Cross-lingual Structure Transfer for Relation and Event Extraction

Ananya Subburathinam1*, Di Lu1*, Heng Ji2, Jonathan May3, Shih-Fu Chang4, Avirup Sil5, Clare Voss6
1Computer Science Department, Rensselaer Polytechnic Institute
2Computer Science Department, University of Illinois at Urbana-Champaign
3University of Southern California Information Sciences Institute
4Electrical Engineering Department, Columbia University
5IBM T.J. Watson Research Center
6Army Research Laboratory
hengji@illinois.edu

Abstract
The identification of complex semantic structures such as events and entity relations, already a challenging Information Extraction task, is doubly difficult from sources written in under-resourced and under-annotated languages. We investigate the suitability of cross-lingual structure transfer techniques for these tasks. We exploit relation- and event-relevant language-universal features, leveraging both symbolic (including part-of-speech and dependency path) and distributional (including type representation and contextualized representation) information. By representing all entity mentions, event triggers, and contexts into this complex and structured multilingual common space, using graph convolutional networks, we can train a relation or event extractor from source language annotations and apply it to the target language. Extensive experiments on cross-lingual relation and event transfer among English, Chinese, and Arabic demonstrate that our approach achieves performance comparable to state-of-the-art supervised models trained on up to 3,000 manually annotated mentions: up to 62.6% F-score for Relation Extraction, and 63.1% F-score for Event Argument Role Labeling. The event argument role labeling model transferred from English to Chinese achieves similar performance as the model trained from Chinese. We thus find that language-universal symbolic and distributional representations are complementary for cross-lingual structure transfer.

1 Introduction
Advanced Information Extraction (IE) tasks entail predicting structures, such as relations between entities, and events involving entities. Given a pair of entity mentions, Relation Extraction aims to identify the relation between the mentions and classify it by predefined type. Event Extraction aims to identify event triggers and their arguments in unstructured texts and classify them respectively by predefined types and roles. As Figure 1 illustrates, both tasks entail predicting an information network (Li et al., 2014) for each sentence, where the entity mentions and event triggers are nodes, and the relations and event-argument links are edges labeled with their relation and argument roles, respectively.

There are certain relations and events that are of primary interest to a given community and so are reported predominantly in the low-resource language data sources available to that community. For example, though English language news will occasionally discuss the Person Aung San Suu Kyi, the vast majority of Physical-Located relations and Meeting events involving her are only reported locally in Burmese news, and thus, without Burmese relation and event extraction, a knowledge graph of this person will lack this available information. Unfortunately, publicly-available gold-standard annotations for relation and event extraction exist for only a few languages (Doddington et al., 2004; Getman et al., 2018), and Burmese is not among them. Compared to other IE tasks such as name tagging, the annotations for Relation and Event Extraction are also more costly to obtain, because they are structured and require a rich label space.

Recent research (Lin et al., 2017) has found that relational facts are typically expressed by identifiable patterns within languages and has shown
that the consistency in patterns observed across languages can be used to improve relation extraction. Inspired by their results, we exploit language-universal features relevant to relation and event argument identification and classification, by way of both symbolic and distributional representations. For example, language-universal POS tagging and universal dependency parsing is available for 76 languages (Nivre et al., 2018), entity extraction is available for 282 languages (Pan et al., 2017), and multi-lingual word embeddings are available for 44 languages (Bojanowski et al., 2017; Joulin et al., 2018). As shown in Figure 2, even for distinct pairs of entity mentions (colored pink and blue, in both English and Russian), the structures share similar language-universal symbolic features, such as a common labeled dependency path, as well as distributional features, such as multilingual word embeddings.

Based on these language-universal representations, we then project all entity mentions, event triggers and their contexts into one multilingual common space. Unlike recent work on multilingual common space construction that makes use of linear mappings (Mikolov et al., 2013; Rothe et al., 2016; Zhang et al., 2016; Lazaridou et al., 2015; Xing et al., 2015; Smith et al., 2017) or canonical correlation analysis (CCA) (Ammar et al., 2016; Faruqui and Dyer, 2014; Lu et al., 2015) to transfer surface features across languages, our major innovation is to convert the text data into structured representations derived from universal dependency parses and enhanced with distributional information to capture individual entities as well.

Figure 2: Multilingual common semantic space and cross-lingual structure transfer.
as the relations and events involving those entities, so we can share structural representations across multiple languages.

Then we construct a novel cross-lingual structure transfer learning framework to project source language (SL) training data and target language (TL) test data into the common semantic space, so that we can train a relation or event extractor from SL annotations and apply the resulting extractor to TL texts. We adopt graph convolutional networks (GCN) (Kipf and Welling, 2017; Marcheggiani and Titov, 2017) to encode graph structures over the input data, applying graph convolution operations to generate entity and word representations in a latent space. In contrast to other encoders such as a Tree-LSTM (Tai et al., 2015), GCN can cover more detailed contextual information from dependency parses because, for each word, it captures all parse tree neighbors of the word, rather than just the child nodes of the word. Using this shared encoder, we treat the two tasks of relation extraction and event argument role labeling as mappings from the latent space to relation type and to event type and argument role, respectively.

Extensive experiments on cross-lingual relation and event transfer among English, Chinese, and Arabic show that our approach achieves highly promising performance on both tasks.

2 Model

2.1 Overview

Our cross-lingual structure transfer approach (see Figure 2) consists of four steps: (1) Convert each sentence in any language into a language-universal tree structure based on universal dependency parsing. (2) For each node in the tree structure, create a representation from the concatenation of multilingual word embedding, language-universal POS embedding, dependency role embedding and entity-type embedding, so that all sentences, independent of their language, are represented within one shared semantic space. (3) Adopt GCN to generate contextualized word representations by leveraging information from neighbors derived from the dependency parsing tree, and (4) Using this shared semantic space, train relation and event argument extractors with high-resource language training data, and apply the resulting extractors to texts of low-resource languages that do not have any relation or event argument annotations.

2.2 Tree Representations

Most previous approaches regard a sentence as a linear sequence of words, and incorporate language-specific information such as word order. Unlike sequence representations, tree representations such as constituency trees and dependency trees are typically constructed following a combination of syntactic principles and annotation guidelines designed by linguists. The resulting structures, such as the verb–subject relation and the verb–object relation, are found across languages. In this paper, we choose dependency trees as the sentence representations because the community has made great efforts at developing language-universal dependency parsing resources across 83 languages (Nivre et al., 2016).

We define the dependency-based tree representation for a sentence as $G = (V, E)$, where $V = \{v_1, v_2, \ldots, v_N\}$ is a set of words, and $E = \{e_1, e_2, \ldots, e_M\}$ is a set of language-universal syntactic relations. $N$ is the number of words in the sentence and $M$ is the number of dependency relations between words. To make this tree representation language-universal, we first convert each tree node into a vector which is a concatenation of representations between words. To make this tree representation language-universal, we first convert each tree node into a vector which is a concatenation of entity-type embedding, POS embedding (Nivre et al., 2016), entity-type embedding, and dependency relation embedding. More details are reported in Section 3.2.

2.3 GCN Encoder

Structural information is important for relation extraction and event argument role labeling, thus we aim to generate contextualized word representations by leveraging neighbors in dependency trees for each node.

Our GCN encoder is based on the monolingual design by Zhang et al. (2018b). The graphical sentence representation obtained from dependency parsing of a sentence with $N$ tokens is converted into an $N \times N$ adjacency matrix $A$, with added self-connections at each node to help capture information about the current node itself, as in Kipf and Welling (2017). Here, $A_{i,j} = 1$ denotes the presence of a directed edge from node $i$ to node $j$ in the dependency tree. Initially, each node contains distributional information about the $i$th word, including word embedding $x_i^w$, embeddings for symbolic information including its POS tag $x_i^p$, dependency relation $x_i^d$ and entity type $x_i^e$. 
We represent this initial representation \( h^{(0)}_i \) as follows:
\[
h^{(0)}_i = x^{w}_i \oplus x^{p}_i \oplus x^{d}_i \oplus x^{t}_i
\]

Then, at the \( k^{th} \) layer of convolution, the hidden representation is derived from the representations of its neighbors at the \((k - 1)^{th}\) layer. Thus, the hidden representation at the \( k^{th} \) layer for the \( i^{th} \) node is computed as:
\[
h^{(k)}_i = \text{ReLU} \left( \sum_{j=0}^{N} A_{ij} W^{(k)} h^{(k-1)}_j \right)
\]
where \( W \) is the weight vector, \( b \) is the bias term, and \( d_i \) represents the degree of the \( i^{th} \) node. The denominator in the equation denotes the normalization factor to neutralize the negative impact of node degree (Zhang et al., 2018b). The final hidden representation of each node after the \( k^{th} \) layer is the encoding of each word \( h^{(k)}_i \) in our language universal common space, and incorporates information about neighbors up to \( k \) hops away in the dependency tree.

2.4 Application on Relation Extraction
The GCN encoder generates the final hidden representation, \( h^{(k)}_i \), for each of the \( n \) nodes. We perform max-pooling over these final node representations to obtain a single vector representation for the sentence, \( h^s \). Similar to previous work (Zhang et al., 2018b), we also use the following recipe to obtain relation type classification results for each mention pair in a sentence: (1) max-pooling over the final representations of the nodes representing entity mentions, to get a single vector representation for each mention in a pair under consideration, \( h^{m1} \) and \( h^{m2} \), (2) a concatenation of three results of max-pooling: \([h^{m1}; h^s; h^{m2}]\) to combine contextual sentence information with entity mention information (Santoro et al., 2017; Lee et al., 2017; Zhang et al., 2018b), (3) a linear layer to generate a combined representation of these concatenated results, and (4) a Softmax output layer for relation type classification. The objective function used here is as follows:
\[
L^r = \sum_{i=1}^{N} \sum_{j=1}^{L_i} y_{ij} \log (\sigma(U^r : [h^{m1}_i; h^s; h^{m2}_j]))
\]
where \( U^r \) is a weight matrix.

2.5 Application on Event Argument Role Labeling
Event argument role labeling distinguishes arguments from non-arguments and classifies arguments by argument role. To label the role of an argument candidate \( x^c \) for an event trigger \( x^t \), we first generate the sentence representation \( h^s \), argument candidate representation \( h^c \), and trigger representation \( h^t \) by applying pooling on sentence, argument candidate \( x^c \) and event trigger \( x^t \) respectively, which is the same as that for relation extraction. The mapping function from the latent space to argument roles is composed of a concatenation operation \([h^t; h^c; h^s] \), a linear layer \( (U^a) \) and a Softmax output layer:
\[
L^a = \sum_{i=1}^{N} \sum_{j=1}^{L_i} y_{ij} \log (\sigma(U^a : [h^t_i; h^c_j; h^s]))
\]
where \( N \) is the number of event mentions, \( L_i \) is the number of argument candidates for \( i^{th} \) event mention and \( \sigma \) is the Sigmoid function.

3 Experiments
3.1 Data and Evaluation Metrics

<table>
<thead>
<tr>
<th>Relation Mentions</th>
<th>Event Mentions</th>
<th>Event Arguments</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>8,738</td>
<td>5,349</td>
</tr>
<tr>
<td>Chinese</td>
<td>9,317</td>
<td>3,333</td>
</tr>
<tr>
<td>Arabic</td>
<td>4,731</td>
<td>2,270</td>
</tr>
</tbody>
</table>

Table 1: Data statistics.

We choose the Automatic Content Extraction (ACE) 2005 corpus (Walker et al., 2006) for our experiments because it contains the most comprehensive gold-standard relation and event annotations for three distinct languages: English, Chinese and Arabic (see Table 1). Our target ontology includes the 7 entity types, 18 relation subtypes and 33 event subtypes defined in ACE. We randomly choose 80% of the corpus for training, 10% for development and 10% for blind test. We downsample the negative training instances by limiting the number of negative samples to be no more than the number of positive samples for each document.

For data preprocessing, we apply the Stanford CoreNLP toolkit (Manning et al., 2014) for Chinese word segmentation and English tokenization, and the API provided by UDPipe (Straka and...
performance by applying models trained with various combinations of training and test data from these three languages, as shown in Tables 3 and 4. We can see that the results are promising. For both tasks, the models trained from English are best, followed by Chinese, and then Arabic. We find that extraction task performance degrades as the accuracy of language-dependent tools (for sentence segmentation, POS tagging, dependency parsing) degrades.

Using English as training data, our cross-lingual transfer approach achieves similar performance on Chinese event argument role labeling (59.0%), compared to the model trained from Chinese annotations (59.3%), which is much higher than the best reported English-to-Chinese transfer result on event argument role labeling (47.7%) (Hsi et al., 2016).

We also show polyglot results for event argument role labeling in Table 4, by combing the training data from multiple languages. We observe that our model benefits from the combination of training data of multiple languages. The polyglot transfer learning does not provide further gains to relation extraction because the model converges quickly on a small amount of training data.

<table>
<thead>
<tr>
<th>Hyperparameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>word embedding size</td>
<td>300</td>
</tr>
<tr>
<td>POS embedding size</td>
<td>30</td>
</tr>
<tr>
<td>entity embedding size</td>
<td>30</td>
</tr>
<tr>
<td>dependency relation embedding size</td>
<td>30</td>
</tr>
<tr>
<td>hidden dimension size</td>
<td>200</td>
</tr>
<tr>
<td>dropout</td>
<td>0.5</td>
</tr>
<tr>
<td>number of layers</td>
<td>2</td>
</tr>
<tr>
<td>pooling function</td>
<td>max pooling</td>
</tr>
<tr>
<td>mlp layers</td>
<td>2</td>
</tr>
<tr>
<td>learning rate</td>
<td>0.1</td>
</tr>
<tr>
<td>learning rate decay</td>
<td>0.9</td>
</tr>
<tr>
<td>batch size</td>
<td>50</td>
</tr>
<tr>
<td>optimization</td>
<td>SGD</td>
</tr>
</tbody>
</table>

Table 2: GCN hyperparameters.

### 3.3 Overall Performance

In order to fully analyze the cross-lingual learning capabilities of our framework, we assess its performance by applying models trained with various combinations of training and test data from these three languages, as shown in Tables 3 and 4. We can see that the results are promising. For both tasks, the models trained from English are best, followed by Chinese, and then Arabic. We find that extraction task performance degrades as the accuracy of language-dependent tools (for sentence segmentation, POS tagging, dependency parsing) degrades.

Using English as training data, our cross-lingual transfer approach achieves similar performance on Chinese event argument role labeling (59.0%), compared to the model trained from Chinese annotations (59.3%), which is much higher than the best reported English-to-Chinese transfer result on event argument role labeling (47.7%) (Hsi et al., 2016).

We also show polyglot results for event argument role labeling in Table 4, by combing the training data from multiple languages. We observe that our model benefits from the combination of training data of multiple languages. The polyglot transfer learning does not provide further gains to relation extraction because the model converges quickly on a small amount of training data.

<table>
<thead>
<tr>
<th>Hyperparameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>word embedding size</td>
<td>300</td>
</tr>
<tr>
<td>POS embedding size</td>
<td>30</td>
</tr>
<tr>
<td>entity embedding size</td>
<td>30</td>
</tr>
<tr>
<td>dependency relation embedding size</td>
<td>30</td>
</tr>
<tr>
<td>hidden dimension size</td>
<td>200</td>
</tr>
<tr>
<td>dropout</td>
<td>0.5</td>
</tr>
<tr>
<td>number of layers</td>
<td>2</td>
</tr>
<tr>
<td>pooling function</td>
<td>max pooling</td>
</tr>
<tr>
<td>mlp layers</td>
<td>2</td>
</tr>
<tr>
<td>learning rate</td>
<td>0.1</td>
</tr>
<tr>
<td>learning rate decay</td>
<td>0.9</td>
</tr>
<tr>
<td>batch size</td>
<td>50</td>
</tr>
<tr>
<td>optimization</td>
<td>SGD</td>
</tr>
</tbody>
</table>

Table 2: GCN hyperparameters.
### 3.4 Ablation Study

Tables 5 and 6 show the impact of each feature category. We can see that all categories help except Chinese POS features for Relation Extraction and Arabic POS features for Event Argument Role Labeling.

Many Chinese word segmentation errors occur on tokens involved in names. For example, “总统(NOUN, president) 萨姆(PROPN, Samuel)·(PUNCT) 努乔马(PROPN, Nujoma)” is mistakenly tagged as “总(NOUN) 统萨姆(PROPN)·(PUNCT) 努乔(NOUN)马(NOUN)”.

In Arabic, sometimes one word is a combination of Noun and Verb. For example, the single word “الصراع” means “Israeli conflict” in English, including both a trigger and an argument, which are not separated by our tokenizer. In contrast, two entity mentions are unlikely to be combined into one word in Arabic, thus Relation Extraction does not suffer from tokenization errors and corresponding POS features.

### 3.5 Comparison with Supervised Approach

We also compare the results with supervised monolingual models trained from manual annotations in the same language. Figures 3 and 4 show the learning curves of these supervised models. For event argument role labeling, we can see that without using any annotations for the target language, our approach achieves performance comparable to the supervised models trained from more than 3,000 manually annotated event argument mentions, which equal to approximately 1,326 Chinese sentences and 1,141 Arabic sentences based on the statistics of ACE data.

Our model performs particularly well on relation types or argument roles that require deep understanding of wide contexts involving complex syntactic and semantic structures, such as PART-WHOLE:Artifact, PART-WHOLE:Geographical, GEN-AFF:Org-Location and ORG-AFF:Employment relations, and Injure: Victim argument role. Despite only having 14 training instances, our model achieves near 100% F-score on PART-WHOLE:Artifact relations when transferred from English to Chinese. Our model achieves 86% F1 score on PART-WHOLE:Geographical relations when transferred from English to Arabic, and 73% and 79% F1 scores on GEN-AFF:Org-Location relations when transferred from English to Chinese and to Arabic, respectively. In an Injure event, a Person can either be an Agent or a Victim. Surface lexical embedding features are often not sufficient to disambiguate them. Our model is effective at transferring structural information such as dependency relations between words, and obtains 72.97% F1 score on labeling Injure: Victim when transferred from English to Chinese, and 75.43% from English to Arabic.

In addition, our model achieves very high performance on event argument roles for which entity type is a strong indicator. For example, a weapon is much more likely to play as an Instrument rather than a Target in an Attack. Our model achieves 89.9% F1 score on Attack: Instrument and 91.4% F1 score on Personnel: POSITION argument roles when transferred from English to Chinese.

### 3.6 Using System Extracted Name Mentions

Table 7 shows the results of event argument role labeling on Chinese and Arabic using English as training data (with system generated entity mentions)

<table>
<thead>
<tr>
<th>Target Language</th>
<th>F1 score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chinese</td>
<td>56.9</td>
</tr>
<tr>
<td>Arabic</td>
<td>60.1</td>
</tr>
</tbody>
</table>

Table 7: Event Argument Role Labeling results (F1 %) on Chinese and Arabic using English as training data (with system generated entity mentions)
thus decrease the performance of the model, but the overall results are still promising.

4 Related Work

A large number of supervised machine learning techniques have been used for English event extraction, including traditional techniques based on symbolic features (Ji and Grishman, 2008; Liao and Grishman, 2011), joint inference models (Li et al., 2014; Yang and Mitchell, 2016), and recently with neural networks (Nguyen and Grishman, 2015a; Nguyen et al., 2016; Chen et al., 2015; Nguyen and Grishman, 2018; Liu et al., 2018b; Lu and Nguyen, 2018; Liu et al., 2018a; Zhang et al., 2018a, 2019). English relation extraction in the early days also followed supervised paradigms (Li and Ji, 2014; Zeng et al., 2014; Nguyen and Grishman, 2015b; Miwa and Bansal, 2016; Pawar et al., 2017; Bekoulis et al., 2018; Wang et al., 2018b). Recent efforts have attempted to reduce annotation costs using distant supervision (Mintz et al., 2009; Surdeanu et al., 2012; Min et al., 2013; Angeli et al., 2014; Zeng et al., 2015; Quirk and Poon, 2017; Qin et al., 2018; Wang et al., 2018a) through knowledge bases (KBs), where entities and static relations are plentiful. Distant supervision is less applicable for the task of event extraction because very few dynamic events are included in KBs. These approaches, however, incorporate language-specific characteristics and thus are costly in requiring substantial amount of annotations to adapt to a new language (Chen and Vincent, 2012; Blessing and Schütze, 2012; Li et al., 2012; Danilova et al., 2014; Agerri et al., 2016; Hsi et al., 2016; Feng et al., 2016).

Regardless of the recent successes in applying cross-lingual transfer learning to sequence labeling tasks, such as name tagging (e.g., (Mayhew et al., 2017; Lin et al., 2018; Huang et al., 2019)), only limited work has explored cross-lingual relation and event structure transfer. Most previous efforts working with cross-lingual structure trans-
fer rely on bilingual dictionaries (Hsi et al., 2016), parallel data (Chen and Ji, 2009; Kim et al., 2010; Qian et al., 2014) or machine translation (Faruqui and Kumar, 2015; Zou et al., 2018). Recent methods (Lin et al., 2017; Wang et al., 2018b) aggregate consistent patterns and complementary information across languages to enhance Relation Extraction, but they do so exploiting only distributional representations.

Event extraction shares with Semantic Role Labeling (SRL) the task of assigning to each event argument its event role label, in the process of completing other tasks to extract the full event structure (assigning event types to predicates and more specific roles to arguments). Cross-lingual transfer has been very successful for SRL. Early attempts relied on word alignment (Van der Plas et al., 2011) or bilingual dictionaries (Kozhevnikov and Titov, 2013). Recent work incorporates universal dependencies (Prazák and Konopík, 2017) or multilingual word embeddings for Polyglot SRL (Mulcaire et al., 2018). Liu et al. (2019) and Mulcaire et al. (2019) exploit multi-lingual contextualized word embedding for SRL and other Polyglot NLP tasks including dependency parsing and name tagging. To the best of our knowledge, our work is the first to construct a cross-lingual structure transfer framework that combines language-universal symbolic representations and distributional representations for relation and event extraction over texts written in a language without any training data.

GCN has been successfully applied to several individual monolingual NLP tasks, including relation extraction (Zhang et al., 2018b), event detection (Nguyen and Grishman, 2018), SRL (Marcheggiani and Titov, 2017) and sentence classification (Yao et al., 2019). We apply GCN to construct multi-lingual structural representations for cross-lingual transfer learning.

5 Conclusions and Future Work

We show how cross-lingual relation and event argument structural representations may be transferred between languages without any training data for the target language, and conclude that language-universal symbolic and distributional representations are complementary for cross-lingual structure transfer. In the future we will explore more language-universal representations such as visual features from topically-related images and videos and external background knowledge.

Acknowledgement

This research is based upon work supported in part by U.S. DARPA LORELEI Program HR0011-15-C-0115, the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via contract FA8650-17-C-9116, and ARL NS-CTA No. W911NF-09-2-0053. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of DARPA, ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

References


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).


Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP.


Tongtao Zhang, Heng Ji, and Avirup Sil. 2019. Joint entity and event extraction with generative adversarial imitation learning. *Data Intelligence Vol 1 (2).*

