Information Extraction from Microblog for Disaster Related Event

Rishab Singla, Dhirubhai Ambani Institute of Information and Communication Technology, Gandhinagar, Gujarat, India, singlarishab15@gmail.com

Sandip Modha, Dhirubhai Ambani Institute of Information and Communication Technology,Gandhinagar, Gujarat, India,sjmodha@gmail.com

Prasenjit Majumder, Dhirubhai Ambani Institute of Information and Communication Technology, Gandhinagar, Gujarat, India, prasenjit_majumder@gmail.com

Chintak Mandalia, LDRP-ITR, Gandhinagar, Gujarat, India, chintak.soni75@gmail.com

Abstract. This paper presents the participation of Information Retrieval Lab(IRLAB) at DAIICT Gandhinagar ,India in Data challenge track of SMERP 2017. This year SMERP Data challenge track has offered a task called Text Extraction on the Italy earthquake tweet dataset, with an objective to retrieve relevant tweets with high recall and high precision. In this task, three runs were submitted by us and we describe the different approaches adopted. Initially, we have performed query expansion on the topics using Wordnet. In the first run, we have ranked tweets using cosine similarity against the topics. In the second run, relevance score between tweets and the topic is calculated using Okapi BM25 ranking function and in the third run relevance score is calculated using language model with Jelinek-Mercer smoothing
.

Keywords: Microblog, Information Retrieval, Disaster, Wordnet, BM25

1 Introduction

Microblogs like Twitter can play a very important role in any disaster related event. Twitter has a massive registered user base. As of 2016, Twitter[1] had more than 319 million monthly active users. On the day of the 2016 U.S. presidential election, Twitter proved to be the largest source of breaking news, with 40 million tweets sent by 10 p.m. (Eastern Time) that day. Twitter enables humans to act as a social sensor to the real world. It allows its registered users to post short texts called tweets having upto 140 characters.

[1] https://en.wikipedia.org/wiki/Twitter

Many incidents in the past have proved that social media is the first medium through which news related to a disaster like earthquakes reach the people. Recently, many earthquake incidents have been reported first on Twitter and then on any other media [5]. Twitter can be effectively accessed by an NGO/Government agency to assess the ground reality of the disaster area to assist in their rescue operations.

The motivation of the data challenge track is to promote development of IR methodologies that can be used to extract important information from social media during emergency events, and to arrange for comparative evaluation of the methodologies [1]. The Data challenge track offered two tasks namely Text retrieval in two levels. The track organizers have provided tweet-id of the first day of Italy earthquake in the first level. In the second level, tweet-ids of tweet posted during second day of Italy earthquake, were provided. [1] Track organizer also provided the topics in TREC style for which we have to extract and summarize relevant tweets.

The aim of Text Retrieval sub track is to retrieve top relevant tweets with respect to each of the specified topics with high precision and high recall. The paper is organized as follow; we will discuss related work in section 2. In section 3 we describe tweet dataset. In section 4, we describe the problem statement. In section 5 we discuss our methodology. In section 6, we will present the results and analysis. In section 7 we draw conclusions and discuss future work.

2 Related Work

We started our work by referring TREC MICROBLOG 2015 papers. TREC[2] has started Microblog track since 2011 with objective to explore new IR methodology on short text. CLIP[2] has trained their Word2vec model using 4 years tweet corpus. They used Okapi BM25 relevance model to calculate the score. To refine the scores of the relevant tweets, tweets were rescored using the SVM rank package using the relevance score of the previous stage.

University of waterloo [4] implemented the filtering tasks, by building a term vector for each user profile and assigning different weights to different types of terms. To discover the most significant tokens in each user profile, they calculated pointwise KL divergence and ranked the scores for each token in the profile.

3 Tweet Dataset

SMERP 2017 Track organizers have provided dataset of tweets-ids posted on Twitter during the earthquake in Italy in August 2016 along with a Python script that can be used to download the tweets using the Twitter API [1]. The text retrieval track is offered in two levels, tweets posted on first day and day two and three of Italy earth-

[2] http://trec.nist.gov/

quake will be considered in level-1 and level 2 dataset respectively. They have provided 52469 tweet ids in level-1 and 19751 tweet ids in level-2 along with 4 topics in the TREC format.

4 Problem Statement

Given topics Q = {SMERP-T_1, SMERP-T_2, SMERP-T_3, SMERP-T_4}, and Tweets Dataset T = {T_1, T_2,..,T_n} from the dataset, we have to design a ranking function R: (Q,T) → {R_1,..Rn} which ranks tweets against given topic based upon the relevance score. Ri = {T_1,…T_n} where R_i is the set relevant tweet against i^{th} profile.

5 Our Methodology

Track organizers have given 4 topics according TREC format which consists of title description and narrative. Essentially these topics are our query and will be used interchangeably throughout the paper. In this section, we describe our approach.

5.1 Topic Preprocessing

Topics consist of title which describe the general information need, description and narrative which are sentence and paragraph long content which describe the overall picture.

<top>
<num> Number: SMERP-T1

<title> WHAT RESOURCES ARE AVAILABLE

<desc> Description:
Identify the messages which describe the availability of some resources.

<narr> Narrative:
A relevant message must mention the availability of some resource like food, drinking water, shelter, clothes, blankets, blood, human resources like volunteers, resources to build or support infrastructure, like tents, water filter, power supply, etc. Messages informing the availability of transport vehicles for assisting the resource distribution process would also be relevant. Also, messages indicating any services like free wifi, sms, calling facility etc. will also be relevant. In addition, any message or announcement about donation of money will also be relevant.However, generalized statements without reference to any resource would not be relevant.

</top>

To covert topic into query, we have first removed stopwords. We run Stanford POS tagger[3] on topics. All keyword with the noun and verb labels are extracted and added to the query. We believe that the topic are extremely vague so human intervention is required to build the query

5.2 Query Expansion

We have used lexical database WordNet[4] for query/topic expansion. It puts english words into sets of synonyms called synsets. For each term in a query, we have extracted top 2 synonyms from WordNet and added to the query. We have set equal term weight for original term and the expanded term.

5.3 Tweet Preprocessing

After downloading the tweets, non-English tweets were filtered out. Tweet includes smiles, hashtags, and many special characters. We did not consider retweets or tweets with only hashtags, emoticons or special characters. Also, we ignored tweets with less than 5 words and removed all the stopwords and non-ASCII character from the tweet.

5.4 Relevance Score

We have submitted two runs in the first level and three runs in the second level for the Text Retrieval track with different retrieval techniques. Further, we will discuss each technique.

Relevance score using Cosine similarity.

In the first run, we used cosine similarities between tweets and expanded topic to calculate relevance score.

[3] http://nlp.stanford.edu:8080/parser/
[4] https://wordnet.princeton.edu/

$$CosineSimilarity(A,B) = \vec{A} * \vec{B} / ||\vec{A}|| * ||\vec{B}||$$

Tweet Relevance score using Okapi BM25 model

In the second run, to calculate relevance score between tweets and expanded query, we have used. Score is defined as follows.

$$BM25 = \sum_{i=1}^{W} \frac{TF(i)(1+k)}{TF(i)+k(1-b+b\frac{DL}{avgDL})} IDF(i)$$

$$IDF(i) = \frac{\log(\frac{N-n+1}{n})}{\log(N)}$$

We have set BM25 model parameter b=0.75,k1=0.2.

Tweet Relevance score using Language Model.

In the third run, we have indexed all the tweets in Lucene[5].Language model with Jelinek-Mercer smoothing was used to retrieve relevant tweets depending on the query. We set a threshold for finding out if a tweet is relevant to a particular topic. The relevance threshold set was 24. The parameter λ was set to 0.1.

[5] https://lucene.apache.org/core/

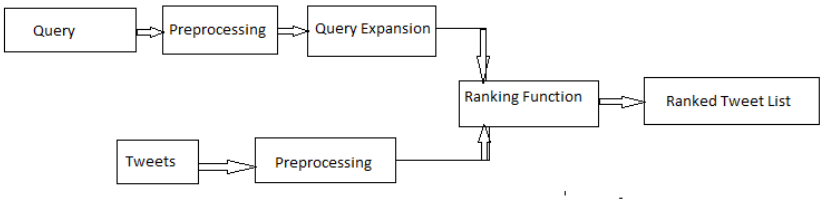

Fig. 1. Methodology Flowchart

6 Results

SMERP Track organizers have used standard TREC metrics like Bpref,Precision@20,Recall@1000 and MAP to evaluate the runs submitted by all teams. Bpref is used as a primary metric to rank all teams. Table 1 and Table 2 show our result in both levels. In level 1, we have achieved higher Recall@1000 compared to top team dcu_ADAPT_run2. However, our Bpref was substantially lower than dcu_ADAPT_run2. In the second run, we have achieved Precision@20,Recall@1000 and MAP better than dcu_ADAPT_run2 but we have reported Bpref substantially lower. We will investigate poor Bpref in future.

Table 1. Task-1 (extraction) result level-1

Sr_no	Run-id	Run type	Bpref	Precision@20	Recall@1000	MAP
1	dai-ict_irlab_2	Semi-automatic	0.3171	0.2250	0.3171	0.0417
2	dai-ict_irlab_1	Semi-automatic	0.3074	0.2125	0.3015	0.0391
3	Toprun dcu_ADAPT_run2	Fully-automatic	0.6170	0.4125	0.1794	0.0517

Table 2. Task-1 (extraction) result level-2

Sr_no	Run-id	Run type	Bpref	Precision@20	Recall@1000	MAP
1	dai-ict_irlab_l2_2	Semi-automatic	0.2869	0.3750	0.2869	0.0635
2	dai-ict_irlab_l2_1	Semi-automatic	0.2869	0.2875	0.2869	0.0571
3	dai-ict_irlab_l2_3	Semi-automatic	0.1204	0.3000	0.1204	0.0433
4	Top run dcu_ADAPT_run2	Fully-automatic	0.7767	0.2125	0.2378	0.0600

7 Conclusions And Future Work

In this paper, we have applied three different retrieval technique namely Okapi BM25, cosine similarities and language model with Jelinek-Mercer smoothing for extraction. Our results show that BM25 model outperforms the other methods in terms of Bpref, Precision@20, Recall@1000 and mean average precision(MAP). We have also concluded that our system has reported poor Bpref score in both the levels which will be investigated further. We also note that topics are more like a question so we have to consider text features like Named entity and verb phrase or relation in the ranking score in addition to raw tweet text. Further on, a ranking system based on deep neural network and logistic regression could be looked at for better results.

8 References

1. SMERP ECIR 2017 guidelines, http://www.computing.dcu.ie/~dganguly/smerp2017/
2. Bagdouri, M., Oard, D.W.: CLIP at TREC 2015: Microblog and LiveQA. In :TREC (2015)

3. Tan, L., Roegiest, A. and Clarke, C.L.: University of Waterloo at TREC 2015 Microblog Track. In : TREC (2015).
4. Tan, L., Roegiest, A., Clarke, C.L. and Lin, J.: Simple dynamic emission strategies for microblog filtering. In : Proc. 39th International ACM SIGIR conference on Research and Development in Information Retrieval , pp. 1009-1012. ACM (2016)
5. Sakaki, T., Okazaki, M. and Matsuo, Y.: Earthquake shakes Twitter users: real-time event detection by social sensors. In: Proc. 19th international conference on World wide web, pp. 851-860. ACM (2010)