SD^{2}SG: A Novel Framework for Detecting Important Subevents from Crisis Events via Dynamic Semantic Graphs

Evangelia Spiliopoulou
Language Technologies Institute, Carnegie Mellon University
espiliop@cs.cmu.edu

Tanay K. Saha **Joel Tetreault** **Alejandro Jaimes**
Dataminr Inc.
{tsaha,jtetreault,ajaimes}@dataminr.com

Abstract

Social media is an essential tool to share information about crisis events, such as natural disasters. Event Detection aims at extracting information in the form of an event, but considers each event in isolation, without combining information across sentences or events. Many posts in Crisis NLP contain repetitive or complementary information which needs to be aggregated (e.g., the number of trapped people and their location) for disaster response. Although previous approaches in Crisis NLP aggregate information across posts, they only use shallow representations of the content (e.g., keywords), which cannot adequately represent the semantics of a crisis event and its sub-events. In this work, we propose a novel framework to extract critical sub-events from a large-scale crisis event by combining important information across relevant tweets. Our framework first converts all the tweets from a crisis event into a temporally-ordered set of graphs. Then it extracts sub-graphs that represent semantic relationships connecting verbs and nouns in 3 to 6 node sub-graphs. It does this by learning edge weights via Dynamic Graph Convolutional Networks (DGCNs) and extracting smaller, relevant sub-graphs. Our experiments show that our extracted structures (1) are semantically meaningful sub-events and (2) contain information important for the large crisis-event. Furthermore, we show that our approach significantly outperforms event detection baselines, highlighting the importance of aggregating information across tweets for our task.

1 Introduction

Social media is widely used for informing humanitarian aid efforts in crisis events (Nazer et al., 2017; Reuter et al., 2017). During a large-scale crisis event, there is a large set of smaller events in duration and impact that are essential components of the larger event, which are called *sub-events*. A sub-event is a structure that represents an action, and thus has a temporal dimension and a list of entities involved. Detecting important sub-events that occur during a crisis (e.g., road blocks, people trapped) can aid authorities to prevent and respond to urgent situations (e.g., rescue efforts) (Nazer et al., 2017). However, this requires connecting information from multiple posts as they contain repetitive or complementary information which needs to be aggregated (e.g., the number of trapped people and their location) for disaster response.

1. #NepalQuake avalanche kills 8 at Nepal's Everest base-camp
2. Obliterated Everest basecamp where at least 10 people were buried alive by avalanche after Nepal earthquake
3. Route to camp1 completely destroyed by avalanche.#NepalQuake
4. Avalanche sweeps Everest base-camp, killing 17: An avalanche triggered by Nepal's massive earthquake...
5. #Everest avalanche more than 100 climbers stuck in camp1 awaiting rescue.! #NepalQuake

Table 1: Example tweets from **April 2015 Nepal Earthquake** crisis event.

Several approaches in crisis NLP aggregate information across multiple tweets in the form of clusters, where each cluster is considered a sub-event (Abhik and Toshniwal, 2013; Pohl et al., 2012; Arachie et al., 2020). However, these methods have several shortcomings. First, the output clusters may not refer to a single sub-event, but to a list of sub-events that share similar information types. For example, consider the tweets *25 people killed in Everest base-camp* and *200 people killed in Gorkha*. They contain the same information type (i.e., number of people killed), but clearly refer to two different sub-events. This results in large, non-interpretable clusters that lack cohesion

(Jiang et al., 2019). Second, most work ignores or uses heuristics to model the temporal dependencies of sub-events, without explicitly modeling time-sensitive information that gets updated, such as the number of injured people. Third, tweet content is represented by shallow semantics, such as bag-of-words or verb-noun pairs. Such representations miss information that distinguishes between different sub-events of the same information type and are inadequate to model semantic dependencies across sub-events. To provide an example, consider the Nepal earthquake in April 2015. In Table 1, we show tweets referring to a deadly avalanche in mountain Everest which was triggered by the earthquake. We may extract *100 climbers were trapped in camp 1 and 2* from one tweet and *the route to camp 1 and camp 2 was completely destroyed* from another. Although these two tweets refer to different sub-events, they are related and an event extraction framework would benefit from modeling their dependencies.

Representing events in text is a complex problem. Most work on event detection relies on the **semantic frame** theory (Fillmore, 2008), according to which an event is represented by the **predicate**, typically a verb or noun denoting an action, and the **event arguments**, a list of entities related to the predicate via a specific set of relations (e.g., agent, patient,..). In this work, we use the same notion of event representations. Thus, we can distinguish different sub-events even if they have the same predicate and/or partially share entities, as in the example discussed earlier.

Recent work on social event understanding proposed a method to model event dependencies. They construct a sequence of graphs representing all the documents (Deng et al., 2019). They preserve temporal dependencies of events by using a Dynamic Graph Convolutional Network (DGCN); a model that learns an expressive graph representation of nodes not only from their connections in a certain time-step, but also from the dynamic context of the previous time-step. In our framework we exploit the expressive power of a DGCN to aggregate tweet content and model large-scale crisis events, by learning graph edge weights. These weights let us identify important nodes and relations; a critical step for sub-event extraction (Meladianos et al., 2018).

We propose SD^{2}SG, (Sub-event Detection via Dynamic Semantic Graphs): a novel framework to extract important sub-events from a temporally-ordered group of tweets related to a crisis event. Our approach combines information across tweets into a set of temporally-ordered graphs, which are used to extract sub-events. Since we have limited data, we impose structural (entities can be connected only via a predicate) and semantic constraints (predicates are defined via the FrameNet ontology) in each graph. Thanks to these constraints, our model learns valid relations instead of coincidental co-occurrences of words via the use of a DGCN model. Finally, we exploit the same structural constraints another time to rank 3 to 6 node subgraphs (sub-events) from the learned graph of a crisis event. The contributions of our proposed framework are summarized as follows:

- This is the first work in Crisis NLP that extracts sub-events in the form of semantic relations (i.e., predicate with a list of arguments); in prior work sub-events correspond to clusters of words or tweets.
- Our framework aggregates information across tweets and models temporal and semantic dependencies between sub-events. This problem is not addressed by event detection approaches, as they treat each sub-event independent.
- We conduct a large-scale human evaluation of the quality of the extracted sub-events, according to which SD^{2}SG outperforms baselines by at least 3% and 6% in terms of the validity and the importance of extracted sub-events respectively.

2 Related Work

Relevant research focuses on two main directions: (*i*) information extraction or classification in the tweet/sentence level and (*ii*) information aggregation across documents / posts.

Information Extraction & Classification in Tweets. Given the large volume of noisy data from social media, most tasks focus on sentence classification problems, where the goal is to filter only the most important posts that might be helpful for first responders. As discussed by Imran et al. (2015); Tapia et al. (2011), there are several types of sentence classification for disaster response, such as determining if a message is related to a specific crisis event (Caragea et al., 2016; Kruspe, 2019; Nguyen et al., 2016; Neubig et al., 2011), if it is actionable (Leavitt and Robinson, 2017) or critical (Mccreadie et al., 2019; Spiliopoulou et al., 2020). Other work classifies tweets with respect to the type of information they contain, a problem that is formulated as a multi-class tweet classification (typically five major information types) (Burel et al., 2017; Nguyen et al., 2017; Imran et al., 2016; Miyazaki et al., 2019; Padhee et al., 2020).

Related work outside of Crisis NLP can also be used to extract information from tweets, in the form of events. Chen et al. (2018) use an encoder-decoder framework to extract sub-events from each tweet, while Rudra et al. (2018) use noun-verb pairs to represent sub-events, where each pair is ranked based on their overlap score in tweets. Some approaches outside the crisis domain that focus on extracting textual sub-events from tweets or documents, in a sequence classification setup (Bekoulis et al., 2019). Other related work includes Open IE methods (open information extraction), which extract tuples of expressions from text that represent the events of the sentence. Such work includes Open IE by AllenNLP (Stanovsky et al., 2018), which uses a deep BiLSTM sequence prediction model and systems that combine BERT embeddings with other neural models, such as a BiLSTM encoder (Kolluru et al., 2020).

Aggregating Information Across Tweets. Research on Crisis NLP that aggregates information across posts aims at extracting sub-events that are important in the context of the larger crisis event. The notion of sub-events varies within this area; a sub-event could correspond to an entire cluster of words / tweets or to a textual span from a single tweet. Earlier work in sub-event extraction forms clusters of tweets during a crisis event based on a set of shallow features, such as tf-idf and metadata (Abhik and Toshniwal, 2013; Pohl et al., 2012). Other approaches use topic clustering to form sets of words (topics) that represent sub-events (Srijith et al., 2017; Xing et al., 2016). Most recent work forms clusters based on verb-noun pairs from individual tweets (Jiang et al., 2019), which are then ranked based on an ontology grounding score (Arachie et al., 2020). In all these methods each cluster is considered to correspond to a different sub-event. However, the elements within each cluster are not necessarily related via temporal or other relations, which raises questions with respect to the interpretability of the cluster/sub-event.

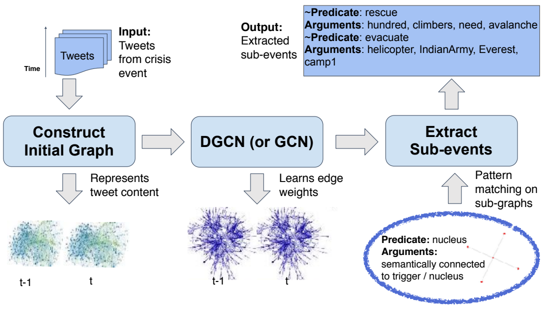

Figure 1: Framework architecture diagram.

In a different direction, a group of temporally ordered messages is used for large-event detection. For example, (Sakaki et al., 2010) use statistical and keyword features in a spatio-temporal model to detect crisis events based on Twitter streams. More recently, Meladianos et al. (2015, 2018) use a graph representation of tweets to extract important sub-events by detecting weight changes, a problem formulated as a summarization task. Early approaches that use text from social media to represent context for social events rely on linear classifiers using topic-related features (Wang et al., 2012), graph features (Keneshloo et al., 2014) or combination of heterogeneous data sources (Korkmaz et al., 2015) and dynamic query expansion models with a static vocabulary (Zhao et al., 2015) or fused with logistic regression (Ramakrishnan et al., 2014). Ning et al. (2018); Zhao et al. (2015) use multi-task models with shared parameters across different locations and events to model spatio-temporal correlations. Most recent work that inspired our approach uses dynamic graphs to represent information from social media, which model temporal constraints from precursor events (Deng et al., 2020, 2019; Ning et al., 2018). A common theme of this work, as further discussed by Ning et al. (2019) underlines the importance of explainability, since it is helpful for experts to analyze which factors led to the development of a large-scale event and potential ways to prevent or mitigate it.

3 Proposed Framework: SD^{2}SG

In this section, we discuss our approach on extracting sub-events from a stream of temporally-ordered tweets related to a crisis event. Our framework consists of the following steps, as shown in Figure 1: (*i*) construct the initial dynamic semantic graph, (*ii*) learn the graph's weights via a graph neural net-

Algorithm 1: Steps of SD^{2}SG

```
Input  : $C$ = A set of crisis events,
         $tweet_1, \ldots, tweet_n$ in a temporal order,
         $n$ = total number of tweets in a crisis event
         $t$ = number of time-steps,
         $k$ = number of sub-events,
         pre-trained embeddings
Output : Extracted sub-events for each time-step $t$:
         $S_1, S_2, \ldots, S_t$ from each crisis event.
1. for each crisis event, $C_i \in C$ do
       for tweets $\in \{relevant, irrelevant\}$ do
           Divide $tweet_1, tweet_2, \ldots, tweet_n$ into
             groups of equal size, $D = D_1, D_2, ..., D_t$;
           for each $D_i \in D$ do
               Construct semantic graph $G_i$;
2. Run learning framework, DGCN on the output
   from Step 1 and extract indicator function $I$ based
   on DGCN's parameters to extract graph weights;
3. for each crisis event, $C_i \in C$ do
       // Only run on graph constructed
          from relevant tweets
       for each $G_i \in G$ do
           a) Extract sub-graph $G_i^{'}$ based on $I$;
           b) Sample sub-events from random graph
              walks in $G_i^{'}$;
           c) Collect sub-events that meet semantic
              constraints (event structure);
           d) Rank extracted sub-events with tf-idf score.
              Choose top $k$;
```

work, and (*iii*) extract sub-events from the learned graph via random walks that satisfy our semantic constraints. The pseudocode of our framework's steps is in Algorithm 1.

3.1 Constructing Initial Dynamic Semantic Graphs

Given a large-scale crisis event, our first step is to represent the content of the related tweets in a graph structure by merging information across all the messages, which we call **initial graph**. An initial graph represents the tweets for a given time-step; it can be used dynamically in a sequence of initial graphs (i.e., one graph per time-step) or as a single graph (i.e., one time-step).

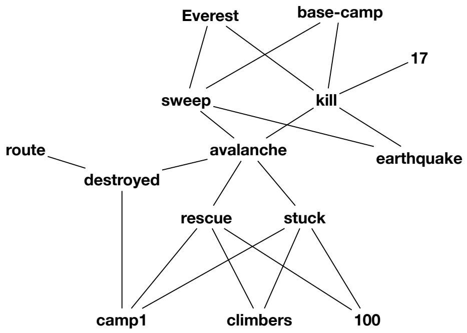

Figure 2: Initial semantic graph, based on subset of Nepal 2015 earthquake tweets (tweets 3, 4, 5)

There are multiple ways to build the initial graph representation of tweets. In SD^{2}SG, we use a sequence of **initial semantic graphs**, where each graph is based on semantic relations from text. Given a set of tweets from a specific time-step, the initial semantic graph is a bipartite graph that connects predicates with their arguments, as they appear together in text. A group of tweets can be represented by a single initial graph, where the same predicate may be connected to different arguments from different sentences. An example of an initial semantic graph is shown in Figure 2, which is constructed based on tweets 3, 4 and 5 in Table 1.

Formally, given a tweet t_i, we use a dependency parser to extract the part-of-speech tags from tweets. Based on these tags we form two groups: (1) verbs and nominalized verbs (i.e., nouns that are derived from verbs, like *explosion*) $V_i = \{v_1, v_2, ...\}$, by matching the tokens to the Lexical Units provided in FrameNet (Baker et al., 1998) and (2) nouns (excluding nominalized verbs) $N_i = \{n_1, n_2, ..\}$. The output graph has as nodes $\cup_i N_i \cup V_i$ for all tweets t_i. We form weighted edges only across the two groups (verbs V and nouns N), which are initialized based on the PMI of each pair (Church and Hanks, 1990). More specifically, for each tweet t_i, we have $(v_i, n_j) \forall i, j$ but no (n_i, n_j) edges. This ensures that we link the sentence's predicate with its arguments, while avoiding to link arguments that appear together under different relations/predicates. As shown in Figure 2, this results in a graph that combines information across tweets in a more explainable way compared to previous approaches (Deng et al., 2019), while maintaining semantic relations from text.

3.2 Learning Edge Weights via a DGCN

The initial semantic graphs mainly capture information mentioned in sentence level, without taking context into account (i.e., information based on neighboring relations). This results in large graphs with noisy relations, where it is hard to extract important information. In order to get a smaller, less noisy graph, we formulate our problem as a classification task using dynamic graph convolutional networks (DGCNs) where the goal is to learn edge weights, a mechanism introduced by Deng et al. to detect social events for news articles. With this setup the model learns which neighbourhoods or sets of nodes in the graph are important and correspond to sub-events that occur during a crisis.

A DGCN model consists of a sequence of GCNs that are linked together by feeding information from the previous time-step to detect important factors in the context of social event understanding. Each initial graph is fed into a different GCN layer by time. For each GCN layer except the first one, the input features are processed by a temporal encoded module, involving the output of the last GCN layer and the current word embeddings, to capture temporal features. Finally, there is a masked nonlinear transformation layer to unify the final output vector from the final GCN layer. The loss is calculated between the model output and ground truth label, which, in our case, is whether a group of tweets is related to a crisis event or not.

Formally, given a sequence of initial semantic graphs, we form their normalized adjacent matrices $A_1, A_2, ..A_t$ for each time-step t. We are also given a matrix of initial node features H_0, which in our case corresponds to the pre-trained word embeddings of the vocabulary. At each time-step, the convolutional layer of the DGCN is computed by: $H_{t+1} = g(A_t \bar{H}_t W^{(t)} + b^{(t)})$, where $W^{(t)}, b^{(t)}$ are model parameters and g is a non-linear activation function. Note that $\bar{H}_t$ does not correspond to the GCN output, but instead to the temporal encoding embeddings, calculated from the last TE layer. The temporal encoding is defined based on the following equations, where W_p, W_e, b_p, b_e are the parameters learned by the model:

$$H_p^{(t)} = H_t W_p^{(t)} + b_p^{(t)} \tag{1}$$

$$H_e^{(t)} = H_0 W_e^{(t)} + b_e^{(t)} \tag{2}$$

$$\bar{H}_t = tanh(|H_p^{(t)}||H_e^{(t)}|) \tag{3}$$

In order to classify the group of tweets as related or not to a crisis event, we set the output feature dimension of the last layer as 1. Due to dynamic graph encoding, the output feature vector of the last GCN layer is a combined representation of all graph nodes, which is different for each large-crisis event (i.e., different graph nodes). To guarantee the consistency of the model across instances, the DGCN uses a masked nonlinear transformation layer to map the final output vector to the prediction of the task. Finally, for each node i in the graph, we use the scalar value $h_{i,t}$ from the last GCN layer and $w_{i,m}$ from the masked nonlinear transformation layer to define an indicator function $I_i = h_{i,t} w_{i,m}$. This indicator function is used to select important nodes and their edges from the graph.

3.3 Extracting Sub-events

Given the sequence of learned graphs, the last part of our framework aims at extracting significantly smaller sub-graphs that represent sub-events (i.e. a typical sub-event contains 3-6 terms, while a graph might have 100-200 nodes). Although we use bipartite graphs to represent tweets, during learning, we treat them as homogeneous with zero edge-weight because of the nature of GCN/DGCN models (they operate on homogeneous graphs, where all nodes are treated equally). In order to generate valid sub-event candidates, we use a pattern matching method based on iterations of random walks on each graph (Bressan et al., 2018; Saha and Hasan, 2015). The patterns used correspond to the typical structure of an event, where the predicate (usually a verb) is linked to a set of arguments (entities/nouns). Similarly to the semantic constraints in each initial semantic graph, we generate sub-events of star-like patterns of variable size (3-6 nodes), where the center node is an event predicate, as defined by the FrameNet lexicon.

After we extract our candidate sub-events, we use a ranking method to remove duplicate or redundant information. To do that, we use a tf-idf scoring scheme, where each sub-event is treated as a single document; the score of each sub-event equals to the average tf-idf score of its words. While other ranking or filtering methods can be used, tf-idf is most appropriate as it retrieves important information (tf) and avoids similar, almost duplicate sub-events (idf).

4 Training & Evaluation

Crisis Event Dataset. To train the model described in Algorithm 1, we use a subset of the combined dataset described in Alam et al., which consists of Twitter data from 59 crisis events, including natural and man-made disasters. The tweets in this dataset were all manually annotated by Alam et al. as either being related or unrelated to their corresponding crisis event. The statistics of the dataset are shown in Table 2.

Crisis Event Type	# Crisis Events	# Related Tweets	# Unrelated Tweets
Hurricane/Typhoon	13	22,154	13,219
Crash/Explosion	11	7,689	9,718
Flood	11	12,366	9,747
Earthquake	10	12,164	6,911
Terrorist Attack	3	5,956	4,205
Tornado	3	6,242	5,034
Wildfire	3	2,842	348
MERS	1	1,113	69
Ebola	1	1,420	210
Volcano	1	104	191
Haze	1	476	136
Landslide	1	364	2,800
Total	59	73,070	52,588

Table 2: Dataset Statistics; number of tweets refers to tweets related to the large-scale crisis event.

Training Details. We execute Step 1 of Algorithm 1 on these 59 crisis events. For the models based on dynamic graphs (SD^{2}SG and Simple Dynamic Graph) we use time-step $t = 3$.

We randomly split the dataset in Table 2 into train, development, and test sets, where an event belongs to only one of these sets. Related and unrelated sub-events are positive and negative examples, respectively. This set-up was used to train the DGCN model (Step 2 of Algorithm 1). Out of 59 crisis events, we use 33, 10 and 16 as training, dev and test sets. For word embeddings we used 100d GloVE (Pennington et al., 2014) pre-trained on Twitter, and the DGCN was trained using the Adam optimizer with learning rate 5e-4, weight decay 5e-4, and dropout rate 0.2.

Once the DGCN is trained, we execute Step 3 in Algorithm 1 and extract sub-graphs of interest for our evaluation. We evaluate our extracted sub-events with respect to two factors: (*i*) validity and (*ii*) importance in the context of a large-scale crisis event.

4.1 Baselines

To verify our assumption that information aggregation is important for our task, we chose baselines consisting of various methods that either aggregate information across tweets or not. We use Open IE as the baseline that does not aggregate information, while the remaining baselines are ablations of our proposed model. Our ablations study lets us verify the impact of every component of the proposed model. Unfortunately, since no prior work in crisis NLP extracts sub-events in the form of semantic relations, we cannot compare with these methods. Here is a brief overview of each baseline:

Large Scale Crisis Events	Extracted Sub-events
2014 India-Pakistan floods	1467
2012 Colorado wildfires	693
2013 Alberta floods	680
2013 Balochistan earthquakes	715
2013 Dhaka garment factory collapse	800
2013 Los Angeles International Airport shooting	687
2013 South Wales bushfires	683
2017 Puebla earthquake	710
2015 Nepal earthquake	939
2019 Covid pandemic	818
Cyclone Oswald	850
2013 Spuyten Duyvil derailment	771
Hurricane Harvey	688
MERS epidemic	894
2014 Typhoon Hagupit	690
West Texas Fertilizer Company explosion	915
Total	13,000

Table 3: Large-scale crisis events and their number of extracted sub-events.

Simple Dynamic-Graph: uses a complete graph (i.e., edges across all pairs of nodes) without any constraints. The weight of each edge is based on the PMI of the two nodes. This model was proposed by Deng et al. (2019) to model social events.

Static Sem-Graph: (1 time step) constructs only one graph for all the tweets, without taking into account their temporal dimensions. The initial graph is constructed in a similar manner as the proposed model, but the weights are learned via a static GCN model.

Init Sem-Graph: uses the sequence of initial semantic graphs as-is (no learning of graph weights)

Open IE: uses the output of an Open IE system for each individual tweet to directly produce sub-event candidates. For this baseline, we use the OpenIE system developed by Stanovsky et al. (2018). Each sub-event is formed by using Open IE's *predicate* as the sub-event predicate and the head nouns of each *argument phrase* as the sub-event arguments. Since the output is already a set of sub-events instead of a graph, we directly rank them based on tf-idf features, similarly to the last step of the proposed model.

Crisis Predicate ⟶	Yes			No				
Event Arguments ⟶	All	Some	None	All	Some	None		
Models ↓	Sub-event Validity Score						**Predicate Accuracy**	**Sub-event Accuracy**
Open IE	0.30%	0.50%	2.70%	1.00%	9.20%	86.30%	4.50%	1.80%
init sem-graph (no learning)	7.70%	14.60%	13.00%	7.00%	18.80%	42.00%	39.20%	29.30%
Simple dynamic	8.00%	10.00%	11.30%	7.00%	13.10%	56.80%	30.00%	18.70%
static sem-graph	5.99%	8.90%	9.85%	1.60%	17.50%	56.10%	26.20%	16.49%
SD^{2}SG (proposed)	9.70%	15.30%	13.70%	7.10%	17.70%	36.50%	**45.40%**	**32.10%**

Table 4: Percentage of valid sub-events. Sub-event Accuracy represents instances that fall under the *Yes and All/Some* and *No and All* categories.

4.2 Validity of Sub-events

We perform a human evaluation of our extracted sub-events based on crowdsourced annotations via Amazon Mechanical Turk.

Data. We collect a total of 13,000 sub-events (details shown in Table 3) by selecting the top 100 sub-events for each baseline per time-step from 16 large-scale crisis events (after removing events with few instances).

Annotation Guidelines. First, we want to assess whether the extracted sub-events are valid. We showed every candidate sub-event $s_i = (t, a_1, a_2, ..)$ (where t is the predicate and $a_1, a_2, ..$ are the event arguments) to three MTurk annotators and asked them the following questions:

1. Does the predicate represent a crisis incident (e.g., outage, collapse, injury) during a major crisis event? Possible answers: *yes* or *no*.
2. How many of the argument words describe a crisis scenario with or without the predicate? Possible answers: *all*, *some*, or *none*.

We estimate the inter-annotator agreement of these judgments via Fleiss' Kappa; the predicate accuracy has $k = 0.5$, while sub-event accuracy $k = 0.37$.

Metric & Eval Summary. In table 4 we show the detailed results of our human evaluation. To estimate the accuracy of the sub-events overall (predicate and event arguments) for each baseline, we merged the answers of three categories; Yes and All, Yes and Some and No and All. We decided this merging for two reasons. First, some sub-events might have predicates that are not clearly related to a crisis (e.g., *keep, go*), but in combination with proper arguments the entire structure is a valid, meaningful sub-event in a crisis scenario (e.g., predicate: *fly*, arguments: *rescuers, climbers, Everest*). Second, some sub-events may be partially valid; the event predicate and some (but not all) of the arguments are valid. Such instances still contain meaningful information for the crisis event and could be used to inform decisions.

Our results show that SD^{2}SG outperforms all baselines. This highlights that all the components of the model (the initial semantic graph, the temporal aspect and the learned weights) contribute to a better model overall. However, we observe that the initial semantic graph is the second best performing model, with only 3% difference. From that, we conclude that the semantic and structural constraints are a crucial component to extract valid sub-events.

4.3 Importance of Sub-events

Determining the importance of a sub-event is a complex task that requires expert annotators, as they need to consider the context of the crisis event. Even though a sub-event may be valid with respect to its structure, we still need to validate if it is important in the context of the large-scale crisis event.

Data. We used the sub-events from the top performing baselines that were previously classified to belong to one of the following categories (sub-event accuracy); *Yes and All*, *Yes and Some* and *No and All*. Out of a total of 1,756 valid sub-events, we randomly select a subset of 300 ($\sim$ 80 sub-events per baseline).

Annotation Guidelines. To evaluate the importance of the sub-events, we conducted another human evaluation, where we asked two expert annotators the following question, for each sub-event:

1. Go to the provided Wikipedia link. Is the proposed sub-event important for this crisis event? Suppose the proposed sub-event did not happen, would the consequences of the major crisis or the humanitarian aid response be different? Possible answers: *yes* or *no*.

We estimate the inter-annotator agreement by Cohen's Kappa and Kappa score for this evaluation task is 0.48.

Models	**Important Sub-event Accuracy**
init sem-graph (no learning)	19%
simple dynamic static	15%
sem-graph	14%
SD^{2}SG (proposed)	**25%**

Table 5: Percentage of important sub-events. The second column is an estimate of the important sub-events that each model extracts.

Metric & Eval Summary. To evaluate sub-event importance we estimate the percentage of all extracted sub-events that are important, per model (important sub-event accuracy). The goal of this metric is to reflect how good each system is in extracting important sub-events.

To estimate the important sub-event accuracy we use the results obtained from both human evaluations. The first evaluation tells us how many valid sub-events each system extracts, while the second how many of these valid sub-events are important, per system. Each annotated sub-event was considered important if any of the two annotators labeled it as such. Thus, the important sub-event accuracy per system is estimated by $\frac{valid\_sub}{extracted\_sub}\frac{important\_sub^*}{valid\_sub^*}$, where $valid\_sub^*$ and $important\_sub^*$ correspond to the number of valid and important sub-events respectively in the annotated sample (i.e., $valid\_sub^*$ = 300).

The results of this evaluation are shown in Table 5. We observe that our proposed model performs substantially better than the baselines (6% higher than the second-best). Although the accuracy of all systems is relatively low, this is due to the low percentage of valid events (i.e., a sub-event must be valid in order to be important).

5 Discussion

In the previous section we show that despite SD^{2}SG performed significantly better than our baselines, our numbers are overall low; 45.5% of our extracted sub-events are valid and only 25% important. In this section we identify and discuss a set of reasons why sub-event extraction of tweets is a challenging problem and how we can improve.

Models	**Predicate Overlap**	**Argument Overlap**
init sem-graph (no learning)	85.10%	65.90%
PMI dynamic static	**92.85%**	58.20%
sem-graph	88.75%	67.60%
SD^{2}SG (proposed)	90.40%	**68.00%**

Table 6: Overlap of extracted sub-events with terms from the EM ontology.

The first step of our analysis is to compare our extracted sub-events to an existing resource manually curated by experts in crisis NLP. We used the EMTerms (Emergency Management Terms) ontology (Temnikova et al., 2015); a resource of 7,000 manually annotated terms that are used in Twitter to describe crisis events, classified into 23 information-specific categories. Based on this lexicon and our extracted sub-events, we estimate the percentage of predicates and arguments that exist in the EM terms by a partial string matching (many EMTerms are phrases of 2-3 words). In Table 6 we show the results of this grounding. We observe that, overall, a large percentage of the predicates can be grounded in the ontology, while the argument overlap is significantly lower. This can be explained by our semantic constraints on the predicates of the extracted sub-events (must exist in FrameNet), while the event arguments had no such constraints. However, given the results of our human evaluation in the previous section, we conclude that, even though a word might be a crisis related keyword, a sub-event formed by such keywords is not necessarily valid. This is due to the fact that sub-events aim to represent relations between several terms, thus grounding to an ontology is not a sufficient metric of the quality of the extracted sub-events.

Crisis Event	Predicate	Arguments
1. Alberta floods	evacuation	flooding, zone, Canada
2. Puebla earthquake	follow	rescuer, victims
3. Typhoon Hagupit	keep	safety, flee
4. Puebla earthquake	school	kill, child, dead
5. MERS	cough	healthcare, surveillance
6. Duyvil derailment	fatality	derailment,helicopter, major, abc, amtrak
7. MERS	emergency	infection, Fukuda discover
8. Dhaka garment factory collapse	rescue	survivors, number, labor,factory

Table 7: Example output from SD^{2}SG

In Table 7, we show a few real-output examples that highlight the complexity of sub-events. These sub-events belong to any of the three accepted categories of valid sub-events (*Yes and All/Some* or *No and All*). Although they were all considered valid by human annotators, we observe a few major challenges. Although some predicates are not crisis words, they could still form a valid crisis sub-event with the appropriate arguments. Such an example is the predicate *follow* in sub-event 2. However, for some other instances (example 4), the predicate might be an entity in the particular context.

A major problem in our framework is that we don't know how the arguments are related to the predicate. Although semantic frames consist of specific relations (e.g., agent, patient, location), our framework provides only a list of the related entities without their relations. An open challenge is to use thematic roles both for tweet representation and for the extracted sub-events, as this will result in more meaningful sub-events that would be easier to evaluate. Given that SD^{2}SGis modularized, it can be modified to represent thematic roles by using heterogeneous graph neural networks (HGNNs) instead of DGCNs. HGNNs are a type of network that consists of multiple types of edges or nodes. However, extracting thematic roles from text (first step of SD^{2}SG) would still be a particularly hard task due to the nature of social media text, which does not always conform to proper syntax and grammar.

6 Conclusion

In this paper, we propose a novel framework to extract sub-events from a large-scale crisis event. Contrary to earlier views of sub-events as clusters of unrelated words or phrases, our methodology aims at extracting sub-events in the form of a predicate and its arguments. Our framework aggregates information across a set of tweets into dynamic graph representations, while maintaining semantic constraints. Through an extensive qualitative analysis of our extracted sub-events, we show that our approach performs better than other baselines and highlight the challenges of our task.

References

Dhekar Abhik and Durga Toshniwal. 2013. Sub-event detection during natural hazards using features of social media data. In *Proceedings of the 22nd International Conference on World Wide Web*, pages 783–788.

Firoj Alam, Hassan Sajjad, Muhammad Imran, and Ferda Ofli. 2020. Standardizing and benchmarking crisis-related social media datasets for humanitarian information processing. *arXiv preprint arXiv:2004.06774*.

Chidubem Arachie, Manas Gaur, Sam Anzaroot, William Groves, Ke Zhang, and Alejandro Jaimes. 2020. Unsupervised detection of sub-events in large scale disasters. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 354–361.

Collin F Baker, Charles J Fillmore, and John B Lowe. 1998. The berkeley framenet project. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, pages 86–90.

Giannis Bekoulis, Johannes Deleu, Thomas Demeester, and Chris Develder. 2019. Sub-event detection from twitter streams as a sequence labeling problem. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 745–750.

Marco Bressan, Flavio Chierichetti, Ravi Kumar, Stefano Leucci, and Alessandro Panconesi. 2018. Motif counting beyond five nodes. 12(4).

Grégoire Burel, Hassan Saif, and Harith Alani. 2017. Semantic wide and deep learning for detecting crisis-information categories on social media. In *International semantic web conference*, pages 138–155. Springer.

Cornelia Caragea, Adrian Silvescu, and Andrea H Tapia. 2016. Identifying informative messages in disaster events using convolutional neural networks. In *International Conference on Information Systems for Crisis Response and Management*, pages 137–147.

Guandan Chen, Nan Xu, and Weiji Mao. 2018. An encoder-memory-decoder framework for sub-event detection in social media. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pages 1575–1578.

Kenneth Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational linguistics*, 16(1):22–29.

Songgaojun Deng, Huzefa Rangwala, and Yue Ning. 2019. Learning dynamic context graphs for predicting social events. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1007–1016.

Songgaojun Deng, Huzefa Rangwala, and Yue Ning. 2020. Dynamic knowledge graph based multi-event forecasting. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1585–1595.

Charles J Fillmore. 2008. Frame semantics. In *Cognitive linguistics: Basic readings*, pages 373–400. De Gruyter Mouton.

Muhammad Imran, Carlos Castillo, Fernando Diaz, and Sarah Vieweg. 2015. Processing social media messages in mass emergency: A survey. *ACM Computing Surveys (CSUR)*, 47(4):1–38.

Muhammad Imran, Prasenjit Mitra, and Carlos Castillo. 2016. Twitter as a lifeline: Human-annotated twitter corpora for nlp of crisis-related messages. *arXiv preprint arXiv:1605.05894*.

Shan Jiang, William Groves, Sam Anzaroot, and Alejandro Jaimes. 2019. Crisis sub-events on social media: A case study of wildfires. In *International Conference on Machine Learning AI for Social Good Workshop, Long Beach, United States, July*, volume 1.

Yaser Keneshloo, Jose Cadena, Gizem Korkmaz, and Naren Ramakrishnan. 2014. Detecting and forecasting domestic political crises: a graph-based approach. In *Proceedings of the 2014 ACM conference on Web science*, pages 192–196.

Keshav Kolluru, Samarth Aggarwal, Vipul Rathore, Soumen Chakrabarti, et al. 2020. Imojie: Iterative memory-based joint open information extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5871–5886.

Gizem Korkmaz, Jose Cadena, Chris J Kuhlman, Achla Marathe, Anil Vullikanti, and Naren Ramakrishnan. 2015. Combining heterogeneous data sources for civil unrest forecasting. In *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*, pages 258–265.

Anna Kruspe. 2019. Few-shot tweet detection in emerging disaster events. *arXiv preprint arXiv:1910.02290*.

Alex Leavitt and John J Robinson. 2017. The role of information visibility in network gatekeeping: Information aggregation on reddit during crisis events. In *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*, pages 1246–1261.

Richard Mccreadie, Cody Buntain, and Ian Soboroff. 2019. Trec incident streams: Finding actionable information on social media.

Polykarpos Meladianos, Giannis Nikolentzos, François Rousseau, Yannis Stavrakas, and Michalis Vazirgiannis. 2015. Degeneracy-based real-time subevent detection in twitter stream. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 9.

Polykarpos Meladianos, Christos Xypolopoulos, Giannis Nikolentzos, and Michalis Vazirgiannis. 2018. An optimization approach for sub-event detection and summarization in twitter. In *European Conference on Information Retrieval*, pages 481–493. Springer.

Taro Miyazaki, Kiminobu Makino, Yuka Takei, Hiroki Okamoto, and Jun Goto. 2019. Label embedding using hierarchical structure of labels for twitter classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6318–6323.

Tahora H Nazer, Guoliang Xue, Yusheng Ji, and Huan Liu. 2017. Intelligent disaster response via social media analysis a survey. *ACM SIGKDD Explorations Newsletter*, 19(1):46–59.

Graham Neubig, Yuichiroh Matsubayashi, Masato Hagiwara, and Koji Murakami. 2011. Safety information mining—what can nlp do in a disaster—. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 965–973.

Dat Nguyen, Kamela Ali Al Mannai, Shafiq Joty, Hassan Sajjad, Muhammad Imran, and Prasenjit Mitra. 2017. Robust classification of crisis-related data on social networks using convolutional neural networks. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11.

Dat Tien Nguyen, Shafiq Joty, Muhammad Imran, Hassan Sajjad, and Prasenjit Mitra. 2016. Applications of online deep learning for crisis response using social media information. *arXiv preprint arXiv:1610.01030*.

Yue Ning, Rongrong Tao, Chandan K Reddy, Huzefa Rangwala, James C Starz, and Naren Ramakrishnan. 2018. Staple: Spatio-temporal precursor learning for event forecasting. In *Proceedings of the 2018 SIAM International Conference on Data Mining*, pages 99–107. SIAM.

Yue Ning, Liang Zhao, Feng Chen, Chang-Tien Lu, and Huzefa Rangwala. 2019. Spatio-temporal event forecasting and precursor identification. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 3237–3238.

Swati Padhee, Tanay Kumar Saha, Joel Tetreault, and Alejandro Jaimes. 2020. Clustering of social media messages for humanitarian aid response during crisis.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Daniela Pohl, Abdelhamid Bouchachia, and Hermann Hellwagner. 2012. Automatic sub-event detection in emergency management using social media. In *Proceedings of the 21st international conference on world wide web*, pages 683–686.

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*, pages 1799–1808.

Christian Reuter, Marc-André Kaufhold, Thomas Spielhofer, and Anna Sophie Hahne. 2017. Social media in emergencies: a representative study on citizens' perception in germany. *Proceedings of the ACM on Human-Computer Interaction*, 1(CSCW):1–19.

Koustav Rudra, Pawan Goyal, Niloy Ganguly, Prasenjit Mitra, and Muhammad Imran. 2018. Identifying sub-events and summarizing disaster-related information from microblogs. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 265–274.

Tanay Kumar Saha and Mohammad Al Hasan. 2015. Finding network motifs using mcmc sampling. In *Complex Networks VI*, pages 13–24, Cham. Springer International Publishing.

Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. 2010. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*, pages 851–860.

Evangelia Spiliopoulou, Eduard Hovy, Alexander G Hauptmann, et al. 2020. Event-related bias removal for real-time disaster events. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 3858–3868.

PK Srijith, Mark Hepple, Kalina Bontcheva, and Daniel Preotiuc-Pietro. 2017. Sub-story detection in twitter with hierarchical dirichlet processes. *Information Processing & Management*, 53(4):989–1003.

Gabriel Stanovsky, Julian Michael, Luke Zettlemoyer, and Ido Dagan. 2018. Supervised open information extraction. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 885–895.

Andrea H Tapia, Kartikeya Bajpai, Bernard J Jansen, John Yen, and Lee Giles. 2011. Seeking the trustworthy tweet: Can microblogged data fit the information needs of disaster response and humanitarian relief organizations. In *Proceedings of the 8th International ISCRAM Conference*, pages 1–10. ISCRAM Lisbon, Portugal.

Irina P Temnikova, Carlos Castillo, and Sarah Vieweg. 2015. Emterms 1.0: A terminological resource for crisis tweets. In *ISCRAM*.

Xiaofeng Wang, Matthew S Gerber, and Donald E Brown. 2012. Automatic crime prediction using events extracted from twitter posts. In *International conference on social computing, behavioral-cultural modeling, and prediction*, pages 231–238. Springer.

Chen Xing, Yuan Wang, Jie Liu, Yalou Huang, and Wei-Ying Ma. 2016. Hashtag-based sub-event discovery using mutually generative lda in twitter. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30.

Liang Zhao, Qian Sun, Jieping Ye, Feng Chen, Chang-Tien Lu, and Naren Ramakrishnan. 2015. Multi-task learning for spatio-temporal event forecasting. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1503–1512.