International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 3 Issue: 2 427 - 430

Twitter Event Summarization Using Phrase Reinforcement Algorithm and NLP Features

Mr. G. S. Mane

Research Scholar, Computer Science & Engineering Department
Walchand Institute of Technology, Solapur

Mrs. A. R. Kulkarni

Assistant Professor, Computer Science & Engineering Department,
Walchand Institute of Technology, Solapur

Abstract-Now a day's social networking sites are the fastest medium which delivers news to user as compare to the news paper and television. There so many social networking sites are present and one of them is Twitter. Twitter allows large no. of users to share/post their views, ideas on any particular event. According to recent survey daily 340 million Tweets are sent on Twitter which is on a different topic and only 4% of posts on Twitter have relevant news data. It is not possible for any human to read the posts to get meaningful information related to specific events. There is one solution to this problem, i.e. we have to apply Summarization technique on it. In this paper, we have used an algorithm which uses a frequency count technique along with this we have also used some NLP features to summarize the event specified by the user. This automatic summarization algorithm handles the numerous, short, dissimilar, and noisy nature of tweets. We believe our novel approach helps users as well as researchers.

Keywords: *Phrase Reinforcement Algorithm (PRA), Twitter API, Twitter, Natural Language Processing (NLP), Textual Entailment, Word Sense Disambiguation, WordNet.*

*****

I. INTRODUCTION

Now a day's Twitter has been cited as breaking many important events before traditional media such as the attacks in Mumbai. With important news being sent in real time to Twitter, it is important that users of Twitter are able to find these important tweets as they occur without having to sort through all the other irrelevant information being posted. Summarization of Twitter events helps to fight with this problem.

Social Networking is a lightweight and easiest form of communication. Micro blogging sites such as Facebook, Google+ and Twitter has become omnipresent in its use with over 4 billion mobile devices worldwide of which over 1 billion support smart services.

According to recent survey daily 340 million Tweets are sent on Twitter which is on a different topic such as Sport related event, Tweets by News agencies, Business news, Tweets posted celebrity etc. It is difficult to find Tweets related to a particular topic. It is not possible for humans to read each and every Tweet to get correct and accurate information related to a specified topic.

The texts from News articles, Books, Research paper, etc. are usually formal writings and have the highest language quality. On the other hand, the language of tweets is highly noisy, spelling and grammar mistakes. Tweets contain Typos, abbreviations, phonetic substitutions, ungrammatical structures and emoticons, etc. Due to the above characteristics, text summarization techniques in general may not adapt well to the Twitter text.

Following are some problem the user has to tackle while reading Tweets about any event:

1. Language issue
2. Duplicate and noisy Tweets
3. Some of the users are trying to divert the event
4. Tweets related to specific topic are not in serial order most of the time
5. Spam-tweets

One solution to these problems is Twitter API. Twitter allows only authorized user to access and use functionality provided by the Twitter API. Twitter API allows user to retrieve Tweets related to the specified topic, event etc... It also provides some functions to filter the Tweets such as removing non-English Tweets, getting tweets related to specific event, etc. Spam-tweets can be easily removed since they almost always have a URL in them. The Twitter API allows user to get only 100 tweets per day. But only 100 tweets are not sufficient to summarize any event and to collect more tweets we have to wait for some days. Instead of this we can use Twitter Stream API to download tweets related to specific events because Twitter Stream API allows downloading infinite no. of tweets. If there is a huge amount of tweets are available, then helps to generate a better summary.

II. LITERATURE REVIEW

Xiaobin Li, Stan Szpakowicz and Stan Matwin[10] has presented A WordNet based Algorithm for Word Sense Disambiguation. In this paper they proposed an algorithm for automatic word sense disambiguation based on lexical knowledge contained in WordNet and on the results of surface-syntactic analysis. The algorithm is designed to support text analysis with minimal pre-coded knowledge, although the algorithm is assumed to aim at word sense disambiguation of noun objects in a text; in fact, it can be easily transformed to cover some other parts of speech in a text. Their approach focuses on two parts the full utilization of the important relationships between words in WordNet and the exploration of WSD heuristic rules based on the semantic similarity between words.

Joel Judd and Jugal Kalita[7] has presented the Better Twitter Summaries. In this technique they are trying to improve the summary produced by PRA. The idea behind creating the desired summary is to parse the "raw" summary and build dependencies between the dependent and governor words in each summary. We perform parts of speech tagging and obtain lists of governing and dependent words. The Stanford Core NLP parser was used to build the lists of governor and dependent words. They show the PR Algorithm can be improved by taking into account governor-dependency relationships among the constituents.

Hassan Sayyadi,Matthew Hurst and Alexey Maykov[5] has presented Event Detection and Tracking in Social Streams. In this paper they propose a new algorithm for event detection using the co-occurrence of keywords. In our community detection algorithm, nodes can fall into different communities as a word or phrase can be in keywords list of more than one event. In the current version of our algorithm we count all keywords in one community as keywords for the event, though a subset of keywords may be better, especially in cases where the number of nodes is large.

Gulab R. Shaikh and Digambar M. Padulkar[3] has presented Template Based Abstractive Summarization of Twitter Topic with Speech Act. In this paper they proposed work, speech act-guided summarization approach is used to generate a summary of the twitter trending topic. With the recognized speech acts, the next step is to extract key words and phrases from tweets to generate abstractive summaries. The extracted key terms are then ranked and inserted into special summary templates designed for speech acts, by using ngram selection algorithm.This system is designed to accommodate the numerous, short, dissimilar, and noisy nature of the tweets. This proposed approach makes a good work contribution to the summarization community.

Prodromos Malakasiotis and Ion Androutsopoulos[9] has presented Learning Textual Entailment using SVMs and String Similarity Measures. They proposed a textual entailment recognition system that relies on SVM s whose features correspond to string similarity measures applied to the lexical and shallow syntactic level. They have suggested two additional possible improvements: applying partial matching to all of the string pairs and investigating other feature selection schemes. In future work, they also plan to exploit WordNet to capture synonyms, hypernyms, etc.

Chin-Yew Lin[8] has presented ROUGE : A Package for Automatic Evaluation of Summaries. ROUGE stands for Recall-Oriented Understudy for Gisting Evaluation. In this paper, they introduced ROUGE, an automatic evaluation package for summarization, and conducted comprehensive evaluations of the automatic measures included in the ROUGE package using three years of DUC data. This paper introduces four different R OUGE measures: ROUGE -N, ROUGE-L, ROUGE-W, and ROUGE–S. They had shown how to achieve high correlation with human judgments in multi-document summarization.

III. METHODOLOGY

Phrase Reinforcement Algorithm

The Phase Reinforcement Algorithm works as follows. The algorithm begins with a starting phrase, which is the event for which one desires to generate a summary. Given the starting phrase, the PR algorithm submits a query to Twitter.com for a list of posts that contain the phrase. Given the returned set of posts, the algorithm next filters the posts to remove any spam or irrelevant posts.

Filtering is an important step since spam and other irrelevant posts can mislead the PR algorithm into summarizing the spam instead of the desired content. In this we first remove those tweets which contain bad words. We remove any non-English posts as well as duplicate posts. After this we have to remove stop words from Tweet Collection.

Once we have a set of relevant posts (training posts), the PR algorithm formally begins. The central idea of the PR algorithm is to build an ordered acyclic graph of all the words within the set of training posts. While constructing graph we check that whether that word is already present in the graph or not. If the word is already present we increment the weight of that word otherwise we create a new with initial weight one.

Phrase Reinforcement Algorithm[2] is referred from the paper Automatic Summarization of Twitter Topics and added two NLP features in it.

Following two Natural Language Processing (NLP) we are using along with the algorithm:

Word sense disambiguation[10] (WSD) is the ability to identify the meaning of words in context. Given a set of words (e.g., a sentence or a bag of words), a technique is applied, which makes use of one or more sources of knowledge to associate the most appropriate senses with words in context. To determine the sense of words we can use machine readable dictionaries, semantic networks.

Consider the following sentences:

a) I can hear bass sounds.

b) They like grilled bass.

In above two sentences the word bass is used for different purposes, i.e. a low-frequency tones and a type of fish, respectively. We can use this feature while assigning weight to the words.

Textual entailment[9] (TE) challenge, which focuses on detecting semantic inference, has attracted a lot of attention. Given a text T (sentences) and a hypothesis H (one sentence), the goal is to detect if H can be inferred from T.

Consider the following sentences:

T: The drugs that slow down or halt Alzheimer's disease work best the earlier you administer them.

H: Alzheimer's disease is treated using drugs.

The above two sentences shows a correct entailment pair.

1. Collect The tweets fro Twitter database for event specified by user and store them into a file.
2. After collecting tweets apply filtration techniques, i.e. removes tweets which contain bad words, remove duplicate tweets, spam tweets and remove stop words.
3. Then count the frequency of each word form set of tweets by using Word sense disambiguation.
4. Get top ten words from frequency count and take only those tweets which contain one of the word from those ten words and transfer only those tweets for further processing.
5. Once again repeat step 3and 4.
6. Then apply Textual Entailment technique on summary given by PR algorithm.

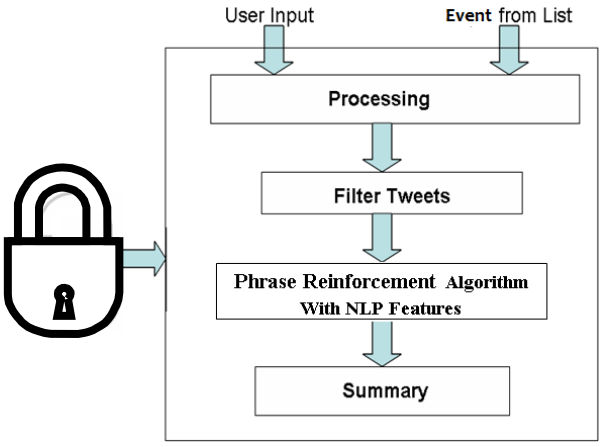

Figure. Conceptual Overview of Proposed System

In the first step we take input from user, i.e. Twitter event name if the user knows the event name and if not, then the user can select can event from the list. The event list contains top 10 current trends from Twitter for the selected region. After selecting input we collect tweets from the Twitter site by using Twitter Stream API and store them into a file for further use.

In the second step we apply filtration technique on collected tweets for removing Duplicate and noisy, irrelevant tweets, Spam-tweets, tweets which contain bad words, etc. After this we remove stop words from the filtered tweets to apply Phrase Reinforcement algorithm (PRA).

In the third step we apply the Phrase Reinforcement algorithm and Natural Language Processing (NLP) features to generate summary related to the specified event. In NLP features we are using Word sense disambiguation and Textual Entailment technique. Word sense disambiguation is used while calculating weight of word. Then we apply Textual Entailment technique on the output of the Phrase Reinforcement algorithm to find out sentences which have strong relation between them. This is done with the help of one of the similarity measure.

The fourth step shows actual output, i.e. summary of a specified event.

REFERENCES

[1] Dehong Gao, Wenjie Li, Xiaoyan Cai, Renxian Zhang, and You Ouyang, Sequential Summarization: A Full View of Twitter Trending Topics in IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 22, NO. 2, FEBRUARY 2014.

[2] B. Sharifi, M.-A.Hutton, and J. K. Kalita, Automatic summarization of Twitter topics in Proc. National Workshop Design Anal. Algorithms, 2010.

[3] Gulab R. Shaikh, Digambar M. Padulkar ,Template Based Abstractive Summarization of Twitter Topic with Speech Act by Asst. Prof., Department of CSE, VPCOE Baramati,Pune, India, India in June 2014.

[4] Renxian Zhang, Wenjie Li, Dehong Gao,and You Ouyang, Automatic Twitter Topic Summarization With Speech Acts in IEEE TRANSACTIONON AUDIO, SPEECH AND LANGUAGE PROCESSING, VOL. 21, NO. 3, MARCH 2013.

[5] Hassan Sayyadi,Matthew Hurst and Alexey Maykov, Event Detection and Tracking in Social Streams. In Proceedings of ICWSM, 2009.

[6] J. Nichols, J. Mahmud, and C. Drews, Summarizing sporting events using Twitter in Proc. IUI-12, 2012.

[7] Joel Judd and Jugal Kalita, Better Twitter Summaries. HLT-NAACL 2013: 445-449.

[8] Chin-Yew Lin, ROUGE : A Package for Automatic Evaluation of Summaries. In Proceedings of Workshop on Text Summarization Branches Out, Post-Conference Workshop of ACL 2004, Barcelona, Spain.

[9] Prodromos Malakasiotis and Ion Androutsopoulos Learning Textual Entailment using SVMs and String Similarity Measures by Department of Informatics, Athens University of Economics and Business Patision 76, GR-104 34 Athens, Greece.

[10] Xiaobin Li, Stan Szpakowicz and Stan Matwin, A WordNet-based Algorithm for Word Sense Disambiguation. In Proceedings of the 14th International Joint Conference on Artificial Intelligence.