30k{\sim}50k$ episodes for the training and around $3k$ episodes for validation and testing sets under 3-Way-2-Doc and 6-Way-2-Doc settings. For the 3-Way-1-Doc setting, since not all documents have at least three unique arguments, we sample about $30k$ for training and $1.5k$ for Val and test. We also show the argument type characteristics of our various N-Way-D-Doc sampling results in Table 3. As we can see, the average sampled number of arguments for both settings is close. For the 3-Way-1-Doc settings, there are around $4.47/3 = 1.5$ arguments per type, while only around 1.14 and 1.10 arguments per type for other settings, demonstrating the diversity of our splitted domains.

4.5 N-Way-D-Doc Choice

Due to the unique nature of few-shot document-level extraction, we only test limited combinations of N-way and D-doc for two reasons. First, there are only limited documents that simultaneously contain the same N argument types, thus we can not find enough samples when extending to more than 6-way. Second, we only keep 1/2-doc for the two reasons: on the one hand, extending to more documents requires doing sequence labelling on much longer documents that consumes much more GPU memory that we struggle to handle even by a 40G GPU. Similar memory issue has also been reported by (Sabo et al., 2021) when handling few-shot learning for relation extraction. On the other hand, adding more documents will also increase the number of NOTA argument types, which will be detrimental for extracting useful real arguments since the vast majority would be None types. Considering the current settings already result in relatively low performance, we do not aim to further increase the challenge.

5 Models

Previous work on document-level EE using BERT_Seq (Du and Cardie, 2020; Tong et al., 2022) demonstrate the success of using a pre-trained BERT model to sequentially label words in the article. And the superior performance of the long document transformer (e.g. Longformer (Beltagy et al., 2020)) has also been proven to improve the argument extraction task (Tong et al., 2022). We thus follow their baseline settings and use BERT or Longformer as document encoders for our task. In order to adapt to our few-shot setting, inspired by the successful applications of the prototypical network (ProtoNet) (Snell et al., 2017) for meta-learning, we assume there exists one prototypical representation for each argument type. Then we implement the models by extending ProtoNet to language encoder with token-level similarity.

Model	Baseline			ProtoNet-BERT			ProtoNet-LongFormer			ProtoNet-MNAV		
	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1 [%]
3-Way-1-Doc	0.88	2.50	1.30	3.31 ± 1.44	15.52 ± 1.45	5.35 ± 1.87	4.83 ± 0.39	15.67 ± 2.66	$\mathbf{7.39 \pm 0.65}$	4.46	11.42	6.42
3-Way-2-Doc	1.22	12.99	2.23	4.77 ± 0.54	15.2 ± 0.56	7.26 ± 0.49	5.88 ± 1.58	14.65 ± 0.39	$\mathbf{8.34 \pm 1.59}$	5.12	13.25	7.38
6-Way-2-Doc	0.64	7.58	1.19	4.42 ± 0.45	11.59 ± 0.46	6.37 ± 0.24	6.17 ± 0.54	14.93 ± 0.41	$\mathbf{8.73 \pm 1.03}$	5.73	15.08	8.30

Table 4: *In domain (small)* results for FewDocAE argument extraction task under three settings.

5.1 Document Encoder

We adopt Longformer (led-base-16384[5]) and BERT (bert-base-uncased[6]) as our encoder for all the experiments. In order to handle long sequences, we split all inputs with a chunk length of 1024 and 512, respectively.

Formally, suppose the document $D = \{w_1, ..., w_{|D|}\}$ where w_i represents the ith token and $|D|$ is the maximum length. By feeding the tokens into the document encoder, we get the contextualized token representation:

$$[h_1, ..., h_{|D|}] = Encoder([w_1, ..., w_{|D|}])$$

5.2 Prototypical Networks (ProtoNet)

The ProtoNet approach (Snell et al., 2017) assumes there exists one prototypical representation for each argument class and learns a metric space where categorization is performed by labeling each query term with the value calculated from the distance between prototype representations that are closest to it. In practice, we use the average representation of all tokens in each argument to represent the contextualized representation of that argument type. Formally, for support set S_{r_i} containing all the arguments of type r_i, following previous step h_t is the representation of token t. By calculating all the prototypes,

$$r_i = \frac{\sum_{t \in S_{r_i}} h_t}{|S_{r_i}|}$$

we get the argument representation set $R = \{r_1, ..., r_N, r_{N+1}\}$ where r_i represent the ith argument prototype and r_{N+1} represents the O type. Then for the query token h_q, the target label is assigned as

$$label = \arg\min_i(d(r_i, h_q))$$

. We use the L2 distance as the distance metric.

Besides, we also add a **Baseline** model based on ProtoNet-LongFormer without any finetuning. The reason is that training on many episodes is very costly as will be shown below. And we are interested on the original performance and how much benefits the finetuning could bring in.

5.3 Nearest Neighbor Tagger (NNShot)

NNShot (Yang and Katiyar, 2020) is a simple but strong system based on token-level nearest neighbor classification for the few-shot sequence tagger task. It first obtain contextual representations for all tokens in their respective documents. Then it assigns the token q a tag r_i corresponding to the most similar token in the support set:

$$label = \arg\min_i(\arg\min_{i' \in S_{r_i}}(d(r'_i, h_q)))$$

where S_{r_i} represents the set of support tokens with r_i tags.

However, simply extending the sentence-level NNShot to a long document is not plausible. For example, for two documents both with $4,096$ tokens and each token has an embedding dimension of 768, the token-level similarity computation requires more than 50GB GPU memory. To make token similarity calculation of long inputs possible, we use another linear layer to reduce the 768 dimensional representation into 32 dimension and use the L1 distance as the distance metric for computational efficiency.

5.4 ProtoNet-MNAV

Here we adjust the Multiple NOTA(None-Of-the-Above) Vectors(MNAV) proposed by (Sabo et al., 2021) for few-shot relation extraction to our FewDocAE since they both face the same issue that the majority labels belong to NOTA. Instead of initializing the NONA vectors by randomly as by (Sabo et al., 2021) or from sampled support sets as in (Popovic and Färber, 2022) and then gradually update them, we adopt a K-means MNAV strategy. Specifically, we perform a K-means clustering for all NOTA representations where K is set to be a hyperparameter. Then for ProtoNet-based models, we still determine the label type by calculating their token similarity. And we attribute all token nearest to the K NOTA vectors as NOTA type. This

[5]https://huggingface.co/allenai/led-base-16384
[6]https://huggingface.co/bert-base-uncased

way we ideally reduce the risk of only representing many NOTA token by one vector as also pointed by (Allen et al., 2019), since there might be multiple NOTA Prototypes. We always use LED as the encoder for ProtoNet-MNAV method.

6 Experiments and Results

We show the experimental results and analysis in this section. More experimental details and configurations can be found in Appendix A. In general, one experiment costs around 15 hours and in total there are more than 100 runs, which also limits more advanced models that we can choose.

In Domain For the *In domain* setting where the training and test examples are sampled from the same domain, the performance is exhibited in Table 4 and Table 5 for the *In domain (small)* and *In domain (base)* results, respectively. As we can see, the baseline without finetuning can not guarantee good performance, and the ProtoNet-Longformer models consistently outperform the ProtoNet-BERT by a large margin under three N-Way-D-Doc settings. On the one hand, this gap can be explained by the superior encoding ability of long documents by LongFormer. On the other hand, the results convince our motivation by extending to document-level argument extraction as a large portion of arguments can only be extracted across sentences, as also confirmed under supervised condition by (Tong et al., 2022). Besides, we also observe a significant performance drop when moving from *In domain (base)* to *In domain (small)* under most results, which clearly manifests the benefits of using training base with broader event types. This validates the intuition that more diverse training base can help train a better argument feature extractor. Also, we can see that as we increase the Ways of N and Docs of D, the overall results continue increasing, demonstrating instances with more ways and more Docs are easier to be predicted. However, ProtoNet-MNAV does not bring performance gain as expected, possibly due to the more unclear decision boundary as illustrated in Appendix C.2. In general, the current models remain relatively low results, demonstrating the challenges of our FewDocAE task in document-level. We additionally show some case study in Appendix C.

Cross Domain The overall arguments extraction results are shown in Table 6 for the *Cross domain* setting. Compared with its *In domain (base)* counterpart, the performance degradation is witnessed under both ProtoNet-Longformer and ProtoNet-BERT models in all settings. This is expected since we split the cross domain to avoid the in domain knowledge and such domain adaptation is more challenging.

We also test the performance using NNShot-Longformer in Table 7 under *Cross domain*. To be compatible with computation memory limits brought by NNShot token-level similarity calculation, we do not take whole documents and split the documents within chunks of 1024 tokens. For the 1-Doc setting, an average F1 score of 5.9%, and for the $3/6$-Way-2-Doc settings, F1 of 3.73% and 4.59% points are observed using NNShot model. The performance decreases a lot by NNShot, which we attribute to the likely underrepresented representation by dimension reduced operation. However, due to the memory limits, we leave more investigation for further research. Besides, how to efficiently adapt NNShot to long documents is also an open challenge.

Overall Analysis We attribute the observed results to three main reasons: First, the document-level argument extraction involves reasoning over a much longer context compared to the sentence-level. Actually, the average number of sentences for DocEE is 30.71 as pointed out in Table 1, which dramatically increases the difficulty of effective encoding over long documents. Even though Longformer can capture attention over long sentences, the ability is still limited. On the other hand, document-level argument extraction faces the new challenge of extremely unbalanced label distribution with more than 95% of label `O`. Compared to its sentence-level counterparts, where the label unbalance already degrades the performance, the document-level sparse distribution of arguments further exacerbates the unbalanced distribution. The majority of `O` labels make it difficult for the model to learn a good representation of arguments among the representation space. Besides, to make the few-shot learning close to a realistic setting, we follow the long-tail arguments distribution of the original DocEE dataset. This extremely unbalanced setting is a good testbed for validating the model ability due to its similar distribution of many real-world few-shot problems, as also pointed by (Sabo et al., 2021). As the results suggest, the long-tail distribution makes it hard for models to

Model	Baseline			ProtoNet-BERT			ProtoNet-LongFormer			ProtoNet-MNAV		
	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1 [%]
3-Way-1-Doc	0.91	5.95	1.58	2.90 ± 0.52	14.12 ± 1.11	4.80 ± 0.78	6.55 ± 0.52	19.13 ± 0.51	$\mathbf{9.76 \pm 0.56}$	5.08	14.49	7.52
3-Way-2-Doc	1.91	7.92	3.08	4.87 ± 0.57	23.05 ± 0.18	8.03 ± 0.76	7.92 ± 0.58	20.89 ± 0.93	$\mathbf{11.48 \pm 0.89}$	7.13	20.35	10.56
6-Way-2-Doc	3.19	14.77	5.25	2.25 ± 0.85	14.32 ± 1.76	4.25 ± 1.61	8.21 ± 0.49	22.76 ± 0.66	$\mathbf{12.07 \pm 0.48}$	8.54	17.68	11.52

Table 5: *In domain (base)* results for FewDocAE argument extraction task under three settings.

Model	Baseline			ProtoNet-BERT			ProtoNet-LongFormer			ProtoNet-MNAV		
	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1 [%]
3-Way-1-Doc	1.12	8.00	1.96	1.94 ± 0.01	12.45 ± 0.02	3.36 ± 0.01	5.65 ± 0.78	19.38 ± 1.52	$\mathbf{8.70 \pm 0.75}$	5.34	17.95	8.23
3-Way-2-Doc	1.03	11.22	1.90	2.81 ± 0.25	22.25 ± 0.34	5.34 ± 0.35	7.09 ± 0.87	20.52 ± 0.90	$\mathbf{10.51 \pm 0.95}$	6.18	19.44	9.38
6-Way-2-Doc	0.73	7.86	1.34	1.67 ± 0.47	17.03 ± 6.37	3.05 ± 0.87	5.91 ± 0.30	17.58 ± 0.20	$\mathbf{8.84 \pm 0.31}$	5.36	16.89	7.49

Table 6: *Cross domain* results for FewDocAE argument extraction task under three settings.

Settings	NNShot-LongFormer			
	val	test		
	F_1	Precision	Recall	F_1 [%]
3-Way-1-Doc	6.02 ± 0.89	3.84 ± 0.67	12.75 ± 0.78	5.9 ± 0.93
3-Way-2-Doc	3.34 ± 0.43	2.59 ± 0.40	6.60 ± 0.43	3.73 ± 0.46
6-Way-2-Doc	4.87 ± 0.68	3.21 ± 0.66	8.07 ± 0.67	4.59 ± 0.86

Table 7: *Cross domain* results for FewDocAE argument detection task under three settings by NNShot model.

	In domain (base)			Cross domain		
	3w1d	3w2d	6w2d	3w1d	3w2d	6w2d
FP	3.68	2.45	4.86	6.42	3.23	4.40
FN	1.99	1.26	1.66	2.60	0.87	1.43

Table 8: The FP means a token with true label `O` is misclassified as a part of an argument, while the FN represents a token within a certain argument is misclassified as `O`.

uniformly focus on all labels. Third, due to the high GPU memory and computation requirements brought by long documents, we only aim at providing benchmark baselines results. More advanced methods might help, but we leave it for future work.

Error Analysis Finally, we report the false positive (FP) and false negative (FN) scores in Table 8 for two domain using ProtoNet-Longformer. The overall FP results demonstrate that most `O` are correctly predicted. However, considering the large number of `O` labels compared with real arguments, even a small portion of FP still leads to a large performance drop of final results. As for real arguments prediction, the main misclassification errors come from assigning one argument type to another type considering FN is low.

7 Conclusion

In order to handle new emerging event arguments with limited annotations and adapt it to the real-world document-level scenario, we propose FewDocAE benchmark to advance the research of few-shot learning for document-level event argument extraction. We conduct comprehensive experiments by extending previous models into our task under in-domain and cross-domain. Our experiments confirm the necessity of moving to document level by showing that current models still witnesses suboptimal performance. We also demonstrate the benefits of using a more diverse training base to learn a good argument feature extractor. The current results show that FewDocAE is challenging due to the long document and limited examples, as well as the intrinsic charisma of few-shot learning. The relatively low extraction score illustrates the difficulty of this novel task, in the meanwhile it also provides new chances for advancing this field. In summary, we hope FewDocAE shed new light on a more realistic but challenging setting for event argument extraction. In the future, we hope to investigate more advanced methods for solving this problem.

8 Limitations

As we mentioned, we only focus on event arguments with the assumption that event type is already provided. However, this is not always true for many applications in real life scenarios. But it would be out of the scope of this work to combine them together, so we leave it for future work. Besides, considering the long input of document-level extraction, the computing memory consumption significantly increase to tens of times compared with its sentence-level counterpart. We only consider the 1/2-Doc cases, although in reality more docs are possible. We believe finding a solution for decreasing the memory requirements would be of great impact for future research in this direction.

9 Acknowledgments

Xianjun Yang was supported by the UC Santa Barbara NSF Quantum Foundry funded via the Q-AMASEi program under NSF award DMR1906325.

References

Kelsey Allen, Evan Shelhamer, Hanul Shin, and Joshua Tenenbaum. 2019. Infinite mixture prototypes for few-shot learning. In *International Conference on Machine Learning*, pages 232–241. PMLR.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *CoRR*, abs/2004.05150.

Jonathan Bragg, Arman Cohan, Kyle Lo, and Iz Beltagy. 2021. Flex: Unifying evaluation for few-shot nlp. *Advances in Neural Information Processing Systems*, 34.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Sarkar Snigdha Sarathi Das, Arzoo Katiyar, Rebecca J Passonneau, and Rui Zhang. 2022. Container: Few-shot named entity recognition via contrastive learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6338–6353.

Shumin Deng, Ningyu Zhang, Jiaojian Kang, Yichi Zhang, Wei Zhang, and Huajun Chen. 2020. Meta-learning with dynamic-memory-based prototypical network for few-shot event detection. In *Proceedings of the 13th International Conference on Web Search and Data Mining*, pages 151–159.

Ning Ding, Guangwei Xu, Yulin Chen, Xiaobin Wang, Xu Han, Pengjun Xie, Haitao Zheng, and Zhiyuan Liu. 2021. Few-nerd: A few-shot named entity recognition dataset. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3198–3213.

George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, and Ralph Weischedel. 2004. The automatic content extraction (ACE) program – tasks, data, and evaluation. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Nanqing Dong and Eric P Xing. 2018. Few-shot semantic segmentation with prototype learning. In *BMVC*, volume 3.

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

Xinya Du, Sha Li, and Heng Ji. 2022. Dynamic global memory for document-level argument extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5264–5275.

Xinya Du, Alexander M Rush, and Claire Cardie. 2021. Template filling with generative transformers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 909–914.

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.

Li Fei-Fei, Robert Fergus, and Pietro Perona. 2006. One-shot learning of object categories. *IEEE transactions on pattern analysis and machine intelligence*, 28(4):594–611.

Rui Feng, Jie Yuan, and Chao Zhang. 2020. Probing and fine-tuning reading comprehension models for few-shot event extraction. *CoRR*, abs/2010.11325.

Tianyu Gao, Xu Han, Zhiyuan Liu, and Maosong Sun. 2019. Hybrid attention-based prototypical networks for noisy few-shot relation classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6407–6414.

Tao Ge, Lei Cui, Baobao Chang, Zhifang Sui, Furu Wei, and Ming Zhou. 2018. Eventwiki: a knowledge base of major events. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Salvatore Giorgi, Vanni Zavarella, Hristo Tanev, Nicolas Stefanovitch, Sy Hwang, Hansi Hettiarachchi, Tharindu Ranasinghe, Vivek Kalyan, Paul Tan, Shaun Tan, Martin Andrews, Tiancheng Hu, Niklas Stoehr, Francesco Ignazio Re, Daniel Vegh, Dennis Atzenhofer, Brenda Curtis, and Ali Hürriyetoğlu. 2021. Discovering black lives matter events in the United States: Shared task 3, CASE 2021. In *Proceedings of the 4th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2021)*, pages 218–227, Online. Association for Computational Linguistics.

Ralph Grishman and Beth Sundheim. 1996. Message Understanding Conference- 6: A brief history. In *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*.

Felix Hamborg, Soeren Lachnit, Moritz Schubotz, Thomas Hepp, and Bela Gipp. 2018. Giveme5w: Main event retrieval from news articles by extraction of the five journalistic W questions. In *Transforming Digital Worlds - 13th International Conference, iConference 2018, Sheffield, UK, March 25-28, 2018, Proceedings*, volume 10766 of *Lecture Notes in Computer Science*, pages 356–366. Springer.

Xu Han, Hao Zhu, Pengfei Yu, Ziyun Wang, Yuan Yao, Zhiyuan Liu, and Maosong Sun. 2018. Fewrel: A large-scale supervised few-shot relation classification dataset with state-of-the-art evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4803–4809.

Ali Hürriyetoğlu, Osman Mutlu, Erdem Yörük, Farhana Ferdousi Liza, Ritesh Kumar, and Shyam Ratan. 2021. Multilingual protest news detection - shared task 1, CASE 2021. In *Proceedings of the 4th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2021)*, pages 79–91, Online. Association for Computational Linguistics.

Viet Lai, Franck Dernoncourt, and Thien Huu Nguyen. 2021. Learning prototype representations across few-shot tasks for event detection. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5270–5277.

Viet Dac Lai, Thien Huu Nguyen, and Franck Dernoncourt. 2020. Extensively matching for few-shot learning event detection. In *Proceedings of the First Joint Workshop on Narrative Understanding, Storylines, and Events*, pages 38–45.

Juanzi Li. 2022. Docee: A large-scale and fine-grained benchmark for document-level event extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 8057–8077.

Sha Li, Heng Ji, and Jiawei Han. 2021. Document-level event argument extraction by conditional generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 894–908.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Kiem-Hieu Nguyen, Xavier Tannier, Olivier Ferret, and Romaric Besançon. 2016. A dataset for open event extraction in english. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1939–1943.

Yingwei Pan, Ting Yao, Yehao Li, Yu Wang, Chong-Wah Ngo, and Tao Mei. 2019. Transferrable prototypical networks for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2239–2247.

Fabio Petroni, Natraj Raman, Tim Nugent, Armineh Nourbakhsh, Žarko Panić, Sameena Shah, and Jochen L Leidner. 2018. An extensible event extraction system with cross-media event resolution. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 626–635.

Nicholas Popovic and Michael Färber. 2022. Few-shot document-level relation extraction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 5733–5746. Association for Computational Linguistics.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Hwee Tou Ng, Anders Björkelund, Olga Uryupina, Yuchen Zhang, and Zhi Zhong. 2013. Towards robust linguistic analysis using ontonotes. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 143–152.

Alan Ritter, Oren Etzioni, and Sam Clark. 2012. Open domain event extraction from twitter. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1104–1112.

Ofer Sabo, Yanai Elazar, Yoav Goldberg, and Ido Dagan. 2021. Revisiting few-shot relation classification: Evaluation data and classification schemes. *Transactions of the Association for Computational Linguistics*, 9:691–706.

Erik Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

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 AAAI Conference on Artificial Intelligence*, volume 32.

Jake Snell, Kevin Swersky, and Richard Zemel. 2017. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30.

Shengli Sun, Qingfeng Sun, Kevin Zhou, and Tengchao Lv. 2019. Hierarchical attention prototypical networks for few-shot text classification. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 476–485.

Meihan Tong, Bin Xu, Shuai Wang, Yixin Cao, Lei Hou, Juanzi Li, and Jun Xie. 2020. Improving event detection via open-domain trigger knowledge. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5887–5897.

Meihan Tong, Bin Xu, Shuai Wang, Meihuan Han, Yixin Cao, Jiangqi Zhu, Siyu Chen, Lei Hou, and Juanzi Li. 2022. Docee: A large-scale and fine-grained benchmark for document-level event extraction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3970–3982.

Runxin Xu, Peiyi Wang, Tianyu Liu, Shuang Zeng, Baobao Chang, and Zhifang Sui. 2022. A two-stream amr-enhanced model for document-level event argument extraction. *arXiv preprint arXiv:2205.00241*.

Bishan Yang and Tom Mitchell. 2016. Joint extraction of events and entities within a document context. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 289–299.

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.

Yi Yang and Arzoo Katiyar. 2020. Simple and effective few-shot named entity recognition with structured nearest neighbor learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6365–6375.

Yuan Yao, Deming Ye, Peng Li, Xu Han, Yankai Lin, Zhenghao Liu, Zhiyuan Liu, Lixin Huang, Jie Zhou, and Maosong Sun. 2019. Docred: A large-scale document-level relation extraction dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 764–777.

Quan Yuan, Xiang Ren, Wenqi He, Chao Zhang, Xinhe Geng, Lifu Huang, Heng Ji, Chin-Yew Lin, and Jiawei Han. 2018. Open-schema event profiling for massive news corpora. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pages 587–596.

Vanni Zavarella, Hristo Tanev, Ali Hürriyetoğlu, Peratham Wiriyathammabhum, and Bertrand De Longueville. 2022. Tracking COVID-19 protest events in the United States. shared task 2: Event database replication, CASE 2022. In *Proceedings of the 5th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE)*, pages 209–216, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D Manning. 2017. Position-aware attention and supervised data improve slot filling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 35–45.

Shun Zheng, Wei Cao, Wei Xu, and Jiang Bian. 2019. Doc2edag: An end-to-end document-level framework for chinese financial event extraction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 337–346.

A Experiments

A.1 Evaluation Metrics

Since we treat this event arguments extraction as a sequence labeling task, we employ `IO` notation, where all tokens within an argument type are labeled as `I-type` while all other tokens are labeled as `O`. Besides, we report all the performance on non `O` types. The prediction is only considered as correct when all tokens within that argument are correctly classified. We use macro precision, recall, and F1 score to measure the performance.

A.2 Experimental Configuration

For all transformer-based models, we employ the released model from HuggingFace[7] and set the learning rate to $[1e-4, 1e-5, 5e-5]$, of which $1e-5$ is the best parameter except for 3w1d. For 3w1d tasks, we use a learning rate of $1e-6$, otherwise, it will not converge. We try the different batch size of $[1, 2, 4, 6]$, of which 4 and 6 does not lead to converging, 2 achieves the best performance. We use AdamW (Loshchilov and Hutter, 2019) as the optimizer and gradient clipping of 1.0. We use the NLTK[8] package for sentence tokenization. All models can be put into four NVIDIA A60 GPUs with an RAM of 48GB each. The training procedure takes around 15 hours and 10 hours for 60k iterations to complete for 2-Doc and 1-Doc settings, respectively. We report the mean and variance for all experiments under two random seeds, except only one run for ProtoNet-MNAV. The validation is done by every 4k interaction on the training episodes, and we use the best checkpoint from the validation results for testing. The number of query instances during testing is always set to 1. For ProtoNet-MNAV, we tried the hyperparameter K ranging from 2 to 6 and did not observe obvious difference. We report the results based on K=6.

B Memory and Computation Issues

Due to the length distribution of this document-level task, we set at least 1,024 tokens as chunk

[7] https://huggingface.co/models

[8] https://www.nltk.org/

length for input to LED-based models. However, a 2-Doc setting results in 4-doc documents coming from both the query and support sets. Considering the argument-extraction is conducted at every token level, the similarity score calculation imposes a severe memory issues for adapting more complicated methodologies. We believe that this is a big challenge for few-shot document-level tasks, which is not only a issue for small language models (Sabo et al., 2021) but also true for large models like GPT-3 (Brown et al., 2020) and leave more exploration for future work.

C Case-study

C.1 Predictions vs. True

Here in Figure 2 and Figure 3 we show two example case study of how our predictions differ from the true labels. The different color corresponds to different event argument types. And we also highlight whether the predictions are accurate or not. All examples are drawn from the test set in in-domain(samll) setting and the predictions are always the ProtoNet-LongFormer model. As we can see, the majority of predictions are wrong with only a few exceptions in Fig. 2. Besides, in Fig. 3 we additionally show the false positive results which also accounts a large protion for the final performance. In general, the current model can not well handle the document-level predictions under few-shot setting, and the prototypical representation for different labels still struggle with token classification.

C.2 Visualizations

Here we plot the two dimensional t-SNE[9] projection of the prototypical embeddings in Fig. 5 and Fig. 4. Ideally 7 or 12 clusters as expected for ProtoNet-LongFormer and ProtoNet-MNAV models. But the results show that the clusters are not always well seperated from each other, which might explains the reason why we still get low performance. Actually, we can also clearly see some clusters indeed include many same argument types, like class 0, 1 and 6 in Fig. 4, but these clusters still spread across the many different locations. For example, we can clearly see 3 big clusters for 6 and 5 big clusters for 1, which indicts that multiple prototypes exist for different argument types. However, when we extend to multiple NONA vectors by ProtoNet-MNAV model, we even see a little performance drop as mentioned in Section 6. Ideally multile NONA vectors can better represent more diverse NONA class but as we can observe from Fig. 5 that the multiple NONA vectors(0 to 5) actually make the overall cluster boundary more obscure. The resulting clusters become even more difficult to be classified.

In general, the current model succeed performing classification for part of the arguments but still fails generating well represented representations for some difficult cases and thus leads to suboptimal resutls.

[9]https://scikit-learn.org/stable/index.html

Support set

Doc1: A Bronx teen was cut loose by a judge in an armed - robbery shooting over prosecutors ' objections — only to proceed to allegedly murder[Charges of imprisonment] an " innocent kid " in a botched gang hit , The Post has learned . Steven Mendez[people Released] , 17 — who was once even busted for allegedly pulling a gun on his own mom — could have been kept behind bars for up to four years[Prison Term] after pleading guilty in the violent armed robbery[Charges of imprisonment] in 2020 , according to officials and law - enforcement sources . ... Steven Mendez was charged with allegedly gunning[Charges of imprisonment] down Saikou Koma on Oct 24[Jail time] . … Mendez had been indicted last year on first - degree assault , first - degree robbery and felony gun - possession[Charges of imprisonment] charges stemming from his role as an accomplice in the July 17 , 2020 , armed robbery in The Bronx . Saikou Koma , 21 , was killed on Oct. 24 . … Steven Mendez allegedly murdered Saikou Koma in a botched gang hit . William Miller Judge Denis Boyle[Judge] granted Steven Mendez probation in May . …. The teen , Alberto Ramirez[People Released] , was bailed out by his family after Boyle reduced his bail from $ 75,000 to $ 10,000 — leaving him free to allegedly gun down [Charges of imprisonment] 34 - year - old dad Eric Velasquez . Last year , Boyle also released Jordon Benjamin[People Released] , 16 , who was facing a manslaughter[Charges of imprisonment] charge , only to have the teen allegedly slash a young woman . And in 1999 , Boyle agreed to a deal that allowed a homeless man to live in a shelter while he awaited sentencing in an attempted sex attack . The suspect , Ishmael Holmes , 22 , was then implicated in a rash of sexual attacks[Charges of imprisonment] on the Upper East Side . Koma 's heartbroken mother , railed that the judge has one job . …

Doc2: A man who had terminal cancer and was released from prison more than a decade early so he could be with his family has died , officials said . Danny Ray Aria[People released] died Wednesday , an official said . He was 49 . Aria served 22 years of a 42 – year[Prison Term] sentence after being convicted of multiple charges for an incident[Charges of imprisonment] that court documents say was fueled by drugs and alcohol . … Earlier this month , Hamilton County Common Pleas Judge Jennifer Branch[Judge] released Aria from prison to home detention – a decision that was strongly criticized by Prosecutor Joe Deters . He viewed it as part of a shift at the courthouse , saying some judges were focusing on defendants instead of victims . Aria 's colon cancer[Reason for release] had spread to other organs ...

Query set

Query: TYLER , Texas (KETK) – The lead defense attorney for William Davis , the former CHRISTUS nurse who was sentenced to death[Prison Term] for killing several patients , was allegedly caught trying to solicit a prostitute[Charges of imprisonment] during the month - long trial . Former CHRISTUS nurse convicted of murder[Charges of imprisonment] moved to death row unit in Livingston According to judicial records , Phillip Hayes[People Released] was arrested Friday , Nov. 5 , and released after posting a $ 2,000 bond[Reason for release] . …by a Smith County jury[Judge] on Oct. 19 and he was sentenced to death on Oct. 27 . A call to Smith County District Attorney Jacob Putman 's office for comment has not been returned as of this writing .

Predictions: death[Prison Term]: ✅, solicit a prostitute[People Released]: ❌, murder[People Released]: ❌, Phillip Hayes[People Released]: ✅, a[Prison Term] Smith County jury[People Released]: ❌

Figure 2: A case study of a 6-Way-2-Doc episode consisting of a support and query set. For simplicity, we only show part of the documents hereinafter. Best viewed in color.

Support set

Doc1: Smith[People Released] was convicted of second - degree murder[Charges of imprisonment] in 1994[Jail time] for luring Derrick Robie into woods near the younger boy 's home and striking his head with a rock . … READ MORE : Robert Durst sentenced to life for murder of best friend Smith 's lawyer unsuccessfully argued that he was mentally ill . Smith was sentenced to nine years to life[Prison Term] in prison . …

Doc2: Prosecutors have argued emphatically against Larry[People Released] getting released on bail[Reason for release]. In a document titled " People 's Request to Deny Bail " filed Oct. 21 , the district attorney 's office laid out its case for keeping Millete in custody . …' Lot of cockroaches and rats ' : Neighbors fed up with California hoarder home In court Thursday , Deputy District Attorney Christy Bowles[Judge] primarily reiterated the arguments laid out in the filings . " He 's a flight risk , " Bowles said , adding , " there 's a public safety concern . "…

Query set

Query: / Updated : Nov 9 , 2021 / 06:08 PM PST A man convicted of murder[Charges of imprisonment] in a 2015[Jail time] case is set to be released after a new state law[Reason for relsease] made him eligible to be retroactively sentenced as a juvenile offender . A lawyer for a victim 's family in the case said a Los Angeles County prosecutor purposefully did not call any witnesses at a crucial hearing . But in a statement to KTLA Tuesday , the DA 's Office said there were no witnesses who could testify in the hearing , and that the judge was " left with no legal option other than to terminate juvenile jurisdiction . " Andrew Cachu[People Released] was originally sentenced to 50 years to life[Prison Term] in prison after being tried and convicted as an adult in the Palmdale homicide of Luis Amela . The defendant was 17 at the time , but he has only served six years of his term . Kathleen Cady , an attorney representing Amela 's family said that the new state law , as well as District Attorney George Gascón 's directive to end the practice of sending youth to the adult court system , " is just not justice . "

Predictions: murder[Reason for release]: ❌, 2015[Reason for release]: ❌, a new state law[Charges of imprisonment]: ❌, Andrew Cachu[Judge]: ❌, 50 years to life[Prison Term]: ✅, (False positive: Nov 9 , 2021 / 06:08 PM[Jail time], A[Charges of imprisonment], man[Reason for release], convicted of[Charges of imprisonment], retroactively sentenced [Prison Term], Los[Jail time] Angeles[Reason for release], juvenile jurisdiction[Prison Term], prison[Prison Term], Kathleen [Judge], Attorney George Gascón [Judge])

Figure 3: Another case study of a 6-Way-2-Doc episode consisting of a support and query set. Best viewed in color.

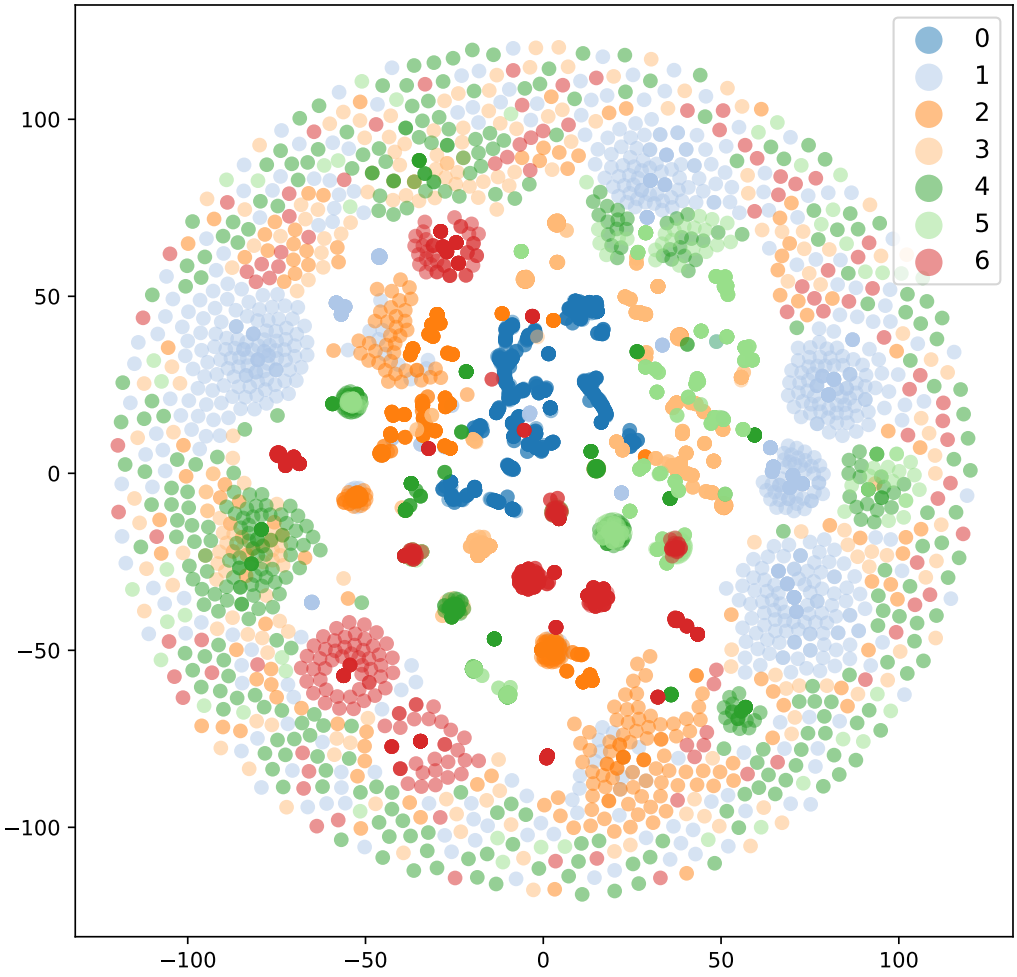

Figure 4: Visualizing prototypical feature vectors using t-SNE. 1 NONA labels(0) and 6 argument types (1 to 5) from trained in-domain(small) checkpoints. Best viewed in color.

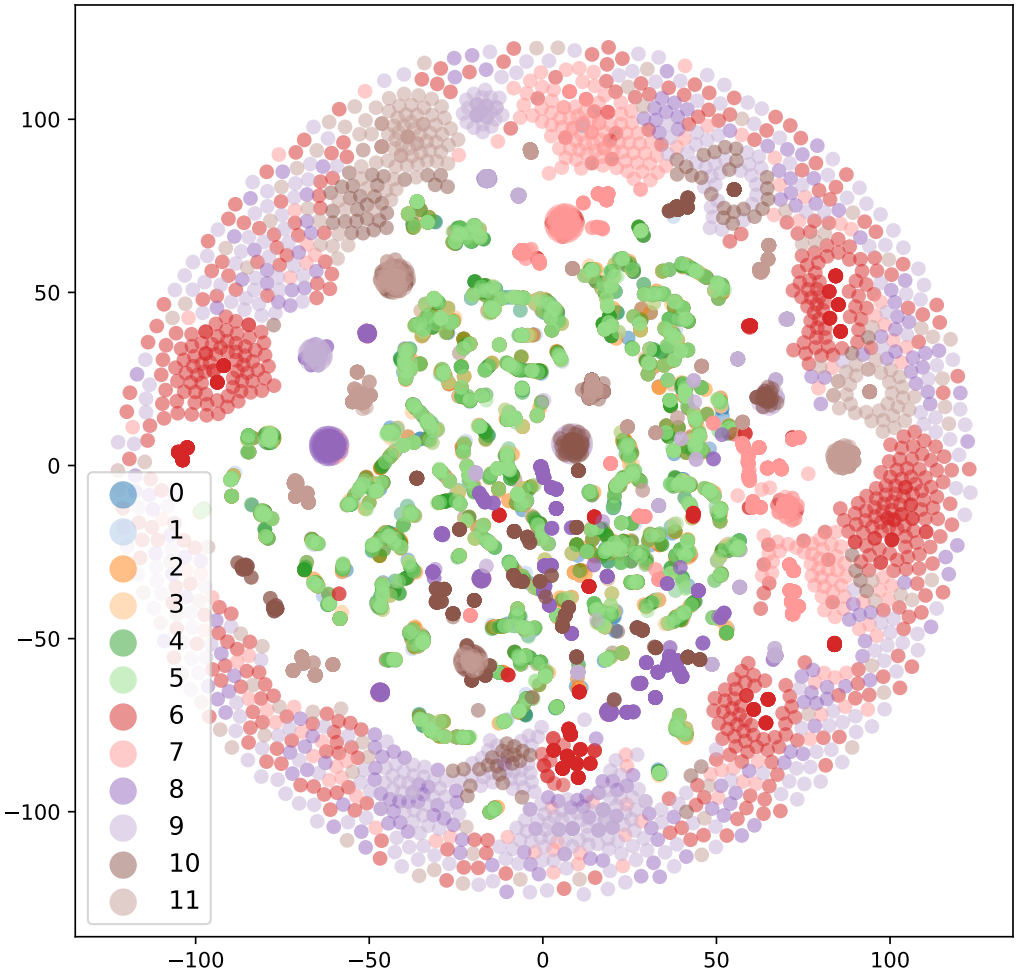

Figure 5: Visualizing prototypical feature vectors using t-SNE. 6 NONA labels (0 to 5) and 6 argument types (6 to 12) from trained in-domain(small) checkpoints. Best viewed in color.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
7

☑ A2. Did you discuss any potential risks of your work?
7

☑ A3. Do the abstract and introduction summarize the paper's main claims?
abstrct and 1

☑ A4. Have you used AI writing assistants when working on this paper?
Grammarly

B ☑ Did you use or create scientific artifacts?

3

☑ B1. Did you cite the creators of artifacts you used?
3

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
3

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
3

☑ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
3

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
3

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
3

C ☑ Did you run computational experiments?

Appendix B

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Appendix B

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Appendix B

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Appendix B

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Appendix B

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.