Interpretable Event Detection and Extraction using Multi-Aspect Attention

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

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

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

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

ABSTRACT

Classical encoding and extraction methods rely on fixed dictionary of keywords and templates or require ground truth labels for phrase/sentences. This prevents wide spread use of these information encoding approaches to large-scale free form (unstructured) text available on the web. Event encoding can be categorized as a hierarchical task where the coarser level task is event detection – identification of documents containing a specific event and the fine-grained task of event encoding – 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 the document representation. 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 use this mechanism within a hierarchical framework to extract event references or extents i.e. event-related sentences and trigger words from the text. We find that our approach outperforms state-of-the-art baselines for event detection on the Civil Unrest and Military Action and Non-State Actor datasets in two different languages. We further provide qualitative examples of the extracted event extents and trigger words giving a peek into what our model learns.

ACM Reference Format:

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 [35], knowledge base construction [32] and named entity recognition [27] to informing critical decisions in domains ranging from national security to cyber security [33] 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 – identification of documents containing a specific event and the fine-grained task of event encoding – identifying key phrases, key sentences describing event related information like type of the event, type of people involved in the event and extracting relationships between those. A variety of event domains, types and definitions combined with the multiple sources of data, scarcity of fine-level labels and unstructured data make both these tasks challenging. Event encoding falls into two categories - i) open information extraction methods and ii) domain specific event encoding methods. Open information extraction methods use predefined frameworks like CAMEO (Conflict and Mediation Event Observations) to encode the extracted entities and relationship. e.g. "Secretary of State John Kerry (entity) complained about (relationship) Russia’s (entity) support of Syria’s Bashar al-Assad" would be coded as US GOVERNMENT / DISAPPROVE / RUSSIAN GOVERNMENT. Domain specific event encoding approaches rely on pre-curated dictionaries and templates that depend on the structure underlying the given text [38]. Specifically, for civil unrest events a pre-defined template will contain fields like participants, purpose, location, and time whereas for insurgency events a template will contain fields like actor and target, violence level, date and casualties. Prior work in event encoding has focused on extracting entities, detecting trigger terms and matching slots on predefined templates [5, 23, 38]. 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 using sentence-level embedding [23] removes contextual information results in false-negatives because event occurrences are not neatly partitioned into unique sentences. Multiple Instance Learning (MIL) approaches have been proposed [14, 34] 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 get 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 the text. 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 [37]. We use a recurrent neural network (RNN) based hierarchical model with attention mechanism at both levels in the hierarchy that constructs sentence and document representations using RNN based encoders. The sentence representation is constructed by attending to all words in the sentence, whereas the document representation is constructed by attending to all sentences. The model inherently weights words and sentences and hence can construct a representation that captures the entire context of a document. To capture the fact that a sentence might capture multiple aspects related to an event (e.g., cause, location, population) we adapt the hierarchical attention network [37] models 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.

We evaluate the proposed hierarchical attention models for the task of event detection on civil unrest (CU) and military action (MANSA) datasets. We also identify key sentences associated with event of interest, the trigger words in our model without any sentence-level or word-level labelings. Identifying key sentences and key words not only improves the efficiency of human-in-the-loop encoding systems such as [28] but also leads to explainable models. 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 was found to improve performance for event detection.
- Our FBMA approach achieves the state of the art results for event detection on three event datasets from two different domains in two different languages.
- We present a workflow that can be used to accurately extract key sentences and trigger words from news articles and augment human-in-the-loop event encoding systems with the corresponding recommendations.

2 RELATED WORK

2.1 Event Extraction

Event extraction has been an active area of research in the past decade. In event extraction supervised approaches usually rely on manually labeled training datasets and handcrafted ontologies. Li et al. [16] utilize the 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 [20]. These approaches are usually limited by the availability of the fine-grained labeled data and required elaborately designed features. Different from these approaches our method uses attention mechanism to implicitly weigh words and sentences and is able to extract the event extents and trigger words with labels only at the document level. This formulation is suitable because labels at document level are easier to obtain than per-sentence level or word level. This makes the task of event extraction also amenable to Multiple Instance Learning (MIL) [8] solutions. In MIL ‘bags’ are groups of ‘instances’ which are to be classified. 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. Kotzias et al. [14] focus on instance-level predictions from group level labels and allow for the application of general aggregation functions for sentiment classification. Wang et al. [34] use a similar idea and have 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.

End2End MemNetworks [30]

2.2 Attention Mechanism

The practice of applying history particularly in computer vision problems [6, 15, 36]. Attention mechanism can be of two types. ‘Hard’ attention model learns to put attention to only specific parts of the sequences or images [22] whereas, ‘soft’ attention learns the attention weights over the entire sequence. Gulchere et al. [9] enhance the attention mechanism by incorporating a planning mechanism. Based on the past actions and current observation the model makes a plan of future alignments. Xu et al. [36] use attention to automatically generate description of images. They have explored both “soft” and “hard” attention mechanisms. Spearheaded by their success in neural machine translation [1, 2, 9, 19, 31] attention mechanisms are used in problems involving sequence to sequence models such as question answering [7, 11, 29] text summarization [4, 24] and visual question answering [38, 39].

In sequence modeling, attention mechanisms allow the decoder to learn which part of the sequence it should “attend” to based on the input sequence and the output it has generated so far [2]. A special case of attention known as self-attention [18] or input attention [37] is used for text classification tasks such as sentiment analysis and natural language inference. A by-product of self-attention mechanism is visualization of words being attended leading to the explainability of the models.

Models have been proposed that compute multiple attention distributions over a single sequence of words. In multi-view networks proposed by Guo et al. [10] they 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. [18] 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. [19] simplify additive attention operation by introducing multiplicative attention which is faster to compute. In multiplicative attention, score between two feature vectors is learned using a bilinear projection matrix. Dot product attention [19] 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. [37] 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 this 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. [13] to factorize the matrix into two
low rank matrices to compute multiple attention distributions over words. Contrary to Guo et al. [10] 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 attentions and performs superior to their approach. We call 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. Our proposed model consists of five main components: 1) A sentence encoder that uses a bi-directional GRU [3] RNN as the sequence encoder. GRU uses a gating mechanism to track the state of sequences. There are two types of gates: the reset gate \( r_t \) and the update gate \( z_t \). The update gate decides how much past information is kept and how much new information is added. At time \( t \), the GRU computes its new state as:

\[
\hat{h}_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tanh(W_h x_t + U_h h_{t-1} + b_h)
\]

and the update gate \( z_t \) is updated as:

\[
z_t = \sigma(W_z x_t + U_z h_{t-1} + b_z)
\]

The RNN candidate state \( \tilde{h}_t \) is computed as:

\[
\tilde{h}_t = \tanh(W_h x_t + r_t \odot (U_h h_{t-1} + b_h))
\]

Here \( r_t \) is the reset gate which controls how much the past state contributes to the candidate state. If \( r_t \) is zero, then it forgets the previous state. The reset gate is updated as follows:

\[
r_t = \sigma(W_r x_t + U_r h_{t-1} + b_r)
\]

Consider a news article containing \( n \) sentences and each sentence contains \( T \) words. Given a sentence consisting of word tokens \( w_{it} \), \( t \in [0, T] \) where every word is converted to a real valued word vector \( x_{it} \) using the pre-trained embedding matrix \( W_e = R^{d \times |V|} \), \( x_{it} = W_e w_{it}, t \in [1, T] \) where \( d \) is the embedding dimension and \( V \) is the vocabulary. The embedding matrix \( W_e \) is fine-tuned during training.

Figure 1: Figure describes the system architecture for event detection.
We encode the sentence using a bi-directional GRU (bi-GRU) 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 $w_{t1}$ to $w_{t}$ whereas the backward GRU denoted by $\overleftarrow{GRU}$ processes it from $w_{t}$ to $w_{t1}$.

$$\mathbf{x}_{it} = \mathbf{W}_c \mathbf{w}_{it}$$

$$\overrightarrow{\mathbf{h}}_{it} = \overrightarrow{GRU}(\mathbf{x}_{it}, \mathbf{h}_{it-1}, \theta)$$

$$\overleftarrow{\mathbf{h}}_{it} = \overleftarrow{GRU}(\mathbf{x}_{it}, \mathbf{h}_{it+1}, \theta)$$

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 [19] 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 $\mathbf{u}_{it}$.

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

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:

$$f_{it} = \mathbf{u}_{it}^\top \mathbf{W}_i \mathbf{u}_w$$

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 [37], $\mathbf{u}_w$ can be seen as a high dimensional representation of the fixed query ‘What is the informative word’. $\mathbf{u}_w \in \mathbb{R}^l$, $\mathbf{u}_{it} \in \mathbb{R}^h$ and $\mathbf{W}_i \in \mathbb{R}^{lx2h}$. The attention weight for the word $\mathbf{x}_{it}$ can be computed through a softmax function.

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

#### 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 (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 $f_{it}$, we need to learn $\mathbf{W} = [\mathbf{W}_1, ..., \mathbf{W}_m] \in \mathbb{R}^{lx2hxm}$ 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 [13] and the matrix factorization approaches proposed in [17, 26] the bilinear projection matrix $\mathbf{W}_i$ can be factorized into two rank 1 matrices $\mathbf{P} \& \mathbf{Q}$. Eq.8 can be written as:

$$f_{it} = \mathbf{u}_{it}^\top \mathbf{W}_i \mathbf{u}_w = \mathbf{P}^\top \mathbf{u}_w \circ \mathbf{Q}^\top \mathbf{u}_{it} = \tilde{\mathbf{u}}_w \circ \tilde{\mathbf{u}}_{it}$$

Here $\mathbf{P} \in \mathbb{R}^{lm}$ and $\mathbf{Q} \in \mathbb{R}^{hxm}$ 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 $\mathbf{u}_{it} \in \mathbb{R}^{2h}$, the word hidden representation and $\mathbf{u}_w \in \mathbb{R}^l$, 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 $\alpha_{it} \in \mathbb{R}^m$ is obtained by computing a softmax function along the sentence.

<table>
<thead>
<tr>
<th>Notation</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>$N$</td>
<td># of articles in a news corpus</td>
</tr>
<tr>
<td>$n$</td>
<td># of sentences in an article</td>
</tr>
<tr>
<td>$T$</td>
<td># of words in a sentence</td>
</tr>
<tr>
<td>$m$</td>
<td>number of aspects</td>
</tr>
<tr>
<td>$f_{it}$</td>
<td>alignment score</td>
</tr>
<tr>
<td>$\alpha_{it}$</td>
<td>attention weight</td>
</tr>
<tr>
<td>$\mathbf{u}_{it}$</td>
<td>word hidden representation</td>
</tr>
<tr>
<td>$\mathbf{u}_w$</td>
<td>word level context vector</td>
</tr>
<tr>
<td>$h$</td>
<td>GRU hidden state dimension</td>
</tr>
<tr>
<td>$l$</td>
<td>word-level context vector dimension</td>
</tr>
<tr>
<td>$S_i$</td>
<td>matrix sentence embedding for $i^{th}$ sentence</td>
</tr>
</tbody>
</table>

Figure 2: An illustration of the proposed Factorized Bilinear Multi-Aspect Attention (FBMA) mechanism.
length:
\[ \alpha_{it} = \frac{\exp(f_{it})}{\sum_{t'} \exp(f_{i't'})} \]  (11)

Before computing softmax, similar to [13] we apply an additional tanh nonlinearity to \( f_{it} \). 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, \( (f_{it} \leftarrow \frac{f_{it}}{||f_{it}||}) \) after the Hadamard product. Fig. 2 illustrates the FBMA approach.

3.2.3 Sentence Representation. Let \( H_i = (h_{i1}, h_{i2}, ..., h_{iT}) \) be a matrix of all word annotations in a sentence; \( H_i \in \mathbb{R}^{T \times 2h} \). Let \( A_i = (\alpha_{i1}, \alpha_{i2}, ..., \alpha_{iT}) \) be the multi-aspect attention matrix for the sentence; \( A_i \in \mathbb{R}^{m \times T} \). The sentence representation for an aspect \( j \) given by \( \alpha_{ij} = \{\alpha_{i1}, \alpha_{i2}, ..., \alpha_{iT}\} \) can be computed by taking a weighted sum of all word annotations.

\[ s_{ij} = \sum_{k=1}^{T} h_{ik} \cdot \alpha_{jk} \]  (12)

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

\[ S_i = A_i H_i \]  (13)

Here \( S_i \in \mathbb{R}^{mx2h} \) 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 [34] we use the global attention mechanism over the sentence annotations to get the document representation. Specifically, given a document containing sentence embeddings \( \{s_1, ..., s_i, ..., s_n\} \) where each \( s_i \) is a flattened representation of the matrix sentence representation \( S_i \) as given by eq. 13 we get the document vector as follows.

\[ \overrightarrow{h}_i = \text{GRU}(s_i, h_{i-1}, \theta) \]  (14a)
\[ \overleftarrow{h}_i = \text{GRU}(s_i, h_{i+1}, \theta) \]  (14b)
\[ h_i = (\overrightarrow{h}_i, \overleftarrow{h}_i) \]  (14c)

The sentence annotation \( h_i \) is obtained by concatenating the forward and backward hidden representations of the bi-GRU. Document representation is obtained by attention over sentences.

\[ u_i = \tanh(W_s h_i + b_s) \]  (15a)
\[ \alpha_i = \frac{\exp(u_i^T u_s)}{\sum_i \exp(u_i^T u_s)} \]  (15b)
\[ d = \sum_i \alpha_i h_i \]  (15c)

Here \( u_s \), the sentence level context vector, is randomly initialized and learned along with other model parameters while training and \( d \) is the document representation that summarizes all the information in the article. Given the document representation, we use a two hidden layer MLP with dropout to get the class scores.

\[ \hat{y} = W_d d + b_c \]  (16)

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

\[ l = -(y \log(\hat{y}) + (1 - y) \log(1 - \hat{y})) \]  (17)

The sentence embedding given by eq. 13 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. [18] that is added to the loss:

\[ P = \frac{\sum_i \|A_i A_i^T - I\|^2_F}{n} \]  (18)

Where \( \| \cdot \|_F \) 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:

\[ L = \sum_d l + \lambda P \]  (19)

The summation is taken over all the documents in the batch and \( \lambda \) is a hyperparameter. We use the mini-batch stochastic gradient descent algorithm [12] with momentum and weight decay for optimizing the loss function and the backpropagation algorithm is used to compute the gradients.

3.3.1 Hyperparameters. We use a word embedding size of 100. The embedding matrix \( W_e \) is pretrained on the corpus using the gensim implementation of the widely used distributed representations models word2vec [21]. All words appearing less than 5 times are discarded. The GRU hidden state is set to \( h = 50 \). In FBMA the dimension of \( u_t \) is given by the dimension of the GRU hidden state, but the dimension of \( u_w \) can be tuned. Empirically we find that setting the dimension of \( u_w \) 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 for the experimental evaluations [25]. 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 the paper (ii) The AutoGSR dataset – This dataset comes from the Embers AutoGSR system [28] which is a web based

1 https://radimrehurek.com/gensim/
Table 2: First row indicates population classes participating in a protest in the CU dataset. Second row indicates the causes of protest.

<table>
<thead>
<tr>
<th>Event Population</th>
<th>General Population, Business, Legal, Labor, Agricultural, Education, Medical, Media</th>
</tr>
</thead>
</table>

Figure 3: An illustration of the model architecture. The FBMA component is used at the word level to get the matrix sentence embedding.

Table 3: Dataset statistics. Total number of news articles, average number of sentences per article and average number of words per article in the datasets.

<table>
<thead>
<tr>
<th>Dataset</th>
<th># articles</th>
<th># sents/article</th>
<th># words/article</th>
</tr>
</thead>
<tbody>
<tr>
<td>MANSAM</td>
<td>1,05,858</td>
<td>6.3</td>
<td>250.3</td>
</tr>
<tr>
<td>CU Spanish</td>
<td>24,110</td>
<td>11.4</td>
<td>337.1</td>
</tr>
<tr>
<td>CU English</td>
<td>32,019</td>
<td>10.1</td>
<td>358.2</td>
</tr>
</tbody>
</table>

Table 4: Various baselines and their key characteristics.

<table>
<thead>
<tr>
<th>Feature Method</th>
<th>hierarchical</th>
<th>key words</th>
<th>key sents</th>
<th>bilinear attention</th>
<th>multi-aspect attention</th>
</tr>
</thead>
<tbody>
<tr>
<td>MI-CNN</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>MI-GCNN</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>HAN</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>HSA</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>HSA</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>FBMA</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

Finally, we evaluate on iii) the Military Action and Non-State Actor (MANSAM) GSR dataset which is in English and Arabic. This contains event encodings from Gulf countries namely, Bahrain, Egypt, 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 MANSAM 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 3 for the overview of our datasets.

4.2 Comparative Methods

Table 4 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. [34]. We consider the MI-CNN model proposed by Wang et al. and its variants as our baselines. The MI-CNN approach first constructs a sentence vector by applying convolution in the temporal dimension followed by $k$ – maxpooling. Then a document vector is formed from sentence vectors in a similar way. Instance representation for each sentence is constructed by concatenating the sentence representation and the document representation. Finally probabilities at instance level are estimated and aggregated to compute the document level probability. This model extracts key sentences but does not extract trigger words from the document nor attends to those words while constructing a sentence representation. 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.
Table 5: Results 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. 8, 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.

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

Since, the proposed model is a multi-aspect attention mechanism, we compare it with another multi-aspect attention mechanism proposed by Lin et al. [18]. 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. [37], 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. 8. Having bilinear attention not only enables the model to learn better pairwise similarity between the word level context vector \( u \) and the the word hidden representations \( l \), but it also gives us the flexibility to vary the dimension of \( u \). We refer to 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. In our experiments all models were implemented using the Pytorch\(^6\) open source framework.

## 5 RESULTS

### 5.1 Event Detection

Table 5 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 \( u \). We find that models with \( l \leq 2h \) where \( l \) is the dimension of \( u \) 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 HAN 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. [18].

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.

#### 5.1.1 Multi-Aspect vs. Single-Aspect.

To understand multi-aspect vs. single aspect attention mechanisms we compare the top K sentences picked by FBMA, a multi-aspect attention model vs. HAN, a single aspect attention model based on their attention weights. For FBMA we average the attention weights of \( m = 5 \) aspect for each word and visualize the resulting distribution. Since FBMA attends to different parts of a sentence in each aspect FBMA visualizations are dense. It can be seen from Tables 6 & 7 that the sentence attention weights for FBMA are more evenly distributed as compared to HAN which assigns high weight to a single sentence. Uniformly distributing attention among a select few key sentences, produces a more contextual document representation, rather than relying on only the top sentence.

\(^6\)https://pytorch.org/
Figure 4: 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).

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 4. 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.

5.3 Event Extraction

One of the advantages of our approach is the ability to identify key sentences and trigger words. Sentences and trigger words with high attention weights can be used to enhance the recommendation models of semi-automated event encoding systems such as AutoGSR [28] and also increase the efficiency of human encoders. In this section we demonstrate how we use the event extents or the key sentences identified in the event detection stage to extract more relevant information about the event such as ‘who’ and ‘why’ of the event. Specifically, we demonstrate the use of our model to identify the population class and the event type from the summaries of the articles for the Civil Unrest (English) dataset. Each trigger word corresponding to the event type and population class might have multiple references throughout the article but might not be contributing to the event. Our hypothesis is that the key sentences extracted by our model would be more informative of event related information and hence would do a better job in identifying the relevant mentions.

For a subset of articles in the CU English dataset, we have labels for the population type that participated in the protest, and the type of event. The population classes corresponding to one of the 11 classes...
Table 6: Comparison between multi-aspect and single aspect attentions. Top 3 sentences visualized as a heat map of word attention weights. Darker shades correspond to higher attention weights. Visualization for FBMA is average attention weights of 5 aspects. ‘Attn’ refers to sentence attention weight. HAN misclassifies the protest article as non-protest because of assigning highest weight to sentence 1 not containing protest related information. FBMA assigns highest weight to sentence 3 containing words like ‘protest’; and correctly classifies it.

<table>
<thead>
<tr>
<th>Attn</th>
<th>FBMA</th>
<th>HAN</th>
<th>Attn</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.24</td>
<td>police on friday arrested separatist leaders syed ali shah geelani and mirwaiz umar farooq as they tried to march from their residences to jamia masjid in srinagar nowhatta area</td>
<td>police on friday arrested separatist leaders syed ali shah geelani and mirwaiz umar farooq as they tried to march from their residences to jamia masjid in srinagar nowhatta area</td>
<td>0.60</td>
</tr>
<tr>
<td></td>
<td>geelani was arrested by police as he tried to defy the house arrest orders a spokesperson of the hardline hurriyat conference said</td>
<td>geelani was arrested by police as he tried to defy the house arrest orders a spokesperson of the hardline hurriyat conference said</td>
<td>0.08</td>
</tr>
<tr>
<td>0.224</td>
<td>a spokesperson of the moderate hurriyat conference said its chairman mirwaiz umar farooq was also arrested and taken to nigeen police station as he tried to march to jamia masjid for friday prayers</td>
<td>a spokesperson of the moderate hurriyat conference said its chairman mirwaiz umar farooq was also arrested and taken to nigeen police station as he tried to march to jamia masjid for friday prayers</td>
<td>0.19</td>
</tr>
<tr>
<td>0.289</td>
<td>the separatist camp have jointly called for a protest march against the recent civilian deaths in security forces action during the clashes in the wake of the killing of hizbul mujahideen commander burhan wani on july 8</td>
<td>the separatist camp have jointly called for a protest march against the recent civilian deaths in security forces action during the clashes in the wake of the killing of hizbul mujahideen commander burhan wani on july 8</td>
<td>0.19</td>
</tr>
<tr>
<td></td>
<td>authorities have imposed strict curfew in the city to thwart any such attempts</td>
<td>authorities have imposed strict curfew in the city to thwart any such attempts</td>
<td></td>
</tr>
</tbody>
</table>

are listed in Table 2, including ‘Education’, ‘Labor’, ‘Legal’ etc. The event type refers to the cause of the event and belong to one of the 5 event types like ‘Government Policies’, ‘Economic Policies’ etc. listed in Table 2.

Each news article is first summarized by selecting $K_i$ sentences based on the article length, where $K_i$ is dynamically computed as $K_i = \text{max}(1, \lceil x_i + \eta \rceil)$ where $\eta$ is a predefined value in (0, 1). We set $\eta = 0.3$ which selects top 2 or 3 sentences from each article. Our goal is to predict the population class and event type labels from the summary of the article. This is a multi-class classification problem and we pick the HAN classifier for this task.

We train all models using 3-fold cross-validation with early stopping and use the same set of hyperparameters as for detection. Performance is evaluated based on the summaries chosen by our model against other baselines. We compare this approach to the following baselines: (i) Random, where we randomly select $K_i$ sentences from an article as our summary. (ii) Start/End, News articles tend to organize key information in the beginning and end of the article so another baseline uses $K_i$ sentences from the top and $K_i$ sentences from the bottom of the article. (iii) MI-CNN We use the top $K$ sentences ranked by the MI-CNN model and the HAN models as other baselines. Figure 5 shows the accuracy of the classifier trained on FBMA summaries vs. baselines. Classifier trained on FBMA summaries outperforms other baselines indicating that our model picks the most informative key sentences containing event related information.

For qualitative visualization we highlight the words assigned the highest attention weight in each sentence, by the binary attention-based classifier, the protest classifier and the population class classifier in Table 8. The words highlighted in pink indicate the population specific words, the words highlighted in yellow show the protest trigger words and the words highlighted in green show the protest type specific words or the ‘why’ of the event. As can be seen in Table 8 words such as ‘teachers’ for education population, ‘residents’ for general population, ‘workers’ for labor population indicating the ‘who’ of the protest are assigned the highest attention.
Table 7: Comparison between multi-aspect and single aspect attentions. Top 3 sentences visualized as a heat map of word attention weights. Darker shades correspond to higher attention weights. Visualization for FBMA is average attention weights of 5 aspects. 'Attn' refers to sentence attention weight. HAN misclassifies the non-protest article as protest because of relying on the top key sentence containing words such as 'rally' which are protest trigger words. FBMA attends to the context of 'rally' containing words such as 'torch', 'fireworks' determining that it refers to a peaceful rally and correctly classifies it.

<table>
<thead>
<tr>
<th>Attn</th>
<th>FBMA</th>
<th>HAN</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.40</td>
<td>the city governments social media team has created an event on facebook with regard to the celebrations, the city government will organize a torch bearing rally fireworks and an exhibition to mark the independence day</td>
<td>the city governments social media team has created an event on facebook with regard to the celebrations, the city government will organize a torch bearing rally fireworks and an exhibition to mark the independence day</td>
</tr>
<tr>
<td>0.23</td>
<td>a fireworks show will be held at gaddafi stadium early on august 14</td>
<td>a fireworks show will be held at gaddafi stadium early on august 14</td>
</tr>
<tr>
<td>0.22</td>
<td>the ravi town administration will take out a torch bearing rally from the tmas office to masti gate chowk</td>
<td>the ravi town administration will take out a torch bearing rally from the tmas office to masti gate chowk</td>
</tr>
</tbody>
</table>

Table 8: Event Visualization. Green color highlights refer to Event Type keywords, Pink Color highlights refer to Population Keywords and Yellow Color highlights refer to Protest Related Keywords. These are top attended key words by an attention-based classifier picked from the top sentences selected by our model indicating that our model picks the most informative keywords.

<table>
<thead>
<tr>
<th>Sentence</th>
<th>Population Class</th>
<th>Event Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>congress and left workers at a rally in kolkata to protest against post poll violence in the state</td>
<td>Labor</td>
<td>Government Policies</td>
</tr>
<tr>
<td>however traffic was affected in certain pockets within the city limit owing to road picketing members of trade unions staged demonstration picketed roads and attempted rail roko to show their protest against the central government</td>
<td>Labor</td>
<td>Employment and Wages</td>
</tr>
<tr>
<td>traffic movement in and around majestic area was disrupted for several hours following a rally organized by karnataka rajya raitha sangha demanding land for the poor and landless farmers the rally comprising over 5,000 members was flagged off by freedom fighter doreswamy</td>
<td>Agriculture</td>
<td>Energy and Resources</td>
</tr>
<tr>
<td>hundreds of teachers of delhi university took to the streets on friday again to protest a university grants commissions age notification that could lead to around 5,000 temporary teachers losing their jobs</td>
<td>Education</td>
<td>Employment and Wages (Government Policies)</td>
</tr>
<tr>
<td>Rawalpindi: The residents staged a protest demonstration against prolonged power outages and blocked Adiala road yesterday. The commuters and pedestrians faced hardships and vehicular movement remained suspended for many hours as violent protestors placed trolleys and trucks in the middle of the road. According to details a large number of residents...</td>
<td>General Population</td>
<td>Energy and Resources</td>
</tr>
<tr>
<td>dalit organizations have called for a bandh on wednesday to protest the brutal thrashing of the community youths in una</td>
<td>Ethnic</td>
<td>Government Policies</td>
</tr>
</tbody>
</table>

weights. The event type classifier attends to the words indicative of the reason for protest such as ‘poll violence’ for Government Policies, ‘trade’ for Employment and Wages, ‘power outages’ for Energy and Resources. The event detection classifier overwhelmingly attends to protest trigger words such as ‘protest’, ‘rally’, ‘strikes’, ‘demonstrations’.
With the hierarchical models we used our proposed FBMA mechanism which leads to contextual sentence and document representations. 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.

REFERENCES


[36] K. Xu et al. 2015. Show, attend and tell: Neural image caption generation with visual attention. In ICML. 2048–2057.


