Event Detection via Gated Multilingual Attention Mechanism

Jian Liu[1,2], Yubo Chen[1], Kang Liu[1,2], Jun Zhao[1,2]
[1] National Laboratory of Pattern Recognition, Institute of Automation,
Chinese Academy of Sciences, Beijing, 100190, China
[2] University of Chinese Academy of Sciences, Beijing, 100049, China
{jian.liu, yubo.chen, kliu, jzhao}@nlpr.ia.ac.cn

Abstract

Identifying event instance in text plays a critical role in building NLP applications such as Information Extraction (IE) system. However, most existing methods for this task focus only on monolingual clues of a specific language and ignore the massive information provided by other languages. Data scarcity and monolingual ambiguity hinder the performance of these monolingual approaches. In this paper, we propose a novel multilingual approach — dubbed as **G**ated **M**ulti**L**ingual **Att**ention (GMLATT) framework — to address the two issues simultaneously. In specific, to alleviate data scarcity problem, we exploit the consistent information in multilingual data via context attention mechanism. Which takes advantage of the consistent evidence in multilingual data other than learning only from monolingual data. To deal with monolingual ambiguity problem, we propose gated cross-lingual attention to exploit the complement information conveyed by multilingual data, which is helpful for the disambiguation. The cross-lingual attention gate serves as a sentinel modelling the confidence of the clues provided by other languages and controls the information integration of various languages. We have conducted extensive experiments on the ACE 2005 benchmark. Experimental results show that our approach significantly outperforms state-of-the-art methods.

Introduction

The goal of Event Detection (ED) is to recognize event instance of particular type in plain text. Specifically, given a sentence, ED requires to decide whether the sentence contains events of interest. If so, it also needs to identify the specific event type and locate the event trigger for each event. Take the following sentence for example:

*In Baghdad, a cameraman **died** when an American tank **fired** on the Palestine hotel.*

According to the ACE 2005 annotation guideline[1], two events are mentioned here: a *Die* event triggered by "***died***", and an *Attack* event triggered by "***fired***". An ED system should be able to identify all of them. Building a robust ED system is challenging. According to (Ji and Grishman 2008), human annotator achieves only about 73% of F_1 score on the ACE 2005 evaluation task.

[1]https://goo.gl/C9T6Fg

To date, many methods (Li, Ji, and Huang 2013; Chen et al. 2015; 2017; Nguyen and Grishman 2015; 2016; Liu et al. 2016; 2017; Feng et al. 2016) have been proposed and obtain state-of-the-art performance. However, most of these methods are indeed monolingual approaches and focus only on exploiting textual clues in monolingual data. We argue that at least two problems hinder their performance:

Data scarcity. We analysis the widely used ACE 2005 corpus and show the statistics in Figure 1. The corpus defines 33 event types, however, nearly 70% of them only have instances fewer than 100. (Three types have instances even fewer than ten). The inadequacy of training data hinders the performance of the currently existing methods, which are under supervised learning paradigm and rely on huge amount of training examples to guarantee good performance. Besides, obtaining extra instances for training is also difficult. Valid data for training must be in accordance with the ACE 2005 annotation guideline, which is a 77-page document where the detailed event schemas and annotation specifications are elaborately illustrated.

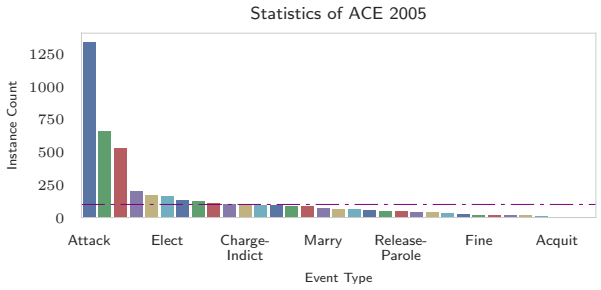

Figure 1: Statistics of ACE 2005 corpus. Dashed line indicates the threshold of 100.

Monolingual ambiguity. Monolingual ambiguity is another problem appears in ED task. On the one hand, the identical event can be described by completely different expressions which have different trigger words. On the other hand, the same word can express completely different events. To illustrate, consider the following sentences:

s1: *An American tank **fired** on the Palestine hotel.*

s2: *Two airline pilots were reportedly **fired** for stripping down.*

Both of s1 and s2 contain a word "***fired***" indicating an event happening. However, in s1, "***fired***" means "*discharge a gun (or other weapon)*[2]" and triggers an *Attack* event. In s2, "***fired***" means "*dismiss (an employee) from a job*", which triggers an event of *End-Position*. The specific event types are totally different. Ambiguity is a common phenomenon in English literature. For quantitative measurement, we examine the annotated events in ACE 2005 and observe that 57% of the trigger words are ambiguous, which makes identifying and categorizing events challenging.

In this paper, we propose a novel multilingual approach called **G**ated **M**ulti**L**ingual **Att**ention (GMLATT) framework to tackle with the above two issues simultaneously. We argue that multilingual approach can take the advantage of two types of information compared with monolingual approach:

1. ***Multilingual consistency***. Sentences conveying identical idea but in different languages usually have same or similar semantic components (McDonald et al. 2013). For example, the parallel Chinese sentence of s1 is: "*YiLiang MeiGuo TanKe Xiang Palestine LvDian KaiHuo*.", where "*MeiGuo TanKe*" is the corresponding counterpart of English expression "*American tank*", which refers to a military weapon. No matter in English or Chinese, weapon word tend to appear along with *Attack* event, which provides important clue for identifying the *Attack* event. The evidence among different languages is consistent and coherent. The augment clues can be leveraged to alleviate data scarcity in source language.

2. ***Multilingual complementation***. Different languages usually have distinct characteristics and idioms. Ambiguous expression of one language might have non-ambiguous counterpart in other language, which can provide complement information for disambiguation (Lin, Liu, and Sun 2017). For example, when projected s1 and s2 to Chinese, the corresponding translations of "*fired*" are *"KaiHuo"* and *"JieGu"* respectively. The two Chinese words share no common character or semantic and express *Attack* event and *End-Position* event without ambiguity. The information in Chinese side are helpful for disambiguating the event types in English side, which we referred as multilingual complementation.

Guided by these motivations, our framework, GMLATT employs two attention mechanisms for exploiting the two types of information to identify and categorize events:

(1) To exploit the consistent evidence among different languages, we adopt monolingual context attention for each language side. It is intuitive that the words in a sentence are of different importance. For example, in s1, "*American tank*" provides more important clue that an event might happen in conflict scenario than other words. In s2, "*airline pilots*" and "*stripping down*" are the salient parts implying an *End-Position* event happening. These words should get more attention than other words. Guided by this, we employ context attention mechanism to model the context texts surrounding the candidate trigger. Attention weights indicate the importance of different words in the sentence to predict the event type.

[2]http://www.dictionary.com/browse/fire

(2) To exploit the complement information, we adopt gated cross-lingual attention. Texts from other languages indeed provide valuable clues. However, how to combine them with the source features is a problem. We employ gated cross-lingual attention to model the confidence of the features provided by other languages. The attention gate serves as a sentinel to control the information flow from other languages to source side. Thus controls the information integration of various languages.

Our main contributions are three folds: (1) We propose a novel multilingual framework for ED task, which tackles with data scarcity and monolingual ambiguity problems simultaneously. (2) The framework contains two attention mechanisms: monolingual context attention and gated cross-lingual attention. To our best, this is the first work to introduce attention mechanism for modelling multilingual data in ED task. (3) We have conducted extensive experiments on the ACE 2005 corpus, and the experimental results show that our method achieves significant improvement over existing methods and sets a new state-of-the-art on this evaluation task.

Task Description

In ACE 2005, an event is defined as a specific occurrence involving one or more participants. The standard Event Extraction (EE) task requires certain types of events, which are mentioned in the source language text to be detected.

We introduce some ACE terminologies to facilitate the understanding of the task:

Entity: an object in one of the semantic categories of interests. **Entity mention:** a reference to an entity (typically, a noun phrase). **Event mention:** a phrase or sentence within which an event is described, including the trigger and arguments. **Event trigger:** the word most clearly expresses the event mention, most often a single verb or noun. **Event argument:** an entity mention, temporal expression or value (e.g. Job-Title) that serves as a participant or attribute with a specific role in an event mention.

For the sentence: "*He **died** in the hospital.*" An event extractor should detect a *Die* event, along with the event trigger "***died***" and the event arguments "*He*" (Role=*Victim*) and "*hospital*" (Role=*Place*). In this paper, instead of doing the overall standard EE task, we concentrate only on ED task — namely Event Trigger Identification and Event Type Classification. That is, for the previous example, our goal is to identify that the token "***died***" is an event trigger and the event type is *Die*.

Methodology

In this section, we illustrate the details of the GMLATT framework. The representation of GMLATT is given in Figure 2. GMLATT works in four steps:

Multilingual Projection — Project monolingual text to parallel multilingual texts. We leverage Machine Translation (MT) to bootstrap the source data.

Sentence Representation — Transfer symbolic representation of sentence to unified distributed representation.

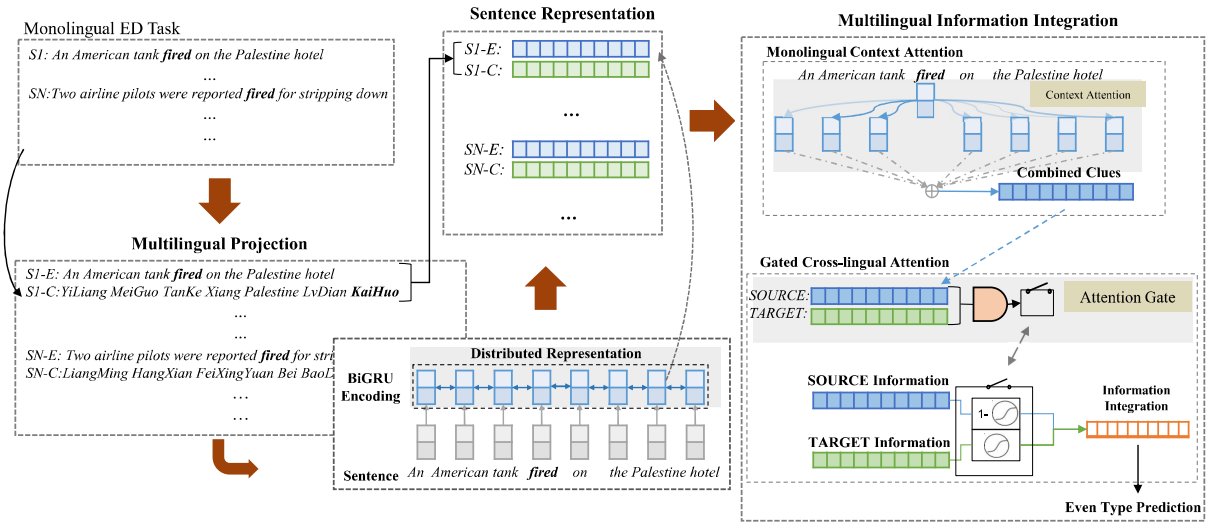

Figure 2: The overall architecture of the GMLATT framework. After multilingual projection, the framework leverages BiGRU encoder to encode sentence to distributed representation. Monolingual context attention and gated cross-lingual attention are employed to do multilingual information integration.

Recurrent Neural Networt (RNN) is leveraged to model sentence of each language.

Multilingual Information Integration — Assemble the information from different languages. It refers to: (1) Use monolingual context attention to assemble information in each language. (2) Use gated cross-lingual attention to integrate information of different languages.

Event Type Prediction — Do the fine-grained event type classification.

In this paper, we focus on English event detection, namely the SOURCE language is set to English. And we use Chinese as the sole TARGET language to bootstrap the source data. Note that in theory, our framework is not limited by specific language, and can even use various languages as TARGET languages simultaneously, which we leave for the future work.

Multilingual Projection

Since ACE 2005 only has annotated monolingual events, to apply our framework, we first use online machine translation service[3] to obtain the parallel text in TARGET language.

For trigger projection, instead of using heuristic rules or external dictionary, we learn the alignments in an unsupervised way. We employ GIZA++[4] — a commonly used tool in MT — to learn multilingual alignments. GIZA++ treats word alignment as hidden variables and use EM algorithm to find the best alignments unsupervisedly (Och and Ney 2003). We concatenate the translated multilingual data with a 200k parallel English-Chinese corpus[5] released by (Eisele and Chen 2010) to learn the alignments together.

[3]http://fanyi.baidu.com/, in our experiment
[4]http://www.fjoch.com/GIZA++.html
[5]http://opus.lingfil.uu.se/MultiUN.php

Note that, the alignments learned by GIZA++ are directed. We learn both SOURCE-to-TARGET and TARGET-to-SOURCE alignments and then leverage *grow-diag-final-and* algorithm published in (Och and Ney 2003) to merge them. Below is one example of multilingual projection:

SOURCE: *An American tank **fired** on the Palestine hotel*
TARGET: *YiLiang MeiGuo TanKe Xiang Palestine LvDian **KaiHuo** .*

Figure 3: An example of multilingual projection. Straight lines indicate alignments. The English trigger word "***fired***" is correctly projected to the Chinese word "***KaiHuo***".

Sentence Representation

Recurrent Neural Network (RNN) shows promising result in sequence modelling. We adopt a particular implementation of RNN called Gated Recurrent Units (GRU) (Chung et al. 2014) to model word sequence to represent sentence.

At each step t, the vanilla GRU accepts a current input x_t and previous hidden state h_{t-1} to compute the current hidden state $h_t = GRU(x_t, h_{t-1})$ as:

$$z_t = \sigma(W_z x_t + U_z h_{t-1} + b_z) \quad (1)$$
$$r_t = \sigma(W_r x_t + U_r h_{t-1} + b_r) \quad (2)$$
$$\widetilde{h_t} = tanh(W_h x_t + r_t \circ U_h h_{t-1} + b_h) \quad (3)$$
$$h_t = z_t \circ h_{t-1} + (1 - z_t) \circ \widetilde{h_t} \quad (4)$$

A more advanced model is bidirectional GRU (Bi-GRU). It maintains two hidden states at each time-step t, one for the left-to-right propagation $\overrightarrow{h_t} = GRU_{ltr}(x_t, \overrightarrow{h}_{t-1})$ and another for the right-to-left propagation $\overleftarrow{h_t} =$

$GRU_{rtl}(x_t, \overleftarrow{h}_{t+1})$. It combines the two states as the current state: $h_t = [\overrightarrow{h_t}; \overleftarrow{h_t}]$.

We adopt Bi-GRU to encode sentence. Each token in the sentence is first transferred to lexical features by three look-up tables:

Word Embedding Table: Word embeddings are able to capture the meaningful semantic regularities of words (Turian, Ratinov, and Bengio 2010). We use word embeddings as the basic features. We use the Skip-gram model to learn word embeddings of different languages — NYT corpus for English and CN Wikipedia for Chinese.

Entity Embedding Table: Following existing work (Li, Ji, and Huang 2013; Chen et al. 2015; Nguyen and Grishman 2015; Liu et al. 2016), we exploit the annotated entity information as additional features. We randomly initialize embedding vectors for each entity type (including the NA type) and update them during training procedure. Entity embedding table is shared by different languages.

Position Embedding Table: In the task of relation extraction, (Zeng et al. 2014) uses position embedding to represent the distance between context words and entities, which brings huge performance improvement. Similarly, we also use position embeddings specified by the relative distances between context words and the candidate trigger. We randomly initialize embedding vectors and update them during training procedure.

We concatenate the above three features as the input to Bi-GRU. The hidden state sequence in Bi-GRU is used as the representation of sentence. We denote the sentence representation in SOURCE side as $S_{src} = \{hs_1, hs_2, ..., hs_L\}$, and in TARGET side as $S_{tgt} = \{ht_1, ht_2, ..., ht_N\}$, where L and N are sentence lengths respectively.

Multilingual Information Integration

This is the core part of the GMLATT framework. The framework leverages monolingual context attention to model context texts for exploiting the consistent evidence among different languages. And it leverages gated cross-lingual attention to model the confidence of the complement clues provided by TARGET side, which controls the information integration from various languages.

Monolingual Context Attention Mechanism

Monolingual context attention mechanism is performed on each side. We only use SOURCE side for illustration to avoid duplication. Give a SOURCE sentence and its representation $S_{src} = \{hs_1, hs_2, ..., hs_L\}$, we first treat each token as a *candidate trigger* with its representation denoted as h_{c_src}. We then leverage attention mechanism to exploit the context texts surrounding it to find evidence for deciding its type. The importance of each token in the context texts is computed as:

$$a_i = \frac{exp(m_i)}{\sum_{l=1}^{L} exp(m_l)} \tag{5}$$

where m_i is the relatedness between the *candidate trigger* representation h_{c_src} and the context token representation hs_i, modelled by bi-linear attention as:

$$m_i = tanh(h_{c\_src}^T W_{Att\_src} hs_i + b_{Att\_src}) \tag{6}$$

where W_{Att_src} is the weight matrix and b_{Att_src} is the bias term. Given all the importance weights, the comprehensive information conveyed by S_{src} with respect to the *candidate trigger* is obtained by weighted sum:

$$R_{src} = \sum_{i=1}^{L} a_i * hs_i \tag{7}$$

For TARGET side, we first find the aligned token of the *candidate trigger* — we denote its representation as h_{c_tgt} — and then do the same process as SOURCE side to exploit the relation between the *aligned candidate trigger* and the TARGET sentence representation. R_{tgt} denotes the comprehensive information of the corresponding sentence in TARGET side.

Gated Cross-Lingual Attention Mechanism

Given R_{src} and R_{tgt}, the next step in to combine them to get the integrated information. We come up with two combination strategies:

STG1: Averaged Sum. We assume that the information from SOURCE side and TARGET side are of same importance. The integrated information is computed as:

$$R_{integrated} = 0.5 * R_{tgt} + 0.5 * R_{src} \tag{8}$$

STG2: Weighted Sum. We assume that information from SOURCE side and TARGET side are of different importance. We leverage gated cross-lingual attention to model the confidence of clues provided by target side R_{tgt}. The value of the attention gate is computed as:

$$G_{cl} = \sigma(W_{cl}[R_{src}; R_{tgt}] + b_{cl}) \tag{9}$$

where W_{cl} is the weight matrix and b_{cl} is the bias term. σ is the multivariate *sigmoid* function accepting vector as input, and compute the output as:

$$\sigma(x) = 1/(1 + \exp(-x)) \tag{10}$$

We use 1 - G_{cl} and G_{cl} as the combination weights of SOURCE side and TARGET side to assemble R_{src} and R_{tgt}. Since our original intention is do ED in the SOURCE side, the attention gate can be seen as a sentinel to control information flow from TARGET side to SOURCE side. Note that the dimension of the attention gate is same as R_{src} and R_{tgt}, and the integrated information is computed by weighted sum as:

$$R_{itgd} = (G_i \circ R_{tgt}) + ((1 - G_i) \circ R_{src}) \tag{11}$$

where $\circ$ stands for element-wise multiplication operation. R_{itgd} represents the final integrated information of various languages.

Event Type Prediction

Follow previous works, we formulate ED as a multiclass classification problem. For each token in SOURCE sentence, we predict whether is an event trigger and decide event type for it. We combine h_{c_src}, h_{c_tgt} and R_{itgd} as input to a softmax classifier:

$$O = softmax(tanh(W_{cls}[h_{c\_src}; h_{c\_tgt}; R_{itgd}] + b_{cls})) \quad (12)$$

where W_{cls} is the weight matrix, and b_{cls} is the bias term. The output O is a real-valued vector indicating the predicted probabilities of different types. The probability of the *candidate trigger* t belongs to type j is:

$$P(j|t, \Theta) = O_{(j)} \quad (13)$$

where Θ represents all the parameters of GMLATT. $O_{(j)}$ indicates the jth element of O.

Training and Optimization

Here we introduce the learning and optimization details of the GMLATT framework. The optimization objective function is defined as multi-class cross-entropy loss:

$$J(\Theta) = -\sum_{i=1}^{K} lnP(y_i|t_i, \Theta) + \lambda(\Theta) \quad (14)$$

where K indicates the number of all tokens in the training data. y_i is the true event type of token t_i. λ is the regularization parameter and Θ indicates all parameters.

For optimization method, we adopt mini-batch stochastic gradient descent (SGD) to minimize the objective function. We add dropout layer to prevent the co-adaptation of the parameters to fight against overfitting problem (Srivastava et al. 2014). Negative sampling is adopted to tackle with data imbalance problem. (The ratio of Non-Type event to typed event is about 200:1 in training data).

Experiments

Dataset and Evaluation Metrics

We conduct experiments on the widely used ACE 2005 dataset. This corpus contains 599 documents annotated with 8 types and 33 subtypes of events. Following previous works (Liao and Grishman 2010; Li, Ji, and Huang 2013; Chen et al. 2015; Nguyen and Grishman 2015; Liu et al. 2016), we simply treat them as 33 separate event types and ignore the hierarchical structure. We use the same data separation as the previous works: 40 particular articles are used as the blind test set; 30 articles are used as the development set; and the remaining 529 articles are used for training.

We use the following criteria for evaluation: (1) A trigger is correctly identified if its offset matches a reference trigger (Event Trigger Detection). (2) A trigger is correctly classified if both its offset and event type match a reference trigger (Event Type Classification). Precision (P), recall (R), and F_1 score (F_1) are used as the evaluation metrics. Same as all the previous works for meaningful comparison.

Comparison with Existing Methods

We compare our model with many state-of-the-art methods. We classify these methods into three types:

Feature-based approach. *MaxEnt*: The method in (Li, Ji, and Huang 2013), which only employs human-designed features. *CrossEvent*: The method in (Liao and Grishman 2010), which leverages document information for complex features. *Combined-PSL*: The method in (Liu et al. 2016), which uses probabilistic soft logic model to exploit global information, the best reported feature-based system.

Representation-based approach. *DMCNN*: The method in (Chen et al. 2015), which uses CNN to do automatical feature extraction. *JRNN*: The method in (Nguyen, Cho, and Grishman 2016), which uses more complicated structure to model event inter-dependencies. *NC-CNN*: The method in (Nguyen and Grishman 2016), which models non-continue n-grams to achieve higher performance.

External resource based approach. *DMCNN-DS*: The method in (Chen et al. 2017), which uses *FreeBase* to label new training data by Distance Supervision. *ANN-FN*: The method in (Liu et al. 2016), which exploits events in *FrameNet* to bootstrap training. *ANN-AugATT*: The method in (Liu et al. 2017), which leverages additional arguments information and *FrameNet*.

The performance is shown in Table 1. From the results, we have several observations:

Among all these methods, representation-based approaches beat feature-based approaches (by *1.9%* of F_1 on average) and external resource based approaches achieve the best results (*1.0%* of F_1 on average over the represent-based approaches). This phenomenon is not surprising. Representation-based approaches achieve better performance by avoid complex feature engineering, and automatically learn salient features in the data. While external resource based methods combine feature learning with much more data from external resource to achieve further improvement.

Besides, among all the approaches, our method behaves best. It achieves the best performance (*72.4%* of F_1) on Event Type Classification and beats the best reported external resource based approach *ANN-AugATT* by *0.7%* on F_1. For Event Type Identification, it achieves comparative performance (*74.1%* vs *74.3%*) compared with the best system.

We make a reasonable explanation: compared with representation-based approaches, our method can leverage the augment clues provided by multilingual data to enhance feature learning. Compared with external resource based approaches, the result is interesting: Our model tends to achieve higher precise score (highest both in Event Type Classification and Event Type Identification) but lower recall. We argue it is the different types of additional information causes this phenomenon. Our model use multilingual parallel data to learn features, and multilingual consistency and multilingual complementation provide more coherent and related clues attribute for the higher precision. While for external resource based methods, they usually need to transfer external data to fit the current event schemas by some heuristic rules (Chen et al. 2017; Liu et al. 2017). The data they leveraged is of huge amount but is indeed imprecise and noisy, which is responsible for the the higher recall but lower precision.

Models	Event Trigger Identification (%)			Event Type Classification (%)		
	P	R	F1	P	R	F1
MaxEnt (Li, Ji, and Huang 2013)	76.2	60.5	67.4	74.5	59.1	65.9
CrossEvent (Liao and Grishman 2010)	n/a	n/a	n/a	68.7	68.9	68.8
Combined-PSL (Liu et al. 2016)	n/a	n/a	71.7	75.3	64.4	69.4
DMCNN (Chen et al. 2015)	80.4	67.7	73.5	75.6	63.6	69.1
JRNN (Nguyen, Cho, and Grishman 2016)	68.5	**75.7**	71.9	66.0	**73.0**	69.3
NC-CNN (Nguyen and Grishman 2016)	n/a	n/a	n/a	n/a	n/a	71.3
DMCNN-DS (Chen et al. 2017)	79.7	69.6	**74.3**	75.7	66.0	70.5
ANN-FN (Liu et al. 2016)	n/a	n/a	n/a	77.6	65.2	70.7
ANN-AugATT (Liu et al. 2017)	n/a	n/a	n/a	78.0	66.3	71.7
GMLATT (Ours)	**80.9**†	68.1	74.1	**78.9**†	66.9	**72.4**†

Table 1: Performance of all the methods. Bold denotes the best result. † denotes signification improvement.

Detailed Analysis

We conduct extra experiments to do detailed analysis. **EI** and **EC** stand for F_1 score of Event Trigger Identification and Event Type Classification respectively.

Impact of Different Attention Strategies

We come up with five models comparing with two baselines to exploit the effects of different attention strategies.

Monolingual Setting

DNN-ED is the basic fast-forword neural network model proposed in (Liu et al. 2016). *DMCNN* is the model in (Chen et al. 2015) leveraging CNN to do automatically feature capturing. *GMLATT-Mon* is GMLATT without any attention mechanisms. *GMLATT-MonATT* is GMLATT only with monolingual context attention.

Multilingual Setting

GMLATT-TA is GMLATT without gated cross-lingual attention but leverages the aligned candidate trigger as multilingual information. *STG1* and *STG2* are two different strategies have been described previously.

Results are shown in Table 2:

Models		EI (%)	EC (%)
Monolingual	*DNN-ED*	68.1	66.9
	DMCNN	73.5	69.1
	GMLATT-Mon	72.7	68.3
	GMLATT-MonATT	**73.6**	**70.4**
Multilingual	*GMLATT-TA*	73.6	71.2
	GMLATT with STG1	73.6	71.7
	GMLATT with STG2	**74.1**	**72.4**

Table 2: Performance on different attention strategies.

We find that: 1) Even the simplest model in multilingual setting beats the best monolingual approach (*0.8%* of F_1 on EC), which justifies the intuition of exploiting multilingual data to detect event. 2) Among monolingual approaches, *GMLATT-MonATT* beats two baselines and achieves the best result, showing the effectiveness of monolingual context attention mechanism. 3) Among multilingual approaches, both STG1 and STG2 achieve better results than *GMLATT-TA*, which only uses aligned triggers as multilingual data. Besides, STG2 beats STG1 by a large margin (*0.7%* on EI, and *0.8%* on EC), showing the advantages of using gated cross-lingual attention to assemble information.

Impact of Feature Combinations

We exploit the effects of feature combinations. The experiments are conducted in monolingual setting to eliminate the effect of multilingual data. *DNN-ED* is treated as simple model, and *GMLATT-MonATT* serves as complex model. Results are shown in table 3:

Models	Features	EI (%)	EC (%)
DNN-ED	Word Embedding	66.6	65.7
	+ Entity Embedding	67.5	66.2
	+ Position Embedding	67.2	66.1
	All	**68.1**	**66.9**
GMLATT-MonATT	Word Embedding	72.4	68.4
	+ Entity Embedding	73.1	69.2
	+ Position Embedding	72.9	68.8
	All	**73.6**	**70.4**

Table 3: Performance on features combination.

According to the results, entity embedding and position embedding provide complement improvement both in simple model and complex model. It seems that, entity embedding is more effective than position embedding (by *0.2%* on average). An intuition explanation is that: when using RNN to encode sentence, position information might have been encoded by some extent to the distributed representations. Which reduces the effect of adding position embedding.

Performance on the Projected Data

We examine the performance on the projected data. Results are shown in Table 4.

Models	EI (%)	EC (%)
DNN-ED	64.4	62.1
DMCNN	67.0	64.7
GMLATT-MonATT	67.3	65.8

Table 4: Performance on the projected Chinese data.

Compared with ED in English, the performance on the projected Chinese data is relatively lower. We attribute the poor performance to the noise introduced by MT. The noise includes translation error and alignment error. Nevertheless,

it still demonstrates the consistence performance of English ED — GMLATT with attention mechanism outperforms the two baselines.

Attention Visualization

Attention weights of two examples are visualized in Figure 4.

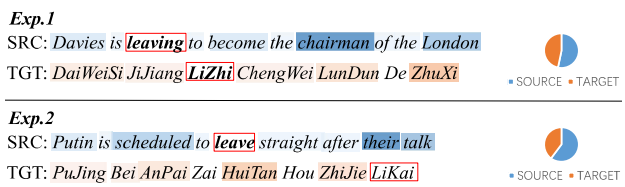

Figure 4: Attention visualization. The heatmap indicates the context attention. Blue for SOURCE language, and orange for TARGET language. The pie chart indicates the *average value* of the gated cross-lingual attention.

In Exp.1, "chairman" and "London" in SOURCE side and "ZhuXi" and "LunDun" in TARGET side are of higher attention weights. Which implies an event related to profession changing and decides the event type of "leaving" is *End-Position*. In Exp.2, "scheduled" and "their talks" are the words with much attentions in SOURCE side. They imply a common life scenario and yield out an *Movement* event for "leave". Note that, the attention weights in TARGET side of Exp.2 are messy. Which is in accordance with the lower value of the averaged gated cross-lingual attention weight of TARGET side compared with Exp.1.

We sample out 100 examples to exploit the gated cross-lingual attention weight, and find that the ratio of SOURCE to TARGET is 0.59:0.41. The lower weight of TARGET side implies that — the noise leaded by MT makes TARGET side less reliable than SOURCE side.

Error Analysis

For that the F_1 score is the compromise of precision and recall, for each metric, we sample out 50 examples to examine the reasons. We give them a brief summary in Table 5.

Metric	Reasons	Count
Precision	Highly ambiguous triggers	29
	Annotation ambiguity	16
	Others	5
Recall	Rare triggers	24
	Nouns and adjectives triggers	16
	Others	10

Table 5: Error analysis for precision and recall

For precision, highly ambiguous trigger (word triggers too many events, especially support verbs such as "go", "take") is the main reasons affect the performance. Annotation ambiguity is also a serious problem. For example, does "hacked to death" express an *Attack* event or *Die* event? ACE 2005 only assigns one event for each event mention, which leads to annotation ambiguity.

For recall, rare trigger (with training instances less than five), nouns and adjectives based triggers are the main reasons. Other reasons include indirect triggers such as "this", "what", which needs deep semantic analysis of the sentence to identify them out.

Related Work

Various methods have been proposed for ED. Feature-based approaches rely on discriminative features to build statistical models. (Ahn 2006) leverages lexical features and syntactic features. More advanced features include cross-document features (Ji and Grishman 2008), cross-event features (Gupta and Ji 2009; Liao and Grishman 2010), etc. (Li, Ji, and Huang 2013) presents a joint framework based on structured perceptron and beam search to do event trigger and event argument prediction jointly. (Liu et al. 2016) uses probabilistic soft logic based approach to combine local and global features. Representation-based approaches have been introduced into ED very recently. It includes (Chen et al. 2015; Nguyen and Grishman 2015) using convolutional neural network (CNN) to avoid complicated feature engineering, modelling non-continue skip-grams (Nguyen and Grishman 2016). Representation-based methods achieve relatively high performance due to their ability of automatic features capturing and modelling complicated hidden interactions in data. However, as we mentioned before, data scarcity in ED limits their further performance. External resource based model tackles data scarcity problems by exploiting additional information. (Chen et al. 2017) uses *FreeBase* to label new training data by Distance Supervision. (Liu et al. 2016) exploits events in *FrameNet* to obtain more available data. (Liu et al. 2017) leverages additional information from arguments and *FrameNet* to detect event, which achieves a better performance.

For multilingual approaches, (Ji 2009; Li et al. 2012; Wei et al. 2017) are excellent works, belong to traditional feature-based approaches. (Agerri et al. 2016; Danilova, Alexandrov, and Blanco 2014; Feng et al. 2016) illustrate technics for building multilingual event extraction systems, however, the "multilingual" in their works indeed means "language independent", which is orthogonal to our work.

Attention-based deep learning is also related to our work. It has attracted a lot of interests of researchers for its ability to learn implicit relationship between source and target. Attention-based models have been applied to various areas such as machine translation (Bahdanau, Cho, and Bengio 2014), machine comprehension (Hermann et al. 2015), image caption generation (Xu et al. 2015), etc. To the best of our knowledge, this is the first effort to adopt attention-based neural network for tackling with ED task.

Conclusion

In conclusion, we leverage two attention mechanisms in a novel multilingual framework to tackle with data scarcity and monolingual ambiguity problems appear in ED simultaneously. Experimental results show that our method achieves significant improvement over existing methods and sets a new state-of-the-art performance.

Acknowledgements

The research work is supported by the National Key Research and Development Program of China under Grant No. 2017YFB1002101 and the National Science Foundation of China under Grant No. 61533018. This work was also supported in part by Ant Financial Services Group and supported by Alibaba Group through Alibaba Innovative Research (AIR) Program.

References

Agerri, R.; Aldabe, I.; Laparra, E.; Rigau, G.; Fokkens, A.; Huijgen, P.; Izquierdo Bevia, R.; van Erp, M.; Vossen, P.; Minard, A.-L.; and Magnini, B. 2016. *Multilingual Event Detection using the NewsReader Pipelines*.

Ahn, D. 2006. The stages of event extraction. In *Proceedings of the Workshop on Annotating and Reasoning About Time and Events*, ARTE '06, 1–8.

Bahdanau, D.; Cho, K.; and Bengio, Y. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

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*, 167–176.

Chen, Y.; Liu, S.; Zhang, X.; Liu, K.; and Zhao, J. 2017. Automatically labeled data generation for large scale event extraction. In *Proceedings of ACL*, 409–419.

Chung, J.; Gulcehre, C.; Cho, K.; and Bengio, Y. 2014. *Empirical evaluation of gated recurrent neural networks on sequence modeling*.

Danilova, V.; Alexandrov, M.; and Blanco, X. 2014. *A Survey of Multilingual Event Extraction from Text*. 85–88.

Eisele, A., and Chen, Y. 2010. Multiun: A multilingual corpus from united nation documents. In Tapias, D.; Rosner, M.; Piperidis, S.; Odjik, J.; Mariani, J.; Maegaard, B.; Choukri, K.; and Chair), N. C. C., eds., *Proceedings of the Seventh conference on International Language Resources and Evaluation*, 2868–2872.

Feng, X.; Huang, L.; Tang, D.; Ji, H.; Qin, B.; and Liu, T. 2016. A language-independent neural network for event detection. In *Proceedings of ACL*, 66–71.

Gupta, P., and Ji, H. 2009. Predicting unknown time arguments based on cross-event propagation. In *Proceedings of the ACL-IJCNLP*, 369–372.

Hermann, K. M.; Kociský, T.; Grefenstette, E.; Espeholt, L.; Kay, W.; Suleyman, M.; and Blunsom, P. 2015. Teaching machines to read and comprehend. In *NIPS*.

Ji, H., and Grishman, R. 2008. Refining event extraction through cross-document inference. In *Proceedings of ACL-08: HLT*, 254–262.

Ji, H. 2009. Cross-lingual predicate cluster acquisition to improve bilingual event extraction by inductive learning. In *Proceedings of the Workshop on Unsupervised and Minimally Supervised Learning of Lexical Semantics*, 27–35.

Li, P.; Zhou, G.; Zhu, Q.; and Hou, L. 2012. Employing compositional semantics and discourse consistency in chinese event extraction. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 1006–1016.

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.

Lin, Y.; Liu, Z.; and Sun, M. 2017. Neural relation extraction with multi-lingual attention. In *Proceedings of ACL*, 34–43.

Liu, S.; Liu, K.; He, S.; and Zhao, J. 2016. A probabilistic soft logic based approach to exploiting latent and global information in event classification. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA.*, 2993–2999.

Liu, S.; Chen, Y.; Liu, K.; and Zhao, J. 2017. Exploiting argument information to improve event detection via supervised attention mechanisms. In *Proceedings of ACL*, 1789–1798.

McDonald, R. T.; Nivre, J.; Quirmbach-Brundage, Y.; Goldberg, Y.; Das, D.; Ganchev, K.; Hall, K. B.; Petrov, S.; Zhang, H.; Täckström, O.; et al. 2013. Universal dependency annotation for multilingual parsing. In *Proceedings of ACL*, 92–97.

Nguyen, T. H., and Grishman, R. 2015. Event detection and domain adaptation with convolutional neural networks. In *Proceedings of EMNLP*, 365–371.

Nguyen, T. H., and Grishman, R. 2016. Modeling skip-grams for event detection with convolutional neural networks. In *Proceedings of EMNLP*, 886–891.

Nguyen, T. H.; Cho, K.; and Grishman, R. 2016. Joint event extraction via recurrent neural networks. In *Proceedings of NAACL: Human Language Technologies*, 300–309.

Och, F. J., and Ney, H. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics* 29(1):19–51.

Srivastava, N.; Hinton, G. E.; Krizhevsky, A.; Sutskever, I.; and Salakhutdinov, R. 2014. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research* 15(1):1929–1958.

Turian, J.; Ratinov, L.-A.; and Bengio, Y. 2010. Word representations: A simple and general method for semi-supervised learning. In *Proceedings of ACL*, 384–394.

Wei, S.; Korostil, I.; Nothman, J.; and Hachey, B. 2017. English event detection with translated language features. In *Proceedings of ACL*, 293–298.

Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A. C.; Salakhutdinov, R.; Zemel, R. S.; and Bengio, Y. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, volume 14, 77–81.

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.