Information Retrieval from Microblogs during Natural Disasters

Roshni Chakraborty
Indian Institute of Technology
Patna, Bihar
India
roshni.pcs15@iitp.ac.in

Maitry Bhavsar
Indian Institute of Technology
Patna, Bihar
India
bhavsar.mtcs15@iitp.ac.in

ABSTRACT

In this paper, we devise an information retrieval system which can filter and rank tweets according to relevance to the query. We devise methods to understand relationships among entities and action verbs from a small set of manually annotated tweets. We further use these relationships to filter tweets and rank them accordingly. Our results (as published by FIRE Microblog Track) show that we have high precision score in detection of topmost 20 tweets.

1. INTRODUCTION

FIRE 2016 Microblog track [1] provided us with about 50,000 tweets related to Nepal Earthquake of April 2015. In this paper, we segregate tweets into different categories, namely, availability of resources, requirement of resources, availability of medical resources and facilities, requirement of medical resources and facilities and information related to infrastructure destruction or restoration. We devised a mechanism to learn text attributes of tweets to segregate them into specific categories.

We manually annotate a random sample of 1000 tweets into specified categories of information, a tweet can also belong to multiple groups. For example, a tweet of *destruction of a bridge* also might convey information about the *requirement of basic amenities*. Hence, different categories of tweets had different text attributes that pertain to a specific information related to that query. We aimed at identifying those text attributes, i.e, combination of different words for any particular query. We further created networks of each query's text attributes' combinations. The edges represent the interrelationships among these text attributes which aid in segregation of tweets according to different queries. We will describe the methodology in details in later sections.

Tweets are informal, so a *vocabulary gap* exists even among tweets of same strata. So, we did not depend only on text analysis of named entities, like food packets but rather combined them with the set of important verbs that identifies a correct relationship among those. We weighed the different identified keywords of each category according to their relevance to the query. We, thereby, could identify tweets due to their presence of relevant keywords for a query. The published results from FIRE suggest we could accurately identify tweets of high relevance of different categories with good precision and recall.

We have divided the paper into following sections. We discuss about data collection and pre-processing in the next section, followed by our procedure of identification of tweets in section 3 and finally results and discussion in section 4 and conclusion in section 5.

2. DATA COLLECTION AND PRE-PROCESSING

FIRE Microblog Track provided a set of about 50,000 tweet-ids which we used to the access the tweets. We filtered only the relevant tweet information from these tweets, that consist of tweet text, tweet id, etc. We further filtered some tweets from the whole set of tweets. For example, during disasters, there are a number of tweets that express grief, urge people to pray or help. These messages are general messages, hence we made a bag of words that express only *urge*, *request*, *pray* etc. and removed those tweets that contain words from this bag.

3. METHODOLOGY

In this section, we discuss our procedure. We do not use any external source of information. We use NLTK toolkit[1] to perform text based analysis on tweets. We rely on tweet text attributes to filter tweets of relevance. In order to understand the text attributes of tweets, we select a random sample of 1000 tweets from the whole set of 50,000 tweets. We manually group tweets according to different queries by FIRE, a tweet can nevertheless belong to different categories.

We perform a set of operations on tweet text for every group (as specified before). Firstly, we remove the *stopwords* from these tweets. *Stopwords* hardly represent any special characteristic of an entity. After removal of the stop words, we use POS Tagger to select only the nouns and verbs from the tweets. We then rank the entities of all tweets according to frequency. We select a subset of these entities according to the ranks, we also include the entities specified in the query itself by FIRE. This step gives us a list of the important entities for a specific query.

Often, an entity to entity matching fails to resolve tweets of different genre, i.e., a tweet containing information of *medical aids* can either highlight availability or requirement of the same. So, we identify the different set of possible actions of any entities, to understand the underlying relationships. We further rank the bigrams to identify the set of *working verbs* to highlight specific actions. Thus, this set of related *working verbs* and *entities* signify tweets of a particular category.

However, there remains a *vocabulary gap* among different tweets of even same category due to their informal structure.

[1] www.nltk.org

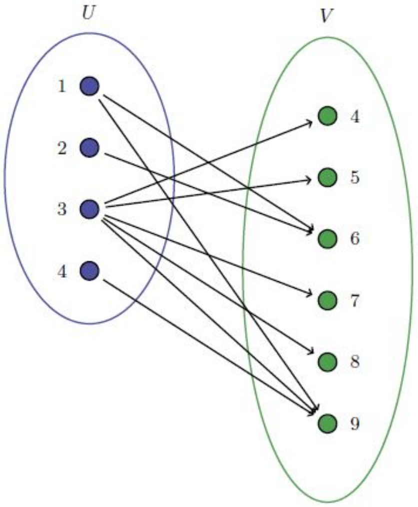

Figure 1: Graph Relationships of Resource Availability Information

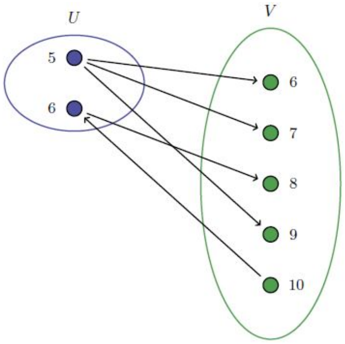

Figure 2: Graph Relationships of Resource Requirement Information

Tweets of both requirement and availability of medical resources may contain entities, like *blood* and working verbs like *donate* but are completely different in meaning. Hence, segregation only on the basis of keywords fails to differentiate these relationships. We analyze the context of those keywords relationships, which reflects the actual meaning, as in the absence of question tags (like, *where*, *how*, *what*, etc), or request tags (*please*, etc) in availability based tweets. The segregation of tweets into different categories thus requires identification of proper *entities*, *actions*, and *context* to understand it's relevance.

We further have ranked an entity and the action verbs according to their importance, which we will explain later. We formulate separate bipartite graphs for each query, that represents the relationships among the entities, actions and context. While a set of nodes represent entities' names, another set of nodes represent the names of verbs (i.e., actions). These relationships were formulated from the manually annotated tweets. We give a brief overview of the specific words and their relationships for each query in the next section.

We select the different types of action verbs from our manually annotated tweets into similar groups. The main action verbs represent *donations*, $transport$, $relief\ information$, *build*. We represent the relationships between these different set of action verbs with different set of entities in the graphs 1 and 2 and the set of keywords of each group in table 1. Thus a new tweet is selected if it contains the existing relationship, as mentioned by the arrow, i.e., it must contain at least an entity and verb from the nodes the arrow connects.

Node Name	Words Representing the Nodes
$Green_4$	off to nepal
$Green_5$	survivor, victim, affect
$Green_6$	food, water, cloth, blanket, biscuit, power, plane, bus, material, beef, equipment
$Green_7$	volunteer, helicopter, item, tool, app
$Green_8$	team
$Green_9$	shelter, tent, house, home
$Green_1 0$	relief, rescue
$Blue_1$	donate
$Blue_2$	transfer, sell, distribut, suppl, send, sent, deliver, dispatch, offer, land, deploy, transport, prepar
$Blue_3$	relief, rescue, working, aid, support, engage, rush
$Blue_4$	build
$Blue_5$	need, want, require

Table 1: Word Dictionary of Resource Availability and Requirement Related Information

3.1 Requirement of Resources

In this section, we intend to filter all tweets that mention the requirement or need of some resource, like human resources or infrastructure like tents, water filter, power supply, etc. We studied our manually annotated tweets, and found the main action verbs that denote requirement of resources are, *need* related or *relief* related. We highlight the different relationships among these various entities in figure 3 and include details of the different terms in table 1. Thus, we later select those tweets from the total list of tweets if it contains the relationship represented by the arrow, i.e., it contains at least an entity and action verb from the list of keywords that the arrow connects.

3.2 Availability of Medical Resources

In this section, we identify messages that mention the availability of some medical resources like blood, blood bank, medicine, etc. Firstly, we distinguish different action verbs from the manually annotated tweets that contain information related to this query, the verbs are namely *donation*,

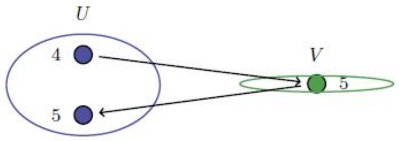

Figure 3: Graph II Relationships of Medical Resource Availability Information

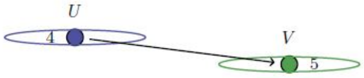

Figure 4: Graph Relationships of Medical Resource Requirement Information

transport, *rescue* etc. There are some action verbs that are ambiguous in meaning, example *need* reflects both the need and the availability of resources. On further analysis of *need* mentioned tweets, we found *need* is used in availability of resources tweets only in conditional statements (example, if is a conditional clause).

We represent the actions and their corresponding entities in the next two graphs, namely graph 4 and graph 2, the arrows represent the relationships among the both. The table 2 represents the set of keywords for each entity or action. Thus, we filter all those tweets from the whole set of fifty thousand tweets which contain the relationships, i.e., at least a keyword from both the nodes of an arrow. There are also some stringent relationships, that comprise of more than just an entity and action name, as illustrated in graph 4.

Node Name	Words Representing the Nodes
$Green_5$	blood, bloodbank, medicine, medical, doctor
$Green_6$	healthcare, hospital, patient, diabities
$Green_7$	provide, survivor, victim,affect
$Blue_1$	donate, donated
$Blue_2$	reach, transfer, sell, distribut, suppl, send, sent, deliver, dispatch, offer, land, deploy, transport, prepar, continu
$Blue_3$	rescue, relief, support, engag, rush
$Blue_4$	need, want require
$Blue_5$	call, contact, helpline

Table 2: Word Dictionary of Medical Resource Availability and Requirement Information

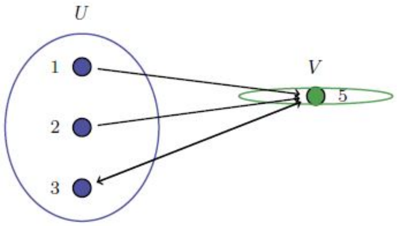

Figure 5: Graph Relationships for Devastation Related Information

3.3 Requirement of Medical Resources

In this section, we identify messages that mention the requirement of some medical resources like blood, blood bank, medicine, etc. We represent the actions and their corresponding entities in graph 5, the arrows represent the relationships among the both. The table 2 represents the set of keywords for each entity or action. Thus, we filter all those tweets from the whole set of fifty thousand tweets which contain the relationships, i.e., at least a keyword from both the nodes of an arrow.

3.4 Infrastructure Damage And Report of Restoration

In this section, we identify messages that mention the damage or restoration of any communication or structural infrastructures. However, the general statements about any structure is not relevant. We filter the possible set of infrastructure names from our manual annotated tweets and the different set of actions related to them. After detection of the relationships among the action verb and entity name from the manually annotated tweets, we select only those tweets that contain We visualize the different relationships among different set of entities in Figure 6, and highlight the set of keywords in table 3.

Node Name	Words Representing the Nodes
$Green_1$	hotel, debris, building, temple, rubble, tower, road, bridge, house, railway, dam, tent, heritage, monument, power grid, engineer, equipment, electricity
$Blue_1$	reduce, flatten, destroy, devastat, avalanche, damage, restore, capture, collapse, build, builds
$Blue_2$	devastat, terrif, heartbreak
$Blue_3$	footage, image, picture

Table 3: Word Dictionary for Devastation Related Information

Node Name	Words Representing the Nodes	Score
$Action_1$	relief, rescue, aid	0.10
$Action_2$	build, transfer, sell, distribut, send, sent, deliver, supply, donat, need	0.4
$Action_3$	deploy, dispatch, lad, transport, fly	0.3
$Action_4$	prepare, offer, launch, allow, provide, make, support, engag, rush, help, working, in action	0.2
$Entity_1$	volunteer, food, biscuit, shelter, tent, house, home, cloth, blanket	0.7
$Entity_2$	power, equipment, material, item, team, helicopter, bus, plane, call, helpline,contact	0.5

Table 4: KeyWord Relevance Score

4. SELECTION OF TWEETS

The above graphs represent different entities, and their set of actions for a particular query. For a given query, we match the relationships among the new tweet with the prescribed relationships. Thus, a tweet is selected if it contains the specified relationships of entities of that query. We further rank those tweets according to it's relevance to the query in the next section.

4.1 Score of Tweets

In this section, we rank the selected tweets by their relevance to query. In order to rank the tweets, we score the different keyword relationships of a query. The keywords are segregated into two different sections, entities and action verbs. We give importance to words that signify better temporal relevance than others, i.e., there is a major difference between tweets like *food items sent to affected areas by Indian government*, *India dispatched 500 packets of rice to Nepal* and *India will dispatch food packets by saturday*. We give a brief description of our scoring mechanism.

1. $Temporal\ Importance$: An action verb is given more importance if it highlights immediate action rather than future. This is illustrated by $Action_2$ and $Action_3$, $Action_4$.

2. $Relevance$: Some action verbs, represent greater relevance in times of calamity, as expressed in $Action_1$. Similarly, there are some *entities* (as in $Entity_1$), which are the basis needs of human livelihood, like food and shelter which are more important than information related to other *entity* (as in $Entity_2$).

The different scores of the keywords are given in table 4. Thus, a tweet's score is the summation of it's keywords' scores. We hereby, could rank the tweets by their relevance score accordingly.

Metric Name	Result
$Precision$@20	0.770
$Recall$@1000	0.4344
MAP@1000	0.2186
$OverallMAP$	0.2208

Table 5: Result

5. RESULTS

In this section, we highlight our results, FIRE Microblog Track matched our selected tweets with a manual annotator's results. We briefly give an explanation of the metrics and our results are depicted in Table 5. The metrics are.

1. Precision at rank 20, i.e., considering up to the top 20 tweets for each topic.

2. Recall at rank 1000.

3. Mean Average Precision at rank 1000.

4. MAP overall, i.e., considering all tweets retrieved in the run.

6. CONCLUSION

In this paper, we devise a mechanism to extract the contextual, content relationships of entities. We are able to filter tweets of high relevance for different queries by matching these relationships. We require a small number of manual annotated tweets to attain our results.

7. REFERENCES

[1] S. Ghosh and K. Ghosh. Overview of the FIRE 2016 Microblog track: Information Extraction from Microblogs Posted during Disasters. In *Working notes of FIRE 2016 - Forum for Information Retrieval Evaluation, Kolkata, India, December 7-10, 2016*, CEUR Workshop Proceedings. CEUR-WS.org, December 2016.