Volume 5, Issue 10, October-2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Automatic Twitter Summarization

Sudhanshu Shiwarkar, Indraneel Deshmukh, Nikhil Patil, Akash Thanke
Savitri Bai Phule Pune University, B.E.Department of Computer Engineering, Rmd Sinhgad School of Engineering, Warje, Pune, Maharashtra, India

Abstract- Now a days social networking sites are the fastest medium for sharing information among the users as compared to television, phone calls and newspapers. Twitter is one of the many social networking sites present. Twitter allows large number of users to share their views, ideas on any particular events. Recent studies indicate that daily 340 million tweets are posted on twitter on different topics and only 4% of posts on twitter have specific or relevant data. It is not possible for any human to read the post to get meaningful data. The solution to this problem is that we have to apply summarization technique on it. To address this challenge, we propose a novel speech act guided summarization approach. In this paper we use speech act recognition technique and data sets. For extraction of key words and phrases we propose a round robin algorithm to generate template basedsummaries.

Keywords- abstractive, summarization, key word/phrase extraction, speechact

I. INTRODUCTION

Recent studies show that social networking website plays the necessary role in human life like Twitter. Twitter offers additional info of current world events like News or current affairs. With the assistance of Twitter, a user will produce and share concepts and knowledge instantly, with none barrier. however there's one drawback daily five hundred million Tweets are posted on Twitter and that they don't seem to be in serial order. summarisation of Twitter events helps to fight with thisdrawback.

Tweets below a Trending Topic contain a large kind of helpful info from several views regarding necessary events going down within the world. Basically, a outline that gives representative info of topics with no redundancy and literate sentences would be mostpopular.

In this paper, we have a tendency to specialise in the matter of topic report in Twitter, that aims to produce a brief and compact outline for a set of tweets on similar or similar topics. we have a tendency to take individual tweets because the basic constituents to compose the outline. Here, tweets square measure to an exact extent analogous to sentences in ancient extractive document report, that has been extensively studied in pastdecades.

The most original a part of our approach is that the use of speech acts, that capture the common grounds of tweets from a communicative perspective. once act with tweets, users could share info, raise queries, create suggestions, specific sentiments, etc. that square measure allinstancesof ―speechacts‖. Every tweet gives information about different human activities like the ―question‖ or ―suggestion‖for the different posts. Unless in a very few cases (e.g., employing a human activity hash tag like #question), users don't report the speech acts they're playacting oncetwittering.

II. LITERATURE SURVEY

Sr. No.	Title of Existing System or Paper	Author & Publication withYear	Problem Statement	Demerits
1.	Wordnet based algorithm	Xiaobin Li, Stan Szpakowicz and Stan Matwin, 14th International Joint Conference on Artificial Intelligence	Lacks Governor dependency relationship	Designed to support text analysis with minimal pre coded knowledge
2.	Phrase raw summary algorithm	Joel Judd and JugalKalita, Better Twitter Summaries. HLTNAACL 2013	Phrase summarization produced by PRA	Lacks Governor dependency relationship

3.	Event detection algorithm	Hassan Sayyadi, Matthew Hurst and Alexey Maykov, Event Detection and Tracking in Social Streams. In Proceedings of ICWSM, 2009.	Event detection using co-occurrence of keywords	Word phrase can be keywords of more than one event

III. PROPOSED SYSTEM

3.1 Disadvantages of ExistingSystems:

- Many tweets may go unseen by audience its intendedfor.
- Difficult to identify spammers orspams.
- Elaborated summarization is not obtained due to pre codedknowledge.
- Time elapsed isirritating.
- No support for Speechacts.

3.2 Proposed SystemIntroduction:

3.2.1 Modules:

Collection Module - This module helps in collecting input form screen i.e. tweets. This module stores the input into datasets.

Identification Module – This module helps in identifying key word or phrases from the input tweets. This module includes identification of events, re-tweets, andsymbols.

Classification Module – This module includes classification of inputs into normal or hash-tag form. It also helps in classification of events and topics.

Recognition Module - This module includes recognition of speech acts. Type of speech acts as per the tweets can be classified with the help of this module.

Display Module - This module is used to display the results obtained by the application. e.g. The tweets like _suggestions‘ or _questions‘ are classified anddisplayed.

3.2.2 SAR

SAR (Speech Act Recognition) is one of the main algorithm used in Twitter Summarization. SAR for Twitter texts helps in keyword or phrase extraction andsummarization.

SAR (Speech Act Recognition):

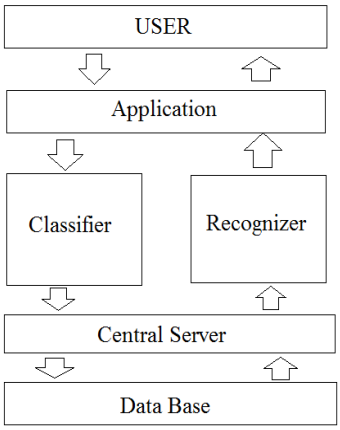
According to SAR the Speech Acts can be classified as

1. Assertive i.e. aStatement
2. Directives i.e. a Question or aSuggestion
3. Expressive i.e. aComment
4. Commensive i.e. miscellaneoustype.

SAR algorithm works with the help of word- based features and symbol-basedfeatures.

- Word basedfeatures:
  1. Cue Words andPhrases
  2. Non-cue Words (Abbreviations, opinion words, vulgar words, emoticons, etc.)
- Symbol-Based Features (#, @, RT,etc.)

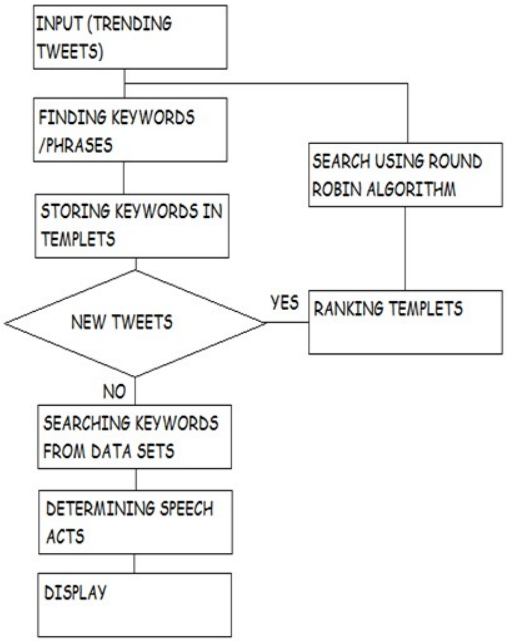
3.2.3 System Architecture



3.2.4.Algorithm

1. Start.
2. Take the dataset_D‘.
3. Taking the input _X‘ form dataset _D‘ Where, X= Comments , tags,etc.
4. Collecting the number of tweets _X‘ andtags_X‘
5. Adding the overall input which is collected fromthedataset_D‘
6. Creating the new dataset _Nd‘ including the collectedinputfromdataset_D‘
7. Takingtheinputfromnewdataset_Nd‘
8. Sorting the collected input according to the rules, Speech acts, tags, grammer,etc.
9. If(X ==Nd)
   {
   Sort(X);
   }
   Else
   {
   Go to step5;
   }
10. Browse the new sorteddataset
11. Analyze the comments, Speechacts
12. Checking the conditions of tags, Extra words, Grammerrules.
13. If _X‘ satisfies theconditions ThenModelingthe_X‘
14. Modeling and annotedTweets
15. Display the SummarizedTweets.
16. Stop.

3.2.5. Flow chart:



X1: True if grouping of tweets is done correctly. X2: True if testing is true. X3: True if no test cases are fail.

```
functionsearch()
{
for(i=0;i<n;i++)
{
 ifstrcmp(,)==true; case :Sucsess;
else
case:failure;
}
}
```

Since the problem becomes a search problem, it is NP-Complete. Now we cannot determine how many test scenarios will get generated and how many test cases will get failed. Moreover there may be some test case in the set of test cases that can be analysed too and those test cases output generated may be failure or success report, which is fixed i.e. In any case report will be generated in any polynomial time. This proves that the problem statement comes under NP Completecategory.

3.2. 6. Feasibilitystudy:

Automatic summarization fortwitter:

Given a failure case viz. invalid grouping of words, we derive an algorithm for this problem asfollows:

For a Problem P to be NP- SAT P ; Let there be 3 constraints X1, X2,X3 Where,

3.2.7. Mathematical Model:

Let S be the solution perspective of the class book suchthat

S= s, e, I, O, SAR, DD, NDD, Fs, success, failure.

where,

s=start state

e=endstate

DD=deterministic data(the data which is taken by our program as input i.e.tweets)

NDD=non-deterministicdata

Input Analysis

Let I be the set of inputparameters

I=test data(the data which is taken by our program as input i.e. information like tweets,keywords,etc)

Output Analysis

The report generated by the function are nothing but the elements of outputset.

Let O be the set of outputparameters. Functions

Functions used in thisproject:-

1) SAR

It is responsible for taking the input parameters which are require for the Output generationpurpose.

2) GenericFunction

SAR=A,K,S,f1,f2,P,F

Where,

A=Abstractive Summarization K=Keyword /phrasesummarization S=speechacts.

f1=accept input() f2=generateoutput(P,F) P=PASS

F=FAIL

Success Case:- It is the case when all the inputs are sortedcorrectly.

Failure Case:- It is the case when the tweets are not summarized according totopics.

3.2.8. Advantages of the proposedSystem:

- Our abstractive summaries are significantly more explanatory, informative andreadable.
- Efficient processing of the massive tweeted information.
- This application handles the numerous, short, dis similar and noisy nature oftweets.
- The rapid proliferation of twitter posts handled by the application decreases and the efficient information acquisitionincreases.

IV. FEATURES

- Application runs efficiently on large number oftweets.
- Classification according to speech acts provides a proper twitter topic summarization.
- The application provides summarization to recently appeared tweets or upcoming tweets.
- A simple graphical representation of summarized twitter topics is provided for user friendlyenvironment.
- Complete analysis of user's input tweets is summarized taking into consideration the speech acts recognitiontechnique.

V. CONCLUSION

In this paper, we propose summarizing tweet streams with regard to topics along time line to produce an overview of topic evolution, which is expressed by sub-topics. Here we have taken a new initiative for Twitter topic summarization—speech act-guided summarization. To automatically recognize speech acts in tweets, we treat SAR as a multi-class classification problem and propose a set of word- based and symbol-based features that can be easily harvested from raw data or freeresources.

VI. FUTURE SCOPE

In the future, we are going to improve Twitter SAR by experimenting with different classifiers, especially the inherent multi-class types such as Naïve Bayes and Decision Tree. As human labeling is expensive and time-consuming, research in a semi-supervised approach is also underway. The summarization framework can also be improved, especially in summary readability. A promising venue is to incorporate the context of key words and phrases during their extraction and count contextual similarity or co-occurrence frequencies into the ranking and template-filling of the extractedterms.

ACKNOWLEDGEMENT

We are very grateful to the help by the faculties of RMD SINHGAD School of Engineering,Pune.

REFERENCES

[1] DehongGao, Wenjie Li, XiaoyanCai, Renxian Zhang, and You Ouyang, Sequential Summarization: A Full View of Twitter Trending Topics in IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 21, NO. 3, MARCH 2013.

[2] J. Searle, P. Cole and J. Morgan,Eds., ―Indirectspeechacts,‖in *Syntax and Semantics*. New York: Academic, 1975, vol. iii, pp. 59–82, Speechacts.

[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 June2014.

[4] Mr. Ganesh Mane, Mrs. Anita Kulkarni, a Event Summarization technique in International Journal of Advanced Research in Computer and Communication Engineering Vol. 4, Issue 5, May2015

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

[6] Joel Judd and JugalKalita, Better Twitter Summaries. HLTNAACL 2013:445-449.

[7] Xiaobin Li, Stan Szpakowicz and Stan Matwin, AWordNetbased Algorithm for Word Sense Disambiguation. In Proceedings of the 14th International Joint Conference on ArtificialIntelligence.

[8] *DUAN YaJuanCHEN ZhuMin WEI FuRu ZHOU Ming Heung − Yeung SHUM for* Summarization by Ranking Tweets in *Proceedings of COLING 2012: Technical Papers*, pages 763–780, COLING 2012, Mumbai, December2012.

AUTHOR PROFILE

Sudhanshu Shiwarkar is currently pursuing the degree of B.E in Computer Engineering from the RMD Sinhgad School of Engineering which is affiliated to the Savitribai Phule Pune University.
Email id: shiwarkars@gmail.com

Indraneel Deshmukh is currently pursuing the degree of B.E in Computer Engineering from the RMD Sinhgad School of Engineering which is affiliated to the Savitribai Phule Pune University.
Email id: indraneeldeshmukh2@gmail.com

Nikhil Patil is currently pursuing the degree of B.E in Computer Engineering from the RMD Sinhgad School of Engineering which is affiliated to the Savitribai Phule Pune University.
Email id: neekhil.patil@gmail.com

Akash Thanke is currently pursuing the degree of B.E in Computer Engineering from the RMD Sinhgad School of Engineering which is affiliated to the Savitribai Phule Pune University.
Email id: thankeakash@gmail.com

