Summarizing Disaster Related Event from Microblog

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

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

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

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

Abstract. The Information Retrieval Lab at DA-IICT India participated in text summarization of the Data Challenge track of SMERP 2017. SMERP 2017 track organizers have provided the Italy earthquake tweet dataset along with the set of topics which describe important information required during any disaster related incident. The main goal of this task is to gather how well the participant's system summarizes important tweets which are relevant to a given topic in 300 words. We have anticipated Text summarization as a clustering problem. Our approach is based on extractive summarization. We have submitted runs in both the levels with different methodologies. We have done query expansion on the topics using Wordnet. In the first level, we have calculated the cosine similarity score between tweets and expanded query. In the second level, we have used language model with Jelinek-Mercer smoothing to calculate relevance score between tweets and expanded query. We have selected tweets above a relevance threshold which are the initial candidate tweets for the summarization of each query. To ensure novelty, Jaccard Similarity is used to create a cluster for each topic. We have reported results in terms of ROGUE-1, ROGUE-2, ROGUE-L and ROGUE-SU4.

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

1 Introduction

Microblogs, like Twitter, provide a unique crowdsourcing platform where people across the world can post their opinions or observations about real world events. Twitter is the real time data source which has massive user-generated content. Since tweets are posted by multiple users with diverse views, many tweets have redundant content. Due to enormous volume of the tweets, tweet visualization is the biggest challenge. We can address this challenge by creating a summary from relevant tweet with respect to given topic.

The aim of the Text summarization Data Challenge Track is to evaluate and benchmark different summarization systems on standard social media dataset. The text summarization track is offered in two levels. In the first level, tweets which are posted on the first day of the earthquake in Italy were provided. Tweets posted on second and third day of the Italy earthquake were provided in the second level.

2 Related Work

Summarization methods can be divided into two types (i) Extractive Summarization (ii) Abstractive summarization. We have focused on extractive summarization. Basically, Extractive Summarization methods are further divided into 3 types which are (i)) graph based (ii) cluster based (iii) Centroid based.

TREC[1] has started Microblog track since 2011 with an adhoc retrieval task and converged it into real time summarization in 2016. CLIP [2] used a word embedding technique to expand query. They have used BM25 model to calculate relevance score between tweets and query. For summarization, they used jaccard similarity across relevant tweets. Luchenet. al [4] used simple keyword matching technique which assigns more weight to the original term compared to the expanded term. For summarization, they have used simple word overlap.

3 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 need to compute the relevance score between tweets and topics in order to create topic-wise summary S = <S_{Q1}....S_{Qn}>.Where S_{Qi} is the set of topic-wise relevant and novel tweets. We can model topic specific summary as below.

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

$SQ1=<T_1,T_2,\ldots..,T_n>$ where $T_i,T_j \in T$

For given topic, Relevance score between tweet and topic must be greater than specified threshold T_{rel}. In addition to this, these tweets should be novel i.e. similarity between all tweet of the summary should less that the novelty threshold T_{nov} .if any tweet T_i is included in the summary for a particular topic then it should satisfy the following constraints.

- Length of summary of profile(S_{Qi}) <= 300 word
- Relevance score(t_i, Q_i) >T_{rel}
- Sim(t_i ,t_j) <T_{nov} for all $t_j \in S_{Qi}$

4 Our Methodology

4 topics have been provided in TREC format by the track organizers. The topics consist of a title, description and a narrative. The topics might be referred to as queries in the paper. Further, we elaborate our approach.

4.1 Topic Preprocessing

Topics consist of a title in which the general information needed is given. A description, which is sentence long and a narrative, the content of which is paragraph long gives an elaborate picture of the topic.

<top><num> Number: SMERP-T4

<title>WHAT ARE THE RESCUE ACTIVITIES OF VARIOUS NGOs / GOVERNMENT ORGANIZATIONS

<desc> Description:
Identify the messages which describe on-ground rescue activities of different NGOs and Government organizations.

<narr> Narrative:
A relevant message must contain information about relief-related activities of different NGOs and Government organizations engaged in rescue and relief operation. Messages that contain information about the volunteers visiting different geographical locations would also be relevant. Messages indicating that organizations are accumulating money and other resources will also be relevant. However, messages that do not contain the name of any NGO / Government organization would not be relevant.

</top>

The topic to query conversion starts with removal of stopwords. We run Stanford POS tagger[2].The noun and verb labeled keywords are extracted and added to the query. We believe that topics are vague so by human intervention, the query is built.

4.2 Topic Expansion

We have used a lexical database WordNet[3] for topic expansion which puts English words into sets of synonyms, synsets. The top two synonyms are extracted and added to the query using Wordnet.

4.3 Tweet Filtering

After downloading the tweets, only English tweets were worked on. Further, retweets and tweets with only hashtags, emoticons or special characters were not considered. Also, tweets with less than 5 words were ignored. We removed all the stopwords and non-ASCII character from the tweets.

4.4 Relevance Score

We have used cosine similarity to calculate the relevance score between tweet and expanded query in the first level. In the second level we have retrieved relevant tweets using language model with Jelinek-Mercer smoothing with parameter λ=0.1.

4.5 Novelty Detection

Tweets are posted by many users at different times from different parts of the world. To create the text summary from the tweets is a challenging task. Ideally, summary should include all relevant tweets with constraint that it should not include redundant

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

information. Tweet summarization is a multiple document summarization problem. Each tweet can be considered as a single document.

To create the summary, we have selected top tweets from each topic whose relevance score is greater than specified relevance threshold T_{rel}. We have empirically set value of T_{rel}. Now for the next eligible tweet, we calculate it's similarity with tweets already added in the summary so as to ensure novelty between them. Again a Jaccard-threshold t_{nov}=0.6 was decided empirically and tweets below it were added into summary. Lower the similarity score, greater is the dissimilarity ensuring more novelty.

$$J(A,B) = (A \cap B)/(A \cup B)$$

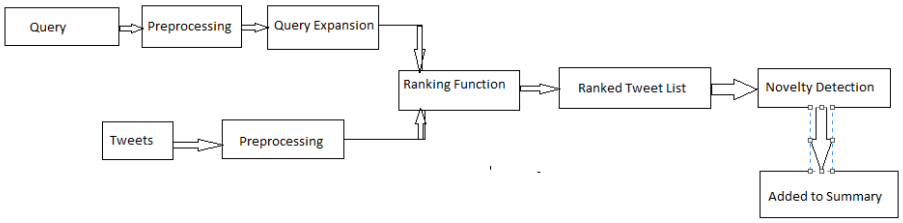

Fig.1. Methodology Flowchart

5 Results

In both levels, our summarization method remain same. However, we have used different tweet retrieval techniques. SMERP 2017 track organizers have considered ROUGE-L as primary metric to evaluate performance of all the runs. The following tables show our results in comparison with the top run.

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

Sr no	Run-id	Run type	Re-call(ROUGE-1)	Re-call(ROUGE-2)	Re-call(ROUGE-L)	Re-call(ROUGE-SU4)
1	daiict_irlab_2	Semi-automatic	.3309	.1543	.3085	.1055
2	Top run IIEST	Semi-automatic	.5109	.2824	.4885	.2329

Table 2. Task-2 (summarization) result level-2

Sr no	Run-id	Run type	Re-call(ROUGE-1)	Re-call(ROUGE-2)	Re-call(ROUGE-L)	Re-call(ROUGE-SU4)
1	dai-ict_irlab_sum m_l2	Semi-automatic	.3515	.1297	.3254	.1194
2	Top run IIEST	Semi-automatic	.5540	.2436	.5142	.2864

6 Conclusions And Future Work

In this paper, we have implemented a method based on extractive summarization. Table 1 and Table 2 show that our results are comparatively lower than IIEST. In the future we will investigate our underperformance and will carry out post-hoc/ error analysis. We would like to design a summarization system based on deep neural network and logistic regression.

7 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)