

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

Wasi Uddin Ahmad¹, Nanyun Peng^{1,2}, Kai-Wei Chang¹

{wasiahmad, violetpeng, kwchang}@cs.ucla.edu

¹University of California, Los Angeles, ²University of Southern California

Abstract

Prevalent approaches in cross-lingual relation and event extraction use graph convolutional networks (GCNs) with universal dependency parses to learn language-agnostic representations such that models trained on one language can be applied to other languages. However, GCNs lack in modeling long-range dependencies or disconnected words in the dependency tree. To address this challenge, we propose to utilize the self-attention mechanism where we explicitly fuse structural information to learn the dependencies between words at different syntactic distances. We introduce GATE, a **Graph Attention Transformer Encoder**, and test its cross-lingual transferability on relation and event extraction tasks. We perform rigorous experiments on the widely used 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. Prevailing approaches in cross-lingual learning for relation and event extraction requires learning a universal encoder that embeds relation and event mentions in a sentence into contextualized representations. Recent works (Huang et al., 2018; Subburathinam

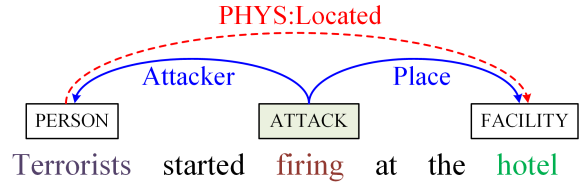


Figure 1: A relation (red dashed) between two entities and an event of type *Attack* (triggered by “firing”) including two arguments (blue) are highlighted.

et al., 2019) suggested embedding universal dependency structure into contextual representations to improve cross-lingual transfer for IE.

There are a couple of advantages of utilizing the dependency structure. First, the syntactic distance between two words¹ 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 2, respectively. Therefore, encoding syntax structure helps in capturing long-range dependencies (Liu et al., 2018b). Second, languages have different word order, e.g., adjective precedes noun (“green apple”) in English but follows in French (“pomme rouge”). As a result, processing sentences sequentially may suffer from the word order difference issue (Ahmad et al., 2019). However, modeling sentence structure can mitigate the problem and thus improves cross-lingual transfer (Liu et al., 2019).

A common way to derive structured representations for cross-lingual NLP tasks is the use of universal dependency parses². A large pool of recent works in IE (Liu et al., 2018b; Zhang et al., 2018b; Subburathinam et al., 2019; Fu et al., 2019; Sun et al., 2019; Liu et al., 2019) employed Graph Convolutional Networks (GCNs) (Kipf and Welling,

¹The shortest path distance in the dependency graph.

²<https://universaldependencies.org/>

2017) to learn sentence representations based on their universal dependency parses. 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 8). 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 et al., 2019a; 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-aware self-attention and found effective for machine translation (Deguchi et al., 2019; Bugliarello and Okazaki, 2020). The key idea is to allow attention between connected words as in the dependency tree and gradually aggregate information across layers. However, IE tasks are relatively low-resourced and thus stacking more layers is not feasible. Hence, we propose to allow attention between all words but use the pairwise syntactic distances as a parameter to retrofit the attention weights. Besides, our preliminary analysis indicates that syntactic distance between entities could characterize certain relation and event types.³ This further motivates us to model the pairwise distance between words in the self-attention mechanism.

In this work, we introduce a **Graph Attention Transformer Encoder (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 at different heads in the multi-head attention structure

based on syntactic distances, which allows to learn the correlation between different mention types and the target label space.

We conduct experiments on cross-lingual transfer among English, Chinese, and Arabic languages using ACE 2005 benchmark (Walker et al., 2006). The experimental results demonstrate that GATE outperforms three recently proposed relation and event extraction methods by a notable margin. We perform a thorough ablation and analysis, and our findings show that GATE is less sensitive towards source language 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, their notations, and define the problem as well as 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 $r \in R \cup \{\text{None}\}$ 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 an `Attack` event.

Event argument role labeling (EARL) is defined as predicting whether words or phrases participate in events (arguments) 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 $r \in R \cup \{\text{None}\}$, 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.

³In 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.

⁴After tokenization, on average, ACE 2005 English and Arabic sentences have approx. 30 and 210 words, 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 (in any language) into a language-universal tree structure using an off-the-shelf universal dependency parser, e.g., UDPipe (Straka and Straková, 2017).
2. Embed the words in the sentence into a shared multilingual space. 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 produced by GATE.

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 and introduce the notations (§ 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.

3.1 Transformer Encoder

Unlike recent works (Zhang et al., 2018b; Subburathinam et al., 2019) that use GCNs (Kipf

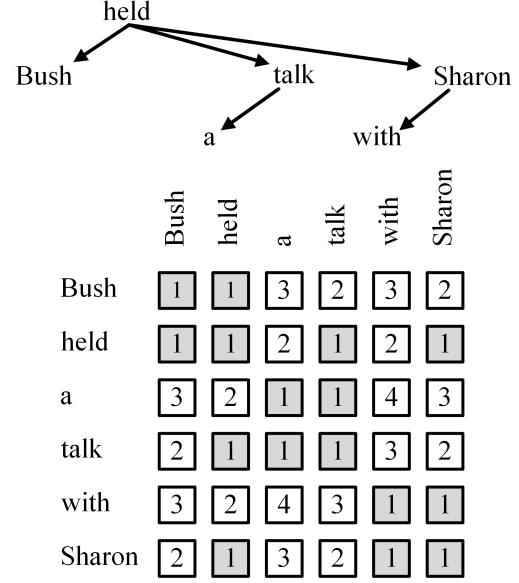


Figure 2: Distance matrix showing the shortest path distances between all pairs of words. The dependency arc direction is ignored while computing pairwise distances. The diagonal value is set to 1, indicating a self-loop. If we set the values in white cells (with value > 1) to 0, the distance matrix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

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, $x = [x_1, \dots, x_{|x|}]$ where $x_i \in \mathbb{R}^d$ are packed into a matrix $H^0 = [x_1, \dots, x_{|x|}]$. Then an L -layer Transformer Encoder $H^l = \text{Transformer}_l(H^{l-1})$, $l \in [1, L]$ takes H^0 as input and generates different levels of latent representations $H^l = [h_1^l, \dots, h_{|x|}^l]$. 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 $H^{l-1} \in \mathbb{R}^{|x| \times d_{\text{model}}}$ is first linearly projected to queries Q , keys K , and values V using parameter matrices $W_l^Q, W_l^K \in \mathbb{R}^{d_{\text{model}} \times d_k}$ and $W_l^V \in \mathbb{R}^{d_{\text{model}} \times d_v}$, respectively.

$$Q = H^{l-1}W_l^Q, K = H^{l-1}W_l^K, V = H^{l-1}W_l^V.$$

Finally, the output of a self-attention head A_l is computed as follows.

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

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

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

By default, the matrix M is a *zero-matrix*. In the next section, we discuss how we manipulate the mask matrix M to model syntactic depth and distance information when encoding a sentence.

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).

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

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 $\{\text{PER}, \text{ORG}\}$ 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.

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

Here, softmax produces an attention matrix $P \in \mathbb{R}^{|x| \times |x|}$ 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 follows.

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

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 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 $h_1^l, \dots, h_{|x|}^l$ where $h_i^l \in \mathbb{R}^{d_{\text{model}}}$. 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 $e_s = [h_{bs}^l, \dots, h_{es}^l]$ and $e_o = [h_{bo}^l, \dots, h_{eo}^l]$, we obtain single vector representations \hat{e}_s and \hat{e}_o by performing max-pooling. Following Zhang et al. (2018b); Subburathinam et al. (2019), we also obtain a vector representation for the sentence, \hat{s} by applying max-pooling over $[h_1^l, \dots, h_{|x|}^l]$ 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 . During training, we optimize the relation extractor on the following objective function.

$$\mathcal{L}_r = - \sum_{s=1}^N \sum_{o=1}^N \sum_{r \in R} y_{so}^r \log(\sigma(\mathbf{U}^r \cdot [\hat{e}_s; \hat{e}_o; \hat{s}])),$$

where N is the number of entity mentions, R is a pre-defined set of relation types, y_{so}^r is a binary indicator of whether e_s and e_o holds a relation in the ground truth, \mathbf{U}^r is a weight matrix, and σ is the Sigmoid function.

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 $e_a = [h_{ba}^l, \dots, h_{ea}^l]$ for an event trigger $e_t = [h_{bt}^l, \dots, h_{et}^l]$ in sentence $s = [h_1^l, \dots, h_{|x|}^l]$, 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 output the argument role label. The event argument role labeler is trained on the following objective function.

$$\mathcal{L}_a = \sum_{t=1}^N \sum_{a=1}^{C_t} \sum_{r \in R} y_{ta}^r \log(\sigma(U^a \cdot [\hat{e}_t; \hat{e}_a; \hat{s}])),$$

where C_t is the number of argument candidates for the t -th event mention, and other notations are similar as that for relation extractor’s objective.

4 Experiment

In this section, we detail our experiment on cross-lingual relation extraction and event argument role labeling to evaluate our proposed approach.

4.1 Setup

Dataset and Evaluation Criteria We conduct experiments using 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 followed their preprocessing steps. We refer the readers to Subburathinam et al. (2019) for the dataset preprocessing details.

Following the previous works (Ji and Grishman, 2008; Li et al., 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 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 is not publicly available at the time this research is conducted. Therefore, we implemented 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. In this baseline, Graph Convolutional Networks (GCNs) (Kipf and Welling, 2017) is used as the contextual encoder to cope with syntactic differences between source and target languages. We use multilingual word embeddings (Joulin et al., 2018) as the continuous representations of tokens and do not use any additional language-universal features.⁵ Since this baseline specifically depends on the target language, we train this baseline for each combination of source and target languages.

- **CL_GCN** (Subburathinam et al., 2019) uses GCN 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) along with language-universal feature embeddings including part-of-speech (POS) tag embedding, dependency relation label embedding, and entity type embedding. 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 recurrent 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.

Implementation Details To embed words into vector representations, we use multilingual BERT (M-BERT) (Devlin et al., 2019). We do not fine-tune M-BERT, only use it as a feature extractor.⁶ We use the universal part-of-speech (POS)

⁵Due to the design principle of Liu et al. (2019), we cannot use multilingual contextual encoders in CL_Trans_GCN.

⁶We could not fine-tune M-BERT because it was computationally infeasible to perform for Chinese and Arabic languages. After performing tokenization for M-BERT, the

Model	Event Argument Role Labeling						Relation Extraction					
	En	En	Zh	Zh	Ar	Ar	En	En	Zh	Zh	Ar	Ar
	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓
	Zh	Ar	En	Ar	En	Zh	Zh	Ar	En	Ar	En	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
GATE	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 1: 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}	{En, Ar}	{Zh, Ar}
	↓	↓	↓
	Ar	Zh	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
GATE	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
GATE	67.0	57.9	74.1

Table 2: 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.

tags⁷, dependency relation labels⁸, 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 Appendix C). We provide details of the dataset, hyperparameters, and training in Appendix A and B.

We implement all the baselines and our approach based on publicly available implementation of Zhang et al. (2018b)⁹ and OpenNMT (Klein et al., 2017). We used transformers¹⁰ to extract M-BERT and XLM-RoBERTa features. We

length of the longest Chinese and Arabic example was 1097 and 2261, respectively.

⁷<https://universaldependencies.org/u/pos/>

⁸<https://universaldependencies.org/u/dep/>

⁹<https://github.com/qipeng/gcn-over-pruned-trees>

¹⁰<https://github.com/huggingface/transformers>

trained all the models three times with different initialization and reported average scores.

4.2 Main Results

We compare our proposed model, GATE with three baseline approaches on event argument role labeling (EARL) and relation extraction (RE) tasks, and the results are presented in Table 1 and 2.

Single-source transfer In the single-source transfer setting, all the models are individually trained on *one* language (source), e.g., English and then directly evaluated on the other two languages (target), e.g., Chinese and Arabic. Table 1 shows that GATE outperforms all the three baselines by a large margin on EARL. On RE, GATE is competitive to CL_RNN if not surpassing its result. To our surprise, CL_RNN performs better than CL_GCN in most settings, although CL_RNN uses a BiLSTM that is not suitable to transfer across syntactically different languages (Ahmad et al., 2019). However, we noted that GCNs lack in capturing long-range dependencies, which is crucial for the tasks at hand. As a result, CL_RNN outperforms CL_GCN in most settings. In comparison, due to modeling distance-based pairwise relationships among words, GATE excels in cross-lingual transfer in both the tasks.

Multi-source transfer In multi-source transfer setting, 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 in comparison to the single-source transfer setting. In Table 2, we see GATE outperforms the baselines in multi-source transfer settings too, except on RE for the source: {English, Arabic} and target: Chinese language setting. The overall result indicates that GATE learns better transferable representations than the baseline approaches.

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 3: 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 (in EARL) and 5 (in RE).

4.3 Analysis and Discussion

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 notion of encoding the dependency structure to emphasize the promise of modeling syntactic distances.

To this end, we compare GATE with [Wang et al. \(2019\)](#) that proposed structural position encoding using the dependency structure of sentences. Results are presented in Table 3. 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 8 in Appendix). As a result, a model trained on English may learn only to attend closer tokens, thus fails on Arabic.

Moreover, we anticipate that different order of subject and verb in English and Arabic¹¹ causes [Wang et al. \(2019\)](#) to transfer poorly on the EARL (as event triggers are mostly verbs) task. To verify our anticipation, we modify the relative structural position encoding ([Wang et al., 2019](#)) by dropping the directional information ([Ahmad et al., 2019](#)),

¹¹ 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 4: 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.

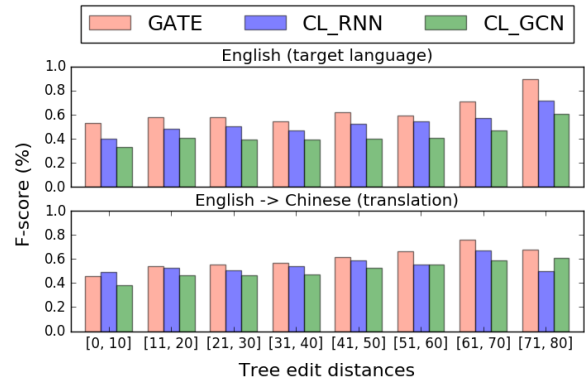


Figure 3: Models trained on the Chinese language perform on event argument role labeling task in English and their parallel Chinese sentences. The parallel sentences have the same meaning but different syntax structure. To quantify the structural difference between the two parallel sentences, we compute the tree edit distances. The language-universal features are not used in this experiments, so the models only rely on multilingual word representations.

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.

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 a model 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 Chinese test set examples into English using Google Cloud

Translate. We detail the process in Appendix E. 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 structure of an English and its Chinese translation sentences, we compute *edit distance* between two dependency tree structures using the APTED¹² algorithm (Pawlik and Augsten, 2015, 2016).

The results are presented in Table 4. We can see that CL_GCN and CL_RNN predicts the target label correctly for more examples if translated (Chinese) sentences are provided, instead of 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. In Figure 3, we illustrate how do 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 than in source language. The overall findings from this experiment signal that GATE is less sensitive towards source language characteristics, 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 on language-universal features and sources of word features to examine their individual impact on cross-lingual transfer. The results are presented in Table 5 and 6. Overall, we found that M-BERT and XLM-RoBERTa 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 6 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.

To study the impact of syntax-based self-attention, we compare GATE with the standard self-attention mechanism (Vaswani et al., 2017),

¹²<https://pypi.org/project/apted/>

Source of word features	Chinese	Arabic
Event Argument Role Labeling		
Multilingual Word Embedding	35.9	43.7
M-BERT	57.1	54.8
XLM-RoBERTa	51.8	61.7
Relation Extraction		
Multilingual Word Embedding	41.0	54.9
M-BERT	55.1	66.8
XLM-RoBERTa	51.4	68.1

Table 5: Contribution of multilingual word embeddings (Joulin et al., 2018), M-BERT (Devlin et al., 2019), and XLM-RoBERTa (Conneau et al., 2020) as a source of word features; using English as source and Chinese, Arabic as the target languages, respectively.

Input features	Chinese	Arabic
Event Argument Role Labeling		
Multilingual BERT	52.5	47.4
+ Part-of-speech embedding	49.3	47.5
+ Dep. relation embedding	49.7	51.0
+ Entity embedding	57.8	60.2
Relation Extraction		
Multilingual BERT	44.0	49.7
+ Part-of-speech embedding	44.1	47.0
+ Dep. relation embedding	48.6	47.0
+ Entity embedding	56.3	63.0

Table 6: Ablation on the language-universal features in GATE (F-score (%)); using English as source and Chinese, Arabic as the target languages, respectively.

and the results are presented in Appendix D. The significant improvements in most transfer directions validates the conjecture that use of universal syntax structure helps in cross-lingual transfer. We further perform ablation on GATE to examine how much distance based attention benefits the IE tasks (paying more attention to tokens that are closer and less attention to tokens that are faraway in the parse tree). We observed consistent improvements (notably on the event argument role labeling) regardless of transfer directions. This finding corroborates our hypothesis that distance based attention modeling helps IE tasks.

5 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, 2011; Hong et al., 2011; Li et al.,

2013; Li and Ji, 2014) or distributional features (Nguyen et al., 2016; Miwa and Bansal, 2016; Liu et al., 2018a; Zhang et al., 2018a; Lu and Nguyen, 2018; Chen et al., 2015; Nguyen and Grishman, 2015a,b; Zeng et al., 2014; Nguyen and Grishman, 2018; Zhang et al., 2018b; Subburathinam et al., 2019; Liu et al., 2019) from a large number of annotations. Joint learning or inference (Bekoulis et al., 2018; Li et al., 2014; Zhang et al., 2019b; Liu et al., 2018b; Nguyen et al., 2016; Yang and Mitchell, 2016) 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 to improve information extraction is also explored in prior works (Lin et al., 2017; Wang et al., 2018; Liu et al., 2018a).

In contrast to these approaches, Subburathinam et al. (2019); Liu et al. (2019) proposed to learn multi-lingual structural representations and employed graph convolutional networks (GCNs) (Kipf and Welling, 2017) to learn such representations. In NLP literature, GCN has been successfully used for many tasks, including sentence classification (Yao et al., 2019), semantic role labeling (Marcheggiani and Titov, 2017), named entity recognition (Cetoli et al., 2017), dependency parsing (Ji et al., 2019), event detection (Nguyen and Grishman, 2018), and relation extraction (Zhang et al., 2018b; Subburathinam et al., 2019; Liu et al., 2019).

However, GCN does not embed finer-grained syntactic information of sentences. To overcome the limitation, we use the multi-head attention mechanism (Vaswani et al., 2017), where we use the syntactic structure to control which sentence words should be attended while encoding the sentence into contextualized representations.

6 Conclusion

In this paper, we proposed to model fine-grained syntactic structural information based on the dependency parse of a sentence. We developed a Graph Attention Transformer Encoder (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.

References

- Wasi Ahmad, Zhisong Zhang, Xuezhe Ma, Eduard Hovy, Kai-Wei Chang, and Nanyun Peng. 2019. [On difficulties of cross-lingual transfer with order differences: A case study on dependency parsing](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2440–2452, Minneapolis, Minnesota. Association for Computational Linguistics.
- David Ahn. 2006. [The stages of event extraction](#). In *Proceedings of the Workshop on Annotating and Reasoning about Time and Events*, pages 1–8, Sydney, Australia. Association for Computational Linguistics.
- Giannis Bekoulis, Johannes Deleu, Thomas Demeester, and Chris Develder. 2018. [Adversarial training for multi-context joint entity and relation extraction](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2830–2836, Brussels, Belgium. Association for Computational Linguistics.
- Emanuele Bugliarelli and Naoaki Okazaki. 2020. [Enhancing machine translation with dependency-aware self-attention](#). In *Proceedings of ACL*, pages 1618–1627.
- Alberto Cetoli, Stefano Bragaglia, Andrew O’Harney, and Marc Sloan. 2017. [Graph convolutional networks for named entity recognition](#). In *Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories*, pages 37–45, Prague, Czech Republic.
- Yubo Chen, Liheng Xu, Kang Liu, Daojian Zeng, and Jun Zhao. 2015. [Event extraction via dynamic multi-pooling convolutional neural networks](#). In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 167–176, Beijing, China. Association for Computational Linguistics.
- Zheng Chen and Heng Ji. 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*, pages 66–74, Boulder, Colorado. Association for Computational Linguistics.
- Zi-Yuan Chen, Chih-Hung Chang, Yi-Pei Chen, Jijnasa Nayak, and Lun-Wei Ku. 2019. [UHop: An unrestricted-hop relation extraction framework for](#)

- knowledge-based question answering. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 345–356, Minneapolis, Minnesota. Association for Computational Linguistics.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.
- Hiroyuki Deguchi, Akihiro Tamura, and Takashi Nomiya. 2019. Dependency-based self-attention for transformer NMT. In *Proceedings of RANLP*, pages 239–246.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Matthew S. Dryer and Martin Haspelmath, editors. 2013. *WALS Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.
- Manaal Faruqui and Shankar Kumar. 2015. Multilingual open relation extraction using cross-lingual projection. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1351–1356, Denver, Colorado. Association for Computational Linguistics.
- Tsu-Jui Fu, Peng-Hsuan Li, and Wei-Yun Ma. 2019. GraphRel: Modeling text as relational graphs for joint entity and relation extraction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1409–1418, Florence, Italy. Association for Computational Linguistics.
- Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.
- Yu Hong, Jianfeng Zhang, Bin Ma, Jianmin Yao, Guodong Zhou, and Qiaoming Zhu. 2011. Using cross-entity inference to improve event extraction. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1127–1136, Portland, Oregon, USA. Association for Computational Linguistics.
- Andrew Hsi, Yiming Yang, Jaime Carbonell, and Ruochen Xu. 2016. Leveraging multilingual training for limited resource event extraction. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1201–1210, Osaka, Japan. The COLING 2016 Organizing Committee.
- Lifu Huang, Heng Ji, Kyunghyun Cho, Ido Dagan, Sebastian Riedel, and Clare Voss. 2018. Zero-shot transfer learning for event extraction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2160–2170, Melbourne, Australia. Association for Computational Linguistics.
- Heng Ji and Ralph Grishman. 2008. Refining event extraction through cross-document inference. In *Proceedings of ACL-08: HLT*, pages 254–262, Columbus, Ohio. Association for Computational Linguistics.
- Tao Ji, Yuanbin Wu, and Man Lan. 2019. Graph-based dependency parsing with graph neural networks. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2475–2485, Florence, Italy. Association for Computational Linguistics.
- Armand Joulin, Piotr Bojanowski, Tomas Mikolov, Hervé Jégou, and Edouard Grave. 2018. Loss in translation: Learning bilingual word mapping with a retrieval criterion. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2984, Brussels, Belgium. Association for Computational Linguistics.
- Seokhwan Kim, Minwoo Jeong, Jonghoon Lee, and Gary Geunbae Lee. 2010a. A cross-lingual annotation projection approach for relation detection. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 564–571. Association for Computational Linguistics.
- Seokhwan Kim, Minwoo Jeong, Jonghoon Lee, and Gary Geunbae Lee. 2010b. A cross-lingual annotation projection approach for relation detection. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 564–571, Beijing, China. Coling 2010 Organizing Committee.
- Seokhwan Kim and Gary Geunbae Lee. 2012. A graph-based cross-lingual projection approach for weakly supervised relation extraction. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 48–53, Jeju Island, Korea. Association for Computational Linguistics.
- Thomas N Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*.

- Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. [OpenNMT: Open-source toolkit for neural machine translation](#). In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada. Association for Computational Linguistics.
- Qi Li and Heng Ji. 2014. [Incremental joint extraction of entity mentions and relations](#). In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 402–412, Baltimore, Maryland. Association for Computational Linguistics.
- Qi Li, Heng Ji, Yu Hong, and Sujian Li. 2014. [Constructing information networks using one single model](#). In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1846–1851, Doha, Qatar. Association for Computational Linguistics.
- Qi Li, Heng Ji, and Liang Huang. 2013. [Joint event extraction via structured prediction with global features](#). In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 73–82, Sofia, Bulgaria. Association for Computational Linguistics.
- Shasha Liao and Ralph Grishman. 2010. [Using document level cross-event inference to improve event extraction](#). In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 789–797, Uppsala, Sweden. Association for Computational Linguistics.
- Shasha Liao and Ralph Grishman. 2011. [Acquiring topic features to improve event extraction: in pre-selected and balanced collections](#). In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, pages 9–16, Hissar, Bulgaria. Association for Computational Linguistics.
- Yankai Lin, Zhiyuan Liu, and Maosong Sun. 2017. [Neural relation extraction with multi-lingual attention](#). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 34–43, Vancouver, Canada. Association for Computational Linguistics.
- Yankai Lin, Zhiyuan Liu, Maosong Sun, Yang Liu, and Xuan Zhu. 2015. Learning entity and relation embeddings for knowledge graph completion. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, AAAI’15*, page 2181–2187. AAAI Press.
- Jian Liu, Yubo Chen, Kang Liu, and Jun Zhao. 2018a. Event detection via gated multilingual attention mechanism. In *Thirty-Second AAAI Conference on Artificial Intelligence*.
- Jian Liu, Yubo Chen, Kang Liu, and Jun Zhao. 2019. [Neural cross-lingual event detection with minimal parallel resources](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 738–748, Hong Kong, China. Association for Computational Linguistics.
- Xiao Liu, Zhunchen Luo, and Heyan Huang. 2018b. [Jointly multiple events extraction via attention-based graph information aggregation](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1247–1256, Brussels, Belgium. Association for Computational Linguistics.
- Weiyi Lu and Thien Huu Nguyen. 2018. [Similar but not the same: Word sense disambiguation improves event detection via neural representation matching](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4822–4828, Brussels, Belgium. Association for Computational Linguistics.
- Diego Marcheggiani and Ivan Titov. 2017. [Encoding sentences with graph convolutional networks for semantic role labeling](#). In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1506–1515, Copenhagen, Denmark. Association for Computational Linguistics.
- Makoto Miwa and Mohit Bansal. 2016. [End-to-end relation extraction using LSTMs on sequences and tree structures](#). In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1105–1116, Berlin, Germany. Association for Computational Linguistics.
- Thien Huu Nguyen, Kyunghyun Cho, and Ralph Grishman. 2016. [Joint event extraction via recurrent neural networks](#). In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 300–309, San Diego, California. Association for Computational Linguistics.
- Thien Huu Nguyen and Ralph Grishman. 2015a. [Event detection and domain adaptation with convolutional neural networks](#). In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 365–371, Beijing, China. Association for Computational Linguistics.
- Thien Huu Nguyen and Ralph Grishman. 2015b. [Relation extraction: Perspective from convolutional neural networks](#). In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pages 39–48, Denver, Colorado. Association for Computational Linguistics.
- Thien Huu Nguyen and Ralph Grishman. 2018. Graph convolutional networks with argument-aware pooling for event detection. In *Thirty-second AAAI conference on artificial intelligence*.

- Jian Ni and Radu Florian. 2019. [Neural cross-lingual relation extraction based on bilingual word embedding mapping](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 399–409, Hong Kong, China. Association for Computational Linguistics.
- Mateusz Pawlik and Nikolaus Augsten. 2015. Efficient computation of the tree edit distance. *ACM Transactions on Database Systems (TODS)*, 40(1):1–40.
- Mateusz Pawlik and Nikolaus Augsten. 2016. Tree edit distance: Robust and memory-efficient. *Information Systems*, 56:157–173.
- Longhua Qian, Haotian Hui, Ya’nan Hu, Guodong Zhou, and Qiaoming Zhu. 2014. [Bilingual active learning for relation classification via pseudo parallel corpora](#). In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 582–592, Baltimore, Maryland. Association for Computational Linguistics.
- Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. 2018. [Self-attention with relative position representations](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 464–468, New Orleans, Louisiana. Association for Computational Linguistics.
- Milan Straka and Jana Straková. 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*, pages 88–99, Vancouver, Canada. Association for Computational Linguistics.
- Ananya Subburathinam, Di Lu, Heng Ji, Jonathan May, Shih-Fu Chang, Avirup Sil, and Clare Voss. 2019. [Cross-lingual structure transfer for relation and event extraction](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 313–325, Hong Kong, China. Association for Computational Linguistics.
- Changzhi Sun, Yeyun Gong, Yuanbin Wu, Ming Gong, Daxin Jiang, Man Lan, Shiliang Sun, and Nan Duan. 2019. [Joint type inference on entities and relations via graph convolutional networks](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1361–1370, Florence, Italy. Association for Computational Linguistics.
- Hao Tang, Donghong Ji, Chenliang Li, and Qiji Zhou. 2020. Dependency graph enhanced dual-transformer structure for aspect-based sentiment classification. In *Proceedings of ACL*, pages 6578–6588.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.
- Christopher Walker, Stephanie Strassel, Julie Medero, and Kazuaki Maeda. 2006. [Ace 2005 multilingual training corpus](#). *Linguistic Data Consortium, Philadelphia*, 57.
- Xiaozhi Wang, Xu Han, Yankai Lin, Zhiyuan Liu, and Maosong Sun. 2018. [Adversarial multi-lingual neural relation extraction](#). In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1156–1166, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
- Xing Wang, Zhaopeng Tu, Longyue Wang, and Shuming Shi. 2019. [Self-attention with structural position representations](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1403–1409, Hong Kong, China. Association for Computational Linguistics.
- Bishan Yang and Tom M. Mitchell. 2016. [Joint extraction of events and entities within a document context](#). In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 289–299, San Diego, California. Association for Computational Linguistics.
- Liang Yao, Chengsheng Mao, and Yuan Luo. 2019. Graph convolutional networks for text classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7370–7377.
- Daojian Zeng, Kang Liu, Siwei Lai, Guangyou Zhou, and Jun Zhao. 2014. [Relation classification via convolutional deep neural network](#). In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 2335–2344, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.
- Chen Zhang, Qiuchi Li, and Dawei Song. 2019a. Aspect-based sentiment classification with aspect-specific graph convolutional networks. In *Proceedings of EMNLP-IJCNLP*, pages 4568–4578.
- Jingli Zhang, Wenxuan Zhou, Yu Hong, Jianmin Yao, and Min Zhang. 2018a. Using entity relation to improve event detection via attention mechanism. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 171–183. Springer.
- Tongtao Zhang, Heng Ji, and Avirup Sil. 2019b. Joint entity and event extraction with generative adversarial imitation learning. *Data Intelligence*, 1(2):99–120.

Yuhao Zhang, Peng Qi, and Christopher D. Manning. 2018b. [Graph convolution over pruned dependency trees improves relation extraction](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2205–2215, Brussels, Belgium. Association for Computational Linguistics.

Zhu Zhu, Shoushan Li, Guodong Zhou, and Rui Xia. 2014. [Bilingual event extraction: a case study on trigger type determination](#). In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 842–847, Baltimore, Maryland. Association for Computational Linguistics.

Bowei Zou, Zengzhuang Xu, Yu Hong, and Guodong Zhou. 2018. [Adversarial feature adaptation for cross-lingual relation classification](#). In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 437–448, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

A Dataset Details

We conduct experiments on the ACE 2005 dataset, which can be downloaded from [here](#). We list the dataset statistics in Table 7. In Table 8, we present the statistics of sequential and shortest path distances between relations mentions and event mentions and their arguments in ACE05.

	English	Chinese	Arabic
Relations Mentions	8,738	9,317	4,731
Event Mentions	5,349	3,333	2,270
Event Arguments	9,793	8,032	4,975

Table 7: Statistics of the ACE 2005 dataset.

Language	Sequential Distance			Structural Distance		
	English	Chinese	Arabic	English	Chinese	Arabic
Relation mentions	4.8	3.9	25.8	2.2	2.6	5.1
Event mentions and arguments	9.8	21.7	58.1	3.1	4.6	12.3

Table 8: Average sequential and structural (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.

B Hyper-parameter Details

We detail the hyper-parameters for all the baselines and our approach in Table 9.

Hyper-parameter	CL_Trans_GCN	CL_GCN	CL_RNN	GATE
word embedding size	300	768	768	768
part-of-speech embedding size	30	30	30	30
entity type embedding size	30	30	30	30
dependency relation embedding size	30	30	30	30
encoder type	GCN	GCN	BiLSTM	Self-Attention
encoder layers	2	2	1	1
encoder hidden size	200	200	300	512
pooling function	max-pool	max-pool	max-pool	max-pool
mlp layers	2	2	2	2
dropout	0.5	0.5	0.5	0.5
optimizer	Adam	SGD	Adam	SGD
learning rate	0.001	0.1	0.001	0.1
learning rate decay	0.9	0.9	0.9	0.9
decay start epoch	5	5	5	5
batch size	50	50	50	50
maximum gradient norm	5.0	5.0	5.0	5.0

Table 9: Hyper-parameters of CL_Trans_GCN ([Liu et al., 2019](#)), CL_GCN ([Subburathinam et al., 2019](#)), CL_RNN ([Ni and Florian, 2019](#)), and our approach, GATE.

C Tuning δ (shown in Eq. (3))

During our initial experiments, we observed that setting $\delta = \infty$ in four attention heads provide consistently better performances. We tune δ in the range $[1, 2, 4, 8]$ on the validation set based on the statistics of the shortest path distances between relations mentions and event mentions and their arguments in ACE05 (shown in Table 8). we set $\delta = [2, 2, 4, 4, \infty, \infty, \infty, \infty]$ and $\delta = [1, 1, 2, 2, \infty, \infty, \infty, \infty]$ for the event argument role labeling and relation extraction tasks, respectively, in all our experiments. This hyper-parameter choice provides us comparably better performances (on test sets), as shown in Table 10.

δ for Attention Heads	En \Rightarrow Zh	En \Rightarrow Ar	Zh \Rightarrow En	Zh \Rightarrow Ar	Ar \Rightarrow En	Ar \Rightarrow Zh	Avg.
Event Argument Role Labeling							
[1, 1, 1, 1, ∞ , ∞ , ∞ , ∞]	63.1	65.9	57.3	67.1	53.5	57.2	60.7
[2, 2, 2, 2, ∞ , ∞ , ∞ , ∞]	64.3	69.6	58.9	69.4	52.7	56.2	61.9
[4, 4, 4, 4, ∞ , ∞ , ∞ , ∞]	62.1	69.8	58.9	70.5	53.0	56.1	61.7
[8, 8, 8, 8, ∞ , ∞ , ∞ , ∞]	63.6	69.4	57.9	71.4	54.0	54.9	61.9
[1, 1, 2, 2, ∞ , ∞ , ∞ , ∞]	63.2	68.5	58.7	69.5	52.7	53.7	61.1
[2, 2, 4, 4, ∞ , ∞ , ∞ , ∞]	65.0	69.6	60.2	69.2	53.9	57.8	62.6
[4, 4, 8, 8, ∞ , ∞ , ∞ , ∞]	63.6	70.5	58.3	70.8	53.4	57.6	62.4
[1, 2, 4, 8, ∞ , ∞ , ∞ , ∞]	64.3	69.6	57.8	69.7	52.5	55.5	61.6
Relation Extraction							
[1, 1, 1, 1, ∞ , ∞ , ∞ , ∞]	54.8	63.7	70.7	62.3	69.8	50.6	62.0
[2, 2, 2, 2, ∞ , ∞ , ∞ , ∞]	55.1	64.1	70.4	59.4	68.7	50.2	61.3
[4, 4, 4, 4, ∞ , ∞ , ∞ , ∞]	55.5	64.5	71.6	61.2	68.7	51.5	62.2
[8, 8, 8, 8, ∞ , ∞ , ∞ , ∞]	55.5	65.5	71.1	61.7	67.5	53.4	62.5
[1, 1, 2, 2, ∞ , ∞ , ∞ , ∞]	56.4	63.5	70.4	63.1	69.4	51.9	62.5
[2, 2, 4, 4, ∞ , ∞ , ∞ , ∞]	55.6	62.0	70.6	61.6	67.2	51.2	61.4
[4, 4, 8, 8, ∞ , ∞ , ∞ , ∞]	55.8	63.9	71.5	63.0	68.5	50.6	62.2
[1, 2, 4, 8, ∞ , ∞ , ∞ , ∞]	55.4	65.0	70.3	61.1	69.6	50.7	62.0

Table 10: Event Argument Role Labeling (EARL) and Relation Extraction (RE) *single-source transfer* results (F-score %) of our proposed approach GATE with different distance threshold δ using perfect event triggers and entity mentions. En, Zh, and Ar denotes English, Chinese, and Arabic languages, respectively. In “X \Rightarrow Y”, X and Y denotes the source and target language, respectively.

D GATE vs. Self-Attention

Our proposed approach GATE is a revision of the self-attention mechanism (Vaswani et al., 2017) and close to the concept of relation-aware self-attention (Shaw et al., 2018), so we compare them on both event argument role labeling and relation extraction tasks in single-source transfer setting. The results are presented in Table 11.

Model	En \Rightarrow Zh	En \Rightarrow Ar	Zh \Rightarrow En	Zh \Rightarrow Ar	Ar \Rightarrow En	Ar \Rightarrow Zh
Event Argument Role Labeling						
Self-Attention	61.5	55.0	58.0	57.7	54.3	57.0
Shaw et al. (2018)	62.3	60.8	57.3	66.3	57.5	59.8
GATE	63.2	68.5	59.3	69.2	53.9	57.8
Relation Extraction						
Self-Attention	57.1	63.4	69.6	60.6	67.0	52.6
Shaw et al. (2018)	58.0	59.9	70.0	55.6	66.5	56.5
GATE	55.1	66.8	71.5	61.2	69.0	54.3

Table 11: Event Argument Role Labeling (EARL) and Relation Extraction (RE) *single-source transfer* results (F-score %) of our proposed approach GATE and the Self-Attention mechanism (Transformer Encoder) using perfect event triggers and entity mentions. En, Zh, and Ar denotes English, Chinese, and Arabic languages, respectively. In “X \Rightarrow Y”, X and Y denotes the source and target languages, respectively.

E Translation Experiment

We perform English to Arabic and Chinese translations using Google Cloud Translate.¹³ During translation, we use special symbols to identify relation mentions and event mentions and their argument candidates in the sentences, as shown in Figure 4. We drop the examples ($\approx 10\%$) in which we cannot identify the mentions after translation.

¹³<https://cloud.google.com/translate/docs/basic/setup-basic>

English sentence: her *stockbroker* was also charged .
 Chinese translation: 她的*股票经纪人*也**被起诉**。
 Arabic translation: *لها* **اتهم** *سمسار الأوراق المالية* كما .

Figure 4: Translation of an English sentence in Chinese and Arabic with an event trigger (surrounded by ****) and a candidate argument (surrounded by *<i></i>*).

F Error Analysis

We compare our proposed approach GATE and the self-attention mechanism (Vaswani et al., 2017) by analyzing their predictions on the event argument role labeling (EARL) and relation extraction (RE) tasks. We consider the models trained on English language and evaluate them on Chinese language. We do not use the event trigger type as features while training models for the EARL task. We present the confusion matrices of these two models in Figure 5, 6, 7, and 8. In general, GATE makes more correct predictions. We noticed that in transferring from English to Chinese on the EARL task, GATE improves notably on Destination, Entity, Person, Place relation types. The syntactic distance between event triggers and their argument mentions that share those types corroborates with our hypothesis that distance-based dependency relations help in cross-lingual transfer.

However, we observed that GATE makes more *false positive* and less *false negative* predictions than the self-attention mechanism. We summarize the prediction rates on EARL in Table 12. There are several factors that may be associated with these wrong predictions. To shed light on those factors, we manually inspect 50 examples and our findings suggests that wrong predictions are due to three primary reasons. First, there are errors in the ground truth annotations in the ACE dataset. Second, the knowledge required for prediction is not available in the input sentence. Third, there are entity mentions, event triggers, and contextual phrases in the test data that rarely appear in the training data.

Model	True Positive	True Negative	False Positive	False Negative
Self-Attention	386	563	179	300
GATE	585	493	249	157

Table 12: Comparing GATE and Self-Attention on the EARL task using English and Chinese as the source and target languages, respectively. The rates are aggregated from confusion matrices shown in Figure 5 and 6.

G Reproducibility Checklist

We provide a few details related to our experiments below.

- Number of parameters
 - CL_Trans_GCN (Liu et al., 2019) 3.73M
 - CL_GCN (Subburathinam et al., 2019) 382k
 - CL_RNN (Ni and Florian, 2019) 1.59M
 - GATE (this work) 4.65M
- Average training time
 - CL_Trans_GCN (Liu et al., 2019) 15 mins
 - CL_GCN (Subburathinam et al., 2019) 12 mins
 - CL_RNN (Ni and Florian, 2019) 12 mins
 - GATE (this work) 15 mins
- Computing infrastructure: two GeForce GTX 1080 GPU.
- We manually tune the hyper-parameters on the validation set of each source language(s).
- We will release the source code on Github upon acceptance.
- We adopt the evaluation metric (F-score %) implementation from [here](#).
- We adopt the GCN implementation from [here](#) for the baseline methods.

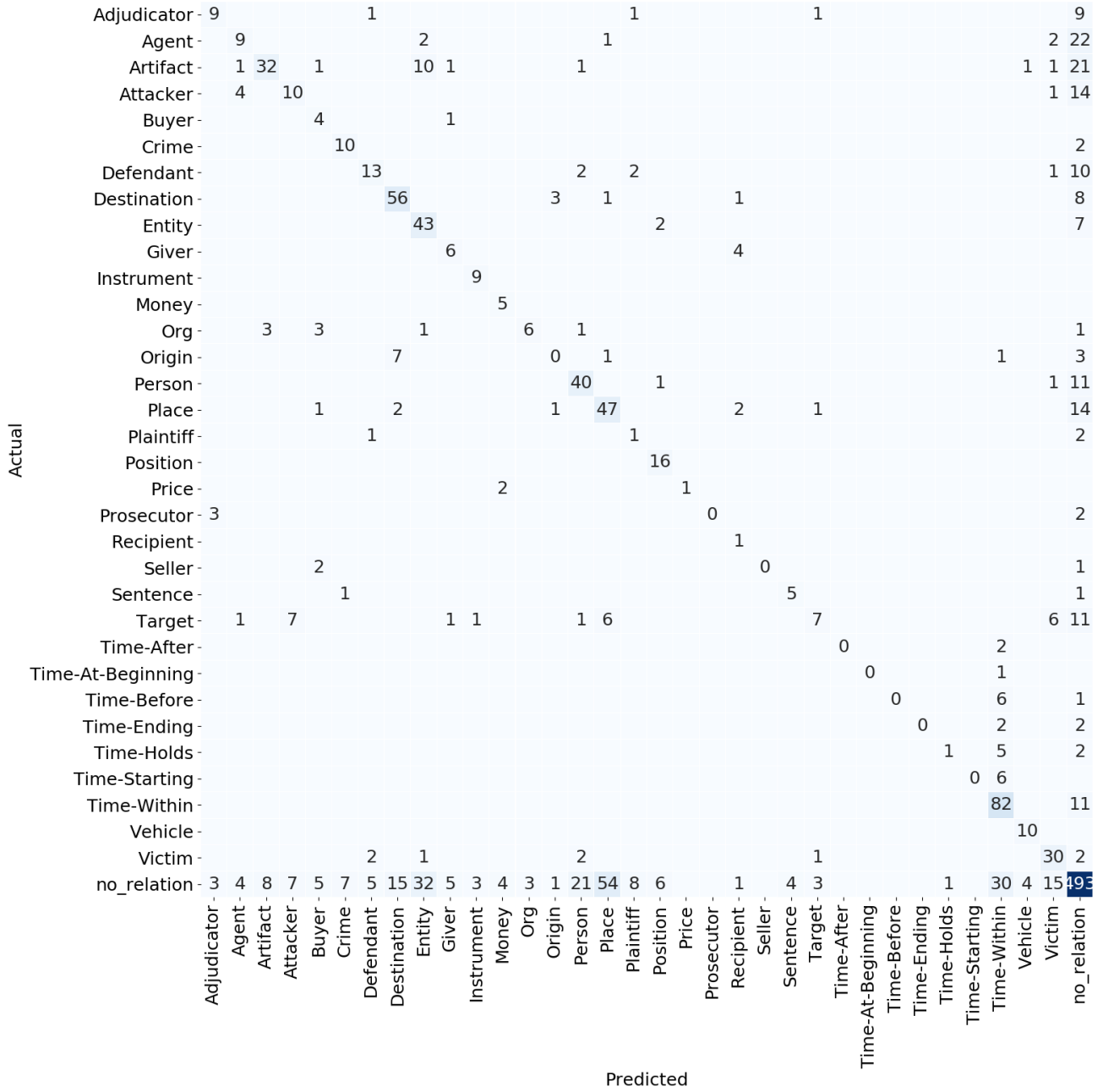


Figure 5: Event argument role labeling confusion matrix (on test set) based on our proposed approach **GATE** using English and Chinese as the source and target languages, respectively. The diagonal values indicate the number of correct predictions, while the other values denote the incorrect prediction counts.

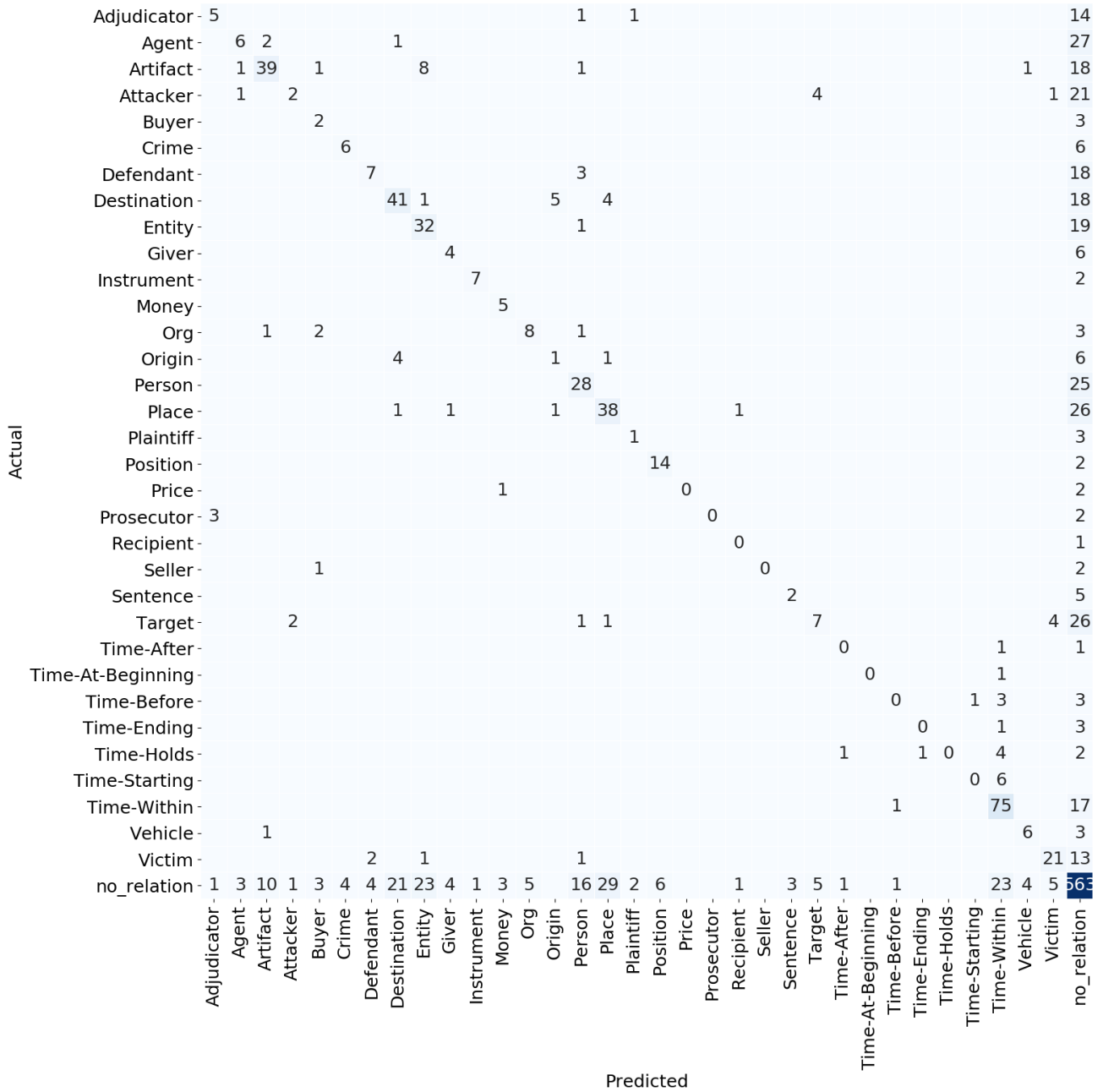


Figure 6: Event argument role labeling confusion matrix (on test set) based on the **Self-Attention (Transformer Encoder)** using English and Chinese as the source and target languages, respectively. The diagonal values indicate the number of correct predictions, while the other values denote the incorrect prediction counts.

Actual	ART:User-Owner-Inventor-Manufacturer	30	1					1										3	3
	GEN-AFF:Citizen-Resident-Religion-Ethnicity	43	6					2										5	2
	GEN-AFF:Org-Location		79					11											
	ORG-AFF:Employment	5	61	15	1													1	1
	ORG-AFF:Investor-Shareholder			0				7											2
	ORG-AFF:Membership		6	16															
	ORG-AFF:Ownership		2		1														
	ORG-AFF:Sports-Affiliation		1			2													1
	ORG-AFF:Student-Alum		1				0												
	PART-WHOLE:Artifact							2											
	PART-WHOLE:Geographical		1					93											1
	PART-WHOLE:Subsidiary		15					55											
	PER-SOC:Business									3			2						2
	PER-SOC:Family									1	15	4							2
	PER-SOC:Lasting-Personal									1	1	4							
	PHYS:Located	2	3					1									65	1	
	PHYS:Near							19									3	0	
	no_relation	6	41	5		13		8	138	89	121	36	35	1	3				198
	ART:User-Owner-Inventor-Manufacturer																		
	GEN-AFF:Citizen-Resident-Religion-Ethnicity																		
	GEN-AFF:Org-Location																		
	ORG-AFF:Employment																		
	ORG-AFF:Investor-Shareholder																		
	ORG-AFF:Membership																		
	ORG-AFF:Ownership																		
	ORG-AFF:Sports-Affiliation																		
	ORG-AFF:Student-Alum																		
	ORG-AFF:Artifact																		
	PART-WHOLE:Geographical																		
	PART-WHOLE:Subsidiary																		
	PER-SOC:Business																		
	PER-SOC:Family																		
	PER-SOC:Lasting-Personal																		
	PHYS:Located																		
	PHYS:Near																		
	no_relation																		
		Predicted																	

Figure 7: Relation extraction labeling confusion matrix (on test set) based on our proposed approach **GATE** using English and Chinese as the source and target languages, respectively. The diagonal values indicate the number of correct predictions, while the other values denote the incorrect prediction counts.

Actual	ART:User-Owner-Inventor-Manufacturer -	26	1					1								7	3
	GEN-AFF:Citizen-Resident-Religion-Ethnicity -	33	1	5					1							16	2
	GEN-AFF:Org-Location -		79						11								
	ORG-AFF:Employment -	5		61		16										1	1
	ORG-AFF:Investor-Shareholder -				0				9								
	ORG-AFF:Membership -			4		18											
	ORG-AFF:Ownership -			3			0										
	ORG-AFF:Sports-Affiliation -			3		1		0									
	ORG-AFF:Student-Alum -			1					0								
	PART-WHOLE:Artifact -								1								1
	PART-WHOLE:Geographical -		1	1					92								1
	PART-WHOLE:Subsidiary -		17							53							
	PER-SOC:Business -										3		2				2
	PER-SOC:Family -										2	17					3
	PER-SOC:Lasting-Personal -												3				3
	PHYS:Located -		1		2					1						68	
	PHYS:Near -									19						3	0
	no_relation -	4	43	1	1		14				1	127	89	79	23	6	8
	ART:User-Owner-Inventor-Manufacturer -																
	GEN-AFF:Citizen-Resident-Religion-Ethnicity -																
	GEN-AFF:Org-Location -																
	ORG-AFF:Employment -																
	ORG-AFF:Investor-Shareholder -																
	ORG-AFF:Membership -																
	ORG-AFF:Ownership -																
	ORG-AFF:Sports-Affiliation -																
	ORG-AFF:Student-Alum -																
	ORG-AFF:Artifact -																
	PART-WHOLE:Geographical -																
	PART-WHOLE:Subsidiary -																
	PER-SOC:Business -																
	PER-SOC:Family -																
	PER-SOC:Lasting-Personal -																
	PHYS:Located -																
	PHYS:Near -																
	no_relation -																
		Predicted															

Figure 8: Relation extraction confusion matrix (on test set) based on the **Self-Attention (Transformer Encoder)** using English and Chinese as the source and target languages, respectively. The diagonal values indicate the number of correct predictions, while the other values denote the incorrect prediction counts.