Event Detection using Hierarchical Multi-Aspect Attention

Sneha Mehta
Virginia Tech, USA
snehamehta@cs.vt.edu

Mohammad Raihanul Islam
Virginia Tech, USA
raihan8@cs.vt.edu

Huzefa Rangwala
George Mason University, USA
rangwala@cs.gmu.edu

Naren Ramakrishnan
Virginia Tech, USA
naren@cs.vt.edu

ABSTRACT

Classical event encoding and extraction methods rely on fixed dictionaries of keywords and templates or require ground truth labels for phrase/sentences. This hinders widespread application of information encoding approaches to large-scale free form (unstructured) text available on the web. Event encoding can be viewed as a hierarchical task where the coarser level task is event detection, i.e., identification of documents containing a specific event, and where the fine-grained task is one of event encoding, i.e., identifying key phrases, key sentences. Hierarchical models with attention seem like a natural choice for this problem, given their ability to differentially attend to more or less important features when constructing document representations. In this work we present a novel factorized bilinear multi-aspect attention mechanism (FBMA) that attends to different aspects of text while constructing its representation. We find that our approach outperforms state-of-the-art baselines for detecting civil unrest, military action, and non-state actor events from corpora in two different languages.

CCS CONCEPTS

• **Computing methodologies → Artificial intelligence**; **Natural language processing**; **Supervised learning**; **Neural networks**.

KEYWORDS

Neural Networks, Hierarchical Attention, Multi-Aspect Attention, Event Encoding

ACM Reference Format:
Sneha Mehta, Mohammad Raihanul Islam, Huzefa Rangwala, and Naren Ramakrishnan. 2019. Event Detection using Hierarchical Multi-Aspect Attention. In *Proceedings of the 2019 World Wide Web Conference (WWW '19), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3308558.3313659

1 INTRODUCTION

Given the large volumes of text available on the web in the form of news, social media, blogs and discussion forums, it is crucial to identify and extract meaningful nuggets of information. A wide range of applications from question answering [24], knowledge base construction [21] and named entity recognition [19] to informing critical decisions in domains ranging from national security to cyber security [22] rely on the event extraction and encoding process. Event analysis refers to the extraction of specific information about certain events from text. It can be categorized as a hierarchical task where the coarser level task is event detection, i.e., identification of documents containing a specific event and the fine-grained task is one of event encoding, i.e., identifying key phrases and sentences describing event related information such as the type of the event, type of people involved in the event, and relationships between these aspects. The sheer diversity of applicable event domains and types combined with the multitude of data sources, and scarcity of fine-level labels make both these tasks challenging.

Prior work in event encoding has focused on extracting entities, detecting trigger terms, and matching slots on predefined templates [3, 16, 26]. However there are a few shortcomings of these approaches. The first drawback is that they rely on fine-grained labeled training data which is hard to obtain for a variety of domains and different types of events. On the other hand, labels at the document level are easier to obtain. Second, they use of sentence-level embedding [16] removes contextual information resulting in false negatives because event occurrences do not neatly partition into unique sentences. Multiple instance learning (MIL) approaches have been proposed [9, 23] as a solution to partly alleviate these problems. MIL approaches view documents as bags of sentences, make predictions at the sentence level, aggregate sentence level probabilities for a select few sentences to obtain document-level predictions. In this work we model the tasks of event detection and key sentence identification from a news article in a unified framework without explicit labels at the sentence level by leveraging the implicit hierarchy in textual corpora. Hierarchical attention based networks seem like a natural choice for this problem, given their ability to differentially attend to more or less important words and sentences when constructing document representations [25].

We use a recurrent neural network (RNN) based hierarchical model (with attention mechanisms at both levels in the hierarchy) that constructs sentence and document representations. The sentence representation is constructed by attending to all words in the sentence, whereas the document representation is constructed by attending to all sentences. To capture the fact that a sentence might capture multiple aspects related to an event (e.g., cause, location, population) we adapt hierarchical attention network models [25] by using a novel multi-aspect attention mechanism that allows for multiple attention distributions over words for a single sentence where each distribution can be thought of as attending to a different aspect of a sentence.

WWW '19, May 13–17, 2019, San Francisco, CA, USA

ACM ISBN 978-1-4503-6674-8/19/05.
https://doi.org/10.1145/3308558.3313659

We evaluate the proposed hierarchical attention models for the task of event detection on civil unrest (CU) and military action/non-state actor (MANSA) datasets.

Our contributions in this work can be succinctly summarized as follows:

- We present a novel factorized bilinear multi-aspect attention mechanism (FBMA) that constructs a sentence representation using multiple attention distributions and that when used with hierarchical models improves performance for event detection.
- Our FBMA approach achieves state-of-the-art results for event detection on three event datasets from two different domains in two different languages.

2 RELATED WORK

2.1 Event Extraction

In event extraction supervised approaches usually rely on manually labeled training datasets and handcrafted ontologies. Li et al. [10] utilize annotated arguments and specific keyword triggers in text to develop an extractor. Supervised approaches have also been studied using dependency parsing by analyzing the event-argument relations and discourse of event interactions [14]. These approaches are usually limited by the availability of fine-grained labeled data and require elaborately designed features. In contrast to these approaches our method uses attention mechanisms to implicitly weigh words and sentences and is able to extract event extents and trigger words with labels provided only at the document level. This formulation is suitable because labels at document level are easier to obtain than at the per-sentence level or at the word level. This makes the task of event extraction also amenable to multiple instance learning (MIL) [5] solutions. In MIL 'bags' are groups of 'instances' which are to be classified. In a 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. Kotzias et al. [9] focus on instance-level predictions from group level labels and allow for the application of general aggregation functions for sentiment classification. Wang et al. [23] use a similar idea and hold the previous state-of-the-art results on one of the datasets we evaluate on. Contrary to these approaches our method is hierarchical and computes the feature representation for the next level in the hierarchy using a weighted average of feature representations in the current layer.

2.2 Attention for Structured Representations

Models have been proposed that compute multiple attention distributions over a single sequence of words. The multi-view networks proposed by Guo et al. [6] use a different set of parameters for each view which leads to a large increase in the number of parameters with increasing number of views. Lin et al. [12] alleviate this problem by producing a matrix embedding from a single set of parameters. Both these methods use a special case of additive attention proposed by Bahdanau et al. [1] in the context of neural machine translation. Luong et al. [13] simplify additive attention operations by introducing the notion of multiplicative attention which is faster to compute. In multiplicative attention, the score between two feature vectors is learned using a bilinear projection matrix. Dot product attention [13] is a special case of multiplicative attention where the score between two features vectors is computed by a simple dot product between them. Yang et al. [25] use dot product attention to compute the similarity of word hidden representation to a word-level context vector which is learned with the rest of the model. In our work we compute the score between the context vector and the word hidden representation using a bilinear projection matrix and then we use an approach inspired by multi-modal low rank bilinear pooling proposed by Kim et al. [8] to factorize the matrix into two low rank matrices to compute multiple attention distributions over words. Contrary to Guo et al. [6] we use matrix factorization to alleviate the problem of increasing parameters with increasing views and our approach uses fewer parameters than Lin et al. to compute multiple attention calculations and performs superior to their approach. We refer to this as multi-aspect attention as it attends to different aspects or parts of a sentence for constructing a sentence embedding.

3 PROPOSED MODEL

The event detection problem can be defined as follows: Given a corpus containing N news articles $\{x_1, x_2, ..., x_N\}$, each article is associated with an event label $y \in \{0, 1\}$, with 1 corresponding to articles containing an event. For each news article we aim to predict its label, indicating if it contains an event or not.

3.1 Sequence Encoder

Consider a news article containing n sentences with each sentence containing T words. A sentence consists of word tokens w_{it}, $t \in [0, T]$ where every word is converted to a real valued word vector $\mathbf{x}_{it}$ using the pre-trained embedding matrix $\mathbf{W}_e = R^{dx|V|}$, $\mathbf{x}_{it} = \mathbf{W}_e\mathbf{w}_{it}$, $t \in [1, T]$ where d is the embedding dimension and V is the vocabulary. We encode the sentence using a bi-directional Gated Recurrent Unit [4](bi-GRU) RNN that summarizes information in both directions along the sentence to get a contextual annotation of a word. In a bi-GRU the hidden state at time step t is represented as a concatenation of hidden states in the forward and backward direction. The forward GRU denoted by $\overrightarrow{GRU}$ processes the sentence from w_{i1} to w_{iT} whereas the backward GRU denoted by $\overleftarrow{GRU}$ processes it from w_{iT} to w_{i1}.

$$\mathbf{x}_{it} = \mathbf{W}_e\mathbf{w}_{it} \tag{1}$$

$$\overrightarrow{\mathbf{h}_{it}} = \overrightarrow{GRU}(\mathbf{x}_{it}, \mathbf{h}_{i(t-1)}, \boldsymbol{\theta}) \tag{2a}$$

$$\overleftarrow{\mathbf{h}_{it}} = \overleftarrow{GRU}(\mathbf{x}_{it}, \mathbf{h}_{i(t+1)}, \boldsymbol{\theta}) \tag{2b}$$

Here the word annotation $\mathbf{h}_{it}$ is obtained by concatenating the forward hidden state $\overrightarrow{\mathbf{h}_{it}}$ and the backward hidden state $\overleftarrow{\mathbf{h}_{it}}$.

3.2 Word-Level Attention

For event extraction the presence of certain words increases the probability of a sentence containing the event. Such words should be given higher weight than other words while computing a sentence representation. Since an event related trigger word can occur anywhere in a sentence we choose the global attention mechanism [13] in which the sentence representation is computed by attending to all words in the sentence.

3.2.1 Bilinear Attention. Let $\mathbf{h}_{it}$ be the annotation corresponding to the word $\mathbf{x}_{it}$. First we transform $\mathbf{h}_{it}$ using a one layer Multi-Layer Perceptron (MLP) to obtain its hidden representation $\mathbf{u}_{it}$.

$$\mathbf{u}_{it} = tanh(\mathbf{W}_w \mathbf{h}_{it} + \mathbf{b}_w) \tag{3}$$

We measure the importance of word by computing an alignment score of $\mathbf{u}_{it}$ to a word level context vector $\mathbf{u}_w$ using a bilinear model:

$$f_{it} = \mathbf{u}_w^T \mathbf{W}_i \mathbf{u}_{it} \tag{4}$$

Here, $\mathbf{W}_i$ is a bilinear projection matrix, $\mathbf{u}_w$ is randomly initialized and jointly learned with other parameters during training. Similar to [25], $\mathbf{u}_w$ can be seen as a high dimensional representation of the fixed query 'What is the informative word'. $\mathbf{u}_w \in \mathbb{R}^l$, $\mathbf{u}_{it} \in \mathbb{R}^{2h}$ and $\mathbf{W}_i \in \mathbb{R}^{2hx2h}$. The attention weight for the word $\mathbf{x}_{it}$ can be computed through a softmax function.

$$\alpha_{it} = \frac{exp(f_{it})}{\sum_{t'} exp(f_{it'})} \tag{5}$$

3.2.2 Factorized Bilinear Multi-Aspect Attention. The attention distribution above usually focuses on a specific component of the sentence, like a special set of trigger words or phrases. So it is expected to reflect an aspect, or component of the semantics in a sentence. However there can be multiple aspects that describe an event like who were involved in the event, what were the causes of the event or where did the event occur. For this we introduce the novel factorized bilinear multi-aspect attention (FBMA) mechanism. Suppose m aspects are to be extracted from a sentence, we need m alignment scores between each word hidden representation $\mathbf{u}_{it}$ and the context vector $\mathbf{u}_w$. To obtain an m dimensional output $\mathbf{f}_{it}$, we need to learn $\mathbf{W} = [\mathbf{W}_1, ..., \mathbf{W}_m] \in \mathbf{R}^{lx2hxm}$ as demonstrated in previous works. Although bilinear model might be effective in capturing pairwise interaction it introduces a huge number of parameters that may lead to a high computational cost. Inspired by multi-modal low rank bilinear pooling approach proposed by Kim et. al [8] and the matrix factorization approaches proposed in [11, 18] the bilinear projection matrix $\mathbf{W}_i$ can be factorized into two rank 1 matrices $\mathbf{P}$ & $\mathbf{Q}$. Eq.4 can be written as:

$$\mathbf{f}_{it} = \mathbf{u}_w^T \mathbf{W}_i \mathbf{u}_{it} = \mathbf{P}^T \mathbf{u}_w \circ \mathbf{Q}^T \mathbf{u}_{it} = \tilde{\mathbf{u}}_w \circ \tilde{\mathbf{u}}_{it} \tag{6}$$

Here $\mathbf{P} \in \mathbb{R}^{lxm}$ and $\mathbf{Q} \in \mathbb{R}^{2hxm}$ are two rank 1 matrices, m is the number of aspects to extract and $\circ$ is the Hadamard product or elementwise multiplication. This brings the two feature vectors $\mathbf{u}_{it} \in \mathbf{R}^{2h}$, the word hidden representation and $\mathbf{u}_w \in \mathbf{R}^l$, the word level context vector in the common space and are given by $\tilde{\mathbf{u}}_{it}$ and $\tilde{\mathbf{u}}_w$ respectively. $\mathbf{f}_{it} \in \mathbb{R}^m$ now is a multi-aspect alignment vector for the word $\mathbf{x}_{it}$. The multi-aspect attention vector $\boldsymbol{\alpha}_{it} \in \mathbb{R}^m$ is obtained by computing a softmax function along the sentence length:

$$\boldsymbol{\alpha}_{it} = \frac{exp(\mathbf{f}_{it})}{\sum_{t'} exp(\mathbf{f}_{it'})} \tag{7}$$

Before computing softmax, similar to [8] we apply an additional tanh nonlinearlity to $\mathbf{f}_{it}$. Since elementwise multiplication is introduced the values of neurons may vary a lot so we apply an l_2 normalization layer across the m dimension, ($\mathbf{f}_{it} \leftarrow \frac{\mathbf{f}_{it}}{||\mathbf{f}_{it}||}$) after the Hadamard product.

3.2.3 Sentence Representation. Let $\mathbf{H}_i = (\mathbf{h}_{i1}, \mathbf{h}_{i2}, ...\mathbf{h}_{iT})$ be a matrix of all word annotations in a sentence; $\mathbf{H}_i \in \mathbb{R}^{Tx2h}$. Let $\mathbf{A}_i = (\boldsymbol{\alpha}_{i1}, \boldsymbol{\alpha}_{i2}, ...\boldsymbol{\alpha}_{iT})$ be the multi-aspect attention matrix for the sentence; $\mathbf{A}_i \in \mathbb{R}^{mxT}$. The sentence representation for an aspect j given by $\boldsymbol{\alpha}_{ij} = \{\alpha_{j1}, \alpha_{j2}, ..\alpha_{jT}\}$ can be computed by taking a weighted sum of all word annotations.

$$\mathbf{s}_{ij} = \sum_{k=1}^{T} \mathbf{h}_{ik} * \alpha_{jk} \tag{8}$$

Similarly, sentence representation can be computed for all aspects and is given in a compact form by:

$$\mathbf{S}_i = \mathbf{A}_i \mathbf{H}_i \tag{9}$$

Here $\mathbf{S}_i \in \mathbb{R}^{mx2h}$ is a matrix sentence embedding and contains as many rows as the number of aspects. Each row contains an attention distribution for a new aspect. It can be flattened by concatenating all rows for further processing.

3.3 Sentence-Level Attention

News articles consist of many sentences with a few of them describing the event and the rest describing the supporting facts. The ones containing event related information should be assigned higher weights. Instead of hard selecting top K sentences and aggregating their probabilities as in prior MIL approaches such as [23] we use the global attention mechanism over the sentence annotations to get the document representation. Specifically, given a document containing sentence embeddings $\{\mathbf{s}_1, ..., \mathbf{s}_i, ..., \mathbf{s}_n\}$ where each $\mathbf{s}_i$ is a flattened representation of the matrix sentence representation $\mathbf{S}_i$ as given by eq. 9 we get the document vector as follows.

$$\overrightarrow{\mathbf{h}_i} = \overrightarrow{GRU}(\mathbf{s}_i, \mathbf{h}_{i-1}, \boldsymbol{\theta}) \tag{10a}$$

$$\overleftarrow{\mathbf{h}_i} = \overleftarrow{GRU}(\mathbf{s}_i, \mathbf{h}_{i+1}, \boldsymbol{\theta}) \tag{10b}$$

$$\mathbf{h}_i = \{\overrightarrow{\mathbf{h}_i}, \overleftarrow{\mathbf{h}_i}\} \tag{10c}$$

The sentence annotation $\mathbf{h}_i$ is obtained by concatenating the forward and backward hidden representations of the bi-GRU. Document representation is obtained by attention over sentences.

$$\mathbf{u}_i = tanh(\mathbf{W}_s \mathbf{h}_i + \mathbf{b}_s), \tag{11a}$$

$$\alpha_i = \frac{exp(\mathbf{u}_i^T \mathbf{u}_s)}{\sum_t' exp(\mathbf{u}_{t'}^T \mathbf{u}_s)} \tag{11b}$$

$$\mathbf{d} = \sum_i \alpha_i \mathbf{h}_i \tag{11c}$$

Here $\mathbf{u}_s$, the sentence level context vector, is randomly initialized and learned along with other model parameters while training and $\mathbf{d}$ is the document representation that summarizes all the information

Table 1: **First row indicates population classes participating in a protest in the CU dataset. Second row indicates the causes of protest.**

Event Population	General Population, Business,Legal, Labor, Agricultural, Education, Medical, Media
Event Type	Government Policies, Employment and Wages, Energy and Resources, Economic Policies, Housing

in the article. Given the document representation, we use a two hidden layer MLP with dropout to get the class scores.

$$\hat{y} = \mathbf{W}_c \mathbf{d} + \mathbf{b}_c. \tag{12}$$

Loss for the document is computed using the standard cross entropy.

$$l = -(y \log(\hat{y}) + (1 - y) \log(1 - \hat{y})) \tag{13}$$

The sentence embedding given by eq. 9 can suffer from redundancy issues if the attention mechanism always provides similar weights for all the m aspects. To attend to a small set of trigger words in each aspect and to encourage diversity in different aspects we use penalization as described by Lin et al. [12] that is added to the loss:

$$P = \frac{\sum_i \left\| \mathbf{A}_i \mathbf{A}_i^T - \mathbf{I} \right\|_F^2}{n} \tag{14}$$

Where $\|\bullet\|_F$ stands for the Frobenius norm of a matrix and the summation is taken over all sentences in the document. The final training loss is given by:

$$L = \sum_d l + \lambda P \tag{15}$$

The summation is taken over all the documents in the batch and λ is a hyperparameter. We use the mini-batch stochastic gradient descent algorithm [7] with momentum and weight decay for optimizing the loss function and the backpropogation algorithm is used to compute the gradients.

3.3.1 Hyperparameters. We use a word embedding size of 100. The embedding matrix $\mathbf{W}_e$ is pretrained on the corpus using the gensim [1] implementation of the widely used distributed representations model word2vec [15]. All words appearing less than 5 times are discarded. The GRU hidden state is set to $h = 50$. In FBMA the dimension of $\mathbf{u}_{it}$ is given by the dimension of the GRU hidden state, but the dimension of $\mathbf{u}_w$ can be tuned. Empirically we find that setting the dimension of $\mathbf{u}_w$ to 32 gives us the best results. We set the classifier MLP hidden state to 512 and apply a 0.4 dropout to the hidden layer. We use a batch size of 64 for training and an initial learning rate of 0.05. For early stopping we use $patience = 5$.

4 EXPERIMENTS

4.1 Datasets

To evaluate our approach we use three event datasets - (i) the Civil Unrest Gold Standard Report labeled manually by analysts from MITRE corporation [17]. It contains encodings of civil unrest events from 10 Latin American countries in Argentina, Brazil, Chile, Colombia, Ecuador, El Salvador, Mexico, Paraguay, Uruguay, and Venezuela. The encodings are obtained from major national newspapers as identified by 4imn.com. It contains a total of 24,110 news articles out of which 18% mention a protest event while the rest are non-protest articles. All the articles are in Spanish and we refer to this dataset as CU (Spanish) in this paper. (ii) The AutoGSR dataset – This dataset comes from the EMBERS AutoGSR system [20] which is a web based system that generates validated civil unrest events extracted from news articles. It contains a total of 32,019 news articles out of which 18% describe protest events (protest articles) and the rest do not describe any protest events (non-protest articles). We refer to this dataset as CU (English) in the paper. For each protest article, the CU English & Spanish datasets contain the population type and protest event type. The population type indicates which class of population are involved in the protest and the event type indicates the main reason behind the protest. These are listed in table 1.

Table 2: **Dataset statistics. Total number of news articles, average number of sentences per article and average number of words per article in the datasets.**

Datset	**# articles**	**# sents**	**# words**
MANSA	105858	6.3	250.3
CU Spanish	24110	11.4	337.1
CU English	32019	10.1	358.2

Table 3: **Various baselines and their key characteristics. The proposed FBMA model is hierarchical, with bilinear multi-aspect attention.**

Feature Method	**hierarchical**	**key words**	**key sents**	**bilinear attention**	**multi-aspect attention**
MICNN	✓	×	✓	×	×
MIGCNN	✓	✓	✓	×	×
HAN	✓	✓	✓	×	×
BSA	✓	✓	✓	✓	×
HSA	✓	✓	✓	×	✓
FBMA	✓	✓	✓	✓	✓

Finally, we evaluate on iii) the Military Action and Non-State Actor (MANSA) GSR dataset which is in English and Arabic. This contains event encodings from gulf countries namely, Bahrain, Egpyt, Iraq, Jordan, Lebanon, Qatar, Saudi Arabia and Syria. The event types include 'Military Actions' (MA) which are actions by military, police, or security organization and 'Non-State Actor'(NSA) which are actions initiated by non-governmental groups or individuals to further political, social, religious or ideological objectives. These events are encoded from news articles collected from the web and print media. Event collection techniques include Google Advanced Search (limited to the newspaper website), Nexis queries, and IHS Janes. Google Advanced Search is used to collect events in online media. Nexis and IHS Janes are used to collect events in print media. About 34% articles describe an event and rest are non-event articles. We refer to this dataset as MANSA dataset in the paper. MA & NSA events are further divided into subtypes. In this work we combine NSA & MA events together for detection. Please refer to table 2 for the overview of our datasets.

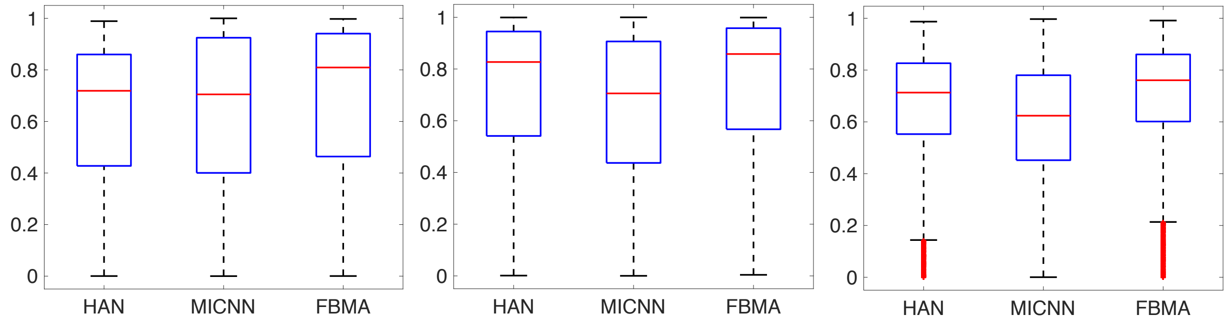
4.2 Comparative Methods

Table 3 shows the different approaches that were evaluated in this study along with their key characteristics. CNNs within a multiple instance learning (MIL) framework have been used by Wang et al. [23]. We consider the MI-CNN model proposed by them and its

[1] https://radimrehurek.com/gensim/

Table 4: Results of Event Detection. FBMA refers to the proposed Factorized Bilinear Multi-aspect Attention mechanism. BSA refers to the Bilinear Single-aspect Attention mechanism presented in eq. 4, MI-GCNN refers to our MIL model, where the CNN encoder is replaced by the RNN encoder followed by simple dot product attention to extract key words. HAN, HSA & MI-CNN are other baselines.

Dataset	**Method**	**Precision** (std.)	**Recall** (std.)	**F1** (std.)
MANSA	MI-CNN **(Wang et al.)**	0.731 (0.012)	0.785 (0.004)	0.686 (0.021)
	MI-GCNN (This paper)	**0.793 (0.003)**	0.622 (0.013)	0.697 (0.007)
	HSA **(Lin et al.)**	0.733 (0.001)	0.823 (0.009)	0.775 (0.003)
	HAN **(Yang et al.)**	0.740 (0.007)	0.831(0.007)	0.783 (0.004)
	BSA (This paper)	0.737 (0.007)	**0.834 (0.003)**	0.782 (0.003)
	FBMA (This paper)	0.747(0.003)	0.831 (0.003)	**0.787(0.002)**
CU (Spanish)	MI-CNN **(Wang et al.)**	0.742 (0.036)	**0.813 (0.041)**	0.775 (0.006)
	MI-GCNN (This paper)	**0.834 (0.011)**	0.721 (0.009)	0.773 (0.005)
	HSA **(Lin et al.)**	0.763 (0.009)	0.745 (0.016)	0.754 (0.011)
	HAN **(Yang et al.)**	0.811 (0.011)	0.775 (0.011)	0.793 (0.005)
	BSA (This paper)	0.812 (0.015)	0.779 (0.011)	0.795 (0.007)
	FBMA (This paper)	0.816 (0.017)	0.784 (0.010)	**0.800 (0.008)**
CU (English)	MI-CNN **(Wang et al.)**	**0.824 (0.01)**	0.644(0.009)	0.723 (0.009)
	MI-GCNN (This paper)	0.815 (0.006)	0.68 (0.017)	0.742 (0.011)
	HSA **(Lin et al.)**	0.746 (0.008)	0.710(0.017)	0.727 (0.010)
	HAN **(Yang et al.)**	0.779 (0.012)	0.746 (0.024)	0.762 (0.015)
	BSA (This paper)	0.786 (0.008)	**0.757 (0.007)**	**0.771 (0.006)**
	FBMA (This paper)	0.785 (0.006)	0.745 (0.007)	0.764 (0.005)
NB	HAN **(Yang et al.)**	0.779 (0.012)	0.746 (0.024)	0.762 (0.015)
	BSA (This paper)	0.786 (0.008)	**0.757 (0.007)**	**0.771 (0.006)**
	FBMA (This paper)	0.785 (0.006)	0.745 (0.007)	0.764 (0.005)



(a) English

(b) Spanish

(c) MANSA

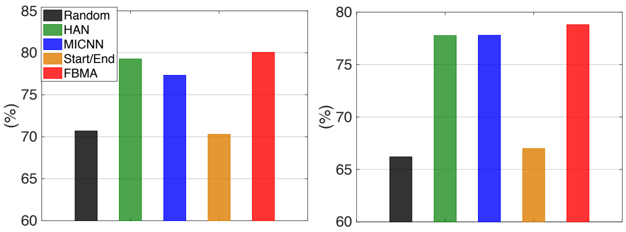
Figure 1: Comparison of event probabilities assigned by FBMA, MICNN and HAN on the CU English, Spanish and MANSA datasets. We can clearly see that mean probability is greater for FBMA in all the datasets for the event class. This depicts that FBMA is usually more confident than other methods in classifying the event articles. We also observe MANSA dataset contains some outliers (shown in red dots at the bottom).

variants as our baselines. The MI-CNN approach first constructs a sentence vector by applying convolution in the temporal dimension followed by $k - maxpooling$. A document vector is formed from sentence vectors in a similar way. Instance representation for each sentence is constructed by concatenating sentence and document representations. Finally probabilities at instance level are aggregated to compute the document level probability. In our first baseline we replace the convolutional sentence encoder by a GRU sentence encoder followed by a simple attention mechanism that attends to words while constructing a sentence representation. We refer to this model as MI-GCNN and it extracts both trigger words and key sentences.

Since, the proposed model is a multi-aspect attention mechanism, we compare it with another multi-aspect attention mechanism proposed by Lin et al. [12]. We refer to as Hierarchical Self Attention (HSA) where we replace the word-level attention mechanism by the Self-Attentive mechanism proposed by them.

We also evaluated the Hierarchical Attention Network (HAN) model proposed by Yang et. al. [25], which attends to both trigger words and keys sentences while constructing sentence and document representations respectively but it only consists of a single aspect attention mechanism at both levels.

Finally, we replace the word level attention in the HAN model with the bilinear attention mechanism given by eq. 4 which has shown promising results in question answering tasks [2, 8].We refer to



(a) Population

(b) Event

Figure 2: Event Extraction accuracy for event type detection and population class detection from key sentences extracted by different models. FBMA outperforms all the methods.

this model as Bilinear Single Aspect Attention model (BSA). We compare the performance of these models with the proposed FBMA model. In all the datasets, the event class is a rare class and hence we report the precision, recall and F1 score of that class for the test set. For each dataset we trained our model using 5-fold cross-validation with an 80/20 train/test split and employed early stopping. Models took less than an hour on a single Tesla P100 GPU to train.

5 RESULTS

5.1 Event Detection

Table 4 reports the performance of our models compared to the baseline approaches. On average, the FBMA model outperforms the MIL-based MICNN and MIGCNN models by 14.2 % & 12.9% on MANSA, 32.2 % & 3.5 % on CU Spanish and 6.6% & 3.9% on CU English datasets respectively. The FBMA model also outperforms HAN across the three datasets. Moreover, unlike HAN our model attends to different aspects while constructing a sequence representation, which results in an increased model size due to added parameters. One key aspect of our model is the ability to tune the dimensionality of the word level context vector $\mathbf{u}_w$. We find that models with $l \leq 2h$ where l is the dimension of $\mathbf{u}_w$ and $2h$ is the dimension of the word hidden representation tend to outperform models with $l > 2h$. The FBMA model is forced to learn a more compact representation of the word-level context vector and thus, retains only the most relevant information.

FBMA also beats HSA on CU English, CU Spanish & MANSA datasets by 5.1 %, 6.1 % and 1.5% respectively. Moreover our attention mechanism uses fewer parameters than the Self-Attentive model proposed by Lin et al. [12].

We also observe that the simple bilinear attention models outperform the dot product based attention models in HAN. For the CU English dataset, the BSA approach outperforms HAN & MIL-based baselines by 1.2% and 6.4% & 3.9% respectively. For the CU English and CU Spanish datasets we set $l = 2h$, whereas for MANSA dataset we set $l = 64$. These values were empirically found to perform best on the corresponding datasets.

5.2 Event Probabilities

For event encoding it is important that models have a high precision otherwise it may lead to false-positives and incorrect representations. This is also an issue with traditional detection approaches that detect events based on occurrence of certain set of trigger words from a pre-curated list, without taking into account the sentence context or relationship between different entities. Hence, it is important for the models to assign a high confidence to positive articles and a low confidence to negative articles. We present a distribution of event probabilities assigned to the positive articles by the FBMA, HAN and MICNN models for the test sets for all the three event datasets in Figure 1. We observe that generally the average probability assigned by FBMA is higher than MICNN and HAN confirming our hypothesis that having multi-aspect attention for each sentence increases the model confidence for a positive article.

6 CONCLUSION

In this paper we presented a framework for event detection. We used hierarchical attention based models for the task because of their ability to attend to words and sentences while constructing sentence and document representations respectively. With the hierarchical models we used our proposed FBMA mechanism which computes multiple attention distributions over words which leads to contextual sentence and document representations. Our results showed that this mechanism performed better than several other approaches and especially single-aspect mechanisms that miss out on the context because there is only one attention distribution. The proposed attention mechanism could easily be used for other tasks where document context is important. Moreover, it uses less number of parameters than other similar approaches, and hence can be scaled for larger datasets. An important open question to investigate in the future is that if the attention weights in each aspect can be constrained using certain rules to capture co-occurrence patterns of event arguments. This will lead to a wider distinction between different aspects captured and more precise event extraction.

7 ACKNOWLEDGEMENTS

This research is based upon work supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via contract 2018-18030100003 and by the U.S. Department of Homeland Security under grant award number 2017-ST-061-CINA01. 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 ODNI, IARPA, US Department of Homeland Security, 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

[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural Machine Translation by Jointly Learning to Align and Translate. *CoRR* abs/1409.0473 (2014).

[2] Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading wikipedia to answer open-domain questions. *arXiv preprint arXiv:1704.00051* (2017).

[3] Yubo Chen, Liheng Xu, Kang Liu, Daojian Zeng, and Jun Zhao. 2015. Event extraction via dynamic multi-pooling convolutional neural networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*, Vol. 1. 167–176.

[4] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555* (2014).

[5] Thomas G Dietterich, Richard H Lathrop, and Tomás Lozano-Pérez. 1997. Solving the multiple instance problem with axis-parallel rectangles. *Artificial intelligence* 89, 1-2 (1997), 31–71.

[6] Hongyu Guo, Colin Cherry, and Jiang Su. 2017. End-to-end multi-view networks for text classification. *arXiv preprint arXiv:1704.05907* (2017).

[7] Jack Kiefer, Jacob Wolfowitz, et al. 1952. Stochastic estimation of the maximum of a regression function. *The Annals of Mathematical Statistics* 23, 3 (1952), 462–466.

[8] Jin-Hwa Kim, Kyoung-Woon On, Woosang Lim, Jeonghee Kim, Jung-Woo Ha, and Byoung-Tak Zhang. 2016. Hadamard product for low-rank bilinear pooling. *arXiv preprint arXiv:1610.04325* (2016).

[9] Dimitrios Kotzias, Misha Denil, Nando De Freitas, and Padhraic Smyth. 2015. From group to individual labels using deep features. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 597–606.

[10] Qi Li, Heng Ji, and Liang Huang. 2013. Joint event extraction via structured prediction with global features. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vol. 1. 73–82.

[11] Yanghao Li, Naiyan Wang, Jiaying Liu, and Xiaodi Hou. [n. d.]. Factorized bilinear models for image recognition. ([n. d.]).

[12] Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. 2017. A structured self-attentive sentence embedding. *arXiv preprint arXiv:1703.03130* (2017).

[13] Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025* (2015).

[14] David McClosky, Mihai Surdeanu, and Christopher D Manning. 2011. Event extraction as dependency parsing. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*. 1626–1635.

[15] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. 3111–3119.

[16] Thien Huu Nguyen and Ralph Grishman. 2015. Event detection and domain adaptation with convolutional neural networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*, Vol. 2. 365–371.

[17] Naren Ramakrishnan, Patrick Butler, Sathappan Muthiah, Nathan Self, Rupinder Khandpur, Parang Saraf, Wei Wang, Jose Cadena, Anil Vullikanti, Gizem Korkmaz, et al. 2014. 'Beating the news' with EMBERS: forecasting civil unrest using open source indicators. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 1799–1808.

[18] Steffen Rendle. 2010. Factorization machines. In *Data Mining (ICDM), 2010 IEEE 10th International Conference on*. IEEE, 995–1000.

[19] Giuseppe Rizzo and Raphaël Troncy. 2012. NERD: A Framework for Unifying Named Entity Recognition and Disambiguation Extraction Tools. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*. 73–76.

[20] Parang Saraf and Naren Ramakrishnan. 2016. EMBERS AutoGSR: Automated Coding of Civil Unrest Events. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 599–608.

[21] Daisy Zhe Wang, Yang Chen, Sean Goldberg, Christan Grant, and Kun Li. 2012. Automatic Knowledge Base Construction Using Probabilistic Extraction, Deductive Reasoning, and Human Feedback. In *Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction*. 106–110.

[22] Wei Wang. 2018. *Event Detection and Extraction from News Articles*. Ph.D. Dissertation. Virginia Tech.

[23] Wei Wang, Yue Ning, Huzefa Rangwala, and Naren Ramakrishnan. 2016. A Multiple Instance Learning Framework for Identifying Key Sentences and Detecting Events. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*. 509–518.

[24] Jason Weston, Antoine Bordes, Sumit Chopra, and Tomas Mikolov. 2015. Towards AI-Complete Question Answering: A Set of Prerequisite Toy Tasks. *CoRR* abs/1502.05698 (2015).

[25] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 1480–1489.

[26] Mo Yu, Matthew R Gormley, and Mark Dredze. 2015. Combining word embeddings and feature embeddings for fine-grained relation extraction. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 1374–1379.