Neural Cross-Lingual Event Detection with Minimal Parallel Resources

Jian Liu[12]*, **Yubo Chen**[1]*, **Kang Liu**[12], **Jun Zhao**[12]
[1] National Laboratory of Pattern Recognition, Institute of Automation
Chinese Academy of Sciences, Beijing, 100190, China
[2] University of Chinese Academy of Sciences
{jian.liu, yubo.chen, kliu, jzhao}@nlpr.ia.ac.cn

Abstract

The scarcity in annotated data poses a great challenge for event detection (ED). Cross-lingual ED aims to tackle this challenge by transferring knowledge between different languages to boost performance. However, previous cross-lingual methods for ED demonstrated a heavy dependency on parallel resources, which might limit their applicability. In this paper, we propose a new method for cross-lingual ED, demonstrating a minimal dependency on parallel resources. Specifically, to construct a lexical mapping between different languages, we devise a context-dependent translation method; to treat the word order difference problem, we propose a shared syntactic order event detector for multilingual co-training. The efficiency of our method is studied through extensive experiments on two standard datasets. Empirical results indicate that our method is effective in 1) performing cross-lingual transfer concerning different directions and 2) tackling the extremely annotation-poor scenario.

1 Introduction

Event detection (ED) is a crucial natural language processing problem that aims to identify event triggers in texts (Ahn, 2006; Nguyen and Grishman, 2015). For example, in the sentence: "*A man* **died** *when a tank* **fired** *on the hotel*", ED requires a system to identify two event triggers, **died** and **fired**, along with their types `Die` and `Attack`.

Generally, training an ED system requires to obtain a considerably large amount of labeled data. However, owing to the complexity and high costs of annotation, existing event resources are scarce and unbalanced across languages (Hsi et al., 2016), which prevents us from building an ED system in languages with insufficient training data. Cross-lingual ED (Ji, 2009; Chen and Ji, 2009; Zhu et al., 2014; Hsi et al., 2016; Liu et al., 2018a) aims to tackle this challenge by transferring knowledge cross languages to boost performance. However, previous cross-lingual ED methods rely on either high-performance machine translation (MT) systems trained on large numbers of parallel sentences or manually aligned documents to achieve a decent performance — the required parallel resources may only exist for a small fraction of language pairs (Koehn et al., 2007), which greatly limits the applicability of these methods.

In this paper, we propose a new simple but effective method for cross-lingual ED, which can overcome the data scarcity problem in annotation-poor languages by jointly training with resources in other languages. Compared with previous methods, our approach demonstrates a minimal dependency on parallel resources, which may fit with language pairs that do not have large bitexts.

To achieve cross-lingual transfer, two challenges exist: 1) how to build lexical mappings between different languages, and 2) how to handle the word order difference problem (Xie et al., 2018). For the first challenge, previous studies (Guo et al., 2015; Ni et al., 2017; Mayhew et al., 2017; Xie et al., 2018; Lample et al., 2018) have investigated embedding projection-based method in cross-lingual applications and achieved promising results. For example, (Xie et al., 2018) proposed a novel "cheap translation" based method which has greatly advanced the performance in zero-shot named entity recognition (NER). However, these methods may not directly fit with cross-lingual ED, as the lexical mapping in ED is usually context-dependent but not deterministic as in other tasks. Consider the following English-to-Chinese lexical mapping examples. To preserve its original meaning, the trigger word "fire" in "*gangsters* fire *at a policeman*" (which evokes an `Attack`

*Equal contribution.

Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pages 738–748, Hong Kong, China, November 3–7, 2019.

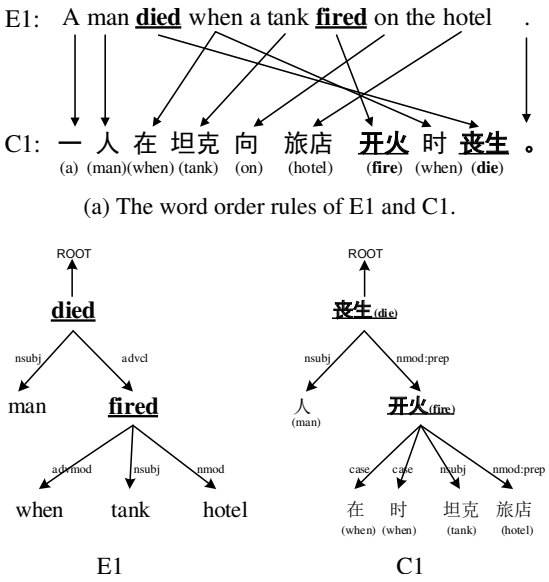
event), and "*the house caught* fire" (which evokes an NA event) should be translated as different Chinese words "开 火(open fire)" and "着 火(be on fire)" respectively. But in previous lexical mapping methods, the "fired" is always having the same transferred representation irrespective of its contexts. This problem might introduce noise for cross-lingual ED.

To address the above issues, in this paper, we devise a content-dependent lexical mapping method for cross-lingual ED. Similar to (Xie et al., 2018), for each source word, we first project it into a shared embedding space, but instead of adopting a deterministic word-to-word translation, we retrieve different translation candidates by looking for its nearest neighbors, and we then adopt a context-aware selective attention mechanism to rank these candidates to find the best-suited translation. Compared with previous methods, our approach can obtain content-dependent translations for each word in the source training data, which may be more suitable for cross-lingual ED.

Considering the word order difference, to our best knowledge, (Xie et al., 2018) is the only work which adopted a self-attention mechanism (Vaswani et al., 2017) to tackle this problem (in cross-lingual NER). Difference with them, we propose a shared syntactic order event detector for cross-lingual ED, which explores the syntactic similarity of resources in different languages and circumvents the word order difference problem in performing multilingual co-training.

To illustrate our motivation, consider an English example E1 and its parallel Chinese translation C1 in Figure 1. As shown, E1 and C1 have rather different word orders (Figure 1a), but they share a similar syntactic structure which captures enough generality for identifying event triggers (Figure 1b). This observation motivates us to explore the syntactic similarity to achieve multilingual co-training. To achieve this goal, we propose a shared syntactic order event detector based on Graph Convolutional Networks (GCNs) (Kipf and Welling, 2016), which can provide each word a contextual feature based on its immediate neighbors in the syntactic graph irrespective of its original position in the sentence. This decoder allows us to train a model on multilingual resources effectively, which circumvents the word order different problem.

To estimate our method, we have conducted



(a) The word order rules of E1 and C1.

(b) The syntactic structures of E1 and C1.

Figure 1: The comparison of word orders and syntactic structures of an annotated English sentence E1 and its parallel Chinese sentence C1.

extensive experiments on two standard datasets, using English, Chinese, and Spanish as experimental languages respectively. The experimental results demonstrate that: 1) our model can perform cross-lingual transfer between different language pairs. Especially, the improvement in Chinese ED is large, with an absolute 3.8% in F1 over the monolingual approaches. 2) Our model is robust in the extreme annotation-poor scenario where a language has very limited training data, which demonstrates a definite advantage over previous monolingual models. Additionally, compared with MT-based methods, our model achieves competitive results but requires much less parallel resource.

This paper is organized as follows: Section 2 briefly introduces the task description and terminologies used in ED; Section 3 elaborates details of our approach; Section 4 reports on our experimental results and other analysis; Section 5 reviews related work; Section 6 concludes the paper and illustrates future work.

2 Task Description

Event detection (ED) is a subtask defined in the overall Event Extraction (EE) evaluation in Automatic Content Extraction (ACE) 2005 program. We first introduce some ACE terminologies to facilitate the understanding of ED task.

In ACE, 1) an **event mention** refers to a phrase/sentence within which an event is described. 2) **Event trigger** refers to a specific word in an event mention which is considered the most representative of the event. Each event trigger has a certain type corresponding to the event mention. 3) **Event arguments** are participants of the event.

With these definitions, the goal of ED is to locate event triggers and categorize their types. For example, in sentence "*The old man* **died** in the *hospital*", ED requires a system to detect a `Die` event along with the event trigger **died**. The detection of event arguments *The old man* (role=`Person`) and *hospital* (role=`Place`) is not involved in the ED task.

Following previous work (Nguyen and Grishman, 2015; Liu et al., 2016), we formulate ED as a token-level multi-class classification task. Namely, given a sentence, we treat every token in it as a trigger candidate, and we aim to classify each candidate into one of 34 categories (33 event types defined in ACE in addition to an `NA` type indicating "not an event trigger").

3 Methodology

This study focuses on cross-lingual ED, which aims to transfer knowledge from a source language with abundant labeled data to a target language with insufficient training data.

Figure 2 visualizes the overall architecture of our model, which consists of three main components: (1) Monolingual embedding layer, which transforms each token into a continuous vector representation. (2) Context-dependent lexical mapping, which maps each word in the source language to its best-suited translation in the target language, by examining its contextual representation and imposing a selective attention over different translation candidates. (3) Shared syntactic order event detector, which employs a Graph Convolutional Networks (GCNs) to explore syntactic similarity of resources of different languages, in order to achieve multilingual co-training.

For the sake of convenience, in the following illustrations, we assume the source language is English and the target language is Chinese, and we use an English sentence $s = \{w_1, w_2, \ldots, w_n\}$ to illustrate our idea.

3.1 Monolingual Embedding Layer

In the monolingual embedding layer, each word is assigned to a distributed vector as its representation. Specifically, we first train English/Chinese word embeddings on the corresponding Wikipedia dumps via Skip-gram model (Mikolov et al., 2013) with a dimensional size $d = 300$. And then we transform each token into its word embedding as its vectorized feature representation.

In this way, s is transformed into an embedding matrix $\boldsymbol{E}_s = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n]^T$, where $\boldsymbol{x}_i \in \mathbb{R}^d$ indicates the word embedding of the token w_i.

3.2 Context-Dependent Lexical Mapping

For each token w_i in s, context-dependent lexical mapping aims to search for its best-suited word translation according to its contextual representation. This process involves: 1) learning multilingual alignment, 2) retrieving translation candidates, and 3) ranking translation words via a selective attention mechanism.

3.2.1 Learning Multilingual Alignment

Let $\boldsymbol{X}$ and $\boldsymbol{Y}$ be the English and Chinese embedding spaces. In order to achieve multilingual alignment, we learn a mapping $\boldsymbol{W} \in \mathbb{R}^{d \times d}$ from $\boldsymbol{X}$ to $\boldsymbol{Y}$ via a seed dictionary with a size of m, by optimizing:

$$\boldsymbol{W}^* = \underset{\boldsymbol{W} \in M_d(\mathbb{R})}{\arg\min} ||\boldsymbol{W}\boldsymbol{X}_{dic} - \boldsymbol{Y}_{dic}||_F \quad (1)$$

where $M_d(\mathbb{R})$ is the space of $d \times d$ matrices; $\boldsymbol{X}_{dic}$, $\boldsymbol{Y}_{dic} \in \mathbb{R}^{d \times m}$ are two matrices containing the aligned embeddings of words in the seed dictionary; $||\cdot||_F$ indicates the Frobenius norm. To get a closed form solution, following (Xing et al., 2015; Lample et al., 2018), we impose an orthogonality constraint on $\boldsymbol{W}$ (i.e.,$\boldsymbol{W}\boldsymbol{W}^T$=$\boldsymbol{W}^T\boldsymbol{W}$=$\boldsymbol{I}$), and in this way, the optimized solution of $\boldsymbol{W}$ corresponds to the singular value decomposition (SVD) of $\boldsymbol{Y}_{dic}\boldsymbol{X}_{dic}^T$, i.e.,

$$\boldsymbol{W}^* = \underset{\boldsymbol{W} \in O_d(\mathbb{R})}{\arg\min} ||\boldsymbol{W}\boldsymbol{X}_{dic} - \boldsymbol{Y}_{dic}||_F = \boldsymbol{U}\boldsymbol{V}^T \quad (2)$$

where $\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T = \mathrm{SVD}(\boldsymbol{Y}_{dic}\boldsymbol{X}_{dic}^T)$.

3.2.2 Retrieving Translation Candidates

Next, we retrieve translation candidates for each token w_i in s. Specifically, we first project w_i into the aligned embedding space (i.e., by applying $\boldsymbol{W}$ on $\boldsymbol{x}_i$), and then we explore its neighborhood to find the nearest Chinese words as its translation candidates. In order to measure the distance

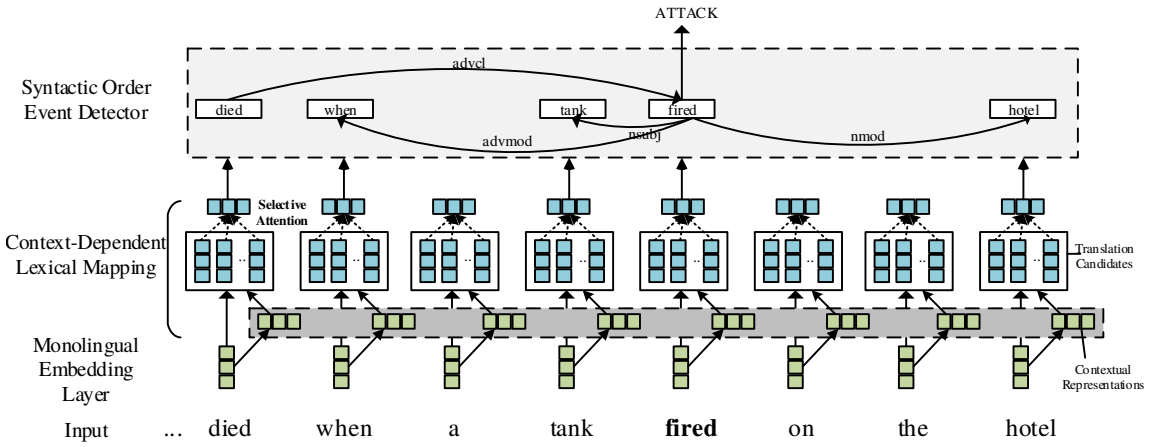


Figure 2: The overview architecture of our model. The figure illustrates the process of performing cross-lingual transfer for an English sentence "A man died when a tank fired on the hotel" into Chinese and using the shared syntactic order event detector to predict the event type for the word "fired".

of w_i and a Chinese word y in the aligned space, we adopt the cross-domain similarity local scaling (CSLS) metric (Lample et al., 2018):

$$\begin{aligned}\text{CSLS}(\boldsymbol{W}\boldsymbol{x}_i, \boldsymbol{y}) =& 2\cos(\boldsymbol{W}\boldsymbol{x}_i, \boldsymbol{y}) \\ & - r_{\boldsymbol{Y}}(\boldsymbol{W}\boldsymbol{x}_i) - r_{\boldsymbol{X}}(\boldsymbol{y})\end{aligned} \tag{3}$$

where $\boldsymbol{y}$ denotes the (Chinese) word embedding of y; $r_{\boldsymbol{Y}}(\boldsymbol{W}\boldsymbol{x}_i)$ indicates the mean cosine similarity between $\boldsymbol{W}\boldsymbol{x}_i$ and its K neighbors in $\boldsymbol{Y}$, which is defined as: $r_{\boldsymbol{Y}}(\boldsymbol{W}\boldsymbol{x}_i) = \frac{1}{K}\sum_{\boldsymbol{y}' \in \mathcal{N}_{\boldsymbol{Y}}(\boldsymbol{W}\boldsymbol{x}_i)} \cos(\boldsymbol{W}\boldsymbol{x}_i, \boldsymbol{y}')$ where $\mathcal{N}_{\boldsymbol{Y}}(\boldsymbol{W}\boldsymbol{x}_i)$ denotes the neighborhood associate with $\boldsymbol{W}\boldsymbol{x}_i$ in $\boldsymbol{Y}$. In our method, for w_i, we take J Chinese words which have the smallest CSLSs as its translation candidates. We denote by $T^{(w_i)}$ the set of translation candidates for w_i, where $T_j^{(w_i)}$ indicates the jth element of $T^{(w_i)}$.

3.2.3 Content-Aware Selective Attention

Finally, for each token w_i, we perform a context-aware selective attention mechanism to weigh each translation candidate in $T^{(w_i)}$ and get the best-suited translation for it.

Learning Contextual Representation. We employ the self-attention mechanism (Vaswani et al., 2017) to learn context representation of w_i. Specifically, given $\boldsymbol{E}_s = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n]^T$, we use different single-layer neural networks to learn queries $\boldsymbol{Q}$, keys $\boldsymbol{K}$, values $\boldsymbol{V}$ respectively. For example, $\boldsymbol{Q} = \tanh(\boldsymbol{E}_s\boldsymbol{W}_m + \boldsymbol{b}_m)$, where $\boldsymbol{W}_m \in \mathbb{R}^{d\times d}$ and $\boldsymbol{b} \in \mathbb{R}^d$ are parameter matrix and bias respectively. Then, we compute a self-attention matrix by computing:

$$\boldsymbol{C}_s = \text{softmax}(\frac{\boldsymbol{Q}\boldsymbol{K}^\mathsf{T}}{\sqrt{d}})\boldsymbol{V} \tag{4}$$

$$= [\boldsymbol{c}_1, \boldsymbol{c}_2, \ldots, \boldsymbol{c}_n]^T \tag{5}$$

where d indicates the word embedding dimension. We take $\boldsymbol{c}_i$ as the contextual representation of w_i.

Learning Selective Attention. For each token w_i, after obtaining its translation candidates list $T^{(w_i)}$ and contextual representation $\boldsymbol{c}_i$, we impose a selective attention mechanism to automatically weigh each candidate. Specifically, the weight of the jth candidate $T_j^{(w_i)}$ is computed as:

$$\alpha_j = \frac{\exp(m_j)}{\sum_{i=1}^{J}\exp(m_i)} \tag{6}$$

where m_j measures the semantic relatedness of $\boldsymbol{c}_i$ and $T_j^{(w_i)}$, which is computed by:

$$m_j = \tanh([\boldsymbol{c}_i; \boldsymbol{y}_j^{(w_i)}]\boldsymbol{W}_r + b_r) \tag{7}$$

where [;] indicates the concatenation operations; $\boldsymbol{y}_j^{(w_i)}$ denotes the Chinese word embedding of $T_j^{(w_i)}$; $\boldsymbol{W}_r \in \mathbb{R}^{d\times 1}$ and $b_r \in \mathbb{R}$ are parameter matrix and bias respectively. Finally, we select the candidate which has the maximal attention weight as the best-suited translation for w_i, which is denoted by w_i'. In this way, the original sentence s is transfer into a Chinese word sequences $t = \{w_1', w_2', \ldots, w_n'\}$ with a same length.

3.3 Shared Syntactic Order Event Detector

As English and Chinese usually have different word orders, the transferred result t might be seen

as a corrupted sentences from Chinese, which could introduce noise for multilingual co-training. We tackle this problem by proposing a Graph Convolutional Neural Networks (GCNs) (Kipf and Welling, 2016) based syntactic order event detector, which provides each word with a feature vector based on its immediate neighbors in the syntactic graph irrespective of its position in the sentence. This allows our model to train with the translated data t and the other labeled data in Chinese indiscriminately.

3.3.1 Extracting Graph Convolution Feature

Specifically, for each token w_i, our model computes a graph convolution feature vector based on its immediate neighbors in the syntactic graph. Figure 3 illustrates the process of extracting the feature for "fired".

Let $\mathcal{N}(w_i)$ denote the set of neighbors of w_i in the syntactic graph, and $L(w_i, v)$ indicate the label of the dependency arc $(w_i \rightarrow v)$ (For example, L("fired", "hotel") = nmod in the example in Figure 3). The original GCNs compute a graph convolution vector for w_i at (k+1)th layer by:

$$\boldsymbol{h}_{w_i}^{k+1} = g\Big(\sum_{v \in \mathcal{N}(w_i)} (\boldsymbol{W}_{L(w_i,v)}^{k} \boldsymbol{h}_{v}^{k}) + \boldsymbol{b}_{L(w_i,v)}^{k}\Big) \tag{8}$$

where g denotes the ReLU function; $\boldsymbol{W}_{L(w_i,v)}^{k}$ and $\boldsymbol{b}_{L(w_i,v)}^{k}$ are parameters of the dependency label $L(w_i, v)$ in the kth layer. However, retaining parameters for every dependency label is space-consuming and compute-intense (there are approximately 50 labels), in our model, we limit $L(w_i, v)$ to have only three types of labels 1) an original edge, 2) a self loop edge, and 3) an added inverse edge, as suggested in (Nguyen and Grishman, 2018). Additionally, since the generated syntactic parsing structures usually contain noise, we apply attention gates on the edges to weigh their individual importances:

$$\alpha_{(w_i,v)}^{k} = \sigma(\boldsymbol{h}_{u}^{k} \boldsymbol{U}_{L(w_i,v)}^{k} + \boldsymbol{d}_{L(w_i,v)}^{k}) \tag{9}$$

where σ is the logistic sigmoid function. $\boldsymbol{U}_{L(w_i,v)}^{k}$ and $\boldsymbol{d}_{(w_i,v)}^{k}$ are the weight matrix and the bias of the gate. With this gating mechanism, the final syntactic GCNs computation in our model is:

$$\boldsymbol{h}_{\boldsymbol{u}}^{k+1} = g\Big(\sum_{\boldsymbol{v} \in \mathcal{N}(\boldsymbol{u})} \alpha_{\boldsymbol{u},\boldsymbol{v}}^{k} (\boldsymbol{W}_{L(\boldsymbol{u},\boldsymbol{v})}^{k} \boldsymbol{h}_{\boldsymbol{v}}^{k}) + \boldsymbol{b}_{L(\boldsymbol{u},\boldsymbol{v})}^{k}\Big) \tag{10}$$

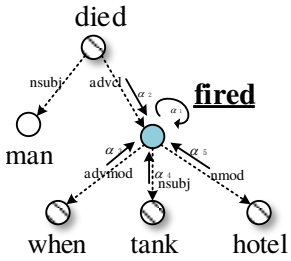


Figure 3: The illustration of using GCNs to compute the order-invariant feature for the word "fired".

We set the initial vectors $\boldsymbol{h}_{w_i}^{0}$ for w_i as the Chinese word embedding of w_i' (its translated word), and we stack 2 layers of GCNs (i.e., k = 2) to obtain the final feature for w_i, denoted as $\mathbf{f_i}$.

3.4 Event Type Classification

Our model incorporates a logistic regression classifier to predict w_i's event type. Specifically, we compute a prediction vector for w_i by taking $\boldsymbol{f}_i$ as the input:

$$\boldsymbol{out} = \text{softmax}(\boldsymbol{W}_o \boldsymbol{f}_i + \boldsymbol{b}_o) \tag{11}$$

where $\boldsymbol{W}_o \in \mathbb{R}^{d \times c}$ and $\boldsymbol{b}_o \in \mathbb{R}^{c}$ are parameters, and c is the total number of event types (i.e., 34 in this study). The probability of t-th class type is denoted as $P(t|w_i)$, which corresponds to the t-th element of $\boldsymbol{out}$.

3.5 Multilingual Co-Training

To enable multilingual co-training, we adopt the cross-entropy loss, and we use λ to balance the contribution of multilingual resources (which is set as 0.7 through a grid search):

$$J(\Theta) = -\lambda \sum_{w_e'} \log P(l_{w_e'}|w_e') - \sum_{w_c} \log P(l_{w_c}|w_c) \tag{12}$$

where Θ denotes all the parameters in our model; w_e' ranges over each token in the translated examples and w_c enumerate each token in the original Chinese training set; $l_{w_e'}$ and l_{w_c} denote the ground-truth event types of w_e' and w_c respectively. We adopt Adam rules (Kingma and Ba, 2014) to update our model's parameters and add dropout layers to prevent over-fitting.

4 Experiments

4.1 Datasets and Evaluation

Our main experiments are conducted on ACE 2005 and TAC KBP 2017, two widely used ED

datasets which contain annotated documents in English and Chinese (The documents are not parallel). For ACE English ED, we split the dataset to training/development/test set as suggested in (Li et al., 2013; Nguyen and Grishman, 2015). For ACE Chinese ED, we perform a 10-fold cross-validation as suggested in (Chen and Ji, 2009; Lin et al., 2018). For TAC KBP 2017 English and Chinese, we use the official test sets for testing, and we split the remaining data with a ratio of 9:1 for training and developing. The bilingual dictionary is obtained from the MUSE project[1] with a size of $5k$. The number of candidate translations J (in Section 3.2.2) is set as 3. We use the Stanford CoreNLP (Manning et al., 2014) to obtain syntactic trees for each language.

Precision (Pre.), Recall (Rec.), and F1-score (F1) are used as evaluation metrics, same as previous ED studies to ensure compatibility. Significant tests (with p=0.05) were conducted using method described in (Yeh, 2000).

4.2 Experimental Results

We conduct two groups of experiments to investigate the ability of our model in 1) performing cross-lingual transfer concerning different language directions, and 2) handling the annotation-poor scenario.

4.2.1 Cross-Lingual Transfer Concerning Different Language Directions

We investigate both English-to-Chinese and Chinese-to-English transfers to investigate whether cross-lingual transfer is feasible concerning different language directions. In these experiments, we jointly train on the translated data with all the labeled data in target language.

We compare our cross-lingual approach (*CL_Trans*) with our monolingual approach (*Monolingual*, which only uses the training data of the target language) and the existing state-of-the-art monolingual ED models (*Monolingual_SOTA*). In ACE ED, we take models proposed in (Nguyen and Grishman, 2015) and (Lin et al., 2018) as the SOTA English and Chinese ED systems respectively. In TAC KBP ED, we take the top systems reported in the official evaluation (Mitamura et al., 2017) as *Monolingual_SOTA*. We also include a vanilla embedding projection based method for comparison (denoted as *Embedding_proj*).

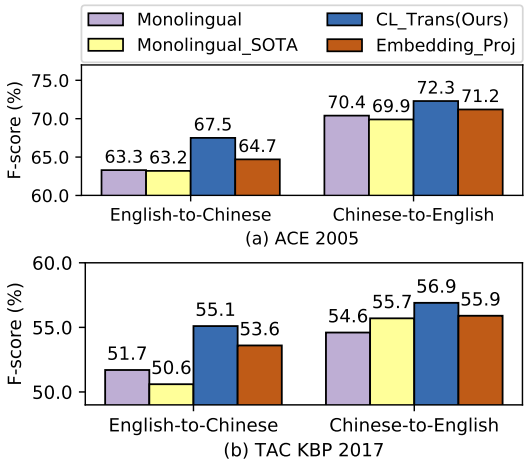


Figure 4: Experimental results on investigating cross-lingual transfer concerning different language directions.

Figure 4 summarizes the results. From the results, 1) our cross-lingual approach *CL_Trans* outperforms two monolingual systems (+2.95% on F1 on the average) and the vanilla embedding projection based method (+1.60% on F1), in the four evaluations. This justifies the effectiveness of our approach concerning cross-lingual in different language directions. 2) Additionally, our cross-lingual approach is more effective for Chinese (+3.80% on F1) than for English (+2.10% on F1). This is understandable as the number of English examples is much larger than that of Chinese examples (5,285 v.s. 2,710 in ACE 2005, and 24,979 v.s. 10,630 in TAC KBP 2017). 3) We obtain interesting findings by investigating each event type. For example, in TAC KBP 2017, the type of "contact/correspondence" has only 167 samples in Chinese but 996 samples in English. By adopting cross-lingual training, our approach leads to an improvement of 15.3% (from 10.3% to 25.6%) in Chinese ED for this type, compared with the monolingual approach. This proves that our method can handle the annotation sparseness problem in the target language.

4.2.2 Exploring the Annotation-Scarce Scenario

We next investigate the annotation-poor scenario, where the source language is set as English and the target language is set as Chinese to compare with previous works. In this scenario, we assume that only a few of annotated documents are available in

[1] https://github.com/facebookresearch/MUSE

Chinese.

In Comparison to Monolingual ED Models. We first compare our cross-lingual approach with the existing monolingual ED models, including *CNN* (Nguyen and Grishman, 2015) which employs Convolutional Neural Networks for the task and *Hbrid* (Feng et al., 2016) which combines CNN with Recurrent Neural Networks (RNN) for ED. Figure 5 presents the experimental results, where the number of the available Chinese documents ranges from 0 to 50.

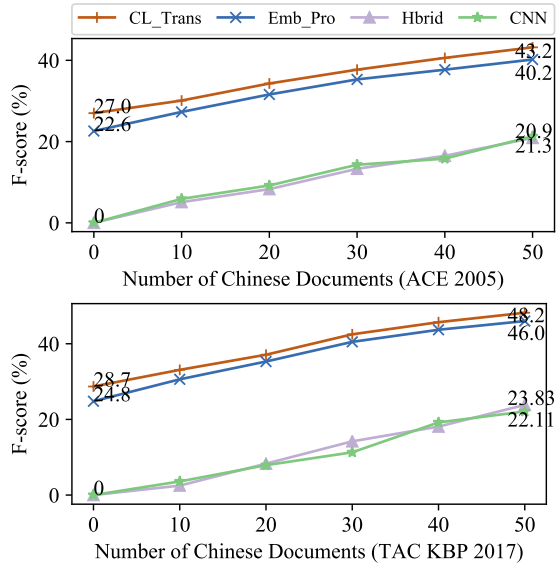


Figure 5: The comparison to monolingual ED models for Chinese ED in the annotation-poor scenario.

From the results, our approach demonstrates a definite advantage over monolingual ED models in the annotation-poor scenario. Particularly, when there is no Chinese training document available (i.e., in the unsupervised cross-lingual transfer scenario), our model achieves a performance of 27.0% on F1 in ACE, and 28.7% on F1 in TAC KBP, while supervised ED methods completely fail. Additionally, our approach can consistently outperform the embedding-projection method.

In Comparison to Cross-Lingual Models. We next compare our model to the existing cross-lingual ED methods, including 1) LexMap (Hsi et al., 2016), which combines embedding projection method with multilingual feature extraction to perform cross-lingual ED, and 2) MTED (Zhu et al., 2014), which uses a MT system to translate the training examples in source language to obtain additional data for training. In our re-implementation, we employ OpenNMT (Klein et al.) as the translation model and we use OpenSubtitles (Lison and Tiedemann, 2016) to train it. To ensure comparability of results, we use the setting of (Hsi et al., 2016) (i.e., using one-fold data (64 annotated documents) for training) to conduct the experiments.

Method	Pre.	Rec.	F1
LexMap (2016) (5k dict.)	**63.5**	28.8	39.6
MTBased (50k sent.)	11.3	16.8	13.5
MTBased (200k sent.)	42.4	26.1	32.3
MTBased (400k sent.)	54.7	**36.2**	43.7
CL_Trans (5k dict.)	62.5	35.7	**45.4**

Table 1: The comparison to cross-lingual ED models for Chinese ED in ACE 2005. (50k sent.) means using 50k parallel sentences to train the MT system.

Table 1 gives the results. From the results, 1) our method outperforms LexMap by a rather large margin (+5.8% on F1). The poor performance of LexMap might be attributed to its feature engineering process, which is often very difficult and requires expert knowledge. 2) Our model behaves competitively to machine translation based method (which are trained on 400k parallel sentences) yet relies on much less parallel resources (a dictionary with a size of 5k).

4.2.3 Ablation Study

We conduct ablation study to explore the effects of our different model components. We limit our study in the extremely annotation-poor scenario, that is, we assume there is no training data in the target language (Chinese).

Exploring Lexical Mapping Method. To explore our lexical mapping method, we compare the performance of several variant systems retrieving a different number of candidates (ranging from 1 to 5) and the embedding-projection method (Embedding_proj). Note the system retrieving only one candidate actually takes the nearest Chinese neighbor as the word translation. The lexical mapping in it is still context-independent. Table 2 summarizes the results.

From the results, we observe that 1) even though both of CL_Trans (1 cand.) and Embedding_proj are content-independent mapping methods, the former outperforms the latter by a margin (+3.2% on F1). This implies that the embedding-projection method might suffer from the misalignment in the shared embedding space, and enforcing a word-to-word alignment (as in CL_Trans (1 cand.)) could alleviate this problem to some ex-

Method	Pre.	Rec.	F1
Embbeding_Proj	26.0	20.0	22.6
CL_Trans (1 cand.)	31.2	21.4	25.4
CL_Trans (2 cand.)	31.7	22.3	26.2
CL_Trans (3 cand.)	**32.0**	23.4	**27.0**
CL_Trans (4 cand.)	30.7	**23.6**	26.7
CL_Trans (5 cand.)	30.2	**23.6**	26.5

Table 2: Experimental results in exploring different lexical mapping methods.

Model	Pre.	Rec.	F1
CL_Trans_MLP	20.3	16.3	18.1
CL_Trans_CNN	32.5	16.3	21.7
CL_Trans_Hbrid	30.4	17.6	22.3
CL_Trans_Self.	**34.9**	18.3	24.0
CL_Trans_GCN (ours)	32.0	**23.4**	**27.0**
CL_Trans_GCN_Self	32.1	23.0	26.8

Table 3: Experimental results in exploring the shared syntactic order event detector.

English	Chinese Translation Candidates
36,000	5.76, 98.8%, 2.53
people	**人(people)**, 人们(folk), 青年人(youngs)
died	**死(die)**, 死亡(death), 死去(dying)
every	**每个(every)**, 所有(all), 都(all)
year	**年(year)**, 年代(years), 年龄(age)
from	以外(beyond), **从(from)**, 外(except)
the	这(this), 整个(total), 完全(completely)
flu	感染(inflection), **流感(flu)**, 疾病(disease)

Table 4: Translation candidates of each English word. Bold indicates the best-suited translation.

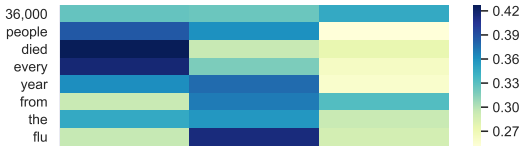


Figure 6: The learned attention weights. Darker color indicates higher weight.

tent. 2) Retrieving more translation candidates could consistently improve Recall. But when too many candidates (e.g., 5) are added, the Precision drops, which harms the overall F1 measure.

Exploring the Syntactic Order Event Detector. We compare our syntactic order event detector (CL_Trans_GCN) with several event detectors, including 1) *CL_Trans_MLP*, which employs a feed-forward network as event detector; 2) CL_Trans_CNN, which uses CNNs as the event detector; 3) CL_Trans_Hbrid, which use a hybrid network (Feng et al., 2016) for event detection. We also compared our model with several variants including 4) CL_Trans_Self., which replaces the GCNs with a self-attention network, and 5) CL_Trans_GCN_Self, which combines GCNs with a self-attention network. We train these models on the same translated English data. Table 3 shows the results.

From the results, 1) CL_Trans_MLP, CL_Trans_CNN, and CL_Trans_Hbrid behave poorly, as expected. The reason might be that these models usually employ order-sensitive structures (e.g., CNNs) for ED, which would suffer the word order inconsistency problem when trained on the translated data. 2) CL_Trans_Self. yields relatively good performance. The reason might be that self-attention network could provide each word with a feature vector based on all the words of a sentence, which is also irrespective of the words' positions in a sentence. This could address the word order difference to some extent. 3) Our syntactic order event detector yields the best performance. While, we do not observe salient advantages by combining GCNs with a self-attention network (by comparing CL_Trans_GCN with CL_Trans_GCN_Self).

4.3 Beyond English-Chinese Pair

We conduct additional experiments on Spanish to investigate cross-lingual transfer beyond English-Chinese transfer. The Spanish corpus is in TAC-KBP solely, with much smaller size and fewer publish evaluations. Experimental results demonstrate that, our model, without any modification, surpasses the best-reported Spanish system (42.8 (Mitamura et al., 2017) with F1 scores of 44.0 and 43.8 concerning EN $\rightarrow$ SP and CH $\rightarrow$ SP transfer. The value changes to 20.8 and 18.9 in zero transfer scenarios. This demonstrates that our approach is language-independent.

4.4 Case Study

We give a case study on the cross-lingual transfer process of a real example in ACE: "*36,000 people* **died** *every year from the flu*". Table 4 and Figure 6 gives the Chinese translation candidates and the learned attention weights respectively.

From the results, the best-suited translations indeed often correspond to larger attention weights, which implies the validity of our approach. In the above example, the Chinese words "从(from)" and "流感(flu)" do not correspond to the nearest

neighbor of “from” and ”flu”, but our content-dependent lexical mapping method enable us to successfully obtain them as the translations.

The case study also poses several future directions for this work. For example, one is how to address the one-to-many mapping between different languages. In the above example, the correct Chinese translation of “every year” should be one single word “每年”, not the combination of two words “每个(every)” and “年(year)”. This calls for more advanced lexical mapping methods.

5 Related Work

Event detection (ED) is a hot topic in natural language processing, which has attracted extensive attention in the past few years. Traditionally, the study of ED has focused on monolingual training. The proposed models can be divided into feature-based methods which employ fine-grained features (Ahn, 2006; Ji and Grishman, 2008; Liao and Grishman, 2010; Hong et al., 2011; Li et al., 2013; Li and Ji, 2014), and deep learning-based methods which employ neural networks to automatically learn features for the task (Chen et al., 2015; Nguyen and Grishman, 2015; Nguyen et al., 2016; Liu et al., 2018b; Orr et al., 2018; Liu et al., 2019). Usually, their performance is limited by the amount of labeled data in a specific language.

Cross-lingual ED attempts transfer knowledge between different languages to boost performance. To name a few, (Chen and Ji, 2009) used an English detector to label events on parallel documents to obtain additional data for boosting Chinese ED; (Zhu et al., 2014; Liu et al., 2018a) used machine translation to obtain additional labeled data for training; (Hsi et al., 2016) combined the embedding-projection method with multilingual feature extraction for bilingual ED. Nevertheless, the heavily dependency on parallel resources often limits the applicability of these methods.

Our study also relates to cross-lingual studies in other applications (Guo et al., 2015; Ni et al., 2017; Mayhew et al., 2017; Xie et al., 2018; Lample et al., 2018). These approaches adopted embedding projection based method to achieve cross-lingual transfer and achieved prosing results. However, since the lexical mapping in these methods is usually deterministic and irrespective of contexts, they might not directly fit with cross-lingual ED, where the cross-lingual transfer should be context-dependent.

6 Conclusions and Future Work

In this paper, we propose a new cross-lingual approach for event detection, which demonstrates a minimal dependency on parallel resources. Specifically, we propose a context-dependent lexical mapping method to obtain content-dependent translation, and we devise a shared syntactic order event detector to explore the syntactic similarity for multilingual co-training. Experiments demonstrate the effectiveness of our method.

Currently, as our approach is predicated on the availability of syntax trees of training examples, it might not fit with languages which lack syntactic parsers. In the future, we plan to investigate more language-independent patterns in cross-lingual transfer to circumvent this dependency.

Acknowledgments

This work is supported by the National Key R&D Program of China under Grant 2018YFB1005100, the National Natural Science Foundation of China (No.61533018), the National Natural Science Foundation of China (No.61806201) and the independent research project of National Laboratory of Pattern Recognition. This work is also supported by a grant from Ant Financial Services Group and the CCF-Tencent Open Research Fund. We would like to thank the anonymous reviewers for their valuable feedback.

References

David Ahn. 2006. The stages of event extraction. In *Proceedings of the Workshop on Annotating and Reasoning About Time and Events*, ARTE ’06, pages 1–8, Stroudsburg, PA, USA. Association for Computational Linguistics.

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. 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*, SemiSupLearn ’09, pages 66–74, Stroudsburg, PA, USA. Association for Computational Linguistics.

Xiaocheng Feng, Lifu Huang, Duyu Tang, Heng Ji, Bing Qin, and Ting Liu. 2016. A language-independent neural network for event detection. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 66–71, Berlin, Germany. Association for Computational Linguistics.

Jiang Guo, Wanxiang Che, David Yarowsky, Haifeng Wang, and Ting Liu. 2015. Cross-lingual dependency parsing based on distributed representations. 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 1234–1244. Association for Computational Linguistics.

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. The COLING 2016 Organizing Committee.

Heng Ji. 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*, UMSLLS '09, pages 27–35, Stroudsburg, PA, USA. 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.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Thomas N. Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *CoRR*, abs/1609.02907.

G. Klein, Y. Kim, Y. Deng, J. Senellart, and A. M. Rush. OpenNMT: Open-Source Toolkit for Neural Machine Translation. *ArXiv e-prints*.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, pages 177–180, Stroudsburg, PA, USA. Association for Computational Linguistics.

Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *International Conference on Learning Representations*.

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. 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. Association for Computational Linguistics.

Hongyu Lin, Yaojie Lu, Xianpei Han, and Le Sun. 2018. Nugget proposal networks for chinese event detection. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1565–1574. Association for Computational Linguistics.

Pierre Lison and Jörg Tiedemann. 2016. Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles. In *LREC*.

Jian Liu, Yubo Chen, and Kang Liu. 2019. Exploiting the Ground-Truth: An Adversarial Imitation Based Knowledge Distillation Approach for Event Detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):6754–6761.

Jian Liu, Yubo Chen, Kang Liu, and Jun Zhao. 2018a. Event detection via gated multilingual attention mechanism. *Proceedings of the AAAI Conference on Artificial Intelligence*.

Shaobo Liu, Rui Cheng, Xiaoming Yu, and Xueqi Cheng. 2018b. Exploiting contextual information via dynamic memory network for event detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1030–1035. Association for Computational Linguistics.

Shulin Liu, Yubo Chen, Shizhu He, Kang Liu, and Jun Zhao. 2016. Leveraging framenet to improve automatic event detection. In *Proceedings of the*

54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 2134–2143, Berlin, Germany. Association for Computational Linguistics.

Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J Bethard, and David McClosky. 2014. The {Stanford} {CoreNLP} Natural Language Processing Toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

Stephen Mayhew, Chen-Tse Tsai, and Dan Roth. 2017. Cheap translation for cross-lingual named entity recognition. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2536–2545, Copenhagen, Denmark. Association for Computational Linguistics.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, pages 3111–3119, USA. Curran Associates Inc.

Teruko Mitamura, Zhengzhong Liu, and Eduard H. Hovy. 2017. Events detection, coreference and sequencing: What's next? overview of the tac kbp 2017 event track. In *TAC*.

Thien Nguyen and Ralph Grishman. 2018. Graph convolutional networks with argument-aware pooling for event detection. *Proceedings of the AAAI Conference on Artificial Intelligence*.

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. 2015. 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.

Jian Ni, Georgiana Dinu, and Radu Florian. 2017. Weakly supervised cross-lingual named entity recognition via effective annotation and representation projection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1470–1480. Association for Computational Linguistics.

Walker Orr, Prasad Tadepalli, and Xiaoli Fern. 2018. Event detection with neural networks: A rigorous empirical evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 999–1004, Brussels, Belgium. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Jiateng Xie, Zhilin Yang, Graham Neubig, Noah A. Smith, and Jaime Carbonell. 2018. Neural cross-lingual named entity recognition with minimal resources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 369–379. Association for Computational Linguistics.

Chao Xing, Dong Wang, Chao Liu, and Yiye Lin. 2015. Normalized word embedding and orthogonal transform for bilingual word translation. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1006–1011. Association for Computational Linguistics.

Alexander Yeh. 2000. More accurate tests for the statistical significance of result differences. In *COLING 2000 Volume 2: The 18th International Conference on 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.