Contextualized Cross-Lingual Event Trigger Extraction with Minimal Resources

Meryem M'hamdi , Marjorie Freedman and Jonathan May
Information Sciences Institute & Computer Science Department
University of Southern California (USC)
{meryem, mrf, jonmay}@isi.edu

Abstract

Event trigger extraction is an information extraction task of practical utility, yet it is challenging due to the difficulty of disambiguating word sense meaning. Previous approaches rely extensively on hand-crafted language-specific features and are applied mainly to English for which annotated datasets and Natural Language Processing (NLP) tools are available. However, the availability of such resources varies from one language to another. Recently, contextualized Bidirectional Encoder Representations from Transformers (BERT) models have established state-of-the-art performance for a variety of NLP tasks. However, there has not been much effort in exploring language transfer using BERT for event extraction.

In this work, we treat event trigger extraction as a sequence tagging problem and propose a cross-lingual framework for training it without any hand-crafted features. We experiment with different flavors of transfer learning from high-resourced to low-resourced languages and compare the performance of different multilingual embeddings for event trigger extraction. Our results show that training in a multilingual setting outperforms language-specific models for both English and Chinese. Our work is the first to experiment with two event architecture variants in a cross-lingual setting, to show the effectiveness of contextualized embeddings obtained using BERT, and to explore and analyze its performance on Arabic.

1 Introduction

Event trigger extraction, as defined the Automatic Content Extraction multilingual evaluation benchmark (ACE2005) (Walker, 2006), is a subtask of event extraction which requires systems to detect and label the lexical instantiation of an event, known as a *trigger*. As an example, in the sentence "John **traveled** to NYC for a meeting", **traveled** is a trigger of a *Movement-Transport* event. Trigger detection is typically the first step in extracting the structured information about an event (e.g. the time, place, and participant arguments; distinguishing between past, habitual, and future events). This definition of the task restricts it to events that can be triggered explicitly by actual words and makes it context-vulnerable: the same event might be expressed by different triggers and a specific trigger can represent different event types depending on the context.

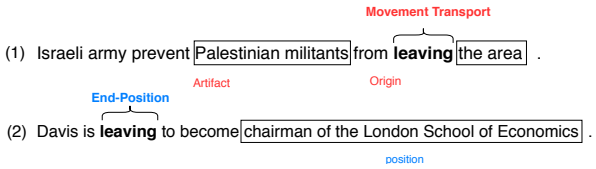

Figure 1: Examples of importance of context in trigger disambiguation.

Event trigger extraction is challenging as it involves understanding the context in order to be able to identify the event that the trigger refers to. Figure 1 shows two examples where context plays a crucial role in disambiguating the word sense of **leaving**, which is a trigger for a *Movement-Transport* event in the first sentence and for an *End-Position* event in the second sentence.

Due to the complexity of the task and the difficulty in constructing a standard annotation scheme, there exists limited labeled data, for only a few languages. The earliest work has focused mainly on English, for which there are relatively many annotated sentences, and relies extensively on language-specific linguistic tools to extract the lexical and syntactic features that need to be computed as a pre-requisite for the task (Ji and Grishman, 2008; Liao and Grishman, 2010; Hong et al., 2011; Li et al., 2013).

Simply generating annotated corpora for each

Proceedings of the 23rd Conference on Computational Natural Language Learning, pages 656–665
Hong Kong, China, November 3-4, 2019.

language of interest is not only costly and time-consuming, it is also not necessarily guaranteed to address the **great NLP divide**, where performance depends on the language, the ability to generate language-specific features, and the quality tools (in this case, syntactic parsers) available for each language. In an attempt to reduce the great NLP divide, we observe a tendency of practitioners drifting away from linguistic features and more towards continuous distributed features that can be obtained without hand-engineering, based simply on publicly available corpora. Recently, approaches have tried to overcome the limitation of traditional lexical features, which can suffer from the data sparsity problem and inability to fully capture the semantics of the words, by making use of sequential modeling methods including variants of Recurrent Neural Networks (RNN) and Convolutional Neural Networks (CNN), and/or Conditional Random Fields (CRF). (Chen et al., 2015; Liu et al., 2016; Nguyen et al., 2016; Sha et al., 2018; Liu et al., 2018b).

Existing approaches which take into consideration the cross-lingual aspect of event trigger extraction tend to either take inspiration from machine translation, distant supervision or multitasking. Machine translation is used by Liu et al. (2018a) to project monolingual text to parallel multilingual texts to model the confidence of clues provided by other languages. However, this approach suffers from error propagation of machine translation.

Another approach relies on multilingual embeddings, which can be pre-trained beforehand on large monolingual corpora, using no explicit parallel data, and bridging the gap between different languages by learning a way to align them into a shared vector space. The ability of these models to represent a common representation of words across languages makes them attractive to numerous downstream NLP applications. Multilingual Unsupervised and Supervised Embeddings (MUSE) is a framework for training cross-lingual embeddings in an unsupervised manner, which leads to competitive results, even compared to supervised approaches (Conneau et al., 2017). However, there is no prior work leveraging this kind of representation for cross-lingual event trigger extraction.

More recently, BERT, a deep bidirectional representation which jointly conditions on both left and right context (Devlin et al., 2019), was proposed, which unlike MUSE, provides *contextualized* word embeddings, and has been shown to achieve state-of-the-art performance on many NLP tasks. In particular, (Yang et al., 2019) propose a method based on BERT for enhancing event trigger and argument extraction by generating more labeled data. However, it has not been applied in the context of cross-lingual transfer learning.

In this paper, we investigate the possibility of automatically learning effective features from data while relying on zero language-specific linguistic resources. Moreover, we explore the application of multilingual embeddings to the event trigger extraction task in a **direct transfer of annotation scheme** where ground truth is only needed for one language and can be used to predict labels in other languages and **other boosted and joint multilingual schemes**. We perform a systematic comparison between training using monolingual versus multilingual embeddings and the difference in gain on performance with respect to different train/test language pairs. We evaluate our framework using two embedding approaches: type-based unsupervised embeddings (MUSE) and contextualized embeddings (BERT). For the latter, we demonstrate that our proposed model achieves a better (or comparable) performance for all languages compared to some benchmarks for event extraction on the ACE2005 dataset.

The main contributions of the paper can be summarized as follows:

(1) We apply different state-of-the-art NN architectures for sequence tagging on trigger extraction and compare them to feature-based baselines and multilingual projection based models.

(2) We achieve a better performance using contextualized word representation learning in event trigger extraction backed up with both quantitative and qualitative analysis.

(3) We evaluate the effectiveness of a multilingual approach using zero-shot transfer learning, targeted cross-lingual and joint multilingual training schemes.

(4) We investigate event trigger extraction performance when using Arabic.

2 Methodology

We treat trigger extraction as a sequence tagging problem for which we start by designing a ba-

sic state-of-the-art approach for sequence tagging based on bidirectional Long Short Term Memory (bi-LSTM) with word and character embeddings and a CRF layer on top of it. Then, we describe an approach that trains BERT with a CRF layer for the task. In both architectures, the input is in the form of BIO notation used to differentiate between the beginning (B), inside (I) and (O) for no associated trigger labels.

2.1 Bi-LSTM-Char-CRF networks

The Bi-LSTM-Char-CRF for sequence tagging model is a hierarchical neural network model based on three components: character-level using character embeddings, word-level using bi-LSTM over word embeddings and sequence-level using CRF. The architecture of the model is depicted in Figure 2.

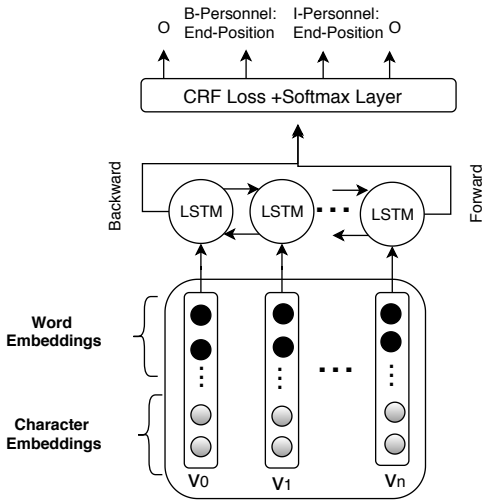

Figure 2: Bidirectional LSTM with character embeddings and CRF layer

2.1.1 Bi-LSTM networks

LSTMs (Hochreiter and Schmidhuber, 1997) are variants of RNNs that help learn long-range dependencies efficiently thanks to their use of memory and forget cells. Those cells help control the amount of the input to be retained/forgotten from previous states.

Given an input character or word embeddings representation x_t for a given time step t, we use bidirectional LSTMs by encoding features in their forward: $fh_i = \overrightarrow{LSTM}(x_i)$ and backward $bh_i = \overleftarrow{LSTM}(x_i)$ directions and concatenating them $h_i = [fh_i, bh_i]$ to capture information from both the past and future.

2.1.2 Character Embeddings

Character embeddings are used to capture orthographic patterns and to deal with out-of-vocabulary words, especially in the cross-lingual setting. We follow the same setup as Lample et al. (2016) to obtain character embeddings using bi-LSTM. Specifically, we concatenate both character and word-level features and use a bi-LSTM on top of that.

2.1.3 CRF Layer

The encoded character and word-level features are fed to a CRF layer to learn inter-dependencies between output trigger tags and find the optimal tag sequence. This layer simulates bi-LSTM in its use of past and future tags to predict the current tag. Following Lafferty et al. (2001), CRF layers define a transition matrix A and use a score A_{ij} to model the transition from the i^{th} state to the j^{th} for a pair of consecutive time steps. The scores $[f_\theta]_{i,t}$ of the matrix is the score output by the network with parameters θ, for the sentence $[x]_1^N$ and for the i^th tag, at the t^h word. The score of a sequence of tags $[y]_1^N$ for a particular sequence of words $[x]_1^N$ is the sum of transition scores and network scores which are computed efficiently using dynamic programming.

$$s([x]_1^N, [i]_1^N) = \sum_{t=1}^{N} ([A]_{[i]_{t-1},[i]_t} + [f_\theta]_{[i]_t,t}) \quad (1)$$

2.2 BERT-CRF

BERT is a multi-layer bidirectional transformer encoder, an extension to the original Transformer model (Vaswani et al., 2017). The input representation consists of a concatenation of WordPiece embeddings (Wu et al., 2016), positional embeddings, and the segment embedding. A special token ([CLS]) is inserted at the beginning of each sentence and another special token ([PAD]) is used to normalize the length of sentences (no ([SEP]) token is used in this case). The pre-trained BERT model provides a powerful contextualized representation which gives the state-of-the-art performance for many NLP tasks. We use BERT-CRF, which adds a CRF layer on top of BERT's contextualized embeddings layer.

3 Experimental Setup

3.1 Dataset

We evaluate our approach on the ACE2005 sentence-level event mention multilingual bench-

mark.[1] This dataset is annotated with 33 event subtypes which, when represented in BIO annotation, results in a 67-way labeling task. For a sound comparison, we use the same data split as the English baseline (as detailed in Section 3.3). To the best of our knowledge, there are no Arabic benchmark systems, so we produced our own split.[2] Statistics of the split for train, validation, and test for the three languages: English (EN), Chinese (ZH) and Arabic (AR) are included in Table 1.

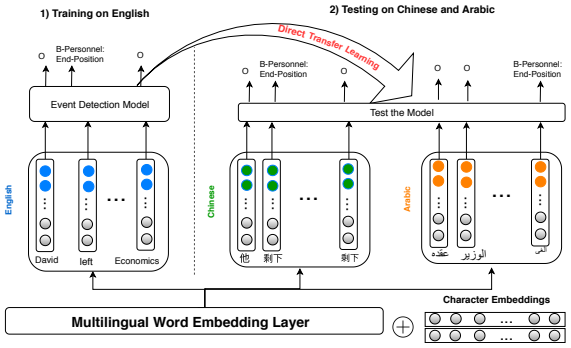

Figure 3: A zero-shot transfer learning architecture for cross-lingual event trigger extraction

3.2 Evaluation

We design different experiments for the evaluation of trigger extraction, where we train several language-specific and multilingual models using different embeddings and sequence labeling architectures. We evaluate the following training schemes:

- *Monolingual Baselines*: We train and fine-tune on EN, ZH or AR using monolingual FastText embeddings and testing on the trained language.
- *Zero-Shot learning experiments*: As depicted in Figure 3, we train and fine-tune on EN using multilingual embeddings (MUSE or BERT(multi)) and test on ZH and AR assuming no resources for those languages. To simplify experiments, we evaluate direct transfer only from EN since it is a high-resourced target language for learning projections needed in multilingual embeddings. We also believe AR and ZH are not good language-pair candidates, so we expect training on AR and testing on ZH and EN or training on ZH and testing on AR and EN would not lead to improvements.
- *Targeted cross-lingual experiments*: For each test language (ZH and AR), we train and fine-tune using multilingual embeddings on language pairs involving the test language in addition to EN to test to what extent adding training instances from the target language boosts the performance over zero-shot learning from EN only. When testing on EN, we train on EN+AR and EN+ZH.
- *Joint multilingual experiments*: We train and fine-tune on all languages (EN, ZH, and AR) using multilingual embeddings and testing on EN, ZH and AR. The hypothesis to be tested is whether a single language-independent model can work well across languages.

3.3 Baselines

We compare our methodology against different systems based on:

- Discrete Only: hand-crafted features only; *Ji's Cross-Entity'08* (Ji and Grishman, 2008); *Liao's Cross-Event'10* (Liao and Grishman, 2010); and *Li's Joint-Beam'13* (Li et al., 2013).
- Discrete + Continuous: using a combination of both linguistic features and trainable continuous features; *Chen's Dynamic CNN (DMCNN'15)* (Chen et al., 2015); *Nguyen's Joint RNN (JRNN'16)* (Nguyen et al., 2016); *Liu's Jointly Multiple Events (JMEE'18)* (Liu et al., 2018b); and *Zhang's Generative Adversarial Imitation Learning (GAIL'19)* (Zhang and Ji, 2018).
- Continuous Only: language-independent features only; *Feng's Hybrid Neural Network (HNN)'16* (Feng et al., 2016) and *Liu's Gated Multilingual Attention (GMLATT)'18* (Liu et al., 2018a).

For cross-lingual results, we include a comparison with ZH baselines; *Li's Maximum Entropy (MaxEnt)'13* (Li et al., 2013); *Chen's Rich-C'12* (Chen and Ng, 2012); *Feng's HNN'16* (Feng et al.,

[1] https://catalog.ldc.upenn.edu/LDC2006T06

[2] Our document splits are available in: https://github.com/meryemmhamdi1/cross-ling-ev-extr

Lang	Training		Validation		Test	
	#doc	#triggers	#doc	#triggers	#doc	#triggers
EN	529	4,420	30	505	40	424
ZH	557	2,213	32	111	44	197
AR	354	1,986	21	112	28	169

Table 1: Number of training, dev, test triggers and documents per language in ACE2005 dataset

2016); and *Hsi's Multi'16* (Hsi et al., 2016). More decriptions are included in Sections 5.1 and 5.2.

3.4 Hyper-parameters and Embeddings

We describe the hyper-parameters leading to the best attainable performance for each event trigger extraction architecture. They are selected based on random search and performance on the validation dataset. For Bi-LSTM-Char-CRF, we train character embeddings using a single bi-LSTM layer with 100 hidden units and use another single layer of bi-LSTM with 300 hidden units to train on the concatenated word and char embeddings. We use a dropout rate of 0.5. We optimize using Adam with learning rate of 0.01, weight decay rate of 0.9, $\beta_1 = 0.7$, $\beta_2 = 0.999$ and $\epsilon = 1\text{e}{-8}$.

For monolingual embeddings, we use 300-dimensional word embeddings for EN, ZH, and AR from fastText (Bojanowski et al., 2017). For multilingual experiments, we use MUSE library [3] to train unsupervised alignments from ZH and AR to EN resulting in a unified vector space for the three languages. We use the same training hyper-parameters across monolingual and multilingual training to ensure a fair comparison.

For BERT-CRF, we train monolingual EN and ZH using cased BERT-Base and BERT-ZH models[4] respectively and for all multilingual experiments, we use the recommended multi-cased BERT-Base model.[5] All models were trained using 12 layers with 768 hidden size and 12 self-attention heads and 110 Million parameters. We fine tune all BERT models with their default parameters. We use Adam with learning rate of 0.01, weight decay rate of 0.9, $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\epsilon = 1\text{e}{-6}$.

For all experiments, we use a batch size of 32 and limit the maximum sequence length of sentences to 128, padding or cutting otherwise. In the end, we report F1 for both trigger identification and classification tasks computed using the seqeval[6] framework for sequence labeling evaluation based on the CoNLL-2000 shared task, complying with previous work. Trigger classification doesn't assume the identification is correct but rather gives a stricter performance metric for measuring whether the trigger is not only identified but also correctly classified.

4 Results

Table 2 shows F1 scores of trigger identification and classification tested on EN, ZH and AR across two event architectures: Bi-LSTM-Char-CRF and BERT-CRF and using different embeddings and training schemes (fine-grained performance analysis based on precision, recall scores can be found in Appendix A).

4.1 Comparison with Feature-Based State-of-the-art

Before digging deeper into the comparison of our results with previous state-of-the-art methodology, it is worth comparing the different approaches taken by the prior work. For both EN and ZH, we observe that the best F1 scores over trigger identification and classification are obtained by Liu's JMEE in the first place and Feng's HNN with a close performance (with scores of 73.7% and 73.4% on trigger classification). For the multilingual case (ZH), it is clear that Feng's HNN is very competitive, whereas models relying on machine translation, namely Liu's GMLATT and Hsi's multi, lag behind the rest of models.

It is not surprising that a neural-based system outperforms other hand-crafted architectures since the former can capture richer sequence information beyond sentence-level than traditional NLP pre-processing such as dependency parsing and avoid errors propagated from such tools.

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

[4] No pre-trained BERT model exists for Arabic.

[5] https://github.com/google-research/bert

[6] https://github.com/chakki-works/seqeval

Model	Train Lang	Embed -dings	Test Lang					
			EN		ZH		AR	
			Ident	Class	Ident	Class	Ident	Class
† *Ji's Cross-Entity'08*	EN	-	N/A	68.3	-	-	-	-
† *Liao's Cross-Event'10*		-	N/A	68.8	-	-	-	-
† *Li's Joint-Beam'13*		-	70.4	67.5	-	-	-	-
‡ *Chen's DMCNN'15*		Skip-Gram	73.5	69.1	-	-	-	-
‡ *Nguyen's JRNN'16*		C-BOW	71.9	69.3	-	-	-	-
‡ *Lius's JMEE'18*		Glove	75.9	73.7	-	-	-	-
‡ *Zhang's GAIL'19*		ELMo	73.9	72.0	-	-	-	-
+ *Feng's HNN'16*		Skip-Gram	75.9	73.4	-	-	-	-
+ Liu's GMLATT'18		Skip-Gram	74.1	72.4	-	-	-	-
† Li's *MaxEnt'13*	ZH	-	-	-	60.6	57.6	-	-
† *Chen's Rich-C'12*		-	-	-	66.7	63.2	-	-
‡ *Hsi's Multi'16*		multi_proj	-	-	N/A	39.6	-	-
+ *Feng's HNN'16*		Skip-Gram	-	-	68.2	63.0	-	-
* **Bi-LSTM-Char-CRF**	Test Lang	FastText	67.5	63.2	86.6	69.5	54.9	52.8
	Test Lang	MUSE	68.9	62.5	29.6	25.0	20.3	18.7
	EN		-	-	61.3	48.8	53.0	42.2
	EN+ZH		69.5	65.8	77.2	70.6	-	-
	EN+AR		70.6	66.9	-	-	56.1	53.2
	All		66.5	61.6	72.6	64.3	69.4	62.3
* **BERT-CRF**	Test Lang	Bert(Base)	79.2	**75.3**	84.4	79.9	-	-
	Test Lang	BERT (multi)	77.8	73.1	83.7	79.8	69.8	66.7
	EN		-	-	76.8	68.5	37.8	30.9
	EN+ZH		**79.8**	75.2	84.7	81.2	-	-
	EN+AR		79.3	74.5	-	-	**74.9**	**69.5**
	All		79.2	73.5	**87.7**	**83.2**	73.2	68.9

Table 2: Comparison of performance with different train/test language pairs using prior work baselines in the 1^{st} half and our method using Bi-LSTM-Char-CRF and BERT-CRF in the 2^{nd} half. †, + and ‡ denote baseline approaches using Discrete Only, Discrete + Continuous and Continuous Only features respectively, whereas * denotes our own approaches. A '-' designates that the experiment doesn't apply for that test language.

We observe that in general our language-independent (monolingual) Bi-LSTM-Char-CRF and BERT-CRF methods are on par with or outperform best attainable results. In particular, **BERT-CRF trained monolingually using BERT(Base) embeddings outperforms other baselines** for both EN and ZH, with F1-scores of 79.2 and 75.3 on trigger identification and classification, respectively, amounting to a 3.3% and 1.6% gain for EN. For ZH, we obtain F1-scores of 84.4% and 79.9%, amounting to an increase of 16.2% and 16.9% over the previous state-of-the-art. On the other hand, although results using Bi-LSTM-Char-CRF lag behind state-of-the-art for EN, incurring a loss of 10.5% over trigger classification, they are competitive for ZH, with scores of 86.6% and 69.5% and gains of 17.9% and 6.2% over Feng's HNN for trigger identification and classification respectively.

4.2 Comparison between MUSE and BERT Embeddings

We observe a significant difference in performance **in favor of BERT-CRF compared to Bi-LSTM-Char-CRF with a gain of 12.1%, 10.4%, and 13.9%** on the classification task. The better performance of BERT-CRF compared to Bi-LSTM-Char-CRF can be attributed to the fact that BERT is able to learn contextualized representation and long-range dependencies at different levels of granularity. Table 3 provides some examples where the surface form of the trigger is hard to dis-

ambiguate without context information. Reconsidering the second example from the introduction, we notice that a Bi-LSTM-Char-CRF fails to effectively associate it with position context clues.

4.3 Cross-lingual Event Trigger Extraction

In general, we observe that multilingual training leveraging multilingual embeddings provides a boost in performance for both event architectures, especially for EN and AR. More precisely, there is a **gain of 3.1% and 4.0% for EN and ZH respectively** on the identification performance of multilingual over monolingual models. We notice that AR benefits the most from multilingual training with an **improvement of 9.5% and 2.8%** on the classification score with BERT-CRF and Bi-LSTM-Char-CRF respectively. This supports our claim about the effectiveness of multilingual models which are efficient to train and are more robust than monolingual models.

Although F1-scores for **zero-shot transfer learning**(train: EN, test: ZH/AR) are not the best among multilingual experiments, they are still promising and exceed prior published work given the fact that no data from the target language was used to fine-tune. In particular, **training with EN using BERT-CRF was helpful for ZH** with a performance not far from monolingual performance. The same can be observed in the case of EN→ZH and EN→AR using MUSE. The lower performance of EN→AR using BERT-CRF raises questions about the quality of BERT(multi) embedding model training for Arabic.

Not surprisingly, training given a reasonable amount of language-specific resources from the test language under a **targeted cross-lingual scheme**(train: EN+ZH/EN+AR, test: EN/ZH/AR), boosts (with rare exceptions) the performance over both monolingual training and zero-shot learning: EN+AR>EN, EN+ZH>ZH>EN and EN+AR>EN>AR when testing on ZH, AR and even for EN for which we have a strong monolingual baseline.

When all languages are used to train **one single joint multilingual model** (train: All, test: EN/AR/ZH) we don't always notice improvements over monolingual models. To gain more insight into why multilingual training boosts performance over monolingual models, we include some examples of when EN is complementary to ZH and AR and without which the model fails to identify some events. In the Chinese example, there are only 12 "nearby" Chinese words to the trigger word 解散 (Jiěsàn) in ZH training data, whereas there are 4 times as many nearby words in EN (e.g. disband, dissolve, shut, cease, etc).

5 Related Work

Since this work is at the intersection of (i) event extraction and (ii) multilingual event extraction, we present previous work in relation to each domain separately in addition to (iii) a description of cross-lingual approaches for other tasks which motivate our current work.

5.1 Event Extraction in English

Previous works in event extraction on ACE2005 benchmark dataset are mostly focused on English and can be categorized based on the degree of hand-crafted features used and whether they are trained in a pipelined or joint fashion. While some systems such as Cross-Document (Ji and Grishman, 2008) and Cross-Event (Liao and Grishman, 2010) leverage document-level information to enhance the performance of event extraction in a pipelined fashion, others propose a more structured framework for joint training of both trigger labeling and argument extraction (Li et al., 2013).

Other approaches explore neural networks on top of linguistic features employing architectures like Dynamic Multi-Pooling CNNs (DM-CNN) (Chen et al., 2015) and bidirectional RNNs (JRNN) with manually crafted features (Nguyen et al., 2016). A joint approach was proposed by Liu et al. (2018b) to extract multiple events based on syntactic graph convolution network. More recently, Zhang and Ji (2018) propose an approach based on inverse reinforcement learning using Generative Adversarial Networks (GAN) to alleviate mistakes related to ambiguous labels making the model less vulnerable to biased, supervised datasets like ACE2005.

However, the majority of the described approaches involves to some degree the use of linguistic features. This is labor intensive and requires rich external resources, which are not necessarily available for low-resource languages.

5.2 Cross-lingual Event Extraction

Previous works for cross-lingual event extraction conducted in a semi-supervised way range from purely supervised approaches to those using ma-

	MUSE	BERT
Davies is **leaving** to become chairman of the London School of Economics	Movement: Transport	Personnel: End-Position
The EU is set to release 20 million euros (US 21) million in immediate humanitarian aid ...	Justice: Release-Parole	Transaction: Transfer-Money
Palestinian **uprising** as Isreal removed all major checkpoints in the coastal territory.	Conflict: Demonstrate	Conflict: Attack
	BERT(mono)	**BERT(multi_all)**
لم يسلم ارسنال من الغرامة حيث فرضت عليه اللجنة ... "Arsenal has not been **released** from the fine ..."	O	Justice:Fine
ينبغي فورا ان تتحول الثورة الى نضال "The stone **revolution** must immediately turn into a fight."	O	Conflict:Attack
由于月之海已经宣布年底前要解散，所以使得... "Since 'the sea of the moon' has been announced to be **disbanded** before the end of the year, ... "	B-Business: Declare-Bankruptcy	Business: End-Org

Table 3: Examples of trigger extraction mislabeled by MUSE but correctly labeled by BERT and those missed/mislabeled with monolingual training only and corrected with multilingual BERT model.

chine translation techniques or word alignment data. Feng et al. (2016) propose a language-independent approach that doesn't require any linguistic feature engineering. However, this approach still requires equally abundant labeled data for different languages and implies the need to train a new model for each language independently.

Hsi et al. (2016) exploit both language-dependent and language-independent features in the form of universal features such as universal dependencies, limited bilingual dictionaries and aligned multilingual word embeddings to train a model with multiple languages. However, this work lags behind in terms of the neural approach used and doesn't investigate the effectiveness of leveraging multiple source languages.

Liu et al. (2018b) propose gated cross-lingual attention as a mechanism to exploit the inherent complementarity of multilingual data which helps with data scarcity and trigger disambiguation. However, this approach relies on machine translation which suffers from error propagation.

5.3 Cross-lingual Tasks

Cross-lingual embeddings are of practical usefulness in many tasks in natural language processing (NLP) and information extraction (IE). In each case, a model is trained on one language and transferred to unseen languages. Downstream applications on which they are applied include part-of-speech (POS) tagging (Cohen et al., 2011), cross-lingual document classification ((Klementiev et al., 2012); (Schwenk and Li, 2018)) named entity recognition (Xie et al., 2018). More recently, BERT was developed as an extension to the transformer architecture and achieved significant improvement in performance for many NLP tasks.

The gain in performance associated with multilingual training is what encouraged us to explore this methodology on event trigger extraction. To the best of our knowledge, there is no prior work adopting conventional or contextualized multilingual embeddings for event trigger detection.

6 Conclusion

In this work, we propose a cross-lingual approach for event trigger extraction using a direct transfer of annotation framework based on multilingual embeddings. Compared to previous approaches, our approach doesn't rely on hand-crafted linguistic features or machine translation.

We evaluate this approach using event trigger extraction architectures with type-based unsupervised embeddings (FastText and MUSE) and supervised embeddings tuned to the context (BERT). Our results for both English and Chinese show competitive performance with baselines on the ACE2005 benchmark even in the zero-shot learning scheme. Although results using MUSE are lower for English, they are on par with Chinese baselines and better for Arabic compared to BERT.

We observe a generous boost in performance when English is added to the target language, and when all languages are combined together to train one cross-lingual model, especially for Arabic. Our result are promising compared to both feature-based approaches and cross-lingual approaches based on machine translation.

Acknowledgment

This material is based on research sponsored by DARPA under agreement number FA8750-18- 2-0014. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of DARPA or the U.S. Government.

References

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Chen Chen and Vincent Ng. 2012. Joint modeling for Chinese event extraction with rich linguistic features. In *Proceedings of COLING 2012*, pages 529–544, Mumbai, India. The COLING 2012 Organizing Committee.

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.

Shay B. Cohen, Dipanjan Das, and Noah A. Smith. 2011. Unsupervised structure prediction with non-parallel multilingual guidance. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 50–61, Stroudsburg, PA, USA. Association for Computational Linguistics.

Alexis Conneau, Guillaume Lample, MarcÁurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word translation without parallel data. *CoRR*, abs/1710.04087.

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.

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.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.*, 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. 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.

Heng Ji and Ralph Grishman. 2008. Refining event extraction through cross-document inference. In *Proceedings of ACL-08: HLT*, pages 254–262. Association for Computational Linguistics.

Alexandre Klementiev, Ivan Titov, and Binod Bhattarai. 2012. Inducing crosslingual distributed representations of words. In *Proceedings of COLING 2012: Technical Papers*, pages 1459–1474, Mumbai, India.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270, San Diego, California. 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.

Jian Liu, Yubo Chen, Kang Liu, and Jun Zhao. 2018a. Event detection via gated multilingual attention mechanism. In *AAAI*, pages 4865–4872. AAAI Press.

Shulin Liu, Kang Liu, Shizhu He, and Jun Zhao. 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*, AAAI'16, pages 2993–2999. AAAI Press.

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.

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. Association for Computational Linguistics.

Holger Schwenk and Xian Li. 2018. A corpus for multilingual document classification in eight languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*, Miyazaki, Japan. European Languages Resources Association (ELRA).

Lei Sha, Feng Qian, Baobao Chang, and Zhifang Sui. 2018. Jointly extracting event triggers and arguments by dependency-bridge RNN and tensor-based argument interaction. In *AAAI*, pages 5916–5923. AAAI Press.

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.

Christopher Walker. 2006. Ace 2005 multilingual training corpus ldc2006t06. In *Linguistic Data Consortium*, Philadelphia, United States of America.

Yonghui Wu, Mike Schuster, and Zhifeng Chen et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144.

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, Brussels, Belgium. Association for Computational Linguistics.

Sen Yang, Dawei Feng, Linbo Qiao, Zhigang Kan, and Dongsheng Li. 2019. Exploring pre-trained language models for event extraction and generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5284–5294, Florence, Italy. Association for Computational Linguistics.

Tongtao Zhang and Heng Ji. 2018. Event extraction with generative adversarial imitation learning. *CoRR*, abs/1804.07881.