Hybrid Knowledge Transfer for Improved Cross-Lingual Event Detection via Hierarchical Sample Selection

Luis F. Guzman-Nateras[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,thien}@cs.uoregon.edu,
franck.dernoncourt@adobe.com

Abstract

In this paper, we address the Event Detection task under a zero-shot cross-lingual setting where a model is trained on a source language but evaluated on a distinct target language for which there is no labeled data available. Most recent efforts in this field follow a direct transfer approach in which the model is trained using language-invariant features and then directly applied to the target language. However, we argue that these methods fail to take advantage of the benefits of the data transfer approach where a cross-lingual model is trained on target-language data and is able to learn task-specific information from syntactical features or word-label relations in the target language. As such, we propose a hybrid knowledge-transfer approach that leverages a teacher-student framework where the teacher and student networks are trained following the direct and data transfer approaches, respectively. Our method is complemented by a hierarchical training-sample selection scheme designed to address the issue of noisy labels being generated by the teacher model. Our model achieves state-of-the-art results on 9 morphologically-diverse target languages across 3 distinct datasets, highlighting the importance of exploiting the benefits of hybrid transfer.

1 Introduction

Event Detection (ED) is a sub-task of the encompassing Information Extraction (IE) Natural Language Processing (NLP) task. The main objective of ED is to detect and categorize the *event triggers* in a sentence, i.e., the words that most clearly indicate the occurrence of an event. Event triggers are known to be frequently related to the verb in a sentence (Majewska et al., 2021). However, they can also be other parts of speech such as nouns or adjectives. For instance, in the sentence *"The ceremony was chaired by the **former** Secretary of State"*, an ED system should recognize *former* as the trigger of a `Personnel:End-Position` event[1].

Generating labeled data for IE tasks such as ED can be a long and expensive endeavor. As such, most labeled ED datasets pertain to a small set of popular languages (e.g., English, Chinese, Spanish). In turn, labeled data is scarce or non-existent for a vast majority of languages. This imbalance in annotated data availability has prompted many research efforts into zero-shot cross-lingual transfer learning which attempts to transfer knowledge obtained from annotated data in a high-resource *source* language to a low-resource *target* language for which no labeled data is available. There are two predominant knowledge-transfer paradigms employed by such cross-lingual methods: *Data transfer* and *Direct transfer*.

Approaches that adhere to the *data transfer* paradigm generate pseudo-labeled data in the target language and then train a model on such data. This pseudo-training data can be constructed by mapping the gold source labels into parallel, or translated, versions of the source data, or by leveraging source-trained models to annotate unlabeled target data. Since models in this category are trained on the target language, they can directly exploit word-label relations and other target-language-specific information such as word order and lexical features (Xie et al., 2018). However, annotated parallel corpora are extremely scarce, and misaligned or incorrect translations introduce noise that affects the model performance.

In contrast, *direct-transfer-based* approaches aim at creating cross-lingual models by training them with delexicalized, language-independent features obtained from the labeled, source-language data. The resulting language-agnostic models can then be applied directly to unlabeled data in the target language.

In recent years, direct transfer has become the favored transfer paradigm as such models have less

[1]Event type taken from ACE05 dataset.

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 5414–5427

need for cross-lingual resources and can be applied to a broader range of languages. As such, previous research efforts on Cross-Lingual Event Detection (CLED) have mostly focused on the direct transfer approach (M'hamdi et al., 2019; Majewska et al., 2021) and, in consequence, have failed to exploit the aforementioned advantages of training with target-language data.

More recent approaches have attempted to address this issue by incorporating unlabeled target-language data into the training process. For example, Nguyen et al. (2021) propose a class-aware, cross-lingual alignment mechanism where they align examples from the source and target languages based on class information. Guzman-Nateras et al. (2022) instead propose to improve standard Adversarial Language Adaptation (ALA) (Joty et al., 2017; Chen et al., 2018) by only presenting the language discriminator with *informative* samples. Despite their improved results, these models only learn task-related information from the source language and fail to make use of the potentially useful information contained in word-label relations in the target language. Furthermore, previous studies on similar tasks have shown that, even for direct transfer methods, lexical features are useful if the source and target languages are close to each other (Tsai et al., 2016).

Given that the data transfer and direct transfer paradigms are orthogonal, in this paper we present a *hybrid transfer* approach for cross-lingual event detection that (1) exploits the desirable features of both and (2) minimizes their respective shortcomings. For this purpose, we propose a *knowledge distillation* framework which has already been proven effective on similar cross-lingual tasks (Wu et al., 2020a,b; Liang et al., 2021; Chen et al., 2021). In our proposed framework, a teacher model is trained using a direct transfer approach (i.e., with language-invariant features obtained from annotated source data) and applied to unlabeled target-language data. Then, this pseudo-labeled data is utilized to train a student model so that it benefits from the advantages of the data transfer paradigm.

Nonetheless, we recognize that the pseudo-labels obtained from the teacher model are prone to containing noisy predictions which can be hurtful for student training. To address this issue, we argue that the teacher model should produce more dependable predictions on target-language examples that share some similarities with their source-language counterparts. As such, we propose to improve the teacher-student learning process by restricting student training to samples with such desirable characteristics. We perform our training-sample selection in a hierarchical manner: First, we leverage Optimal Transport (OT, Villani, 2008) to compute similarity scores between batch samples in the source and target languages. Only samples with similarity scores above a certain threshold are selected in this first step. OT has already been shown to be effective at estimating cross-lingual similarities for sample selection (Phung et al., 2021; Guzman-Nateras et al., 2022). Then, in the second step, we make use of Cross-domain Similarity Local Scaling (CSLS, Conneau et al., 2018) to refine our sample selection. CSLS provides an enhanced measure to obtain reliable matches between samples in the source and target languages by addressing the *hubness* phenomenon that plagues nearest-neighbor-based pair-matching methods. The student model is then trained on the hierarchically-selected target-language samples exclusively.

In order to validate our approach, we compare our model's performance against current state-of-the-art models for CLED. For this purpose, we report our results on the most commonly used CLED benchmarking datasets: ACE05 (Walker et al., 2006) and ACE05-ERE (Song et al., 2015). These datasets, in conjunction, contain ED annotations for 3 distinct target languages. Our experimental results show that our approach consistently outperforms such state-of-the-art CLED models. Additionally, we further evaluate the flexibility and applicability of our model by leveraging the recently released MINION dataset (Pouran Ben Veyseh et al., 2022) which contains ED annotations for 8 typologically different languages.

The remainder of this document is organized as follows: section 2 presents the definition of the ED task and an in-depth description of our model and approach, section 3 includes the main results from our experiments and related analysis, section 4 provides a review of previous relevant work, and finally, section 5 presents our conclusions.

2 Model

2.1 Event Detection: Problem Definition

We follow a similar approach to previous CLED efforts (M'hamdi et al., 2019; Majewska et al., 2021; Guzman-Nateras et al., 2022) and model the ED task as a sequence labeling problem.

Given a group of sentences $\mathcal{S} = \{s_1, s_2, \ldots, s_n\}$ where each of such sentences is considered as a sequence of tokens $s_i = \{t_{i1}, t_{i2}, \ldots, t_{im}\}$ accompanied by a corresponding label sequence $y_i = \{y_{i1}, y_{i2}, ..., y_{im}\}$, the main idea is to train a model to generate token-level contextualized representations which can then be used to predict token-level labels.

In broad terms, a sequence labeling model consists of an encoder $\mathcal{E}$ and a classifier $\mathcal{C}$. The encoder consumes a sequence of input tokens t_i and outputs a sequence of contextualized representations h_i (Eq. 1). These representations are then fed to the classifier which produces a probability distribution over all of the possible types. A candidate label is selected by choosing the type with the largest probability. The model loss $\mathcal{L}_\mathcal{C}$ is then computed via negative log-likelihood with the classifier-selected labels and the expected *gold* labels (Eq. 2).

$$h_{i1}, h_{i2}, \ldots, h_{im} = \mathcal{E}(t_{i1}, t_{i2}, \ldots, t_{im}) \quad (1)$$

$$\mathcal{L}_\mathcal{C} = -\frac{1}{n * m} \sum_{i=1}^{n} \sum_{j=1}^{m} \log \mathcal{C}(y_{ij}|h_{ij}) \quad (2)$$

2.1.1 Zero-shot Cross-lingual Event Detection

In a cross-lingual setting, different languages are utilized during the training and testing phases. The language utilized during training is referred to as the *source* language. Once training is complete, the model is tested on the so-called *target* language.

A zero-shot setting further assumes that there is no labeled data in the target language to be leveraged during training. Nonetheless, raw, unlabeled target-language text can usually be collected without major difficulties. As such, in our work, we assume the availability of two distinct sets of sentences during training: the labeled source sentences $\mathcal{S}_{src}$ and unlabeled target sentences $\mathcal{S}_{tgt}^{unl}$. For model evaluation purposes, we leverage a set of labeled target-language sentences.

2.2 Hybrid Knowledge Transfer

As mentioned in Section 1, we propose to combine the direct transfer and data transfer approaches by leveraging a *Knowledge Distillation* framework. Knowledge distillation was originally proposed as a way to compress models by transferring knowledge from a larger *teacher* model onto a smaller *student* model (Bucilua et al., 2006). However, knowledge distillation has since been applied to several different tasks such as machine translation (Weng et al., 2020), automated machine learning (Kang et al., 2020), cross-modal learning (Hu et al., 2020), and cross-lingual named entity recognition (Wu et al., 2020a,b; Liang et al., 2021; Chen et al., 2021).

To the best of our knowledge, our approach is the first effort into leveraging a knowledge-distillation framework for CLED. The following sections present the details of our teacher and student models as well as our hierarchical data-sample selection strategy for student-model training.

2.2.1 Teacher Model

Our teacher model architecture follows that of previous direct-transfer-based models for CLED (M'hamdi et al., 2019; Majewska et al., 2021; Guzman-Nateras et al., 2022). We leverage a transformer-based pre-trained multilingual language model as the encoder $\mathcal{E}_T$. In particular, we make use of XLM-R (Conneau et al., 2019) as it often outperforms multilingual BERT (Devlin et al., 2019) on the CLED task (Pouran Ben Veyseh et al., 2022). For the classifier $\mathcal{C}_T$, we employ a simple Feed-Forward Neural Network (FFNN) with 2 hidden layers (Eq. 3). A softmax operation is applied to the resulting predictions to obtain a probability distribution over the event types.

$$\mathcal{C}_T(y_{ij}) = \text{softmax}(W^{C_T 2} \text{ ReLU}(W^{C_T 1} h_{ij})) \quad (3)$$

where $W^{C_T 1}$ and $W^{C_T 2}$ are parameter matrices to be learned and $C_T(y_{ij}) \in \mathbb{R}^{|\mathbb{C}|}$ is the probability distribution over the event type set $\mathbb{C}$ for token $t_{ij} \in \mathcal{S}_{src}$.

Some related works use a Conditional Random Field (CRF) layer on top of the FFNN classifier in an attempt to capture the interactions between the label sequences (M'hamdi et al., 2019). However, we did not find substantial performance differences when using a CRF layer and choose not to include it to keep our model as simple as possible.

2.2.2 Teacher Adversarial Training

Pre-trained multilingual language models such as mBERT or XLM-R provide contextualized representations for word sequences in multiple languages by embedding the words into a shared multilingual latent space. However, several studies have shown that, in such multilingual latent space, words from the same language group together, creating language clusters (Nguyen et al., 2021; Yarmohammadi et al., 2021). As such, the word representa-

tions generated by these encoders are not language invariant. For a cross-lingual model, however, it is beneficial for similar words in the source and target languages to have similar (i.e. close) representations in the latent space. For instance, an English-trained Spanish-tested cross-lingual model would benefit if the representations for the words *dog* and *perro* were similar to each other as then the model could adequately handle the Spanish sample provided it learns how to handle its English counterpart during training.

A technique that has been frequently used to promote the generation of such language-invariant representations is Adversarial Language Adaptation (ALA) (Joty et al., 2017; Chen et al., 2018). ALA introduces a *language discriminator* network $\mathcal{D}$ whose objective is to differentiate between the source and target languages. It learns language-dependent features that allow it to classify word representations as belonging to either the source or target languages. Concurrently, the encoder network is trained in an adversarial manner: it attempts to fool the discriminator by generating language-independent representations that are difficult to classify. A key feature of ALA is that it only requires unlabeled target-language data and, as such, it can be applied in a zero-shot setting using the available $\mathcal{S}_{tgt}^{unl}$ sentence set.

Other works that have leveraged ALA perform adversarial training at the sequence level (Guzman-Nateras et al., 2022). That is, they only present the discriminator with sequence-level representations (e.g., the representation for the `[CLS]` token in mBERT). However, in this work we leverage token-level adversarial training which has been found to be more effective at generating language-invariant representations (Chen et al., 2021)

We again use a two-layer FFNN for the discriminator network $\mathcal{D}$. Instead of a softmax operation to generate a probability distribution, we employ a sigmoid function σ to predict the associated language l (Eq. 4).

$$\mathcal{D}(l_i) = \sigma(W^{D2}\,\text{ReLU}(W^{D1}h_{ij})) \quad (4)$$

where W^{D1} and W^{D2} are parameter matrices to be learned and $\mathcal{D}(l_{ij})$ is a scalar $\in [0, 1]$ that indicates how likely it is that the current token representation h_{ij} belongs to the source ($l_i = 0$) or target ($l_i = 1$) languages.

Thus, besides the ED classification loss $\mathcal{L}_\mathcal{C}$ described in Equation 2, adversarial training introduces the discriminator loss $\mathcal{L}_\mathcal{D}$ (Eq. 5) as an additional training signal.

$$\mathcal{L}_\mathcal{D} = \quad (5)$$
$$\frac{1}{n*m}\sum_{i=1}^{n}\sum_{j=1}^{m} l_i \cdot \mathcal{D}(h_{ij}) + (1-l_i)(1-\mathcal{D}(h_{ij}))$$

Our adversarial training is achieved by minimizing the following term:

$$\arg\min_{\mathcal{E},\mathcal{C}} \sum_{i=1}^{n}\sum_{j=1}^{m} (\mathcal{L}_\mathcal{C}(y_{ij}|h_{ij}) - \lambda\mathcal{L}_\mathcal{D}(l_i|h_{ij})) \quad (6)$$

We leverage a Gradient-Reversal Layer (GRL) (Ganin and Lempitsky, 2015) to implement Equation 6 by applying the GRL to the discriminator input vectors h_{ij}. A GRL acts as the identity function during the forward pass and reverses the direction of the gradients during the backward pass. As such, the encoder parameters are trained in the opposite direction to those of the discriminator, effectively learning to generate token representations with language-invariant features.

Figure 1 shows the architecture of the Teacher model.

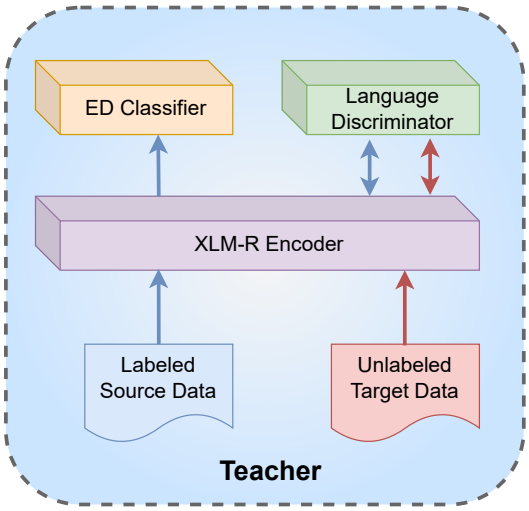

Figure 1: Adversarially-trained Teacher model. Source and target (unlabeled) data is passed through the encoder and fed at a token-level to the language discriminator. The discriminator gradients are then used to update the encoder parameters in an adversarial manner. The ED classifier is trained with the labeled source samples exclusively.

2.2.3 Student Model

As described in the previous section, the teacher model is trained using a direct transfer approach: it learns to generate language-independent representations from the labeled source-language data so that it can be directly applied to unlabeled target-language data. However, in our proposed hybrid knowledge transfer approach, we expect the student model to reap the benefits of the data transfer paradigm. Hence, we train the student model using target-language data so that it may learn from syntactical features and word/label relations.

First, we apply the teacher model *Teach* to the unlabeled target dataset $\mathcal{S}_{tgt}^{unl}$ to obtain a pseudo-labeled training set S_{tgt}^{Teach}. Afterward, the student model $Student$ is trained in a supervised manner using the obtained pseudo-labels.

The model architecture of our student model mirrors the one of the teacher model: a pre-trained multilingual language model as the encoder $\mathcal{E}_{STU}$ and a two-layer FFNN for a classifier $\mathcal{C}_{STU}$.

$$\mathcal{C}_{STU}(y_{ij}) = \text{softmax}(W^{Cs2}\ \text{ReLU}(W^{Cs1}h_{ij})) \tag{7}$$

Previous works on knowledge distillation have found that using soft labels (i.e., probability distributions over class types) is beneficial for student learning as they contain richer and more helpful information than hard labels (Hinton et al., 2015). As such, we train the student model to minimize the Mean Squared Error (MSE) between the student-predicted and teacher-generated event-type distributions (Eq. 8).

$$\mathcal{L}_{Student} = \tag{8}$$
$$\frac{1}{n*m}\sum_{i=1}^{n}\sum_{j=1}^{m}(\mathcal{C}_{STU}(\mathcal{E}_{STU}(t_{ij})) - \mathcal{C}_{T}(\mathcal{E}_{T}(t_{ij})))^2$$

2.3 Student-Training Sample Selection

An important challenge in our teacher-student framework is that the target pseudo-labels obtained from the teacher model are prone to contain noisy predictions. The teacher model is trained with a direct transfer approach and, even though its word representations are encouraged to be language-independent through adversarial training, it learns task-related information exclusively from the source-language labels. We argue this prevents the teacher from learning task-specific information in the target language as it is unable to exploit the word-label relations specific to such language. Furthermore, even though the student model should be able to benefit from being trained in the target language, any potential benefits can be nullified if the quality of the teacher-generated pseudo-labels is too poor.

To address the aforementioned issue, we argue that the teacher model should produce more reliable pseudo-labels on target-language examples that share some similarities (structural or otherwise) with the source-language examples. Hence, we suggest improving the knowledge-distillation process by restricting student-model training to target-language examples with such desirable characteristics. We implement this idea by designing a two-step hierarchical sample-selection scheme: First, we leverage Optimal Transport (OT) (Villani, 2008) to generate an alignment score between source and target samples and select samples above a defined alignment threshold. Then, using the selected source and target samples, we compute their pairwise Cross-domain Similarity Scaling scores (CSLS, Conneau et al., 2018) and only keep the pairs with the highest similarities. The following subsections describe each step in further detail.

Figure 2 presents an overview of our teacher-student framework.

2.3.1 Optimal-Transport-based Selection

Recent research efforts have successfully leveraged OT for cross-lingual language adaptation (Phung et al., 2021; Guzman-Nateras et al., 2022) and word-label alignment for event detection (Pouran Ben Veyseh and Nguyen, 2022). OT relies on a distance-based cost function to compute the most cost-effective transformation between two discrete probability distributions by solving the following optimization problem:

$$\pi^*(x,z) = \min_{\pi \in \prod(x,z)} \sum_{x \in \mathcal{X}} \sum_{z \in \mathcal{Z}} \pi(x,z)\ D(x,z) \tag{9}$$
$$\textbf{s.t.}\ x \sim P(x) \text{ and } z \sim P(z)$$

In Eq. 9, D is a cost function that maps $\mathcal{X}$ to $\mathcal{Z}$, $D(x,z), \mathcal{X} \times \mathcal{Z} \longrightarrow \mathbb{R}_+$, $P(x)$ and $P(z)$ are probability distributions for the $\mathcal{X}$ and $\mathcal{Z}$ domains, and $\pi^*(x,z)$ is the optimal joint distribution over the set of all joint distributions $\prod(x,z)$ (i.e., the optimal transformation between $\mathcal{X}$ and $\mathcal{Z}$).

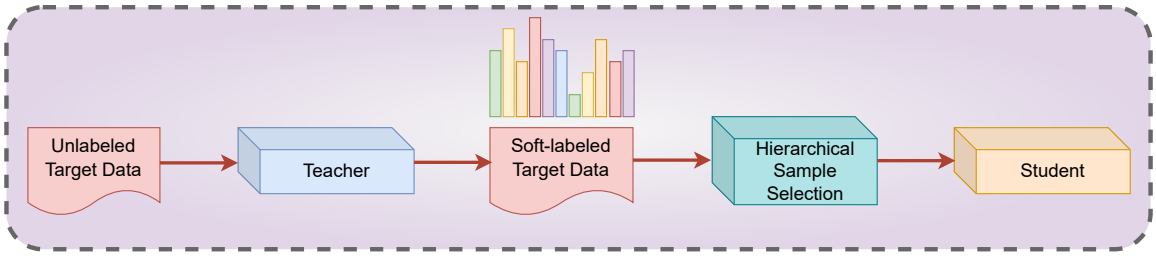

Figure 2: Teacher-student framework. The adversarially trained Teacher is used to annotate unlabeled target samples. Our hierarchical sample selection process picks a subset of samples to be used to train the Student model.

For our work, we consider the source and target languages as the $\mathcal{X}$ to $\mathcal{Z}$ domains to be aligned. Each training sample corresponds to a data point in a distribution and is represented by its sentence-level encoding h_0. Following prior work (Pouran Ben Veyseh and Nguyen, 2022), we estimate probability distributions $P(x)$ and $P(z)$ using a single-layer FFNN and use Euclidean distance as the cost function:

$$D(h^x_{i0}, h^z_{j0}) = ||h^x_{i0} - h^z_{j0}||^2_2 \quad (10)$$

where h^x_{i0} is the i-th source-language sample and h^z_{j0} is the j-th target-language sample.

Once the OT algorithm converges, we leverage the solution matrix π^* to compute an overall similarity score $k\_$ for each sample h_0 by averaging the optimal cost of transforming it to the other domain:

$$k^x_i = \frac{\sum_j^m \pi^*(h^x_{i0}, h^z_{j0})}{m} \quad (11)$$

Finally, a hyperparameter α determines the proportion of samples with the highest similarity scores k to be selected for use in the next step.

2.3.2 CSLS-based Selection

The OT-based similarity score described previously captures the *global* alignment of a sample with the alternate language, e.g., how well a source-language sample aligns with the target language and vice versa. Nonetheless, we propose to further refine our sample selection by considering the *pairwise* similarity between source and target samples.

To this end, we make use of the CSLS similarity measure which was originally designed to improve word-matching accuracy in word-to-word translation (Wu et al., 2020b). CSLS addresses a fundamental issue of pair-matching methods based on Nearest Neighbors (NN): NNs are asymmetric by nature, i.e. if a is a NN of b, b is not necessarily a NN of a. In high-dimensional spaces, this asymmetry leads to *hubness*, a detrimental phenomenon for pair matching: samples in dense areas have high probabilities of being NN to many others, while samples that are isolated will not be a NN to any other sample (Conneau et al., 2018).

As such, when computing the similarity between a pair of samples, CSLS (Eq. 12) computes mean similarity $r\_$ of a sample to its neighborhood $\mathcal{N}\_$ (i.e., its K nearest neighbors) in the alternate language and leverages it to increase the similarity scores of isolated samples while decreasing the scores of so-called *hub* samples. For example, the mean similarity r_Z for source sample h^x_i is computed with its target neighborhood $\mathcal{N}_Z$ (Eq. 13).

$$\text{CSLS}(h^x_i, h^z_j) = \quad (12)$$
$$2\cos(h^x_i, h^z_j) - r_Z(h^x_i) - r_X(h^z_j)$$

$$r_Z(h^x_i) = \frac{1}{|\mathcal{N}_Z|} \sum_{\mathcal{N}_Z} \cos(h^x_i, h^z_j) \quad (13)$$

$$r_X(h^z_j) = \frac{1}{|\mathcal{N}_X|} \sum_{\mathcal{N}_X} \cos(h^z_j, h^x_i) \quad (14)$$

where *cos* is the cosine similarity. In our work, the source $\mathcal{N}_X$ and target $\mathcal{N}_Z$ neighborhoods are defined as the corresponding sample sets kept by the previous selection step. Again, we keep a proportion of the samples with the best pairwise similarity scores determined by a hyperparameter β.

Figure 3 presents an overview of our proposed hierarchical sample-selection strategy.

3 Experiments

3.1 Datasets and Hyperparameters

For our experiments, we leverage the ACE05 (Walker et al., 2006) and ACE05-ERE (Song et al., 2015) datasets as they are the most commonly used datasets for CLED. ACE05 contains ED annotations in 3 languages:

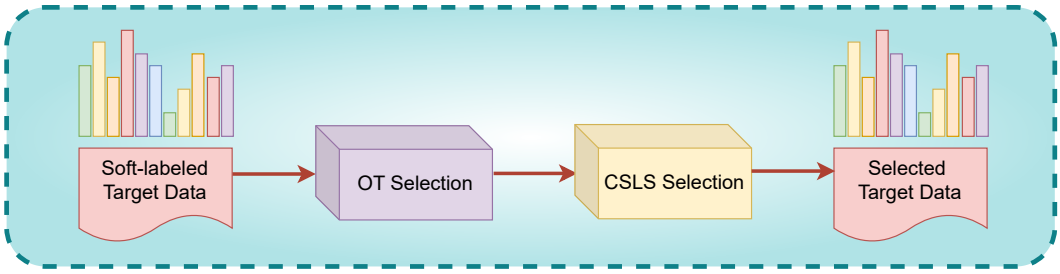

Figure 3: Hierarchical sample selection scheme. The target-language samples annotated by the Teacher model are first filtered by OT-based selection. The remaining samples are then further refined via CSLS. The final subset of samples is used to train the Student model.

English (En), Chinese (Zh), and Arabic (Ar) while ACE05-ERE annotates data for English and Spanish (Es). In addition, we evaluate our model on the recently released MINION dataset (Pouran Ben Veyseh et al., 2022), which contains annotations for 8 morphologically and syntactically distinct languages: English, Spanish, Hindi (Hi), Japanese (Ja), Korean (Ko), Polish (Pl), Portuguese (Pt), and Turkish (Tr). For a fair comparison, we follow the same train/val/test splits as prior work (M'hamdi et al., 2019; Pouran Ben Veyseh et al., 2022).

We tune all hyperparameters on the validation sets. In particular, we use AdamW (Loshchilov and Hutter, 2017) as the optimizer. We approximate the solution to the intractable problem described by Equation 9 by solving its entropy-based relaxation via the Sinkhorn iterative algorithm (Cuturi, 2013). Following prior works (Wu et al., 2020b), we freeze the embeddings and first three layers of the XLM-R encoder for student training. Learning rates for the transformer and non-transformer parameters are set at $2e^{-5}$ and $1e^{-4}$ respectively. The α and β hyperparameters are set at 0.5 and 0.75 respectively. We employ a batch size of 32 for the experiments on ACE05 and a batch size of 16 for the experiments on MINION. The size of the hidden feed-forward layers is 300. We use a learning rate linear scheduler with 5 warm-up epochs for teacher models and 10 warm-up epochs for student models. We use a parameter weight decay of 0.5 for transformer parameters and $1e^{-4}$ for non-transformer parameters. Finally, we train the teacher model for 20 epochs and the student model for 100 epochs.

3.2 Main results

In order to evaluate our Hybrid Knowledge Transfer for Cross-Lingual Event Detection (HKT-CLED) model, we first present our results on the ACE05 and ACE05-ERE datasets in Table 1. We compare against 6 recent CLED efforts including the current state-of-the-art model (Guzman-Nateras et al., 2022). All the baseline results are taken directly from the original papers and our model's results are the average of 5 runs with different seeds. English is used as the sole source language and Arabic, Chinese, and Spanish are employed as target languages. Following previous works, we report F1 scores.

Model	Target Language		
	Zh	Ar	Es
Liu et al. (2019)	27.0	-	-
M'hamdi et al. (2019)	68.5	30.9	-
Lu et al. (2020)	-	-	41.77
Majewska et al. (2021)	46.9	29.3	-
Nguyen et al. (2021)	72.1	42.7	-
Guzman-Nateras et al. (2022)	74.64	44.86	47.69
HKT-CLED (Ours)	**75.22**	**46.37**	**48.58**

Table 1: Cross-lingual event detection model performance comparison. English is used as the source language. ACE05 is used for Chinese (Zh) and Arabic (Ar), ACE05-ERE is used for Spanish (Es).

Our proposed approach obtains new state-of-the-art performance across all 3 target languages with improvements of $+0.58$, $+1.51$, and $+0.89$ F1 points for Chinese, Arabic, and Spanish, respectively. We believe these results demonstrate the importance of hybrid knowledge transfer as it gives HKT-CLED an edge over previous works that follow a direct transfer approach (M'hamdi et al., 2019; Majewska et al., 2021; Nguyen et al., 2021; Guzman-Nateras et al., 2022).

To validate the effectiveness and general applicability of our approach, Table 2 presents the performance of our HKT-CLED model on the more diverse MINION dataset. Once again, we employ

Model	Target Language						
	Es	Hi	Ja	Ko	Pl	Pt	Tr
Baseline*	62.83	58.19	35.12	56.78	60.13	72.77	47.21
HKT-CLED	66.03	68.63	61.84	58.24	61.35	77.28	53.85
Improvement	+3.2	+10.44	+26.72	+1.46	+1.22	+4.51	+6.64

Table 2: Cross-lingual ED performance on the MINION dataset. F1 scores are reported. English is used as the source language. Baseline* performance was obtained directly from the original MINION paper (Pouran Ben Veyseh et al., 2022). HKT-CLED results are the average of 3 runs.

English as the source language and test our model's performance on the remaining 7 languages. For a fair comparison, we use their best XLM-R results. Our model consistently outperforms their reported baseline with an average performance improvement of +7.74 F1 points for all target languages (+5.25 if the highest and lowest improvements are not considered). In the case of Japanese, HKT-CLED obtains a massive performance improvement of over 25 F1 points. Also of note is that HKT-CLED performance is a lot more uniform across target languages than the baseline. There is a difference of 23.43 F1 points between the best-performing (Pt, 77.28) and the worst-performing (Tr, 53.85) target languages, as opposed to a 37.65 point difference in the baseline case (Pt, 72.77 and Ja, 35.12).

3.3 Analysis

3.3.1 Ablation Study

We first explore the contribution of each model component by performing an ablation study (Table 3). In particular, we evaluate the impact of three aspects: teacher adversarial training, OT-based sample selection, and CSLS-based sample selection. The *Teacher (Vanilla)* results were obtained with a standard sequence-labeling model without any adversarial training. Its performance leaves room for improvement as its word representations do not display any language-invariant qualities. A considerable improvement is achieved when training the teacher model with token-level adversarial training (*Teacher + Adv*). Then, the *Student (Vanilla)* row shows the result of training a student network on the teacher-generated pseudo-labels without any sample selection. We argue its performance is worse than the adversarially-trained teacher due to the noisy pseudo-labels. By incorporating OT-based selection, *Student + OT* is able to outperform its teacher. However, it is only by performing our hierarchical sample selection that the student model achieves new state-of-the-art performance.

Model	Target Language		
	Zh	Ar	Es
HKT-CLED	75.22	46.37	48.58
Student + OT	74.37	45.53	47.63
Student (Vanilla)	73.48	44.10	46.81
Teacher + Adv	73.85	44.42	47.37
Teacher (Vanilla)	70.51	43.59	46.75

Table 3: Ablation experiment results.

3.3.2 Impact of Sample-Selection Ratios

Figure 4 shows the impact of hyperparameter α on model performance. α determines the proportion of student-training samples kept by the OT-based selection step. An $\alpha = 1$ value performs no sample selection and $\alpha = 0.25$ only keeps a fourth of the batch samples with the highest similarity scores.

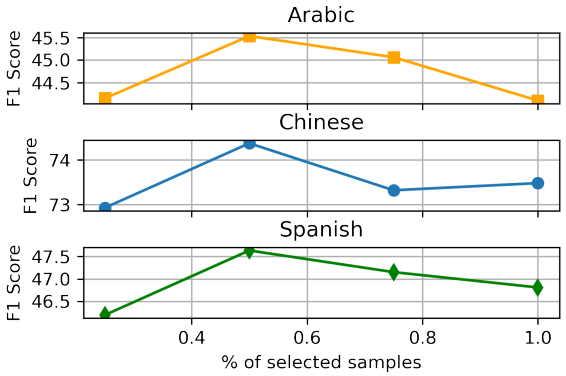

Figure 4: Performance impact of hyperparameter α.

Best results are obtained when half of the samples are kept ($\alpha = 0.5$) exemplifying the importance of removing training examples with potentially noisy pseudo-labels. However, if too few samples are chosen (e.g., $\alpha = 0.25$) the student performance drops below its *vanilla* version ($\alpha = 1$).

Similarly, Figure 5 presents the effect on performance of hyperparameter β which defines the

proportion of samples kept by the CSLS-selection step. A $\beta = 1$ value uses all of the samples selected by the previous step.

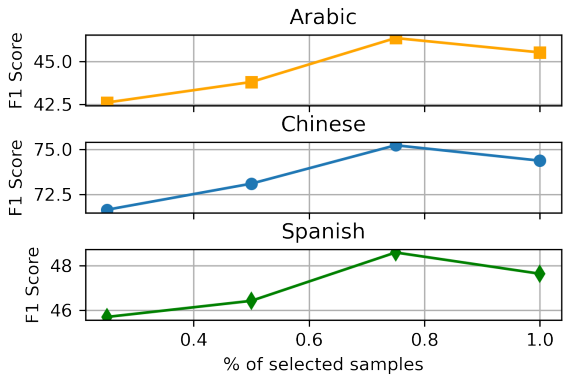

Figure 5: Performance impact of hyperparameter β.

Removing about a quarter ($\beta = 0.75$) of the previously-selected samples improves performance across all languages. Of note is the fact that the OT and CSLS similarity scores complement each other. From Figure 4 it would seem that removing more than half of the training samples would only hurt performance. However, given CSLS pairwise focus, it is able to effectively remove some remaining noisy samples and obtain better results.

4 Related Work

Event detection (ED) is an active research area in NLP (Nguyen and Grishman, 2015, 2018; Pouran Ben Veyseh et al., 2021), featuring cross-lingual ED as a recent direction with growing interests. The work by Liu et al. (2019) presents a data transfer method that learns a mapping between monolingual word embeddings, translates the source training data on a word-by-word basis and uses a graph convolutional network to generate order-independent representations. M'hamdi et al. (2019) leverage mBERT as an encoder to perform zero-shot transfer learning and a CRF layer to account for label dependency. Lu et al. (2020) present a cross-lingual structure transfer approach that represents sentences as language-universal structures (trees, graphs). In their work, Majewska et al. (2021) argue that event triggers are usually related to the verb in a sentence and propose to incorporate external verb knowledge by pre-training their encoder to classify whether two verbs belong to the same class according to two distinct ontologies VerbNet, (Kipper et al., 2006) and FrameNet, (Baker et al., 1998). Model *priming* (Fincke et al., 2021) is a simple, yet effective method that consists in augmenting the encoder inputs by concatenating a candidate trigger to the input sentence so that the encoder learns to generate task-specific representations. Nguyen et al. (2021) leverage class information and word categories as language-independent sources of information and condition their encoder to generate representations that are consistent in both the source and target languages. Finally, Guzman-Nateras et al. (2022) propose to optimize standard adversarial language adaptation by restricting the language discriminator training to *informative* examples.

Our approach is also closely related to knowledge distillation models for cross-lingual Named Entity Recognition (NER). Wu et al. (2020a) were the first to train a NER student model on the label distributions obtained from a teacher model. Wu et al. (2020b) improved upon this initial approach with a multi-step training method that involved fine-tuning the teacher model with pseudo-labeled data and generating hard labels that were later used for student training. More recent proposals improve the knowledge distillation with either reinforcement learning (Liang et al., 2021) or adversarial training (Chen et al., 2021). Nonetheless, our approach is the first to leverage a knowledge distillation framework for CLED, and our novel hierarchical training-sample selection scheme further differentiates our work from previous efforts.

5 Conclusion

In this work, we present the first effort to leverage a hybrid knowledge-transfer approach for the cross-lingual event detection task. We propose a teacher-student framework complemented by a hierarchical training-sample selection scheme that effectively constrains the student-training process to pseudo-labeled target-language samples that are similar to their source-language counterparts. Our HKT-CLED model sets a new state-of-the-art performance on the most popular benchmarking datasets ACE05 and ACE05-ERE, and obtains substantial performance improvements on the recently-released, and more diverse, MINION dataset with an average improvement of $+7.74$ F1 points across 7 distinct target languages. We believe these results demonstrate our model's robustness and applicability and validate our claim that combining the benefits of the direct transfer and data transfer approaches is beneficial for cross-lingual learning.

Limitations

We strived to make this work as accessible and applicable as possible. However, as with any other research effort, it suffers from several limitations stemming from preconceived assumptions. We believe that the most important limitation of our work is the assumption of the existence of a pre-trained multilingual language model, to be used as an encoder, that supports both the desired source and target languages. Though most modern multilingual language models support over a hundred languages, with over 7000 spoken languages in the world, the vast majority of languages remain unsupported. That being said, language models are trained in an unsupervised manner, meaning that only unlabeled data is required for training purposes. As such, a suitable encoder could be trained provided there is access to enough unlabeled data. This leads to what we consider to be the second biggest limitation of our work: the assumption of the availability of unlabeled target-language data. In general, raw unlabeled data is easy to obtain for most languages. However, it can represent a challenge for extremely low-resource languages. In these special cases, training an effective encoder can be an impossibility which, in turn, limits the applicability of our approach. Other limitations stem from our constrained time and computational resources. Our method requires a GPU with a large-enough memory to fit the transformer-based encoder which is usually more than what a personal computer GPU provides. Depending on the dataset and selected batch size, our model requires between 15 and 32 GB of GPU memory. We performed all our experiments on a Tesla V100 GPU with 32GB. Finally, additional experiments on a more diverse set of source/target language pairs could certainly provide a more comprehensive overview of our method's strengths and weaknesses.

Acknowledgement

This research has been supported by the Army Research Office (ARO) grant W911NF-21-1-0112, the NSF grant CNS-1747798 to the IUCRC Center for Big Learning, and the NSF grant # 2239570. This research is also supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via the HIATUS Program contract 2022-22072200003. 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 ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

References

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *COLING 1998 Volume 1: The 17th International Conference on Computational Linguistics*.

Cristian Bucilua, Rich Caruana, and Alexandru Niculescu-Mizil. 2006. Model compression. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Weile Chen, Huiqiang Jiang, Qianhui Wu, Börje Karlsson, and Yi Guan. 2021. AdvPicker: Effectively Leveraging Unlabeled Data via Adversarial Discriminator for Cross-Lingual NER. 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 743–753, Online. Association for Computational Linguistics.

Xilun Chen, Yu Sun, Ben Athiwaratkun, Claire Cardie, and Kilian Weinberger. 2018. Adversarial Deep Averaging Networks for Cross-Lingual Sentiment Classification. In *Transactions of the Association for Computational Linguistics*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. In *CoRR*.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *CoRR*.

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

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

Steven Fincke, Shantanu Agarwal, Scott Miller, and Elizabeth Boschee. 2021. Language model priming for cross-lingual event extraction. In *Association for the Advancement of Artificial Intelligence (AAAI)*.

Yaroslav Ganin and Victor Lempitsky. 2015. Unsupervised domain adaptation by backpropagation. In *Proceedings of the 32nd International Conference on Machine Learning*.

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.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. In *CoRR*.

Hengtong Hu, Lingxi Xie, Richang Hong, and Qi Tian. 2020. Creating something from nothing: Unsupervised knowledge distillation for cross-modal hashing. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Shafiq Joty, Preslav Nakov, Lluís Màrquez, and Israa Jaradat. 2017. Cross-language learning with adversarial neural networks. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL)*, pages 226–237.

Minsoo Kang, Jonghwan Mun, and Bohyung Han. 2020. Towards oracle knowledge distillation with neural architecture search. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Karin Kipper, Anna Korhonen, Neville Ryant, and Martha Palmer. 2006. Extending VerbNet with novel verb classes. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

Shining Liang, Ming Gong, Jian Pei, Linjun Shou, Wanli Zuo, Xianglin Zuo, and Daxin Jiang. 2021. Reinforced iterative knowledge distillation for cross-lingual named entity recognition. In *CoRR*.

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. In *CoRR*.

Di Lu, Ananya Subburathinam, Heng Ji, Jonathan May, Shih-Fu Chang, Avi Sil, and Clare Voss. 2020. Cross-lingual structure transfer for zero-resource event extraction. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1976–1981, Marseille, France. European Language Resources Association.

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 (Volume 1: Long Papers)*, pages 6952–6969, Online. Association for Computational Linguistics.

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.

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

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

Thien Huu Nguyen and Ralph Grishman. 2018. Graph convolutional networks with argument-aware pooling for event detection. In *Association for the Advancement of Artificial Inteligence (AAAI)*.

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.

Amir Pouran Ben Veyseh, Minh Van Nguyen, Franck Dernoncourt, and Thien Nguyen. 2022. MINION: a large-scale and diverse dataset for multilingual event detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2286–2299, Seattle, United States. Association for Computational Linguistics.

Amir Pouran Ben Veyseh, Minh Van Nguyen, Nghia Ngo Trung, Bonan Min, and Thien Huu Nguyen. 2021. Modeling document-level context for event detection via important context selection. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5403–5413, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Amir Pouran Ben Veyseh and Thien Nguyen. 2022. Word-label alignment for event detection: A new perspective via optimal transport. In *Proceedings of the 11th Joint Conference on Lexical and Computational Semantics*, pages 132–138, Seattle, Washington. Association for Computational Linguistics.

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

Chen-Tse Tsai, Stephen Mayhew, and Dan Roth. 2016. Cross-lingual named entity recognition via wikification. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 219–228, Berlin, Germany. Association for Computational Linguistics.

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*.

Rongxiang Weng, Heng Yu, Shujian Huang, Shanbo Cheng, and Weihua Luo. 2020. Acquiring knowledge from pre-trained model to neural machine translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34.

Qianhui Wu, Zijia Lin, Börje Karlsson, Jian-Guang Lou, and Biqing Huang. 2020a. Single-/multi-source cross-lingual NER via teacher-student learning on unlabeled data in target language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6505–6514, Online. Association for Computational Linguistics.

Qianhui Wu, Zijia Lin, Börje F. Karlsson, Biqing Huang, and Jian-Guang Lou. 2020b. Unitrans : Unifying model transfer and data transfer for cross-lingual named entity recognition with unlabeled data. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*.

Jiateng Xie, Zhilin Yang, Graham Neubig, Noah A. Smith, and Jaime Carbonell. 2018. Neural cross-lingual named entity recognition with minimal resources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 369–379, Brussels, Belgium. Association for Computational Linguistics.

Mahsa Yarmohammadi, Shijie Wu, Marc Marone, Haoran Xu, Seth Ebner, Guanghui Qin, Yunmo Chen, Jialiang Guo, Craig Harman, Kenton W. Murray, Aaron Steven White, Mark Dredze, and Benjamin Van Durme. 2021. Everything is all it takes: A multipronged strategy for zero-shot cross-lingual information extraction. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Section 6 Limitations

☐ A2. Did you discuss any potential risks of your work?
Not applicable. We currently do not identify any potential risks inherently associated with our work.

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

☒ A4. Have you used AI writing assistants when working on this paper?
We did not use any AI writing assistants.

B ☑ Did you use or create scientific artifacts?

Section 2 describes our model which we created using the Pytorch library. The appendix B contains our hyperparameter values and discusses additional implementation details.

☑ B1. Did you cite the creators of artifacts you used?
Not limited to a specific section. We cite all the original papers from artifacts such as the multilingual pre-trained language models we use as econders.

☒ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
We did not explicitly discuss the license or terms of use of the artifacts as they are publically available on the original sites.

☒ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
We did not explicitly discuss in or work on its intended use. We intend to release our work publicly under Apache License 2.0.

☒ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Our datasets are widely used in previous research for Multilingual Event Detection. We do not observe concerns for private information or offensive content in our datasets in previous work.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Section 2 Model, Section 3 Experiments, Section 4 Analysis, Appendix A and Appendix B

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

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

C ☑ Did you run computational experiments?

Section 3 Experiments

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

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

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Section 3 Experiments, Section 4 Analysis

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Section 3 Experiments, Appendix B

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

We did not use any human annotators or research with human objects.

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

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

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

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

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