Learning Cross-lingual Representations for Event Coreference Resolution with Multi-view Alignment and Optimal Transport

Duy Phung[1], **Hieu Minh Tran**[1], **Minh Van Nguyen**[2], **and Thien Huu Nguyen**[2]
[1] VinAI Research, Vietnam
[2] Dept. of Computer and Information Science, University of Oregon, Eugene, OR, USA
{v.duypv1,v.hieutm4}@vinai.io, {minhnv,thien}@cs.uoregon.edu

Abstract

We study a new problem of cross-lingual transfer learning for event coreference resolution (ECR) where models trained on data from a source language are adapted for evaluations in different target languages. We introduce the first baseline model for this task based on XLM-RoBERTa, a state-of-the-art multilingual pre-trained language model. We also explore language adversarial neural networks (LANN) that present language discriminators to distinguish texts from the source and target languages to improve the language generalization for ECR. In addition, we introduce two novel mechanisms to further enhance the general representation learning of LANN, featuring: (i) multi-view alignment to penalize cross coreference-label alignment of examples in the source and target languages, and (ii) optimal transport to select close examples in the source and target languages to provide better training signals for the language discriminators. Finally, we perform extensive experiments for cross-lingual ECR from English to Spanish and Chinese to demonstrate the effectiveness of the proposed methods.

1 Introduction

Event coreference resolution (ECR) aims to link event-trigger expressions (event mentions) in a document that refer to the same event in real world. Technically, the core problem in ECR involves predicting if two event mentions in a document corefer to each other or not (i.e., a binary classification problem). For instance, consider the following text:

*With national outrage boiling over, Bangladeshi paramilitary officers tracked down and **arrested** Sohel Rana. When loudspeakers at the rescue site announced his **capture**, local news reports said, the crowd broke out in cheers.*

An ECR system in information extraction (IE) should be able to recognize the coreference of the two event mentions associated with the trigger words "*arrested*" and "*capture*" in this text.

Prior work on ECR assumes the monolingual setting where training and test data are presented in the same languages. Current state-of-the-art ECR systems thus rely on large monolingual datasets to train advanced models (Nguyen et al., 2016; Choubey and Huang, 2018; Lu and Ng, 2017, 2018; Huang et al., 2019) that are only annotated for popular languages (e.g., English). As document annotation for ECR is an expensive process, porting ECR models for English to other languages is crucial and appealing to enhance the accessibility of ECR systems. To this end, this paper explores cross-lingual transfer learning for ECR where models are trained on annotated documents in English (source language) and tested on documents from other languages (target languages). To be clear, our work considers zero-resource cross-lingual learning that requires no labeled data for ECR in the target languages as well as human or machine generated parallel text. The systems in this work only have access to unlabeled text in the target languages to aid the cross-lingual learning for ECR. To our knowledge, this is the first work on cross-lingual transfer learning for event coreference resolution in the literature.

Recent advances in contextualized word embeddings have featured multilingual pre-trained language models, e.g., multilingual BERT (Devlin et al., 2019), XLM-RoBERTa (Conneau et al., 2019), that overcome the vocabulary difference of languages and produce language-universal representations for cross-lingual transfer learning in different NLP tasks (Wu and Dredze, 2019; Subburathinam et al., 2019a). In fact, such pre-trained language models have set a new standard for multilingual learning in NLP (Wu and Dredze, 2020; Nguyen et al., 2021a), serving as the baseline models for our cross-lingual transfer learning problem for ECR in this work.

How can we improve the cross-lingual performance of ECR models over multilingual language

Proceedings of the 1st Workshop on Multilingual Representation Learning, pages 62–73

model baselines? Treating the source and target languages as the source and target domains in domain adaptation (Chen et al., 2018a, 2019; Keung et al., 2019), one can borrow the popular technique of domain adversarial neural networks (DANN) (Ganin et al., 2016; Fu et al., 2017) to induce better language-general representations for ECR, called language adversarial neural networks (LANN) to make it consistent with our language generalization problem. As such, in addition to traditional learning objectives (e.g., cross-entropy), the key idea of LANN is to introduce a language discriminator that seeks to differentiate representation vectors for text inputs from the source and target languages. To enhance the language generalization, models will attempt to generate representation vectors so the language discriminator is fooled, i.e., its performance is minimized to align the source and target languages (Chen et al., 2018a; Keung et al., 2019). However, there are two major limitations with LANN that will be addressed to improve the cross-lingual performance for ECR models in this work.

First, taking the binary classification setting for ECR, inputs to the language discriminator in LANN involve two pairs of event mentions in the source and target languages. As coreference labels for pairs of event mentions in target languages are not available, the language discriminator will thus only aim to align marginal distributions of event mention pairs (called examples) in the source and target languages (without considering the coreference labels for the pairs). This is less optimal as the lack of coreference labels in the alignment might unexpectedly cause coreferring examples in the source language to be mapped or aligned with non-coreferring examples in the target languages and vice versa, thus impairing the discriminative nature of representation vectors for ECR. To overcome this issue, we propose to use two network architectures to obtain two complementary representation vectors for each example in both source and target languages. Representation vectors from each network will be first aligned between source and target languages using the usual LANN technique. In addition, representation vectors from the two networks will be regularized to agree with each other over same examples in target languages. As demonstrated later, this regularization helps to penalize the alignment between coreferring examples in the source language and non-coreferring examples in the target languages (and vice versa) in LANN, thus improving the representation quality.

Second, as LANN attempts to discriminate all examples in the source language from all examples in the target languages, it also employs training signals from examples whose representations are far away from each other in the source and target languages. However, it is intuitive that the most useful information for model training comes from close examples in the source and target language spaces. Including long-distance examples might even introduce noise and hurt the models' performance. Consequently, instead of using all examples for LANN, we propose to only leverage examples with close representation vectors for the language discriminator in ECR models. As such, our approach involves measuring distances between representation vectors of examples in the source and target languages to determine which examples are used for the language discriminator. To access the distance between two examples in the source and target languages, instead of only relying on the similarity of learned representations, we propose to additionally consider coreference likelihoods of examples that assign higher similarity if two examples have similar coreference likelihoods (i.e., examples with the same coreference labels are more similar to each other than others in ECR). Accordingly, our model employs Optimal Transport, a method to determine the cheapest transformation between two data distributions, as a natural solution to simultaneously incorporate both representation vectors and coreference likelihoods of examples into the distance estimation for example selection in the language discriminator of LANN. We conduct cross-lingual ECR evaluation for English, Spanish and Chinese that demonstrates the benefits of the proposed methods by significantly outperforming the baseline models. We will release experiment setups and code to push forward research on cross-lingual ECR in the future.

2 Related Work

Regarding coreference resolution, our work is related to studies in entity coreference resolution that aim to resolve nouns phrases/mentions for entities (Raghunathan et al., 2010; Ng, 2010; Durrett and Klein, 2013; Lee et al., 2017; Joshi et al., 2019). This work focuses on event coreference resolution that is often considered as a more challenging task than entity resolution due to the more complex

structures of event mentions (Yang et al., 2015).

For event coreference resolution, although there have been works on cross-document resolution (Lee et al., 2012a; Kenyon-Dean et al., 2018; Barhom et al., 2019; Phung et al., 2021), this work is more related to prior work on within-document ECR (Lu and Ng, 2018; Tran et al., 2021). In particular, previous within-document ECR methods have applied feature-based models for pairwise classifiers (Ahn, 2006; Chen et al., 2009; Cybulska and Vossen, 2015; Peng et al., 2016), spectral graph clustering (Chen and Ji, 2009b), information propagation (Liu et al., 2014), markov logic networks (Lu et al., 2016), end-to-end modeling with event detection (Araki and Mitamura, 2015; Lu et al., 2016; Chen and Ng, 2016; Lu and Ng, 2017), and recent deep learning models (Nguyen et al., 2016; Choubey and Huang, 2018; Huang et al., 2019; Choubey et al., 2020; Tran et al., 2021). Our work is different from such prior work as we investigate a novel setting of cross-lingual transfer learning for ECR.

Cross-lingual transfer learning has been studied for other NLP and IE tasks, including sentiment analysis (Chen et al., 2018b), relation extraction (Lin et al., 2017; Zou et al., 2018; Wang et al., 2018; Nguyen and Nguyen, 2021), event extraction (Chen and Ji, 2009a; Hsi et al., 2016; Subburathinam et al., 2019b; Nguyen et al., 2021b), and entity coreference resolution (Rahman and Ng, 2012; Hardmeier et al., 2013; Martins, 2015; Kundu et al., 2018; Urbizu et al., 2019). Compared to such prior work, this paper presents two novel approaches to improve the language generalization of representation vectors based on multi-view alignment and OT. Finally, our work involves LANN that bears some similarity with DANN models in domain adaptation research of machine learning (Ganin et al., 2016; Bousmalis et al., 2016; Fu et al., 2017; Kumar et al., 2018; Naik and Rose, 2020; Ngo et al., 2021). Compared to such work, our work explores a new dimension of adversarial networks for language-invariant representation learning for texts in ECR.

3 Model

We formalize our ECR problem using a pairwise approach (Lu and Ng, 2018; Choubey and Huang, 2018; Barhom et al., 2019). Let $W = w_1, w_2, \ldots, w_n$ be a document (with n words) that contains two input event mentions with event triggers located at w_{e_1} and w_{e_2} in W ($1 \leq e_1 < e_2 \leq n$). As such, the core problem in ECR is to perform a binary prediction to determine whether two event mentions w_{e_1} and w_{e_2} refer to the same event or not. An example in our ECR task thus involves an input tuple $X = (W, e_1, e_2)$ and a binary output variable y to indicate the coreference of w_{e_1} and w_{e_2}. This work focuses on cross-lingual transfer learning for ECR where training data involve input documents W in English (the source language) while sentences in test data are presented in another language (the target language). To enable the zero-resource cross-lingual setting for ECR, our model takes two following inputs: $\mathcal{D}^{src} = \{(X_i = (W^i, e_1^i, e_2^i), y_i)\}_{i=1..N_{src}}$ as the training set with N_{src} labeled examples in the source language (English), and $\mathcal{D}^{tar} = \{X_i = (W^i, e_1^i, e_2^i)\}_{i=N_{src}+1..N_{src}+N_{tar}}$ as the unlabeled set in the target language with N_{tar} examples.

3.1 Baseline Model

As this is the first work on cross-lingual transfer learning for ECR, this section aims to establish a baseline method for further research. In particular, recent work has shown that multilingual pre-trained language models with deep stacks of transformer layers, e.g., multilingual BERT (Devlin et al., 2019), XLM-RoBERTa (Conneau et al., 2019), can provide strong baselines with competitive performance for zero-shot cross-lingual transfer for a variety of NLP tasks (Wu and Dredze, 2019). As such, we utilize XLM-RoBERTa[1] to obtain language-general representation vectors for a cross-lingual baseline model of ECR in this work. Given the input document and event mentions $X = (W, e_1, e_2)$ (in the source or target language), we first prepend the special token `[CLS]`, and insert two special tokens `<e>` and `</e>` right before and after the trigger words w_{e_1} and w_{e_2} in W to mark their positions, leading to a new document $W' =$ `[CLS]` $w_1 \ldots w_{e_1-1}$`<e>` w_{e_1}`</e>` $w_{e_1+1} \ldots w_{e_2-1}$`<e>` w_{e_2}`</e>` $w_{e_2+1} \ldots w_n$. Afterward W' is fed into the base version of XLM-RoBERTa to obtain hidden vectors for their word-pieces. Let h_{cls} be the hidden vector for the special token `[CLS]`, h_s^1 and h_e^1 be the hidden vectors for the special tokens `<e>` and `</e>` surrounding w_{e_1}, and h_s^2 and h_e^2 be the hidden vectors for the special tokens `<e>` and `</e>` surrounding w_{e_2}

[1] XLM-RoBERTa is chosen due to its better performance than multilingual BERT in our experiments.

in W. Note that as the number of word-pieces in W might exceed the maximum length of 512 in XLM-RoBERTa, we divide the word-piece sequence for W into chunks of lengths equal to or smaller than 512; these chunks are then processed separately by XLM-RoBERTa. In the next step, an overall representation vector $V(X)$ for X, i.e., $V(X) = [h_{cls}, h_s^1, h_e^1, h_s^2, h_e^2]$, is formed and sent into an one-layer feed-forward network $\texttt{FF}$ with softmax in the end to compute a distribution $P(.|X) = \texttt{FF}(V(X))$ over possible coreference labels (i.e., two possible labels) for the input X. Finally, the negative log-likelihood function $\mathcal{L}_{pred}$ over labeled examples in the source language $\mathcal{D}^{src}$ is employed to train the baseline model in this work: $\mathcal{L}_{pred} = -\sum_{i=1}^{N_{src}} \log P(y_i|X_i)$.

In the test time, we use the trained model to predict coreference labels for every pair of event mentions in a document. We then form a graph for each document where event mentions serve as the nodes and two event mentions are connected if their coreference label is positive. As such, connected components in this graph will be returned as event mention clusters for the document in ECR.

3.2 Language Adversarial Networks

To further improve the language generalization for the baseline, we explore the adaptation of domain adversarial neural networks (DANN) in domain adaptation (Ganin et al., 2016) for zero-resource cross-lingual learning (i.e., treating source and target languages as source and target domains). In language adversarial neural networks (LANN), a language discriminator D is introduced to discriminate examples from the source and target languages. As such, the overall representation vector $V(X)$ for each input example X is sent into the language discriminator D (i.e., a two-layer feed-forward network with the sigmoid function in the end) to obtain a scalar score $D(V(X))$ to indicate whether X belongs to the source language or not. The discriminator loss $\mathcal{L}_{disc}$ is then computed over both source and target language data (i.e., $\mathcal{D}^{src}$ and $\mathcal{D}^{tar}$):

$$\mathcal{L}_{disc} = \sum_{i=1}^{N_{src}+N_{tar}} -l_i \log D(V(X_i)) - (1-l_i)\log(1-D(V(X_i))) \quad (1)$$

where l_i is the language indicator (i.e., $l_i = 1$ if $1 \leq i \leq N_{src}$; and 0 otherwise). The overall loss to train the model in this case is thus: $\mathcal{L} = \mathcal{L}_{pred} + \alpha\mathcal{L}_{disc}$ where α is a trade-off parameter. Note that as LANN aims to prevent the language discriminator from recognizing languages from input representation vectors, we insert the Gradient Reversal Layer (GRL) (Ganin et al., 2016) between $V(X)$ and D to reverse the gradients during the backward pass from $\mathcal{L}_{disc}$. Overall, fooling the language discriminator in LANN with GRL helps eliminate language-specific features to improve generalization across languages.

3.3 Multi-view Alignment

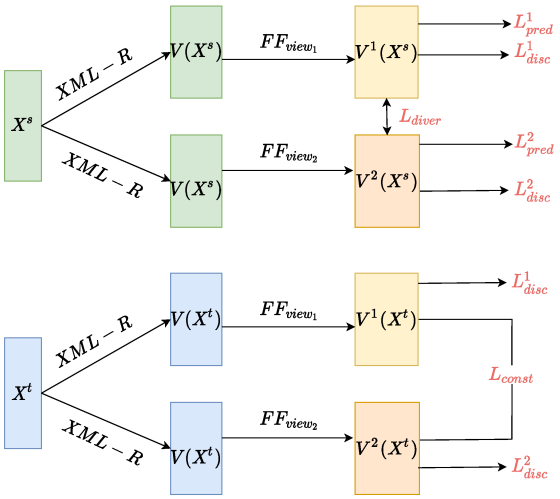

Figure 1: Multi-view alignment mechanism.

One limitation of LANN is that it only attempts to align marginal distributions of examples for ECR in the source and target languages (due to the lack of coreference labels for target examples), causing the unexpected cross-language alignment of coreferential and non-coreferential examples between two languages. To address this issue, instead of relying on one representation vector $V(X)$ for X, we propose to obtain two complementary representation vectors $V^1(X)$ and $V^2(X)$ for X (two views) by sending $V(X)$ into two feed-forward networks with two layers f_1 and f_2: $V^1(X) = f_1(V(X))$ and $V^2(X) = f_2(V(X))$ (preserving the dimensionality of $V(X)$). Afterward, several loss and regularization terms are proposed to penalize the alignment of coreferential and non-coreferential examples across languages in LANN as follows.

First, to ensure that representation vectors $V^1(X)$ and $V^2(X)$ include discriminative information for coreference prediction, we predict the coreference label y_i from both vectors using two feed-forward networks (one layer) with softmax in the end $\texttt{FF}_1$ and $\texttt{FF}_2$ to obtain distributions $P^1(.|X) = \texttt{FF}_1(V^1(X))$ and

$P^2(.|X) = \texttt{FF}_2(V^2(X))$. Negative log-likelihood loss functions $\mathcal{L}^i_{pred}$ from $P^i(.|X)$ $(i = 1, 2)$ are then utilized for the loss function: $\mathcal{L}^i_{pred} = -\sum_{i=1}^{N_{src}} \log P^i(y_i|X_i)$.

Second, representation vectors of source-language examples from each view ($V^1(X)$ or $V^2(X)$) are also aligned their counterparts in the target language based on LANN with language discriminators, i.e., D^1 or D^2 respectively (two-layer feed-forward networks). As such, the discriminator loss $\mathcal{L}^k_{disc}$ for the view $V^k(X)$ $(k = 1, 2)$ is:

$$\begin{aligned}\mathcal{L}^i_{disc} = \sum_{i=1}^{N_{src}+N_{tar}} & -l_i \log D(V^k(X_i)) \\ & - (1 - l_i) \log(1 - D(V^k(X_i)))\end{aligned} \tag{2}$$

Third, to encourage the diversity or complementary nature of the information captured by two views V^1 and V^2, we seek to increase the difference between representation vectors $V^1(X)$ and $V^2(X)$ over the same source-language examples X in $\mathcal{D}^{src}$ by including their negative distance $\mathcal{L}_{diver}$ into the overall loss function:

$$\mathcal{L}_{diver} = -\frac{1}{N_{src}} \sum_{i=1}^{N_{src}} ||V^1(X) - V^2(X)||_2^2 \tag{3}$$

Fourth, representation vectors from two views $V^1(X)$ and $V^2(X)$ will be constrained to be consistent with each other for the same examples $X \in \mathcal{D}^{tar}$ in the target language. This is done by introducing the difference $\mathcal{L}_{const}$ between $V^1(X)$ and $V^2(X)$ over target-language examples in $\mathcal{D}^{tar}$ into the overall loss function for minimization:

$$\mathcal{L}_{const} = \frac{1}{N_{tar}} \sum_{i=N_{src}+1}^{N_{src}+N_{tar}} ||V^1(X_i) - V^2(X_i)||_2^2 \tag{4}$$

As such, consider an unexpected alignment by LANN where a set of coreferential examples $S^{src} \subset \{(X_i, y_i) \in \mathcal{D}^{src} | y_i = 1\}$ is aligned a set of non-coreferential examples $T^{tar} \subset \mathcal{D}^{tar}$ by view $V^1(X)$ ($V^1(S^{src}) \longleftrightarrow V^1(T^{tar})$). Our prediction consistency regularization $\mathcal{L}_{const}$ between two views will hep to penalize this unexpected alignment as it incorporates the difference between representation vectors from two views V^1 and V^2 over the target examples in T^{tar} (i.e., $V^1(T^{tar})$ and $V^2(T^{tar})$) into the loss function. Due to the alignment $V^1(S^{src}) \longleftrightarrow V^1(T^{tar})$, this implicitly translates into injecting the difference between representation vectors in $V^1(S^{src})$ and $V^2(T^{tar})$ into the loss function. However, this difference is expected to be high to prevent the alignment between $V^1(S^{src})$ and $V^1(T^{tar})$ for two reasons: (i) V^1 and V^2 are regularized to encode different information via $\mathcal{L}_{diver}$, and (ii) S^{src} and T^{tar} contain examples with different coreference labels, implying the large distance between their representation vectors for ECR. Consequently, the overall loss function to train models in our two-view model is: $\mathcal{L} = \mathcal{L}_{pred} + \alpha^1_{disc}\mathcal{L}^1_{disc} + \alpha^2_{disc}\mathcal{L}^2_{disc} + \alpha_{diver}\mathcal{L}_{diver} + \alpha_{const}\mathcal{L}_{const}$ where α^1_{disc}, α^2_{disc}, α_{diver}, and α_{const} are trade-off parameters.

3.4 Optimal Transport

Another limitation of LANN is that it employs all examples of the source and target language data in $\mathcal{D}^{src}$ and $\mathcal{D}^{tar}$ for the language discriminators. This is unexpected as faraway examples might not provide useful training signals for the language discriminators in general representation learning. As such, we aim to only apply the language discriminators to examples in the source and target language data that are close to each other. Given that, the major question is how to effectively estimate the distance between examples in the source and target languages in ECR for this example selection. To this end, as motivated in the introduction, our intuition is to simultaneously consider representations and coreference likelihoods of examples in $\mathcal{D}^{src}$ and $\mathcal{D}^{tar}$ to compute this distance function.

In particular, we directly use the vector $V(X)$ obtained before as the representation vector of X for our example selection purpose in LANN. Afterward, to obtain a coreference likelihood score u^X for an example X, we compute the average of the probabilities for being coreferential of X from the two view's coreference distributions $P^1(y = 1|X)$ and $P^2(y = 1|X)$: $u^X = \frac{P^1(y=1|X)+P^2(y=1|X)}{2}$. Consequently, to exploit both $V(X)$ and u^X of examples X for distance estimation between examples, we seek to find an optimal alignment between examples in the source and target language data $\mathcal{D}^{src}$ and $\mathcal{D}^{tar}$ such that two examples with closer representation vectors and coreference likelihoods have better chance to be aligned to each other. As such, this problem can be solved naturally with optimal transport (OT) methods that facilitate the computation of the optimal mapping between two probability distributions.

Formally, given two probability distributions $p(s)$ and $q(t)$ over domains $\mathcal{S}$ and $\mathcal{T}$, and a cost function $C(s, t) : \mathcal{S} \times \mathcal{T} \to \mathbb{R}_+$ for mapping $\mathcal{S}$ to

$\mathcal{T}$, OT finds the optimal joint distribution $\pi^*(s,t)$ (over $\mathcal{S} \times \mathcal{T}$), which has marginals $p(s)$ and $q(t)$ and achieves cheapest transportation from $p(s)$ to $q(t)$, by solving the following problem:

$$\pi^*(s,t) = \min_{\pi \in \Pi(s,t)} \sum_{s \in \mathcal{S}} \sum_{t \in \mathcal{T}} \pi(s,t) C(s,t) ds dt \quad \text{s.t. } s \sim p(s) \text{ and } t \sim q(t) \tag{5}$$

where $\Pi(s,t)$ is the set of all joint distributions with marginals $p(s)$ and $q(t)$. Here, π^* represents a matrix whose entry (s,t) represents the probability of transforming the data point $s \in \mathcal{S}$ to $t \in \mathcal{T}$ to convert the distribution $p(s)$ to $q(t)$. To this end, our model defines the domains $\mathcal{S}$ and $\mathcal{T}$ in OT via representation vectors for examples in the source and target language data $\mathcal{D}^{src}$ and $\mathcal{D}^{tar}$ respectively: $\mathcal{S} = \{V(X_i) | X_i \in \mathcal{D}^{src}\}$, $\mathcal{T} = \{V(X_i) | X_i \in \mathcal{D}^{tar}\}$. As such, the cost function $C(X_i, X_j)$ ($X_i \in \mathcal{D}^{src}$, $X_j \in \mathcal{D}^{tar}$) is computed by the Euclidean distance between representation vectors of corresponding elements, i.e., $C(X_i, X_j) = ||V(X_i) - V(X_j)||_2^2$. Also, the probability distributions $p(X_i)$ and $q(X_j)$ ($X_i \in \mathcal{D}^{src}$, $X_j \in \mathcal{D}^{tar}$) are defined over the normalized likelihood scores u^{X_i} and u^{X_j}, i.e., $p(X_i) = \textit{softmax}(u^{X_i} | X_i \in \mathcal{D}^{src})$ and $p(X_j) = \textit{softmax}(u^{X_j} | X_j \in \mathcal{D}^{tar})$. Based on these definitions, the element (X_i, X_j) of the OT solution matrix π^*, which is obtained by solving Equation 5, can be used as the distance between the example X_i and X_j ($X_i \in \mathcal{D}^{src}$, $X_j \in \mathcal{D}^{tar}$), aggregating the information from both representation vectors $V(X)$ and coreference likelihoods u^X.

To facilitate the example selection, we leverage $\pi^*(X_i, X_j)$ to compute an overall score v_i for each example $X_i \in \mathcal{D}^{src}$ to capture the closeness of X_i w.r.t examples in the target language using the average distance: $v_i = \frac{\sum_{X_j \in \mathcal{D}^{tar}} \pi^*(X_i, X_j)}{|\mathcal{D}^{src}|}$. Similarly, we obtain an overall score v_j for each example $X_j \in \mathcal{D}^{tar}$: $v_j = \frac{\sum_{X_i \in \mathcal{D}^{src}} \pi^*(X_i, X_j)}{|\mathcal{D}^{tar}|}$. Finally, based on the overall scores v_i and v_j, we only select γ percents of examples in $\mathcal{D}^{src}$ and γ percents of examples in $\mathcal{D}^{tar}$ that have smallest scores in their corresponding sets to participate into the loss functions $\mathcal{L}^1_{disc}$ and $\mathcal{L}^2_{disc}$ of the language discriminators for representation learning (i.e., the unselected examples are not included in the discriminators' loss functions). Here, γ is a hyper-parameter of the model. Note that as solving the OT problem in Equation 5 is intractable, we employ the entropy-based approximation of OT and solve it with the Sinkhorn algorithm (Peyre and Cuturi, 2019).

4 Experiments

Datasets and Hyper-parameters: We leverage the multilingual KBP datasets annotated by NIST (Mitamura et al., 2015, 2016, 2017) to perform cross-lingual evaluation for ECR models in this work. In particular, we use the KBP 2015 dataset (Mitamura et al., 2015) that provides annotation for 360 documents in English to train ECR models. For test and development data, we employ annotated articles for ECR in English, Spanish and Chinese of the KBP 2016 and KBP 2017 datasets. Here, KBP 2016 (Mitamura et al., 2016) involves 85 articles for each language English, Spanish and Chinese (i.e., $3 * 85 = 255$ documents in total) while the number of articles for each language in KBP 2017 (Mitamura et al., 2017) is 83 (i.e., $3 * 83 = 249$ documents). As such, for each language (English, Spanish or Chinese), when the models are tested on KBP 2016, we use a half of the KBP 2017 articles for the development data and the other half for unlabeled data in the language discriminators. Similarly for the testing on KBP 2017, articles in KBP 2016 will be used for development and unlabeled data. Finally, to focus the evaluation of cross-lingual transfer learning, we employ golden event mentions in documents in this work.

Following (Choubey and Huang, 2018; Huang et al., 2019), we employ the official KBP 2017 scorer (version 1.8) to obtain the coreference resolution performance for models. This evaluation script reports common performance metrics for ECR, including MUC (Vilain et al., 1995), B^3 (Bagga and Baldwin, 1998) and CEAF-e (Luo, 2005), BLANC (Lee et al., 2012b) and Average CoNLL (the average of four prior metrics).

Hyper-parameters for the models are fine-tuned by Average CoNLL scores over development data. The suggested values from the fine-tuning involve: $5e$-5 for the learning rate with the Adam optimizer (selected from [$1e$-5, $2e$-5, $3e$-5, $4e$-5, $5e$-5]); 512 for the numbers of hidden units in the middle layers of the feed-forward language discriminator D, D_1 and D_2 (selected from $[64, 128, 256, 512, 1024]$); $\alpha = 0.1$, $\alpha^1_{disc} = 0.1$, $\alpha^2_{disc} = 0.1$, $\alpha_{diver} = 0.01$, $\alpha_{const} = 0.01$ for the trade-off parameters in the loss functions of the models (selected from $[0.01, 0.05, 0.1, 0.5, 1]$); and $\gamma = 50\%$ for the percentage of selected examples for the language discriminators in the optimal transport (selected

	KBP 2016 Spanish					KBP 2017 Spanish				
Model	B^3	$CEAF_e$	MUC	BLANC	AVG-CoNLL	B^3	$CEAF_e$	MUC	BLANC	AVG-CoNLL
Baseline	86.67	80.17	71.30	80.47	79.65	84.24	78.17	60.86	73.93	74.30
LANN	86.61	79.86	71.80	80.63	79.72	84.70	78.83	61.56	74.60	74.92
CLMAOT	88.65	82.59	74.72	82.22	**82.05**	85.68	81.10	63.98	75.69	**76.61**
	KBP 2016 Chinese					**KBP 2017 Chinese**				
Model	B^3	$CEAF_e$	MUC	BLANC	AVG-CoNLL	B^3	$CEAF_e$	MUC	BLANC	AVG-CoNLL
Baseline	85.44	78.01	64.35	77.88	76.42	81.38	74.77	63.64	70.39	72.54
LANN	86.87	79.80	64.59	78.65	77.48	81.56	74.95	63.82	70.68	72.75
CLMAOT	89.03	84.17	66.75	79.02	**79.74**	83.01	77.85	62.97	69.64	**73.37**

Table 1: Cross-lingual performance on the test sets of KBP 2016 and 2017 for Spanish and Chinese. Models are trained on English documents of KBP 2015. The performance improvement of CLMAOT is significant with $p < 0.01$ over all datasets.

from $[10\%, 30\%, 50\%, 70\%, 90\%]$). Finally, we use the base version of XLM-RoBERTa for the models that has 768 dimensions for hidden vectors of word-pieces, leading to the dimensionality of $768*5 = 3840$ for the representation vectors $V(X)$ and determining the shape of the feed-forward networks (e.g., FF, FF_1, FF_2, f_1, f_2, D, D_1, D_2).

Model Performance: We compare the proposed model for ECR with cross-lingual multi-view alignment and optimal transport (called CLMAOT), the baseline model with XLM-RoBERTa in Section 3.1 (called Baseline), and the Baseline model introduced with LANN (called LANN) in Section 3.2. Table 1 reports the cross-lingual performance of the models on the KBP 2016 and 2017 test datasets for Spanish and Chinese (the models are trained in English documents in KBP 2015). As can be seen, LANN improves the cross-lingual performance of Baseline over different target languages and datasets (although the improvements are not significant for some datasets, i.e., KBP 2016 Spanish and KBP 2017 Chinese), thus suggesting the benefits of language discriminators for language generalization for ECR. More importantly, comparing with CLMAOT, we find that CLMAOT significantly outperforms other baseline models over different performance measures and target languages (i.e., Spanish and Chinese). In particular, for Spanish, CLMAOT is 2.33% and 1.70% better than LANN on the Average CoNLL scores over KBP 2016 and KBP 2017 respectively. For Chinese, the performance gaps between CLMAOT and LANN are 2.26% and 0.62% for KBP 2016 and KBP 2017 (with the Average CoNLL scores), thus demonstrating the effectiveness of the proposed cross-lingual model with multi-view alignment and optimal transport for representation learning in ECR.

Interestingly, we have also evaluated the ECR models (trained on English documents of KBP 2015) on the English documents of KBP 2016 and KBP 2017. The AVG-CoNLL scores of the Baseline, LANN, and CLMAOT models on KBP 2016 from our experiments are 68.64, 69.21, and 71.14 respectively while the corresponding scores for KBP 2017 involve 70.68, 71.75, and 73.48 (respectively). As such, CLMAOT is also significantly better than Baseline and LANN in English, thus highlighting the advantages of CLMAOT for ECR. Note that the worse performance of the models on English (compared to those on Spanish and Chinese) is potentially due to the larger number of event mentions in English documents in KBP 2016 and KBP 2017 (e.g., KBP 2016 has 2505, 1261, and 1390 event mentions in English, Spanish, and Chinese documents respectively).

Ablation Study: Two major components in the proposed model CLMAOT involve the multi-view alignment for representation vectors and the OT to select examples for LANN. This section evaluates ablated versions and variants of such components to reveal their contributions for CLMAOT. First, to highlight the importance of the proposed regularization terms in the loss function $\mathcal{L}$ for the multi-view alignment component, the following ablated models are considered: (i) **CLMAOT - LANN**: this model eliminates the language discriminators D_1 and D_2 with the loss terms $\mathcal{L}^1_{disc}$ and $\mathcal{L}^2_{disc}$ from CLMAOT; (ii) **CLMAOT - Diversity**: this model does not apply the diversity regularization over source-language examples $\mathcal{L}_{diver}$ in CLMAOT; and (iii) **CLMAOT - Consistency**: this model excludes the consistency regularization over target-language examples $\mathcal{L}_{const}$ from CLMAOT. In addition, we evaluate the variant (iv) **CLMAOT_OneView** of CLMAOT where the two-view representations $V^1(X)$ and $V^2(X)$ are not

Model	KBP 2016 Spanish					KBP 2016 Chinese				
	B^3	$CEAF_e$	MUC	BLANC	AVG-CoNLL	B^3	$CEAF_e$	MUC	BLANC	AVG-CoNLL
CLMAOT	88.65	82.59	74.72	82.22	**82.05**	89.03	84.17	66.75	79.02	**79.74**
CLMAOT - LANN	87.21	80.39	71.94	81.39	80.23	86.79	79.94	65.00	78.81	77.64
CLMAOT - Diversity	87.27	80.52	71.94	81.39	80.28	87.47	81.07	64.84	78.96	78.08
CLMAOT - Consistency	87.40	80.64	72.47	81.46	80.49	87.61	80.84	65.01	79.44	78.22
CLMAOT_OneView	86.66	80.29	71.74	80.75	79.86	86.65	79.64	65.53	79.15	77.74
CLMAOT - OT	87.50	80.84	72.41	81.65	80.60	87.22	80.48	64.95	78.85	77.88
CLMAOT - OT^{rep}	87.75	81.03	73.13	81.84	80.94	88.26	82.15	66.02	79.90	79.08
CLMAOT - OT^{coref}	87.79	81.45	73.26	81.75	81.06	88.29	82.47	66.09	79.21	79.02

Table 2: Ablation study for CLMAOT over the KBP 2016 datasets for Spanish and Chinese.

employed, thus directly using $V(X)$ for the language discriminator and avoiding the diversity and consistency regularization $\mathcal{L}_{diver}$ and $\mathcal{L}_{const}$. Note that the OT for example selection is still preserved in **CLMAOT_OneView**.

Second, for the optimal transport component, we evaluate the following variants for CLMAOT: (v) **CLMAOT - OT**: this model removes the optimal transport component and utilize all examples in the source and target languages for the language discriminators in CLMAOT ($\gamma = 100\%$); (vi) **CLMAOT - OT^{rep}**: this variant retains the OT component; however, instead of computing the cost function $C(X_i, X_j)$ based on the representation vectors for X_i and X_j, this version assumes a constant cost function $C_{X_i,X_j} = 1$, aiming to demonstrate the necessity of induced representation vectors for OT-based example selection for language discriminators; and (vii) **CLMAOT - OT^{coref}**: instead of relying on coreference likelihood scores to obtain the probability distributions $p(X_i)$ and $p(X_j)$, this model assumes uniform distributions for $p(X_i)$ and $p(X_j)$ in the OT computation. The motivation for this variant is to emphasize the importance of introducing coreference likelihood scores into OT for ECR.

Table 2 presents the performance of the models on the KBP 2016 test sets for Spanish and Chinese. It is clear from the table that the proposed regularization terms in the multi-view alignment component are helpful for CLMAOT as excluding any of them would significantly hurt the performance. We attribute this to the fact that the regularization terms in multi-view alignment might prevent the alignment of examples with different coreference labels in the source and target languages for the language discriminators. In addition, Table 2 shows that the performance of CLMAOT degrades when the optimal transport component or its elements (i.e., representation vectors for cost computation and coreference likelihoods for distributions) are removed. This thus proves the benefits of the designed optimal transport component in CLMAOT for cross-lingual ECR in this work.

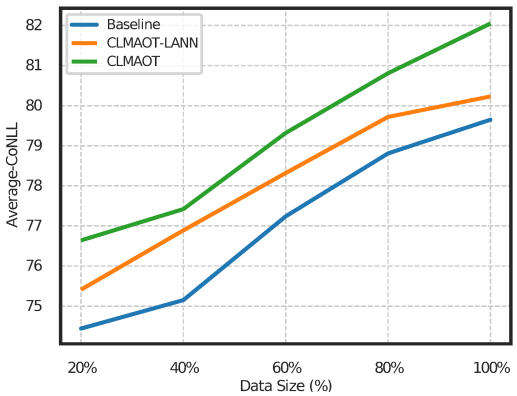

Figure 2: Learning curves over KBP 2016 for Spanish.

Finally, Figures 2 and 3 show the learning curves of Baseline, **CLMAOT - LANN**, and CLMAOT over Spanish and Chinese where we vary the size of the training data in English (from KBP 2015) and test the models' performance on the KBP 2016 test data. As can be seen, CLMAOT demonstrates better cross-lingual performance than the baseline models over different sizes of the training data, thus further confirming the effectiveness of our proposed model CLMAOT for ECR.

5 Conclusion

This paper presents the first study on cross-lingual transfer learning for event coreference resolution. We introduce the first baseline models for this problem, leveraging a state-of-the-art pre-trained language models for multilingual NLP (i.e., XLM-RoBERTa) and LANN for language-invariant representation learning. We propose two novel techniques for cross-lingual transfer learning based multi-view alignment to avoid cross-label align-

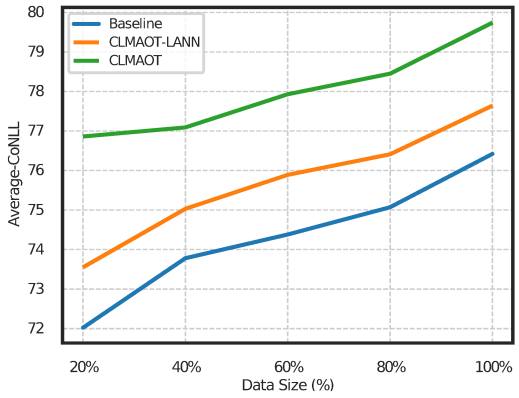

Figure 3: Learning curves over KBP 2016 for Chinese.

ment in the source and target languages and optimal transport for example selection in LANN. Our experiments provide baselines for future research and demonstrate the benefits of the proposed methods for cross-lingual transfer learning for ECR.

Acknowledgments

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

References

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

Jun Araki and Teruko Mitamura. 2015. Joint event trigger identification and event coreference resolution with structured perceptron. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

A. Bagga and B. Baldwin. 1998. Entity-based cross-document coreferencing using the vector space model. In *Proceedings of the International Conference on Computational Linguistics (COLING)*.

Shany Barhom, Vered Shwartz, Alon Eirew, Michael Bugert, Nils Reimers, and Ido Dagan. 2019. Revisiting joint modeling of cross-document entity and event coreference resolution. In *ACL*.

Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan. 2016. Domain separation networks. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*.

Chen Chen and Vincent Ng. 2016. Joint inference over a lightly supervised information extraction pipeline: Towards event coreference resolution for resource-scarce languages. 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 Annual Meeting of the Association for Computational Linguistics (ACL)*.

Xilun Chen, Yu Sun, Ben Athiwaratkun, Claire Cardie, and Kilian Weinberger. 2018a. Adversarial deep averaging networks for cross-lingual sentiment classification. In *Transactions of the Association for Computational Linguistics (TACL)*.

Xilun Chen, Yu Sun, Ben Athiwaratkun, Claire Cardie, and Kilian Weinberger. 2018b. Adversarial deep averaging networks for cross-lingual sentiment classification. *Transactions of the Association for Computational Linguistics (TACL)*.

Zheng Chen and Heng Ji. 2009a. Can one language bootstrap the other: a case study on event extraction. In *Workshop on Semi-Supervised Learning for Natural Language Processing*.

Zheng Chen and Heng Ji. 2009b. Graph-based event coreference resolution. In *Proceedings of the Workshop on Graph-based Methods for Natural Language Processing*.

Zheng Chen, Heng Ji, and Robert Haralick. 2009. A pairwise event coreference model, feature impact and evaluation for event coreference resolution. In *Proceedings of the Workshop on Events in Emerging Text Types*.

Prafulla Kumar Choubey and Ruihong Huang. 2018. Improving event coreference resolution by modeling correlations between event coreference chains and

document topic structures. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

Prafulla Kumar Choubey, Aaron Lee, Ruihong Huang, and Lu Wang. 2020. Discourse as a function of event: Profiling discourse structure in news articles around the main event. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Agata Cybulska and Piek T. J. M. Vossen. 2015. "bag of events" approach to event coreference resolution. supervised classification of event templates. In *Int. J. Comput. Linguistics Appl.*

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 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*.

Greg Durrett and Dan Klein. 2013. Easy victories and uphill battles in coreference resolution. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Lisheng Fu, Thien Huu Nguyen, Bonan Min, and Ralph Grishman. 2017. Domain adaptation for relation extraction with domain adversarial neural network. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (IJCNLP)*.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. In *Journal of Machine Learning Research*.

Christian Hardmeier, Jörg Tiedemann, and Joakim Nivre. 2013. Latent anaphora resolution for cross-lingual pronoun prediction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Andrew Hsi, Yiming Yang, Jaime Carbonell, and Ruochen Xu. 2016. Leveraging multilingual training for limited resource event extraction. In *Proceedings of the International Conference on Computational Linguistics (COLING)*.

Yin Jou Huang, Jing Lu, Sadao Kurohashi, and Vincent Ng. 2019. Improving event coreference resolution by learning argument compatibility from unlabeled data. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*.

Mandar Joshi, Danqi Chen, Y. Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2019. Spanbert: Improving pre-training by representing and predicting spans. In *Transactions of the Association for Computational Linguistics (TACL)*.

Kian Kenyon-Dean, Jackie Chi Kit Cheung, and Doina Precup. 2018. Resolving event coreference with supervised representation learning and clustering-oriented regularization. In *Proceedings of *SEM*.

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 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Abhishek Kumar, Prasanna Sattigeri, Kahini Wadhawan, Leonid Karlinsky, Rogerio Feris, Bill Freeman, and Gregory Wornell. 2018. Co-regularized alignment for unsupervised domain adaptation. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*.

Gourab Kundu, Avi Sil, Radu Florian, and Wael Hamza. 2018. Neural cross-lingual coreference resolution and its application to entity linking. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

Heeyoung Lee, Marta Recasens, Angel Chang, Mihai Surdeanu, and Dan Jurafsky. 2012a. Joint entity and event coreference resolution across documents. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Heeyoung Lee, Marta Recasens, Angel Chang, Mihai Surdeanu, and Dan Jurafsky. 2012b. Joint entity and event coreference resolution across documents. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. End-to-end neural coreference resolution. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Yankai Lin, Zhiyuan Liu, and Maosong Sun. 2017. Neural relation extraction with multi-lingual attention. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

Zhengzhong Liu, Jun Araki, Eduard Hovy, and Teruko Mitamura. 2014. Supervised within-document event coreference using information propagation. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*.

Jing Lu and Vincent Ng. 2017. Joint learning for event coreference resolution. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

Jing Lu and Vincent Ng. 2018. Event coreference resolution: A survey of two decades of research. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI)*.

Jing Lu, Deepak Venugopal, Vibhav Gogate, and Vincent Ng. 2016. Joint inference for event coreference resolution. In *Proceedings of the International Conference on Computational Linguistics (COLING)*.

Xiaoqiang Luo. 2005. On coreference resolution performance metrics. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

André F. T. Martins. 2015. Transferring coreference resolvers with posterior regularization. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

Teruko Mitamura, Zhengzhong Liu, and Eduard H. Hovy. 2015. Overview of TAC-KBP 2015 event nugget track. In *Proceedings of the Text Analysis Conference (TAC)*.

Teruko Mitamura, Zhengzhong Liu, and Eduard H. Hovy. 2016. Overview of TAC-KBP 2016 event nugget track. In *Proceedings of the Text Analysis Conference (TAC)*.

Teruko Mitamura, Zhengzhong Liu, and Eduard H. Hovy. 2017. Events detection, coreference and sequencing: What's next? overview of the TAC KBP 2017 event track. In *Proceedings of the Text Analysis Conference (TAC)*.

Aakanksha Naik and Carolyn Rose. 2020. Towards open domain event trigger identification using adversarial domain adaptation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

Vincent Ng. 2010. Supervised noun phrase coreference research: The first fifteen years. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

Nghia Trung Ngo, Duy Phung, and Thien Huu Nguyen. 2021. Unsupervised domain adaptation for event detection using domain-specific adapters. In *Findings of the Association for Computational Linguistics (ACL-IJCNLP)*.

Minh Van Nguyen, Viet Dac Lai, Amir Pouran Ben Veyseh, and Thien Huu Nguyen. 2021a. Trankit: A light-weight transformer-based toolkit for multilingual natural language processing. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations (EACL)*.

Minh Van Nguyen and Thien Huu Nguyen. 2021. Improving cross-lingual transfer for event argument extraction with language-universal sentence structures. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*.

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 (EMNLP)*.

Thien Huu Nguyen, , Adam Meyers, and Ralph Grishman. 2016. New york university 2016 system for kbp event nugget: A deep learning approach. In *Proceedings of the Text Analysis Conference (TAC)*.

Haoruo Peng, Yangqiu Song, and Dan Roth. 2016. Event detection and co-reference with minimal supervision. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing: Findings (EMNLP)*.

Gabriel Peyre and Marco Cuturi. 2019. Computational optimal transport: With applications to data science. In *Foundations and Trends in Machine Learning*.

Duy Phung, Tuan Ngo Nguyen, and Thien Huu Nguyen. 2021. Hierarchical graph convolutional networks for jointly resolving cross-document coreference of entity and event mentions. In *Proceedings of the Fifteenth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-15)*.

Karthik Raghunathan, Heeyoung Lee, Sudarshan Rangarajan, Nathanael Chambers, Mihai Surdeanu, Dan Jurafsky, and Christopher Manning. 2010. A multi-pass sieve for coreference resolution. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Altaf Rahman and Vincent Ng. 2012. Translation-based projection for multilingual coreference resolution. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*.

Ananya Subburathinam, Di Lu, Heng Ji, Jonathan May, Shih-Fu Chang, Avirup Sil, and Clare Voss. 2019a. Cross-lingual structure transfer for relation and event extraction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Ananya Subburathinam, Di Lu, Heng Ji, Jonathan May, Shih-Fu Chang, Avirup Sil, and Clare Voss. 2019b. Cross-lingual structure transfer for relation and event extraction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Hieu Minh Tran, Duy Phung, and Thien Huu Nguyen. 2021. Exploiting document structures and cluster consistencies for event coreference resolution. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL)*.

Gorka Urbizu, Ander Soraluze, and Olatz Arregi. 2019. Deep cross-lingual coreference resolution for less-resourced languages: The case of Basque. In *Proceedings of the Second Workshop on Computational Models of Reference, Anaphora and Coreference*.

Marc Vilain, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. 1995. A model-theoretic coreference scoring scheme. In *Sixth Message Understanding Conference (MUC-6)*.

Xiaozhi Wang, Xu Han, Yankai Lin, Zhiyuan Liu, and Maosong Sun. 2018. Adversarial multi-lingual neural relation extraction. In *Proceedings of the International Conference on Computational Linguistics (COLING)*.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Shijie Wu and Mark Dredze. 2020. Are all languages created equal in multilingual BERT? In *Proceedings of the 5th Workshop on Representation Learning for NLP*.

Bishan Yang, Claire Cardie, and Peter Frazier. 2015. A hierarchical distance-dependent Bayesian model for event coreference resolution. In *Transactions of the Association for Computational Linguistics (TACL)*.

Bowei Zou, Zengzhuang Xu, Yu Hong, and Guodong Zhou. 2018. Adversarial feature adaptation for cross-lingual relation classification. In *Proceedings of the International Conference on Computational Linguistics (COLING)*.