GATE: Graph Attention Transformer Encoder for Cross-lingual Relation and Event Extraction

Wasi Uddin Ahmad, Nanyun Peng, Kai-Wei Chang
University of California, Los Angeles
{wasiahmad, violetpeng, kwchang}@cs.ucla.edu

Abstract

Recent progress in cross-lingual relation and event extraction use graph convolutional networks (GCNs) with universal dependency parses to learn language-agnostic sentence representations such that models trained on one language can be applied to other languages. However, GCNs struggle to model words with long-range dependencies or are not directly connected in the dependency tree. To address these challenges, we propose to utilize the self-attention mechanism where we explicitly fuse structural information to learn the dependencies between words with different syntactic distances. We introduce GATE, a **G**raph **A**ttention **T**ransformer **E**ncoder, and test its cross-lingual transferability on relation and event extraction tasks. We perform experiments on the ACE05 dataset that includes three typologically different languages: English, Chinese, and Arabic. The evaluation results show that GATE outperforms three recently proposed methods by a large margin. Our detailed analysis reveals that due to the reliance on syntactic dependencies, GATE produces robust representations that facilitate transfer across languages.

1 Introduction

Relation and event extraction are two challenging information extraction (IE) tasks; wherein a model learns to identify semantic relationships between entities and events in narratives. They provide useful information for many natural language processing (NLP) applications such as knowledge graph completion (Lin et al. 2015) and question answering (Chen et al. 2019). Figure 1 gives an example of relation and event extraction tasks. Recent advances in cross-lingual transfer learning approaches for relation and event extraction learns a universal encoder that produces language-agnostic contextualized representations so the model learned on one language can easily transfer to others. Recent works (Huang et al. 2018; Subburathinam et al. 2019) suggested embedding *universal dependency structure* into contextual representations improves cross-lingual transfer for IE.

There are a couple of advantages of leveraging dependency structures. First, the syntactic distance between two words[1] in a sentence is typically smaller than the sequential distance. For example, in the sentence *A fire in a Bangladeshi garment factory has left at least 37 people dead and 100 hospitalized*, the sequential and syntactic distance between “fire” and “hospitalized” is 15 and 4, respectively. Therefore, encoding syntax structure helps capture long-range dependencies (Liu, Luo, and Huang 2018). Second, languages have different word order, e.g., adjectives precede or follow nouns as (“red apple”) in English or (“pomme rouge”) in French. Thus, processing sentences sequentially suffers from the word order difference issue (Ahmad et al. 2019a), while modeling dependency structures can mitigate the problem in cross-lingual transfer (Liu et al. 2019).

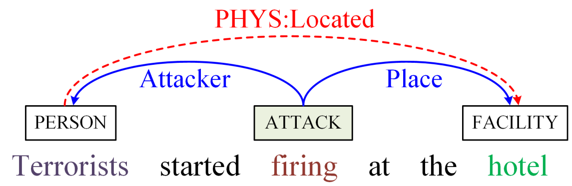


Figure 1: A relation (red dashed) between two entities and an event of type *Attack* (triggered by “firing”) including two arguments and their role labels (blue) are highlighted.

A common way to leverage dependency structures for cross-lingual NLP tasks is using universal dependency parses.[2] A large pool of recent works in IE (Liu, Luo, and Huang 2018; Zhang, Qi, and Manning 2018; Subburathinam et al. 2019; Fu, Li, and Ma 2019; Sun et al. 2019; Liu et al. 2019) employed Graph Convolutional Networks (GCNs) (Kipf and Welling 2017) to learn sentence representations based on their universal dependency parses, where a k-layers GCN aggregates information of words that are k hop away. Such a way of embedding structure may hinder cross-lingual transfer when the source and target languages have different path length distributions among words (see Table 1). Presumably, a two-layer GCN would work well on English but may not transfer well to Arabic.

Moreover, GCNs have shown to perform poorly in modeling long-distance dependencies or disconnected words in the dependency tree (Zhang, Li, and Song 2019; Tang et al. 2020). In contrast, the self-attention mechanism (Vaswani et al. 2017) is capable of capturing long-range dependencies. Consequently, a few recent studies proposed dependency-

[1] The shortest path in the dependency graph structure.

[2] https://universaldependencies.org/

aware self-attention and found effective for machine translation (Deguchi, Tamura, and Ninomiya 2019; Bugliarello and Okazaki 2020). The key idea is to allow attention between connected words in the dependency tree and gradually aggregate information across layers. However, IE tasks are relatively low-resource (the number of annotated documents available for training is small), and thus stacking more layers is not feasible. Besides, our preliminary analysis indicates that syntactic distance between entities could characterize certain relation and event types.[3] Hence, we propose to allow attention between all words but use the *pairwise syntactic distances* to weigh the attention.

We introduce a **G**raph **A**ttention **T**ransformer **E**ncoder (GATE) that utilizes self-attention (Vaswani et al. 2017) to learn structured contextual representations. On one hand, GATE enjoys the capability of capturing long-range dependencies, which is crucial for languages with longer sentences, e.g., Arabic.[4] On the other hand, GATE is agnostic to language word order as it uses syntactic distance to model pairwise relationship between words. This characteristic makes GATE suitable to transfer across typologically diverse languages, e.g., English to Arabic. One crucial property of GATE is that it allows information propagation among different heads in the multi-head attention structure based on syntactic distances, which allows to learn the correlation between different mention types and target labels.

We conduct experiments on cross-lingual transfer among English, Chinese, and Arabic languages using the ACE 2005 benchmark (Walker et al. 2006). The experimental results demonstrate that GATE outperforms three recently proposed relation and event extraction methods by a significant margin.[5] We perform a thorough ablation study and analysis, which shows that GATE is less sensitive to source language's characteristics (e.g., word order, sentence structure) and thus excels in the cross-lingual transfer.

2 Task Description

In this paper, we focus on *sentence-level* relation extraction (Subburathinam et al. 2019; Ni and Florian 2019) and event extraction (Subburathinam et al. 2019; Liu et al. 2019) tasks. Below, we first introduce the basic concepts, the notations, as well as define the problem and the scope of the work.

Relation Extraction is the task of identifying the relation type of an ordered pair of entity mentions. Formally, given a pair of entity mentions from a sentence s - $(e_s, e_o; s)$ where e_s and e_o denoted as the subject and object entities respectively, the relation extraction (RE) task is defined as predicting the relation $r \in R \cup \{\texttt{None}\}$ between the entity mentions, where R is a pre-defined set of relation types. In the example provided in Figure 1, there is a `PHYS:Located` relation between the entity mentions "Terrorists" and "hotel".

Event Extraction can be decomposed into two sub-tasks, *Event Detection* and *Event Argument Role Labeling*. Event detection refers to the task of identifying *event triggers* (the words or phrases that express event occurrences) and their types. In the example shown in Figure 1, the word "firing" triggers the `Attack` event.

Event argument role labeling (EARL) is defined as predicting whether words or phrases (arguments) participate in events and their roles. Formally, given an event trigger e_t and a mention e_a (an entity, time expression, or value) from a sentence s, the argument role labeling refers to predicting the mention's role $r \in R \cup \{\texttt{None}\}$, where R is a pre-defined set of role labels. In Figure 1, the "Terrorists" and "hotel" entities are the arguments of the `Attack` event and they have the `Attacker` and `Place` role labels, respectively.

In this work, we focus on the EARL task; we assume event mentions (triggers) of the input sentence are provided.

Zero-Short Cross-Lingual Transfer refers to the setting, where there is no labeled examples available for the *target* language. We train neural relation extraction and event argument role labeling models on one (single-source) or multiple (multi-source) *source* languages and then deploy the models in target languages. The overall cross-lingual transfer approach consists of four steps:

1. Convert the input sentence into a language-universal tree structure using an off-the-shelf universal dependency parser, e.g., UDPipe[6] (Straka and Straková 2017).
2. Embed the words in the sentence into a shared semantic space across languages. We use off-the-shelf multilingual contextual encoders (Devlin et al. 2019; Conneau et al. 2020) to form the word representations. To enrich the word representations, we concatenate them with *universal* part-of-speech (POS) tag, dependency relation, and entity type embeddings (Subburathinam et al. 2019). We collectively refer them as *language-universal* features.
3. Based on the word representations, we encode the input sentence using the proposed GATE architecture that leverages the syntactic depth and distance information. Note that this step is the main focus of this work.
4. A pair of classifier predicts the target relation and argument role labels based on the encoded representations.

3 Approach

Our proposed approach GATE revises the multi-head attention architecture in Transformer Encoder (Vaswani et al. 2017) to model syntactic information while encoding a sequence of input vectors (represent the words in a sentence) into contextualized representations. We first review the standard multi-head attention mechanism (§3.1). Then, we introduce our proposed method GATE (§3.2). Finally, we describe how we perform relation extraction (§3.3) and event argument role labeling (§3.4) tasks.

[3] In ACE 2005 dataset, the relation type `PHYS:Located` exists among {`PER, ORG, LOC, FAC, GPE`} entities. The average syntactic distance in English and Arabic sentences among `PER` and any of the {`LOC, FAC, GPE`} entities are approx. 2.8 and 4.2, while the distance between `PER` and `ORG` is 3.3 and 1.5.

[4] After tokenization, on average, ACE 2005 English and Arabic sentences have approximately 30 and 210 words, respectively.

[5] Code available at https://github.com/wasiahmad/GATE

[6] http://ufal.mff.cuni.cz/udpipe

3.1 Transformer Encoder

Unlike recent works (Zhang, Qi, and Manning 2018; Subburathinam et al. 2019) that use GCNs (Kipf and Welling 2017) to encode the input sequences into contextualized representations, we propose to employ Transformer encoder as it excels in capturing long-range dependencies. First, the sequence of input word vectors, $x = [x_1, \ldots, x_{|x|}]$ where $x_i \in R^d$ are packed into a matrix $H^0 = [x_1, \ldots, x_{|x|}]$. Then an L-layer Transformer Encoder $H^l = Transformer_l(H^{l-1})$, $l \in [1, L]$ takes H^0 as input and generates different levels of latent representations $H^l = [h^l_1, \ldots, h^l_{|x|}]$, recursively. Typically the latent representations generated by the last layer (L-th layer) are used as the contextual representations of the input words. To aggregate the output vectors of the previous layer, multiple (n_h) self-attention heads are employed in each Transformer layer. For the l-th Transformer layer, the output of the previous layer $H^{l-1} \in \mathbb{R}^{|x| \times d_{model}}$ is first linearly projected to queries Q, keys K, and values V using parameter matrices $W^Q_l, W^K_l \in \mathbb{R}^{d_{model} \times d_k}$ and $W^V_l \in \mathbb{R}^{d_{model} \times d_v}$, respectively.

$$Q_l = H^{l-1} W^Q_l, K_l = H^{l-1} W^K_l, V_l = H^{l-1} W^V_l.$$

The output of a self-attention head A_l is computed as:

$$A_l = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}} + M\right) V_l, \tag{1}$$

where the matrix $M \in \mathbb{R}^{|x| \times |x|}$ determines whether a pair of tokens can attend each other.

$$M_{ij} = \begin{cases} 0, & \text{allow to attend} \\ -\infty, & \text{prevent from attending} \end{cases} \tag{2}$$

The matrix M is deduced as a *mask*. By default, the matrix M is a *zero-matrix*. In the next section, we discuss how we manipulate the mask matrix M to incorporate syntactic depth and distance information in sentence representations.

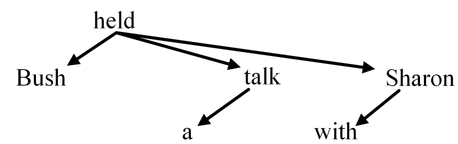
3.2 Graph Attention Transformer Encoder

The self-attention as described in §3.1 learns how much attention to put on words in a text sequence when encoding a word at a given position. In this work, we revise the self-attention mechanism such that it takes into account the syntactic structure and distances when a token attends to all the other tokens. The key idea is to manipulate the mask matrix to impose the graph structure and retrofit the attention weights based on pairwise syntactic distances. We use the universal dependency parse of a sentence and compute the syntactic (shortest path) distances between every pair of words. We illustrate an example in Figure 2.

We denote distance matrix $D \in \mathbb{R}^{|x| \times |x|}$ where D_{ij} represents the syntactic distance between words at position i and j in the input sequence. If we want to allow tokens to attend their adjacent tokens (that are 1 hop away) at each layer, then we can set the mask matrix as follows.

$$M_{ij} = \begin{cases} 0, & D_{ij} = 1 \\ -\infty, & \text{otherwise} \end{cases}$$

We generalize this notion to model a distance based attention; allowing tokens to attend tokens that are within distance δ (hyper-parameter).



	Bush	held	a	talk	with	Sharon
Bush	1	1	3	2	3	2
held	1	1	2	1	2	1
a	3	2	1	1	4	3
talk	2	1	1	1	3	2
with	3	2	4	3	1	1
Sharon	2	1	3	2	1	1

Figure 2: Distance matrix showing the shortest path distances between all pairs of words. The dependency arc direction is ignored while computing pairwise distances. The diagonal value is set to 1, indicating a self-loop. If we set the values in white cells (with value > 1) to 0, the distance matrix becomes an adjacency matrix.

$$M_{ij} = \begin{cases} 0, & D_{ij} \leq \delta \\ -\infty, & \text{otherwise} \end{cases} \tag{3}$$

During our preliminary analysis, we observed that syntactic distances between entity mentions or event mentions often correlate with the target label. For example, if an `ORG` entity mention appears closer to a `PER` entity than a `LOC` entity, then the {`PER, ORG`} entity pair is more likely to have the `PHYS:Located` relation. We hypothesize that modeling syntactic distance between words can help to identify complex semantic structure such as events and entity relations. Hence we revise the attention head A_l (defined in Eq. (1)) computation as follows.

$$A_l = F\left(\text{softmax}\left(\frac{QK^T}{\sqrt{d_k}} + M\right)\right) V_l. \tag{4}$$

Here, softmax produces an attention matrix $P \in R^{|x| \times |x|}$ where P_i denotes the attentions that i-th token pays to the all the tokens in the sentence, and F is a function that modifies those attention weights. We can treat F as a parameterized function that can be learned based on distances. However, we adopt a simple formulation of F such that GATE pays more attention to tokens that are closer and less attention to tokens that are faraway in the parse tree. We define the (i, j)-th element of the attention matrix produced by F as:

$$F(P)_{ij} = \frac{P_{ij}}{Z_i D_{ij}}, \tag{5}$$

where $Z_i = \sum_j \frac{P_{ij}}{D_{ij}}$ is the normalization factor and D_{ij} is the distance between i-th and j-th token. We found this formulation of F effective for the IE tasks.

3.3 Relation Extractor

Relation Extractor predicts the relationship label (or `None`) for each mention pair in a sentence. For an input sentence s, GATE produces contextualized word representations $h_1^l, \ldots, h_{|x|}^l$ where $h_i^l \in \mathbb{R}^{d_{model}}$. As different sentences and entity mentions may have different lengths, we perform max-pooling over their contextual representations to obtain fixed-length vectors. Suppose for a pair of entity mentions $e_s = [h_{bs}^l, \ldots, h_{es}^l]$ and $e_o = [h_{bo}^l, \ldots, h_{eo}^l]$, we obtain single vector representations $\hat{e}_s$ and $\hat{e}_o$ by performing max-pooling. Following Zhang, Qi, and Manning (2018); Subburathinam et al. (2019), we also obtain a vector representation for the sentence, $\hat{s}$ by applying max-pooling over $[h_1^l, \ldots, h_{|x|}^l]$ and concatenate the three vectors. Then the concatenation of the three vectors $[\hat{e}_s; \hat{e}_o; \hat{s}]$ are fed to a linear classifier followed by a Softmax layer to predict the relation type between entity mentions e_s and e_o as follows.

$$\mathcal{O}_r = \mathrm{softmax}(\boldsymbol{W}_r^T[\hat{e_s}; \hat{e_o}; \hat{s}] + \boldsymbol{b}_r),$$

where $\boldsymbol{W}_r \in R^{3d_{model} \times r}$ and $\boldsymbol{b}_r \in R^r$ are parameters, and r is the total number of relation types. The probability of t-th relation type is denoted as $P(r_t|s, e_s, e_o)$, which corresponds to the t-th element of $\mathcal{O}_r$. To train the relation extractor, we adopt the cross-entropy loss.

$$\mathcal{L}_r = -\sum_{s=1}^{N}\sum_{o=1}^{N} \log(P(y_{so}^r|s, e_s, e_o)),$$

where N is the number of entity mentions in the input sentence s and y_{so}^r denotes the ground truth relation type between entity mentions e_s and e_o.

3.4 Event Argument Role Labeler

Event argument role labeler predicts the argument mentions (or `None` for non-argument mentions) of an event mention and assigns a role label to each argument from a pre-defined set of labels. To label an argument candidate $e_a = [h_{ba}^l, \ldots, h_{ea}^l]$ for an event trigger $e_t = [h_{bt}^l, \ldots, h_{et}^l]$ in sentence $s = [h_1^l, \ldots, h_{|x|}^l]$, we apply max-pooling to form vectors $\hat{e}_a$, $\hat{e}_t$, and $\hat{s}$ respectively, which is same as that for relation extraction. Then we concatenate the vectors ($[\hat{e_t}; \hat{e_a}; \hat{s}]$) and pass it through a linear classifier and Softmax layer to predict the role label as follows.

$$\mathcal{O}_a = \mathrm{softmax}(\boldsymbol{W}_a^T[\hat{e_t}; \hat{e_a}; \hat{s}] + \boldsymbol{b}_a),$$

where $\boldsymbol{W}_a \in R^{3d_{model} \times r}$ and $\boldsymbol{b}_a \in R^r$ are parameters, and r is the total number of argument role label types. We optimize the role labeler by minimizing the cross-entropy loss.

4 Experiment Setup

Dataset We conduct experiments based on the Automatic Content Extraction (ACE) 2005 corpus (Walker et al. 2006) that includes manual annotation of relation and event mentions (with their arguments) in three languages: English (En), Chinese (Zh), and Arabic (Ar). We present the data statistics in Appendix. ACE defines an ontology that includes 7 entity types, 18 relation subtypes, and 33 event subtypes. We add a class label `None` to denote that two entity mentions or a pair of an event mention and an argument candidate under consideration do not have a relationship belong to the target ontology. We use the same dataset split as Subburathinam et al. (2019) and follow their preprocessing steps. We refer the readers to Subburathinam et al. (2019) for the dataset preprocessing details.

	Sequential			Syntactic		
	En	Zh	Ar	En	Zh	Ar
Relation Mention	4.8	3.9	25.8	2.2	2.6	5.1
Event Mention & Arg.	9.8	21.7	58.1	3.1	4.6	12.3

Table 1: Average sequential and syntactic (shortest path) distance between relation mentions and event mentions and their candidate arguments in ACE05 dataset. Distances are computed by ignoring the order of mentions.

Evaluation Criteria Following the previous works (Ji and Grishman 2008; Li, Ji, and Huang 2013; Li and Ji 2014; Subburathinam et al. 2019), we set the evaluation criteria as, (1) a relation mention is correct if its predicted type and the head offsets of the two associated entity mentions are correct, and (2) an event argument role label is correct if the event type, offsets, and argument role label match any of the reference argument mentions.

Baseline Models To compare GATE on relation and event argument role labeling tasks, we chose three recently proposed approaches as baselines. The source code of the baselines are not publicly available at the time this research is conducted. Therefore, we reimplemented them.

• CL_Trans_GCN (Liu et al. 2019) is a context-dependent lexical mapping approach where each word in a source language sentence is mapped to its best-suited translation in the target language. We use multilingual word embeddings (Joulin et al. 2018) as the continuous representations of tokens along with the language-universal features embeddings including part-of-speech (POS) tag embedding, dependency relation label embedding, and entity type embedding.[7] Since this model focuses on the target language, we train this baseline for each combination of source and target languages.

• CL_GCN (Subburathinam et al. 2019) uses GCN (Kipf and Welling 2017) to learn structured common space representation. To embed the tokens in an input sentence, we use multilingual contextual representations (Devlin et al. 2019; Conneau et al. 2020) and the language-universal feature embeddings. We train this baseline on the source languages and directly evaluate on the target languages.

• CL_RNN (Ni and Florian 2019) uses a bidirectional Long Short-Term Memory (LSTM) type recurrnet neural networks (Hochreiter and Schmidhuber 1997) to learn contextual representation. We feed language-universal features for words in a sentence, constructed in the same way as Subburathinam et al. (2019). We train and evaluate this baseline in the same way as CL_GCN.

In addition to the above three baseline methods, we compare GATE with the following two encoding methods.

[7] Due to the design principle of Liu et al. (2019), we cannot use multilingual contextual encoders in CL_Trans_GCN.

Model	Event Argument Role Labeling						Relation Extraction					
	En ⇓ Zh	En ⇓ Ar	Zh ⇓ En	Zh ⇓ Ar	Ar ⇓ En	Ar ⇓ Zh	En ⇓ Zh	En ⇓ Ar	Zh ⇓ En	Zh ⇓ Ar	Ar ⇓ En	Ar ⇓ Zh
CL_Trans_GCN	41.8	55.6	41.2	52.9	39.6	40.8	56.7	65.3	65.9	59.7	59.6	46.3
CL_GCN	51.9	50.4	53.7	51.5	50.3	51.9	49.4	58.3	65.0	55.0	56.7	42.4
CL_RNN	60.4	53.9	55.7	52.5	50.7	50.9	53.7	63.9	70.9	57.6	67.1	55.7
Transformer	61.5	55.0	58.0	57.7	54.3	57.0	57.1	63.4	69.6	60.6	67.0	52.6
Transformer_RPR	62.3	60.8	57.3	66.3	**57.5**	**59.8**	**58.0**	59.9	70.0	55.6	66.5	**56.5**
GATE (this work)	**63.2**	**68.5**	**59.3**	**69.2**	53.9	57.8	55.1	**66.8**	**71.5**	**61.2**	**69.0**	54.3

Table 2: Single-source transfer results (F-score % on the test set) using perfect event triggers and entity mentions. The language on top and bottom of ⇓ denotes the source and target languages, respectively.

Model	{En, Zh} ⇓ Ar	{En, Ar} ⇓ Zh	{Zh, Ar} ⇓ En
Event Argument Role Labeling			
CL_Trans_GCN	57.0	44.5	44.8
CL_GCN	58.9	56.2	57.9
CL_RNN	53.5	62.5	60.8
Transformer	59.5	62.0	60.7
Transformer_RPR	71.1	**68.4**	**62.2**
GATE (this work)	**73.9**	65.3	61.3
Relation Extraction			
CL_Trans_GCN	66.8	54.4	69.5
CL_GCN	64.0	46.6	65.8
CL_RNN	66.5	60.5	73.0
Transformer	**68.3**	59.3	73.7
Transformer_RPR	65.0	**62.3**	73.8
GATE (this work)	67.0	57.9	**74.1**

Table 3: Multi-source transfer results (F-score % on the test set) using perfect event triggers and entity mentions. The language on top and bottom of ⇓ denotes the source and target languages, respectively.

• Transformer (Vaswani et al. 2017) uses multi-head self-attention mechanism and is the base structure of our proposed model, GATE. Note that GATE has the same number of parameters as Transformer since GATE does not introduce any new parameter while modeling the pairwise syntactic distance into the self-attention mechanism. Therefore, we credit the GATE's improvements over the Transformer to its distance-based attention modeling strategy.

• Transformer_RPR (Shaw, Uszkoreit, and Vaswani 2018) uses relative position representations to encode the structure of the input sequences. This method uses the pairwise *sequential* distances while GATE uses pairwise *syntactic* distances to model attentions between tokens.

Implementation Details To embed words into vector representations, we use multilingual BERT (M-BERT) (Devlin et al. 2019). Note that we do not fine-tune M-BERT, but only use it as a feature extractor. We use the universal part-of-speech (POS) tags, dependency relation labels, and seven entity types defined by ACE: person, organization, geo-political entity, location, facility, weapon, and vehicle. We embed these language-universal features into fixed-length vectors and concatenate them with M-BERT vectors to form the input word representations. We set the model size (d_{model}), number of encoder layers (L), and attention heads (n_h) in multi-head to 512, 1, and 8 respectively. We tune the distance threshold δ (as shown in Eq. (3)) in $[1, 2, 4, 8, \infty]$ for each attention head on each source language (more details are provided in the supplementary).

We implement all the baselines and our approach based on the implementation of Zhang, Qi, and Manning (2018) and OpenNMT (Klein et al. 2017). We used `transformers`[8] to extract M-BERT and XLM-R features. We provide a detailed description of the dataset, hyper-parameters, and training of the baselines and our approach in the supplementary.

5 Results and Analysis

We compare GATE with five baseline approaches on event argument role labeling (EARL) and relation extraction (RE) tasks, and the results are presented in Table 2 and 3.

Single-source transfer In the single-source transfer setting, all the models are individually trained on *one* source language, e.g., English and directly evaluated on the other two languages (target), e.g., Chinese and Arabic. Table 2 shows that GATE outperforms all the baselines in four out of six transfer directions on both tasks. CL_RNN surprisingly outperforms CL_GCN in most settings, although CL_RNN uses a BiLSTM that is not suitable to transfer across syntactically different languages (Ahmad et al. 2019a). We hypothesize the reason being GCNs cannot capture long-range dependencies, which is crucial for the two tasks. In comparison, by modeling distance-based pairwise relationships among words, GATE excels in cross-lingual transfer.

A comparison between Transformer and GATE demonstrates the effectiveness of syntactic distance-based self-attention over the standard mechanism. From Table 2, we see GATE outperforms Transformer with an average improvement of 4.7% and 1.3% in EARL and RE tasks, respectively. Due to implicitly modeling graph structure, Transformer_RPR performs effectively. However, GATE achieves an average improvement of 1.3% and 1.9% in EARL and RE tasks over Transformer_RPR. Overall, the significant performance improvements achieved by GATE corroborate our hypothesis that syntactic distance-based attention helps in the cross-lingual transfer.

[8]https://github.com/huggingface/transformers

Model	EARL		RE	
	Chinese	Arabic	Chinese	Arabic
Wang et al. (2019)				
Absolute	61.2	53.5	57.8	65.2
Relative	55.3	47.1	**58.1**	66.4
GATE	**63.2**	**68.5**	55.1	**66.8**

Table 4: GATE vs. Wang et al. (2019) results (F-score %) on event argument role labeling (EARL) and relation extraction (RE); using English as source and Chinese, Arabic as the target languages, respectively. To limit the maximum relative position, the clipping distance is set to 10 and 5 for EARL and RE tasks, respectively.

Multi-source transfer In the multi-source cross-lingual transfer, the models are trained on a pair of languages: {English, Chinese}, {English, Arabic}, and {Chinese, Arabic}. Hence, the models observe more examples during training, and as a result, the cross-lingual transfer performance improves compared to the single-source transfer setting. In Table 3, we see GATE outperforms the previous three IE approaches in multi-source transfer settings, except on RE for the source:{English, Arabic} and target: Chinese language setting. On the other hand, GATE performs competitively to Transformer and Transformer_RPR baselines. Due to observing more training examples, Transformer and Transformer_RPR perform more effectively in this setting. The overall result indicates that GATE more efficiently learns transferable representations for the IE tasks.

Encoding dependency structure GATE encodes the dependency structure of sentences by guiding the attention mechanism in self-attention networks (SANs). However, an alternative way to encode the sentence structure is through positional encoding for SANs. Conceptually, the key difference is the modeling of syntactic distances to capture fine-grained relations among tokens. Hence, we compare these two notions of encoding the dependency structure to emphasize the promise of modeling syntactic distances.

To this end, we compare the GATE with Wang et al. (2019) that proposed structural position encoding using the dependency structure of sentences. Results are presented in Table 4. We see that Wang et al. (2019) performs well on RE but poorly on EARL, especially on the Arabic language. While GATE directly uses syntactic distances between tokens to guide the self-attention mechanism, Wang et al. (2019) learns parameters to encode structural positions that can become sensitive to the source language. For example, the average shortest path distance between event mentions and their candidate arguments in English and Arabic is 3.1 and 12.3, respectively (see Table 1). As a result, a model trained in English may learn only to attend closer tokens, thus fails to generalize on Arabic.

Moreover, we anticipate that different order of subject and verb in English and Arabic[9] causes Wang et al. (2019) to transfer poorly on the EARL task (as event triggers are mostly verbs). To verify our anticipation, we modify the relative structural position encoding (Wang et al. 2019) by dropping the directional information (Ahmad et al. 2019a), and observed a performance increase from 47.1 to 52.2 for English to Arabic language transfer. In comparison, GATE is order-agnostic as it models syntactic distance; hence, it has a better transferability across typologically diverse languages.

[9]According to WALS (Dryer and Haspelmath 2013), the order of subject (S), object (O), and verb (V) for English, Chinese and Arabic is SVO, SVO, and VSO.

Model	EARL		RE	
	English	Chinese*	English	Chinese*
CL_GCN	51.5	56.3	46.9	50.7
CL_RNN	55.6	59.3	56.8	**62.0**
GATE	**63.8**	**64.2**	**58.8**	57.0

Table 5: Event argument role labeling (EARL) and relation extraction (RE) results (F-score %); using Chinese as the source and English as the target language. * indicates the English examples are translated into Chinese using `Google Cloud Translate`.

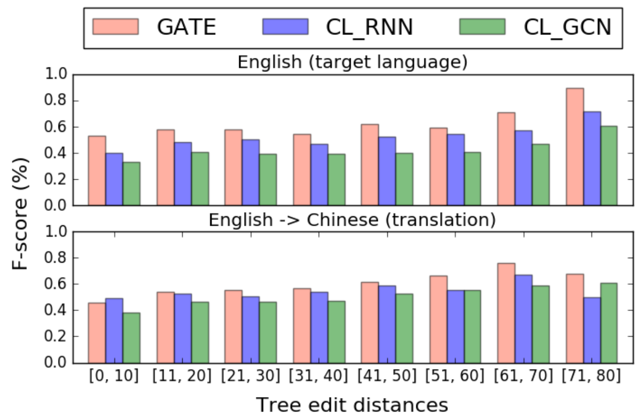


Figure 3: Models trained on the Chinese language perform on event argument role labeling in English and their parallel Chinese sentences. The parallel sentences have the same meaning but a different structure. To quantify the structural difference between two parallel sentences, we compute the tree edit distances.

Sensitivity towards source language Intuitively, an RE or EARL model would transfer well on target languages if the model is less sensitive towards the source language characteristics (e.g., word order, grammar structure). To measure sensitivity towards the source language, we evaluate the model performance on the target language and their parallel (translated) source language sentences. We hypothesize that if a model performs significantly well on the translated source language sentences, then the model is more sensitive towards the source language and may not be ideal for cross-lingual transfer. To test the models on this hypothesis, we translate all the ACE05 English test set examples into Chinese using `Google Cloud Translate`.[10] We train GATE and two baselines on the Chinese and evaluate them on both English (test set) examples and their Chinese translations. To quantify the difference between the dependency

[10]Details are provided in the supplementary.

Word features	EARL		RE	
	Chinese	Arabic	Chinese	Arabic
Multi-WE	35.9	43.7	41.0	54.9
M-BERT	**57.1**	54.8	**55.1**	66.8
XLM-R	51.8	**61.7**	51.4	**68.1**

Table 6: Contribution of multilingual word embeddings (Multi-WE) (Joulin et al. 2018), M-BERT (Devlin et al. 2019), and XLM-R (Conneau et al. 2020) as a source of word features; using English as source and Chinese, Arabic as the target languages, respectively.

Input features	EARL		RE	
	Chinese	Arabic	Chinese	Arabic
M-BERT	52.5	47.4	44.0	49.7
+ POS tag	49.3	47.5	44.1	47.0
+ Dep. label	49.7	51.0	48.6	47.0
+ Entity type	**57.8**	**60.2**	**56.3**	**63.0**

Table 7: Ablation on the use of language-universal features (part-of-speech (POS) tag, dependency relation label, and entity type) in GATE (F-score (%); using English as source and Chinese, Arabic as the target languages, respectively.

structure of an English and its Chinese translation sentences, we compute *edit distance* between two tree structures using the APTED[11] algorithm (Pawlik and Augsten 2015, 2016).

The results are presented in Table 5. We see that CL_GCN and CL_RNN have much higher accuracy on the translated (Chinese) sentences than the target language (English) sentences. On the other hand, GATE makes a roughly similar number of correct predictions when the target and translated sentences are given as input. Figure 3 illustrates how the models perform when the structural distance between target sentences and their translation increases. The results suggest that GATE performs substantially better than the baselines when the target language sentences are structurally different from the source language. The overall findings signal that GATE is less sensitive to source language features, and we credit this to the modeling of distance-based syntactic relationships between words. We acknowledge that there might be other factors associated with a model's language sensitivity. However, we leave the detailed analysis for measuring a model's sensitivity towards languages as future work.

Ablation study We perform a detailed ablation study on language-universal features and sources of word features to examine their individual impact on cross-lingual transfer. The results are presented in Table 6 and 7. We observed that M-BERT and XLM-R produced word features performed better in Chinese and Arabic, respectively, while they are comparable in English. On average M-BERT performs better, and thus we chose it as the word feature extractor in all our experiments. Table 7 shows that part-of-speech and dependency relation embedding has a limited contribution. This is perhaps due to the tokenization errors, as pointed out by Subburathinam et al. (2019). However, the use of language-universal features is useful, particularly when we have minimal training data. We provide more analysis and results in the supplementary.

[11]https://pypi.org/project/apted/

6 Related Work

Relation and event extraction has drawn significant attention from the natural language processing (NLP) community. Most of the approaches developed in past several years are based on supervised machine learning, using either symbolic features (Ahn 2006; Ji and Grishman 2008; Liao and Grishman 2010; Hong et al. 2011; Li, Ji, and Huang 2013; Li and Ji 2014) or distributional features (Liao and Grishman 2011; Nguyen, Cho, and Grishman 2016; Miwa and Bansal 2016; Liu et al. 2018; Zhang et al. 2018; Lu and Nguyen 2018; Chen et al. 2015; Nguyen and Grishman 2015; Zeng et al. 2014; Peng et al. 2017; Nguyen and Grishman 2018; Zhang, Qi, and Manning 2018; Subburathinam et al. 2019; Liu et al. 2019; Huang, Yang, and Peng 2020) from a large number of annotations. Joint learning or inference (Bekoulis et al. 2018; Li et al. 2014; Zhang, Ji, and Sil 2019; Liu, Luo, and Huang 2018; Nguyen, Cho, and Grishman 2016; Yang and Mitchell 2016; Han, Ning, and Peng 2019; Han, Zhou, and Peng 2020) are also among the noteworthy techniques.

Most previous works on cross-lingual transfer for relation and event extraction are based on annotation projection (Kim et al. 2010a; Kim and Lee 2012), bilingual dictionaries (Hsi et al. 2016; Ni and Florian 2019), parallel data (Chen and Ji 2009; Kim et al. 2010b; Qian et al. 2014) or machine translation (Zhu et al. 2014; Faruqui and Kumar 2015; Zou et al. 2018). Learning common patterns across languages is also explored (Lin, Liu, and Sun 2017; Wang et al. 2018; Liu et al. 2018). In contrast to these approaches, Subburathinam et al. (2019); Liu et al. (2019) proposed to use graph convolutional networks (GCNs) (Kipf and Welling 2017) to learn multi-lingual structured representations. However, GCNs struggle to model long-range dependencies or disconnected words in the dependency tree. To overcome the limitation, we use the syntactic distances to weigh the attentions while learning contextualized representations via the multi-head attention mechanism (Vaswani et al. 2017).

Moreover, our proposed syntax driven distance-based attention modeling helps to mitigate the word order difference issue (Ahmad et al. 2019a) that hinders cross-lingual transfer. Prior works studied dependency structure modeling (Liu et al. 2019), source reordering (Rasooli and Collins 2019), adversarial training (Ahmad et al. 2019b), constrained inference (Meng, Peng, and Chang 2019) to tackle word order differences across typologically different languages.

7 Conclusion

In this paper, we proposed to model fine-grained syntactic structural information based on the dependency parse of a sentence. We developed a **G**raph **A**ttention **T**ransformer **E**ncoder (GATE) to generate structured contextual representations. Extensive experiments on three languages demonstrates the effectiveness of GATE in cross-lingual relation and event extraction. In the future, we want to explore other sources of language-universal information to improve structured representation learning.

Acknowledgments

This work was supported in part by National Science Foundation (NSF) Grant OAC 1920462 and the Intelligence Advanced Research Projects Activity (IARPA) via Contract No. 2019-19051600007.

References

Ahmad, W.; Zhang, Z.; Ma, X.; Hovy, E.; Chang, K.-W.; and Peng, N. 2019a. On Difficulties of Cross-Lingual Transfer with Order Differences: A Case Study on Dependency Parsing. In *Proceedings of NAACL*, 2440–2452.

Ahmad, W. U.; Zhang, Z.; Ma, X.; Chang, K.-W.; and Peng, N. 2019b. Cross-Lingual Dependency Parsing with Unlabeled Auxiliary Languages. In *Proceedings of CoNLL*, 372–382.

Ahn, D. 2006. The stages of event extraction. In *Proceedings of the Workshop on Annotating and Reasoning about Time and Events*, 1–8.

Bekoulis, G.; Deleu, J.; Demeester, T.; and Develder, C. 2018. Adversarial training for multi-context joint entity and relation extraction. In *Proceedings of EMNLP*, 2830–2836.

Bugliarello, E.; and Okazaki, N. 2020. Enhancing Machine Translation with Dependency-Aware Self-Attention. In *Proceedings of ACL*, 1618–1627.

Chen, Y.; Xu, L.; Liu, K.; Zeng, D.; and Zhao, J. 2015. Event Extraction via Dynamic Multi-Pooling Convolutional Neural Networks. In *Proceedings of ACL-IJCNLP*, 167–176.

Chen, Z.; and Ji, H. 2009. Can One Language Bootstrap the Other: A Case Study on Event Extraction. In *Proceedings of the NAACL HLT 2009 Workshop on Semi-supervised Learning for Natural Language Processing*, 66–74.

Chen, Z.-Y.; Chang, C.-H.; Chen, Y.-P.; Nayak, J.; and Ku, L.-W. 2019. UHop: An Unrestricted-Hop Relation Extraction Framework for Knowledge-Based Question Answering. In *Proceedings of NAACL*, 345–356.

Conneau, A.; Khandelwal, K.; Goyal, N.; Chaudhary, V.; Wenzek, G.; Guzmán, F.; Grave, E.; Ott, M.; Zettlemoyer, L.; and Stoyanov, V. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of ACL*.

Deguchi, H.; Tamura, A.; and Ninomiya, T. 2019. Dependency-Based Self-Attention for Transformer NMT. In *Proceedings of RANLP*, 239–246.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL*.

Dryer, M. S.; and Haspelmath, M. 2013. The world atlas of language structures online .

Faruqui, M.; and Kumar, S. 2015. Multilingual Open Relation Extraction Using Cross-lingual Projection. In *Proceedings of NAACL*, 1351–1356.

Fu, T.-J.; Li, P.-H.; and Ma, W.-Y. 2019. GraphRel: Modeling Text as Relational Graphs for Joint Entity and Relation Extraction. In *Proceedings of ACL*, 1409–1418.

Han, R.; Ning, Q.; and Peng, N. 2019. Joint Event and Temporal Relation Extraction with Shared Representations and Structured Prediction. In *Proceedings of EMNLP*, 434–444.

Han, R.; Zhou, Y.; and Peng, N. 2020. Domain Knowledge Empowered Structured Neural Net for End-to-End Event Temporal Relation Extraction. In *Proceedings of EMNLP*, 5717–5729.

Hochreiter, S.; and Schmidhuber, J. 1997. Long short-term memory. *Neural computation* 9(8): 1735–1780.

Hong, Y.; Zhang, J.; Ma, B.; Yao, J.; Zhou, G.; and Zhu, Q. 2011. Using Cross-Entity Inference to Improve Event Extraction. In *Proceedings of NAACL*, 1127–1136.

Hsi, A.; Yang, Y.; Carbonell, J.; and Xu, R. 2016. Leveraging Multilingual Training for Limited Resource Event Extraction. In *Proceedings of COLING*, 1201–1210.

Huang, K.-H.; Yang, M.; and Peng, N. 2020. Biomedical Event Extraction with Hierarchical Knowledge Graphs. In *Findings of ACL: EMNLP 2020*.

Huang, L.; Ji, H.; Cho, K.; Dagan, I.; Riedel, S.; and Voss, C. 2018. Zero-Shot Transfer Learning for Event Extraction. In *Proceedings of ACL*, 2160–2170.

Ji, H.; and Grishman, R. 2008. Refining Event Extraction through Cross-Document Inference. In *Proceedings of ACL*, 254–262.

Joulin, A.; Bojanowski, P.; Mikolov, T.; Jégou, H.; and Grave, E. 2018. Loss in Translation: Learning Bilingual Word Mapping with a Retrieval Criterion. In *Proceedings of EMNLP*, 2979–2984.

Kim, S.; Jeong, M.; Lee, J.; and Lee, G. G. 2010a. A cross-lingual annotation projection approach for relation detection. In *Proceedings of COLING*, 564–571.

Kim, S.; Jeong, M.; Lee, J.; and Lee, G. G. 2010b. A Cross-lingual Annotation Projection Approach for Relation Detection. In *Proceedings of COLING*, 564–571.

Kim, S.; and Lee, G. G. 2012. A Graph-based Cross-lingual Projection Approach for Weakly Supervised Relation Extraction. In *Proceedings of ACL*, 48–53.

Kipf, T. N.; and Welling, M. 2017. Semi-supervised classification with graph convolutional networks. In *ICLR*.

Klein, G.; Kim, Y.; Deng, Y.; Senellart, J.; and Rush, A. 2017. OpenNMT: Open-Source Toolkit for Neural Machine Translation. In *Proceedings of ACL 2017, System Demo*.

Li, Q.; and Ji, H. 2014. Incremental Joint Extraction of Entity Mentions and Relations. In *Proceedings of ACL*, 402–412.

Li, Q.; Ji, H.; Hong, Y.; and Li, S. 2014. Constructing Information Networks Using One Single Model. In *Proceedings of EMNLP*, 1846–1851.

Li, Q.; Ji, H.; and Huang, L. 2013. Joint Event Extraction via Structured Prediction with Global Features. In *Proceedings of ACL*, 73–82.

Liao, S.; and Grishman, R. 2010. Using Document Level Cross-Event Inference to Improve Event Extraction. In *Proceedings of ACL*, 789–797.

Liao, S.; and Grishman, R. 2011. Acquiring Topic Features to improve Event Extraction: in Pre-selected and Balanced Collections. In *Proceedings of RANLP*, 9–16.

Lin, Y.; Liu, Z.; and Sun, M. 2017. Neural Relation Extraction with Multi-lingual Attention. In *Proceedings of ACL*.

Lin, Y.; Liu, Z.; Sun, M.; Liu, Y.; and Zhu, X. 2015. Learning Entity and Relation Embeddings for Knowledge Graph Completion. In *Proceedings of AAAI*, 2181–2187.

Liu, J.; Chen, Y.; Liu, K.; and Zhao, J. 2018. Event detection via gated multilingual attention mechanism. In *Proceedings of AAAI*, 4865–4872.

Liu, J.; Chen, Y.; Liu, K.; and Zhao, J. 2019. Neural Cross-Lingual Event Detection with Minimal Parallel Resources. In *Proceedings of EMNLP-IJCNLP*, 738–748.

Liu, X.; Luo, Z.; and Huang, H. 2018. Jointly Multiple Events Extraction via Attention-based Graph Information Aggregation. In *Proceedings of EMNLP*, 1247–1256.

Lu, W.; and Nguyen, T. H. 2018. Similar but not the Same: Word Sense Disambiguation Improves Event Detection via Neural Representation Matching. In *Proceedings of EMNLP*, 4822–4828.

Meng, T.; Peng, N.; and Chang, K.-W. 2019. Target Language-Aware Constrained Inference for Cross-lingual Dependency Parsing. In *Proceedings of EMNLP-IJCNLP*, 1117–1128.

Miwa, M.; and Bansal, M. 2016. End-to-End Relation Extraction using LSTMs on Sequences and Tree Structures. In *Proceedings of ACL*, 1105–1116.

Nguyen, T. H.; Cho, K.; and Grishman, R. 2016. Joint Event Extraction via Recurrent Neural Networks. In *Proceedings of NAACL*, 300–309.

Nguyen, T. H.; and Grishman, R. 2015. Event Detection and Domain Adaptation with Convolutional Neural Networks. In *Proceedings of EMNLP-IJCNLP*, 365–371.

Nguyen, T. H.; and Grishman, R. 2018. Graph convolutional networks with argument-aware pooling for event detection. In *Proceedings of AAAI*.

Ni, J.; and Florian, R. 2019. Neural Cross-Lingual Relation Extraction Based on Bilingual Word Embedding Mapping. In *Proceedings of EMNLP-IJCNLP*, 399–409.

Pawlik, M.; and Augsten, N. 2015. Efficient computation of the tree edit distance. *ACM Transactions on Database Systems (TODS)* 40(1): 1–40.

Pawlik, M.; and Augsten, N. 2016. Tree edit distance: Robust and memory-efficient. *Information Systems* 157–173.

Peng, N.; Poon, H.; Quirk, C.; Toutanova, K.; and Yih, W.-t. 2017. Cross-sentence N-ary Relation Extraction with Graph LSTMs. *Transactions of ACL* .

Qian, L.; Hui, H.; Hu, Y.; Zhou, G.; and Zhu, Q. 2014. Bilingual Active Learning for Relation Classification via Pseudo Parallel Corpora. In *Proceedings of ACL*, 582–592.

Rasooli, M. S.; and Collins, M. 2019. Low-Resource Syntactic Transfer with Unsupervised Source Reordering. In *Proceedings of NAACL*, 3845–3856.

Shaw, P.; Uszkoreit, J.; and Vaswani, A. 2018. Self-Attention with Relative Position Representations. In *Proceedings of NAACL*, 464–468.

Straka, M.; and Straková, J. 2017. Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, 88–99.

Subburathinam, A.; Lu, D.; Ji, H.; May, J.; Chang, S.-F.; Sil, A.; and Voss, C. 2019. Cross-lingual Structure Transfer for Relation and Event Extraction. In *Proceedings of EMNLP-IJCNLP*, 313–325.

Sun, C.; Gong, Y.; Wu, Y.; Gong, M.; Jiang, D.; Lan, M.; Sun, S.; and Duan, N. 2019. Joint Type Inference on Entities and Relations via Graph Convolutional Networks. In *Proceedings of ACL*, 1361–1370.

Tang, H.; Ji, D.; Li, C.; and Zhou, Q. 2020. Dependency Graph Enhanced Dual-transformer Structure for Aspect-based Sentiment Classification. In *Proceedings of ACL*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *NeurIPS*, 5998–6008.

Walker, C.; Strassel, S.; Medero, J.; and Maeda, K. 2006. ACE 2005 multilingual training corpus. *Linguistic Data Consortium, Philadelphia* 57.

Wang, X.; Han, X.; Lin, Y.; Liu, Z.; and Sun, M. 2018. Adversarial Multi-lingual Neural Relation Extraction. In *Proceedings of COLING*, 1156–1166.

Wang, X.; Tu, Z.; Wang, L.; and Shi, S. 2019. Self-Attention with Structural Position Representations. In *Proceedings of EMNLP-IJCNLP*, 1403–1409.

Yang, B.; and Mitchell, T. M. 2016. Joint Extraction of Events and Entities within a Document Context. In *Proceedings of NAACL*, 289–299.

Zeng, D.; Liu, K.; Lai, S.; Zhou, G.; and Zhao, J. 2014. Relation Classification via Convolutional Deep Neural Network. In *Proceedings of COLING*, 2335–2344.

Zhang, C.; Li, Q.; and Song, D. 2019. Aspect-based Sentiment Classification with Aspect-specific Graph Convolutional Networks. In *Proceedings of EMNLP-IJCNLP*.

Zhang, J.; Zhou, W.; Hong, Y.; Yao, J.; and Zhang, M. 2018. Using entity relation to improve event detection via attention mechanism. In *CCF International Conference on NLPCC*.

Zhang, T.; Ji, H.; and Sil, A. 2019. Joint entity and event extraction with generative adversarial imitation learning. *Data Intelligence* 99–120.

Zhang, Y.; Qi, P.; and Manning, C. D. 2018. Graph Convolution over Pruned Dependency Trees Improves Relation Extraction. In *Proceedings of EMNLP*, 2205–2215.

Zhu, Z.; Li, S.; Zhou, G.; and Xia, R. 2014. Bilingual Event Extraction: a Case Study on Trigger Type Determination. In *Proceedings of ACL*, 842–847.

Zou, B.; Xu, Z.; Hong, Y.; and Zhou, G. 2018. Adversarial Feature Adaptation for Cross-lingual Relation Classification. In *Proceedings of COLING*, 437–448.