A Multiple Instance Learning Framework for Identifying Key Sentences and Detecting Events

Wei Wang*, Yue Ning*, Huzefa Rangwala†, Naren Ramakrishnan*
* Discovery Analytics Center, Department of Computer Science, Virginia Tech, Arlington, VA 22203
† Department of Computer Science, George Mason University, Fairfax, VA 22030
* tskatom, yning, naren@vt.edu
† rangwala@cs.gmu.edu

ABSTRACT

State-of-the-art event encoding approaches rely on sentence or phrase level labeling, which are both time consuming and infeasible to extend to large scale text corpora and emerging domains. Using a multiple instance learning approach, we take advantage of the fact that while labels at the sentence level are difficult to obtain, they are relatively easy to gather at the document level. This enables us to view the problems of event detection and extraction in a unified manner. Using distributed representations of text, we develop a multiple instance formulation that simultaneously classifies news articles and extracts sentences indicative of events without any engineered features. We evaluate our model in its ability to detect news articles about civil unrest events (from Spanish text) across ten Latin American countries and identify the key sentences pertaining to these events. Our model, trained without annotated sentence labels, yields performance that is competitive with selected state-of-the-art models for event detection and sentence identification. Additionally, qualitative experimental results show that the extracted event-related sentences are informative and enhance various downstream applications such as article summarization, visualization, and event encoding.

Keywords

MIL; CNN; Deep Learning; Event Detection; Information Extraction

1. INTRODUCTION

Identifying and extracting relevant information from large volumes of text articles play a critical role in various applications ranging from question answering [40], knowledge base construction [38] and named entity recognition [33]. With the rise and pervasiveness of digital media such as news, blogs, forums and social media, automatically detecting the occurrence of events of societal importance, further categorizing them by performing event classification (i.e., type of event) and automatically/manually encoding them are popular areas of research. As a case in point, Muthiah et al. [26] use a dictionary-based approach to first identify news articles pertaining to planned civil unrest events, extract key indicators of these events and later use this information to predict the onset of protests. Other applications include event forecasting [31], social media monitoring [13] and detecting financial events [1].

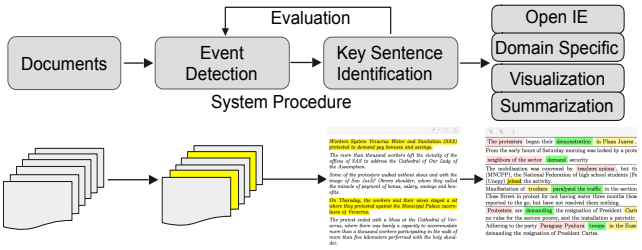

Figure 1: System Overview

We decompose prior work on event analysis into two interrelated subproblems: (1) event detection (or recognition): identification of the documents describing a specific event; (2) event encoding (or extraction): identification of the phrases, tokens or sentences (with relationships) that provide detailed information about the event e.g., type of event, location of event, people involved, and time of event. Event detection and encoding pose multitude of challenges due to the variety of event domains, types, definitions and expectations of these algorithms.

In general, the efforts on event encoding (extraction) can be categorized into two groups: open information extraction and domain-specific event extraction. Open information extraction methods normally take text as input and output tuples that include two entities and the relationship between them (e.g., Teachers (entity), government (entity), protest against (relationship)). Domain-specific event extraction approaches rely on templates, dictionaries, or presence of a specific structure within the input text. These input templates for events vary dramatically based on different situations. For instance, an earthquake event template might contain location, magnitude, missing people, damaged infrastructure, and time. Whereas, a civil unrest event template might contain fields like participants, purpose, location, and time. Most prior event extraction research [5, 28, 42] has focused on extracting entities, detecting trigger terms (or keywords), and matching up event slots on pre-defined templates. Huang et. al. [12] propose a boot-

CIKM'16 , October 24-28, 2016, Indianapolis, IN, USA

DOI: http://dx.doi.org/10.1145/2983323.2983821

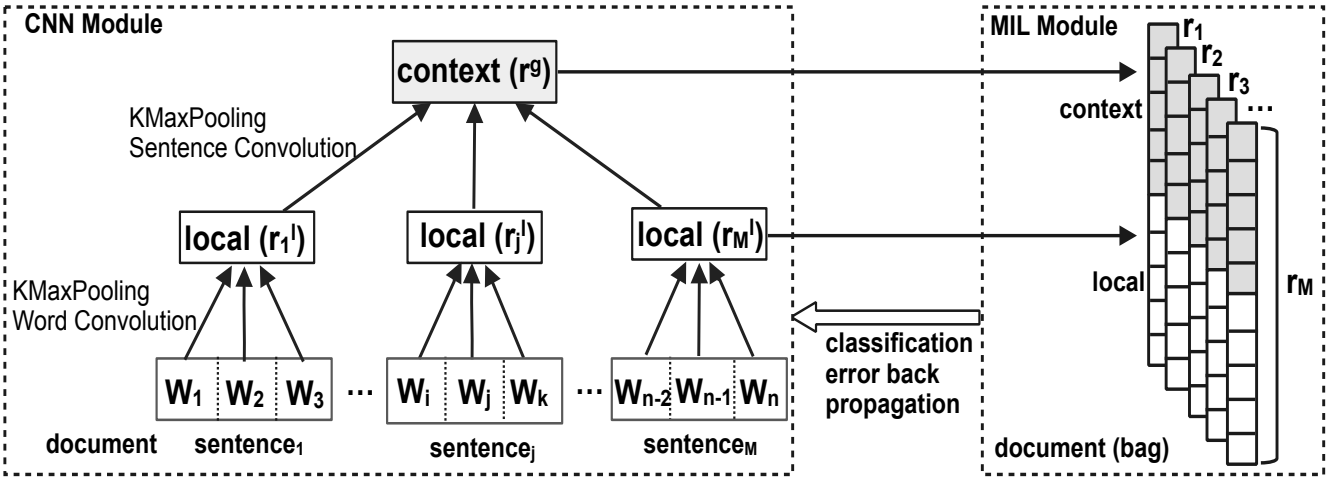

Figure 2: MI-CNN Model Overview.

strapping approach to learn event phrase, agent term, and purpose phrase for event recognition. Entity-driven probabilistic graphical models [27, 4] were proposed to jointly learn the event templates and align the template slots to identified tokens.

In this study, we view the twin problems of event detection and extracting key sentences to enable event encoding and classification in a unified manner as a form of multiple instance learning (MIL) [7]. This enables us to identify salient event-related sentences from news articles without training labels at the sentence level and the use of manually defined dictionaries. Our motivation stems from the practical contexts in which event extraction systems operate (see Figure 1). In a typical news article, there exist a small set of key sentences that provide detailed information for a specific event. Identifying these sentences automatically is useful for succinctly summarizing the news article. Highlighting such key sentences within a visualization tool will enable a human analyst to quickly locate important information, rapidly scan the contents of the article, and make timely decisions. Additionally, as we will demonstrate, key sentences can form the basis of automated event encoding and we can extract the final event based on the identified salient sentences. Figure 1 provides an overview of the methods developed in this paper.

In more detail, we propose an MIL approach based on convolutional neural networks (CNN) that incorporates a distributed representation of documents to extract event-related sentences. Specifically, we consider each individual sentence within a document to be an instance and the collection of instances within a document as a bag. We are provided with labels at only the bag (document) level. A positive label indicates that the news article refers to a protest related event. Our model seeks to predict the labels at the document and sentence levels but with no available sentence-level labels during training. Traditional MIL formulations [7, 2, 21] treat each instance (sentence) within a bag (document) as independent of each other. Our model relaxes this strong assumption by combining local and global context information to construct a continuous sentence level representation. We evaluate our proposed model on the specific domain of civil unrest events such as protests, strikes and "occupy events", with data obtained from ten Latin American Countries.

The major contributions of our work can be summarized as follows:

- We propose a novel framework which views event detection and identification of key sentences as a form of multiple instance learning.
- We develop a novel sentence representation that combines local and global information using convolutional neural network formalisms.
- We propose a new MIL-based loss function that encourages selection of a small set of salient sentences for the protest articles.

2. PROBLEM DEFINITION

Given a set of N news articles, $\{x_i\}, i = 1..N$, each news article is associated with a label $y_i \in \{0, 1\}$ indicating whether the article refers to a protest event or not.

Our goals here are twofold. First, we aim to predict the event label $\hat{y_i}$ for each news article x_i. This is the standard text classification formulation for solving the event detection (recognition) problem. Our second goal is to extract a small set of salient sentences that are considered as indicative (key) of event related information. The dynamic number $k = |x_i| \times \eta$ of sentences to extract is decided by the length of the article, where η in $(0, 1]$ is a predefined value. We define the second task as **key sentences extraction** problem. The extracted key sentences are helpful for related tasks such as event detection, classification, encoding, summarization, and information visualization.

3. PROPOSED MODEL

We propose a multiple instance learning (MIL) model based on convolutional neural networks (MI-CNN) for our task. Each text article is considered as a *bag* and sentences within the bag are individual instances. We have labels only for the article-level (bags) and do not have individual ground truth labels available for each sentence (instances). Similar to MIL formulations [21, 17], we seek to predict the

document-level labels and transfer the labels from the bag-level to individual sentences to identify the key sentences summarizing the protest-related information.

We utilize CNN to construct a distributed representation for each instance (sentence), that are the input to the MIL framework. Using the feedback from MIL training process, the CNN module updates the instance representation. For every sentence within an article, our model estimates a sentence-level probability that indicates the belief of the sentence indicating event related information. The MI-CNN applies an aggregation function over the sentences to compute a probability estimate for an article referring to a protest. Figure 2 provides an overview of our proposed model.

3.1 Instance Representation

As seen in Figure 2, the raw word tokens from the article are input into the network. Given that a sentence s consists of D words $s = \{w_1, w_2, ..., w_D\}$, every word w is converted to a real value vector representation using a pretrained word embedding matrix W. The individual word representations are then concatenated for every sentence. The embedding matrix $W \in R^{d \times |V|}$, where d is the embedding dimension and V is a fixed-sized vocabulary, will be fine-tuned during the training process.

The first convolution and k-max pooling layer are used to construct the local vector representations for every sentence referred by $\mathbf{r_j}^l$ for the j-th sentence. The convolutional layer scans over the text, produces a local feature around each word and captures the patterns regardless of their locations. The k-max pooling layer only retains the k-most significant feature signals and discards the others. It creates a fixed-sized local vector for each sentence.

Given a sentence, s, the convolution layer applies a sliding window function to the sentence matrix. The sliding window is called a kernel, filter, or feature detector. Sliding the filter over the whole matrix, we get the full convolution and form a feature map. Each convolution layer applies different filters, typically dozens or hundreds, and combines their results. The k-max pooling layer applied after the convolution layer output k values for each feature map. In addition to providing a fixed-size output matrix, the pooling layer reduces the representation dimensionality but tends to keep the most salient information. We can think of each filter as detecting a specific feature such as detecting if the sentence contains a protest keyword. If this protest-related phrase occurs somewhere in the sentence, the result of applying the filter to that region will produce a large value, and small values in other regions. By applying the max operator we are able to keep information about whether or not the feature appears in the sentence.

The local features, $\mathbf{r_j}^l$, aim to capture the semantic information embedded within the scope of the j-th sentence. These local representations are then transformed using another convolution and k-max pooling layer above to construct the article-level context representation, denoted by $\mathbf{r}^g$. The context features $\mathbf{r}^g$ capture the information across all the sentences within the article and are shared by all the sentences. For every sentence, it's specific local representation is concatenated with the context representation and used for the MIL-based optimization. This combined representation is denoted by $\mathbf{r_j}$ for the j-th sentence.

$$\mathbf{r_j} = \mathbf{r_j}^l \oplus \mathbf{r}^g, \tag{1}$$

where $\oplus$ is the concatenation operator. Intutively, the context feature vector encodes topic information of the document and is useful for distinguishing the theme and disambiguating polysemy encoded in local features, $\mathbf{r_j}^l$. For instance, a sentence containing the token *strike* may refer to a civil unrest event, but it is also often related to a military activity. Without context information, it is very hard to make this decision.

3.2 Sentence- and Document-Level Estimates

Given the distributed representation $\mathbf{r_j}^i$ of the j-th sentence in the document x_i, we compute a probabilistic score p_j^i using a **sigmoid** function:

$$p_j^i = \sigma(\theta^T \mathbf{r_j}^i + b_s) \tag{2}$$

where θ is the coefficient vector for sentence features and b_s is the bias parameter. Intuitively, p_j^i is the probability that the j-th sentence within article x_i refers to information pertaining to a protest. Aggregating these estimated probabilities over these indicative sentences will provide an estimate for a document to indicate a protest event. To alleviate the bias of varying lengths of different articles, we choose a predefined ratio, η (set to 0.2), to choose the dynamic number of key sentences. We choose the set of top highly ranked sentences K_i as key sentences in each article x_i. $|K_i| = max(1, \lfloor |x_i| \times \eta \rfloor)$. Generally, we will select one or two sentences each article given η as 0.2 in our dataset.

We compute the probability P_i of an article referring to a civil unrest event as the average score of the key sentences:

$$\text{Prob}(y_i = 1) = P_i = \frac{1}{|K_i|} \sum_{k \in K_i} p_k^i \tag{3}$$

There are several other common options to aggregate the instance probabilities to bag-level probability. Averaging is one of the most common aggregation functions. It is suitable for the cases where the bag label is decided by majority rule. Another common option is the max function. In this case, the bag probability is decided by the most significant instance. Noise-OR is also a aggregation function used often in MIL. It tends to predict bags to be positive due to its natural property. In protest news articles, there often exists a small set of sentences indicating the occurrence of a protest event and remaining sentences are often related to the background or discussion about that event. In this case, using the average over all sentences makes the salient sentences indistinguishable from the background sentences. However, using the *max* function makes the model sensitive to longer documents. We ran preliminary experiments based on these different aggregation functions.

3.3 Multiple Instance Learning (MIL)

During training, the input to the MIL module is a document x_i consisting of individual sentences; label $y_i \in \{0, 1\}$ provided for the document. To encourage the model to select meaningful key sentences, we design a compositional cost function that consists of four components: bag-level loss, instance ratio, instance-level loss, and an instance-level manifold propagation term. The loss function is given by:

$$L(x, y; \theta, W, F, b) = \underbrace{\frac{1}{N}\sum_{n}^{N}(1-y_n)\log P_n + y_n \log(1-P_n)}_{\text{bag-level loss}}$$

$$+ \underbrace{\frac{\alpha}{N}\sum_{n}^{N} y_n \max(0, |K_n| - Q_n) + (1-y_n)Q_n}_{\text{instance ratio control loss}}$$

$$+ \underbrace{\frac{\beta}{N}\sum_{n}^{N}\frac{1}{M_n}\sum_{m}^{M_n}\max(0, m_0 - \text{sgn}(p_m^n - p_0)\theta^T r_m^n)}_{\text{The instance-level loss}}$$

$$+ \underbrace{\frac{\gamma}{(\sum_n M_n)^2}\sum_{n}^{N}\sum_{i}^{N}\sum_{m}^{M_n}\sum_{j}^{M_i}(p_m^n - p_j^i)^2 e^{(-\|r_m^n - r_j^i\|_2^2)}}_{\text{instance-level manifold propagation}}$$

where $Q_n = \sum_m 1(p_m^n > 0.5)$ is an indicator function that returns the number of instances with a probability score greater than 0.5. N is the number of documents and M_n is the number of sentences in n-th document. Hyper-parameters α, β, and γ control the weights of different loss components. Dropout layers are applied on both word embedding and sentence representation to regularize the model, and Adadelta [43] is used as the model optimization algorithm. We used a dropout rate of 0.2 in the word convolutional layer and 0.5 in the sentence convolutional layer. α, β and γ are 0.5, 0.5 and 0.001, and are chosen by cross-validation on training set, respectively.

- **Bag Level Loss**: this component is the classical cross-entropy loss for classification which penalizes the difference between predictions and the true labels for bags.
- **Instance Ratio Control Loss**: this component encourages no sentence in the negative article to have a high probabilistic estimate and pushes the model to assign high probabilistic estimates to a smaller set of sentences in the positive articles.
- **The Instance-Level Loss**: this part is a standard hinge loss that encourages wider margin (m_0) between positive and negative samples. Here sgn is the sign function. The hyper parameter m_0 and p_0 control the sensitivity of the model. p_0 determines positiveness of instance. We set m_0 as 0.5 and p_0 as 0.6 in our case.
- **Instance-level Manifold Propagation**: Inspired by [17], the manifold propagation term encourages the similar sentence representations to have similar predictions/estimates.

To optimize the cost function we use mini-batch stochastic gradient descent. This approach was found to be scalable and insensitive to the different parameters within the proposed model. A backpropagation algorithm is used to compute the gradient in our model. In our experiments, the MI-CNN model was implemented using the Theano framework [3].

Table 1: Event population and Type

Event Population
General Population
Business
Legal
Labor
Agricultural
Education
Medical
Media

Event Type
Government Policies
Employment and Wages
Energy and Resources
Economic Policies
Housing

4. EXPERIMENTS

4.1 Dataset

In our experiments, we use a manually labeled dataset (GSR; Gold Standard Report) of Spanish protest events from ten Latin America countries [1] from October 2015 to Jan 2016. The dataset consists of 19795 news articles that do not refer to a protest (negatives) and 3759 articles that are protest-related (positives). For each positive article, the GSR provides the population and event type of the protest event. The event population indicates the type of participants involved in the protest. The event type identifies the main reason behind the protest. The set of event population and event types are listed in Table 1. Each annotated sample is checked by three human analysts and the labels are confirmed if two of them agree on the assignment. We use 5-fold cross validation for evaluation. On average, we have 18844 articles for training and 4710 for test in each fold. Since the dataset is imbalanced we report precision, recall, and F1 score computed for the positive class for our experiments.

During the data pre-processing phase, we augment a special token (*padding*) by $\frac{T-1}{2}$ times to the beginning and the end of the sentence, where T is the window size of filter in word convolution layer. For the mini-batch setting in Theano, we define two variables max_s and max_d to control the maximum number of tokens for each sentence and maximum number of sentences for each document. The special token (*padding*) is appended to the end of each sentence until max_s is achieved. Likewise, the *padding* sentences are attached to the end of a document until max_d achieved. We set max_s as 70 and 30 for max_d in our experiments.

Pretrained Word Embedding For the initial word embeddings, we use a training corpus consisting of 5.7 million Spanish articles ingested from thousands of news and blog feeds covering Latin America area during the time period of Jan 2013 to April 2015. The open source tool *word2vec* [24] is used for pretraining word embeddings in our experiments. We set the word embedding dimension as 100. Tokens appearing less than ten times are removed and we use the skip-gram structure to train the model.

[1]Argentina, Brazil, Chile, Colombia, Ecuador, El Salvador, Mexico, Paraguay, Uruguay and Venezuela

Table 2: Hyperparameters for MI-CNN model

N	Batch size	50
max_w	Max number of words in sentence	70
max_s	Max number of sentences in a article	30
f_w	Number of feature maps in word Conv layer	50
f_s	Number of feature maps in sentence Conv layer	100
k_w	k-max pooling parameter in word Conv layer	3
k_s	k-max pooling parameter in sentence Conv layer	2
T_w	Filter window size in word Conv layer	5
T_s	Filter window size in sentence Conv layer	3
η	The ratio of choose key sentences	0.2
α	The control parameter of Instance ratio control loss	0.5
β	The control parameter of Instance level loss	0.5
γ	The control parameter of Instance level manifold propogation	0.001
$drop_w$	Dropout rate in word Conv Layer	0.2
$drop_s$	Dropout rate in sentence Conv Layer	0.5
d	Pretrained word embedding dimension	100

4.2 Comparative Methods

Support vector machines (SVM) are known to be effective for the standard text classification problem [14, 37]. We use SVM as one of the baseline models for the article classification problem. We remove Spanish stop words, apply lemmatization on tokens, and use TF-IDF features.

The second comparative approach used in our study is a CNN with a softmax classifier. The CNN model first constructs a sentence vector by applying convolution and k-max pooling over word representations. Then a document vector is formed over sentence vectors in a similar way. Finally, the softmax layer uses the document vector as input and predicts the final label.

Although the SVM and CNN model can classify whether an article refers to a protest or not, they do not directly output the key sentences referring to the events. Both SVM and CNN models construct a document level representation (global information) and use it as input to final classifier; we refer to them as global methods.

As opposed to global methods, local methods assign credit to individual sentences and make the final decisions based on an aggregation function applied over the individual sentences. As such, these approaches can extract the set of significant sentences along with an article-level label prediction. The multiple instance support vector machine (MISVM) [8], group instance cost function (GICF) [17] (discussed in Section 4.3.1), and our proposed approach (MI-CNN) are all local methods. To train the GICF and MISVM models, we use the sentence representation learned from the CNN model as instance features. Table 2 shows the hyper-parameters used in the model MI-CNN.

4.3 Experimental Results

4.3.1 Event Detection (Article Classification)

Table 3 shows the classification results for MI-CNN and comparative approaches for identifying whether a news article is "protest-related" or not. We report the mean precision, recall and F1 score along with standard deviation across five folds. The MI-CNN approach outperforms all other baseline methods. Both MISVM and GICF models have relatively poor performance on this dataset. Specifically, the MI-CNN model outperforms GICF by 40% and MISVM by 20% with respect to the F1 score. One possible explanation for the poor performance of MISVM and GICF is that the sentence vectors learned from CNN model only capture the local information (sentence level) but ignore the contextual information important for article classification. In contrast to the GICF model, which uses fixed sentence representation learned from the CNN model, MI-CNN updates the sentence representation during the training process according to the feedback from the multiple instance classification.

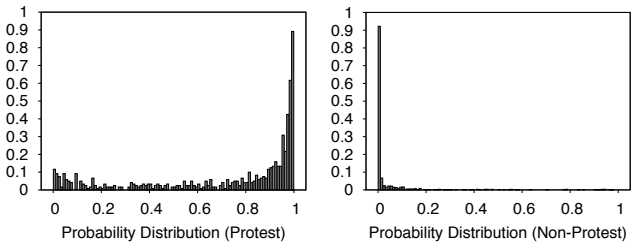

Figure 3: The histogram of probability estimates for protest and non-protest articles for test set

Importance of Context Information To show that context information is helpful when encoding the sentence representation, we performed a set of experiments based on the variants of our MI-CNN model. We trained a model referred by MI-CNN (max), which does not add context information to the sentence vector. The maximum score of sentences in the article is used as the probability of an article to be positive. Different from MI-CNN (max) model, the second variant MI-CNN (avg) model infers the probability of a positive article as the average score over all the sentences. In the model referred by MI-CNN (context + k-max), the context information encoded into the sentence level representation and the dynamic top "k" sentences are used to infer the probability a given article to be positive (i.e., protest).

As shown in Table 3, the MI-CNN (max) model has worse performance when compared with SVM, CNN and two other MI-CNN models, which all use the global information to some extent. This experiment shows that exclusively using the local information is not beneficial for the classification task. Further, MI-CNN (context + k-max) achieves the best performance confirming the importance of context information.

Probability Distributions Figure 3 presents the distribution of the estimated document level probability estimates for protest and non-protest articles based on the aggregation of key sentence-level probability estimates. Within the MIL formulation, the sentence-level (instances within each bag) loss function attempts to separate the margin between the positive and negative sentences. The results show the stability of our predictions, because the majority of estimated probabilities for the protest articles are greater than 0.8, whereas for the non-protest articles are smaller than 0.2.

4.3.2 Identifying Key Sentences

In addition to classifying whether an article is reporting a civil unrest event or not, our model also extracts the most indicative sentences for each article. We perform a qualitative and quantitative evaluation of the indicative sentences.

Quantitative Evaluation

Since we do not have available ground truth data for the key sentences, we evaluate the quality of our identified sentences by comparing with sentences selected by several methods. We assume that key sentences should be discriminative about protest references. If we only use the selected sentences to represent the whole document and apply an article label classifier on these documents, we expect

Table 3: Event detection performance. comparison based Precision, Recall and F-1 score w.r.t to state-of-the-art methods. The proposed MI-CNN method outperform state-of-the-art methods

	Precision(Std.)	**Recall(Std.)**	**F1(Std.)**
SVM	0.818 (0.019)	0.720 (0.008)	0.765 (0.009)
MISVM	0.724 (0.030)	0.584 (0.017)	0.646 (0.018)
CNN Model	0.732 (0.033)	0.783 (0.026)	0.756 (0.007)
GICF	**0.833 (0.019)**	0.421 (0.09)	0.553 (0.086)
MI-CNN (max)	0.685 (0.030)	0.730 (0.029)	0.706 (0.018)
MI-CNN (avg)	0.731 (0.069)	0.789 (0.042)	0.759 (0.026)
MI-CNN (context + k-max)	0.742 (0.036)	**0.813 (0.041)**	**0.775(0.006)**

Table 4: Event detection performance using key sentences only.

	Prec.(Std.)	Recall(Std.)	F1(Std.)
Keywords Protest	0.755 (0.021)	**0.638 (0.017)**	0.692 (0.018)
Random Sentences	0.681 (0.026)	0.433 (0.019)	0.551 (0.018)
Start/End Sentences	0.751 (0.022)	0.555 (0.026)	0.638 (0.019)
MI-CNN	**0.761 (0.015)**	0.635 (0.024)	**0.693 (0.019)**

that the selected sentences with higher quality will have better classification performance. In our experiment, we try three other methods for extracting the same number of sentences and apply the SVM classifier. The first baseline method (**Random**) randomly chooses sentences from a given article. News articles generally organize important information at the start and end of a document. As such, we select the first $\frac{k}{2}$ sentences from the start and $\frac{k}{2}$ from the end of an article as another baseline (**Start/End**). The third method (**Keywords**) selects sentences containing protest-related keywords such as *demonstration, march, protest* based on an expert-defined dictionary.

Table 4 shows the comparative results of the above outlined approaches. The MI-CNN approach outperforms all other methods with respect to F1 score. As expected, all methods show better performance than randomly choosing sentences. Using the sentences with protest-related keywords has the highest recall. However, this approach has a higher chance of false positives due to polysemy. For example, the term *march* can refer to the protest movement as well as the month of year. A significant strength of our proposed model compared to the keyword approach is that our model does not require any domain experts to curate a dictionary of protest keywords and is easier to adapt to new and unknown domains with minimal effort.

In addition to using the classifier to evaluate the quality of the sentences extracted by our model, we randomly choose 100 protest articles represented by key sentences for manual evaluation. We ask three annotators to determine whether the extracted sentences refer to a protest event. If the sentence contains the participant and protest action information, we consider that the method correctly identified a sentence referring to a protest event. In case of inconsistencies amongst the human evaluators, the final decision is decided by a simple majority. The annotators agreed with each other 95% of the time in our labeling process. Figure 4 presents this human-based evaluation result. Our model has the highest average accuracy and least variance. The average accuracy that our model achieves is approximately 10% higher than **keywords** approach and 80% than **Start/End** approach.

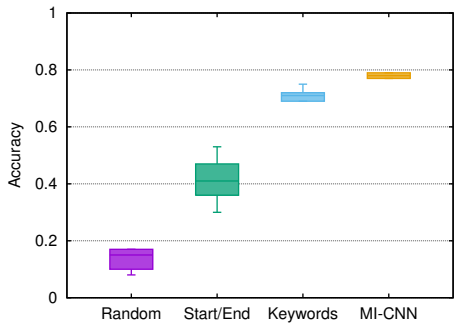

Figure 4: Event Reference Accuracy for Protest Articles

Workers System Veracruz Water and Sanitation (SAS) protested to demand pay bonuses and savings.

The more than thousand workers left the vicinity of the offices of SAS to address the Cathedral of Our Lady of the Assumption.

Some of the protesters walked without shoes and with the image of San Jose Obrero shoulder, whom they called the miracle of payment of bonus, salary, savings and benefits.

On Thursday, the workers and their wives staged a sit where they protested against the Municipal Palace caceroLazos of Veracruz.

The protest ended with a Mass at the Cathedral of Veracruz, where there was barely a capacity to accommodate more than a thousand workers participating in the walk of more than five kilometers performed with the holy shoulder.

Angelica Navarrete, general secretary of the Union of SAS, insisted on Tuesday that if they do not receive what they owe, they will strike.

During the march, at the height of Zamora Park, a passenger bus of the coastline they were pounced on protesters, upset because he wanted to spend and the march went through, but no injuries.

According to the protesters, the SAS, owed to workers 85 thousand 300 million pesos.

.....

Case Study 1: Key sentences are highlighted within a protest news article.

Qualitative Evaluation and Case Studies

A useful application of identifying the key sentences is text summarization and visualization. Our model can assist a human analyst in quickly identifying the key information about an event without reading an entire document. Case

Table 5: List of positive and negative sentences selected by our model sorted by score: The positive sentences show common patterns that include location references and purpose-indicating terms. The negative sentences may contain protest keywords, but are not related to a specific civil unrest event. The third and forth columns show whether the titled methods also select the same sentence as our approach as the key sentence. The pink color highlights the protest participant, green for protest keyword and yellow for location

Positive Sentences	Score	Keywords	Start/End
The protesters began their demonstration in Plaza Juarez, advanced by 16 September to Hidalgo.	0.9992	Yes	No
From the early hours of Saturday morning was locked by a protest the Francisco Fajardo highway from Caricuao, neighbors of the sector demand security	0.9991	Yes	No
The mobilization was convened by teachers unions, but the national March of public colleges and private (MNCPP), the National Federation of high school students (Fenaes) and the Center Union of secondary students (Unepy) joined the activity.	0.9991	No	No
Manifestation of truckers paralyzed the traffic in the section clean- Roque Alonso	0.9991	Yes	Yes
Close Street in protest for not having water three months those who protested pointed out that the problem was reported to the go, but have not resolved them nothing.	0.9991	Yes	Yes
Protesters are demanding the resignation of President Cartes, since they consider that - as they understand - no rules for the sectors poorer, and the installation a patriotic junta in power.	0.9991	Yes	No
Adhering to the party Paraguay Pyahura troops in the Eusebio Ayala Avenue heading to downtown Asuncion, demanding the resignation of President Cartes.	0.9991	No	Yes
From 09:00 hours, tens of inhabitants of the municipal head were concentrated at the entrance of Arcelia and almost 10 o'clock began a March toward the Center, which showed banners against staff of the PF.	0.999	Yes	No
Nurses were stationed opposite the hospital with placards to demand to the authorities of the IPS that their claims are solved immediately.	0.9989	No	No
A group of taxi drivers protested this Monday morning in the central town of el Carrizal municipality, in Miranda State, according to @PorCarrizal the demonstration is due to that, he was denied the circulation to the drivers who benefited from the transport mission.	0.9988	Yes	Yes
Negative Sentences	**Score**	**Keywords**	**Start/End**
Bled some guardians, also protesters, friends and family that went with them.	0.172	Yes	No
The parade by the 195 years of independence of Ambato yesterday (November 12) had a different connotation.	0.0125	Yes	No
This morning, the situation is similar, as already record barricades and demonstrations in the same place, by what police is already around the terminal.	0.0109	Yes	No
The young man asked that they nicely other costume to so participate in the parade.	0.0097	No	No
Employees announced that they will be inside until you cancel them owed assets.	0.0093	No	No
Workers arrived Thursday to the plant where the only person who remained on duty in the place who has not claimed his salary joined the protest.	0.0088	No	No

Study 1 shows a demonstration of this practical application where key sentences within a news article are highlighted. From the highlighted sentences, we can easily find key information such as the which entity (*who*) against which entity, the details and reason behind the protest (*what, why*) and the location and time of the protest if available (*where, when*).

Table 5 shows the set of top positive sentences ordered by probability scores, as well as the set of negative sentences. Different event roles are also being highlighted with different colors in the text. [2] We report common patterns among the positive sentences. For instance, most of them contain the location information such as *in Plaza Juarez, in the Eusebio Ayala Avenue*. Another common pattern is that the indicative sentences often contain some purpose-indicating words such as *demand, against*. From analyzing the negative sentences, we find that they may include some protest related words such as $protest, protestor, parade$, but are assigned lower scores because of the lack of protest action pattern and contextual information.

Further, in the last two columns of Table 5, we show whether the keywords and start/end methods also select our high ranked sentences as key sentences. We find that the keywords method has a high overlap with our method for the positive sentences. However it also introduces false positives as shown for the negative sentences.

4.3.3 Event Type and Population-Specific Tokens

For every protest article, the GSR provides a specific classification as it relates to the event "population" and "type". Representing the protest articles by the identified key sentences we extract the most frequent words within these sentences and report them in Table 6 in descending order of the normalized frequency score. Specifically, for each class c_p and c_e in event population and event type, we assign a score to each word w to evaluate it's contribution to a given class. The score function is a normalized word frequency given by:

$$\text{Score}_c(w) = f_{c,w} \log \frac{N}{n_w}, \tag{4}$$

where, $c \in \{c_p, c_e\}$, $f_{c,w}$ is the frequency of token w for class c, n_w is number of documents containing w. N is the total set of articles. From Table 6, we see that many of these terms are recognizable as terms about Business, Media and Education (event population) and Housing and Economic (event type). For instance, terms such as "sellers" and "commercial" have been chosen as top words in the key sentences in business articles. "Students", "education" and "teachers" are selected with higher weights in news articles in education category although some neutral words such as "national" are also identified.

4.3.4 Event Encoding

As a downstream application, we explored the capability of encoding (extracting event information) events from the identified key sentences. Since, the event encoding task is not the main focus of our work, we try previously developed open information extraction tools for this purpose.

[2] The text examples listed in this section are translated from Spanish to English using Google Translate Tool.

We use ExtrHech [44], a state-of-the-art open information extraction tool. ExtrHech is a Spanish Open IE tool based on syntactic constraints over parts-of-speech (POS) tags within sentences. It takes sentences as input and outputs the relations in the form of tuples (argument 1; relation; argument 2). Table 7 shows a list of events extracted by ExtrHech. We notice that ExtrHech is good at capturing the event population and action information, but not good for the "event type" information. The reason might be that the syntactic rules in ExtrHech are more suitable for capturing the pattern (Subject, Predicate, Object). For instance, ExtrHech captured entity words such as "campus" (indicating education), "pension" , "producers" (indicating business), "mayor", "gendarmes" (indicating Legal).

5. RELATED WORK

5.1 Event Extraction

Event detection/extraction with online open source datasets has been a large and active area of research in the past decades. In political science field, there have been several systems such as GDELT [19], ICEWS [30], and EL:DIABLO [34] working on extracting political events from online media. Supervised, unsupervised, and distant supervision learning techniques have been developed to tackle different domains and challenges.

Supervised learning approaches often focus on handcrafted ontologies and heavily rely on manually labeled training datasets at the sentence, phrase, and token levels. Chen and Li *et al.* [5, 20] utilize the annotated arguments and specific keyword triggers in text to develop an extractor. Leveraging social network datasets, Social Event Radar [11] is a service platform that provides alerts for any merchandise flaws, food-safety related issues, unexpected eruption of diseases, or campaign issues towards the government through keyword expansion. Social streams such as Twitter [41, 32] have been used for event records extraction and event detection. Event structure in open domains are mostly complex and nested. Supervised event extraction [23],[18] has been studied by analyzing the event-argument relations and discourse aspects of event interactions with each other. Even though, these methods often achieve high precision and recall, they do not scale to large datasets due to the limited availability of low level labeled data. Different from these approaches, our method utilizes the multi-instance formulation to propagate the labels from article level to sentence and phrase level. The proposed method is suitable because training data is easily available at the document level rather than per-sentence level.

In the unsupervised setting, approaches have been developed [22, 29] that model the underlying structure by jointly modeling the role of multiple entities within events or modeling an event with past reference events and context. Approaches [4, 27] extract the event without templates based on probabilistic graphical models. The advantage of unsupervised approaches is that they don't require any labeled data and might be able to use the large quantities of unlabeled data, available online. The disadvantage of unsupervised methods is that they might suffer due to noisy information and concept drift.

Between supervised and unsupervised approach, distant supervision methods try to mitigate their disadvantages and often utilize the public knowledge base to generate training samples. Mintz et al. [25] use Freebase relations and find sentences which contain entities appearing in these relations. From these sentences, they extract text features and train a classifier for relation classification.

5.2 Multiple Instance Learning

Multiple Instance Learning (MIL) [7] is developed for classifying groups of instances called "bags". In standard MIL formulation, individual instance level labels are not available and labels are provided only at the group/bag level. Each *bag* is labeled positive if it contained at least one positive instance and negative otherwise. This MIL formulation makes strong assumptions regarding the relationship between the bag and instance-level labels. There are approaches that extend Support Vector Machines (SVM) for the MIL problem [2, 9] which include: (i) modifying the maximum margin formulation to discriminate between bags rather than individual instances and (ii) developing kernel functions that operate directly on bags (MI-SVM, evaluated in this paper). Specifically, the generalized MIL [39] assumes the presence of multiple concepts and a bag is classified as positive if there exists instances from every concept. Relevant to our work, besides predicting bag labels, Liu *et al.* [21] seek to identify the key instances within the positively-labeled bags using nearest neighbor techniques. The recent work of [17] focuses on instance-level predictions from group level labels (GICF) and allows for the application of general aggregation functions while inferring the sentiment associated with sentences within reviews. Similar to our idea, Hoffmann and Surdeanu *et al.* [10, 36] utilize external knowledge base to extract relation from text in MIL framework. Different from traditional distant supervision, they assume that if two entities participate in a relation, then at least one sentence that contains these two entities might express that relation. Different from these work, we don't have an external source to determine the involved entities in the events.

5.3 Convolutional Neural Networks

Convolutional Neural Networks (CNN) have found success in several natural language processing (NLP) applications such as part-of-speech tagging, chunking, named entity recognition, and semantic role labeling [6]. Kim [16] applies CNN to text classification for sentiment analysis. Kalchbenner *et al.* [15] propose a CNN approach to model the sentence vector based on dynamic k-max pooling and folding operation. Shen *et al.* [35] propose a latent semantic model based on CNN to learn a distributed representation for search queries. In our work, we use CNN to learn sentence representations by combining both local and global information and couple this representation within a relaxed MIL formulation.

6. CONCLUSION

We propose a novel method to extract event-related sentences from news articles without explicitly provided sentence-level labels for training. Our approach integrates a convolution neural network model into the multi-instance learning framework. The CNN model provides a distributed sentence representation which combines local and global information to relax the independence assumptions of standard MIL formulations. We perform a comprehensive set of experiments to demonstrate the effectiveness of our proposed model in terms of classifying a news document as a

Table 6: Top scored terms in different categories of event populations and event types. All the articles are represented by the MI-CNN model selected key sentences.

EventPopluation					EventType				
Business	**Media**	**Medical**	**Legal**	**Education**	**Housing**	**Energy**	**Economic**	**Employment**	**Government**
sellers	communicators	health	grant	students	housing	water	producers	worker	national
commercial	journalists	medical	congress	education	neighborhood	energy	mobilization	official	march
drivers	express	hospital	judges	national	service	company	route	drivers	government
strike	agreement	unemployment	specialties	government	terms	sector	budget	payment	demand
transport	exhibited	doctor	reprogramming	teachers	family	neighbors	carriers	wages	square
measure	profession	nursing	budget	college	group	lack	association	unemployment	city
carriers	legislation	clinics	explanation	professor	transfers	supply	ministry	guild	front
public	guards	patients	deny	faculty	place	population	cooperators	employee	hours
municipal	intervened	welfare	approve	school	mutual	authority	peasants	company	demonstration
strength	collaboration	power	exist	dean	bill	organization	PLRA	job	students

Table 7: List of events extracted using ExtrHech

Argument 1	**Relation**	**Argument 2**
the retired	require pension	Social Security Institute Servers
the protesters	complain	Guerrero campus
the manifestation	cause trouble	passangers
the district	organize	carnival
the protesters	are required	councilors
Antorcha Campesina organization	agglutinated	the capital
the situation	annoy	producers
the mayor	demand expulsion	colonists
gendarmes	ensure	conflicts

protest or not and extracting the indicative sentences from the article. Using the identified sentences to represent a document, we show strong classification results in comparison to baselines without use of expert-defined dictionaries or features. The strengths of our proposed model is highlighted by integrating with visualization and summarization applications as well as detection of finer patterns that are associated with an event type and population.

Acknowledgments

Supported by the Intelligence Advanced Research Projects Activity (IARPA) via DoI/NBC contract number D12PC000337, the US Government is authorized to reproduce and distribute reprints of this work for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: 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 IARPA, DoI/NBC, or the US Government.

Huzefa Rangwala was partially supported by a George Mason University Provost Grant.

7. REFERENCES

[1] A. Adi, D. Botzer, G. Nechushtai, and G. Sharon. Complex event processing for financial services. In *SCW*, pages 7–12. IEEE, 2006.

[2] S. Andrews, I. Tsochantaridis, and T. Hofmann. Support vector machines for multiple-instance learning. In *NIPS*, pages 561–568. MIT Press, 2003.

[3] F. Bastien, P. Lamblin, R. Pascanu, J. Bergstra, I. J. Goodfellow, A. Bergeron, N. Bouchard, and Y. Bengio. Theano: new features and speed improvements. NIPS, 2012.

[4] N. Chambers. Event schema induction with a probabilistic entity-driven model. In *Proc. EMNLP*, volume 13, pages 1797–1807, 2013.

[5] Y. Chen, L. Xu, K. Liu, D. Zeng, and J. Zhao. Event extraction via dynamic multi-pooling convolutional neural networks. In *Proc. ACL*, volume 1, pages 167–176, 2015.

[6] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa. Natural language processing (almost) from scratch. *The Journal of Machine Learning Research*, 12:2493–2537, 2011.

[7] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artif. Intell.*, 89(1-2):31–71, 1997.

[8] G. Doran and S. Ray. A theoretical and empirical analysis of support vector machine methods for multiple-instance classification. *Machine Learning*, 97(1-2):79–102, 2014.

[9] T. Gartner, P. A. Flach, A. Kowalczyk, and A. J. Smola. Multi-instance kernels. In *Proc. ICML*, pages 179–186, San Francisco, CA, USA, 2002. Morgan Kaufmann Publishers Inc.

[10] R. Hoffmann, C. Zhang, X. Ling, L. Zettlemoyer, and D. S. Weld. Knowledge-based weak supervision for information extraction of overlapping relations. In *Proc. ACL*, pages 541–550. Association for Computational Linguistics, 2011.

[11] W.-T. Hsieh, T. Ku, C.-M. Wu, and S.-c. T. Chou. Social event radar: A bilingual context mining and sentiment analysis summarization system. In *Proc. ACL*, ACL '12, pages 163–168, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.

[12] R. Huang and E. Riloff. Multi-faceted event recognition with bootstrapped dictionaries. In *HLT-NAACL*, pages 41–51, 2013.

[13] S. Jiang, H. Chen, J. F. Nunamaker, and D. Zimbra. Analyzing firm-specific social media and market: A stakeholder-based event analysis framework. *Decision Support Systems*, 67:30–39, 2014.

[14] T. Joachims. *Text categorization with support vector*

machines: Learning with many relevant features. Springer, 1998.

[15] N. Kalchbrenner, E. Grefenstette, and P. Blunsom. A convolutional neural network for modelling sentences. *arXiv preprint arXiv:1404.2188*, 2014.

[16] Y. Kim. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*, 2014.

[17] D. Kotzias, M. Denil, N. de Freitas, and P. Smyth. From group to individual labels using deep features. In *Proc. SIGKDD*, KDD '15, pages 597–606, New York, NY, USA, 2015. ACM.

[18] K. Lee, Y. Artzi, Y. Choi, and L. Zettlemoyer. Event detection and factuality assessment with non-expert supervision. In L. MÃărquez, C. Callison-Burch, J. Su, D. Pighin, and Y. Marton, editors, *Proc. EMNLP*, pages 1643–1648. The Association for Computational Linguistics, 2015.

[19] K. Leetaru and P. A. Schrodt. Gdelt: Global data on events, location, and tone, 1979–2012. In *ISA Annual Convention*, volume 2, page 4, 2013.

[20] Q. Li, H. Ji, and L. Huang. Joint event extraction via structured prediction with global features. In *ACL (1)*, pages 73–82, 2013.

[21] G. Liu, J. Wu, and Z.-H. Zhou. Key instance detection in multi-instance learning. In S. C. H. Hoi and W. L. Buntine, editors, *ACML*, volume 25 of *JMLR Proceedings*, pages 253–268. JMLR.org, 2012.

[22] W. Lu and D. Roth. Automatic event extraction with structured preference modeling. In *Proc. ACL*, ACL '12, pages 835–844, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.

[23] D. McClosky, M. Surdeanu, and C. D. Manning. Event extraction as dependency parsing. In *Proc. ACL*, HLT '11, pages 1626–1635, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[24] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[25] M. Mintz, S. Bills, R. Snow, and D. Jurafsky. Distant supervision for relation extraction without labeled data. In *Proc. ACL*, pages 1003–1011. Association for Computational Linguistics, 2009.

[26] S. Muthiah, B. Huang, J. Arredondo, D. Mares, L. Getoor, G. Katz, and N. Ramakrishnan. Planned protest modeling in news and social media. In *AAAI*, pages 3920–3927, 2015.

[27] K.-H. Nguyen, X. Tannier, O. Ferret, and R. Besançon. Generative event schema induction with entity disambiguation. In *Proc. ACL*, 2015.

[28] T. H. Nguyen and R. Grishman. Event detection and domain adaptation with convolutional neural networks. *Volume 2: Short Papers*, page 365, 2015.

[29] J. Nothman, M. Honnibal, B. Hachey, and J. R. Curran. Event linking: Grounding event reference in a news archive. In *Proc. ACL*, ACL '12, pages 228–232, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.

[30] S. P. O'brien. Crisis early warning and decision support: Contemporary approaches and thoughts on future research. *International Studies Review*, 12(1):87–104, 2010.

[31] N. Ramakrishnan, P. Butler, S. Muthiah, N. Self, R. Khandpur, P. Saraf, W. Wang, J. Cadena, A. Vullikanti, G. Korkmaz, et al. 'beating the news' with embers: Forecasting civil unrest using open source indicators. In *Proc. SIGKDD*, pages 1799–1808. ACM, 2014.

[32] A. Ritter, O. Etzioni, S. Clark, et al. Open domain event extraction from twitter. In *Proc. SIGKDD*, pages 1104–1112. ACM, 2012.

[33] G. Rizzo and R. Troncy. Nerd: a framework for unifying named entity recognition and disambiguation extraction tools. In *Proc. ACL*, pages 73–76. Association for Computational Linguistics, 2012.

[34] P. A. Schrodt, J. Beieler, and M. Idris. ThreeâĂŹsa charm?: Open event data coding with el: Diablo, petrarch, and the open event data alliance. In *ISA Annual Convention*, 2014.

[35] Y. Shen, X. He, J. Gao, L. Deng, and G. Mesnil. Learning semantic representations using convolutional neural networks for web search. In *Proc. WWW*, pages 373–374. International World Wide Web Conferences Steering Committee, 2014.

[36] M. Surdeanu, J. Tibshirani, R. Nallapati, and C. D. Manning. Multi-instance multi-label learning for relation extraction. In *Proc. EMNLP*, pages 455–465. Association for Computational Linguistics, 2012.

[37] S. Tong and D. Koller. Support vector machine active learning with applications to text classification. *The Journal of Machine Learning Research*, 2:45–66, 2002.

[38] D. Z. Wang, Y. Chen, S. Goldberg, C. Grant, and K. Li. Automatic knowledge base construction using probabilistic extraction, deductive reasoning, and human feedback. In *Proc. ACL*, pages 106–110. Association for Computational Linguistics, 2012.

[39] N. Weidmann, E. Frank, and B. Pfahringer. A two-level learning method for generalized multi-instance problems. In *Proc. EMNLP*, pages 468–479, 2003.

[40] J. Weston, A. Bordes, S. Chopra, and T. Mikolov. Towards ai-complete question answering: A set of prerequisite toy tasks. *arXiv preprint arXiv:1502.05698*, 2015.

[41] D. Wurzer, V. Lavrenko, and M. Osborne. Twitter-scale new event detection via k-term hashing. In L. MÃărquez, C. Callison-Burch, J. Su, D. Pighin, and Y. Marton, editors, *Proc. EMNLP*, pages 2584–2589. The Association for Computational Linguistics, 2015.

[42] M. Yu, M. R. Gormley, and M. Dredze. Combining word embeddings and feature embeddings for fine-grained relation extraction. In *Proc. NAACL*, 2015.

[43] M. D. Zeiler. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*, 2012.

[44] A. Zhila and A. F. Gelbukh. Open information extraction for spanish language based on syntactic constraints. In *ACL (Student Research Workshop)*, pages 78–85, 2014.