Real Time Information Extraction from Microblog

Sandip Modha
DAIICT Gandhinagar
Gujarat-382007, India

Chintak Mandalia
LDRP Gandhinagar
Gujarat-382015, India

Krati Agrawal
DAIICT Gandhinagar
Gujarat-382007, India

Deepali Verma
DAIICT Gandhinagar
Gujarat-382007, India

Prasenjit Majumder
DAIICT Gandhinagar
Gujarat-382007, India

ABSTRACT

This paper present the participation of Information Retrieval Lab(IR LAB DA-IICT Gandhinagar) in FIRE 2016 Microblog Track. The main objective of the track is to identify Information Retrieval methodologies to retrieve important information from Twitter posted during the disasters. We have submitted two runs for this track. In the first run, daiict_irlab_1, we have expanded topic term using Word2vec model trained by the tweet corpus provided by the organizer. Relevance score between tweet and corpus are calculated by Okapi BM25 model. Precision@20 ,primary metric, for this run, is 0.3143. In the second run,daiict_irlab_2, we have set different weight for original term and expanded topic term, we achieve precision@20 around 0.30.

1. INTRODUCTION

Social media, like Twitter, is a massive source of real-time information. Twitter is, one of the popular micro blogging website, which has massive user-generated content due to its large number of registered user. During the disaster, Twitter proved its importance on many occasions.

In the FIRE 2016 Microblog track [2], a large set of micro blogs (tweets), posted during a Nepal earthquake, was made available by track organizer, along with a set of topics (in TREC format). Each 'topic' identified by a broad information need during a disaster, such as âĂŞ what resources are needed by the population in the disaster affected area, what resources are available, what resources are required / available in which geographical region, and so on. Specifically, each topic will contain a title, a brief description, and a more detailed narrative on what type of tweets will be considered relevant to the topic.

2. RELATED WORK

We started our work by referring TREC MICROBLOG 2015 papers [1, 5, 4].

CLIP [1] 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. Then Novelty Detection is done, where the tweets which are not useful are discarded, this is done using Jaccard similarity.

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.

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

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

Figure 1: BM25

3. PROBLEM STATEMENT

Given a topic Q_v=$\{FMT_1, ..., FMT_7\}$, representing different information needs, given corpus of tweets T=$\{t_1, t_2, ..t_n\}$, we need to compute the relevance score between tweets and topics.

$$R\_score = f(T, Q)$$

4. OUR APPROACH

In this section, we discuss the architecture of the proposed system.

4.1 Topic Pre-processing

FIRE 2016 Microblog track has given 7 topics. Essentially these topics are our query. We converted these topics into the query by removing stop words and consider Noun proper noun and verb using Stanford POS tagger.

4.2 Topic (Query) expansion

We have trained Word2vec model [3] using the corpus provided by an organizer to expand topic term. We find 5 similar words and hash tag. We set equal weight for each term in the first run(daiict_irlab_1). In the second run, we have set weight different weight for original terms and expanded terms. Words like required and available have been expanded with their synonyms using WordNet and assigned more weights.

4.3 Tweet Pre-processing

Run Id	Precisio@20	Recall@1000	MAP@1000	Overall MAP
daiict_irlab_1	0.3143	0.0729	0.0275	0.0275
daiict_irlab_2	0.3000	0.0704	0.0250	0.0250

Table 1: Official Results as declared by track organizer

Run Id	Precisio@20	Recall@1000	MAP@1000	Overall MAP
daiict_irlab_1	0.3143	0.1499	0.0638	0.0638
daiict_irlab_2	0.3000	0.1528	0.0625	0.0625

Table 2: Post Evaluation results on top 1000 tweet

In this step, non-English tweets were filtered out. Tweet includes smileys, hashtags, and many special characters. We did not consider retweets and tweet with only hashtag or emoticon or special characters. We also ignored the tweet with less than 5 words and removed all the stop words from the tweet.

4.4 Query Normalization

In this step, Title and description were merged so as to make topics more informative. To increase the relevance, topics were also pre-processed topics by converting all alphabets to small case and expanding the abbreviations. Example: NYC- New York City. Also, topics were stemmed. Eg: behaving was converted to behave.

4.5 Relevance Score

In this phase, we have calculated relevance score between tweets and topics, In the first run, we kept same weight for original term and expanded term. In the second run, we set weight 2 for the original term in the topics and 1 for expanded term. We have used Okapi BM25 model for calculating relevance score between expanded topics and tweets.

$$R\_score = BM25\_Sim(Q_exp, T)$$

5. RESULT

We misunderstand the track guideline. We had sent only top 100 tweets for each topic. So our Precision@20 is on the line with other participant but other metric were substantially lower. Table 1 represents the result declared by track organizer. After Getting Gold Standard data from track organizer, we again perform experiments. Table 2 shows the result of top 1000 tweets for each topic.

6. CONCLUSION

We have submitted 2 runs in FIRE Microblog track. In the first run, we expanded topic term by training Word2vec model with corpus provided by track organizer. We have calculated relevance a score between expanded topic term and tweet using Okapi BM25 model. We have kept the same weight of the original term and expanded term. In the second run, we set weight of the original term and expanded the term in the ratio 2 :1. We have put more weight on a word like âĂIJavailableâĂİ, âĂIJrequiredâĂİ. After analyzing the results, we conclude that by changing the weight for the original term and the expanded term does not improve Precision@20 but actually have some adverse effect. However, Recall@1000 improves approximately 2%.

7. REFERENCES

[1] M. Bagdouri and D. W.Oard. CLIP at TREC 2015: Microblog and LiveQA. In *Proc. TREC 2015*, 2015.

[2] 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, 2016.

[3] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *CoRR, abs/1301.3781*, 2013.

[4] L. Tan, A. Roegiest, and C. L. Clarke. University of Waterloo at TREC 2015 Microblog Track. In *Proc. TREC 2015*, 2015.

[5] X. Zhu et al. NUDTSNA at TREC 2015 Microblog Track. In *Proc. TREC 2015*, 2015.