Microblog Retrieval During Disasters: Comparative Evaluation of IR Methodologies

Moumita Basu[1,2(✉)], Kripabandhu Ghosh[3], Somenath Das[1], Somprakash Bandyopadhyay[4], and Saptarshi Ghosh[1,5]

[1] Indian Institute of Engineering Science and Technology, Shibpur, Howrah, India
moumitabasu0979@gmail.com
[2] University of Engineering and Management, Kolkata, India
[3] Indian Institute of Technology, Kanpur, Kanpur, India
[4] Indian Institute of Management, Calcutta, Kolkata, India
[5] Indian Institute of Technology, Kharagpur, Kharagpur, India
saptarshi@cse.iitkgp.ernet.in

Abstract. Microblogging sites are important sources of situational information during any natural or man-made disasters. Hence, it is important to design and test Information Retrieval (IR) systems that retrieve information from microblogs during disasters. With this perspective, a track was organized at the 8th meeting of Forum for Information Retrieval Evaluation (FIRE) 2016, focused on microblog retrieval during disaster events. A collection of about 50,000 microblogs posted during the Nepal Earthquake in April 2015 was released, along with a set of seven pragmatic information needs during a disaster situation. The task was to retrieve microblogs relevant to these information needs. Ten teams participated in the task, and fifteen runs were submitted. Evaluation of the performances of various microblog retrieval methodologies, as submitted by the participants, revealed several challenges associated with microblog retrieval. In this chapter, we describe our experience in organizing the FIRE track on microblog retrieval during disaster events. Additionally, we propose two novel methodologies for the said task, which perform better than all the methodologies submitted to the FIRE track.

Keywords: Microblog retrieval · FIRE microblog track · Disasters
Word embedding · Word2vec

1 Introduction

In a disaster situation, access to situational information is crucial [29], since such information helps in reducing casualties, preventing secondary disasters, economic losses, and social disruption [3]. In recent times, microblogging sites such as Twitter have played important roles as information media in disaster management [12,27,28]. However, in such forums, important information is often obscured by a lot of personal opinions, emotion, and sentiment (e.g., prayers

P. Majumder et al. (Eds.): FIRE 2016 Workshop, LNCS 10478, pp. 20–38, 2018.
https://doi.org/10.1007/978-3-319-73606-8_2

and sympathy for the victims of the disaster). Thus, automated IR techniques are sought to extract precise, meaningful situational information from a large amount of social media text. There have been a few recent studies on identifying specific types of microblogs (or tweets) during disaster situations [1,27] (see Sect. 2). All these works have used their own datasets, and there has not been much effort till now to develop a standard test collection on which microblog retrieval methodologies can be compared empirically.

The FIRE (Forum for Information Retrieval Evaluation) 2016 Microblog track [10] was motivated by the TREC Microblog Track [19] which aims to evaluate microblog retrieval methodologies. In contrast, the FIRE 2016 Microblog Track focused on microblog retrieval in a disaster situation. The goal of FIRE 2016 Microblog track were two-fold – (i) to develop a benchmark dataset for evaluation of microblog retrieval methodologies during disaster situations, and (ii) to evaluate and compare the performances of various IR methodologies on the benchmark dataset.

In the FIRE 2016 Microblog track, a collection of about 50,000 microblogs posted during a recent disaster – the Nepal Earthquake in April 2015 – was released along with topics reflecting seven realistic information needs during disaster situations (obtained through discussion with agencies who respond to disasters). Minutiae about the collection are specified in Sect. 3. The task was to retrieve microblogs pertinent to the information requirements. Ten teams took part in the task and 15 runs were submitted that are described in Sect. 4. Standard measures of Precision and MAP were used to evaluate the run against the gold standard developed by human annotators. This book chapter describes our experience in organizing the FIRE 2016 Microblog track.

Additionally, in this chapter, we also propose two novel word embedding based retrieval models (using word2vec [17]) for the said task. We also apply two strategies for query expansion via pseudo relevance feedback – the standard Rocchio expansion, and an expansion strategy based on word embeddings (details in Sect. 5). The proposed methodologies perform better than all the methodologies submitted to the FIRE 2016 Microblog track (evaluation results in Sect. 6).

To summarize, our primary contributions are – (i) we develop a test collection for evaluating IR systems for microblog retrieval in disaster situations. (ii) we perform comparative evaluation of several diverse microblog retrieval methodologies, including methodologies involving Natural Language Processing, word embedding, and so on, and (iii) we also propose two novel methodologies for the said task, based on word embeddings, and demonstrate that our proposed methodologies perform better than all the submitted runs in the FIRE 2016 track. In general, we establish that word embedding based retrieval is a promising approach for dealing with the short, noisy nature of microblogs.

2 Related Work

2.1 Developing Test Collections for Evaluating IR Strategies

The Text REtrieval Conference (TREC – http://trec.nist.gov/) was perhaps the first endeavour to present testbeds for standard evaluation of IR systems. They advocated the Cranfield style [6] which states that an IR test collection should comprise three components – (1) Text representations of information needs, called *topics*, (2) A static set of documents, and (3) Relevance status of the documents with respect to a query, called *relevance assessments*. We also adopt this style while preparing our dataset.

The TREC Microblog Track [19] (introduced in 2011) focuses on the evaluation of microblog retrieval strategies in general. Various microblog retrieval tasks have also been organized as part of CLEF[1] and NTCIR.[2] However, to our knowledge, the FIRE 2016 task described in this chapter is the first task designed specifically for microblog retrieval in a real-life disaster situation.

2.2 IR on Microblogs Posted During Disasters

There has been lot of recent interest in addressing various challenges on microblogs posted during disaster events, such as classification, summarization, event detection, and so on. The reader is referred to [12] for a survey on these prior works.

Some datasets of social media posts during disasters have also been developed [7], but they are primarily meant for evaluating methodologies for classification among different types of posts (and not for retrieval methodologies).

Few methodologies for retrieving specific types of microblogs have also been proposed, such as tweets asking for help, and tweets reporting infrastructure damage [1,27]. However, all such studies have used different datasets. To our knowledge, there is no standard test collection for evaluating strategies for microblog retrieval in a disaster scenario; this chapter describes attempts to develop such a test collection.

3 Dataset

In this section, we discuss the procedure for developing the test collection. As stated earlier, we follow the Cranfield style [6] for developing the test collection. The formation of topics (information needs), document set (here, microblogs or tweets) collection and relevance assessment to prepare the gold standard necessary for evaluation of IR methodologies are explained in this section.

[1] http://clef2016.clef-initiative.eu/.
[2] http://research.nii.ac.jp/ntcir/index-en.html.

Table 1. Seven topics of interest to agencies who respond to disasters such as earthquakes or floods. Each topic is written following the format used in TREC tracks. The challenge is to retrieve microblogs relevant to each of these topics.

<num> **Number: FMT1** <title> **What resources were available** <desc> Identify the messages which describe the availability of some resources. <narr> A relevant message must mention the availability of some resource like food, drinking water, shelter, clothes, blankets, human resources like volunteers, resources to build or support infrastructure, like tents, water filter, power supply and so on. Messages informing the availability of transport vehicles for assisting the resource distribution process would also be relevant. However, generalized statements without reference to any resource or messages asking for donation of money would not be relevant.
<num> **Number: FMT2** <title> **What resources were required** <desc> Identify the messages which describe the requirement or need of some resources. <narr> A relevant message must mention the requirement / need of some resource like food, water, shelter, clothes, blankets, human resources like volunteers, resources to build or support infrastructure like tents, water filter, power supply, and so on. A message informing the requirement of transport vehicles assisting resource distribution process would also be relevant. However, generalized statements without reference to any particular resource, or messages asking for donation of money would not be relevant.
<num> **Number: FMT3** <title> **What medical resources were available** <desc> Identify the messages which give some information about availability of medicines and other medical resources. <narr> A relevant message must mention the availability of some medical resource like medicines, medical equipments, blood, supplementary food items (e.g., milk for infants), human resources like doctors/staff and resources to build or support medical infrastructure like tents, water filter, power supply, ambulance, etc. Generalized statements without reference to medical resources would not be relevant.
<num> **Number: FMT4** <title> **What medical resources were required** <desc> Identify the messages which describe the requirement of some medicine or other medical resources. <narr> A relevant message must mention the requirement of some medical resource like medicines, medical equipments, supplementary food items, blood, human resources like doctors/staff and resources to build or support medical infrastructure like tents, water filter, power supply, ambulance, etc. Generalized statements without reference to medical resources would not be relevant.
<num> **Number: FMT5** <title> **What were the requirements / availability of resources at specific locations** <desc> Identify the messages which describe the requirement or availability of resources at some particular geographical location. <narr> A relevant message must mention both the requirement or availability of some resource, (e.g., human resources like volunteers/medical staff, food, water, shelter, medical resources, tents, power supply) as well as a particular geographical location. Messages containing only the requirement / availability of some resource, without mentioning a geographical location would not be relevant.
<num> **Number: FMT6** <title> **What were the activities of various NGOs / Government organizations** <desc> Identify the messages which describe on-ground activities of different NGOs and Government organizations. <narr> A relevant message must contain information about relief-related activities of different NGOs and Government organizations in rescue and relief operation. Messages that contain information about the volunteers visiting different geographical locations would also be relevant. However, messages that do not contain the name of any NGO / Government organization would not be relevant.
<num> **Number: FMT7** <title> **What infrastructure damage and restoration were being reported** <desc> Identify the messages which contain information related to infrastructure damage or restoration. <narr> A relevant message must mention the damage or restoration of some specific infrastructure resources, such as structures (e.g., dams, houses, mobile tower), communication infrastructure (e.g., roads, runways, railway), electricity, mobile or Internet connectivity, etc. Generalized statements without reference to infrastructure resources would not be relevant.

3.1 Topics for Retrieval

In this track, our goal was to develop a test collection to evaluate IR methodologies for extracting information (from microblogs) that can essentially help Government agencies and other responding organizations to undertake any disaster

situation such as an earthquake or a flood more competently. To this end, we conferred with members of several NGOs like, Doctors For You[3] and SPADE,[4] who regularly work in disaster-affected regions. Our aim was to know what are the conventional information requirements during a humanitarian relief operation. The NGO members helped us to figure out some specific information requirements, such as, what resources are required/available (especially medical resources), what infrastructure damages are being reported, the situation at specific geographical locations, the ongoing activities of various NGOs and government agencies (so that the operations of various responding agencies can be coordinated), and so on. Based on their opinion, *seven* pertinent topics (information needs) were identified.

The topics, developed as a part of the test collection, are written in the format typically used for TREC topics in Table 1. Each topic contains an identifying number (num), a textual representation of the information need (title), a brief description (desc) of the same and a more detailed narrative (narr) explaining what type of documents (tweets) will be considered relevant to the topic, and what type of tweets would not be considered relevant.

3.2 Tweet Dataset

A destructive earthquake occurred in Nepal and parts of India on 25th April 2015. Tweets related to the Nepal earthquake, posted during the two weeks following the earthquake, were extracted through the Twitter Search API [26], using the keyword 'nepal'. In total, we collected 100 K tweets written in English (where the language recognition is done by Twitter itself).

It is known that some tweets are usually retweeted/reposted by multiple users [25], especially during events like a disaster. Thus, there is likely to be a considerable presence of duplicates and near-duplicates in the set of tweets. However, the presence of duplicates can result in over-estimation of the performance of an IR methodology. Hence, it is preferred that duplicates should not be present in a test collection for IR. Moreover, while developing the gold standard, the occurrence of duplicate documents also leads to overwork for human annotators [14]. Therefore, we removed duplicate and near-duplicate tweets using a simplified version of the methodologies illustrated in [25], as follows. Initially, tweets are pre-processed by excluding standard English stopwords and URLs. Subsequently considering each tweet as a bag of words, the similarity between two tweets was measured as the Jaccard similarity between the two corresponding bags (sets) of words. Two tweets were considered as near-duplicates if the Jaccard similarity between two tweets was measured to be greater than a threshold value (0.7). The longer tweet (potentially more informative) was retained in the collection.

After eliminating duplicates and near-duplicates, we obtained a set of 50,068 tweets, which was used as the test collection for the track.

[3] http://doctorsforyou.org/.
[4] http://www.spadeindia.org/.

3.3 Developing Gold Standard for Retrieval

The gold standard or ground truth judgment of relevance is a principal requirement in the evaluation of any IR methodology. We involved a set of three human annotators in developing the gold standard. Each of the annotators is a regular user of Twitter, has proficiency in English and has preceding experience of working with social media content posted during disasters. The gold standard was developed in three phases.

Phase 1: Each annotator was given the set of 50,068 tweets, and the seven topics (in TREC format, as stated in Table 1). Each annotator was instructed to identify all tweets relevant to each topic, separately, i.e., without conferring with the other annotators. The tweets were indexed using the well-known Indri IR system [24], to aid the annotators to search for tweets containing specific terms. For each topic, The annotators were instructed to figure out appropriate search terms for each topic, retrieve tweets containing those search terms (using Indri), and to manually judge the relevance of the retrieved tweets.

After the first phase, it is observed that the set of tweets identified to be relevant to the same topic by different annotators was notably dissimilar. This difference was because different annotators used different search terms to retrieve tweets.[5] Thus, a second phase was conducted.

Phase 2: In the second phase, for each topic, we considered the tweets that were judged relevant in the first phase by *at least one* of the annotators. The judgment on the topic-wise relevance of the tweet was confirmed through discussion among all the annotators and mutual agreement.

Phase 3: This phase used standard pooling [23] over all runs submitted to the FIRE 2016 track (as commonly done in TREC tracks) - for each topic the top ranked 30 tweets of all the submitted runs were pooled and annotated. In the third phase, all annotators were judging a common set of tweets; hence inter-annotator agreement could be measured. There was agreement among all annotators for over 90% of the tweets; for the rest, the relevance was decided through discussion among all the annotators and mutual agreement.

The final ground truth judgment of relevance includes the tweets pertinent to the seven topics. For each topic, number of relevant tweets and corresponding examples are illustrated in Table 2.

3.4 Insights from the Gold Standard Development Process

Through the process of developing the ground truth, we realize that for any of the topics, there are numerous tweets which are definitely relevant to the topic, but hard to retrieve even by manual annotation. This is apparent from the fact that, many of the relevant tweets could initially be retrieved by only one out of the three annotators (in the first phase), but when the tweets were shown to the

[5] Since the different annotators potentially judged different sets of tweets, reporting inter-annotator agreement would not be meaningful under these circumstances.

Table 2. Final ground truth judgment of relevance pertinent to the seven topics.

Topic	Relevant tweets	Example tweets
FMT1	589	Bharat Sevashram Sangha started relief work in earthquake hit Nepal. Cooked food being distributed in the outskirts of Kathmandu
FMT2	301	#SoulVultures: Pickaxes, shovels and earth-moving equipment required in Nepal. [url]
FMT3	334	mount of supplies may be used for more than 7 days. Mobile Hospital is equipped to even perform a surgery. #earthquake #Nepal #Kathmandu
FMT4	112	THT - Hospitals may face shortage of oxygen and medicine [url]
FMT5	189	RT @Goal_Nepal Sahara Club Sends Bottled Water, Noodles To Gorkha, Lamjung For Earthquake Victims READ: [url]
FMT6	377	We @CFCT_INDIA are collecting Blankets for Nepal earthquake victims. please contact .....[mobile-no] [url]
FMT7	254	#NBC #News Runway Damage Closes Nepal Airport to Heavy Planes: Earthquake-struck Nepal has been forced to clos... [url]

other annotators (in the second phase), they unanimously agreed that the tweet was relevant. These observations highlight the challenges in microblog retrieval.

It is to be noted that our process of developing the gold standard is dissimilar from the approach used in TREC tracks, where the gold standard is normally developed by annotating few top-ranked documents, retrieved by different submitted systems. In other words, only the standard pooling phase (as described in Phase 3) is applied in TREC tracks.

Given that it is challenging to identify many of the tweets relevant to a topic (as discussed above), annotating only a relatively small pool of documents retrieved by IR methodologies has the potential risk of missing many of the relevant documents which are more difficult to retrieve. We believe that our approach, where the annotators searched through the entire dataset instead of a relatively small pool, is likely to be more robust, and is expected to have resulted in the development of a complete gold standard which is irrespective of the performance of any IR methodology.

4 Baseline Methodologies Submitted to the FIRE 2016 Track

The participants of the FIRE 2016 track were given the tweet collection and the seven topics described earlier.[6] The participants were invited to develop IR

[6] Note that the Twitter terms and conditions prohibit direct public sharing of tweets. Hence, only the tweet-ids of the tweets were distributed among the participants, along with a Python script using which the tweets can be downloaded via the Twitter API.

methodologies for retrieving tweets relevant to each of the seven topics, and submit a ranked list of tweets that they judge relevant to each topic. The ranked list was evaluated based on the gold standard (developed as described earlier).

The track invited three types of methodologies – (i) *Automatic*, where both query formulation and retrieval are automated, and (ii) *Semi-automatic*, where manual intervention is involved in the query formulation stage (but not in the retrieval stage), and (iii) *Manual*, where manual intervention is involved in both query formulation and retrieval stages.

Ten teams participated in the FIRE 2016 Microblog track, submitting 15 runs in total, out of which, one run was fully automatic, while the others were semi-automatic. A summary of the methodologies used by each team is discussed in this section.

dcu_fmt16 [13]**:** This team participated from ADAPT Centre, School of Computing, Dublin City University, Ireland and applied WordNet[7] for query expansion. It submitted the following two runs:

1. *dcu_fmt16_1:* This is an *Automatic* run. The initial query was created from the words in *title* and *narr*, from which the stopwords were removed. For each word in the query thus formed, the expanded query was frormed by adding the synonyms using WordNet. This expanded query was used for retrieval based on the BM25 model [21].
2. *dcu_fmt16_2:* This is a *Semi-automatic* run. An initial ranked list was created from the original topic. From the top retrieved 30 tweets, 1–2 relevant tweets were manually identified from which query expansion was done. This expanded query was expanded further using WordNet like *dcu_fmt16_1* and used for retrieval.

iiest_saptarashmi_bandyopadhyay [2]**:** This team participated from Indian Institute of Engineering Science and Technology, Shibpur, India. It submitted one *Semi-automatic* run described below:

- *iiest_saptarashmi_bandyopadhyay_1:* The relevance score for a given topic-tweet pair was determined by the correlation between the topic words and the tweet. The Stanford NER tagger[8] was applied to identify the LOCATION, ORGANIZATION and PERSON names in the tweets. For each topic, a number of tools (e.g., PyDictionary, NodeBox toolkit etc.) was applied on some manually selected keywords. These were used to find the corresponding synonyms, inflectional variants etc. The bag of words were converted into a vector using Word2Vec package.[9] The relevance score was computed from the correlation between the vector representations of the topic words and the tweet text.

[7] https://wordnet.princeton.edu/.
[8] nlp.stanford.edu/software/Stanford-ner-2015-04-20.zip.
[9] https://deeplearning4j.org/word2vec.

JU_NLP [8]**:** This team participated from Jadavpur University, India. It submitted three *Semi-automatic* runs described as below:

1. *JU_NLP_1:* For each topic, relevant words were hand-picked. Query expansion was done with the synonyms obtained from NLTK WordNet toolkit. Past, past participle and present continuous forms of verbs were obtained using the NodeBox library for Python. For the topics FMT5 and FMT6, Stanford NER tagger was used to identify the location and organization information. GloVe [20] model was trained on the twitter collection. For a tweet-query pair, a tweet vector and a query vector was formed by taking the normalized summation of the GloVe vector of the constituent words. For each pair, the similarity score was computed by the cosine similarity of the corresponding vectors.
2. *JU_NLP_2:* This run is similar to *JU_NLP_1* except that here word bags were split categorically. Then, average similarity between the tweet vector and the split topic vectors was calculated.
3. *JU_NLP_3:* This is identical to *JU_NLP_2*.

iitbhu_fmt16 [22]**:** This team participated from Department of Computer Science and Engineering, Indian Institute of Technology (BHU) Varanasi, India. It submitted one *Semi-automatic* run – *iitbhu_fmt16_1* described as follows:

- *iitbhu_fmt16_1:* The run was generated using the Lucene[10] default similarity model, which combines Vector Space Model (VSM) and probabilistic models (e.g., BM25). StandardAnalyzer was used to handle names, email address, lowercasing of each token and removal of stopwords and punctuations. Manual intervention was done in the the query formulation stage.

daiict_irlab [18]**:** This team participated from DAIICT, Gandhinagar, India and LDRP, Gandhinagar, India. It submitted two *Semi-automatic* runs described as follows:

1. *daiict_irlab_1:* In this run, the 5 most similar words and hashtags from the Word2vec model, trained on the tweet corpus, were added to the original query. Equal weight was assigned to each term.
2. *daiict_irlab_2:* This run was produced in the same way as *daiict_irlab_1* except that different weights were assigned to the expanded terms than the original terms. Higher weights were assigned to the words like *required* and *available*. Query expansion was also done using WordNet.

trish_iiest [9]**:** This team participated from Indian Institute of Engineering Science and Technology, Shibpur, India. It submitted two *Semi-automatic* runs described below:

[10] https://lucene.apache.org/ (2016, August 20).

1. *trish_iiest_ss:* The similarity score between a query and a tweet was computed from the word-overlap between them and then normalized by the query length. For each topic, the query was formed by the nouns, identified by the Stanford Part-Of-Speech Tagger. Higher weight was assigned to the words like *availability* or *requirement*.
2. *trish_iiest_ws:* In this run, the overlap score was calculated using the synsets of each term obtained from WordNet.

nita_nitmz [4]**:** This team participated from National Institute of Technology, Agartala, India and National Institute of Technology, Mizoram. It submitted one *Semi-supervised* run described as below:

- *nita_nitmz_1:* Apache Nutch 0.9 was used for this run to perform search using the different combination of words present in the query. The final result was obtained by merging the results generated by different combinations of query.

Helpingtech [5]**:** This team participated from Indian Institute of Technology, Patna, Bihar, India and submitted the following *Semi-automatic* run (on 5 topics only):

- *Helpingtech_1:* For this run, relationships entities and action verbs were defined through manual inspection. The ranking score was calculated on the basis of the presence of these relationships in the tweet for a given query. Higher consideration was given to a tweet which indicated immediate action than a one which suggested a proposed action for future.

GANJI [15]**:** This team participated from Èvora University, Portugal. It submitted retrieval results for the first three topics only using *Semi-automatic* methodology, described below:

- *GANJI:* Keywords were extracted using Part-of-speech tagger, Word2Vec (to obtain the *nouns*) and WordNet (to obtain the *verbs*). Retrieval was performed on Terrier[11] using the BM25 model. SVM classifier was used to classify the retrieved tweets into *available*, *required* and *other* classes.

relevancer_ru_nl [11]**:** This team participated from Radboud University, the Netherlands and submitted the following *Semi-automatic* run:

- *relevancer_ru_nl:* A tool *Relevancer* was used to generate this run. First, the tweet collection was clustered to identify *coherent* clusters. Each *coherent* cluster was manually labelled by some experts as "relevant" or "not relevant". This training data was fed into a Naive Bayes classifier. The test tweets, which were predicted as relevant by the classifier, were submitted.

From the above descriptions, it is evident that a wide variety of methodologies were applied, including traditional IR methodologies, methodologies using NLP, as well as methodologies based on recent advances like word embeddings.

[11] http://terrier.org.

5 Proposed Methodology

In addition to the methodologies submitted by the FIRE 2016 participants (described in the previous section), we now propose a novel methodology for the microblog retrieval task.

We start by observing that the topics FMT1, FMT2, FMT3, FMT4 and FMT7 are more general, and any tweet containing suitable information would be relevant to these topics. However, the topics FMT5 and FMT6 need some special consideration. In case of FMT5, relevant tweets should have a reference of location (in addition to the context of need and availability), and for FMT6, the relevant tweets must contain the reference to an organization (Government organization/NGO). Hence, we first describe a general IR methodology for the topics FMT1, FMT2, FMT3, FMT4 and FMT7, and then later describe the distinct methodology for FMT5 and FMT6.

5.1 Proposed Methodology for Topics FMT1, FMT2, FMT3, FMT4 and FMT7

For the topics FMT1, FMT2, FMT3, FMT4 and FMT7, we consider three stages in the retrieval – (i) generating a query from the topic, (ii) retrieving and ranking microblogs with respect to the query, and (iii) expanding the query, and subsequently retrieving and ranking microblogs with respect to the expanded query. These steps are described next.

Query generation from topics: For a given topic, we consider the query to be a set of terms (unigrams) extracted from the text of the topic (stated in Table 1). Specifically, terms are extracted from the narrative part of the topic, as follows.

For all the topics, a standard set of English stopwords are ignored, and terms which appear in the narrative of at least four out of the five topics (e.g., 'message', 'mention') are ignored. Additionally, the last sentence of the narrative, which mentions what type of tweets would *not* be relevant, is ignored. Next, an English part-of-speech (POS) tagger[12] is applied on the rest of the narrative text, and *nouns*, *verbs*, and *adjectives* (terms which are tagged as 'NN', 'VB', and 'JJ' by the POS tagger) are extracted.[13] The selected terms are stemmed using the standard Porter stemmer, and the query is considered as a bag of the stemmed terms. Table 3 shows the automatically generated queries (showing the terms obtained after stemming) for the five topics.

Microblog retrieval and ranking: We now describe the proposed methodology of retrieving microblogs for a given query. The tweets are pre-processed by removing a standard set of English stopwords, URLs and punctuation symbols, and all terms whose frequency in the entire corpus is less than 5 are also ignored. The remaining terms are then stemmed using the standard Porter stemmer. Thus, each tweet is considered as a bag/set of terms (stemmed).

[12] The POS tagger included in the Python Natural Language Toolkit was used.

[13] We also tried retrieval with other parts of speech, and observed that forming the query out of nouns, verbs, and adjectives, gives the best retrieval performance.

Table 3. Queries for the five topics FMT1, FMT2, FMT3, FMT4 and FMT7. Each query is a set of unigrams selected from the text of the topics, and then stemmed.

Topic	Query terms
FMT1: What resources were available	Avail, blanket, cloth, distribut, drink, process, shelter, transport, vehicl, volunt
FMT2: What resources were required	Blanket, cloth, distribut, need, process, requir, shelter, transport, vehicl, volunt
FMT3: What medical resources were available	Ambul, avail, blood, doctor, equip, infant, item, medic, medicin, milk, staff, supplementari
FMTT4: What medical resources were required	Ambul, blood, doctor, equip, item, medic, medicin, requir, staff, supplementari
FMT7: What infrastructure damage and restoration were being reported	Commun, connect, dam, damag, electr, hous, internet, mobil, railwai, restor, road, runwai, specif, structur, tower

We employ a simplistic word2vec [17] based retrieval model which is suitable for short documents like tweets. We first trained word2vec over the pre-processed set of tweets.[14] Specifically, the continuous bag of words model is used for the training, along with Hierarchical softmax, and the following parameter values – Vector size: 2000, Context size: 5, Learning rate: 0.05. The word2vec model gives a vector (of dimension 2000, as decided by the parameters) for each term in the corpus, which we refer to as *term-vectors*. The term-vector for a particular term is expected to capture the context in which the term has been frequently used.

For a given query (which we consider as a set of terms, as described above), we derive a *query-vector* by summing the term-vectors of all terms in the query. Since the word2vec term-vectors are additive – i.e., they can be added or subtracted to respectively include or exclude the semantic representations of the corresponding terms [17] – we expect the query-vector to capture the collective context of all the terms in the query. For instance, the sum of the vectors for the terms 'need' and 'food' will capture the context of food being needed.

Similarly, for each tweet, we derive a *tweet-vector* by summing the term-vectors of all terms contained in the pre-processed tweet (as described above). For retrieving tweets for a given query, we compute the cosine similarity between the corresponding query-vector and each tweet-vector, and rank the tweets in decreasing order of the cosine similarity.

Query expansion: Query expansion is the process of adding some relevant terms to the query, in an attempt to improve retrieval performance (especially the recall). We consider the *pseudo (or blind) relevance feedback* setting [16] – after documents are retrieved using a particular query, a small number of the

[14] The Gensim implementation for word2vec was used – https://radimrehurek.com/gensim/models/word2vec.html.

top-ranked documents are assumed to be relevant, and certain terms are selected from the top retrieved documents to add to the query.

We consider two approaches for query expansion – Rocchio expansion with pseudo relevance feedback [16], and the other based on word2vec, as described below. For both approaches, we consider the 10 top-ranked tweets retrieved by the original query, and select $p = 5$ terms from these 10 top-ranked tweets to expand the query.

Rocchio expansion: For each distinct term in the 10 top-ranked tweets retrieved by the original query, the $tf \times idf$ Rocchio scores are computed, where tf is the frequency of the term among the 10 top-ranked tweets, and idf is the inverse document frequency of the term over the entire corpus. and the top p terms in the decreasing order of Rocchio scores are selected for expanding the query.

Query expansion using word2vec: As stated earlier, the query-vector is expected to capture the overall context of a query. After retrieving tweets using the initial query, we identify a set of terms within the 10 top-ranked tweets, which are most related to the context of the query, and use these terms to expand the query. Specifically, we compute the cosine similarity of the query-vector with the term-vector of every distinct term in the 10 top-ranked tweets, and select those p terms for which the term-vector has highest cosine similarity with the query-vector.

Table 4 states the expansion terms (stemmed) obtained by the two strategies for some of the topics.

Table 4. Expansion terms (stemmed) for the initial queries (stated in Table 3), obtained through Rocchio expansion and word2vec-based expansion.

Topic: FMT1	Topic: FMT2	Topic: FMT7
Query expansion by Rocchio method		
Food, medicin, shelter, tent, water	Food, medicin, need, sanit, water	Beyond, damag, hous, repair, road
Query expansion using word2vec		
Biscuit, hygien, medicin, sanit, tem	Biscuit, hygien, medicin, necess, sanit	Cheaper, inspect, partial, scout, sprint

5.2 Proposed Methodology for Topics FMT5 and FMT6

As stated earlier, we adopted a different approach for the topics FMT5 (what were requirement/availability of resources at specific location) and FMT6 (what were the activities of various NGOs/Government organizations). From Table 1, it is evident that the sets of tweets relevant to these two topics are actually subsets of the combined set of tweets relevant to the topics FMT1, FMT2, FMT3 and FMT4. Specifically, if we consider the combined set of tweets relevant to the topics FMT1, FMT2, FMT3 and FMT4, then the subset of these tweets that

contain a specific location would be relevant to FMT5. Similarly, the subset of tweets which contain the name of an organization would be relevant to FMT6.

Hence, we adopted the following strategy for FMT5 and FMT6. First, we considered the combined set of retrieved tweets judged as relevant to the topics FMT1, FMT2, FMT3 and FMT4, by any of the methodologies described earlier, and obtained the top-ranked 4,000 tweets from this set.[15]

Next, we applied the Stanford Named Entity Recognition (NER) tagger[16] on the selected tweets, to identify location and organization references. The Stanford NER tagger labels locations present in a text as LOCATION, and organization names as ORGANIZATION. The tweets which were found to contain a location reference were considered relevant to FMT5, and the tweets which were found to contain the name of an organization were considered relevant to FMT6.

Finally, we rank the tweets that were judged relevant to FMT5/FMT6. A query-vector was formed for FMT5/FMT6 by summing the query-vectors for FMT1, FMT2, FMT3 and FMT4 (considering the expanded queries, following the Rocchio expansion or the word2vec-based expansion strategy). The tweets were ranked according to the cosine similarity of the tweet-vectors and the query-vector, as described earlier.

6 Experimental Results

In this section, we discuss the performance of the methodologies described in the previous sections – the methodologies submitted to the FIRE 2016 Microblog track, and the two proposed methodologies. Ideally, for a given topic, an IR methodology should retrieve only the relevant microblogs (i.e., high precision) and all the relevant microblogs (i.e., high recall). So, we consider the following measures to evaluate the performance of an IR methodology – (i) *Precision at 20* (Prec@20), i.e., what fraction of the top-ranked 20 results are actually relevant according to the gold standard, (ii) *Mean Average Precision* (MAP) considering the full retrieved ranked list.

Table 5 reports the retrieval performance for all the submitted runs, along with our proposed methodologies. It is evident that, among the methodologies which attempted retrieval for all seven topics, our proposed methodologies have outperformed all the other methodologies submitted to the FIRE 2016 Microblog track.

Table 6 represents topic-wise MAP score of submitted and proposed methodologies. The MAP score of proposed methodologies, for a large majority of the topics, are significantly better than all the other methodologies as evident from the Table 6.

In general, we observe that most of the methodologies which performed well, applied query expansion and word embedding techniques (like word2vec and glove). The better performance of such methodologies can probably be explained

[15] We had many ties among the rankings, e.g., the top-ranked tweet for FMT1 and the top-ranked tweet for FMT2 both had same rank.

[16] nlp.stanford.edu/software/Stanford-ner-2015-04-20.zip.

Table 5. Comparison among all the methodologies submitted to FIRE 2016 track, and the two proposed methodologies. Methodologies which attempted retrieval only for a subset of the topics are listed separately at the end of the table.

Methodology	Prec@20	MAP	Type	Method summary
Proposed Methodology 1	0.4428	0.1800	Automatic	Word2vec Expansion and Word2vec Ranking, NER tagger
Proposed Methodology 2	0.4357	0.1829	Automatic	Rocchio Expansion and Word2vec Ranking, NER tagger
dcu_fmt16_1	0.3786	0.1103	Automatic	WordNet, Query Expansion
iiest_saptarashmi_bandyopadhyay_1	0.4357	0.1125	Semi-automatic	Correlation, NER, Word2vec
JU_NLP_1	0.4357	0.1079	Semi-automatic	WordNet, Query Expansion, NER, GloVe
dcu_fmt16_2	0.4286	0.0815	Semi-automatic	WordNet, Query Expansion, Relevance Feedback
JU_NLP_2	0.3714	0.0881	Semi-automatic	WordNet, Query Expansion, NER, GloVe, word bags split
JU_NLP_3	0.3714	0.0881	Semi-automatic	WordNet, Query Expansion, NER, GloVe, word bags split
iitbhu_fmt16_1	0.3214	0.0827	Semi-automatic	Lucene default model
relevancer_ru_nl	0.3143	0.0406	Semi-automatic	Relevancer system, Clustering, Manual labelling, Naive Bayes classification
daiict_irlab_1	0.3143	0.0275	Semi-automatic	Word2vec, Query Expansion, equal term weight
daiict_irlab_2	0.3000	0.0250	Semi-automatic	Word2vec, Query Expansion, unequal term weights, WordNet
trish_iiest_ss	0.0929	0.0203	Semi-automatic	Word-overlap, POS tagging
trish_iiest_ws	0.0786	0.0099	Semi-automatic	WordNet, POS tagging
nita_nitmz_1	0.0583	0.0031	Semi-automatic	Apache Nutch 0.9, query segmentation, result merging
Helpingtech_1 (only 5 topics)	0.7700	0.2208	Semi-automatic	Entity and action verbs relationships, Temporal Importance
GANJI (only 3 topics)	0.8500	0.2420	Semi-automatic	Keyword extraction, Part-of-speech tagger, Word2vec, WordNet, Terrier, SVM classification

Table 6. MAP score of different methodologies, for each of the seven topics.

Methodology	FMT1	FMT2	FMT3	FMT4	FMT5	FMT6	FMT7
Proposed Methodology 1	0.3448	0.2073	0.4217	0.0735	0.0369	0.1022	0.0850
Proposed Methodology 2	0.3322	0.2255	0.4151	0.0768	0.0368	0.0934	0.1004
dcu_fmt16_1	0.0569	0.1730	0.2677	0.0599	0.0306	0.0087	0.1753
iiest_saptarashmi _bandyopadhyay_1	0.1571	0.1234	0.2158	0.1212	0.0290	0.0365	0.1046
JU_NLP_1	0.1530	0.1151	0.2374	0.0905	0.0211	0.0369	0.1014
dcu_fmt16_2	0.0596	0.0853	0.2198	0.0791	0.0274	0.0722	0.0269
JU_NLP_2	0.1055	0.1146	0.1468	0.1047	0.0198	0.0196	0.1057
JU_NLP_3	0.1055	0.1146	0.1468	0.1047	0.0198	0.0196	0.1057
iitbhu_fmt16_1	0.1036	0.2102	0.0275	0.1856	0.0212	0.0054	0.0257
relevancer_ru_nl	0.0913	0.0459	0.0586	0.0036	0.0027	0.0414	0.0409
daiict_irlab_1	0.0257	0.0649	0.0281	0.0502	0.0033	0.0030	0.0176
daiict_irlab_2	0.0190	0.0702	0.0086	0.0415	0.0175	0.0004	0.0175
trish_iiest_ss	0.0234	0.0404	0.0064	0.0270	0.0171	0.0200	0.0079
trish_iiest_ws	0.0134	0.0128	0.0088	0.0074	0.0088	0.0137	0.0042
nita_nitmz_1	0.0000	0.0007	0.0087	0.0090	0.0000	-	0.0000
Helpingtech_1	0.1824	0.2516	0.2201	0.2852	-	-	0.1648
GANJI	0.2270	0.3401	0.1588	-	-	-	-

as follows. Short, informally-written microblogs that are relevant to a topic often do *not* contain key terms that are seemingly important for the topic. For instance, we found that a considerable fraction of the microblogs relevant to a topic *do not contain any of the terms in the corresponding query.* In such scenarios, query expansion is useful for inclusion of relevant terms in the query. Additionally, word embedding methods can better match the semantic context of a microblog with the context of the query, even if the same terms are not present in the two.

7 Conclusion

The primary contribution of the FIRE 2016 Microblog track was the creation of a benchmark collection of microblogs posted during disaster events, and comparison among the performance of various IR methodologies over the collection. Additionally, in this book chapter, we have proposed IR methodologies that have performed better than all the methodologies submitted to the FIRE track.

In future, several extensions of the work can be considered:

- Instead of just considering binary relevance (where a tweet is either relevant to a topic or not), graded relevance can be considered, e.g., based on factors

like how actionable the information contained in the tweet is, how useful the tweet is likely to be to the agencies responding to the disaster, and so on.
- The work described in this chapter considered a static set of microblogs. But in reality, microblogs are obtained in a continuous stream. The challenge can be extended to retrieve relevant microblogs dynamically, e.g., as and when they are posted.

Furthermore, it can be noted that even the best performing methodology described in this book chapter achieved relatively low MAP scores, which highlights the difficulty and challenges in microblog retrieval during a disaster situation. We hope that the test collection developed in this work will help development of better models for microblog retrieval in future.

Acknowledgement. We thank the FIRE organizing committee for allowing us to run the track, and all participating teams for their participation. This research was partially supported by a grant from the Information Technology Research Academy (ITRA), MeITY, Government of India (Ref. No.: ITRA/15 (58)/Mobile/DISARM/05).

References

1. AIDR - Artificial Intelligence for Disaster Response. https://irevolutions.org/2013/10/01/aidr-artificial-intelligence-for-disaster-response/
2. Bandyopadhyay, S.: Correlation distance based information extraction system at FIRE 2016 Microblog Track. In: Working notes for the 2016 Conference of the Forum for Information Retrieval Evaluation (FIRE), CEUR Workshop Proceedings. CEUR-WS.org, December 2016
3. Basu, M., Bandyopadhyay, S., Ghosh, S.: Post disaster situation awareness and decision support through interactive crowdsourcing. In: Proceedings of International Conference on Humanitarian Technology: Science, Systems and Global Impact (HumTech), Procedia Engineering, pp. 167–173. Elsevier (2016)
4. Bhardwaj, P., Pakray, P.: Information extraction from Microblogs. In: Working Notes for the 2016 Conference of the Forum for Information Retrieval Evaluation (FIRE), CEUR Workshop Proceedings. CEUR-WS.org, December 2016
5. Chakraborty, R., Bhavsar, M.: Information Retrieval from Microblogs during natural disasters. In: Working Notes for the 2016 Conference of the Forum for Information Retrieval Evaluation (FIRE), CEUR Workshop Proceedings. CEUR-WS.org, December 2016
6. Cleverdon, C.: The cranfield tests on index language devices. In: Sparck Jones, K., Willett, P. (eds.) Readings in Information Retrieval, pp. 47–59. Morgan Kaufmann Publishers Inc., San Francisco (1997)
7. CrisisLex: Crisis-related Social Media Data and Tools. http://crisislex.org/
8. Dasgupta, S., Kumar, A., Das, D., Naskar, S.K., Bandyopadhyay, S.: Word embeddings for information extraction from tweets. In: Working notes for the 2016 Conference of the Forum for Information Retrieval Evaluation (FIRE), CEUR Workshop Proceedings. CEUR-WS.org, December 2016
9. Ghorai, T.: An information retrieval system for FIRE 2016 Microblog Track. In: Working Notes for the 2016 Conference of the Forum for Information Retrieval Evaluation (FIRE), CEUR Workshop Proceedings. CEUR-WS.org, December 2016

10. Ghosh, S., Ghosh, K.: 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, pp. 56–61. 7–10 December 2016. http://ceur-ws.org/Vol-1737/T2-1.pdf
11. Hürriyetoğlu, A., van den Bosch, A., Oostdijk, N.: Relevant tweet detection in Nepal earthquake with relevancer. In: Working notes for the 2016 Conference of the Forum for Information Retrieval Evaluation (FIRE), CEUR Workshop Proceedings. CEUR-WS.org, December 2016
12. Imran, M., Castillo, C., Diaz, F., Vieweg, S.: Processing social media messages in mass emergency: a survey. ACM Comput. Surv. **47**(4), 67:1–67:38 (2015)
13. Li, W., Ganguly, D., Jones, G.J.F.: Using WordNet for query expansion: ADAPT@ FIRE 2016 Microblog Track. In: Working Notes for the 2016 Conference of the Forum for Information Retrieval Evaluation (FIRE), CEUR Workshop Proceedings. CEUR-WS.org, December 2016
14. Lin, J., Efron, M., Wang, Y., Sherman, G., Voorhees, E.: Overview of the TREC-2015 Microblog Track. In: Proceedings of Text Retrieval Conference (TREC) (2015). http://trec.nist.gov/pubs/trec24/papers/Overview-MB.pdf
15. Lkhagvasuren, G., Gonçalves, T., Saias, J.: Semi-automatic keyword based approach for FIRE 2016 Microblog Track. In: Working Notes for the 2016 Conference of the Forum for Information Retrieval Evaluation (FIRE), CEUR Workshop Proceedings. CEUR-WS.org, December 2016
16. Manning, C.D., Raghavan, P., Schütze, H.: Introduction to Information Retrieval. Cambridge University Press, New York (2008)
17. Mikolov, T., Yih, W., Zweig, G.: Linguistic regularities in continuous space word representations. In: NAACL HLT 2013 (2013)
18. Modha, S., Mandalia, C., Agrawal, K., Verma, D., Majumder, P.: Real time information extraction from Microblog. In: Working Notes for the 2016 Conference of the Forum for Information Retrieval Evaluation (FIRE), CEUR Workshop Proceedings. CEUR-WS.org, December 2016
19. Ounis, I., Macdonald, C., Lin, J., Soboroff, I.: Overview of the TREC-2011 Microblog Track. In: Proceedings of Text Retrieval Conference (TREC) (2011). http://trec.nist.gov/pubs/trec20/papers/MICROBLOG.OVERVIEW.pdf
20. Pennington, J., Socher, R., Manning, C.D.: Glove: global vectors for word representation. In: Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543 (2014). http://www.aclweb.org/anthology/D14-1162
21. Robertson, S.E., Zaragoza, H.: The probabilistic relevance framework: BM25 and beyond. Found. Trends Inf. Retr. **3**(4), 333–389 (2009)
22. Soni, R., Pal, S.: IIT BHU at FIRE 2016 Microblog Track: a semi-automatic Microblog retrieval system. In: Working Notes for the 2016 Conference of the Forum for Information Retrieval Evaluation (FIRE), CEUR Workshop Proceedings. CEUR-WS.org, December 2016
23. Sparck Jones, K., van Rijsbergen, C.: Report on the need for and provision of an ideal information retrieval test collection. Technical report 5266, Computer Laboratory, University of Cambridge, UK (1975)
24. Strohman, T., Metzler, D., Turtle, H., Croft, W.B.: Indri: a language model-based search engine for complex queries. In: Proceedings of ICIA (2004). http://www.lemurproject.org/indri/
25. Tao, K., Abel, F., Hauff, C., Houben, G.J., Gadiraju, U.: Groundhog day: near-duplicate detection on Twitter. In: Proceedings of World Wide Web (WWW) (2013)

26. Twitter Search API. https://dev.twitter.com/rest/public/search
27. Varga, I., et al.: Aid is out there: looking for help from tweets during a large scale disaster. In: Proceedings of ACL (2013)
28. Vieweg, S., Hughes, A.L., Starbird, K., Palen, L.: Microblogging during two natural hazards events: what Twitter may contribute to situational awareness. In: Proceedings of ACM SIGCHI (2010)
29. World Disasters Report 2013 - Focus on technology and the future of humanitarian action (2013). http://www.ifrc.org/PageFiles/134658/WDR2013complete.pdf

http://www.springer.com/978-3-319-73605-1

Text Processing
FIRE 2016 International Workshop, Kolkata, India,
December 7–10, 2016, Revised Selected Papers
Majumder, P.; Mitra, M.; Mehta, P.; Sankhavara, J. (Eds.)
2018, VIII, 219 p. 56 illus., Softcover
ISBN: 978-3-319-73605-1