Few-Shot Cross-Lingual Learning for Event Detection

Luis F. Guzman-Nateras[1], **Viet Dac Lai**[1],
Franck Dernoncourt[2], **and Thien Huu Nguyen**[1]
[1] Department of Computer Science, University of Oregon, Eugene, OR, USA
[2] Adobe Research, Seattle, WA, USA
{lfguzman,vietl,thien}@cs.uoregon.edu,
franck.dernoncourt@adobe.com

Abstract

Cross-Lingual Event Detection (CLED) models are capable of performing the Event Detection (ED) task in multiple languages. Such models are trained using data from a *source* language and then evaluated on data from a distinct *target* language. Training is usually performed in the standard supervised setting with labeled data available in the source language. The Few-Shot Learning (FSL) paradigm is yet to be explored for CLED despite its inherent advantage of allowing models to better generalize to unseen event types. As such, in this work we study the CLED task under an FSL setting. Our contribution is threefold: first, we introduce a novel FSL classification method based on Optimal Transport (OT); second, we present a novel regularization term to incorporate the global distance between the support and query sets; and third, we adapt our approach to the cross-lingual setting by exploiting the alignment between source and target data. Our experiments on three, syntactically-different, target languages show the applicability of our approach and its effectiveness at improving the cross-lingual performance of few-shot models for event detection.

1 Introduction

Event Detection (ED) is a significant sub-task within the larger task of Information Extraction (IE) in Natural Language Processing (NLP). Its core purpose is to identify the words, or phrases, that most clearly express the occurrence of an event, known as event *triggers*, and to correctly categorize them into a discrete set of classes. For instance, in the sentence:

*Frank **purchased** his dream house yesterday.*

the word "**purchased**" should be identified by an ED system as the trigger of a `Transaction:Transfer-Ownership` event type[1]. Event detection is a highly active research area which has been lately dominated by deep-learning-based approaches (Sha et al., 2018; Wadden et al., 2019; Zhang et al., 2019a; Yang et al., 2019a; Nguyen and Nguyen, 2019; Zhang et al., 2020; Liu et al., 2020; Lu et al., 2021). Most of these works use the standard supervised learning paradigm in which lots of labeled data is required during training. However, a significant limitation of models trained in this manner is their inability to properly generalize to new event types that were unobserved during training (Lai et al., 2020b).

Few-Shot Learning: In contrast to the supervised approach, Few-Shot Learning (FSL) proposes a training setting in which a model must quickly learn new concepts from just a few examples, similar to how humans can learn to detect and identify new objects after having observed only a couple of instances. During an FSL training iteration, a model is given a *support* set and a *query* set, each of which contains only a handful of examples for a set of classes. Then, the model is trained to predict the classes for the query samples based on the labeled support samples. Under these constrained training settings, supervised training easily results in model overfitting due to the limited availability of training data. Furthermore, in FSL, a model is evaluated on its ability to generalize to new, unobserved types. To achieve this, during testing an FSL model is provided with new support and query sets whose samples belong to entirely new classes never observed during training.

Typical FSL approaches consist of obtaining a vector representation for each sample and then performing classification based on the distance between such vectors, e.g., Matching Networks (Vinyals et al., 2016), Relation Networks (Sung et al., 2018), and Prototypical Networks (Snell et al., 2017). The key differences between these approaches often come down to the way the sample representations are generated, and how the distance between such representations is

[1] Event type example taken from ACE05 dataset.

Proceedings of the The 2nd Workshop on Multi-lingual Representation Learning (MRL), pages 16 - 27

determined.

FSL training allows a model to easily extend to new classes as it only needs to see a few labeled examples in order to successfully classify them. FSL has been applied successfully for many tasks. Recently, there have been several efforts that explore event detection under a few-shot learning setting (FSLED) (Lai et al., 2020a,b; Deng et al., 2020; Lai et al., 2021a,b; Cong et al., 2021; Shen et al., 2021; Chen et al., 2021).

Cross-Lingual Event Detection: Cross-Lingual Learning (CLL) is a paradigm that aims at transferring the knowledge from one language to another (Pikuliak et al., 2021). CLL can help overcome the lack of data availability that plagues many languages and allow for the creation of NLP-based tools that can benefit their communities.

As such, Cross-lingual Event Detection (CLED) aims at detecting and classifying event triggers with the added complexity of operating on two separate languages. These two languages are referred to as *source* and *target*, respectively. In standard *zero-shot* training, a CLED model is trained using labeled data belonging to the source language exclusively. Then, at testing time, data from the target language is used to evaluate the model's performance (M'hamdi et al., 2019; Majewska et al., 2021; Nguyen et al., 2021; Guzman-Nateras et al., 2022).

Contributions: A proper effort on CLED under FSL conditions has yet to be explored despite the potential advantages it could contribute to cross-lingual models. Hence, we recognize this opportunity and propose the novel Few-Shot Cross-Lingual Event Detection (FSCLED) task to integrate these two settings. We consider the following as our main contributions:

- To the best of our knowledge, this is the first effort at integrating the few-shot and cross-lingual settings for the event detection task. To provide foundation for future research, we first evaluate the performance of representative FSL methods (Vinyals et al., 2016; Snell et al., 2017; Sung et al., 2018) in this task.
- We propose a novel optimal-transport-based method for FSL classification that leverages the optimal alignment between the support and query samples.
- We address a limitation of traditional FSL methods by incorporating a novel regularization term that considers the global distance between the support and query sets.
- To adapt our approach to the cross-lingual setting, we promote language-invariant representation learning by integrating the distance between source and target data into our model.
- Our experiments on three diverse target languages (Arabic, Chinese, and Spanish) show that our approach improves the best-performing FSL methods in the new FSCLED setting and that our proposed training signals can be seamlessly incorporated with other FSL models to improve their performance on the challenging FSCLED task.

The rest of the paper is organized as follows: Section 2 provides a formal definition for FSCLED task, Section 3 describes the details our proposed approach, Section 4 presents the results of our experiments, and finally, we present our conclusions in Section 6.

2 Problem Definition

2.1 Few-shot Event Detection

We follow the same problem formulation as in prior work for few-shot ED (Lai et al., 2020b; Deng et al., 2020; Lai et al., 2021a). In particular, we cast event detection as a token classification task in which a model must learn to correctly classify the trigger tokens. In a standard FSL setting, an iteration involves a support set $\mathcal{S}$ and a query set $\mathcal{Q}$ that cover sample sentences for N distinct classes; each class is represented by $K \in [1, 10]$ examples. Additionally, for event detection, $\mathcal{S}$ and $\mathcal{Q}$ are extended with an additional negative, or non-event, type *NULL* (also with K examples) (Lai et al., 2021a). In this manner, given an input sentence along with an trigger candidate, an FSL model for ED should be able to predict whether the candidate is an event trigger as well as which event type is evoked by the trigger (if any).

Hence, the formal definition of the FSL task is as follows. The $\mathcal{S}$ and $\mathcal{Q}$ sets are defined by:

$$\mathcal{S} = \{(s_i^{j(\mathcal{S})}, t_i^{j(\mathcal{S})}, y_i^{j(\mathcal{S})})\}$$
$$\mathcal{Q} = \{(s_i^{j(\mathcal{Q})}, t_i^{j(\mathcal{Q})}, y_i^{j(\mathcal{Q})})\}$$

where $i \in [1, K]$[2], $j \in [0, N]$ ($j = 0$ is used for the non-event type), and a single sample

[2]We use the same number of samples for each class in both the support and query sets.

$(s_i^{j(\cdot)}, t_i^{j(\cdot)}, y_i^{j(\cdot)})$ contains a sentence $s_i^{j(\cdot)}$, a trigger candidate word $t_i^{j(\cdot)}$ in $s_i^{j(\cdot)}$, and an event label type $y_i^{j(\cdot)}$. As per FSL requirements, the label set used when training the model must be disjoint from those used when evaluating the model to properly assess the model's ability to generalize to unobserved classes.

2.2 Few-shot Cross-lingual Event Detection

Cross-Lingual Learning (CLL) methods (Pikuliak et al., 2021) emerged from the need to create NLP models for low-resource *target* languages that lack the required labeled data to perform supervised learning. The core idea is to train models using available labeled data from a high-resource *source* language with techniques that allow them to learn task-specific language-invariant features. The models are then evaluated on the desired target language without access to target-language labeled data during training. This setting is known as *zero-shot* cross-lingual transfer learning[3].

As such in the zero-shot cross-lingual ED task, the labeled samples used during training $\mathcal{D}_{train}$ and development $\mathcal{D}_{dev}$ belong to the source language while the ones used for testing $\mathcal{D}_{test}$ correspond to the target languages (M'hamdi et al., 2019; Majewska et al., 2021).

In this work, we combine the aforementioned *zero-shot* approach to cross-lingual evaluation with the added intricacy of the standard few-shot setting. During training, the models are presented with a support set $\mathcal{S}^{src}$ and a query set $\mathcal{Q}^{src}$ that belong to the source language. Then, at testing time, the support set $\mathcal{S}^{tgt}$ and query set $\mathcal{Q}^{tgt}$ are taken from the target language for evaluation. Furthermore, given the FSL setting, the label set used during training is disjoint from the label set for development and testing. We designate this novel task as Few-Shot Cross-Lingual Event Detection (FSCLED).

3 Model

As done in prior FSL models for ED (Lai et al., 2021a), our model for FSCLED involves two main components: an encoder E and a classifier C.

3.1 Encoder

The encoder's purpose is to obtain a representation vector $v_i^{j(\cdot)}$ for each sample in the support $\mathcal{S}$ and query $\mathcal{Q}$ sets:

$$v_i^{j(\cdot)} = E(s_i^{j(\cdot)}, t_i^{j(\cdot)}) \in \mathbb{R}^d$$

where d is the vector size, and $\cdot$ can be either $\mathcal{S}$ or $\mathcal{Q}$.

Following recent work on CLED, we leverage the pretrained multilingual language model (mLM) mBERT (Devlin et al., 2019) for our encoder to take advantage of its ability to induce language-invariant representations (Majewska et al., 2021). Additionally, we stack a Multi-Layer Perceptron (MLP) layer on top of the transformer outputs to create our multilingual encoder, called BERTMLP (Yang et al., 2019b). Then, we employ the vector representation for $t_i^{j(\cdot)}$ generated by BERTMLP to serve as the representation $v_i^{j(\cdot)}$.

3.2 Classifier

For convenience, let v^s and v^q be the representation vectors for the sample $s \in \mathcal{S}$ and $q \in \mathcal{Q}$, and $V^{(\mathcal{S})}$ and $V^{(\mathcal{Q})}$ be the sets of representation vectors for all samples in the support and query sets, respectively.

The classifier C aims to predict a label y^q for each instance q in the query set based on its representation v^q and the representations of the samples in the support set $V^{(\mathcal{S})}$:

$$y^q = C(v^q, V^{(\mathcal{S})})$$

Given the multilingual representations $v_i^{j(\cdot)}$, a feasible approach is to employ existing FSL models (e.g., Matching, Relation, or Prototypical networks) to perform classification in FSCLED. The models can then be trained using the standard cross-entropy loss.

3.2.1 Optimal Transport

We recognize, nonetheless, a potential issue with traditional FSL models in that they only consider local distances between individual pairs of samples in the support and query sets. In the case of Prototypical Networks (Snell et al., 2017), for example, the distance is between a query sample and a class prototype. Hence, if the overall global distance between the support and query sets is large, a small difference between the distances of two individual samples becomes less reliable to determine the label assignments. In turn, we argue that the global distances between $\mathcal{S}$ and $\mathcal{Q}$ should be minimized to improve the reliability of the distances between individual pairs for accurate FSCLED.

[3]Not to be confused with standard zero-shot learning where zero data for a new class is used by models to perform prediction.

To this end, we propose utilizing Optimal Transport (OT) (Villani, 2008) to estimate the distance between the support $\mathcal{S}$ and query $\mathcal{Q}$ sets for FSCLED. In broad terms, OT aims to find the most cost-effective transformation between two discrete probability distributions. Optimal transport employs a cost function to compute the cost of transforming data points from one distribution to the other. If a distance function (Euclidean, Cosine, etc.) is used as such cost function, the obtained minimum cost is known as the Wasserstein distance. Formally, OT solves the following optimization problem:

$$\pi^*(x,z) = \min_{\pi \in \prod(x,z)} \sum_{x \in \mathcal{X}} \sum_{z \in \mathcal{Z}} \pi(x,z)\ D(x,z)$$
$$\textbf{s.t. } x \sim P(x) \text{ and } z \sim P(z)$$

where $P(x)$ and $P(z)$ are probability distributions for the $\mathcal{X}$ and $\mathcal{Z}$ domains, and D is a distance-based cost function for mapping $\mathcal{X}$ to $\mathcal{Z}$, $D(x,z) : \mathcal{X} \times \mathcal{Z} \longrightarrow \mathbb{R}_+$. Finally, $\pi^*(x,z)$ is the optimal joint distribution over the set of all joint distributions $\prod(x,z)$ (i.e., the optimal transformation between $\mathcal{Z}$ and $\mathcal{X}$). The described OT optimization problem is, however, intractable as it requires optimizing over the infinite set $\prod(x,z)$. In practice, we instead solve an entropy-based relaxation of the discrete OT problem using the Sinkhorn algorithm (Cuturi, 2013).

3.2.2 Few-Shot Classification via OT

To adapt FSL classification into an OT formulation we consider the support $\mathcal{S}$ and query $\mathcal{Q}$ sets as the two domains to be transformed. Each sample in $\mathcal{S}$ and $\mathcal{Q}$ represents a data point in the corresponding distribution. The probability distributions $P(\mathcal{S})$ and $P(\mathcal{Q})$ are estimated using an *event-presence* module F. In our work, F is a feed-forward neural network (FFNN) with a single output and sigmoid activation that scores the likelihood that a trigger candidate word is actually an event trigger. F receives as input the vector representation of a trigger $v^{(\cdot)}$ from either $\mathcal{S}$ or $\mathcal{Q}$, and outputs a scalar in the range [0-1]. Then, the probability distributions for $\mathcal{S}$ and $\mathcal{Q}$ are obtained by computing the Softmax over F's outputs for the samples in each set:

$$P(\mathcal{S}) = \textit{Softmax}(F(V^{(\mathcal{S})}))$$
$$P(\mathcal{Q}) = \textit{Softmax}(F(V^{(\mathcal{Q})}))$$

To supervise the event-presence module F, we include the cross-entropy loss for event identification into the overall loss function:

$$\mathcal{L}_{ident} = \sum_{s \in \mathcal{S}} i^s \cdot \sigma(F(v^s)) + (1 - i^s) \cdot \sigma(1 - F(v^s))$$

where i^s is the golden binary variable to indicate if s corresponds to an event trigger or not, and σ is the sigmoid function.

In our model, the distance $D(q,s)$ between a sample in $q \in \mathcal{Q}$ and a sample $s \in \mathcal{S}$ is based on the Euclidean distance between their representation vectors v^s and v^q:

$$D(q,s) = \sqrt{\sum_{i \in d} (v_i^q - v_i^s)^2}$$

Once the OT algorithm converges, or the maximum number of iterations is reached, the obtained optimal alignment matrix π^* is a squared matrix with dimensions $((N+1)*K) \times ((N+1)*K)$ where each entry $\pi^*_{r,c}$ represents the alignment score between the r-th query sample and c-th support sample.

The conversion from matrix index (r, c) to event type (j) and sample number (i) can be computed in a straightforward manner as all samples from the same class (event type) are contiguous: $j = r//K, i = r\ \%\ K$ where // and % are the integer division and modulo operators.

To perform sample classification and train our FSCLED model, we first use the optimal alignment matrix π^* to compute a likelihood vector α for each query sample (i.e., the r-th) by performing class-based pooling with respect to the $N+1$ classes:

$$\alpha_r^j = \sum_{i \in [0, K-1]} \pi^*_{r,(j*K)+i}$$

where $j \in [0, N]$. As such, the resulting α_r vectors have $N+1$ dimensions. And the complete α matrix has a $((N+1)*K) \times (N+1)$ size. We then apply a Softmax operation over α_r to obtain a class distribution P_r for the r-th query sample: $P_r = \textit{Softmax}(\alpha_r)$. P_r will then be used for training and inference in our model. In particular, we use the negative log-likelihood loss as the main term of our overall training loss:

$$\mathcal{L}_{class} = -\sum_r P_r(y_r)$$

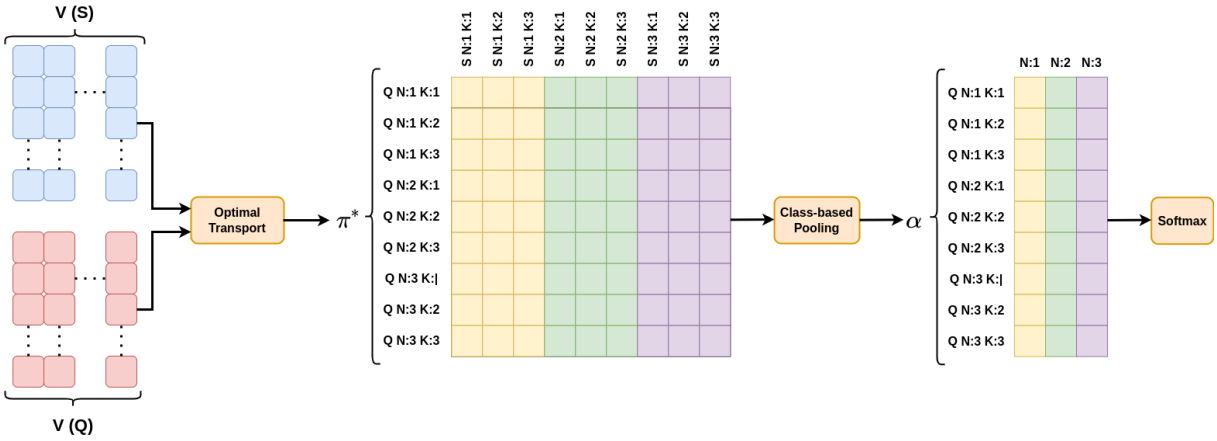

Figure 1: OT-based classification procedure example for a 3-way, 3-shot setting.

where y_r is the golden class for the r-th query example. Figure 1 shows a visualization of the described procedure for a 3-way, 3-shot setting. As such, a key distinction is that the class distribution P_r in our FSL method is obtained from the support-query alignment scores π^* in optimal transport. This is in contrast to previous FSL models where the class distributions tend to be computed directly from sample representations.

3.3 Support-Query Distance

In addition to our optimal-transport-based FSL classifier, we propose computing the Wasserstein distance between $\mathcal{S}$ and $\mathcal{Q}$ and including it into the loss function as a regularization term to minimize the overall distance between the support and query sets for reliable predictions. We obtain the aforementioned Wasserstein distance using the optimal alignment matrix π^*:

$$\mathcal{L}_{dist} = \sum_{s \in \mathcal{S}} \sum_{q \in \mathcal{Q}} \pi^*_{r,c} \, D(q, s)$$

where r and c are the matrix indexes for q and s, respectively.

3.4 Cross-Lingual Distance

To adapt our approach to the cross-lingual setting, we aim to encourage language-invariant representation learning by regularizing our model so the representation vectors of samples in the source and target languages are closer to each other in the embedding space.

Following the work by Guzman-Nateras et al. (2022), which leveraged OT to successfully align samples taken the source and target languages to improve adversarial language adaptation, we propose to further use OT to estimate the distance between samples in the source and target languages so that it can be included in the overall loss function as an additional regularization term for minimization.

To this end, given the unavailability of labeled data in the target language, we make use of unlabeled data – often readily available for most languages – instead. For convenience, let $\mathcal{R}$ and $\mathcal{T}$ represent the source-language and target-language data set respectively. In any given FSL training iteration, the support $\mathcal{S}$ and the query $\mathcal{Q}$ sets comprise the $\mathcal{R}$ set for the source language. To constitute the set representing the target language $\mathcal{T}$, we collect enough unlabeled samples to match the size of $\mathcal{R}$.

Thus, similarly to the OT formulation described in section 3.2.2 that computes the optimal alignment between two domains $\mathcal{S}$ and $\mathcal{Q}$, in this context we consider the source- and target-language data set $\mathcal{R}$ and $\mathcal{T}$ as the domains to be transformed. Subsequently, we employ our BERTMLP multilingual encoder to obtain representation vectors for the samples in both $\mathcal{R}$ and $\mathcal{T}$ that will serve as the inputs for the OT algorithm.

It is important to note that, due to the unavailability of the class information for the target-language samples $\mathcal{T}$ for training, it is less reliable to estimate the probability distribution $P(\mathcal{T})$ for the target language using the event-presence prediction module F as performed for $P(\mathcal{S})$ and $P(\mathcal{Q})$. Hence, we initialize $P(\mathcal{R})$ and $P(\mathcal{T})$ as uniform distributions for the OT computation in this case.

Under this setting, we solve the OT equation to obtain the optimal alignment matrix ρ^* between $\mathcal{R}$ and $\mathcal{T}$. The Wasserstein distance $\mathcal{L}_{cross}$ is then computed and integrated into the overall loss func-

tion for regularization:

$$\mathcal{L}_{cross} = \sum_{r \in \mathcal{R}} \sum_{t \in \mathcal{T}} \rho^*_{n,m} \, D(r,t)$$

where n and m are the matrix indexes for r and t, respectively.

3.4.1 Full Model

Finally, the overall loss function $\mathcal{L}$ used to train our Optimal-Transport-based Event Detection (OTED) model is: $\mathcal{L} = \mathcal{L}_{class} + \alpha\mathcal{L}_{ident} + \beta\mathcal{L}_{dist} + \gamma\mathcal{L}_{cross}$ where α, β, and γ are trade-off hyperparameters.

4 Experiments

4.1 Datasets

We use the ACE05 (Walker et al., 2006) and ERE05 (Song et al., 2015) datasets, which are frequently used as the standard benchmarks in cross-lingual event detection efforts (M'hamdi et al., 2019; Majewska et al., 2021; Nguyen et al., 2021; Guzman-Nateras et al., 2022), to evaluate our FSCLED models. In particular, we utilize data in three languages (English, Chinese, and Arabic) from ACE05 and two languages (English and Spanish) from ERE05. Both ACE05 and ERE05 organize their event classes in a hierarchical structure of types and subtypes. For example, in the `Transaction:Transfer-Ownership` class, `Transaction` is the main event type and `Transfer-Ownership` is the subtype. The two datasets have distinct label sets as ACE05 includes 33 event subtypes and ACE05-ERE has 38 event subtypes. Each language in the datasets has its own training/development/test split.

4.1.1 FSL Preprocessing

Standard datasets used for supervised learning, such as ACE05 and ERE05, can also be exploited for FSL by simulating a limited-data-availability setting via *episodic training* (Lai et al., 2021a). An *episode* is created by sampling a set of K examples from a small subset of classes N out of the total number of classes in the dataset. This setting is referred to as *N-way, K-shot* and N and K are usually selected in the range of 1 to 10.

Following previous work on FSL for ED (Lai et al., 2020b), we further truncate the training, development, and testing portions of the datasets for each language to satisfy the conditions for FSL: (1) the set of event types in the training data must be disjoint from those for the development and test data; (2) the types in each set must contain at least 5 samples (to facilitate 5+1-way 5-shot learning with the additional +1 class being used for non-triggers); and (3) the training set should have as many samples as possible.

Dataset	# Types	Removed Types
ACE05-English (train)	19	Justice:Extradite Justice:Pardon
ACE05-English (dev)	12	
ACE05-Chinese (test)	11	Life:Divorce
ACE05-Arabic (test)	9	Life:Be-Born Life:Divorce Personnel:Nominate
ERE05-English (train)	22	Business:Bankrupcy
ERE05-English (dev)	15	
ERE05-Spanish (test)	14	Personnel:Nominate

Table 1: Dataset preparation for FSCLED. The total number of remaining types is shown for each data section alongside the removed subtypes without a sufficient number of samples for episodic training.

Adapting these criteria to cross-lingual FSL, we separate the samples belonging to the `Business`, `Contact`, `Conflict`, and `Justice` types to be used for training purposes. Meanwhile, we leave the samples belonging to the `Life`, `Movement`, `Personnel`, and `Transaction` event types for development and testing. Furthermore, we remove any subtypes that do not contain enough samples to construct an episode (5 samples minimum). Table 1 shows the total number of remaining classes for each portion of data in different languages for our FSCLED setting. We also list the event subtypes that are removed to meet the criteria in each dataset portion. Note that, while the training label set must be disjoint from the development and testing label sets, there is no requirement for the latter two to be disjoint as done in (Lai et al., 2020b).

As the final step in our data preprocessing, we obtain the samples for the non-event type by selecting words, other than the actual triggers, from annotated sentences similar to the approach taken by Lai et al. (2020b).

4.2 Training Details

4.2.1 Episode Composition

In all our experiments, English is considered the sole source language as it is often used as the benchmark source language in cross-lingual efforts. As such, training and development episodes are constructed from English data. However, given the FSL constraints, their samples must come from

	Target Language								
	Chinese			Arabic			Spanish		
Model Version	P	R	F1	P	R	F1	P	R	F1
Relation	78.62	79.1	78.86	52.89	53.35	53.12	48.53	48.77	48.65
Matching	85.44	85.79	85.64	66.21	65.92	66.06	56.77	56.95	56.86
Prototypical	85.81	86.12	85.96	70.02	70.44	70.23	60.87	61.17	61.02
OTED (ours)	**86.05**	**86.29**	**86.17**	**70.66**	**70.98**	**70.82**	**62.25**	**62.49**	**62.37**

Table 2: Performance for cross-lingual few-shot event detection. English is the source language used for training. The experiments for Chinese and Arabic are done over ACE05 while ERE05 is used for Spanish.

disjoint label sets. Hence, in any training iteration, the samples used for both the support $\mathcal{S}$ and query $\mathcal{Q}$ sets are in English and belong to the training subtypes of the `Business, Contact, Conflict`, or `Justice` types. In contrast, during validation, $\mathcal{S}$ and $\mathcal{Q}$ will still be in English but their samples belong to the validation subtypes of the `Life, Movement, Personnel`, or `Transaction` types.

Furthermore, as cross-lingual models are evaluated on the target language, during testing, episodes are created from target-language data and their samples belong to the same types as the development episodes, i.e., the `Life, Movement, Personnel`, or `Transaction` types.

4.2.2 Additional Settings

We utilize a fixed 6-way (5 event types plus the non-event), 5-shot setting for all the experiments. We initialize our encoder E with the pre-trained `bert-base-multilingual-cased` transformer model (Devlin et al., 2019) and add a single linear layer followed by a hyperbolic tangent non-linearity on top. Our final encoder representations have 512 dimensions. All hyperparameters were tuned on the development data of the source language, and all reported values are the average obtained from five runs with different random seeds. Our fine-tuning process suggests the following values:

- AdamW (Loshchilov and Hutter, 2017) as the optimizer.
- Using 5 warm up epochs.
- Learning rate is set to $3e^{-4}$.
- The α, β and γ hyper-parameters are set to $0.1, 0.01$, and 0.01 respectively.
- The batch size is set to 16.
- 512 for the dimensionality of the layers in the feed-forward networks.
- A dropout of 10% for added regularization during training.

4.3 Results

We compare our Optimal-Transport-based Event Detection (OTED) model, against three typical FSL models adapted to FSCLED as the baselines: Matching networks (Vinyals et al., 2016), Prototypical networks (Snell et al., 2017), and Relation networks (Sung et al., 2018). All models utilize the same mBERT-based encoder for a fair comparison. We use English as the source language during training as it is recurrently utilized the source-language benchmark (M'hamdi et al., 2019; Majewska et al., 2021) due to its high-resource availability.

Our main experiment results are presented in Table 2 which shows that our OTED model consistently outperforms the best-performing baselines in every target language: Chinese ($+0.21\%$), Arabic ($+0.59\%$), and Spanish ($+1.35\%$). We believe these results validate OTED as a suitable and effective alternative for FSCLED.

Furthermore, an additional benefit of OTED's training signals (i.e., the loss terms $\mathcal{L}_{ident}$, $\mathcal{L}_{dist}$, and $\mathcal{L}_{cross}$) is that they can be directly integrated into any existing FSL methods. Thus, we conduct a supplementary set of experiments where we integrate the loss function terms from OTED into Relation, Matching, and Prototypical networks (i.e., combining our training signals in OTED with the standard cross-entropy losses of such FSL baselines). The performance for these integrated models are presented in Table 3. Comparing the corresponding performance in Tables 2 and 3, it is evident that integrating OTED with traditional FSL methods leads to overall performance improvement across different target languages and FSL models, further demonstrating the benefits and applicability of OTED for FSCLED.

Model Version	Target Language		
	Chinese	**Arabic**	**Spanish**
Relation + OTED	79.36	53.41	48.89
Matching + OTED	85.88	66.21	56.97
Prototypical + OTED	**86.42**	**71.11**	**62.43**

Table 3: Model performance for integrating OTED into traditional FSL methods. F1 scores are reported.

Model	Target Language		
	Chinese	Arabic	Spanish
OTED (full)	86.17	70.82	62.37
-$\mathcal{L}_{dist}$	85.63	70.57	61.85
-$\mathcal{L}_{cross}$	85.45	70.22	61.78
-$\mathcal{L}_{dist}$ -$\mathcal{L}_{cross}$	85.25	69.44	61.19
-$\mathcal{L}_{ident}$ - $\mathcal{L}_{dist}$ - $\mathcal{L}_{cross}$	84.67	68.21	60.65

Table 4: Ablation results over the test data.

4.4 Ablation study

To evaluate the contribution of different proposed components (i.e., $\mathcal{L}_{ident}$, $\mathcal{L}_{dist}$, and $\mathcal{L}_{cross}$), we perform an ablation study whose outcomes are presented in Table 4. The left-most column indicates the components being removed from the overall loss $\mathcal{L}$. The first two rows show the performance when either the Wasserstein-distance loss term, i.e., $\mathcal{L}_{dist}$ or $\mathcal{L}_{cross}$ is removed. As expected, removing any of them hurts the performance of OTED across different target languages. This demonstrates the importance of considering the global distances between query and support sets, and the necessity of adapting to the cross-lingual setting by leveraging unlabeled target-language data. Furthermore, the performance of OTED suffers even more when both $\mathcal{L}_{dist}$ and $\mathcal{L}_{cross}$ are excluded.

Similarly, when $\mathcal{L}_{ident}$ is removed in the last row, the performance is also further reduced, dropping significantly by more than 1.5% for Chinese and Arabic compared to the full model. Note that removing $\mathcal{L}_{ident}$ has deeper implications as, in such case, the event-presence module F is not trained. In turn, the $P(\mathcal{S})$ and $P(\mathcal{Q})$ distributions for the support and query sets cannot be estimated reliably and are instead initialized using uniform distributions in the OT computation. These results thus confirm the usefulness of the event identification loss to support the OT computation in our model.

5 Related Work

Event detection has been thoroughly studied over the years. Early ED efforts were based on handcrafted features (Ahn, 2006; Ji and Grishman, 2008; Patwardhan and Riloff, 2009; Liao and Grishman, 2010a,b; Hong et al., 2011; McClosky et al., 2011; Li et al., 2013; Miwa et al., 2014; Yang and Mitchell, 2016). More recently, deep learning techniques such as recurrent neural networks (Nguyen et al., 2016a; Sha et al., 2018; Nguyen and Nguyen, 2019), convolutional neural networks (Nguyen and Grishman, 2015; Chen et al., 2015; Nguyen et al., 2016b), graph convolutional networks (Nguyen and Grishman, 2018a; Yan et al., 2019), adversarial networks (Hong et al., 2018)(Zhang et al., 2019b), pre-trained language models (Wadden et al., 2019; Zhang et al., 2019a; Yang et al., 2019a; Zhang et al., 2020; Liu et al., 2020), and generative models (Lu et al., 2021) have been prevalent. Nevertheless, these works study ED under a supervised or semi-supervised setting.

Alternatively, ED was recently formulated as a few-shot task (Lai et al., 2021a). In a short time, several methods have been proposed using a variety of techniques such as meta-learning(Deng et al., 2020; Shen et al., 2021), cross-task prototyping (Lai et al., 2021a), dependency graphs (Lai et al., 2021b), causal modeling (Cong et al., 2021), and label dependency via conditional random fields (Chen et al., 2021).

Previous works on cross-lingual ED generally make use of cross-lingual resources such as bilingual dictionaries or parallel corpora (Muis et al., 2018; Liu et al., 2019) to address the differences between languages. More recent approaches exploit the language-invariant characteristics of pre-trained multilingual language models (Hambardzumyan et al., 2020) along with complementary features such as label dependency (M'hamdi et al., 2019), verb-class knowledge (Majewska et al., 2021), and class-aware cross-lingual alignment (Nguyen et al., 2021).

Optimal transport has also been recently used in cross-lingual settings for information extraction tasks such as event co-reference resolution (Phung et al., 2021) and event detection (Guzman-Nateras et al., 2022). However, the amalgamation of the few-shot and cross-lingual settings creates unique challenges that have not been tackled by any related work. Consequently, our proposed use of OT differs from related works as it addresses the global alignment between the support and query sets for few-shot learning and between source and target languages for the cross-lingual setting.

6 Conclusion

We explore a novel few-shot cross-lingual setting for event detection that combines the limited training-data conditions of FSL with zero-shot cross-lingual transfer learning. We provide the performance of typical FSL models as the foundations for future research. More importantly, we introduce a novel method for FSCLED that leverages the optimal alignment between query and support sets obtained via OT to perform FSL classification. Our method is complemented by two additional regularization terms that aim at integrating the global distance between support and query sets and fostering language-invariant representations by leveraging unlabeled data in the target language. Our experiments on three target languages demonstrate the advantages of our approach and its general applicability to traditional FSL models. As future work, we intend to extend our method to other related tasks in IE such as relation extraction.

7 Limitations

As is the case for any research effort, the scale of our work is restricted by time and resource limitations. Supplementary experiments with diverse source/target language pairs could provide a more comprehensive overview of our method's performance and additional insight into its strengths and weaknesses. Episode composition also plays an key role during few-shot training which can introduce some variance in the results. Furthermore, the cross-lingual setting and casting the problem as a token classification task places some important restrictions as prior knowledge of event triggers is required even for target-language data (only the trigger is required, not its label) which could limit the applicability of our method for some low-resource languages. Finally, considerable GPU resources are required to be able to train our model, particularly in order to fit the multilingual transformer encoder.

Acknowledgement

This research has been supported by the Army Research Office (ARO) grant W911NF-21-1-0112 and the NSF grant CNS-1747798 to the IUCRC Center for Big Learning. This research is also based upon work supported by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via IARPA Contract No. 2019-19051600006 under the Better Extraction from Text Towards Enhanced Retrieval (BETTER) Program. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ARO, ODNI, IARPA, the Department of Defense, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein. This document does not contain technology or technical data controlled under either the U.S. International Traffic in Arms Regulations or the U.S. Export Administration Regulations.

References

David Ahn. 2006. The stages of event extraction. In *Proceedings of the Workshop on Annotating and Reasoning about Time and Events*.

Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. 2021. Honey or poison? solving the trigger curse in few-shot event detection via causal intervention. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8078–8088, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yubo Chen, Liheng Xu, Kang Liu, Daojian Zeng, and Jun Zhao. 2015. Event extraction via dynamic multi-pooling convolutional neural networks. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

Xin Cong, Shiyao Cui, Bowen Yu, Tingwen Liu, Wang Yubin, and Bin Wang. 2021. Few-Shot Event Detection with Prototypical Amortized Conditional Random Field. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 28–40, Online. Association for Computational Linguistics.

Marco Cuturi. 2013. Sinkhorn distances: Lightspeed computation of optimal transport. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*.

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*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*.

Luis Guzman-Nateras, Minh Van Nguyen, and Thien Nguyen. 2022. Cross-lingual event detection via

optimized adversarial training. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5588–5599, Seattle, United States. Association for Computational Linguistics.

Karen Hambardzumyan, Hrant Khachatrian, and Jonathan May. 2020. The role of alignment of multilingual contextualized embeddings in zero-shot cross-lingual transfer for event extraction. In *Collaborative Technologies and Data Science in Artificial Intelligence Applications*.

Yu Hong, Jianfeng Zhang, Bin Ma, Jianmin Yao, Guodong Zhou, and Qiaoming Zhu. 2011. Using cross-entity inference to improve event extraction. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

Yu Hong, Wenxuan Zhou, Jingli Zhang, Guodong Zhou, and Qiaoming Zhu. 2018. Self-regulation: Employing a generative adversarial network to improve event detection. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 515–526, Melbourne, Australia. Association for Computational Linguistics.

Heng Ji and Ralph Grishman. 2008. Refining event extraction through cross-document inference. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

Viet Lai, Franck Dernoncourt, and Thien Huu Nguyen. 2021a. 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, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Viet Dac Lai, Franck Dernoncourt, and Thien Huu Nguyen. 2020a. Exploiting the matching information in the support set for few shot event classification. In *Advances in Knowledge Discovery and Data Mining*.

Viet Dac Lai, Minh Van Nguyen, Thien Huu Nguyen, and Franck Dernoncourt. 2021b. *Graph Learning Regularization and Transfer Learning for Few-Shot Event Detection*.

Viet Dac Lai, Thien Huu Nguyen, and Franck Dernoncourt. 2020b. 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.

Qi Li, Heng Ji, and Liang Huang. 2013. Joint event extraction via structured prediction with global features. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

Shasha Liao and Ralph Grishman. 2010a. Filtered ranking for bootstrapping in event extraction. In *Proceedings of the International Conference on Computational Linguistics (COLING)*.

Shasha Liao and Ralph Grishman. 2010b. Using document level cross-event inference to improve event extraction. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

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

Jian Liu, Yubo Chen, Kang Liu, and Jun Zhao. 2019. Neural cross-lingual event detection with minimal parallel resources. 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 738–748, Hong Kong, China. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2017. Fixing weight decay regularization in adam.

Yaojie Lu, Hongyu Lin, Jin Xu, Xianpei Han, Jialong Tang, Annan Li, Le Sun, Meng Liao, and Shaoyi Chen. 2021. Text2Event: Controllable sequence-to-structure generation for end-to-end event extraction. 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 2795–2806, Online. Association for Computational Linguistics.

Olga Majewska, Ivan Vulić, Goran Glavaš, Edoardo Maria Ponti, and Anna Korhonen. 2021. Verb knowledge injection for multilingual event processing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*.

David McClosky, Mihai Surdeanu, and Christopher Manning. 2011. Event extraction as dependency parsing. In *BioNLP Shared Task Workshop*.

Meryem M'hamdi, Marjorie Freedman, and Jonathan May. 2019. Contextualized cross-lingual event trigger extraction with minimal resources. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*.

Makoto Miwa, Paul Thompson, Ioannis Korkontzelos, and Sophia Ananiadou. 2014. Comparable study of event extraction in newswire and biomedical domains. In *Proceedings of the International Conference on Computational Linguistics (COLING)*.

Aldrian Obaja Muis, Naoki Otani, Nidhi Vyas, Ruochen Xu, Yiming Yang, Teruko Mitamura, and Eduard Hovy. 2018. Low-resource cross-lingual event type detection via distant supervision with minimal effort. In *Proceedings of the 27th International Conference on Computational Linguistics*.

Minh Van Nguyen, Tuan Ngo Nguyen, Bonan Min, and Thien Huu Nguyen. 2021. Crosslingual transfer learning for relation and event extraction via word category and class alignments. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5414–5426, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Thien Huu Nguyen, Kyunghyun Cho, and Ralph Grishman. 2016a. Joint event extraction via recurrent neural networks. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*.

Thien Huu Nguyen, Lisheng Fu, Kyunghyun Cho, and Ralph Grishman. 2016b. A two-stage approach for extending event detection to new types via neural networks. In *Proceedings of the 1st ACL Workshop on Representation Learning for NLP (RepL4NLP)*.

Thien Huu Nguyen and Ralph Grishman. 2015. Event detection and domain adaptation with convolutional neural networks. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

Thien Huu Nguyen and Ralph Grishman. 2018a. Graph convolutional networks with argument-aware pooling for event detection. In *AAAI*.

Trung Minh Nguyen and Thien Huu Nguyen. 2019. One for all: Neural joint modeling of entities and events. In *Association for the Advancement of Artificial Inteligence(AAAI)*.

Siddharth Patwardhan and Ellen Riloff. 2009. A unified model of phrasal and sentential evidence for information extraction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Duy Phung, Hieu Minh Tran, Minh Van Nguyen, and Thien Huu Nguyen. 2021. Learning cross-lingual representations for event coreference resolution with multi-view alignment and optimal transport. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 62–73, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Matúš Pikuliak, Marián Šimko, and Mária Bieliková. 2021. Cross-lingual learning for text processing: A survey. In *Expert Systems with Applications*.

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 Association for the Advancement of Artificial Intelligence (AAAI)*.

Shirong Shen, Tongtong Wu, Guilin Qi, Yuan-Fang Li, Gholamreza Haffari, and Sheng Bi. 2021. Adaptive knowledge-enhanced Bayesian meta-learning for few-shot event detection. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2417–2429, Online. Association for Computational Linguistics.

Jake Snell, Kevin Swersky, and Richard Zemel. 2017. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*.

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.

Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip H. S. Torr, and Timothy M. Hospedales. 2018. Learning to compare: Relation network for few-shot learning.

C. Villani. 2008. *Optimal Transport: Old and New*. Grundlehren der mathematischen Wissenschaften. Springer Berlin Heidelberg.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, koray kavukcuoglu, and Daan Wierstra. 2016. Matching networks for one shot learning. In *Advances in Neural Information Processing Systems*.

David Wadden, Ulme Wennberg, Yi Luan, and Hannaneh Hajishirzi. 2019. Entity, relation, and event extraction with contextualized span representations. 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)*.

Christopher Walker, Stephanie Strassel, Julie Medero, and Kazuaki Maeda. 2006. Ace 2005 multilingual training corpus. In *Technical report, Linguistic Data Consortium*.

Haoran Yan, Xiaolong Jin, Xiangbin Meng, Jiafeng Guo, and Xueqi Cheng. 2019. Event detection with multi-order graph convolution and aggregated attention. 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 5766–5770, Hong Kong, China. Association for Computational Linguistics.

Bishan Yang and Tom M. Mitchell. 2016. Joint extraction of events and entities within a document context. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*.

Sen Yang, Dawei Feng, Linbo Qiao, Zhigang Kan, and Dongsheng Li. 2019a. Exploring pre-trained language models for event extraction and generation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

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

Junchi Zhang, Yanxia Qin, Yue Zhang, Mengchi Liu, and Donghong Ji. 2019a. Extracting entities and events as a single task using a transition-based neural model. In *IJCAI*.

Tongtao Zhang, Heng Ji, and Avirup Sil. 2019b. Joint Entity and Event Extraction with Generative Adversarial Imitation Learning. *Data Intelligence*, 1(2):99–120.

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