Cross-Lingual Event Detection via Optimized Adversarial Training

Anonymous ACL submission

Abstract

In this work, we focus on Cross-Lingual Event Detection (CLED) where a model is trained on data from a *source* language but its performance is evaluated on data from a second, *target*, language. Most recent works in this area have harnessed the language-invariant qualities displayed by pre-trained Multi-lingual Language Models (MLM). Their performance, however, reveals there is room for improvement as they mishandle delicate cross-lingual instances. We leverage the use of unlabeled data to train a Language Discriminator (LD) to discern between the source and target languages. The LD is trained in an adversarial manner so that our encoder learns to produce refined, language-invariant representations that lead to improved CLED performance. More importantly, we optimize the adversarial training by only presenting the LD with the most *informative* samples. We base our intuition about *what* makes a sample informative on two disparate metrics: sample similarity and event presence. Thus, we propose using Optimal Transport (OT) as a solution to naturally combine these two distinct information sources into the selection process. Extensive experiments on 8 different language pairs, using 4 languages from unrelated families, show the flexibility and effectiveness of our model that achieves new state-of-the-art results.

1 Introduction

Event Detection (ED) is an important sub-task within the broader Information Extraction (IE) task. ED consists in being able to identify the words, commonly referred to as *triggers*, that denote the occurrence of events in a sentence, and classify them into a discrete set of event types. For example, in the sentence *"Jamie **bought** a car yesterday."*, *bought* is considered the trigger of a TRANSACTION:TRANSFER-OWNERSHIP event type. It is a very well studied task in which there have been lots of previous research efforts (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; Nguyen and Grishman, 2015; Chen et al., 2015; Nguyen et al., 2016a,b; Sha et al., 2018; Zhang et al., 2019; Yang et al., 2019; Nguyen and Nguyen, 2019; Zhang et al., 2020; Xiang and Wang, 2019).

Nonetheless, ED remains quite a challenging task as the context in which a trigger word occurs can change its corresponding type completely, and the same event might also be expressed by entirely different words/phrases. The vast majority of the aforementioned efforts, however, are limited to a monolingual setting, i.e approaches that perform ED on text belonging to a single language. Additionally, most ED-related research focuses on a small set of popular languages, such as Chinese or English. This, in turn, means that most of the available annotated data belongs to these, aptly named, *high-resource languages*. Data scarcity becomes a critical problem for *low-resource languages* for which the amount of available training data is minimal or non-existent. Consequently, some approaches have proposed taking advantage of the widely available unlabeled data in a semi-supervised manner (Muis et al., 2018).

Cross-lingual ED (CLED) proposes the more challenging scenario of creating models that effectively perform ED on data belonging to more than one language. This entails additional challenges for a CLED model. For instance, trigger words present in one language might not exist in another one. An example of this phenomenon are verb conjugations where some tenses only exist in some languages, which is commonplace in ED as event triggers are usually related to the verbs in a sentence. Another problematic issue are trigger words with different meanings that are each distinct words in other languages. For example, the word "*juicio*" in Spanish can be either "*judgement*" or "*trial*" in English, depending on the context. These, and other similar, issues make CLED a challenging task.

A compelling approach to creating a cross-lingual model is to use *transfer learning* which attempts to transfer the performance of a model trained on a *source* language onto a second *target* language. The general idea is leveraging the existing high-quality annotated data available for a high-resource language to train a model in a way that allows it to learn the language-invariant characteristics of the task at hand, ED in this case, so that it also performs effectively on text from a second language. Prior work on transfer learning for CLED has relied on pre-trained Multilingual Language Models (MLMs), such as mBERT (Devlin et al., 2019), to take advantage of their innate language-invariant qualities. Yet, their performance still shows room for improvement as they are unable to handle the difficult instances, unique to cross-lingual settings, mentioned earlier. We identify a significant shortcoming of previous CLED efforts in that they do not exploit the abundant supply of unlabeled data: even though MLMs are trained on immense amounts of it, unlabeled data is not used when fine-tuning for the ED task. It is our intuition that by integrating unlabeled data into the training process, the model is exposed to more language context which should help deal with issues such as verb variation and multiple connotations.

As such, in this work we propose using Adversarial Language Adaptation (ALA), inspired by Adversarial Domain Adaptation (ADA) (Ganin and Lempitsky, 2015), which aims at creating cross-lingual models able to successfully perform ED on both a *source* language and a *target* language. The key idea is to generate language-invariant representations that are not-indicative of language but remain informative for the ED task. A fundamental characteristic of our ALA approach is its lack of requirements for annotated data in the target language. Instead, unlabeled data, from both the source and target languages, is used to train a Language Discriminator (LD) network that learns to discern between the two. The *adversarial* part comes from the fact that the encoder is trained in the reverse direction of the LD: as the LD becomes better at distinguishing between languages, the encoder learns to generate more language-invariant representations in an attempt to *fool* the LD.

Furthermore, contrary to past uses of ADA where the same importance is given to all unlabeled samples, we recognize that such course of action is sub-optimal as certain samples are bound to be more informative for the discriminator than others. For example, we would like to present the LD with the samples that allow it to learn the fine-grained distinctions between the source and target languages, instead of relying on syntactic differences. Moreover, we suggest it would be beneficial for the LD, and the encoder, to be trained with examples containing events, instead of non-event samples, as then the presence of an event can be incorporated into the generated representations.

Hence, we propose refining the adversarial training process by only keeping the most informative examples while disregarding less useful ones. Our intuition as to *what* makes samples more informative for CLED is two-fold: First, we presume that presenting the LD with examples that are too different makes the discrimination task too simple. As mentioned previously, we would like the LD to learn a fine-grained distinction between the source and target languages which, in turn, improves the language-invariance of the encoder's representations. Thus, we suggest presenting the LD with examples that have similar contextual semantics. Second, we consider sentences containing events to be more relevant for the LD. Accordingly, such sentences should have a larger probability of being selected for ALA training.

One challenge of using these two criteria for our ALA sample selection process is that they come with two different measures which are hard to combine. In consequence, we propose using Optimal Transport (OT) (Villani, 2008) as a natural solution to simultaneously incorporate both the similarity between examples and the likelihood of the samples containing an event into a single framework. OT is, in broad terms, the problem of finding out the cheapest transformation between two discrete probability distributions. It requires a cost function to determine the cost of transforming a data point in one distribution into a data point in the second distribution. When the cost function is based on a valid distance function, the minimum cost is known as the Wasserstein distance.

Therefore, we cast sample selection as an OT problem in which we attempt to find the best alignment between the samples from the source and target languages. Similarity between samples is scored through the Euclidean distance of their contextualized vector representations. The probability distributions are obtained by introducing an Event Presence (EP) prediction network trained to determine whether a sentence in the batch contains an event or not, its normalized outputs are used as inputs for the OT alignment algorithm.

For our experiments, we focus on the widely used ACE05 and ACE05-ERE datasets (Walker et al., 2006) which, in conjuction, contain event-annotations in 4 different languages: English, Spanish, Chinese, and Arabic. We work on 8 different language pairs by selecting different languages as the source and target. Our proposed model obtains new state-of-the-art results with considerable performance improvements (+ 2-3% in F1 scores) over competitive baselines and previously published results (M'hamdi et al., 2019). These results demonstrate our model's efficacy and applicability at creating CLED systems.

2 Model

2.1 Problem Definition

Following prior works [cite], we treat ED as a sequence labeling problem. Given a set $\mathcal{D}$ of word sequences $w_i = \{w_{i1}, w_{i2}, ..., w_{in-1}, w_{in}\}$ and their corresponding label sequences $y_i = \{y_{i1}, y_{i2}, ..., y_{in-1}, y_{in}\}$, we use an encoder network E to obtain a contextualized vector representation of the words in the input sequence $\mathbf{h_i} = E(w_i) = \{h_{i1}, h_{i2}, ..., h_{in-1}, h_{in}\}$. Then, we feed the representations h_i into a prediction network P to compute a distribution over the set of possible labels and train it in a supervised manner using the negative log-likelihood function $\mathcal{L}_P$:

$$\mathcal{L}_P = -\sum_{i=1}^{|D|}\sum_{j=1}^{n} \log P(y_{ij}|h_{ij}) \quad (1)$$

In the cross-lingual transfer-learning setting, the data used to train the model and the data on which the model is tested come from different languages known as the *source* and *target*, respectively. As such, we deal with two datasets $\mathcal{D}_{src}$ and $\mathcal{D}_{tgt}$. We assume that we do not have access to the gold labels of the target language y_{tgt}, other than to evaluate our CLED model at testing time.

Our goal is to define a model able to generate language-invariant word representations that are refined enough so that cross-lingual issues, such as the ones described previously, are properly handled.

2.2 Baseline Model

Here we briefly describe the BERT-CRF model (M'hamdi et al., 2019) which was the previous state-of-the-art and serves as our baseline. As its name implies, BERT-CRF uses mBERT (Devlin et al., 2019) as its encoder which generates robust, contextualized representations for words from different languages. For words that are split into multiple word-pieces, the average of the representation vectors for all comprising sub-pieces is used as the representation of the full word.

For classification purposes, instead of assigning the labels of each token independently, BERT-CRF leverages using a Conditional Random Field (CRF) layer on top of the prediction network to better capture the interactions between the label sequences. As such, the representation vectors h_i of the words in the sequence are fed to a CRF layer which finds the optimal label sequence.

2.3 Adversarial Language Adaptation

The pre-trained versions of MLMs like mBERT or XLM-RoBERTa generate contextualized representations with a certain degree of language-invariance. This can be confirmed by their successful application in cross-lingual settings (M'hamdi et al., 2019). However, a problem with these works is that they are unable to learn the nuances of the target language such as verb variations that do not exist in the source language used to train them. It is our intuition, nonetheless, that refining these representations to achieve an even greater level of language-invariance would be ultimately beneficial in a cross-lingual system.

As such, we propose Adversarial Language Adaptation (ALA), a technique inspired by Adversarial Domain Adaptation (ADA) (Ganin and Lempitsky, 2015) which is used to create domain-invariant models. With ALA, we aim to refine our multilingual transformer encoder so that its obtained representations display better language-invariant qualities. Our ALA framework consists in including an additional module called the *Language Discriminator* whose purpose is to learn language-dependant features and be able to differentiate between the samples from either the source or the target languages.

Given that annotated events are not needed to train the LD, we can use data from both $\mathcal{D}_{src}$ and $\mathcal{D}_{tgt}$. An auxiliary dataset D_{aux} =

$\{(w_1, l_1), \ldots, (w_{2m}, l_{2m})\}$ is created where w_i is a text sequence from either $\mathcal{D}_{src}$ or $\mathcal{D}_{tgt}$, and l_i is a language label. The cardinality of D_{aux} is $|D_{aux}| = 2m$, where m is equal to the batch size. Text samples $w_1 \ldots w_m \in \mathcal{D}_{src}$, and samples $w_{m+1} \ldots w_{2m} \in \mathcal{D}_{tgt}$. As described earlier, the encoder E receives the text sequences and produces a sequence of contextualized representations $E(w_i) = h_i = \{h_{i0}, h_{i1}, h_{i2}, \ldots, h_{in}\}$ where h_{i0} is the representation of the *[CLS]* token added at the beginning of every input sequence.

In our work, the LD is a a simple Multi-Layer Perceptron(MLP) network that takes h_{i0} as input and produces a single sigmoid output. It's trained with the usual *binary cross-entropy* loss function objective:

$$LD_{loss} = \arg\min_{LD} \; \mathcal{L}(LD(h_{i0}), l_i) \tag{2}$$

As the LD learns to distinguish between the source and target languages, we want to concurrently train the encoder to "fool" the discriminator. In other words, the encoder must learn to generate representations that are language-invariant enough that the LD is unable to classify them while still remaining predictive for event-trigger classification. We optimize the following loss:

$$\arg\min_{E,C} \sum_{j=1}^{n} (\mathcal{L}(C(h_{ij}), y_{ij})) - \lambda \mathcal{L}(LD(h_{i0}, l_i)) \tag{3}$$

Where C refers to the CRF-based classifier network and λ is a hyperparameter.

Equation 3 is implemented by using a Gradient-Reversal Layer (GRL)(Ganin and Lempitsky, 2015) which acts as the identity during the forward pass, but reverses the direction of the gradients during the backward pass. The first term in Equation 3 can, of course, only be applied for annotated data from the source language.

The GRL is applied to the input vectors, h_{i0}, of the LD. This way, the LD is being trained to differentiate between the two languages while the encoder is trained in the opposite direction, i.e. to generate sequence representations that are harder to discriminate.

2.4 Adversarial Training Optimization

ADA has already been shown to be effective at generating domain-invariant models(Naik and Rose, 2020). However, in regular ADA training, all samples in a batch, from both the source and target domains, are treated equally. That is, all samples are used as examples for the discriminator to learn how to better discern between the two domains. We propose that ADA effectiveness can be further improved by carefully selecting the samples with which to train the discriminator. We argue that some samples might be more informative than others and that, by only using such informative samples during training, better adaptation results can be achieved.

In the context of CLED, where the objective is to create a language-invariant model, we base our notion as to *what* makes a sample more informative on two factors. First, we argue that presenting the LD with examples from the source and target language that are too dissimilar makes its task easier which, in turn, leads to the LD not learning the fine-grained distinctions between the languages. Instead, we propose using samples whose vector representations h_{i0} are close to each other in the embedding space. The intuition for this being that, as representations capture the contextual semantics of the samples, closer representations correspond to more similar examples. Second, we suggest that presenting the LD with samples containing events should make the encoder incorporate task-specific information into its representations.

2.4.1 Optimal Transport

We propose using Optimal Transport (OT) as a natural way to combine our two metrics into a single framework for sample selection. OT can be described as finding the cheapest transportation cost between two discrete probability distributions. Formally, it solves the following optimization problem:

$$\pi^*(s,t) = \min_{\pi \in \prod(s,t)} \sum_{s \in \mathcal{S}} \sum_{t \in \mathcal{T}} \pi(s,t) \; C(s,t) \; ds \; dt \tag{4}$$

$$\textbf{s.t. } s \sim p(s) \text{ and } t \sim q(t)$$

Where $\mathcal{S}$ and $\mathcal{T}$ are two domains with probability distributions $p(s)$ and $q(t)$, and C is a cost function for mapping $\mathcal{S}$ to $\mathcal{T}$, $C(s,t) : \mathcal{S} \times \mathcal{T} \longrightarrow \mathrm{R}_+$. Finally, $\pi^*(s,t)$ is the optimal joint distribution over the set of all joint distributions $\prod(s,t)$. The problem described by Equation 4 is, of course, intractable. Therefore, we use instead the Sinkhorn algorithm (Cuturi, 2013) which is an entropy-based relaxation of the discrete OT problem.

2.4.2 Problem Formulation

We formulate the OT problem as follows: the domains $\mathcal{S}$ and $\mathcal{T}$ are defined as the representation vectors of the text samples in either the source h^s_{i0} or the target h^t_{j0} languages. We use the L2 distance between these representations as the cost function:

$$C(h^s_{i0}, h^t_{j0}) = ||h^s_{i0} - h^t_{j0}||^2_2 \quad (5)$$

To define the marginal probability distributions $p(s)$ and $q(t)$ for the $\mathcal{S}$ and $\mathcal{T}$ domains, we propose including an Event-Presence (EP) prediction module and use its normalized likelihood scores as the probability distributions for $\mathcal{S}$ and $\mathcal{T}$. Thus, the auxiliary dataset D_{aux} is augmented to include an event-presence label e_i for each sample, $D_{aux} = \{(w_1, l_1, e_1), \ldots, (w_{2m}, l_{2m}, e_{2m})\}$, and the EP module is trained to optimize the following loss:

$$EP_{loss} = \arg\min_{EP} \ \mathcal{L}(EP(h_{i0}), e_i) \quad (6)$$

The probability distributions $p(s)$ and $p(t)$ are the computed as follows:

$$p(s) = Softmax(EP(h^s_{i0}) \mid l_i == s) \quad (7)$$

$$p(t) = Softmax(EP(h^t_{i0}) \mid l_i == t) \quad (8)$$

2.4.3 Sample Selection

We use the OT solution matrix π^*, where an entry $\pi^*(s,t)$ represents the optimal cost of transforming data point $s \in \mathcal{S}$ into $t \in \mathcal{T}$, to compute an the overall similarity score v_i of a sample $h_{i0} \in \mathcal{S}$ to the samples in the target domain $\mathcal{T}$ by using the average distance:

$$v_i = \frac{\sum_j^m \pi^*(h^s_{i0}, h^t_{j0})}{m} \quad (9)$$

Correspondingly, we compute an overall similarity score v_j of each sample $h_{j0} \in \mathcal{T}$ to the samples in the source domain $\mathcal{S}$:

$$v_j = \frac{\sum_i^m \pi^*(h^s_{i0}, h^t_{j0})}{m} \quad (10)$$

Lastly, we select a fraction, hyperparameter γ, of samples with the best similarity scores from both the source and target languages, and only use these selected samples during ALA training.

2.5 OACLED Model

We train our Optimized Adversarial Cross-Lingual Event Detection (OACLED) model end-to-end with the following loss objective:

$$L_{full} = CRF_{loss} + \alpha LD_{loss} + \beta EP_{loss} \quad (11)$$

where α and β are trade-off hyperparameters.

3 Experiments

3.1 Datasets

We evaluate our model on the ACE05 (Walker et al., 2006) dataset which includes annotated event-trigger data in 3 languages: English, Chinese and Arabic. To include an additional language in our experiments, we also evaluate on the ERE version of ACE05 which has annotated data in English and Spanish. The ACE05 and ACE05-ERE versions, however, do not share the same label set: ACE05 involves 33 distinct event types while ACE05-ERE involves 38 event types. Dataset characteristics can be found in Appendix A. We follow the same data pre-processing and splits as in previous work(M'hamdi et al., 2019) to ensure a fair comparison.

3.2 Main Results

In our experiments, we work with 8 distinct language pairs by selecting each of the available languages as either the source or target language: *English-Chinese, Chinese-English, English-Arabic, Arabic-English, Chinese-Arabic, Arabic-Chinese, English-Spanish, and Spanish-English*. The *Chinese-Spanish, Spanish-Chinese, Arabic-Spanish*, and *Spanish-Arabic* language combinations are unavailable due the previously mentioned incompatibility between the event type sets in ACE05 and ACE05-ERE.

Tables 1 and 2 show the results of our experiments on the ACE05 and ACE05-ERE datasets, respectively.

We compare our OACLED model against 2 relevant baselines. BERT-CRF (M'hamdi et al., 2019), and XLM-R-CRF which is equivalent in all regards to BERT-CRF except that it uses XLM-RoBERTa as the encoder. The cross-lingual experiments in the original BERT-CRF paper included results for English being used as the source language, and Chinese and Arabic used as targets. The corresponding entries in Table 1 were taken directly from their paper. In our experiments, we use *bert-base-cased*

		Target		
Source	**Model**	**English**	**Chinese**	**Arabic**
English	**BERT-CRF**	X	68.5	30.9
	XLM-R-CRF	X	70.49	43.54
	OACLED	X	**74.64**	**44.86**
Chinese	**BERT-CRF**	37.52	X	**35.05**
	XLM-R-CRF	41.72	X	32.76
	OACLED	**45.77**	X	34.48
Arabic	**BERT-CRF**	40.1	58.78	X
	XLM-R-CRF	45.22	61.76	X
	OACLED	**47.98**	**63.13**	X

Table 1: Results on the ACE05 dataset.

		Target	
Source	**Model**	**English**	**Spanish**
English	**BERT-CRF**	X	43.28
	XLM-R-CRF	X	46.79
	OACLED	X	**47.69**
Spanish	**BERT-CRF**	39.8	X
	XLM-R-CRF	45.61	X
	OACLED	**47.5**	X

Table 2: Results on ACE05-ERE dataset.

and *xlm-roberta-base* for the encoders, parameters are tuned on the development data of the source language, and all entries are the average of five runs.

From Tables 1 and 2, we can observe a substantial performance increase by performing the trivial change of replacing BERT with XLM-RoBERTa as the encoder. Furthermore, our OACLED model clearly and consistently outperforms the baselines for all language pairings, with the exception of the *Chinese-Arabic* pair. We attribute this to the impaired performance of XLM-RoBERTa as the encoder for that specific pair as can be confirmed by the poor performance of the XLM-R-CRF baseline on the same configuration. Most importantly, OACLED's improvement over the XLM-R-CRF baseline is present in every configuration, which confirms the effectiveness of our optimized approach to ALA training.

3.3 Ablation Study

We identify 2 main components in our approach: using ALA to create refined language-invariant representations, and optimizing the adversarial training process by selecting a subset of samples chosen with OT to incorporate our measures of informativeness into the sample selection process. Of course, removing ALA training entirely restores the model to the baseline. However, adversarial training optimization via OT has various aspects to it. In order to understand the contribution of these aspects, we explore four different models: *OACLED-OT* presents the effects of removing sample selection entirely and using all available samples to train the LD; *OACLED-L2* uses a constant distance between the unlabeled samples instead the standard L2 distance used in the Sinkhorn algorithm; *OACLED-EP* completely removes the EP module and a uniform distribution is used as the probability distributions for both languages; finally, *OACLED-ED-Loss* keeps the EP module, but removes its EP_{loss} term from Equation 11. The performance results of these models is presented in Table 3. Due to space limitations, we present the results of experiments using English as the sole source language. We, however, found consistency in the displayed effects for different source/target language configurations.

Model version	**Target Language**		
English	**Chinese**	**Arabic**	**Spanish**
OACLED-OT	70.94	40.55	44.96
OACLED-L2	71.35	41.79	44.39
OACLED-EP	73.08	42.81	46.99
OACLED-EP-Loss	72.93	43.4	46.35
OACLED	**74.64**	**44.86**	**47.69**

Table 3: Ablation experiment results

As expected, removing the sample selection through OT leads to the worst performance drop. This highlights the importance of selecting informative examples for the LD. Furthermore, removing the cost function also hurts performance greatly, which shows that a proper distance function is needed for the OT algorithm to work effectively. While the effects of removing the EP module and its corresponding loss term are not of the same magnitude, they are still significant. These results support our claim for the need and utility of all the components in our approach, showing that their inclusion is crucial in achieving state-of-the-art performance.

3.4 Language Model Finetuning

The key contribution of our approach is to exploit unlabeled data in the target language, which is usually abundant, by introducing it into the training process to improve our model's language-invariant qualities.

To confirm the utility of our approach, Table 4 contrasts our model's performance against a baseline whose encoder has been finetuned with the same unlabeled data using the standard masked language model objective.

Model Version	Target Language		
English	**Chinese**	**Arabic**	**Spanish**
Finetuned XLM-R	71.06	43.71	**47.82**
OACLED	**74.64**	**44.86**	47.69

Table 4: OACLED performance versus a baseline using an encoder finetuned with unlabeled data.

It can be observed that our model outperforms the finetuned baseline in two out of the three target languages. Additionally, the difference in performance in those two instances is considerably larger (3.58% and 1.15%), than the setting in which the baseline performs better (0.13%).

3.5 Analysis

3.5.1 Learned Representation Distances

First, we look at the distance between the sentence-level representations h_{i0} generated by the encoder for different source/target language pairs. Figure 1 shows a plot of such distances using cosine distance as the distance function.

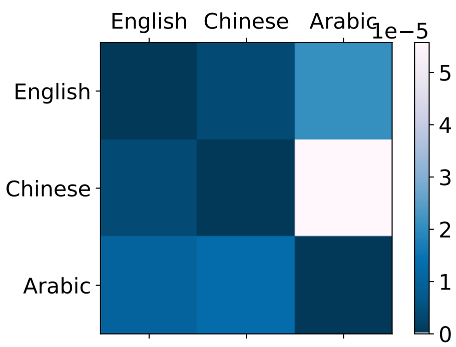

Figure 1: Distance between sentence representations for different language pairs.

When computing the correlation with the performance results in Table 1, we obtain a score $R = -0.6616$, meaning there is moderate negative correlation between the distance of the representations and model performance, i.e. closer representations lead to better performance.

Similarly, Table 5 shows a comparison of the distances between the representations generated by OACLED and those obtained by the XLM-R-CRF baseline.

We observe that OACLED representations are closer, by several orders of magnitude, than those obtained by the baseline. This supports our claim that our model's encoder generates more refined

	Cosine Distance	
Source/Target	**Baseline**	**OACLED**
English/Chinese	3.64e-3	**3.93e-6**
English/Arabic	7.71e-2	**2.08e-5**
English/Spanish	5.4e-3	**5.3e-6**
Chinese/English	3.62e-3	**3.87e-6**
Arabic/English	4.16e-2	**1.02e-5**
Spanish/English	6.87e-3	**1.49e-5**

Table 5: Comparison of representation-vector distances for language pairs between our model and the baseline.

language-invariant representations than those obtained by the default version of XLM-RoBERTa.

3.5.2 Access to Labeled Target Data

Previously, we discussed how a key feature of our approach is that it does not require annotated data in the target language and, instead, leverages the use of unlabeled data which is readily available. Nonetheless, we also explore the performance of our model in the event that there exists a small amount of annotated target data available for training. Figure 2 shows the results of our experiments when using different amounts of labeled target data during training.

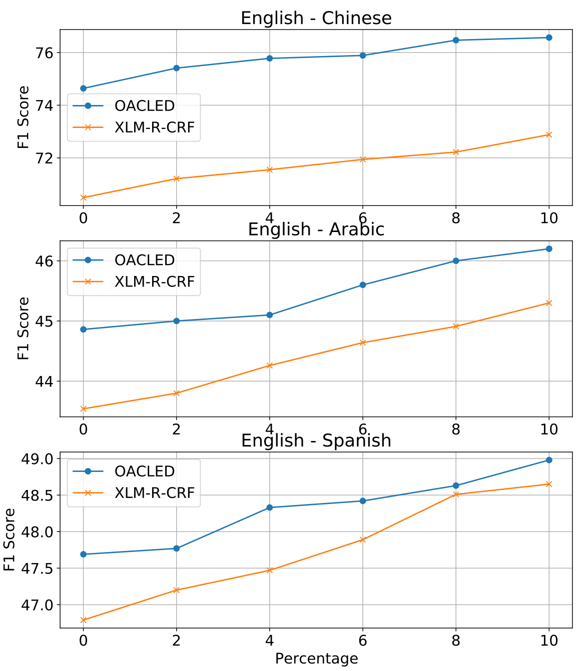

Figure 2: Model performance when training on small quantities of labeled target data. The X axis presents the percentage (0 - 10%) of data used out of the entire training set of the target language.

It can be observed that OACLED consistently outperforms the baseline even when there is some availability of annotated data. Additionally, performance steadily increases as more and more data is used. This conforms to expectations, and confirms that having labeled data in the target language available for training is ultimately beneficial to the model's performance.

3.5.3 Case Study

Next, we look into our model's predictions and analyse instances where it outperforms the baseline to exemplify the advantages of dealing with optimized language-invariant representations. We identify two important patterns.

First, our model is able to better classify events in the target language that involve trigger words that have distinct connotations that depend on context. Specially those that are two distinct words in the source language. For example, the Spanish word *"juicio"* can have two distinct meanings that are different words in English: *"trial"* and *"judgement"*. Our model correctly classifies it as a JUSTICE:TRIAL-HEARING-type trigger in the sentence *"Dos llamados a **juicio** fueron hechos por un jurado federal investigador"*, meanwhile the baseline fails to even recognize it as a trigger. Another example is the word *"detenido"*, an adjective that can mean both *"detained"*, in a criminal context, and *"stopped"*, as in halted. Our model correctly classifies it in the sentence *"Padilla no debería permanecer **detenido** durante meses alejado de otros reos"* as a JUSTICE:ARREST-JAIL trigger while the baseline fails to detect the event.

Second, our model can correctly classify different verb conjugation variants that do not exist in the source language. For instance, our model correctly recognizes the words *"venderlos", "vender", "vendes", and "vendedor"* (variants of the verb *"to buy"*) as TRANSACTION:TRANSFER-OWNERSHIP triggers whereas the baseline incorrectly classifies them as being of the TRANSACTION:TRANSFER-MONEY type. A similar example are the trigger words *"matar", "mató", "homicidio", "asesinato"*, all of which refer to the act of killing or murdering. Our model correctly tags them as LIFE:DIE events while the baseline incorrectly classifies them as CONFLICT:ATTACK.

These findings illustrate how, by introducing additional context in the form of unlabeled data, our model is able to learn fine-grained word representations that better capture the semantics of the words in the target language, and successfully deals with difficult cross-lingual issues.

4 Related Work

Feature-based methods were the basis of early ED approaches (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 recent efforts have primarily made use of deep learning techniques (Nguyen and Grishman, 2015; Chen et al., 2015; Nguyen et al., 2016a,b; Sha et al., 2018; Zhang et al., 2019; Yang et al., 2019; Nguyen and Nguyen, 2019; Zhang et al., 2020),

Works on CLED generally make use of cross-lingual resources employed to address the differences between languages such as bilingual dictionaries or parallel corpora (Muis et al., 2018; Liu et al., 2019) and, more recently, pre-trained MLMs (M'hamdi et al., 2019; Hambardzumyan et al., 2020). Unlike these approaches, our method leverages using unlabeled data to hone the language-invariant qualities of the pre-trained MLMs.

Additional examples of downstream applications of Cross-lingual Learning (CLL) are document classification (Holger and Xian, 2018), named entity recognition (Xie et al., 2018) and part-of-speech tagging (Cohen et al., 2011). For a thorough review on CLL, we refer the reader to (Pikuliak et al., 2021).

Finally, our ALA approach was inspired by models in domain adaptation research (Ganin and Lempitsky, 2015; Naik and Rose, 2020). Our method improves upon these approaches optimizing the adversarial training process by selecting the most informative examples from the unlabeled data.

5 Conclusion

In this work we present OACLED, a new model for cross-lingual event detection that leverages the use of ADA and OT to achieve new state-of-the-art performance. Our experiments on 8 different language pairs demonstrate OACLED's robustness and effectiveness at generating refined language-invariant representations that allow for better event detection results. Our analysis of its intermediate outputs and predictions confirm that OACLED's representations are indeed closer to each other and that this proximity translates into better handling of difficult cross-lingual instances.

References

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

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)*.

Shay B. Cohen, Dipanjan Das, and Noah Smith. 2011. Unsupervised structure prediction with nonparallel multilingual guidance. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Marco Cuturi. 2013. Sinkhorn distances: Lightspeed computation of optimal transport. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, page 2292–2300, Red Hook, NY, USA. Curran Associates Inc.

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

Yaroslav Ganin and Victor Lempitsky. 2015. Unsupervised domain adaptation by backpropagation. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1180–1189, Lille, France. PMLR.

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*.

Schwenk Holger and Li Xian. 2018. A corpus for multiligual document classification in eight languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC)*.

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)*.

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)*.

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, 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.

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)*, pages 656–665, Hong Kong, China. Association for Computational Linguistics.

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*.

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.

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)*.

Trung Minh Nguyen and Thien Huu Nguyen. 2019. One for all: Neural joint modeling of entities and events. In *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)*.

Matúš Pikuliak, Marián Šimko, and Mária Bieliková. 2021. Cross-lingual learning for text processing: A survey. *Expert Systems with Applications*, 165:113765.

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)*.

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

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

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

Jiateng Xie, Zhilin Yang, Graham Neubig, Noah A. Smith, and Jaime G. Carbonell. 2018. Neural cross-lingual named entity recognition with minimal resources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

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. 2019. Exploring pre-trained language models for event extraction and generation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

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

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*.

A Appendix A

A.1 Dataset Characteristics

Dataset	Language	Split	Sentences	Events
ACE05	English	Train	19,240	4,419
		Dev	902	468
		Test	676	424
	Chinese	Train	6,841	2,926
		Dev	526	217
		Test	547	190
	Arabic	Train	2,555	1,793
		Dev	301	230
		Test	262	247
ACE05-ERE	English	Train	14,219	6,419
		Dev	1,162	552
		Test	1,129	559
	Spanish	Train	7,067	3,272
		Dev	556	210
		Test	546	269

Table 6: Dataset statistics.

B Reproducibility Checklist

- **Source Code**: Upon the acceptance, we will release the source code via a public GitHub repository.

- **Computing Infrastructure**: In this work, we use a single Tesla V100-SXM2 GPU with 32GB memory operated by Red Hat Enterprise Linux Server 7.8 (Maipo). PyTorch 1.4.0 is used to implement the models.

- **Evaluation Metric**: We report F1 for trigger classification computed using the seqeval [1] framework for sequence labeling evaluation based on the CoNLL-2000 shared task, complying with previous work (M'hamdi et al., 2019). The reported results are the average performance of 5 model runs with different random seeds.

- **(Hyper-)parameters**: Our full model has 278.5M parameters. However, the vast majority of these come from the XLM-Roberta transformer (278M parameters), the rest of our model accounts for $< 500K$ parameters.

  We fine-tune the hyper-parameters for our OA-CLED model using the development data. We suggest the following values for fine-tuning:

  – AdamW as the optimizer.
  – Using 5 warm up epochs.
  – A learning rate of $1e^{-5}$ for the transformer parameters and of $1e^{-4}$ for the rest of the parameters. We arrived at this values after searching among $[1e^{-6}, 3e^{-6}, 1e^{-5}, 3e^{-5}, 1e^{-4}, 3e^{-4}]$.
  – A batch size of 16, chosen between $[8, 10, 16, 24, 32]$.
  – 300 for the dimensionality of the layers in feed-forwards networks, chosen from [100, 200, 300, 400, 500].
  – A $\gamma = 0.5$ for the percentage of samples used in adversarial training.
  – A $\lambda = 0.001$ as the scaling factor of the GRL layer.
  – An $\alpha = 1$ and $\beta = 0.001$ as the trade-off parameters of the LD loss and ED loss, respectively.
  – A dropout of 10% for added regularization during training.

[1] https://github.com/chakki-works/seqeval