Leveraging Open Information Extraction for More Robust Domain Transfer of Event Trigger Detection

David Dukić[1,†] **Kiril Gashteovski**[2,3] **Goran Glavaš**[4] **Jan Šnajder**[1]
[1]TakeLab, Faculty of Electrical Engineering and Computing, University of Zagreb
[2]NEC Laboratories Europe, Heidelberg, Germany
[3]CAIR, Ss. Cyril and Methodius University, Skopje, North Macedonia
[4]CAIDAS, University of Würzburg, Germany
[1,2,4]name.surname@{fer.hr, neclab.eu, uni-wuerzburg.de}

Abstract

Event detection is a crucial information extraction task in many domains, such as Wikipedia or news. The task typically relies on trigger detection (TD) – identifying token spans in the text that evoke specific events. While the notion of triggers should ideally be universal across domains, domain transfer for TD from high- to low-resource domains results in significant performance drops. We address the problem of negative transfer in TD by coupling triggers between domains using subject-object relations obtained from a rule-based open information extraction (OIE) system. We demonstrate that OIE relations injected through multi-task training can act as mediators between triggers in different domains, enhancing zero- and few-shot TD domain transfer and reducing performance drops, in particular when transferring from a high-resource source domain (Wikipedia) to a low(er)-resource target domain (news). Additionally, we combine this improved transfer with masked language modeling on the target domain, observing further TD transfer gains. Finally, we demonstrate that the gains are robust to the choice of the OIE system.[1]

1 Introduction

Event detection is an important part of the information extraction pipeline in natural language processing (NLP). Event detection systems are typically bound to domain-specific schemes and fill predefined event-specific slots evoked by an event *trigger* – a span of words that evokes a particular type of event. A typical domain-specific event detection workflow consists of trigger detection (TD), which locates the trigger span in the text, and trigger classification (Xiang and Wang, 2019), which assigns one of the predefined event types to the trigger. With triggers identified, the next step is typically to detect the corresponding arguments, e.g., participants, location, and time. The detected events can be leveraged for many downstream tasks, including knowledge graph construction (Zhang et al., 2021), information retrieval (Glavaš and Šnajder, 2013), text summarization (Zhang et al., 2023), and aspect-based sentiment analysis (Tang et al., 2022).

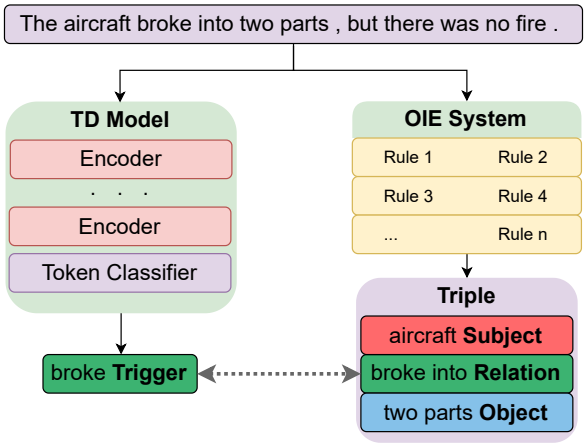


Figure 1: An example of event trigger detection and subject-relation-object extraction with an open information extraction (OIE) system. The detected trigger and extracted OIE relation often overlap to a significant degree, which can be leveraged for creating more robust trigger detection models across domains.

While the notion of an event trigger is intuitive and universal (i.e., events and their triggers exist in all text domains), NLP research has struggled to provide a clear-cut operational definition of an event, giving rise to diverse annotation schemes, e.g., (Doddington et al., 2004; Pustejovsky et al., 2005; Shaw et al., 2009; Cybulska and Vossen, 2014; Song et al., 2015). The differences between annotation schemes, alongside the usual distribution shifts between text domains, make domain transfer of TD very challenging. Empirical evidence has demonstrated massive performance drops in zero- and low few-shot TD transfer from a high-resource source to a low(er)-resource target

[†]Corresponding author: david.dukic@fer.hr
[1]Find code at https://github.com/dd1497/oie-td.

Findings of the Association for Computational Linguistics: EACL 2024, pages 1197–1213
March 17-22, 2024

domain – a phenomenon commonly referred to as *negative transfer* (Wang et al., 2019; Ngo Trung et al., 2021; Meftah et al., 2021). The absence of an effective domain transfer method for TD implies a costly (large-scale) manual annotation of event trigger spans for each domain of interest.

One way to facilitate domain transfer of TD may be by means of a proxy task that (i) exhibits a smaller distributional shift across domains and could thus (ii) mediate representational alignment between triggers of different domains. In principle, all tasks that extract structures that relate to event semantics, such as syntactic or predicate-argument structures, make good candidates for such a mediator (McClosky et al., 2011; Liu et al., 2016). Recent work by Deng et al. (2022) showed that trigger and argument detection could be aligned with the subject-relation-object triples as mediators (in Chinese), with subjects and objects mapped to arguments and relations to triggers. In other words, both events and subject-relation-object triples represent predicate-argument structures, pointing to tasks that extract the latter as potentially good mediators for domain transfer of TD.

Open Information Extraction (OIE) systems (Banko et al., 2007) automatically extract subject-relation-object triples in a domain-independent manner because they discover relations not predefined by any schema (Fader et al., 2011; Wang et al., 2018; Sun et al., 2018; Gashteovski et al., 2019). Although most recent OIE systems are neural models trained in a supervised manner (Kolluru et al., 2020; Kotnis et al., 2022), traditional OIE systems such as Stanford OIE (Angeli et al., 2015) and MinIE (Gashteovski et al., 2017) are rule-based and typically do not require domain-specific pre-processing of the input text (Lauscher et al., 2019). Moreover, recent fact-based evaluation (Gashteovski et al., 2022) renders them more accurate than neural OIE models. Figure 1 illustrates the overlap between the trigger *broke* detected by the trigger detection model and an OIE relation *broke into*, extracted by MinIE. This overlap is the main motivation for our work.

In this paper, we address the challenge of negative transfer in TD by leveraging OIE relations to align representations of event triggers across domains. While annotating event triggers in the target domain is costly, automatic extraction of open relations with a rule-based OIE system is cheap, even at a large scale. With this in mind, we investigate remedies for negative domain transfer of TD based on the automatic extraction of OIE subject-object relations. More precisely, we couple the domain-specific trigger annotations with the relation extractions obtained with a domain-agnostic rule-based OIE system through different (i) multi-task architectures and (ii) zero- and few-shot transfer regimes. The intuition is that, by coupling trigger annotations with OIE relations, we effectively couple event triggers between domains with OIE relations as mediators. Although OIE relations do not always align perfectly with event triggers, we find that they can facilitate and stabilize the domain transfer of TD. We demonstrate that (i) multi-task fine-tuning of a pretrained language model (PLM) for OIE relation extraction and TD and (ii) transfer training regimes adopted from the body of work on language transfer (Lauscher et al., 2020; Schmidt et al., 2022) reduce the trigger distribution shift between domains and consequently improve TD performance in the low-resource target domain.

Contributions. **(1)** We mitigate negative domain transfer of trigger detection by coupling event triggers with subject-object relations extracted by rule-based OIE; we couple the two in different multi-task model designs and investigate the effects in both zero- and few-shot transfer. **(2)** We show that target-domain masked language modeling (MLM), in the vein of Gururangan et al. (2020), as an additional auxiliary objective next to open relation extraction, further improves TD transfer. **(3)** We validate that the gains from the OIE-based proxy are robust and not dependent on the specific OIE system. We believe our work is an important step towards universally more effective event extraction.

2 Background and Related Work

Domain Transfer. Domain transfer has been investigated for numerous structured prediction tasks such as query translation (Yao et al., 2020), term extraction (Hazem et al., 2022), named entity recognition (Jia and Zhang, 2020) and disambiguation (Blair and Bar, 2022), and event argument extraction (Sainz et al., 2022). Existing work on domain transfer for event extraction predominantly resorted to semantic role labeling (SRL) as the vehicle for facilitating the transfer. Lyu et al. (2021) ran SRL to detect predicates as potential event triggers for the domain transfer of event extraction via question answering and textual entailment models. Peng et al. (2016) investigated the use of SRL predicates and arguments to facilitate domain transfer for both

event detection and event co-reference resolution. While SRL is structurally fit to be a proxy task for even extraction, it is also a task that requires domain-specific annotations. More recently, domain adaptation for models based on PLMs has been driven by general self-supervised language modeling on (unlabeled) domain-specific corpora (Gururangan et al., 2020; Hung et al., 2022).

Domain Adaptation for Event Detection. Nguyen and Grishman (2015) were the first to employ a convolutional neural network (CNN) for event detection domain adaptation by learning more universal trigger representations through a CNN architecture and various features such as word, position, and entity type embeddings. Naik and Rose (2020) tackled TD transfer between literature and news domains using adversarial domain adaptation to produce representations predictive for triggers but not predictive of the example's domain, thus forcing the model to learn domain-agnostic trigger representations. Ngo Trung et al. (2021) leveraged domain-specific adapters for event detection domain transfer. More recently, Trung et al. (2022) developed an unsupervised domain adaptation method applicable to text classification tasks, including event detection and sentiment classification, which utilizes meta- and self-paced learning approaches. Other strands of research deal with improving few-shot event detection but are mostly limited to in-domain transfer between different event types (Lai et al., 2020; Li et al., 2020). Examples include improving the zero- and few-shot in-domain event detection performance with cloze-based prompt meta-learning (Yue et al., 2023) and ontology embeddings (Deng et al., 2021).

OIE for NLP tasks. OIE systems are intended to facilitate various downstream tasks, including text summarization (Fan et al., 2019; Ribeiro et al., 2022), question answering (Yan et al., 2018; Nagumothu et al., 2022), incomplete sentence reconstruction (Montella et al., 2020), and event extraction (Chen et al., 2023). Many event-related tasks, such as event schema induction (Balasubramanian et al., 2013) and cross-domain event coreference (Pratapa et al., 2021), benefit from leveraging OIE triples. However, OIE has not yet been employed to improve TD. A step in that direction is the work by Deng et al. (2022), where authors created a dataset named *Title2Event* consisting of Chinese titles designed for *open event extraction* based on OIE triples, subscribing to the idea that events

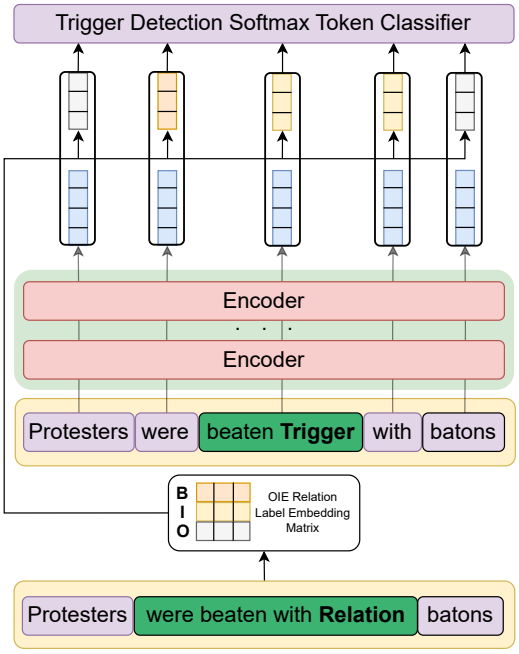


Figure 2: *Implicit* model during training. The input sentence is fed twice: once with trigger IOB2 tags through PLM encoders and once with OIE relation IOB2 tags by indexing the corresponding label embedding matrix. At the *implicit* output, PLM's last hidden state embeddings are concatenated with OIE relation label embeddings per token and passed through the TD softmax classifier.

are well-aligned with the subject-relation-object schema, which we also adopt in this work.

3 OIE for Event Trigger Detection

Following prior work (Naik and Rose, 2020; Ngo Trung et al., 2021), we frame TD as a sequence labeling task where each token is classified as either part of some event trigger span or outside of it. This task formulation is intuitive, given that event triggers are consecutive token sequences, and multiple triggers may appear in the same input sentence. We use the widely adopted IOB2 (inside, outside, begin) tagging scheme (Ratnaparkhi, 1998). Analogously, we model relation extraction (RE) – for which we use OIE relation extractions as ground-truth labels – also as a sequence labeling task with its own set of IOB2 tags. We tackle domain transfer for TD with two different model architectures (based on a PLM) that couple OIE relations with TD annotations, which we refer to as (i) *implicit* and (ii) *explicit* OIE-TD multi-task models. We next describe both variants in detail.

Implicit Multi-Task. In the *implicit* model, we train and use embeddings for token labels of OIE

relations: one randomly initialized vector for each of the three IOB2 tags. The model concatenates the embedding $\mathbf{x}_{\text{OIE}} \in \mathbb{R}^d$ of the OIE relation label of each token embedding to the contextualized token embedding of the token $\mathbf{x}_{\text{PLM}} \in \mathbb{R}^h$ (the output of the last PLM layer), where d is the dimension of the trainable OIE relation label embeddings (hyperparameter of the model), and h is the PLM's hidden size. The final token representation, $\mathbf{x} = [\mathbf{x}_{\text{PLM}}; \mathbf{x}_{\text{OIE}}]$, is fed to the standard softmax classifier, which predicts the IOB2 event trigger label for the token, $softmax(\mathbf{W}_{\text{cl}}^{\text{T}}\mathbf{x} + \mathbf{b}_{\text{cl}})$, with $\mathbf{W}_{\text{cl}} \in \mathbb{R}^{(d+h)\times 3}$ and $\mathbf{b}_{\text{cl}} \in \mathbb{R}^3$ as trainable parameters of the classifier. As is common in multi-class classification, we tune all parameters by minimizing the (multi-class) cross-entropy loss. The *implicit* model is illustrated in Figure 2. We train the model on TD in the source domain, optimizing (1) all of the PLM's parameters, (2) classifier's parameters $\mathbf{W}_{\text{cl}}$ and $\mathbf{b}_{\text{cl}}$, and (3) embedding matrix $\mathbf{X}_{\text{OIE}} \in \mathbb{R}^{3\times d}$ containing the trainable embeddings of the OIE labels. At inference time in the target domain, we run the OIE system on test sentences to obtain the OIE relation labels for tokens and then perform inference using the *implicit* PLM for TD and embeddings of OIE labels obtained in training.

We hypothesize that the *implicit* model is incentivized to establish – within the OIE label embeddings trained via event TD – contextualized associations between the two tasks. Intuitively, this should improve the recall of TD in the target domain as long as the OIE – which is rule-based and thus more domain agnostic – is resilient to distribution shifts between domains. Similar event detection approaches based on training label embeddings exist (Nguyen and Grishman, 2015; Liu et al., 2017; Ji et al., 2019). However, they typically concatenate the label and token embeddings at the encoder's input and rely on encoders shallower than common Transformer-based PLMs.

Explicit Multi-Task. The *explicit* model works with two standard softmax classifiers and a shared PLM encoder. The representation of each token $\mathbf{x}_{\text{PLM}} \in \mathbb{R}^h$, from PLM's last layer, is forwarded to the (i) TD softmax classifier $softmax(\mathbf{W}_{\text{td}}^{\text{T}}\mathbf{x}_{\text{PLM}} + \mathbf{b}_{\text{td}})$, which predicts the IOB2 event trigger label for the token and (ii) RE softmax classifier $softmax(\mathbf{W}_{\text{re}}^{\text{T}}\mathbf{x}_{\text{PLM}} + \mathbf{b}_{\text{re}})$, which predicts the IOB2 relation label for the token, with $\mathbf{W}_{\text{td}}, \mathbf{W}_{\text{re}} \in \mathbb{R}^{h\times 3}$ and $\mathbf{b}_{\text{td}}, \mathbf{b}_{\text{re}} \in \mathbb{R}^3$ as trainable parameters of two classifiers. Based on the predictions, the (multi-class) cross-entropy loss is calculated for each classifier separately on a mini-batch basis. The average of calculated TD and RE losses is used to update PLM's and classifiers' parameters during training. This is where the interaction of knowledge from both tasks occurs. At inference time, we do not use OIE relation labels in any way. The intuition is that if the notion of triggers is universal across domains and the OIE relations are indeed domain-independent, it should be sufficient only to leverage the in-domain trigger-relation connection during training. Considering that the TD and RE tasks have the same number of corresponding labels, we tried to share the softmax classifier between TD and RE, but that led to worse overall performance.

Dataset	Train			Valid			Test		
	#Sent	#Tr	#Re	#Sent	#Tr	#Re	#Sent	#Tr	#Re
MAVEN	25944	24063	15590	6487	6038	3940	8042	7469	4805
ACE 2005	14672	3256	7403	873	340	446	711	292	412
EDNYT	1842	1500	1164	95	74	65	198	155	115
EVEXTRA	8534	7056	5461	1103	902	700	2482	2077	1590

Table 1: Statistics for the four datasets and their splits: the number of sentences (#Sent), the number of sentences with triggers (#Tr), and the number of relations after post-processing of MinIE triple extractions (#Re).

4 Experimental Setup

Our experiments investigate the transfer from a high-resource source domain to a low-resource target domain, which is the common transfer direction. For facilitating few-shot domain transfer of TD, we employ *joint* and *sequential* transfer training regimes in combination with multi-task models.

4.1 Datasets and Preprocessing

As a dataset from a high-resource source domain, we use MAVEN, a dataset of Wikipedia articles with sentence-level trigger annotations. In the low-resource target domain, we use datasets from the news domain – ACE 2005, EDNYT, and the EVEXTRA – which also have sentence-level trigger annotations. Table 1 summarizes the dataset statistics.

MAVEN. The MAssive eVENt detection dataset (Wang et al., 2020) from the English Wikipedia domain is the largest freely available dataset suitable for TD. It covers more than 150 events. The size and coverage of event types make MAVEN an ideal source dataset for the domain transfer of TD. MAVEN comes with tokenized sentences and

a predefined train, validation, and test split. However, since no gold test set labels were published, we use the official validation set as a test set (only to measure the source model performance on it) and randomly sample 20% of sentences from the training data as a new validation set.

ACE 2005. The Automatic Content Extraction dataset (Doddington et al., 2004) is a widely used event detection dataset consisting predominantly of articles from various news sources in multiple languages. We use only the English train, validation, and test split, obtained with the standard ACE preprocessing tool,[2] which we also use to obtain sentences and tokens. Although ACE is a sizable dataset, as noted by Wang et al. (2020), many ACE sentences do not contain any triggers (cf. Table 1).

EDNYT. The event detection dataset of Maisonnave et al. (2022) was compiled from the New York Times articles on financial crises, which makes the dataset more topically focused than the other datasets. The dataset was not tokenized, but it came with a train-test split, with the test set comprising 10% of the data. We obtain a validation set by randomly sampling 5% of the train data. We use spaCy (Honnibal et al., 2020) to tokenize the sentences. We discarded 3% of sentences with trigger spans that could not be aligned with spaCy tokenization.

EVEXTRA. The EVEXTRA dataset (Glavaš and Šnajder, 2015) is an English newspaper corpus annotated with event triggers. It comes tokenized but with no predefined split. We randomly assign sentences to train, validation, and test sets in a 70/10/20 ratio, respectively, ensuring that sentences from the same article end up in the same set. Less than 1% of sentences were dropped because aligning the trigger annotations with tokens was impossible.

Relation Extraction. We use the rule-based OIE system MinIE (Gashteovski et al., 2017) to extract subject-relation-object triples from sentences. MinIE has proven useful for many downstream tasks by the BenchIE benchmark and evaluation framework (Gashteovski et al., 2022). However, it extracts all possible triples from the input text and introduces minor extraction errors, so we use a set of heuristics to post-process the results and improve the alignment of extracted relations and labeled triggers. To verify the alignment, we conduct a χ^2 test of dependence on train sets of both source and target datasets, considering whether the same token is labeled as a relation and as a trigger. The dependence between variables was significant for all datasets ($p < .01$). A detailed description is given in Appendix A.1. First, we remove implicit[3] triple extractions and discard all non-consecutive subject, relation, or object extractions. Further, we remove non-triples, relations with more than five tokens, and extractions not in the subject-relation-object order. Finally, we remove subject and object extraction information from the sentences and drop duplicates, leaving us only with relation extractions. Table 1 shows the final number of sentences containing relations in the post-processed datasets.

4.2 Training Regimes

In addition to using OIE relations with multi-task models to couple triggers with relations, we take inspiration from recent findings in language transfer (Meftah et al., 2021; Schmidt et al., 2022) and experiment with three transfer training regimes: *joint training*, *joint transfer*, and *sequential transfer*. For the sake of completeness, we also consider *in-domain training*, which reduces to fine-tuning each model on few-shot target domain examples.

Joint Training. The *joint training* regime relies on mixed batches, adopted from the work on language transfer (Schmidt et al., 2022). A mixed batch consists predominantly of source trigger examples combined with a much lower fixed share of few-shot target trigger examples. Intuitively, having fewer few-shot examples should contribute to the update of model parameters with equal weight as the abundant source examples and ultimately prevent the model from overfitting on source data. We create mixed mini-batches consisting of $B = n + m$ examples, where n are source examples, m are randomly sampled few-shot target examples, and $n \gg m$. If more than m few-shot examples are available, m are consistently sampled from the few-shot pool. We fix $B = 32$ with $n = 27, m = 5$ in our experiments. Fine-tuning is performed for a fixed number of epochs based on mixed mini-batch loss, calculated as the average of the source loss and m-shot target loss. In our experiments, *joint training* amounts to mixed batch fine-tuning from either single- (TD) or multi-task (TD+RE) PLMs.

[2] https://bit.ly/ace2005-preprocessing

[3] OIE systems often incorporate binding tokens (like the copula *is*), which do not have to be present in the text.

Joint Transfer. Similar to *joint training*, the *joint transfer* regime also uses mixed batches. However, instead of fine-tuning from PLM, we first train each PLM on source training data and then fine-tune with mixed batches in the same manner as in *joint training*. *Joint transfer* applied to multi-task models utilizes source OIE relations twice and target relations once during mixed batch fine-tuning.

Sequential Transfer. Analogously to *joint transfer*, in the *sequential transfer* regime, we fine-tune for a fixed number of epochs from the PLM trained on the source domain training data. However, unlike in *joint transfer*, fine-tuning is done only with target few-shot examples.

4.3 Training Details and Hyperparameters

We briefly describe the training details (see Appendix A.2 for more details). We use the RoBERTa-base (Liu et al., 2019) PLM for token classification, implemented in *Hugging Face* (Wolf et al., 2020). We evaluate TD by micro F1 score on IOB2 tag predictions using strict matching, where the predicted output span must exactly match the expected output span. The models are trained with cross-entropy loss and Adam optimizer (Kingma and Ba, 2014) with the learning rate of 0.00001 for 10 epochs.

When training on the source domain, we use the source validation set to select the best model based on the TD micro F1 score. Specifically, we choose the model from the epoch that yields the highest TD validation performance.[4] Fine-tuning in *joint/sequential transfer* regimes starts from the best model selected on the source validation set. In *joint transfer* with the *implicit* model, we perform mixed batch fine-tuning by averaging the source TD and target few-shot TD losses. Similarly, we average the source TD and RE losses with the target few-shot TD and RE losses in the *joint transfer* with the *explicit* model. Throughout experiments, we use a batch size of $B = 32$. Also, we employ gradient clipping of model parameters to a maximum of 1.0 before each mini-batch update. We do transfer experiments with $0, 5, 10, 50, 100, 250,$ and 500 shots. For MLM and *in-domain training*, we update the models' parameters in an alternate fashion inside each epoch: first, based on target training data MLM loss, and then based on target few-shot loss. The MLM *sequential transfer* is similar as without MLM. The difference is in the starting model, which is obtained by first training in the same described alternate fashion but with updates based on MLM loss on target training data and TD loss on source training data.

[4]We also experimented with selecting the model based on the MLM perplexity on the target validation set, but that led to worse performance than optimizing for TD F1 on the source validation set. The two options present a trade-off between learning TD adequately or adjusting to the target domain at the expense of TD performance.

5 Results and Discussion

Table 2 shows the main results of our experiments, with MinIE as a relation extractor for the multi-task models. *Vanilla* is the sequence labeling PLM fine-tuned only for event TD, i.e., PLM with softmax token classifier on top trained on labeled event trigger spans. This model is trained in the same fashion as our proposed *implicit* and *explicit* variants, but without incorporating in any way the OIE relation information. For all experiments in this section, we average results over three seeds and report micro F1 TD scores on the held-out target test sets. For few-shot experiments, we additionally perform averaging on five different randomly sampled subsets from the target data training set. Moreover, we take precautions to ensure that samples from each draw are consistent across experiments and exclusively contain examples with triggers.

5.1 Main Results

Zero-shot domain transfer of TD from MAVEN as the source to news datasets as targets exhibits noticeable negative transfer. The drops are massive compared to the performance of the models trained on all ACE 2005, EDNYT, or EVEXTRA training data. Even in this worst-case zero-shot setup, multi-task *implicit* and *explicit* models bring gains compared to *vanilla* ones. Some interesting trends emerge when the number of shots increases. On average, relations help achieve higher target domain TD performance for a low-to-moderate number of shots. However, when the number of shots reaches 500 (or even 250 in some cases) target examples, the effects of relations become negligible, except for the EVEXTRA dataset, where the gains from relations are consistent regardless of the number of shots or training regime. When considering all training regimes, the *implicit* model outperforms the *explicit* model. Contrary to the findings from language transfer (Schmidt et al., 2022), *joint* transfer training regimes were almost consistently worse compared to *sequential transfer* and *in-domain training*. These findings are of

Training Regime		ACE 2005 (0.706)			EDNYT (0.702)			EVEXTRA (0.893)		
		Vanilla	Implicit	Explicit	Vanilla	Implicit	Explicit	Vanilla	Implicit	Explicit
	0-Shot	0.234	0.237	**0.240**	0.392	0.399	**0.408**	0.650	0.650	**0.653**
joint training	5-Shot	0.246	0.250	**0.256**	0.451	0.455	**0.457**	0.643	0.643	**0.654**
	10-Shot	0.251	0.253	**0.262**	0.482	**0.484**	**0.484**	0.645	0.645	**0.658**
	50-Shot	0.265	0.268	**0.283**	0.566	**0.575**	0.567	0.679	0.681	**0.687**
	100-Shot	0.286	0.286	**0.310**	0.597	**0.602**	0.596	0.715	0.721	**0.725**
	250-Shot	0.332	0.330	**0.357**	0.628	**0.629**	**0.629**	0.766	**0.767**	0.765
	500-shot	0.382	0.378	**0.398**	**0.649**	**0.649**	0.646	0.793	**0.798**	0.792
joint transfer	5-Shot	0.248	0.248	**0.254**	0.433	0.436	**0.440**	0.631	0.633	**0.636**
	10-Shot	0.251	0.250	**0.256**	0.448	**0.451**	0.450	0.632	0.634	**0.638**
	50-Shot	0.262	0.265	**0.267**	0.524	**0.536**	0.507	0.650	**0.656**	0.648
	100-Shot	0.283	0.283	**0.284**	0.569	**0.573**	0.551	0.676	**0.684**	0.667
	250-Shot	**0.328**	**0.328**	0.318	0.608	**0.611**	0.592	0.727	**0.735**	0.705
	500-Shot	**0.388**	0.381	0.369	0.637	**0.641**	0.621	0.770	**0.777**	0.744
sequential transfer	5-Shot	**0.294**	**0.294**	0.276	0.458	**0.466**	0.448	0.659	**0.661**	0.653
	10-Shot	0.372	**0.374**	0.330	0.512	**0.521**	0.490	0.688	**0.693**	0.680
	50-Shot	**0.511**	0.506	0.463	0.581	**0.592**	0.568	0.750	**0.764**	0.741
	100-Shot	0.538	**0.548**	0.501	0.605	**0.616**	0.584	0.786	**0.795**	0.773
	250-Shot	**0.587**	0.577	0.556	0.631	**0.644**	0.607	0.824	**0.835**	0.813
	500-Shot	**0.610**	0.609	0.586	**0.653**	0.652	0.640	0.852	**0.857**	0.836
in-domain training	5-Shot	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	10-Shot	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	50-Shot	0.464	**0.466**	0.417	**0.607**	0.601	0.597	0.768	**0.774**	0.757
	100-Shot	0.510	**0.529**	0.511	0.626	**0.632**	0.611	0.807	**0.812**	0.801
	250-shot	**0.570**	0.569	0.550	0.649	**0.654**	0.642	0.845	**0.847**	0.835
	500-Shot	0.598	**0.600**	0.584	0.660	0.658	**0.666**	0.858	**0.862**	0.854

Table 2: TD domain transfer micro F1 scores when transferring from MAVEN as a source to ACE 2005, EDNYT, and EVEXTRA as targets (zero-shot, three few-shot transfer training regimes, and in-domain, with six varying numbers of shots). The numbers in parentheses next to the target dataset are the in-domain performance test set scores when using all target training data. *Joint/in-domain training* – target fine-tuning from PLM. *Joint/sequential transfer* – target fine-tuning from PLM trained for TD on MAVEN source training data. The best results by dataset and model per training regime are in **bold**. *Implicit* and *explicit* models leverage MinIE relation labels, unlike the *vanilla* model. All reported results are averages of three runs. We report standard deviations in Appendix A.3.

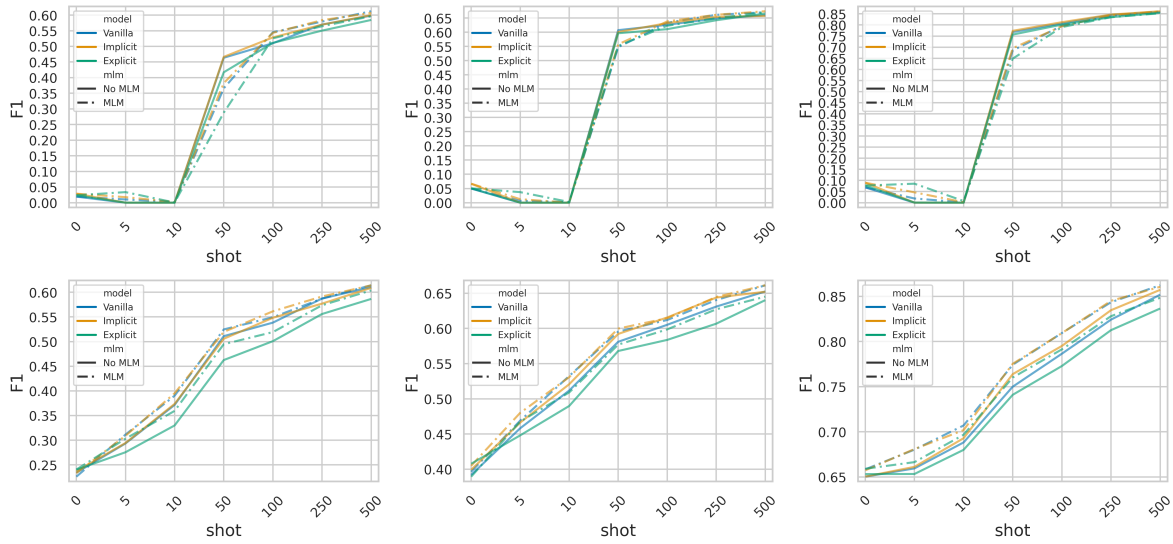
practical interest since *joint* is worse performance-wise and takes far more resources and time to train. With 500 shots, *sequential transfer* and *in-domain training* come close to the full in-domain training performance for each news dataset. For a low number of shots (5 and 10), doing *in-domain training* is useless, and in this case, *sequential transfer* is a better option. However, a higher number of shots in combination with *in-domain training* can lead to a better performance than *sequential transfer*.

5.2 Adding Auxiliary MLM Objective

Building on recent findings from work on PLM domain adaptation (Gururangan et al., 2020), we investigate whether MLM can further boost TD transfer from Wikipedia to the news domain. Since *joint* regimes were consistently worse in main results, we examine the MLM effect only for *in-domain training* and *sequential transfer*. We achieve this by adding token-level MLM as an auxiliary training objective through an extra MLM head in all model variants. The head's parameters are updated during training and not used during inference. Figure 3 gives the results. *Sequential transfer* proved to be more efficient than *in-domain training*. On average, MLM with relations embodied into *implicit* model in *sequential transfer* regime outperforms the best results without MLM. An exception is the EVEXTRA dataset, where using OIE relations in conjunction with MLM and *sequential transfer* does not lead to performance improvements compared to using only MLM.

5.3 The Choice of the OIE System

Finally, to examine if our results are specific to the OIE system, we replace MinIE with Stanford OIE. We post-process the relations in the same manner as for MinIE (cf. Section 4). The experiments are conducted without MLM and for *sequential transfer* and *in-domain training* regimes. Table 3 shows



(a) ACE 2005 (0.706) (b) EDNYT (0.702) (c) EVEXTRA (0.893)

Figure 3: TD domain transfer micro F1 scores when transferring from MAVEN as a source to ACE 2005, EDNYT, and EVEXTRA as targets (zero-shot, *in-domain training*, and *sequential transfer*, with six varying numbers of shots). The numbers in parentheses next to the target dataset are the in-domain performance test set scores when using all target training data. The upper three plots show *in-domain training* results – target fine-tuning starting from PLM. The lower three plots show *sequential transfer* results – target fine-tuning starting from PLM trained for TD on MAVEN source training data. Dash-dotted lines correspond to models with an auxiliary MLM objective on target domain training data. The x-axis shows the number of shots on an ordinal scale. *Implicit* and *explicit* models leverage MinIE relation labels, unlike the *vanilla* model. All reported results are averages of three runs. The corresponding results in tabular form with standard deviations are in Appendix A.3.

the results. The difference between using MinIE and Stanford OIE is negligible for *implicit* model but exists for *explicit* model. Since *explicit* outperformed *implicit* in only five out of 156 cases from Table 3, we conclude that the gains from leveraging OIE relations in multi-task models are not due to the higher quality of MinIE extractions and persist for Stanford OIE. One can achieve similar, if not almost identical, gains using either extractor.

6 Conclusion

We showed that OIE relations can be utilized to improve the domain transfer of trigger detection (TD) in zero- and few-shot setups. The best improvements were achieved with *implicit* multi-task model and *sequential transfer* training regime. We also demonstrated that more substantial gains can be reached when combining OIE relations with MLM as an auxiliary task. This is especially evident for the models pre-trained with TD task on the source domain and with MLM training objective on the target domain in the *implicit* multi-task model. Replacing MinIE with Stanford OIE revealed that gains on the target domain for the TD task persist when using the other OIE extractor.

Future work may further explore the potential of OIE for improving domain transfer of TD on diverse datasets and domains, such as the cybersecurity (Man Duc Trong et al., 2020), literature (Sims et al., 2019), and biomedical (Kim et al., 2009) domains. Applying the coupling concept to other NLP tasks, such as event argument detection or named entity recognition, where OIE extractions might enhance the in- and out-of-domain performance, is another exciting future work direction.

7 Limitations

Our experiments were limited by the available computing resources. For reliability, in our experiments, we report performance scores averaged over three runs (differing in random seeds). Similarly, we sampled the few-shot examples five times. Averaging over larger samples would make the results even more reliable. Furthermore, the results of few-shot experiments can sometimes turn out to be misleading due to the high variance of the sample of examples. Fixing the learning rate and some other hyperparameters across experiments may have resulted in suboptimal adaptation to the trigger detection task in both source and target

Training Regime		ACE 2005 (0.706)				EDNYT (0.702)				EVEXTRA (0.893)			
		MinIE		Stanford OIE		MinIE		Stanford OIE		MinIE		Stanford OIE	
		Implicit	Explicit	Implicit	Explicit	Implicit	Explicit	Implicit	Explicit	Implicit	Explicit	Implicit	Explicit
	0-Shot	0.237	0.240	0.237	**0.242**	0.399	**0.408**	0.401	0.406	0.650	0.653	0.650	**0.657**
sequential transfer	5-Shot	0.294	0.276	**0.296**	0.283	0.466	0.448	**0.468**	0.464	**0.661**	0.653	**0.661**	0.658
	10-Shot	0.374	0.330	**0.375**	0.350	**0.521**	0.490	0.520	0.512	**0.693**	0.680	**0.693**	0.688
	50-Shot	**0.506**	0.463	**0.506**	0.476	**0.592**	0.568	0.591	0.570	**0.764**	0.741	0.763	0.747
	100-Shot	**0.548**	0.501	**0.548**	0.525	**0.616**	0.584	0.615	0.587	0.795	0.773	**0.796**	0.775
	250-Shot	**0.577**	0.556	**0.577**	0.568	0.644	0.607	**0.647**	0.602	**0.835**	0.813	0.834	0.818
	500-Shot	**0.609**	0.586	0.602	0.584	0.652	0.640	**0.653**	0.627	**0.857**	0.836	0.856	0.845
in-domain training	5-Shot	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	10-Shot	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	50-Shot	0.466	0.417	**0.467**	0.446	0.601	0.597	0.601	**0.605**	0.774	0.757	**0.775**	0.765
	100-Shot	**0.529**	0.511	**0.529**	0.515	0.632	0.611	**0.633**	0.615	0.812	0.801	**0.814**	0.805
	250-Shot	**0.569**	0.550	**0.569**	0.557	**0.654**	0.642	0.652	0.638	**0.847**	0.835	0.846	0.840
	500-Shot	**0.600**	0.584	0.598	0.585	0.658	**0.666**	0.657	0.662	**0.862**	0.854	0.861	0.852

Table 3: TD domain transfer micro F1 scores when transferring from MAVEN as a source to ACE 2005, EDNYT, and EVEXTRA as targets w.r.t. MinIE and Stanford OIE systems (zero-shot, *sequential transfer*, and *in-domain training*, with six varying numbers of shots). The numbers in parentheses next to the target dataset are the in-domain performance test set scores when using all target training data. *Sequential transfer* – target fine-tuning from PLM trained for TD on MAVEN source training data. *In-domain training* – target fine-tuning from PLM. The best results by dataset, *implicit* or *explicit* relation-leveraging models, per training regime and OIE system, are in **bold**. All reported results are averages of three runs.

domains. Moreover, all experiments were done only with RoBERTa-base; using a different suitable PLM might yield further insights. Finally, our experiments were limited to datasets in the English language; further insights may be gained by extending to cross-lingual trigger detection domain transfer, more transfer directions, and datasets.

8 Ethical Considerations

Developing models for automated event detection comes with inherent risks, including the potential for misuse and unintended consequences. The ability to autonomously extract events from sensitive data raises possible ethical concerns, especially in the context of enhanced domain transfer. Combining open information systems with trigger detection models for improved domain transfer reduces the effort of event extraction from sensitive data in a novel domain when only a handful of annotated examples from that domain can be obtained.

References

Gabor Angeli, Melvin Jose Johnson Premkumar, and Christopher D. Manning. 2015. Leveraging linguistic structure for open domain information extraction. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 344–354, Beijing, China. Association for Computational Linguistics.

Niranjan Balasubramanian, Stephen Soderland, Mausam, and Oren Etzioni. 2013. Generating coherent event schemas at scale. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1721–1731, Seattle, Washington, USA. Association for Computational Linguistics.

Michele Banko, Michael J Cafarella, Stephen Soderland, Matthew Broadhead, and Oren Etzioni. 2007. Open information extraction for the web. In *International Joint Conference on Artificial Intelligence*, volume 7, pages 2670–2676.

Philip Blair and Kfir Bar. 2022. Improving few-shot domain transfer for named entity disambiguation with pattern exploitation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6797–6810, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Samuel Broscheit, Kiril Gashteovski, Yanjie Wang, and Rainer Gemulla. 2020. Can we predict new facts with open knowledge graph embeddings? A benchmark for open link prediction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2296–2308, Online. Association for Computational Linguistics.

Yi-Pei Chen, An-Zi Yen, Hen-Hsen Huang, Hideki Nakayama, and Hsin-Hsi Chen. 2023. LED: A dataset for life event extraction from dialogs. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 384–398, Dubrovnik, Croatia. Association for Computational Linguistics.

Agata Cybulska and Piek Vossen. 2014. Guidelines for ECB+ annotation of events and their coreference.

Haolin Deng, Yanan Zhang, Yangfan Zhang, Wangyang Ying, Changlong Yu, Jun Gao, Wei Wang, Xiaoling Bai, Nan Yang, Jin Ma, Xiang Chen, and Tianhua

Zhou. 2022. Title2Event: Benchmarking open event extraction with a large-scale Chinese title dataset. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6511–6524, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Shumin Deng, Ningyu Zhang, Luoqiu Li, Chen Hui, Tou Huaixiao, Mosha Chen, Fei Huang, and Huajun Chen. 2021. OntoED: Low-resource event detection with ontology embedding. 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 2828–2839, Online. 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.

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).

Anthony Fader, Stephen Soderland, and Oren Etzioni. 2011. Identifying relations for open information extraction. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1535–1545, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Angela Fan, Claire Gardent, Chloé Braud, and Antoine Bordes. 2019. Using local knowledge graph construction to scale Seq2Seq models to multi-document inputs. 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 4186–4196, Hong Kong, China. Association for Computational Linguistics.

Kiril Gashteovski, Rainer Gemulla, and Luciano del Corro. 2017. MinIE: Minimizing facts in open information extraction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2630–2640, Copenhagen, Denmark. Association for Computational Linguistics.

Kiril Gashteovski, Sebastian Wanner, Sven Hertling, Samuel Broscheit, and Rainer Gemulla. 2019. OPIEC: an open information extraction corpus. *arXiv preprint arXiv:1904.12324*.

Kiril Gashteovski, Mingying Yu, Bhushan Kotnis, Carolin Lawrence, Mathias Niepert, and Goran Glavaš. 2022. BenchIE: A framework for multi-faceted fact-based open information extraction evaluation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4472–4490, Dublin, Ireland. Association for Computational Linguistics.

Goran Glavaš and Jan Šnajder. 2013. Event-centered information retrieval using kernels on event graphs. In *Proceedings of TextGraphs-8 Graph-based Methods for Natural Language Processing*, pages 1–5, Seattle, Washington, USA. Association for Computational Linguistics.

Goran Glavaš and Jan Šnajder. 2015. Construction and evaluation of event graphs. *Natural Language Engineering*, 21(4):607–652.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Amir Hazem, Merieme Bouhandi, Florian Boudin, and Beatrice Daille. 2022. Cross-lingual and cross-domain transfer learning for automatic term extraction from low resource data. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 648–662, Marseille, France. European Language Resources Association.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength natural language processing in Python.

Chia-Chien Hung, Anne Lauscher, Simone Ponzetto, and Goran Glavaš. 2022. DS-TOD: Efficient domain specialization for task-oriented dialog. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 891–904, Dublin, Ireland. Association for Computational Linguistics.

Yuze Ji, Youfang Lin, Jianwei Gao, and Huaiyu Wan. 2019. Exploiting the entity type sequence to benefit event detection. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 613–623, Hong Kong, China. Association for Computational Linguistics.

Chen Jia and Yue Zhang. 2020. Multi-cell compositional LSTM for NER domain adaptation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5906–5917, Online. Association for Computational Linguistics.

Jin-Dong Kim, Tomoko Ohta, Sampo Pyysalo, Yoshinobu Kano, and Jun'ichi Tsujii. 2009. Overview of BioNLP'09 shared task on event extraction. In *Proceedings of the BioNLP 2009 Workshop Companion*

Volume for Shared Task, pages 1–9, Boulder, Colorado. Association for Computational Linguistics.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Keshav Kolluru, Vaibhav Adlakha, Samarth Aggarwal, Mausam, and Soumen Chakrabarti. 2020. OpenIE6: Iterative Grid Labeling and Coordination Analysis for Open Information Extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3748–3761, Online. Association for Computational Linguistics.

Bhushan Kotnis, Kiril Gashteovski, Daniel Rubio, Ammar Shaker, Vanesa Rodriguez-Tembras, Makoto Takamoto, Mathias Niepert, and Carolin Lawrence. 2022. MILIE: Modular & iterative multilingual open information extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6939–6950, Dublin, Ireland. Association for Computational Linguistics.

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, Online. Association for Computational Linguistics.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From zero to hero: On the limitations of zero-shot language transfer with multilingual Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4483–4499, Online. Association for Computational Linguistics.

Anne Lauscher, Yide Song, and Kiril Gashteovski. 2019. Minscie: Citation-centered open information extraction. In *2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pages 386–387. IEEE.

Fayuan Li, Weihua Peng, Yuguang Chen, Quan Wang, Lu Pan, Yajuan Lyu, and Yong Zhu. 2020. Event extraction as multi-turn question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 829–838, Online. Association for Computational Linguistics.

Shulin Liu, Yubo Chen, Shizhu He, Kang Liu, and Jun Zhao. 2016. Leveraging FrameNet to improve automatic event detection. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2134–2143, Berlin, Germany. Association for Computational Linguistics.

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

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Qing Lyu, Hongming Zhang, Elior Sulem, and Dan Roth. 2021. Zero-shot event extraction via transfer learning: Challenges and insights. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 322–332, Online. Association for Computational Linguistics.

Mariano Maisonnave, Fernando Delbianco, Fernando Tohmé, Ana Maguitman, and Evangelos Milios. 2022. Detecting ongoing events using contextual word and sentence embeddings. *Expert Systems with Applications*, 209:118257.

Hieu Man Duc Trong, Duc Trong Le, Amir Pouran Ben Veyseh, Thuat Nguyen, and Thien Huu Nguyen. 2020. Introducing a new dataset for event detection in cybersecurity texts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5381–5390, Online. Association for Computational Linguistics.

David McClosky, Mihai Surdeanu, and Christopher Manning. 2011. Event extraction as dependency parsing. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1626–1635, Portland, Oregon, USA. Association for Computational Linguistics.

Sara Meftah, Nasredine Semmar, Youssef Tamaazousti, Hassane Essafi, and Fatiha Sadat. 2021. On the hidden negative transfer in sequential transfer learning for domain adaptation from news to tweets. In *Proceedings of the Second Workshop on Domain Adaptation for NLP*, pages 140–145, Kyiv, Ukraine. Association for Computational Linguistics.

Sebastien Montella, Betty Fabre, Tanguy Urvoy, Johannes Heinecke, and Lina Rojas-Barahona. 2020. Denoising pre-training and data augmentation strategies for enhanced RDF verbalization with transformers. In *Proceedings of the 3rd International Workshop on Natural Language Generation from the Semantic Web (WebNLG+)*, pages 89–99, Dublin, Ireland (Virtual). Association for Computational Linguistics.

Dinesh Nagumothu, Bahadorreza Ofoghi, Guangyan Huang, and Peter Eklund. 2022. PIE-QG: Paraphrased information extraction for unsupervised question generation from small corpora. In *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)*, pages 350–359, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Aakanksha Naik and Carolyn Rose. 2020. Towards open domain event trigger identification using adversarial domain adaptation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7618–7624, Online. Association for Computational Linguistics.

Nghia Ngo Trung, Duy Phung, and Thien Huu Nguyen. 2021. Unsupervised domain adaptation for event detection using domain-specific adapters. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4015–4025, Online. Association for Computational Linguistics.

Thien Huu Nguyen and Ralph Grishman. 2015. Event detection and domain adaptation with convolutional neural networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 365–371, Beijing, China. Association for Computational Linguistics.

Haoruo Peng, Yangqiu Song, and Dan Roth. 2016. Event detection and co-reference with minimal supervision. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 392–402, Austin, Texas. Association for Computational Linguistics.

Adithya Pratapa, Zhengzhong Liu, Kimihiro Hasegawa, Linwei Li, Yukari Yamakawa, Shikun Zhang, and Teruko Mitamura. 2021. Cross-document event identity via dense annotation. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 496–517, Online. Association for Computational Linguistics.

James Pustejovsky, Robert Ingria, Roser Sauri, José M Castaño, Jessica Littman, Robert J Gaizauskas, Andrea Setzer, Graham Katz, and Inderjeet Mani. 2005. The specification language TimeML.

Adwait Ratnaparkhi. 1998. *Maximum entropy models for natural language ambiguity resolution*. University of Pennsylvania.

Leonardo F. R. Ribeiro, Mengwen Liu, Iryna Gurevych, Markus Dreyer, and Mohit Bansal. 2022. FactGraph: Evaluating factuality in summarization with semantic graph representations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3238–3253, Seattle, United States. Association for Computational Linguistics.

Oscar Sainz, Itziar Gonzalez-Dios, Oier Lopez de Lacalle, Bonan Min, and Eneko Agirre. 2022. Textual entailment for event argument extraction: Zero- and few-shot with multi-source learning. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2439–2455, Seattle, United States. Association for Computational Linguistics.

Fabian David Schmidt, Ivan Vulić, and Goran Glavaš. 2022. Don't stop fine-tuning: On training regimes for few-shot cross-lingual transfer with multilingual language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10725–10742, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ryan Shaw, Raphaël Troncy, and Lynda Hardman. 2009. LODE: Linking open descriptions of events. *ASWC*, 9:153–167.

Matthew Sims, Jong Ho Park, and David Bamman. 2019. Literary event detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3623–3634, Florence, Italy. Association for Computational Linguistics.

Zhiyi Song, Ann Bies, Stephanie Strassel, Tom Riese, Justin Mott, Joe Ellis, Jonathan Wright, Seth Kulick, Neville Ryant, and Xiaoyi Ma. 2015. From light to rich ERE: Annotation of entities, relations, and events. In *Proceedings of the The 3rd Workshop on EVENTS: Definition, Detection, Coreference, and Representation*, pages 89–98, Denver, Colorado. Association for Computational Linguistics.

Mingming Sun, Xu Li, and Ping Li. 2018. Logician and orator: Learning from the duality between language and knowledge in open domain. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2119–2130, Brussels, Belgium. Association for Computational Linguistics.

Siyu Tang, Heyan Chai, Ziyi Yao, Ye Ding, Cuiyun Gao, Binxing Fang, and Qing Liao. 2022. Affective knowledge enhanced multiple-graph fusion networks for aspect-based sentiment analysis. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5352–5362, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Nghia Ngo Trung, Linh Ngo Van, and Thien Huu Nguyen. 2022. Unsupervised domain adaptation for text classification via meta self-paced learning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4741–4752, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Xiaozhi Wang, Ziqi Wang, Xu Han, Wangyi Jiang, Rong Han, Zhiyuan Liu, Juanzi Li, Peng Li, Yankai Lin, and Jie Zhou. 2020. MAVEN: A Massive General Domain Event Detection Dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1652–1671, Online. Association for Computational Linguistics.

Xuan Wang, Yu Zhang, Qi Li, Yinyin Chen, and Jiawei Han. 2018. Open information extraction with meta-pattern discovery in biomedical literature. In

Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics, BCB '18, page 291–300, New York, NY, USA. Association for Computing Machinery.

Zirui Wang, Zihang Dai, Barnabás Póczos, and Jaime Carbonell. 2019. Characterizing and avoiding negative transfer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11293–11302.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Wei Xiang and Bang Wang. 2019. A survey of event extraction from text. *IEEE Access*, 7:173111–173137.

Zhao Yan, Duyu Tang, Nan Duan, Shujie Liu, Wendi Wang, Daxin Jiang, Ming Zhou, and Zhoujun Li. 2018. Assertion-based QA with question-aware open information extraction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Liang Yao, Baosong Yang, Haibo Zhang, Boxing Chen, and Weihua Luo. 2020. Domain transfer based data augmentation for neural query translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4521–4533, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Zhenrui Yue, Huimin Zeng, Mengfei Lan, Heng Ji, and Dong Wang. 2023. Zero-and few-shot event detection via prompt-based meta learning. *arXiv preprint arXiv:2305.17373*.

Zixuan Zhang, Heba Elfardy, Markus Dreyer, Kevin Small, Heng Ji, and Mohit Bansal. 2023. Enhancing multi-document summarization with cross-document graph-based information extraction. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1696–1707, Dubrovnik, Croatia. Association for Computational Linguistics.

Zixuan Zhang, Hongwei Wang, Han Zhao, Hanghang Tong, and Heng Ji. 2021. EventKE: Event-enhanced knowledge graph embedding. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1389–1400, Punta Cana, Dominican Republic. Association for Computational Linguistics.

A Appendix

A.1 Relation Extraction Details

During the relation extraction with the OIE system, implicit triples and long relations can appear. We filter out both implicit triples and long relations (longer than five tokens) as it has been shown that these relations are noisy (Broscheit et al., 2020), and implicit relations cannot be used for token classification since they introduce tokens that are not present in the text. For example, if the OIE system is presented with the sentence: "President Biden right now stands really worried about future economic growth." it might extract (i) implicit triple ("Biden"; "is"; "President") and (ii) triple with long relation ("President Biden"; "right now stands really worried about"; "future economic growth"). Our heuristics would drop both extractions, and the implicit extraction would also be filtered out on account of not being in the order subject-relation-object in the input sentence. Also, we filter out all extractions that are incomplete triples, i.e., are missing either subject, relation, or object. If, after that, there are still multiple relation extractions for the same sentence, we try to merge the remaining relations. The merging process is designed to keep all the relations if the tokens are not shared between them. In the case of shared tokens, we keep only the relation extraction with the highest number of tokens that make up the relation. Finally, subject and object extractions are dropped, only the relations are kept, and if our heuristics filter out all the relation extractions for the sentence, we do not discard it but consider it a sentence without relations and use it for training as an example with all "outside" token labels based on IOB2 tagging scheme. We apply the OIE system, and this described post-processing, to each split of the source and target datasets.[5]

A.2 Experimental Setup Details

Training. The total GPU usage for all the experiments amounts to 1280 hours on *Ampere A100* GPU. We use the RoBERTa-base model with 125 million parameters. The input sequences are not lowercased. Since RoBERTa-base works on input split into subwords, the TD cross-entropy loss is adjusted to take into account only the first token of each tokenized word from the input sequence. Our preliminary experiments found incorporating

[5]Relation extractor is always shared between domains.

a learning rate scheduler is beneficial. We use a multiplicative learning rate scheduler with a multiplying factor of 0.99, which multiplies the learning rate in each epoch, lowering it throughout training. For each mini-batch, padding is applied to match the length of the longest example in the batch.

Hyperparameter Optimization. When training on the source domain, the *implicit* model is additionally optimized on the source validation set (based on the TD micro F1 score) with a simple grid search over the dimension of the trainable OIE-label embeddings d and the learning rate for it. We try dimensions of $10, 50, 100$, and 300 and learning rates of $0.0001, 0.00005$, and 0.00001. When performing target few-shot fine-tuning in *joint transfer* and *sequential transfer*, we fix the dimension to the one that produced the highest source validation set TD micro F1 score. In the *joint training* and *in-domain training* experiments, we arbitrarily fix the embedding size of the *implicit* model to 300 and 10 across all the experiments, respectively.

Auxiliary MLM Objective. We use a token-level masking probability of 15%, and the masking procedure is inherited from Devlin et al. (2019). Specifically, out of 15% of randomly chosen tokens, we mask 80% tokens, replace 10% tokens with random tokens from the vocabulary, and leave the remaining 10% of tokens unchanged.

A.3 Additional Results

Training Regime		ACE 2005 (0.706)			EDNYT (0.702)			EVEXTRA (0.893)		
		Vanilla	**Implicit**	**Explicit**	**Vanilla**	**Implicit**	**Explicit**	**Vanilla**	**Implicit**	**Explicit**
	0-Shot	0.234	0.237	**0.240**	0.392	0.399	**0.408**	0.650	0.650	**0.653**
sequential transfer	5-Shot	**0.294**	**0.294**	0.276	0.458	**0.466**	0.448	0.659	**0.661**	0.653
	10-Shot	0.372	**0.374**	0.330	0.512	**0.521**	0.490	0.688	**0.693**	0.680
	50-Shot	**0.511**	0.506	0.463	0.581	**0.592**	0.568	0.750	**0.764**	0.741
	100-Shot	0.538	**0.548**	0.501	0.605	**0.616**	0.584	0.786	**0.795**	0.773
	250-Shot	**0.587**	0.577	0.556	0.631	**0.644**	0.607	0.824	**0.835**	0.813
	500-Shot	**0.610**	0.609	0.586	**0.653**	0.652	0.640	0.852	**0.857**	0.836
in-domain training	5-Shot	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	10-Shot	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	50-Shot	0.464	**0.466**	0.417	**0.607**	0.601	0.597	0.768	**0.774**	0.757
	100-Shot	0.510	**0.529**	0.511	0.626	**0.632**	0.611	0.807	**0.812**	0.801
	250-shot	**0.570**	0.569	0.550	0.649	**0.654**	0.642	0.845	**0.847**	0.835
	500-Shot	0.598	**0.600**	0.584	0.660	0.658	**0.666**	0.858	**0.862**	0.854

(a) Without MLM.

Training Regime		ACE 2005 (0.706)			EDNYT (0.702)			EVEXTRA (0.893)		
		Vanilla	**Implicit**	**Explicit**	**Vanilla**	**Implicit**	**Explicit**	**Vanilla**	**Implicit**	**Explicit**
	0-Shot	0.226	0.233	**0.241**	0.396	**0.405**	0.389	0.658	0.657	**0.659**
sequential transfer	5-Shot	**0.311**	0.309	0.303	0.469	**0.480**	0.468	0.680	**0.681**	0.666
	10-Shot	0.390	**0.395**	0.359	**0.532**	0.531	0.509	**0.707**	0.702	0.697
	50-Shot	**0.525**	0.520	0.495	0.595	**0.600**	0.577	0.774	**0.775**	0.760
	100-Shot	0.549	**0.561**	0.519	0.612	**0.615**	0.599	**0.809**	**0.809**	0.791
	250-Shot	0.587	**0.591**	0.574	0.640	**0.645**	0.627	0.843	**0.845**	0.828
	500-Shot	**0.614**	**0.614**	0.604	**0.661**	**0.661**	0.645	**0.862**	0.861	0.848
in-domain training	5-Shot	0.010	0.018	**0.034**	0.007	0.012	**0.037**	0.019	0.046	**0.085**
	10-Shot	**0.002**	**0.002**	0.000	0.002	0.000	**0.003**	0.001	0.002	**0.007**
	50-Shot	0.366	**0.383**	0.288	0.548	**0.557**	0.552	0.685	**0.695**	0.649
	100-Shot	**0.545**	0.543	0.526	0.633	**0.638**	0.623	**0.796**	0.794	0.790
	250-shot	0.579	**0.584**	0.564	**0.661**	**0.661**	0.650	0.841	**0.844**	0.835
	500-Shot	**0.612**	0.607	0.596	0.670	**0.674**	0.671	**0.861**	**0.861**	0.852

(b) With MLM.

Table 4: TD domain transfer micro F1 scores when transferring from MAVEN as a source to ACE 2005, EDNYT, and EVEXTRA as targets (zero-shot, *sequential transfer*, and *in-domain training*, with six varying numbers of shots). The numbers in parentheses next to the target dataset are the in-domain performance scores when using all target training data. *In-domain training* results – target fine-tuning starting from PLM. *Sequential transfer* results – target fine-tuning starting from PLM trained for TD on MAVEN source training data. Table (a) shows results without an auxiliary MLM objective, while Table (b) depicts results with an auxiliary MLM training objective on target domain training data. *Implicit* and *explicit* models leverage MinIE relation labels, unlike the *vanilla* model. All reported results are averages of three runs.

Training Regime		ACE 2005 (0.706)			EDNYT (0.702)			EVEXTRA (0.893)		
		Vanilla	**Implicit**	**Explicit**	**Vanilla**	**Implicit**	**Explicit**	**Vanilla**	**Implicit**	**Explicit**
	0-Shot	0.005	0.003	0.003	0.007	0.009	0.005	0.003	0.002	0.004
joint training	5-Shot	0.008	0.001	0.003	0.014	0.015	0.011	0.004	0.001	0.002
	10-Shot	0.003	0.003	0.006	0.014	0.010	0.010	0.003	0.002	0.005
	50-Shot	0.006	0.005	0.011	0.018	0.009	0.012	0.005	0.008	0.007
	100-Shot	0.005	0.002	0.010	0.011	0.003	0.004	0.008	0.005	0.008
	250-Shot	0.009	0.003	0.010	0.010	0.004	0.007	0.010	0.008	0.004
	500-shot	0.013	0.010	0.006	0.009	0.001	0.006	0.008	0.002	0.005
joint transfer	5-Shot	0.010	0.005	0.004	0.018	0.012	0.018	0.004	0.006	0.002
	10-Shot	0.011	0.006	0.005	0.014	0.009	0.018	0.004	0.006	0.003
	50-Shot	0.007	0.007	0.005	0.005	0.002	0.006	0.001	0.005	0.004
	100-Shot	0.005	0.006	0.005	0.005	0.003	0.014	0.004	0.004	0.005
	250-Shot	0.012	0.007	0.014	0.009	0.015	0.009	0.005	0.001	0.011
	500-Shot	0.008	0.008	0.021	0.009	0.006	0.008	0.006	0.005	0.004
sequential transfer	5-Shot	0.014	0.016	0.014	0.022	0.024	0.025	0.012	0.003	0.003
	10-Shot	0.012	0.016	0.020	0.013	0.013	0.017	0.011	0.005	0.003
	50-Shot	0.011	0.006	0.003	0.004	0.010	0.006	0.011	0.010	0.003
	100-Shot	0.003	0.015	0.013	0.003	0.012	0.004	0.009	0.008	0.002
	250-Shot	0.007	0.006	0.012	0.004	0.012	0.013	0.009	0.005	0.005
	500-Shot	0.004	0.010	0.002	0.004	0.009	0.002	0.004	0.003	0.005
in-domain training	5-Shot	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.001
	10-Shot	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.001
	50-Shot	0.013	0.013	0.034	0.007	0.003	0.010	0.009	0.012	0.009
	100-Shot	0.009	0.006	0.012	0.001	0.008	0.010	0.004	0.004	0.007
	250-shot	0.001	0.004	0.017	0.012	0.010	0.007	0.003	0.005	0.006
	500-Shot	0.008	0.004	0.006	0.004	0.010	0.010	0.003	0.006	0.003

Table 5: Standard deviation of TD domain transfer micro F1 scores from Table 2. All reported results are averages of three runs.

Training Regime		ACE 2005 (0.706)			EDNYT (0.702)			EVEXTRA (0.893)		
		Vanilla	**Implicit**	**Explicit**	**Vanilla**	**Implicit**	**Explicit**	**Vanilla**	**Implicit**	**Explicit**
	0-Shot	0.005	0.003	0.003	0.007	0.009	0.005	0.003	0.002	0.004
sequential transfer	5-Shot	0.014	0.016	0.014	0.022	0.024	0.025	0.012	0.003	0.003
	10-Shot	0.012	0.016	0.020	0.013	0.013	0.017	0.011	0.005	0.003
	50-Shot	0.011	0.006	0.003	0.004	0.010	0.006	0.011	0.010	0.003
	100-Shot	0.003	0.015	0.013	0.003	0.012	0.004	0.009	0.008	0.002
	250-Shot	0.007	0.006	0.012	0.004	0.012	0.013	0.009	0.005	0.005
	500-Shot	0.004	0.010	0.002	0.004	0.009	0.002	0.004	0.003	0.005
in-domain training	5-Shot	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.001
	10-Shot	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.001
	50-Shot	0.013	0.013	0.034	0.007	0.003	0.010	0.009	0.012	0.009
	100-Shot	0.009	0.006	0.012	0.001	0.008	0.010	0.004	0.004	0.007
	250-shot	0.001	0.004	0.017	0.012	0.010	0.007	0.003	0.005	0.006
	500-Shot	0.008	0.004	0.006	0.004	0.010	0.010	0.003	0.006	0.003

(a) Without MLM.

Training Regime		ACE 2005 (0.706)			EDNYT (0.702)			EVEXTRA (0.893)		
		Vanilla	**Implicit**	**Explicit**	**Vanilla**	**Implicit**	**Explicit**	**Vanilla**	**Implicit**	**Explicit**
	0-Shot	0.005	0.002	0.008	0.005	0.001	0.012	0.001	0.005	0.001
sequential transfer	5-Shot	0.017	0.018	0.016	0.022	0.019	0.007	0.003	0.003	0.003
	10-Shot	0.028	0.018	0.020	0.012	0.013	0.006	0.003	0.004	0.007
	50-Shot	0.008	0.014	0.021	0.006	0.010	0.005	0.005	0.007	0.003
	100-Shot	0.009	0.012	0.012	0.001	0.006	0.002	0.001	0.004	0.004
	250-Shot	0.013	0.013	0.005	0.005	0.008	0.005	0.004	0.004	0.003
	500-Shot	0.006	0.007	0.015	0.008	0.003	0.007	0.002	0.003	0.002
in-domain training	5-Shot	0.003	0.016	0.032	0.005	0.011	0.032	0.007	0.037	0.072
	10-Shot	0.002	0.002	0.001	0.003	0.000	0.002	0.001	0.002	0.007
	50-Shot	0.042	0.042	0.067	0.008	0.010	0.022	0.026	0.006	0.033
	100-Shot	0.009	0.013	0.004	0.008	0.002	0.002	0.009	0.013	0.002
	250-shot	0.008	0.002	0.004	0.001	0.004	0.002	0.004	0.002	0.003
	500-Shot	0.018	0.008	0.014	0.002	0.002	0.001	0.001	0.002	0.002

(b) With MLM.

Table 6: Standard deviation of TD domain transfer micro F1 scores from Table 4. All reported results are averages of three runs.