Bridging the Gap between Native Text and Translated Text through Adversarial Learning: A Case Study on Cross-Lingual Event Extraction

Pengfei Yu[1]**, Jonathan May**[2]**, Heng Ji**[1]
[1]University of Illinois Urbana-Champaign [2]University of South California
{pengfei4,hengji}@illinois.edu
jonmay@isi.edu

Abstract

Recent research in cross-lingual learning has found that combining large-scale pretrained multilingual language models with machine translation can yield good performance (Phang et al., 2020; Fang et al., 2021). We explore this idea for cross-lingual event extraction with a new model architecture that jointly encodes a source language input sentence with its translation to the target language during training, and takes a target language sentence with its translation back to the source language as input during evaluation. However, we observe significant representational gap between the native texts and translated texts, both in the source language and the target language. This representational gap undermines the effectiveness of cross-lingual transfer learning for event extraction with machine-translated data. In order to mitigate this problem, we propose an adversarial training framework that encourages the language model to produce more similar representations for the translated text and the native text. To be specific, we train the language model such that its hidden representations are able to fool a jointly trained discriminator that distinguishes translated texts' representations from native texts' representations. We conduct experiments on cross-lingual event extraction across three languages. Results demonstrate that our proposed adversarial training can effectively incorporate machine translation to improve event extraction, while simply adding machine-translated data yields unstable performance due to the representational gap.[1]

1 Introduction

There are over 6,000 living languages in the world, and for many of them, too little appropriate data exists to build natural language processing (NLP) models. Cross-lingual learning has been proposed to leverage resources in data-rich languages to train NLP models for data-scarce languages (Ruder et al., 2019). There are two main strategies for building cross-lingual models: (1) train models with multilingual language models and language-universal features that are transferable to the target language (Huang et al., 2019; Hsu et al., 2019; Hu et al., 2020a; Luo et al., 2020; Wei et al., 2021; Ouyang et al., 2021; Liu et al., 2019; Subburathinam et al., 2019; M'hamdi et al., 2019; Ahmad et al., 2021); (2) use machine translation models in a pipeline, either by transforming annotated training data into the desired target language to build target-language models, or by translating data at inference time into the source language and applying source-language models (Cui et al., 2019; Hu et al., 2020a; Yarmohammadi et al., 2021). The first approach relies on the quality of the constructed multilingual semantic space; the discrepancy between source-language training data and target-language evaluation data may cause overfitting. The second approach does not require a perfect multilingual semantic space since models can be trained in a monolingual fashion, but it depends on the quality of machine translation.

A combination of both approaches showed good performance on a variety of tasks such as natural language inference and question answering (Phang et al., 2020; Fang et al., 2021), but is underexplored for event extraction. Compared with previous research in cross-lingual event extraction mainly adopting the first approach (Liu et al., 2019; Subburathinam et al., 2019; M'hamdi et al., 2019; Ahmad et al., 2021), we explore the idea of combining both machine translations and language-universal representations for cross-lingual event extraction in this work. We perform translation by extending the previous effort on cross-lingual reading comprehension (Hsu et al., 2019) and question answering (Hu et al., 2020a) by adding special tags around the trigger and entity spans to translate the annotations. We use a multilingual language model to simultaneously encode a sentence and its corresponding

[1]Code at https://github.com/Perfec-Yu/CrossIE

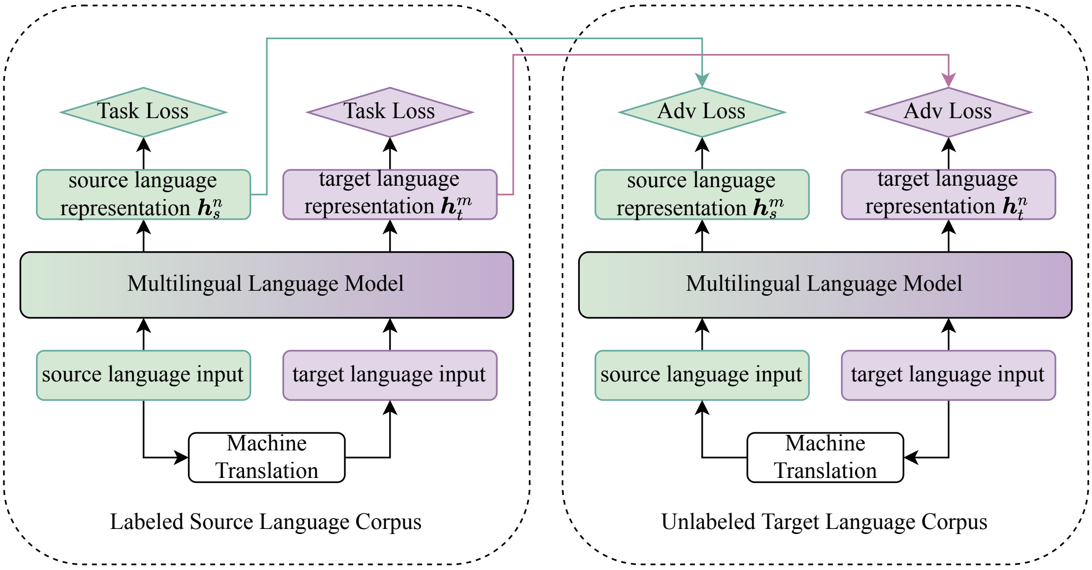


Figure 1: Overall cross-lingual information extraction framework

translation as shown on the left side of Figure 1. For example, in an English-to-Chinese cross-lingual learning setting, we would train a model with English sentences with their Chinese translations as training data, and evaluate our model with Chinese sentences and their English translations as inputs. Since our work includes both cross-lingual learning and machine translation, to avoid ambiguity, we will use "source" language as the one we perform cross-lingual learning from, and "target" language as the one we perform cross-lingual learning to. We will call texts before translation "native" text and text after translation "translated" text for the machine-translation-related descriptions.

We found that one challenge in cross-lingual event learning with machine translations is that the machine-translated text $\mathcal{M}_{\mathcal{K}\to\mathcal{L}}$ from one language $\mathcal{K}$ into another language $\mathcal{L}$ may be different from the native text in the target language $\mathcal{N}_{\mathcal{L}}$. This difference is also introduced and studied as the problem of "translationese" (translated text as a different language) in previous machine translation research (Pylypenko et al., 2021; Riley et al., 2020). In cross-lingual event extraction, we observe from a simple preliminary experiment that there indeed exists a distinguishable gap between representations of native texts $\boldsymbol{H}(\mathcal{N}_{\mathcal{L}})$ and translated text $\boldsymbol{H}(\mathcal{M}_{\mathcal{K}\to\mathcal{L}})$ in some multilingual language model $\boldsymbol{H}$. The pretrained language models appear to be "unaccustomed" to the translated text. The representational gap will negatively impact the cross-lingual learning with machine-translated data. Since we, as introduced above, simultaneously encode a native source language sentence $\mathcal{N}_{\mathcal{S}}$ and its translation into the target $\mathcal{M}_{\mathcal{S}\to\mathcal{T}}$ language during training, and a native target language sentence $\mathcal{N}_{\mathcal{T}}$ and its translation back to the source language $\mathcal{M}_{\mathcal{T}\to\mathcal{S}}$ during evaluation, the problem of representational gap between $\mathcal{N}_{\mathcal{S}}$ and $\mathcal{M}_{\mathcal{T}\to\mathcal{S}}$, as well as $\mathcal{N}_{\mathcal{T}}$ and $\mathcal{M}_{\mathcal{S}\to\mathcal{T}}$ need to be resolved. Here $\mathcal{S}$ and $\mathcal{T}$ refer to the source and the target language respectively.

In order to mitigate the representational gap problem between machine-translated text $\mathcal{M}$ and native text $\mathcal{N}$ in both source and target languages, we propose to take advantage of an unlabeled corpus in the target language and use adversarial training to make the encoder produce more similar representations for $\mathcal{N}_{\mathcal{S}}$ and $\mathcal{M}_{\mathcal{T}\to\mathcal{S}}$, as well as $\mathcal{N}_{\mathcal{T}}$ and $\mathcal{M}_{\mathcal{S}\to\mathcal{T}}$. The adversarial framework trains the language model $\boldsymbol{H}$ such that its hidden representations can fool a jointly trained discriminator that distinguishes translated texts' representations $\boldsymbol{H}(\mathcal{M})$ from native texts' representations $\boldsymbol{H}(\mathcal{N})$. Our complete cross-lingual IE framework is shown in Figure 1, which combines translation-based methods with transfer-based methods, and uses an unlabeled target language corpus to improve the representations in multilingual language models. Our method shows superior performance on event trigger labeling and argument role labeling, and through quantitative studies, we observe that adversarial training indeed makes the multilingual language model generate closer representations for the translated text and the native text. We believe our proposed adversarial training can also be helpful in other NLP tasks where machine translation can boost performance.

To summarize, our contributions are two-fold:

- We observe the gap between representations of the machine-translated text and the native text in multilingual language models.
- We propose an adversarial training method to close the representational gap, which improves event extraction performance.

2 Approach

In this section, we will start with a simple preliminary experiment to validate the problem of the representational gap, and then introduce our approaches to cross-lingual event trigger and argument role labeling. For both tasks, we first design specific methods to use machine translation models to translate source language annotations into the target language. We then use XLM-RoBERTa (Conneau et al., 2020) to encode pairs of parallel sentences simultaneously into hidden representations. Task-specific losses are used on top of the hidden representations. In order to make the multilingual language model produce more similar representations for translated sentences and native sentences, we further use an unlabeled target language corpus for adversarial training.

2.1 Preliminary Experiment on Representational Gap

We translate Chinese sentences from the ACE 2005 Chinese corpus into English and encode the translated English sentences $\mathcal{M}_{\mathrm{ZH}\rightarrow\mathrm{EN}}$ and native English sentences $\mathcal{N}_{\mathrm{EN}}$ in the ACE 2005 English data using the multilingual language model XLM-RoBERTa (Conneau et al., 2020). We then train linear Support Vector Machines (SVMs) (Cortes and Vapnik, 1995) to classify the encoded representations of these two sets of sentences as $\mathcal{N}$ative or $\mathcal{M}$achine-translated. The model achieves 83.4% accuracy on a held-out test set classifying the translated English sentences $\mathcal{M}_{\mathrm{ZH}\rightarrow\mathrm{EN}}$ and native English sentences $\mathcal{N}_{\mathrm{EN}}$. We also perform translation from English to Chinese and achieve 93.4% accuracy classifying native Chinese sentences $\mathcal{N}_{\mathrm{ZH}}$ and translated Chinese sentences $\mathcal{M}_{\mathrm{EN}\rightarrow\mathrm{ZH}}$. Both numbers are significantly higher than the random 50% accuracy, indicating that the translated text and the native text are almost linearly separable in the multilingual language models and hence validating the representational gap between the two types of texts.

2.2 Event Trigger Labeling

In monolingual event trigger labeling, the input to the model is a sequence of text tokens $\{w_0, w_1, \ldots, w_l\}$. The model identifies consecutive text spans as event triggers and classifies the spans into event types. We first obtain the token representations using the text encoder as $\{\boldsymbol{h}_0, \boldsymbol{h}_1, \ldots, \boldsymbol{h}_l\}$. Then we apply a linear layer to classify each token into one of the event types.

For the cross-lingual setting, we first translate the monolingual training data in the source language into the target language together with the trigger annotations. We will explain the translation process in Section 2.4. We encode the source language text sequence $\{w_{s0}, w_{s1}, \ldots, w_{sl}\}$ and its translation $\{w_{t0}, w_{t1}, \ldots, w_{tk}\}$ using the XLM-RoBERTa (Conneau et al., 2020) model. We also adopt a special fusion strategy as introduced in the FILTER (Fang et al., 2021), which adds cross-lingual attention between the source language text and its translation in some hidden Transformer layers. We apply the classification step as in the monolingual setting for both w_s and w_t. The task loss is the summation of losses from w_s and w_t.

$$\mathcal{L} = \mathcal{L}_s + \mathcal{L}_t. \tag{1}$$

In the training phase described above, the input sequences to the multilingual language model consist of a native source language sequence w_s^n and its translations w_t^m. In the evaluation phase, the input sequence becomes a native target language sequence w_t^n and a translated source language sequence w_s^m. Therefore, we need to bridge the representational gap in the multilingual LM between two pairs: (w_s^n, w_s^m) and (w_t^m, w_t^n). In order to encourage the multilingual LM to generate closer representations for w_s^n and w_s^m, as well as for w_t^m and w_t^n, we further propose an adversarial loss using another unlabeled target language corpus. We first translate the unlabeled target language corpus, from which we sample w_t^n, into the source language (w_s^m) to construct an unlabeled parallel corpus. Then parallel sentence pairs (w_s^m, w_t^n) in the unlabeled corpus are encoded by the multilingual LM in the same way as the labeled training sentence pairs (w_s^n, w_t^m). We train two additional two-layer discriminators, D_s and D_t. D_s attempts to distinguish native source language representations $\boldsymbol{w}_s^n$ from translated source language representations $\boldsymbol{w}_s^m$. D_t attempts to distinguish translated target language representations $\boldsymbol{w}_t^m$ from the native

	Trigger Labeling	Argument Role Labeling
Source Language	Now that Enron has ceased to exist, Bechtel and GE are suing the Indian Government for 5.6 billion US dollars.	The electricity that Enron produced was so exorbitant that the government decided it was cheaper not to buy electricity and <a>pay Enron the mandatory fixed charges specified in the contract.
Target Language	现在安然已经不复存在，柏克德和通用电气正在起诉印度政府，要求赔偿56 亿美元	安然生产的电力如此昂贵，以至于政府决定不购买电力并<a>支付安然合同中规定的强制性固定费用更便宜

Table 1: Example of training data translation for trigger labeling and argument role labeling.

target language representations $\boldsymbol{w}_t^n$. The adversarial loss is also illustrated in Figure 1. For adversarial training, we adopt W-GAN (Arjovsky et al., 2017) with gradient penalty (Gulrajani et al., 2017) in this work. Specifically, D_s and D_t are two-layer neural networks with one output unit, i.e., they output single scalars. Optimization targets of the two discriminators are

$$\begin{aligned}\mathcal{L}_{D_s} &= D_s(\boldsymbol{h}_s^m) - D_s(\boldsymbol{h}_s^n;\theta) \\ &\quad + \mathrm{GP}(D_s;\boldsymbol{h}_s^m,\boldsymbol{h}_s^n), \\ \mathcal{L}_{D_t} &= D_t(\boldsymbol{h}_t^m) - D_t(\boldsymbol{h}_t^n) \\ &\quad + \mathrm{GP}(D_t;\boldsymbol{h}_s^m,\boldsymbol{h}_s^n).\end{aligned} \tag{2}$$

Here GP refers to the gradient penalty loss in Gulrajani et al. (2017) to regularize the discriminators. D_s and D_t are both neural networks that output a single value. We use $D_s(\boldsymbol{w}_s^m;\theta)$ to denote the average output value of all token representations in the sequence $\boldsymbol{w}_s^m$, and D_t in an analogous way. We expect our multilingual LM to produce representations that confuse both discriminators. The optimization target for the encoder is,

$$\begin{aligned}\mathcal{L}_G &= D_s(\boldsymbol{h}_s^n) - D_s(\boldsymbol{h}_s^m) \\ &\quad + D_t(\boldsymbol{h}_t^n) - D_t(\boldsymbol{h}_t^m).\end{aligned} \tag{3}$$

The gradients of the loss in Equation (1) are back propagated to both the multilingual language model and the trigger classification layers. The gradients of the discriminator loss in Equation (2) are back propagated to D_s and D_t only. The gradients of the generator loss in Equation (3) are back propagated to the multilingual language model. In practice we find that it is beneficial to back propagate $\mathcal{L}_G$ to only the last layer of the XLM-RoBERTa to match the capacity of the discriminators D_s and D_t.

2.3 Argument Role Labeling

Argument Role Labeling identifies the roles entities play in events. Assuming gold-standard entity spans are provided, the input is a sentence x with a trigger span and an entity span, and the model predicts the argument role of the entity in the event. We use an additional `None` label for the case where the entity does not participate in the event.

For monolingual prediction, we first insert into the sentence two pairs of anchors to specify spans for the trigger and the entity: (“<a>”, “”) around the trigger span and (“”, “”) around the entity span. We encode the modified sentence into hidden representation $\boldsymbol{x}$ by a pretrained language model. We consider the token representation for the CLS token inserted into the beginning of every sentence $\boldsymbol{x}_{CLS}$ as the summarization of the sentence and feed it to a linear layer for classification. For adversarial training, we use a similar loss as in Equations (2) and (3), but use the CLS token representation $\boldsymbol{x}_{CLS}$ as the input to the discriminators.

2.4 Annotation Translation

We show two examples in Table 1 for translating annotations for trigger labeling and argument role labeling respectively. For trigger labeling, we first enclose each trigger span in the source language sentences with special tokens (“”, “”) inspired by previous efforts on question answering (Hu et al., 2020b). The machine translation model is applied to the new sentence. If the paired special tokens (“”, “”) exist in the translated sentence, we label the text span inside the pair as the event trigger. Otherwise we consider the translation as invalid and discard the target language loss $\mathcal{L}_t$ in Equation (1) when training. We still use the invalid translations for the adversarial training loss in Equation (2) and Equation (3) since the computation of these losses doesn't require trigger spans.

For argument role labeling, we take advantage of the anchor tokens used for training and simply translate the sentences with trigger and entity spans enclosed by anchor tokens into the target language. Due to the imperfections in the machine translation model, there are corrupted translated samples missing "<a>" or "<b>" tags. However, since the role labeling model architecture doesn't require the existence of these tags to be runnable, we still consider them as valid inputs and use the corrupted translated samples as training data for both the target language loss $\mathcal{L}_t$ in Equation (1) and the adversarial losses in Equation (2) and Equation (3).

2.5 Evaluation

At inference time, the inputs to the framework are sentences in the target language. We first translate the target language sentence into the source language using the same machine translation model used for the unlabeled target language corpus during training and apply our framework to the sentence pairs. We make predictions using the hidden representations of the target language.

3 Experiments

3.1 Dataset and Machine Translation

We use the ACE[2] 2005 dataset for experiments. We study all six transfer learning settings among the three languages in the dataset: Arabic, Chinese and English. We follow previous work on event extraction (Lin et al., 2020) to split the ACE dataset for the trigger labeling task. For the argument role labeling task, previous work (Subburathinam et al., 2019; Ahmad et al., 2021) has adopted a different split from Lin et al. (2020). We therefore follow the split in (Subburathinam et al., 2019; Ahmad et al., 2021) in this task. However, since their processed version of ACE dataset is not available, we use our own processed version and retrain their models on our version for comparison. We provide basic data statistics in Table 2. We also provide more fine-grained data statistics in Appendix. There are some other competitive cross-lingual event extraction baselines that we are not able to compare due to limited availibity of code or split information. We provide further discussion in the related work section. We use Google Translate for all machine translation components.

[2]https://www.ldc.upenn.edu/collaborations/past-projects/ace

		Trigger		Role	
		#Docs	#Events	#Cands	#Args
EN	Train	529	4,419	14,036	7,018
	Dev	28	468	1,754	719
	Test	40	424	1,756	878
ZH	Train	551	2,926	11,826	5,931
	Dev	40	217	1482	602
	Test	42	190	1484	578
AR	Train	303	1,751	7,918	3,959
	Dev	50	255	990	495
	Test	50	262	990	495

Table 2: Data statistics for ACE 2005 dataset. EN, ZH and AR refer to the English, the Chinese and the Arabic splits respectively. The trigger labeling task (Trigger) and the argument role labeling task (Role) use different splits to compare with previous methods. We present the number of documents and the number of event mentions for Trigger splits. For Role splits, we present the number of candidate trigger-entity pairs for prediction (#Cands) and the total number of pairs that hold some argument role relationship (#Args).

3.2 Experiment Settings

Methods in Comparison We compare the following approaches in evaluation:

Direct, which directly trains a model on the source language with a multilingual language model and evaluates it on the target language. We use XLM-RoBERTa as the multilingual LM to be comparable with our method;

GATE (Ahmad et al., 2021) is a state-of-the-art cross-lingual model for the argument role labeling task. Hence we only compare with GATE in the argument role labeling task;

Trans is a baseline that excludes our proposed adversarial loss but keeps all the remaining components;

Trans+Adv is our proposed framework;

Target Supervision is a mono-lingual IE model trained on the target language data.

Evaluation Settings Except for **Target Supervision**, all cross-lingual models are trained with the source language annotations. We use the target language training corpus without annotations to compute the adversarial loss in our proposed method. We report F1 scores in the following sections and include precision and recall scores in Appendix.

Event Trigger Labeling	AR - EN	ZH - EN	AR - ZH	EN - ZH	ZH - AR	EN - AR
Direct	39.8	44.4	33.4	46.9	36.7	39.0
Trans	39.4	46.3	38.8	47.3	36.6	39.3
Trans+Adv (ours)	**41.5**	**54.6**	**40.1**	**49.3**	**38.4**	**42.3**
Target Supervision	68.5		65.6		56.1	

(a) Event trigger labeling.

Argument Role Labeling	AR - EN	ZH - EN	AR - ZH	EN - ZH	ZH - AR	EN - AR
GATE	50.3	57.0	55.7	63.6	65.1	65.0
Direct	56.8	61.5	64.6	71.7	64.0	62.5
Trans	57.5	60.6	64.9	71.3	63.8	62.2
Trans+Adv (ours)	**58.4**	**62.9**	**65.6**	**72.0**	**68.0**	**65.1**
Target Supervision	77.2		82.0		77.8	

(b) Argument role labeling.

Table 3: F1(%) scores for the cross-lingual event extractions. GATE (Ahmad et al., 2021) is a state-of-the-art method for cross-lingual argument role labeling. Direct,Trans and Target Supervision are introduced in Section 3.2.AR,EN and ZH correspond to Arabic, English and Chinese respectively.

3.3 Experiment Results

We show the evaluation results for trigger labeling in Table 3a. We show results for the argument role labeling task in Table 3b. Our model shows superior performance compared with other cross-lingual baselines in both trigger labeling and role labeling tasks and across all six cross-lingual transfer settings. Our model outperforms the `Trans` baseline that is trained without the adversarial loss. This indicates that our proposed approach effectively narrows the gap between the translations and the original natural language to improve the performance. Moreover, we notice that the `Trans` that uses translated data for training cannot consistently outperform the `Direct` baseline which doesn't use translated data. This shows that the representational gap can have a negative impact on the model performance than the positive impact brought by including the translated data. In the following sections, we provide further analysis on the representational gap, our model's improvements and remaining errors.

3.4 Effect of Adversarial Training

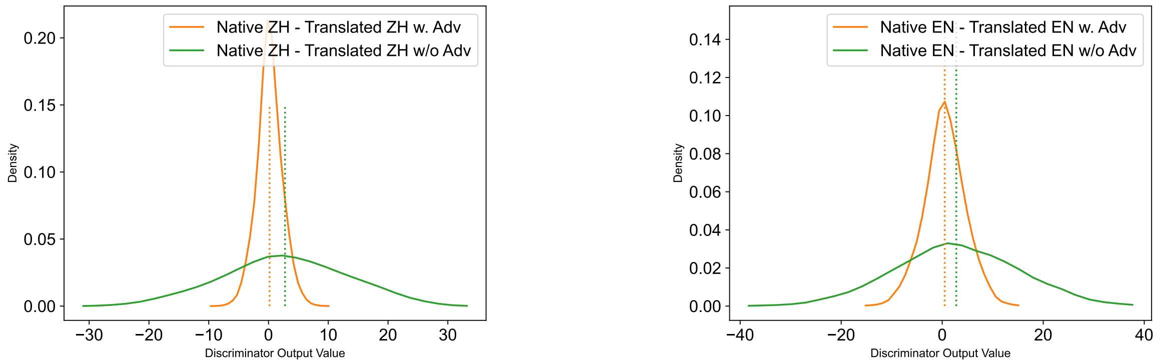
In this section we evaluate the effect of the adversarial training on reducing the representational gap. Hence we compare our model against the `Trans` baseline that doesn't use the adversarial training loss. We take the English-to-Chinese transfer learning setting as a case study in this section.

Argument Role Labeling	EN-to-ZH		
	T-ZH	ZH	Diff
Trans	74.3	71.3	-3.0
Trans+Adv (ours)	74.5	72.0	-2.5

Table 4: F1 scores (in %) of the English-Chinese cross-lingual argument role labeling models on translated Chinese test corpus (from English test corpus), `T-ZH` and the native Chinese test corpus, `ZH`. `Diff` is the performance gap between two test corpora.

A straightforward way to examine the representational gap between the native text and the translated text inside a model is to compare its performance on these two types of texts on role labeling. In Table 4, we report the F1 scores on the native Chinese test set and translated Chinese text from English dataset respectively. The performance on translated Chinese is better than native Chinese since both models use the translated Chinese instead of native Chinese during training. Our adversarial training method shows a smaller performance gap compared with the non-adversarial baseline, indicating that our model indeed reduces the representational gap.

In addition to this evaluation, we further check whether the proposed generator loss helps the model to produce representations that confuse the discriminators. We compare the discriminator out-



(a) Native Chinese v.s. Translated Chinese

(b) Native English v.s. Translated English

Figure 2: Distribution of differences in discriminator outputs between native text and translated text. We compute the density with NumPy[3] histogram function on original data points. *w. Adv* refers to our model with the adversarial training. *w/o Adv* is the output of the additional discriminators trained on the baseline *Trans* without adversarial training. (See Appendix for details on how the additional discriminators are trained)

Task	Sentence	Error
Trigger Labeling	...徐鹏航...支持参与亲属购买内部职工股... (...Penghang Xu... supported and participated in relatives' purchasing internal employee shares)	Baseline model makes a false positive prediction of "支持" (support) as a trigger for `Transfer-Money` event

Table 5: An Example error that the baseline approach fails but our proposed model succeeds.

puts for the native text representations and the translated text representations in Figure 2, for both English and Chinese. Since we use W-GAN (Arjovsky et al., 2017) for adversarial training, the discriminator output for an input sentence is a single scalar. For each language, we plot the distribution of difference in the output scalars $D_{s,t}(\boldsymbol{h}^n_{s,t}) - D(\boldsymbol{h}^m_{s,t})$ between the native test corpus and the translated test corpus. These difference values are closer to 0 if the model fools the discriminators. For comparison we trained additional discriminators for the `Trans` baseline as the `w/o ADV` curves on the plot. The adversarial training makes the difference between the native text and the translated text much smaller for both English and Chinese.

Apart from the quantitative analysis, we show an example error from the baseline model that our proposed framework with adversarial training has managed to avoid in Table 5. The model makes the wrong prediction because in the English training data, "support"(支持) can trigger a `Transfer-Money` event with certain context which is uncommon in Chinese. By aligning the representation spaces with adversarial training, the model will align 支持 in translated text to representations of more common used Chinese words that trigger the `Transfer-Money` event.

3.5 Remaining Challenges

Chinese Sentences	Error
40年来，日本皇室就没有再添男丁。 (For 40 years, the Japanese royal family has not added any more males.)	Misses the trigger 添(add), `Be-Born`
德仁皇太子唯一的弟弟，是[皇室]$_{\text{entity}}$最后一名[出生]$_{\text{trigger}}$的男性 (the only brother of Prince Naruhito was the last male [born]$_{\text{trigger}}$ in the [Royal Family]$_{\text{entity}}$.)	False positive role prediction:`Place`.

Table 6: Remaining error examples of cross-lingual trigger and argument role labeling from our proposed model. We provide Chinese test sentences and English translations on the left and errors on the right.

Our experiments show cross-lingual trigger labeling from English to Chinese is very challenging. In Table 6, the first two examples are from the trigger labeling task. In the first example, the Chinese trigger span has the meaning of "add," which can only trigger a `Born` event under specific context such as "add children." However, this is not a typical English expression, and it appears very rarely in the ACE 2005 English training data. Therefore cross-lingual learning fails on this case.

The second example is from the argument role labeling task. The model makes the wrong prediction because "室" in the entity span has the meaning of "room," making the model to consider the entity as a location. Joint learning of entity typing and role labeling can be helpful for such cases.

4 Related Work

Multilingual Language Representations . Early work on multilingual representations learns aligned word or sentence embeddings from dictionaries (Mikolov et al., 2013; Faruqui and Dyer, 2014; Pan et al., 2019), parallel corpora (Gouws et al., 2015; Luong et al., 2015) or semi-supervised or unsupervised approaches (Artetxe et al., 2017; Zhang et al., 2017; Artetxe et al., 2018; Lample et al., 2018). Recent advances in pretrained language models have inspired research on cross-lingual language models such as mBERT (Devlin et al., 2019), XLM (Conneau and Lample, 2019) and XLM-RoBERTa (Conneau et al., 2020).

Cross-Lingual Learning for NLP There is research in cross-lingual learning for many NLP tasks such as name tagging (Huang et al., 2019), reading comprehension (Cui et al., 2019; Hsu et al., 2019), summarization[4] (Zhu et al., 2019; Cao et al., 2020). XGLUE (Liang et al., 2020), XTREME (Hu et al., 2020a) and XTREME-R (Ruder et al., 2021) present benchmarks covering a wide range of tasks including natural language inference, paraphrase detection, part-of-speech tagging, name tagging, question answering, sentence retrieval and generation, which are followed by (Phang et al., 2020; Fang et al., 2021; Luo et al., 2020; Wei et al., 2021; Ouyang et al., 2021). However these benchmarks don't include event extraction as a subtask. For cross-lingual event extraction, early work utilizes multilingual embeddings and language universal parsing structures for cross-lingual transfer for trigger labeling (Liu et al., 2019) and argument role labeling (Subburathinam et al., 2019). It is worth mentioning that Liu et al. (2019) focus on augmenting the existing supervision in the target language with cross-lingual learning that is different from the setting in this work, which requires no supervision in the target language. M'hamdi et al. (2019) explore using mBERT (Devlin et al., 2019) for direct cross-lingual trigger labeling and find it outperforms previous methods. Our `Direct` baseline can be considered as a re-implementation of their method with XLM-RoBERTa (Conneau et al., 2020). GATE (Ahmad et al., 2021) follows (Subburathinam et al., 2019) and uses a graph convolutional architecture and pretrained knowledge from language models to further improve the performance. Yarmohammadi et al. (2021) first translate the whole sentence and then uses token aligners to get a sub-sentential alignment, which has shown to be beneficial. We use a different translation strategy, and our proposed adversarial training approach may also be helpful with their translations. A more recent and parallel attempt (Guzman-Nateras et al., 2022) proposes to use adversarial training to close the gap between the source language and target language for event trigger labeling, which is different from our approach. (Fincke et al., 2022) uses priming methods to make the model understand the critical information for argument labeling. The performance of these two methods is not directly comparable due to different splits and limited code availability. We will add comparison once they release code. (Huang et al., 2022) proposes a generative approach to directly generate arguments for cross-lingual event argument extraction. However they don't take entity spans as inputs for evaluation and results are not comparable.

5 Conclusions and Future Work

In this paper, we proposed a new cross-lingual event extraction framework and evaluated the framework on the ACE 2005 dataset. Our framework combines the multilingual language models with a machine-translation-based method. Meanwhile, we observe the representational gap between the translated text and the native text in multilingual language models that may affect the performance and propose an adversarial training approach to make the language model produce more similar representations for these two types of text.

One potential reason for remaining errors in cross-lingual transfer learning could be that the

[4] Cross-lingual summarization has a different task formulation than common cross-lingual learning, but it is still related.

source and the target languages may differ in the common expressions of an event type. It will be helpful to detect such differences from pretrained multilingual language models and incorporate them for training. Although we focus on cross-lingual event extraction in this work, our adversarial training approach could be extended to other cross-lingual language understanding tasks.

6 Limitations

Although we have demonstrated our framework's performance in six cross lingual transfer learning directions for both the trigger labeling and argument role labeling, our experiments is mostly on the ACE 2005 dataset due to the availability of multilingual event extraction data. Since the ACE 2005 dataset only contains Arabic, Chinese and English, we were not able to test our framework on some languages with extremely limited resources, which are more common use cases for the cross lingual transfer learning .Besides, although our proposed adversarial loss is a general approach not specific to the event extraction task, we have not validate the effectiveness of it on other cross lingual NLP benchmarks or using other machine translation models. Moreover, our supervised models are trained in the multilingual language model (XLM-RoBERTa) for direct comparison. However, the performance is different from models trained with monolingual language models specific to the target language.

7 Acknowledgement

This research is based upon work supported in part by U.S. DARPA LORELEI Program No.HR0011-15-C-0115, U.S. DARPA AIDA Program No. FA8750-18-2-0014 and KAIROS Program No. FA8750-19-2-1004. 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 DARPA, 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 *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 12462–12470. AAAI Press.

Martín Arjovsky, Soumith Chintala, and Léon Bottou. 2017. Wasserstein GAN. *CoRR*, abs/1701.07875.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2017. Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 451–462, Vancouver, Canada. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 789–798, Melbourne, Australia. Association for Computational Linguistics.

Yue Cao, Hui Liu, and Xiaojun Wan. 2020. Jointly learning to align and summarize for neural cross-lingual summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6220–6231, Online. 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. 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.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 7057–7067.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Mach. Learn.*, 20(3):273–297.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2019. Cross-lingual machine reading comprehension. 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 1586–1595, Hong Kong, China. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Yuwei Fang, Shuohang Wang, Zhe Gan, Siqi Sun, and Jingjing Liu. 2021. FILTER: an enhanced fusion method for cross-lingual language understanding. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 12776–12784. AAAI Press.

Manaal Faruqui and Chris Dyer. 2014. Improving vector space word representations using multilingual correlation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 462–471, Gothenburg, Sweden. Association for Computational Linguistics.

Steven Fincke, Shantanu Agarwal, Scott Miller, and Elizabeth Boschee. 2022. Language model priming for cross-lingual event extraction. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 10627–10635. AAAI Press.

Stephan Gouws, Yoshua Bengio, and Greg Corrado. 2015. Bilbowa: Fast bilingual distributed representations without word alignments. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 748–756. JMLR.org.

Ishaan Gulrajani, Faruk Ahmed, Martín Arjovsky, Vincent Dumoulin, and Aaron C. Courville. 2017. Improved training of wasserstein gans. *CoRR*, abs/1704.00028.

Luis F Guzman-Nateras, Minh Van Nguyen, and Thien Huu Nguyen. 2022. Cross-lingual event detection via optimized adversarial training. In *Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, Seattle, Washington, USA. Association for Computational Linguistics.

Tsung-Yuan Hsu, Chi-Liang Liu, and Hung-yi Lee. 2019. Zero-shot reading comprehension by cross-lingual transfer learning with multi-lingual language representation model. 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 5933–5940, Hong Kong, China. Association for Computational Linguistics.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020a. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 4411–4421. PMLR.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020b. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4411–4421. PMLR.

Kuan-Hao Huang, I-Hung Hsu, Prem Natarajan, Kai-Wei Chang, and Nanyun Peng. 2022. Multilingual generative language models for zero-shot cross-lingual event argument extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4633–4646, Dublin, Ireland. Association for Computational Linguistics.

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.

Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, Xiaodong Fan, Bruce Zhang, Rahul Agrawal, Edward Cui, Sining Wei, Taroon Bharti, Ying Qiao, Jiun-Hung Chen, Winnie Wu, Shuguang Liu, Fan Yang, Rangan Majumder, and Ming Zhou. 2020. XGLUE: A new benchmark dataset for cross-lingual pre-training, understanding and generation. *CoRR*, abs/2004.01401.

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*, pages 7999–8009, Online. Association for Computational Linguistics.

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.

Fuli Luo, Wei Wang, Jiahao Liu, Yijia Liu, Bin Bi, Songfang Huang, Fei Huang, and Luo Si. 2020. VECO: variable encoder-decoder pre-training for cross-lingual understanding and generation. *CoRR*, abs/2010.16046.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Bilingual word representations with monolingual quality in mind. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pages 151–159, Denver, Colorado. 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.

Tomás Mikolov, Quoc V. Le, and Ilya Sutskever. 2013. Exploiting similarities among languages for machine translation. *CoRR*, abs/1309.4168.

Xuan Ouyang, Shuohuan Wang, Chao Pang, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2021. ERNIE-M: Enhanced multilingual representation by aligning cross-lingual semantics with monolingual corpora. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 27–38, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xiaoman Pan, Thamme Gowda, Heng Ji, Jonathan May, and Scott Miller. 2019. Cross-lingual joint entity and word embedding to improve entity linking and parallel sentence mining. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, pages 56–66, Hong Kong, China. Association for Computational Linguistics.

Jason Phang, Iacer Calixto, Phu Mon Htut, Yada Pruksachatkun, Haokun Liu, Clara Vania, Katharina Kann, and Samuel R. Bowman. 2020. English intermediate-task training improves zero-shot cross-lingual transfer too. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 557–575, Suzhou, China. Association for Computational Linguistics.

Daria Pylypenko, Kwabena Amponsah-Kaakyire, Koel Dutta Chowdhury, Josef van Genabith, and Cristina España-Bonet. 2021. Comparing feature-engineering and feature-learning approaches for multilingual translationese classification. *CoRR*, abs/2109.07604.

Parker Riley, Isaac Caswell, Markus Freitag, and David Grangier. 2020. Translationese as a language in "multilingual" NMT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7737–7746. Association for Computational Linguistics.

Sebastian Ruder, Noah Constant, Jan Botha, Aditya Siddhant, Orhan Firat, Jinlan Fu, Pengfei Liu, Junjie Hu, Dan Garrette, Graham Neubig, and Melvin Johnson. 2021. XTREME-R: towards more challenging and nuanced multilingual evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 10215–10245. Association for Computational Linguistics.

Sebastian Ruder, Ivan Vulic, and Anders Søgaard. 2019. A survey of cross-lingual word embedding models. *J. Artif. Intell. Res.*, 65:569–631.

Ananya Subburathinam, Di Lu, Heng Ji, Jonathan May, Shih-Fu Chang, Avirup Sil, and Clare Voss. 2019. Cross-lingual structure transfer for relation and event extraction. 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 313–325, Hong Kong, China. Association for Computational Linguistics.

Xiangpeng Wei, Rongxiang Weng, Yue Hu, Luxi Xing, Heng Yu, and Weihua Luo. 2021. On learning universal representations across languages. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Mahsa Yarmohammadi, Shijie Wu, Marc Marone, Haoran Xu, Seth Ebner, Guanghui Qin, Yunmo Chen, Jialiang Guo, Craig Harman, Kenton 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 *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1950–1967, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Meng Zhang, Yang Liu, Huanbo Luan, and Maosong Sun. 2017. Earth mover's distance minimization for unsupervised bilingual lexicon induction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1934–1945, Copenhagen, Denmark. Association for Computational Linguistics.

Junnan Zhu, Qian Wang, Yining Wang, Yu Zhou, Jiajun Zhang, Shaonan Wang, and Chengqing Zong. 2019. NCLS: neural cross-lingual summarization. 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 2019, Hong Kong, China, November 3-7, 2019*, pages 3052–3062. Association for Computational Linguistics.

A Appendix

A.1 Details for Model Training

For both the trigger labeling and role labeling task, we use batch size of 8 for training. We evaluate performance after each epoch and select the best model based on the development performance. We use early-stop strategy with a patience of 5 epochs. We conduct our experiments on a single Nvidia Tesla V100 GPU with 16GB memory.

The learning rate for both the trigger labeling and role labeling loss is $1e-5$. In adversarial training, the learning rate for the discriminator loss is $1e-5$. For the generator loss, we found in practice it is very likely to confuse the discriminators within a few steps if we finetune the whole XLM-RoBERTa architecture or the learning rate is set too large. Hence the generator learning rate for the generator loss is chosen between $\{1e-5, 1e-6, 1e-7, 1e-8, 1e-9\}$ on the dev set for each cross lingual transfer learning task. We empirically found that the trigger labeling tasks usually take a smaller learning rate $(1e-8, 1e-9)$ and the argument role labeling tasks usually take a larger one $(1e-5, 1e-6)$. We also only finetune the last output layer of the XLM-RoBERTa model for the generator loss to match the capacity of the discriminators. The discriminator and the generator are trained alternatively. We train 5 discriminator steps per generator step.

For the simultaneous encoding of a sentence and its translation, we adopt the special fusion strategy in FILTER (Fang et al., 2021) for the role labeling task. FILTER will select some hidden layers of the XLM-RoBERTa model, for which it will concatenate the hidden representation of the original sentence and its translation together for self-attention computation. We follow FILTER to use the 21st layer for representation fusion. We found this strategy to be more helpful in role labeling task than trigger labeling task. In trigger labeling task, it suffices to simply encode the sentence pairs individually for prediction.

The approximate number of parameters is 3.5 million (mainly parameters of XLM-RoBERTa). We run our model on a single NVIDA V100 with 16 GB memory. Training our framework takes approximately 20-40 minutes/epoch since 16GB memory can only take batch size of 1 for training. We need to accumulate the gradients over multiple runs for larger batch size. However, we notice that our model usually converges much faster than a simple XLM-RoBERTa baseline (`Direct` baseline). Usually we achieve our best model with 2-4 epochs. In total it usually takes around 4-5 hours to train a model. We implement the XLM-RoBERTa model using Transformers[5] Library.

For the back propagation, note that the gradients of the loss in Equation (1) are back propagated to both the language model and the trigger classification layers, the gradients of the loss in Equation (2) are back propagated to D_s and D_t, and the gradients of the loss in Equation (3) are back propagated to the language model. In practice we found that it is beneficial to back propagate loss in Equation (3) to only the last layer of the FILTER model to match the capacity of the discriminators D_s and D_t.

A.2 Details for Machine Translation

We use Google Cloud API[6] for machine translation. For trigger labeling, if a sentence contains multiple triggers, we enclose each of them with "" and "" for translation. After the sentence is translated, we retrieve all trigger spans in the target language one by one, and map them back to the triggers in the source language according the offset in the sentence. For example, the first trigger span in the source language will be mapped to the first trigger span in the target language. If we retrieve less triggers spans in the target language than the source language, we consider this translation invalid and discard this instance for the trigger labeling loss. We still use it for the adversarial training. For argument role labeling, we directly translate the sentence with inserted "<a>", "","", "" and always apply the role labeling loss on the translated sentence even if it may not contain paired special tokens.

For trigger labeling, our translation method retrieved[7] 4,284 event triggers out of 4,419 triggers in

[5] https://huggingface.co/docs/transformers/index

[6] https://cloud.google.com/translate

[7] Here "retrieved" means that after the translation of a source language sentence of the format in Table 1, the trans-

the ACE 2005 English training data. For argument role labeling, there is no simple automatic metric to evaluate our translation method. Therefore, we sampled a small portion of the translation and conduct a small scale manual evaluation. 80.0% of the translations are considered reasonable by human assessors.

The reason behind this translation strategy is that the machine translation model trained on large-scale web-crawled data could have seen some HTML tags during training. "" are HTML tags for displaying bold characters, and "<a>" are tags for the content of reference links. Therefore we expect the model to translate properly if it can translate HTML formatted text.

A.3 ACE 2005 Dataset Details

This dataset is licensed by LDC.[8] Membership is required for access. The dataset can be used for research purpose.

There are three languages in this dataset. For all the languages, we notice a significant long-tailed distribution among event types. We provide number of event mentions for all splits in Table 7. We also notice that the most frequent types for all languages are similar with minor differences.

A.4 Details of Additional Discriminators for Case Study

For fair comparison of the additional discriminators for the `Trans` baseline and the discriminators in our framework, we also jointly train the the discriminators with the `Trans` baseline in the same way as we conduct adversarial training in our framework. The training process can be seen as training our framework with the generator learning rate being 0. Note that the parameters of the discriminators are disjoint of that of the `Trans` baseline model. Therefore the joint training will not affect the learning the `Trans` baseline model.

A.5 Corruption Ratio of Translated Training Data

We provide corruption ratio for the argument role labeling task here for translation of the training data. Due to our strategy of inserting special tokens, a corrupted translation is defined as a translated sentence without either of the special tokens. In sentences translated into Arabic, we noticed that special tokens are sometimes translated as '<a >' or '' with additional spaces. We don't consider them as corrupted and automatically cleaned up such errors. The corruption ratios are as below: EN-ZH, 10%; EN-AR: 22%; ZH-EN: 12%, ZH-AR: 27%; AR-EN: 26%; AR-ZH: 38%.

It is also worth mentioning that Google translate offers the option to respect HTML mark up. However, we didn't adopt this option in our experiments. We believe enabling this function can further reduce the corruption ratio and potentially improve the performance.

A.6 Full Results

We present full results of all six cross-lingual transfer settings across two tasks, including the precision, recall and f1 scores. We include trigger labeling performance in Table 8a-8f. We include role labeling performance in Table 9a-9c.

lated sentence include paired "" and "" tokens and the content between them are not empty. In this sense retrieved triggers are not guaranteed to be correct annotations. This is just a rough estimation of the performance of proposed translation method.

[8] https://www.ldc.upenn.edu

Split	English			Chinese			Arabic		
	train	dev	test	train	dev	test	train	dev	test
Conflict:Attack	1,272	172	93	470	37	17	377	45	55
Movement:Transport	611	59	48	662	54	43	354	46	34
Life:Die	524	53	17	211	18	14	177	33	34
Contact:Meet	200	29	50	163	19	26	152	38	27
Personnel:Elect	162	4	16	28	1	9	31	6	4
Personnel:End-Position	159	19	22	71	5	11	37	14	7
Transaction:Transfer-Money	128	52	14	84	3	5	34	11	3
Life:Injure	127	9	1	149	7	7	92	14	21
Contact:Phone-Write	112	3	8	77	8	2	45	3	8
Justice:Trial-Hearing	103	1	5	79	4	8	58	1	6
Justice:Charge-Indict	96	2	8	50	0	2	45	2	5
Transaction:Transfer-Ownership	92	4	30	84	2	1	9	0	1
Personnel:Start-Position	92	12	13	95	5	2	36	10	0
Justice:Sentence	84	4	11	79	4	7	46	1	4
Justice:Arrest-Jail	78	4	6	115	11	6	82	13	14
Life:Marry	73	0	10	55	0	2	9	7	0
Conflict:Demonstrate	65	9	7	72	3	1	55	8	10
Justice:Convict	64	6	6	13	3	0	3	1	1
Justice:Sue	60	12	4	76	0	3	2	0	0
Life:Be-Born	47	0	3	22	0	6	6	0	0
Justice:Release-Parole	46	0	1	31	5	2	18	6	7
Business:Declare-Bankruptcy	40	1	2	15	0	4	1	0	0
Business:End-Org	31	1	5	16	0	2	6	1	1
Justice:Appeal	30	7	6	35	0	0	12	0	7
Business:Start-Org	29	0	18	77	2	5	12	0	2
Justice:Fine	22	0	6	7	4	2	33	0	0
Life:Divorce	20	0	9	11	0	0	3	2	0
Business:Merge-Org	14	0	0	36	16	1	1	0	0
Justice:Execute	14	5	2	5	0	1	0	0	0
Personnel:Nominate	11	0	1	24	0	1	4	0	3
Justice:Extradite	6	0	1	2	2	0	7	0	0
Justice:Acquit	5	0	1	3	0	0	3	0	0
Justice:Pardon	2	0	0	9	4	0	1	0	1

Table 7: Event type distribution for the event trigger labeling task

Trigger Labeling	P(%)	R(%)	F(%)
Direct	42.3	52.5	46.9
Trans	39.9	58.1	47.3
Trans+Adv (ours)	**42.5**	**58.7**	**49.3**
ZH Supervision	65.2	65.9	65.6

(a) English-to-Chinese.

Trigger Labeling	P(%)	R(%)	F(%)
Direct	32.0	**50.0**	39.0
Trans	33.1	48.1	39.3
Trans+Adv (ours)	**38.1**	47.7	**42.3**
AR Supervision	49.4	64.9	56.1

(b) English-to-Arabic.

Trigger Labeling	P(%)	R(%)	F(%)
Direct	50.6	39.6	44.4
Trans	56.0	39.4	46.3
Trans+Adv (ours)	**63.2**	**48.1**	**54.6**
EN Supervision	63.0	75.0	68.5

(c) Chinese-to-English.

Trigger Labeling	P(%)	R(%)	F(%)
Direct	43.1	**33.3**	39.8
Trans	**57.4**	29.9	39.4
Trans+Adv (ours)	56.0	33.0	**41.5**
EN Supervision	63.0	75.0	68.5

(d) Arabic-to-English.

Trigger Labeling	P(%)	R(%)	F(%)
Direct	30.3	37.3	33.4
Trans	34.5	44.3	38.8
Trans+Adv (ours)	**36.1**	**45.1**	**40.1**
AR Supervision	49.4	64.9	56.1

(e) Chinese-to-Arabic.

Trigger Labeling	P(%)	R(%)	F(%)
Direct	35.3	**38.3**	36.7
Trans	37.6	35.6	36.6
Trans+Adv (ours)	**49.6**	31.4	**38.4**
ZH Supervision	65.2	65.9	65.6

(f) Arabic-to-Chinese.

Table 8: Precision(P), recall(R) and f1(F) scores for the cross-lingual trigger labeling task. *Direct*,*Trans* and *Target Supervision* are introduced in Section 3.2.

Argument Role Labeling	Chinese-to-English			Chinese-to-Arabic		
	Precision(%)	Recall(%)	F1(%)	Precision(%)	Recall(%)	F1(%)
GATE	48.0	**70.0**	57.0	64.1	**66.1**	65.1
Direct	**59.7**	63.4	61.5	68.2	60.3	64.0
Trans	56.6	65.0	60.6	67.9	60.1	63.8
Trans+Adv (ours)	59.1	67.3	**62.9**	**72.4**	64.4	**68.0**
Target Supervision	75.1	79.5	77.2	77.5	78.1	77.8

(a) Chinese as the source language.

Argument Role Labeling	English-to-Chinese			English-to-Arabic		
	Precision(%)	Recall(%)	F1(%)	Precision(%)	Recall(%)	F1(%)
GATE	60.7	66.8	63.6	72.5	**58.9**	65.0
Direct	72.6	70.8	71.7	**81.5**	50.7	62.5
Trans	**73.0**	69.7	71.3	76.3	52.5	62.2
Trans+Adv (ours)	72.2	**71.8**	**72.0**	76.0	57.0	**65.1**
Target Supervision	79.7	84.4	82.0	77.5	78.1	77.8

(b) English as the source language.

Argument Role Labeling	Arabic-to-English			Arabic-to-Chinese		
	Precision(%)	Recall(%)	F1(%)	Precision(%)	Recall(%)	F1(%)
GATE	40.4	70.5	50.3	44.7	**74.1**	55.7
Direct	50.5	64.8	56.8	60.7	69.0	64.6
Trans	50.6	**66.6**	57.5	62.2	67.8	64.9
Trans+Adv (ours)	**54.1**	63.4	**58.4**	**64.1**	67.1	**65.6**
Target Supervision	75.1	79.5	77.2	79.7	84.4	82.0

(c) Arabic as the source language.

Table 9: Precision(P), recall(R) and f1(F) scores for the cross-lingual argument role labeling task. *GATE* (Ahmad et al., 2021) is a state-of-the-art method for cross-lingual argument role labeling. *Direct*,*Trans* and *Target Supervision* are introduced in Section 3.2.