CMU_CS_Event TAC-KBP2016 Event Argument Extraction System

Andrew Hsi
Carnegie Mellon University
Pittsburgh, PA 15213 USA
ahsi@cs.cmu.edu

Jaime Carbonell
Carnegie Mellon University
Pittsburgh, PA 15213 USA
jgc@cs.cmu.edu

Yiming Yang
Carnegie Mellon University
Pittsburgh, PA 15213 USA
yiming@cs.cmu.edu

Abstract

This year, the CMU_CS_Event team participated in the Event Argument Extraction and Linking Task. We utilize a pipeline of classifiers approach to event extraction, with particular focus on leveraging resources from multiple languages in order to train a single cross-lingual model. We apply our system to tri-lingual event extraction, resulting in the top performance on both the Chinese and Spanish document-level sub-tasks.

1 Introduction

As event extraction algorithms continue to improve in performance, a natural question to ask is how to adapt such systems to new languages? To date, the majority of event extraction work has focused on English (Grishman et al., 2005; Ji and Grishman, 2008; Gupta and Ji, 2009; Liao and Grishman, 2010; Liao and Grishman, 2011; Li et al., 2013; Bronstein et al., 2015). Beyond this, only a small subset of research has attempted to explore languages beyond English – primarily in Chinese, but occasionally other languages as well (Chen and Ji, 2009b; Piskorski et al., 2011; Li et al., 2012; Chen and Ng, 2012; Chen and Ng, 2014).

However, even when considering event extraction in non-English languages, the majority of research still focuses on monolingual event extraction; that is, training and testing on the same language. With the major challenges and costs that exist in obtaining training data for event extraction, it is desirable to be able to transfer knowledge between languages in order to improve performance. This has been widely studied within the natural language processing (NLP) community on a variety of tasks (Richman and Schone, 2008; Zeman and Resnik, 2008; Snyder et al., 2009; Cohen et al., 2011; McDonald et al., 2011; Ammar et al., 2016), but there exists only a small amount of work extending cross-lingual NLP to event extraction, such as those by Chen and Ji (2009a), Piskorski et al. (2011), and more recently, Hsi et al. (2016).

Our core algorithm is primarily based on the work of Hsi et al. (2016), with some adaptations to modify this work to match the TAC-KBP Event Argument Extraction specifications. We begin by introducing necessary terminology for the event extraction task in Section 2. We then describe our overall system architecture in Section 3. In Section 4, we show some experimental tri-lingual results on RichERE data, as well as our official results on the 2016 evaluation set. Finally, we offer conclusions and ideas for future work in Section 5.

2 Terminology

We begin by identifying relevant terminology for event extraction.

- An *event* is something that happens in the world at a particular place and time.
- An *event mention* is a particular occurrence of an event in a document. An event may be mentioned multiple times within the same document, or the same event may be mentioned across a set of documents.

- An *event trigger* is a particular word that signifies the existence of an event.
- An *event argument* is an entity that fulfills some role within a particular event. The set of valid roles for an event depends on the type of event, including roles such as Agent, Place, and Time.
- An *event argument mention* is a particular textual instance of an event argument.

3 System Architecture

Our system architecture is as follows. We begin with a preprocessing step, which includes tokenization, part-of-speech tagging, entity mention recognition, and dependency parsing. We then use a series of classifiers to perform event trigger classification, event argument classification, and argument realis classification. The results of these classifiers are used in a postprocessing step to match the specified output format for the TAC-KBP task. The overall pipeline can be seen in Figure 1.

We begin by describing each of our components in turn, focusing on the general case of monolingual training. We begin by first describing our model in general, and then highlight our components enabling cross-lingual learning.

3.1 Preprocessing

We begin by running the Stanford CoreNLP tool on the input texts to obtain segmentation, tokenization, and part-of-speech tags (Manning et al., 2014). For English, we additionally run the CoreNLP dependency parsing module. For Spanish and Chinese, we obtain dependency parses using MaltParser (Nivre et al., 2007), which we found to work much faster in practice than CoreNLP.

For each of the three languages, we train a conditional random field (Lafferty et al., 2001) with the Stanford Named Entity Recognizer (NER) (Finkel et al., 2005) in order to detect entity mention candidates. In the official evaluation, we also utilize results from the TAC KBP Entity Discovery and Linking (EDL) track.

3.2 Event Trigger Extraction

Following the preprocessing stage, we extract features for our event trigger classifier, which may be seen in Table 1. For each word in each document, we classify the word as belonging to one of the event types in the RichERE ontology, or "NONE" if the word is not an event trigger. We train a logistic regression classifier to make the predictions, using LIBLINEAR (Fan et al., 2008).

We obtain word embeddings with word2vec (Mikolov et al., 2013) for all three target languages (English, Chinese, Spanish) using their respective Wikipedia dumps.

3.3 Event Argument Extraction

Given the resulting event triggers from the previous step, we then make classification decisions on event arguments. For each trigger word/entity mention pair within a sentence, we classify the relationship between them as belonging to one of the argument roles in the RichERE ontology, or "NONE" if no such relationship exists. We train our argument classifier using LIBLINEAR.

We extract features for this component from the preprocessed texts and the output triggers from the previous step. A detailed view of our features can be seen in Figure 2.

3.4 Realis Classification

For each argument found by the event argument detection component, we then make a final classification decision to determine the realis value of the argument – one of ACTUAL, GENERIC, or OTHER. We once again train a logistic regression classifier with LIBLINEAR to perform this task. For this component, we use similar features to those of the argument detection component.

3.5 Postprocessing

Finally, once we have obtained our set of extracted event arguments and their corresponding realis labels, we perform some final postprocessing steps in order to match the output format defined by the TAC-KBP Event Argument Extraction and Linking task. Within documents, we link together all arguments that belong to the same event type. We do not perform any additional linking of arguments across documents.

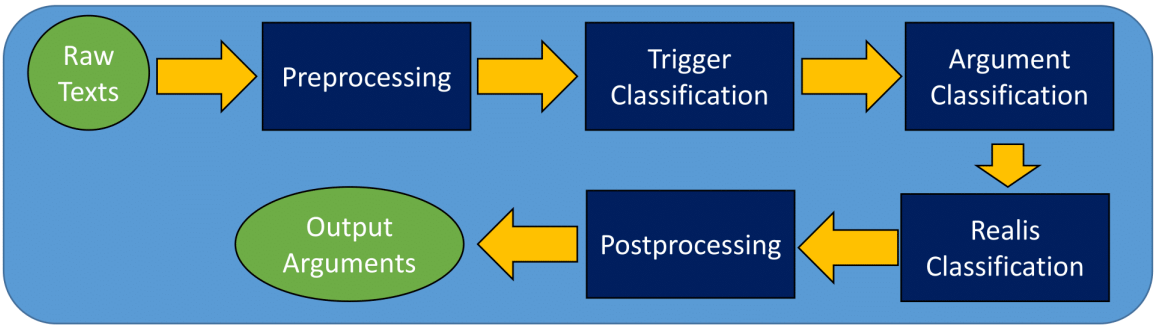

Figure 1: Architecture for our event extraction system.

Event Trigger Extraction Features
Lexical features (e.g. words and lemmas within a context window)
Length of the current word
Language-specific POS tags within a context window
Universal POS tags within a context window
Word embedding vector for current word
Dependent/Governor information from dependency parsing

Table 1: Features used in the Event Trigger Extraction component

Event Argument Extraction Features
Lexical features about the entity phrase
Lexical features for individual words in the entity phrase
Entity type
Event type and subtype of trigger word
Existence of any other candidate entities in the same sentence
Distance between the trigger and entity
Dependent/Governor information from dependency parsing

Table 2: Features used in the Event Argument Extraction component

3.6 Cross-lingual Training

Our overall system architecture described in the previous sections is designed in such a way as to allow both monolingual and cross-lingual usage. In this section, we now describe how our approach enables cross-lingual training.

A major challenge in event extraction research is the lack of available training data. This is especially true in the case of Spanish, which has far fewer annotated documents than either English or Chinese. To overcome this obstacle, we incorporate training data from all three languages in order to create a single cross-lingual model that can make predictions on documents in any of these languages.

In order to facilitate the use of all three languages in training, we use a combination of language dependent and language independent features in training. By using such a combination of features, we allow our model to both focus on the nuances of the individual language being tested on as well as the general behaviors that can be observed across languages. The language dependent features include individual words, language-specific part-of-speech tags, and word embeddings. The language independent features include Universal POS tags (Petrov et al., 2012), trigger type information, entity type information, and Universal Dependencies (McDonald et al., 2013).

4 Experiments

We now present our experimental results. We begin by describing our own internal experiments conducted prior to the TAC-KBP 2016 evaluation. We then move on to discuss our performance during the official evaluation.

4.1 Internal Experiments

For our internal experiments, we utilize the English, Chinese, and Spanish portions of RichERE. We conduct separate experiments on all three target languages to show the improvement in performance that may be obtained via cross-lingual training. For these experiments, we utilize the RichERE gold standard entity mentions and event triggers as input to our system.

For each experiment, we first consider a monolingual baseline. We split the data for the target language into 10 folds, and conduct cross-validation. We then consider the cross-lingual extension of this baseline by adding all of the available training data in the other two (non-target) languages. Results for these experiments may be seen in Tables 3, 4, and 5 for English, Chinese, and Spanish respectively. We report both micro-averaged and macro-averaged results.

Testing on English, we find that the cross-lingual approach suffers from slightly worse precision, but shows slight improvement on recall and F1. On Chinese, we find that all metrics are improved under the cross-lingual setting. Furthermore, on Chinese we see a larger difference in performance between the monolingual baseline and the cross-lingual version.

Spanish suffers from notably worse performance than the other languages in its monolingual baseline – almost 20 points lower on F1 compared to English! This is to be expected however due to the much smaller amount of available training data. This also means that Spanish provides the best opportunity for drawing improvements from other languages. Indeed, when we apply our cross-lingual approach to Spanish, we find that this provides a noticeably larger boost in performance – nearly 10 points of improvement on macro and micro F1 over its monolingual counterpart.

4.2 Official Results

For the official evaluation, we ran our system on all three target languages, submitting 5 different runs:

- Run 1 – trained a single cross-lingual model, used combination of entity mentions from the CMU_CS_EDL team and entity mentions extracted by our mention detection system
- Run 2 – trained a single cross-lingual model, used entity mentions from the CMU_CS_EDL team
- Run 3 – trained three separate monolingual models, used entity mentions from the CMU_CS_EDL team
- Run 4 – trained a single cross-lingual model, used entity mentions extracted by our mention detection system
- Run 5 – trained three separate monolingual models, used entity mentions extracted by our mention detection system

All 5 runs were trained on the union of the ACE 2005 and RichERE data. Results for each of our five runs on English, Chinese, and Spanish may be seen in Tables 6, 7, and 8 respectively.

In English, our best system is from Run 1, which uses cross-lingual training and both entity mention sources. We see mixed results overall about whether cross-lingual training improves over monolingual training – it provided a small boost in performance when using the Stanford NER entity mentions, but a slight drop in performance when using just the EDL mentions. We quite clearly see a boost in performance when using the Stanford NER mentions over the EDL mentions, with the best performance resulting when we take the union of these mentions.

In Chinese, our best system is also Run 1. However, in Chinese we find that using cross-lingual training boosts F1 performance regardless of the entity mention source – Run 2 outperforms Run 3, and Run 4 outperforms Run 5. This is promising to note, and supports our internal experiments on RichERE.

In Spanish, our best system is Run 3, which means that cross-lingual training did not give the best performance on this language. An interesting difference between our Spanish results and our other results is that Spanish is the only language where using

	Macro-Average			Micro-Average		
	Precision	Recall	F1	Precision	Recall	F1
Monolingual approach	**72.3**	50.3	57.6	**78.6**	53.7	63.8
Cross-lingual approach	71.8	**52.1**	**58.6**	78.3	**54.5**	**64.2**

Table 3: English Argument Results on RichERE

	Macro-Average			Micro-Average		
	Precision	Recall	F1	Precision	Recall	F1
Monolingual approach	59.1	39.8	45.7	75.5	46.4	57.4
Cross-lingual approach	**66.6**	**46.1**	**52.3**	**76.4**	**48.6**	**59.3**

Table 4: Chinese Argument Results on RichERE

	Macro-Average			Micro-Average		
	Precision	Recall	F1	Precision	Recall	F1
Monolingual approach	54.5	32.3	37.8	72.7	32.2	44.4
Cross-lingual approach	**58.2**	**42.4**	**46.0**	**73.6**	**40.6**	**52.0**

Table 5: Spanish Argument Results on RichERE

just EDL mentions outperformed using the Stanford NER mentions. This seems to imply that our Spanish entity detection module is insufficient to properly identify candidate arguments. In future, it may be beneficial for us to explore alternative directions for Spanish entity mention detection.

Across all three languages, we find that our system's overall performance is clearly tied to our low recall. This is an area for clear improvement, although we note that this kind of low-recall problem is typical for event extraction systems. This year in particular, we anticipated lower recall than in the past due to the shift from previous years to a gold-standard based evaluation, so the overall results are not too surprising for us.

In general, we have found less improvement with cross-lingual training in the official evaluation than with our internal experiments on RichERE. One possible explanation for this is due to the selected event categories used by the TAC KBP 2016 evaluation. This year's evaluation focused on a smaller set of event types compared to previous years. In contrast, our internal experiments focus on the entire set of categories. This is meaningful because our cross-lingual algorithm is particularly designed to improve performance on rare classes by borrowing additional training examples from other languages. When there is already sufficient training data to do well on a class, adding additional data via cross-lingual training offers diminishing gains. Another factor to consider is that in the evaluation setting, our system suffers from noisy decisions earlier in the pipeline (e.g. entity mentions, trigger classification), particularly in the case of Spanish. In contrast, the internal experiments we described in the previous section assume much less noisy input.

On the official metric for ranking argument scores, we achieve the median rank on English, and the top rank on both Chinese and Spanish. These scores may be seen in Tables 9, 10, and 11, along with the top ranked and median (where available) scores.

5 Conclusion

This year at TAC-KBP, we submitted 5 systems to the Event Argument and Linking Task. We focus in particular on cross-lingual training and leveraging resources from multiple source languages to improve performance on the target language. Our internal results on RichERE show promising results for all three languages when incorporating additional languages during training, with particularly notable improvements on Chinese and Spanish.

In the official document-level results, we achieved the median ranked score for English, and the top

	Precision	Recall	F1	Arg Score	Link Score
CMU_CS_Event1	31.2	**4.9**	**8.4**	**3.0**	**1.3**
CMU_CS_Event2	49.5	2.2	4.2	2.0	0.3
CMU_CS_Event3	**50.9**	2.3	4.3	2.1	0.3
CMU_CS_Event4	28.5	3.9	6.8	2.3	0.9
CMU_CS_Event5	28.6	3.8	6.8	2.3	0.8

Table 6: Document-level English results in official TAC KBP 2016 Evaluation

	Precision	Recall	F1	Arg Score	Link Score
CMU_CS_Event1	12.2	**3.6**	**5.5**	**3.5**	**1.5**
CMU_CS_Event2	**14.7**	1.1	2.0	1.3	0.5
CMU_CS_Event3	12.7	1.0	1.8	1.2	0.3
CMU_CS_Event4	12.2	3.1	5.0	3.2	1.3
CMU_CS_Event5	11.5	3.1	4.9	3.3	1.0

Table 7: Document-level Chinese results in official TAC KBP 2016 Evaluation

	Precision	Recall	F1	Arg Score	Link Score
CMU_CS_Event1	17.0	1.1	2.1	1.3	0.6
CMU_CS_Event2	17.5	1.1	2.0	1.2	0.5
CMU_CS_Event3	14.6	**1.4**	**2.5**	**1.5**	**0.8**
CMU_CS_Event4	**33.3**	0.3	0.5	0.4	0.2
CMU_CS_Event5	21.1	0.3	0.5	0.3	0.1

Table 8: Document-level Spanish results in official TAC KBP 2016 Evaluation

	5%	50%	95%
Our best system	2.5	3.0	3.4
Top-ranked system	8.6	9.7	10.9
Median-ranked system	2.5	3.0	3.4

Table 9: Official measurement for ranking English argument scores. Scores are given as percentiles based on bootstrap resampling.

	5%	50%	95%
Our best system	2.3	3.5	4.4
Top-ranked system	2.3	3.5	4.4

Table 10: Official measurement for ranking Chinese argument scores. Scores are given as percentiles based on bootstrap resampling.

ranked scores for both Chinese and Spanish. We find that cross-lingual training sometimes, but not always boosts performance under this setting, which we hypothesize to be due to the reduced coverage of rare classes in the evaluation data as well as noisy input from classification decisions made earlier in the system pipeline.

	5%	50%	95%
Our best system	1.1	1.5	1.9
Top-ranked system	1.1	1.5	1.9

Table 11: Official measurement for ranking Spanish argument scores. Scores are given as percentiles based on bootstrap resampling.

A possible direction for future work would be the adaptation of cross-lingual training to more sophisticated machine learning models, including the use of structured prediction and neural methods. An additional direction to explore would be to expand our entity mention detector to a cross-lingual model. Currently, we only consider cross-lingual training at the trigger, argument, and realis classification stages of the system, so applying a cross-lingual entity mention detector could offer additional boosts in performance.

Acknowledgments

This research was supported in part by DARPA grant FA8750-12-2-0342 funded under the DEFT program.

References

Waleed Ammar, George Mulcaire, Miguel Ballesteros, Chris Dyer, and Noah A. Smith. 2016. Many languages, one parser. In *TACL*.

Ofer Bronstein, Ido Dagan, Qi Li, Heng Ji, and Anette Frank. 2015. Seed-based event trigger labeling: How far can event descriptions get us? In *ACL*.

Zheng Chen and Heng Ji. 2009a. Can one language bootstrap the other: A case study on event extraction. In *NAACL HLT Workshop on Semi-supervised Learning for Natural Language Processing*.

Zheng Chen and Heng Ji. 2009b. Language specific issue and feature exploration in chinese event extraction. In *HLT-NAACL*.

Chen Chen and Vincent Ng. 2012. Joint modeling for chinese event extraction with rich linguistic features. In *COLING*.

Chen Chen and Vincent Ng. 2014. Sinocoreferencer: An end-to-end chinese event coreference resolver. In *LREC*.

Shay B. Cohen, Dipanjan Das, and Noah A. Smith. 2011. Unsupervised structure prediction with non-parallel multilingual guidance. In *EMNLP*.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. Liblinear: A library for large linear classification. In *JMLR*.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *ACL*.

Ralph Grishman, David Westbrook, and Adam Meyers. 2005. Nyus english ace 2005 system description. In *Proc. ACE 2005 Evaluation Workshop*.

Prashant Gupta and Heng Ji. 2009. Predicting unknown time arguments based on cross-event propagation. In *ACL-IJCNLP*.

Andrew Hsi, Yiming Yang, Jaime Carbonell, and Ruochen Xu. 2016. Leveraging multilingual training for limited resource event extraction. In *COLING*.

Heng Ji and Ralph Grishman. 2008. Refining event extraction through cross-document inference. In *ACL*.

John Lafferty, Andrew McCallum, and Fernando Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML*.

Peifeng Li, Guodong Zhou, Qiaoming Zhu, and Libin Hou. 2012. Employing compositional semantics and discourse consistency in chinese event extraction. In *EMNLP*.

Qi Li, Heng Ji, and Liang Huang. 2013. Joint event extraction via structured prediction with global features. In *ACL*.

Shasha Liao and Ralph Grishman. 2010. Filtered ranking for bootstrapping in event extraction. In *ACL*.

Shasha Liao and Ralph Grishman. 2011. Acquiring topic features to improve event extraction: in pre-selected and balanced collections. In *RANLP*.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *ACL*.

Ryan McDonald, Slav Petrov, and Keith Hall. 2011. Multi-source transfer of delexicalized dependency parsers. In *EMNLP*.

Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, Claudia Bedini, Núria Bertomeu Castelló, and Jungmee Lee. 2013. Universal dependency annotation for multilingual parsing. In *ACL*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *ICLR*.

Joakim Nivre, Johan Hall, Jens Nilsson, Atanas Chanev, Gülsen Eryigit, Sandra Kübler, Svetoslav Marinov, and Erwin Marsi. 2007. Maltparser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering*.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2012. A universal part-of-speech tagset. In *LREC*.

Jakub Piskorski, Jenya Belayeva, and Martin Atkinson. 2011. Exploring the usefulness of cross-lingual information fusion for refining real-time news event extraction: A preliminary study. In *RANLP*.

Alexander E. Richman and Patrick Schone. 2008. Mining wiki resources for multilingual named entity recognition. In *ACL*.

Benjamin Snyder, Tahira Naseem, Jacob Eisenstein, and Regina Barzilay. 2009. Adding more languages improves unsupervised multilingual part-of-speech tagging: A bayesian non-parametric approach. In *NAACL*.

Daniel Zeman and Philip Resnik. 2008. Cross-language parser adaptation between related languages. In *IJCNLP*.