Retrieving Relevant Context to Align Representations for Cross-lingual Event Detection

Chien Van Nguyen[1], **Linh Van Ngo**[2], **and Thien Huu Nguyen**[3]
[1] VinAI Research, Vietnam
[2] Hanoi University of Science and Technology, Hanoi, Vietnam
[3] Department of Computer Science, University of Oregon, Eugene, OR, USA
v.chiennv22@vinai.io, linhnv@soict.hust.edu.vn, thien@cs.uoregon.edu

Abstract

We study the problem of cross-lingual transfer learning for event detection (ED) where models trained on a source language are expected to perform well on data for a new target language. Among a few recent works for this problem, the main approaches involve representation matching (e.g., adversarial training) that aims to eliminate language-specific features from the representations to achieve the language-invariant representations. However, due to the mix of language-specific features with event-discriminative context, representation matching methods might also remove important features for event prediction, thus hindering the performance for ED. To address this issue, we introduce a novel approach for cross-lingual ED where representations are augmented with additional context (i.e., not eliminating) to bridge the gap between languages while enriching the contextual information to facilitate ED. At the core of our method involves a retrieval model that retrieves relevant sentences in the target language for an input sentence to compute augmentation representations. Experiments on three languages demonstrate the state-of-the-art performance of our model for cross-lingual ED.

1 Introduction

As one of the core tasks in Information Extraction (IE), the goal of Event Detection (ED) is to identify and classify the word(s) that most clearly evoke events in text (called event triggers). For instance, in the sentence "*He was* **fired** *from the corporation yesterday.*", an ED system needs to predict "*fired*" as an event trigger of the type *Attack*. Due to its applications, ED has been well studied over the last decade, featuring deep learning as the most recent approach with state-of-the-art performance (Nguyen and Grishman, 2015; Chen et al., 2015; Nguyen et al., 2016; Liu et al., 2017; Lu and Nguyen, 2018; Lin et al., 2020).

However, despite intensive studies, most prior work has focused on monolingual learning settings for ED where models are trained and evaluated on labeled data of the same languages (Nguyen and Grishman, 2018; Wadden et al., 2019; Lai et al., 2020; Yang et al., 2019; Ngo et al., 2020; Liu et al., 2020; Nguyen et al., 2021a). As such, to extend current ED models to another language, monolingual learning will require new annotated data to train the models, which can be very expensive to obtain for different languages. To this end, there has been a growing interest in cross-lingual transfer learning for ED where models trained on a source language are directly applied to data of a new target language (M'hamdi et al., 2019). In this way, labeled data from high-resource languages (e.g., English) can be leveraged to develop effective ED models for other languages (e.g., low-resource ones). In this work, we focus on zero-shot cross-lingual transfer learning for ED to avoid the need for labeled data in the target languages, thus enabling fast adaptation of ED models to multiple languages.

A key strategy for cross-lingual transfer learning is to align input text representations for the source and target languages to facilitate cross-lingual extraction of events. As such, prior work on cross-lingual ED has explored multilingual word embeddings (e.g., MUSE) (Joulin et al., 2018; Liu et al., 2019) or recent multilingual pre-trained language models (e.g., mBERT) (M'hamdi et al., 2019) to represent source- and target-language texts in the same space. Recently, state-of-the-art cross-lingual ED methods have leveraged unlabeled data in the target language with representation matching frameworks to further align text representations for the source and target languages (Nguyen et al., 2021b). Given two sentences in the source and target languages, these methods aim to encode the two sentences to obtain representation vectors for language-universal objects, e.g., sentences, event types, universal dependency relations or parts of

Findings of the Association for Computational Linguistics: ACL 2023, pages 2157–2170
July 9-14, 2023 ©2023 Association for Computational Linguistics

speech (Nguyen et al., 2021b). Afterward, the representations of the same language-universal objects (computed with the source or language data) are regulated to be similar to each other to improve the alignment between the source and target languages for cross-lingual ED (e.g., with adversarial training to fool the language discriminators).

As such, to achieve representation similarity between languages, previous representation matching methods for ED will need to filter information/features that are specific to each language in the representations (Nguyen et al., 2021b). However, as the representations for each language are computed from input sentences, the language-specific information might involve/mix with important contextual structures, which are necessary to reveal event types in the representations. Consequently, removing language-specific information might also eliminate important discriminative context features, thus limiting the performance for ED models. To address this issue, our work explores a novel approach for alignment of language representations for cross-lingual ED that avoids direct similarity regularization and removal of important context information from the representation vectors. Instead, our approach seeks to add relevant context information into original input representations for the source and target language texts to make them closer for effective cross-lingual ED. In particular, starting with the representation vectors S and T to perform ED in the source and target languages (respectively), we aim to induce additional context representations $A(S)$ and $A(T)$ that will be added into S and T (i.e., leading to the augmented representations $S + A(S)$ and $T + A(T)$) to achieve greater similarity between representations for the source and target languages. One the one hand, the additional representations $A(S)$ and $A(T)$ will be obtained over sentences in the target language to bias the prediction representations toward the target space and enhance representation alignment for cross-lingual ED. On the other hand, we will leverage external sentences with relevant/related contexts to the original representations S and T to compute the augmented representations $A(S)$ and $A(T)$. As such, with enriched context information, we expect that the augmented representations $S + A(S)$ and $T + A(T)$ will facilitate the prediction of event types to boost performance for cross-lingual ED. Note that this representation augmentation with relevant context is not possible in previous cross-lingual transfer learning methods for ED, thus highlighting the advantage of our proposed approach for cross-lingual ED in this work.

To implement the representation augmentation idea for cross-lingual ED, we introduce a "retrieve-then-classify" framework with two major steps. In the first step of retrieval, given an input sentence, our model first retrieves relevant/related sentences from an unlabeled dataset (e.g., focusing on sentences with similar event types). Next, the retrieved sentences will be encoded and their representations will be injected into the representation for the input sentence to perform ED. In our method, the unlabeled dataset will be taken from the target language (for input sentences from both the source and target languages) to shift the augmented representations to the target language space, thus implicitly bridging the gap between representations for different languages for ED. In addition, to better customize the context retrieval for our cross-lingual ED task, the retrieval model will be jointly trained with the ED model in an end-to-end fashion to encourage their interactions/feedback for better overall performance. Our framework also introduces a novel regularization mechanism to promote the shared awareness of relevant context sentences for an input sentence between retrieval and ED models to further improve the induced representations for cross-lingual ED. Finally, we conduct extensive experiments on the multilingual ACE 2005 dataset for ED (with three languages: English, Chinese, and Arabic), demonstrating the state-of-the-art performance of the proposed method over different language pairs for cross-lingual transfer learning of ED. To our knowledge, this is the first work on retrieval-based models for cross-lingual ED.

2 Model

We follow prior work (M'hamdi et al., 2019) to formalize the cross-lingual transfer learning (CLTL) task for ED as a sequence labeling problem. Given an input sentence $W = w_1, w_2, \ldots, w_n$ with n words, we need to assign a label y_i for each word $w_i \in W$ using the BIO annotation schema to capture event triggers and their types in W. In CLTL, the input sentence W belongs to the source language in the training time while sentences in the new target language are leveraged for evaluation in test time. Similar to recent methods on CLTL for ED (Nguyen et al., 2021b), our model assumes an unlabeled dataset $U = \{U_1, U_2, \ldots, U_m\}$ that contains m sentences in the target language (U_t is the t-th sentence in U).

Given the input sentence W, the first step in our model involves retrieving a set of relevant sentences R in the unlabeled dataset U (i.e., $R \subset U$) to provide augmented context information for W for cross-lingual ED. Note that the unlabeled set U will be used to retrieve sentences for input texts in both training and testing phases. The representations for the retrieved sentences in R will later be integrated into the representation vectors for the words in W to perform sequence labeling for ED. The benefit of this representation augmentation approach is twofold. First, as the representations for the retrieved sentences R are computed over the target language sentences, during the training time with the source-language input sentence W, the representation augmentation will shift the representations for the words $w_i \in W$ closer to the target language space. This helps to bridge the gap between the source- and target-language representation spaces that enables the training of ED models over source-language data to better generalize to data in the target language (i.e., cross-lingual generalization). Second, during the test time with the target language, incorporating context information from the retrieved relevant sentences R will enrich/strengthen the representations for the words in the original input sentence W, thus facilitating the predictions of labels to boost performance for cross-lingual ED. Our following sections will describe the retrieval and ED models in our method.

2.1 Relevant Context Retrieval

To retrieve relevant sentences for W in U for ED, our intuition is to identify sentences in U that express the same event types using similar context patterns as in W. We expect that such relevant sentences can strengthen the necessary context to predict event triggers in W, and improve the target-language orientation of the representations to boost cross-lingual performance. To this end, our retrieval model first aims to compute a representation vector for W and each sentence $U_t \in U$ to capture their event contexts. For W, we append the special token [CLS] to the beginning and send it into a multilingual pre-trained language model to learn representations for each token. In particular, we leverage miniLM (Wang et al., 2020), a multilingual language model distilled from the pre-trained model XLM-RoBERTa (large version) (Conneau et al., 2020), to obtain representation vectors for the words in W in our retrieval component. Compared to XLM-RoBERTa with 24 transformer layers and 1024 hidden dimensions, the multilingual miniLM version only includes 6 transformer layers with 384 hidden dimensions that can make our retrieval component more efficient for representation computation. As such, the representation vector for [CLS] in the last layer of miniLM will be used as the representation vector $\overline{W}$ for W. Similarly, we also compute the representation vector $\overline{U}_t$ for each sentence U_t in the unlabeled set U with miniLM. Here, we employ two separate versions of the pre-trained miniLM model to encode the input sentence W and the unlabeled sentences U in the target language (called miniLM$_W$ and miniLM$_U$ respectively), thus enabling the flexibility to capture context information for each type of data, i.e., $\overline{W} = \text{miniLM}_W(W)$ and $\overline{U}_t = \text{miniLM}_U(U_t)$.

Given the representation vectors $\overline{W}$ and $\overline{U}_t$, we compute a similarity score between W and each unlabeled sentence $U_t \in U$ using the cosine similarity: $sim(\overline{W}, \overline{U}_t) = \overline{W} \cdot \overline{U}_t / ||\overline{W}|| ||\overline{U}_t||$. Afterward, we select the top K sentences in U that have the highest similarities sim with W to serve as the retrieved set R of relevant sentences: $R = \{R_1, R_2, \ldots, R_K\}$ (i.e., R_k is the k-th sentence in $R \subset U$ and K is a hyper-parameter).

Warm-up Training: The computed representation vectors $\overline{W}$ and $\overline{U}_t$ so far are generic and not customized for our goals of same event types and similar context. To this end, we propose to fine-tune the language models miniLM$_W$ and miniLM$_U$ to adapt their encoding mechanisms to the retrieval problem for ED using contrastive learning (Khosla et al., 2020). Given a sentence W in the training dataset L of the source language, let T_W be the set of event types that are presented in W. We focus on the sentences W with at least one event in this contrastive learning process (e.g., $|T_W| > 0$). As such, to obtain a positive example, we identify another sentence $P \in L$ that involves at least one event type in T_W (i.e., containing the same event types). For negative examples, we leverage a set of sentences $N(W)$ in L that do not express any event type in T_W. In the implementation, we compute $N(W)$ for each sentence using the other sentences in the same minibatch. As such, our contrastive loss to fine-tune the miniLM models for event retrieval is formed via: $\mathcal{L}_{const} = -\log \frac{\exp(sim(\overline{W}, \overline{P}))}{\sum_{N \in N(W)} \exp(sim(\overline{W}, \overline{N}))}$ where

$\overline{W} = \text{miniLM}_W(W)$, $\overline{P} = \text{miniLM}_U(P)$, and $\overline{N} = \text{miniLM}_U(N)$. Note that this contrastive training process is only used as a warm-up step to prepare our retrieval model event types and context in our task; we will later jointly train the retrieval model with the ED model to leverage the training signals for ED to improve the retrieval model.

2.2 Event Detection Model

To solve the cross-lingual ED problem for the input sentence W, our model aims to perform sequence labeling over W conditioning on the retrieved relevant sentence $R \subset U$. For convenience, let $\overline{R}_k$ be the representation vector for $R_k \in R$ induced from miniLM$_U$. The similarity score between W and R_k is thus $sim(\overline{W}, \overline{R}_k)$. Also, let $R_k = r_{k,1}, r_{k,2}, \ldots, r_{k,I_k}$ be the sequence of words for R_k (i.e., I_k is the length of R_k and $r_{k,j}$ is the j-th word in R_k).

To this end, our ED model first feeds W (prepended with [CLS]) into the multilingual pretrained language model XLM-RoBERTa (base version) (Conneau et al., 2020), called XLMR, to obtain representations for the words $w_i \in W$. In particular, using the hidden vectors in the last transformer layer of XLMR, we leverage the average of the hidden vectors for the subtokens of $w_i \in W$ to compute the representation vector $\overline{h}_i$ for w_i, denoted by $\overline{h}_1, \overline{h}_2, \ldots, \overline{h}_n = \text{XLMR}(w_1, w_2, \ldots, w_n)$. In a typical sequence labeling model, the representation $\overline{h}_i$ can be sent into a feed-forward network to produce a distribution over possible BIO tags for w_i for ED. In our model, to augment the representation $\overline{h}_i$ for w_i with the retrieved sentence context in R for cross-lingual ED, we further seek to incorporate context representations for the words $r_{k,j}$ in the sentences $R_k \in R$ to improve $\overline{h}_i$ for cross-lingual prediction. As such, we also feed each sentence R_k into the multilingual model XLMR to generate the representation vectors $\overline{r}_{k,j}$ for the words $r_{k,j} \in R_k$, following the same procedure for $\overline{h}_i$: $\overline{r}_{k,1}, \overline{r}_{k,2}, \ldots, \overline{r}_{k,I_k} = \text{XLMR}(r_{k,1}, r_{k,2}, \ldots, r_{k,I_k})$.

In the next step, using the attention mechanism, we quantify the contribution of each representation vector $\overline{r}_{k,j}$ for the augmentation of $\overline{w}_i$ for W with the attention weight $a_{i,k,j}$. In particular, our motivation for $a_{i,k,j}$ is that the attention weight of $r_{k,j}$ for w_i needs to capture their context similarity within their corresponding sentences R_k and W. In addition, the attention weight $a_{i,k,j}$ should also condition on the retrieval similarity between the corresponding retrieved sentence R_k and the input sentence W (i.e., $sim(\overline{W}, \overline{R_k})$). The rationale is that the words in a retrieved sentence R_k with higher retrieval similarity score with W should be more preferable than the words in other sentences in R for the context augmentation of w_i (i.e., a retrieval bias). To this end, the attention weight $a_{i,k,j}$ of $\overline{r}_{k,j}$ for $\overline{w}_i$ is computed via: $a_{i,k,j} = \frac{b_{i,k,j}}{\sum_{k'=1}^{K} \sum_{j'=1}^{I_{k'}} b_{i,k',j'}}$ where $b_{i,k,j} = \exp(\overline{w}_i A \overline{r}_{k,j} + \alpha sim(\overline{W}, \overline{R_k}))$. Here, α is a trade-off parameter between context and retrieval similarities and A is the learnable matrix. Afterward, the augmentation context representation $\overline{a}_i$ from retrieved sentences R for $\overline{w}_i$ is obtained via the weighted sum: $\overline{a}_i = \sum_{k=1}^{K} \sum_{j=1}^{I_k} a_{i,k,j} \overline{r}_{k,j}$.

Finally, the representation vector for event prediction for w_i is computed by: $v_i = \overline{w}_i + \overline{a}_i$. v_i is then fed into a two-layer feed-forward network FF to compute a score vector to capture the possibilities for w_i to receive the possible BIO labels for ED: $p_i = FF(v_i)$. Next, the score vectors p_i are sent into a Conditional Random Field (CRF) layer to encode the tag dependencies and compute the conditional probability $P(\cdot|W, R)$ for the possible label sequences for W. The negative log-likelihood for the golden label sequence Y^* is then used to train the model: $\mathcal{L}_{seq} = -\log P(Y^*|W, R)$.

In the test time, given an input sentence W in the target language, we also compute the augmentation representations $\overline{a}_i$ for the words in W using the same unlabeled set U. Viterbi decoding with $P(\cdot|W, R)$ is then employed to predict the label sequence for W for ED. As such, the augmentation representations $\overline{a}_i$ are computed over the same unlabeled set U of the target language for both training and testing phases, thus shifting the prediction representations v_i toward the target language space to achieve better cross-lingual alignment for ED.

Joint Training: The inclusion of the retrieval similarity score $sim(\overline{W}, \overline{R_k}))$ (computed from miniLM$_W$ and miniLM$_U$) in the attention weight $a_{i,k,j}$ for the ED model implies that the training signals for ED in $\mathcal{L}_{seq}$ are also back-propagated to the retrieval model, thus better adapting the retrieval model to our problem of similar event context retrieval. However, this back-propagation also entails updating the miniLM$_W$ and miniLM$_U$ models in the retrieval component after each mini-batch in the training process. As such, the retrieval model will also need to recompute the representations $\overline{U}_t$ for each unlabeled sentence $U_t \in U$ after each training step, which can be very expensive and slow down the training. To this end, instead of updating the retrieval model after each training step, in the implementation, we only update miniLM$_W$ and miniLM$_U$ after every Q training steps/mini-batches (Q is a hyper-parameter). In this way, although we cannot leverage the latest updates for the retrieval component, our model can maintain the synchronization between miniLM$_W$ and miniLM$_U$, reduce training time significantly, and still retrieve relevant sentences from U for cross-lingual ED.

2.3 Similarity Regularization

In our model, the retrieved sentences $R_k \in R$ are expected to be relevant/similar to the input sentence W according to the retrieval model with miniML$_W$ and miniML$_U$. As such, to achieve a consistency between the retrieval model and the ED model, we argue that the retrieved sentences R_k should also be similar to W according to the ED model with the XLMR model for sentence encoding. Consequently, we propose to explicitly encourage the the similarities between the representations for R_k and W as computed by the XLMR model for ED, serving as a regularization to improve representation learning in our model. In particular, when W and $R_k \in R$ are encoded by XLMR for the ED model, we also use the hidden vectors for the [CLS] token in the last transformer layer of XLMR represent these sentences, leading to the representation vectors $\overline{W}^{XLMR}$ and $\overline{R_k}^{XLMR}$ for W and R_k respectively. Afterward, we enforce the XLMR-based similarity between W and R_k by minimizing the negative cosine similarity between $\overline{W}^{XLMR}$ and $\overline{R_k}^{XLMR}$: $\mathcal{L}_{reg} = -\sum_{k=1}^{K} sim(\overline{W}^{XLMR}, \overline{R_k}^{XLMR})$. The overall loss function to train our model is thus: $\mathcal{L} = \mathcal{L}_{seq} + \lambda \mathcal{L}_{reg}$ (λ is a trade-off parameter).

During the training time, as W and R_k belong to the source and target languages respectively, the minimization of $\mathcal{L}_{reg}$ also serves to align the representations for the source and target languages, thus similar to the representation matching frameworks in prior work for cross-lingual ED (Nguyen et al., 2021b). However, a key difference is that previous representation matching methods tend to match randomly chosen sentences in the source and target languages that might involve different event contexts. To align the source- and target-language representations, such previous methods might thus learn to exclude those event contexts from the representations, causing poorer discriminative features for ED. In contrast, our cross-lingual similarity regularization with $\mathcal{L}_{reg}$ is performed over the sentences W and R_k with similar event context (due to the retrieval component). As such, our model might be able to learn to only eliminate language-specific features that do not overlap with the common event context features. The event context information is thus preserved to best perform cross-lingual ED.

3 Experiments

Datasets and Hyper-parameters: We evaluate our cross-lingual retrieval-based model for ED (called CLRED) on the multilingual dataset ACE 2005 (Walker et al., 2006), following previous work (M'hamdi et al., 2019; Nguyen et al., 2021b). ACE 2005 provides event trigger annotations for 33 event types in documents of three languages: English (EN), Chinese (ZH) and Arabic (AR). To achieve a fair comparison, we use the exact data split and preprocessing provided by previous work (Nguyen et al., 2021b). The data split includes training, development, and test data for each of the three languages. To perform cross-lingual transfer learning evaluation, we will consider six possible pairs of languages chosen from English, Chinese, and Arabic. For each language pair, we train the models on the training data of one language (i.e., the source language) and evaluate the models on the test data of the other language (i.e., the target language). Similar to previous work (Nguyen et al., 2021b), the unlabeled dataset U in our experiments is obtained from the training data of the target language where the labels are completely removed.

To tune the hyper-parameters for our model, we use the performance over the development data of the source languages. In particular, the selected hyper-parameters from our tuning process involve: $1e$-5 for the learning rate with the AdamW optimizer, 16 for the mini-batch size, 300 dimensions for the hidden layers of the feed-forward network FF, $K = 2$ for the number of retrieved sentences in R, $Q = 30$ for the number of steps to update miniLM$_W$ and miniLM$_U$, $\alpha = 1$ for the trade-off parameter between context and retrieval similarities in the attention weights, and $\lambda = 0.1$ for the trade-off parameter in the overall loss function. Finally, we utilize the base version of XLM-RoBERTa (Conneau et al., 2020) with 768 dimensions for the hidden vectors for our ED model.

Baselines: We consider two groups of baselines for our cross-lingual model CLRED. The first group concerns previous methods that only leverage training data in the source language for learning (i.e., no unlabeled data in the target language). The state-of-the-art model in this group involves the BERT-CRF model in (M'hamdi et al., 2019) that applies a CRF layer on top of multilingual BERT (mBERT) (Devlin et al., 2019). To make it fair, we also report the performance of XLMR-CRF that replaces mBERT[1] in BERT-CRF with our XLM-RoBERTa model. Note that XLRM-CRF is equivalant to our CLRED model when the retrieval component and augmentation context are excluded.

The second group of baselines additionally uses the unlabeled dataset U in the target language to train cross-lingual models for ED. A state-of-the-art model in this group features the BERT-CRF-CCCAR model in (Nguyen et al., 2021b) that utilizes unlabeled data to match representations for universal word categories and event types computed from BERT-CRF. In the experiments, we also provide the performance of XLMR-CRF-CCCAR that is similar to BERT-CRF-CCCAR, but replaces BERT with XLM-RoBERTa. To make it compatible, we obtain the original code and implementation for BERT-CRF-CCCAR from (Nguyen et al., 2021b) to perform the replacement and evaluation.

In addition, we explore the language adversarial training (LAT) method to leverage unlabeled target data to induce language-universal representations for cross-lingual ED. In LAT, a base model for cross-lingual ED is also either BERT-CRF or XLMR-CRF. Further, a language discriminator is introduced to classify whether a representation vector is computed over a sentence in the source or target language (Chen et al., 2019; Huang et al., 2019; Keung et al., 2019). We follow the same implementation of LAT for cross-lingual ED in (Nguyen et al., 2021b) that jointly trains the language discriminator with the sequence labeling model for ED. The Gradient Reversal Layer (GRL) (Ganin and Lempitsky, 2015) is employed to fool the discriminator and eliminate language-specific features from the representations. To this end, we report the performance of LAT for both BERT-CRF and XLMR-CRF, leading to BERT-CRF-LAT and XLMR-CRF-LAT in our experiments.

Motivated by prior work on cross-lingual learning (Pfeiffer et al., 2020), we also evaluate the language model fine-tuning (LMFT) method where a multilingual pre-trained model is first fine-tuned on the unlabeled data U of the target language using mask language modeling (Devlin et al., 2019). The fine-tuned model is then directly employed as the encoder in the base sequence labeling model (e.g., XLMR-CRF) with CRF for cross-lingual ED. Considering both mBERT and XLM-RoBERTa, we also have two versions for this LMFT method, i.e., BERT-CRF-LMFT and XLMR-CRF-LMFT. Here, the *huggingface* library is utilized to fine-tune mBERT and XLM-RoBERTa on unlabeled target data for $100,000$ steps.

Finally, we report the performance of the recent model OACLED (Guzman et al., 2022) that has the best reported performance for cross-lingual ED so far. OACLED is also based on the idea of LAT; however, it introduces a a new component to leverage optimal transport and XLMR to perform data selection for the language discriminator.

Comparison: Table 1 presents the cross-lingual performance for six different language pairs.

The first observation is that the XLMR-based models are significantly better than their corresponding BERT-based models across most language pairs and models (e.g., *-CRF and *-CRF-CCCAR). This demonstrates the advantages of the multilingual language model XLM-RoBERTa over multilingual BERT for cross-lingual ED. Second, comparing the models with and without unlabeled target-language data, we find that the *-CRF-CCCAR and OACLED models substantially outperform the *-CRF models regardless of the multilingual pre-trained models over different language pairs. The *-CRF-LAT and *-CRF-LMFT models are also better than the *-CRF models in most situations (except for some language pairs). As such, it highlights the benefits of using unlabeled data in the target language to improve the language-universal representations and cross-lingual performance for ED if introduced appropriately. Most importantly, Table 1 shows that the proposed model CLRED achieves significantly better performance than all the baseline methods (with $p < 0.01$) across different language pairs. The state-of-the-art performance of CLRED thus clearly demonstrates the advantages of our new retrieval-based approach with representation augmentation for cross-lingual transfer learning for ED.

Ablation Study: Compared to the base model

[1] We also use the base version of mBERT in this work.

Model	Langauge Pairs					
Source	EN	EN	ZH	ZH	AR	AR
Target	ZH	AR	EN	AR	EN	ZH
BERT-CRF	68.5	30.9	37.5	20.1	40.1	58.8
BERT-CRF-LAT	70.0	33.5	41.2	20.3	37.2	55.6
BERT-CRF-LMFT	69.4	33.4	42.9	20.0	36.5	56.3
BERT-CRF-CCCAR	72.1	42.7	45.8	20.7	40.7	59.8
XLMR-CRF	70.5	43.5	41.7	32.8	45.4	61.8
XLMR-CRF-LAT	70.2	43.4	42.3	33.2	45.2	60.9
XLMR-CRF-LMFT	71.1	43.7	42.1	32.9	45.9	62.1
XLMR-CRF-CCCAR	74.4	44.1	49.5	34.3	46.3	62.9
OACLED	74.6	44.9	45.8	35.1	48.0	63.1
CLRED (ours)	**76.6**	**46.4**	**50.8**	**39.2**	**49.2**	**67.3**

Table 1: Performance (F1 scores) on test data for ED in six cross-lingual settings. Each column corresponds to one language pair where source languages are shown above target languages. The proposed model is significantly better than other models with $p < 0.01$.

#	Model	Langauge Pairs					
		EN	EN	ZH	ZH	AR	AR
		ZH	AR	EN	AR	EN	ZH
1	CLRED (full)	76.6	46.4	50.8	39.2	48.2	67.3
2	No retrieval	70.5	43.5	41.7	32.8	45.4	61.8
3	No $sim(\overline{W}, \overline{R_k})$ in $a_{i,k,j}$	72.9	43.7	46.8	37.5	45.9	65.4
4	Not update miniLM$_*$	74.0	44.6	47.4	37.8	45.6	64.3
5	Not update miniLM$_W$	72.5	45.3	46.3	36.6	45.3	61.5
6	Not update miniLM$_U$	73.5	45.7	47.7	38.0	46.2	66.3
7	No warm up	75.5	44.4	46.9	32.9	47.4	63.6
8	No $\mathcal{L}_{reg}$	74.1	45.3	48.4	38.3	47.8	66.7
9	With unlabeled source	74.4	43.8	45.5	38.5	46.7	66.2

Table 2: Ablation study.

XLMR-CRF, the key distinction in our model involves the retrieval model. Table 2 studies the performance of the ablated/varied versions of the retrieval model in CLRED over the test sets of different language pairs. In particular, line 2 "**No retrieval**" completely removes the retrieval component from CLRED (i.e., XLMR-CRF with no augmentation representation $\overline{a}_i$). As the performance is significantly reduced, it demonstrates the benefit of the retrieval model for our CLRED model. In line 3 with "**No** $sim(\overline{W}, \overline{R_k})$ **in** $a_{i,k,j}$", we do not include the retrieval similarity between the retrieved and input sentences in the attention weights $a_{i,k,j}$ for augmentation representation. This model also implies that the retrieval and ED models are disconnected and the retrieval components miniLM$_W$ and minLM$_U$ are freeze during the training process for the ED model. As such, the poorer performance of "**No** $sim(\overline{W}, \overline{R_k})$ **in** $a_{i,k,j}$" in Table 2 clearly confirms the importance of $sim(\overline{W}, \overline{R_k})$ in the attention weights for CLRED.

Next, as we update the two retrieval models miniLM$_W$ and minLM$_U$ after every Q steps in the joint training, lines 4, 5, and 6 explore the variants where we fix the two models (line 4) or only update one of them (lines 5 and 6) during the training of ED. As can be seen, the degraded performance in lines 4, 5, and 6 highlight the necessity to update and synchronize miniLM$_W$ and minLM$_U$ to achieve the best performance for CLRED. In addition, line 7 "**No warm up**" and line 8 "**No** $\mathcal{L}_{reg}$" demonstrate the benefits of our warm up step and XLMR-computed similarity regularization (respectively) for the retrieval model as removing any of them will lead to significant performance drops. Finally, in line 9, instead of using unlabeled data in the target language, the retrieval component retrieves relevant sentences from unlabeled data of the source language that is obtained by removing labels from the training data of the source language (i.e., excluding the input sentence W) for our cross-lingual learning setting. As can be seen, unlabeled data in the source language cannot guarantee the best cross-lingual performance for ED, thus testifying to the importance of using unlabeled sentences in the target language for cross-lingual ED.

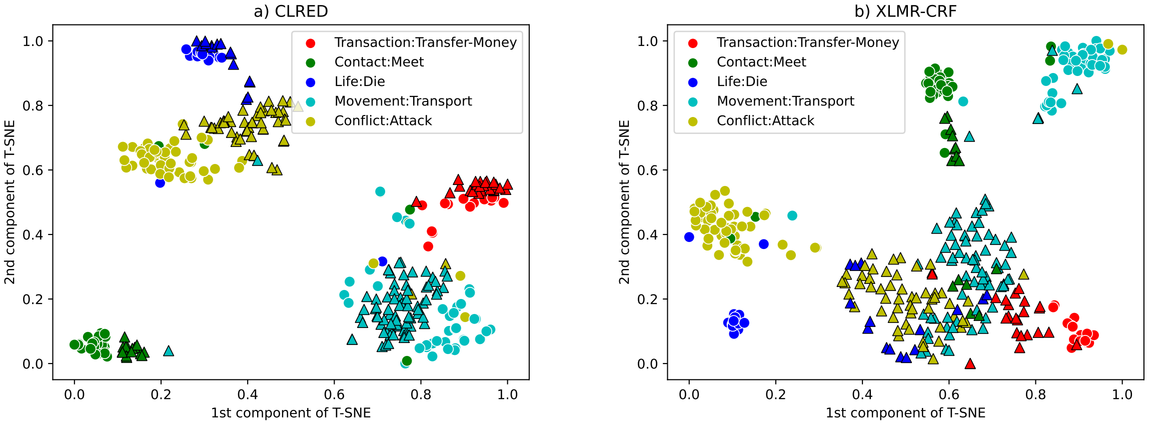

Figure 1: T-SNE visualizations for prediction representations of the words from English (i.e., source) and Chinese (i.e., target) data. Circles and triangles represent English and Chinese examples respectively while colors indicate event types.

Speed Evaluation: Given the retrieval component with representation computation for U with miniLM, we evaluate the running time for our model CLRED. Using the time for XLMR-CRF as the baseline, Table 3 presents the training and inference time for the full model CLRED (averaged over six language pairs). For reference, we also report the time for the variant of CLRED where the retrieval model with miniLM$_W$ and miniLM$_U$ is fixed during the training of the ED model. Overall, the training time of our retrieval-based model is double that for the base model XLMR-CRF; however, our inference time is only increased by 1.18 times. Note that in practice, the FAISS open-source toolkit (Johnson et al., 2021) can be used to precompute and index the representations for the sentences in U. This will allow us to handle larger unlabeled set U and achieve efficient vector search.

Model	Training	Inference
XLMR-CRF	1.00x	1.00x
CLRED with fixed retrieval	1.42x	1.18x
CLRED (full)	2.07x	1.18x

Table 3: Latency cost for our retrieval-based model CLRED. All results are computed with a single Nvidia V100 GPU.

Analysis: To better understand the operation of CLRED, we analyze the examples in the test sets for the target languages that can be correctly predicted by CLRED, but cannot be recognized by the non-retrieval baseline XLMR-CRF. A key insight from our analysis is that XLMR-CRF tends to incorrectly recognize event types in the input texts of the target languages due to the ambiguity of context. CLRED can fix the errors in these cases as the retrieval component is able to return relevant sentences that contains the same correct event types as the inputs. As such, the augmentation representation from the retrieved sentences can strengthen the context information to produce correct type prediction. For instance, consider the language pair ZH→EN (i.e., Chinese is the source and English is the target) with the sentence "*Blasphemy is punishable by death under the Pakistan Penal Code.*" in the target language. XLMR-CRF incorrectly predicts "*death*" as an event trigger of type *Life:Die* while CLRED can correctly identify "*punishable*" as an event trigger of type *Justice:Sentence*. This is understandable given that the two retrieved sentences from CLRED involves: "*Big "snake head" Weng Jinshun sentenced to life imprisonment.*" and "*Roman was sentenced to seven years in prison.*", which clearly express *Justice:Sentence* events.

In addition, to illustrate the impact of augmentation representation from the retrieved target-language sentences for CLRED, Figure 1 presents the t-SNE visualization for the representation vectors that are computed by XLMR-CRF and CLRED to predict event types for the words in the source- and target-language test data. As can be seen, the representations learned by XLMR-CRF for the source language examples are quite separate from those for the target language. In contrast, with augmentation representation, CLRED can better align representations for the source and target examples of the same event types, thus improving cross-lingual performance for ED.

4 Related Work

ED has been studied mostly for monolingual settings, involving feature-based models (Liao and Grishman, 2011; Li et al., 2013; Yang and Mitchell, 2016) and recent deep learning models (Nguyen and Grishman, 2015; Chen et al., 2015; Nguyen and Grishman, 2018; Man Duc Trong et al., 2020; Zhang et al., 2019; Lin et al., 2020; Pouran Ben Veyseh et al., 2021a,b). Cross-lingual transfer learning for ED has gained more interests recently where different resources are leveraged to project the representations for different languages into the same space, including bilingual dictionaries/parallel corpora (Muis et al., 2018; Liu et al., 2019) and multilingual language models (M'hamdi et al., 2019; Ahmad et al., 2021; Majewska et al., 2021). To further bridge the gap between the representations for cross-lingual ED, (Nguyen et al., 2021b) explores adversarial training with language discriminators (Huang et al., 2019; Lange et al., 2020; He et al., 2020; Guzman et al., 2022) and representation matching of similar objects to remove language-specific features. We also note that these methods are motivated from domain adaptation methods that aim to avoid domain-specific features (Ganin and Lempitsky, 2015; Cicek and Soatto, 2019; Tang et al., 2020; Trung et al., 2022; Ngo et al., 2022). In contrast, our model introduces additional augmentation representations from retrieval to achieve language-universal representations.

5 Conclusion

We present a novel method for cross-lingual transfer learning for ED. Instead of removing language-specific features, our model augments the representations for the input sentences with those from relevant sentences in the target language to align the representations for the source and target languages. Our method involves a retrieval component to obtain relevant sentences that is jointly trained with the ED model. Our proposed method demonstrates the state-of-the-art cross-lingual performance over six different language pairs.

Limitations

In this work we present a novel method based on representation augmentation to solve cross-lingual transfer learning for event detection (ED). Although our experiments demonstrate the effectiveness of the proposed method, there are still some limitations that can be improved in future work. First, our current method only leverages sentence-level context in input document to perform ED over different languages. This might not be optimal as document-level context has been shown to be helpful for ED (Pouran Ben Veyseh et al., 2021b) that can be explored in future research to improve our cross-lingual models. Second, the evaluation for our model is limited to only three popular languages (English, Chinese, and Arabic) that are supported by existing pre-trained language models, unlabeled data, and text processing tools. As such, it is unclear whether the method can be adapted to many other languages with limited access to such resources (e.g., low-resource languages). We believe this is an important direction that can be investigated in future work to advance our understanding for ED models. Finally, our method requires joint training with a retrieval model (based on multilingual pre-trained language models) that can impose additional computational costs (as shown in Table 3). Reducing necessary computational costs for our model is an important direction to make it more accessible for different applications and domains.

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

Wasi Uddin Ahmad, Nanyun Peng, and Kai-Wei Chang. 2021. Gate: Graph attention transformer encoder for cross-lingual relation and event extraction. In *Proceedings of the Association for the Advancement of Artificial Intelligence (AAAI)*.

Xilun Chen, Ahmed Hassan Awadallah, Hany Hassan, Wei Wang, and Claire Cardie. 2019. Multi-source cross-lingual model transfer: Learning what to share. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3098–3112, Florence, Italy. Association for Computational Linguistics.

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

Safa Cicek and Stefano Soatto. 2019. Unsupervised domain adaptation via regularized conditional alignment. In *Proceedings of the International Conference on Computer Vision (ICCV)*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*.

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

Luis Nateras Guzman, 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.

Keqing He, Yuanmeng Yan, and Weiran Xu. 2020. Adversarial cross-lingual transfer learning for slot tagging of low-resource languages. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*.

Lifu Huang, Heng Ji, and Jonathan May. 2019. Cross-lingual multi-level adversarial transfer to enhance low-resource name tagging. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3823–3833, Minneapolis, Minnesota. Association for Computational Linguistics.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2021. Billion-scale similarity search with gpus. In *IEEE Transactions on Big Data*.

Armand Joulin, Piotr Bojanowski, Tomas Mikolov, Hervé Jégou, and Edouard Grave. 2018. Loss in translation: Learning bilingual word mapping with a retrieval criterion. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2984, Brussels, Belgium. Association for Computational Linguistics.

Phillip Keung, Yichao Lu, and Vikas Bhardwaj. 2019. Adversarial learning with contextual embeddings for zero-resource cross-lingual classification and NER. 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 1355–1360, Hong Kong, China. Association for Computational Linguistics.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*.

Viet Dac Lai, Tuan Ngo Nguyen, and Thien Huu Nguyen. 2020. Event detection: Gate diversity and syntactic importance scores for graph convolution neural networks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5405–5411, Online. Association for Computational Linguistics.

Lukas Lange, Anastasiia Iurshina, Heike Adel, and Jannik Strötgen. 2020. Adversarial alignment of multilingual models for extracting temporal expressions from text. *arXiv preprint arXiv:2005.09392*.

Qi Li, Heng Ji, and Liang Huang. 2013. Joint event extraction via structured prediction with global features. In *Proceedings of the 51th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Shasha Liao and Ralph Grishman. 2011. Acquiring topic features to improve event extraction: in preselected and balanced collections. In *Proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP)*.

Ying Lin, Heng Ji, Fei Huang, and Lingfei Wu. 2020. A joint neural model for information extraction with global features. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*.

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

Jian Liu, Yubo Chen, Kang Liu, and Jun Zhao. 2019. Neural cross-lingual event detection with minimal parallel resources. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 738–748, Hong Kong, China. Association for Computational Linguistics.

Shulin Liu, Yubo Chen, Kang Liu, and Jun Zhao. 2017. Exploiting argument information to improve event detection via supervised attention mechanisms. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*.

Weiyi Lu and Thien Huu Nguyen. 2018. Similar but not the same: Word sense disambiguation improves event detection via neural representation matching. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4822–4828, Brussels, Belgium. Association for Computational Linguistics.

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

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

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.

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 (COLING)*.

Nghia Ngo, Bonan Min, and Thien Nguyen. 2022. Unsupervised domain adaptation for joint information extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5894–5905, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Nghia Ngo, Tuan Ngo Nguyen, and Thien Huu Nguyen. 2020. Learning to select important context words for event detection. In *Proceedings of the 24th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*.

Minh Van Nguyen, Viet Lai, and Thien Huu Nguyen. 2021a. Cross-task instance representation interactions and label dependencies for joint information extraction with graph convolutional networks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 27–38, Online. Association for Computational Linguistics.

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

Thien Huu Nguyen, Kyunghyun Cho, and Ralph Grishman. 2016. Joint event extraction via recurrent neural networks. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 300–309, San Diego, California. 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 *Proceedings of the Association for the Advancement of Artificial Intelligence (AAAI)*.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020. MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7654–7673, Online. Association for Computational Linguistics.

Amir Pouran Ben Veyseh, Viet Lai, Franck Dernoncourt, and Thien Huu Nguyen. 2021a. Unleash GPT-2 power for event detection. 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 6271–6282, Online. Association for Computational Linguistics.

Amir Pouran Ben Veyseh, Minh Van Nguyen, Nghia Ngo Trung, Bonan Min, and Thien Huu Nguyen. 2021b. 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.

Hui Tang, Ke Chen, and Kui Jia. 2020. Unsupervised domain adaptation via structurally regularized deep clustering. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*.

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

David Wadden, Ulme Wennberg, Yi Luan, and Hannaneh Hajishirzi. 2019. Entity, relation, and event extraction with contextualized span representations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*.

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

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*.

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

ACL 2023 Responsible NLP Checklist

A For every submission:

□ A1. Did you describe the limitations of your work?
Left blank.

□ A2. Did you discuss any potential risks of your work?
Left blank.

□ A3. Do the abstract and introduction summarize the paper's main claims?
Left blank.

□ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B □ Did you use or create scientific artifacts?

Left blank.

□ B1. Did you cite the creators of artifacts you used?
Left blank.

□ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Left blank.

□ 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)?
Left blank.

□ 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?
Left blank.

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

□ 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.
Left blank.

C □ Did you run computational experiments?

Left blank.

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

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

□ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Left blank.

□ 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?
Left blank.

□ 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.)?
Left blank.

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

Left blank.

□ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Left blank.

□ 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)?
Left blank.

□ 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?
Left blank.

□ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Left blank.

□ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Left blank.