I J C T A, 9(15), 2016, pp. 7497-7501

Tweet Segmentation and its Application Using Random Walk and Part-of-speech Methods

Kukku S.*, Reshma Reghu and Gaina K.G.*****

ABSTRACT

Many users share their information in twitter and they produces large amount of data every day. However, the short kind of tweets created many severe problems in the applications of Information retrieval (IR) and Natural Language processing (NLP).In this paper, we put forward an innovative foundation for tweet segmentation in batch mode, known as HybridSeg.The downstream applications are able to easily withdraw and maintain the semantic or context information, if the tweets are broke into meaningful chunksBoosting the total stickiness score of its candidate segments is the method adopted by HybridSeg to achieve the excellent tweet segmentation. Global context and local context are the two factors which influences stickiness score. For the local context, we suggest and appraise two models which consider the grammatical properties and interdependence in a group of tweets. From the experiments conducted on datasets, it shows that the segmentation quality is improved by considering global as well as local contexts. By conducting experiments and comparing the results, we prove that local grammatical traits are more important for assimilate local context compared with term-interdependence.In this paper we illustrate that more excellence in segmentation is possible by applying part-of-speech method.

Keywords: Tweet Segmentation, Random Walk, Part-Of-Speech methods

1. INTRODUCTION

Sites like Twitter provide users the opportunity and platform to express their interests and opinions. Thus these sites enable users to act as a vast network sharing and circulating popular interests. This is vital information that is circulating and is relevant for organizations in terms of business and from many other perspectives also. Twitter has captivated the interest of organizations and academia because of its enormous possibilities in the current scenario. One example for such possibility is capturing public's true opinion from the twitter streams on an organization or a product. It is enormous amount of data that needs to be handled here and it will be inefficient and infeasible to monitor the constantly streaming Twitter data. Tweets are filtered based on user's requirements and targeted Twitter stream is built. This can help in collecting early reactions and identifying crisis at early stages. Named Entity Recognition (NER), event detection, opinion mining and sentiment analysis are some applications and understanding tweet language is one important factor that needs to be processed.

The informal nature of tweets makes it user friendly but the chances for informal abbreviations, errors in grammar and spellings are increased in this environment. This environment in Twitter makes its processing models inconsistent in nature. For example, consider a tweet "I no come, I go there." there is no hint to guess its true theme by disregarding word order (i.e., bag-of-word model).

Misspellings, error correction, tweet segmentation and comparing Part-of-speech and Random walk algorithms are the areas we are focusing in this paper.

* Department of Computer Science & I.T., *Email: kukkusaboo@gmail.com*

** Amrita School of Arts & Sciences, Kochi *Email: reshmareghuraman@gmail.com*

*** Amrita Vishwa Vidyapeetham (Amrita University), India *Email: gainakg@gmail.com*

2. RELATED WORKS

Many researchers had done numerous experiments to rectify the misspellings occur while tweeting in the tweet application. Very few approaches were implemented to uncover error correction while posting a tweet using the NER algorithms. Some of the approaches are reviewed below.

[1]This paper presented NER system for targeted Twitter stream, called TwiNER. TwiNER is unsupervised method and it does not based on the local linguistics characteristics. Instead it Experimental results are favorable for TwiNER.

From the experiment it also shows that state-of-the-art NER systems and TwiNER has the same performance in real-life tweet streams. [2] This approach finds the association between user interest and followed friends and posted tweets. This approach provides a fine basis for a solid tweet application. This theory is making use of named entities withdrew from tweets that have the potential to decide the users interest. [3] This shows excellent tweet segmentation is especially achieved by existing state-of-art algorithm HybridSeg. It also proves named entity recognition is effectively possible with finer segmentation process in tweets. [4] This study is also based on named entities from tweets. Based on the entities withdrew, user modeling and tweet recommendation is formed. This study also shows that for getting named entities, annotated huge amount of training data is not needed, hence overburden of annotation can be avoided. Also this approach does not based on linguistics of the language. Experiments prove that user interest is playing major role for tweet recommendation in this approach. [5] Suggested in order to keep semantic definition of tweets, tweet segmentation really helps. Improved correctness and excellence is achieved by segment based recognition techniques.[6]SCUBA is a model for detecting sarcasm in tweets. This model has two major advantages.1)It considers psychological and behavior features of construct resilient global and local context for tweets from the information from the web. sarcasm 2) It grasps user's former information. These helped to detect whether tweets are sarcastic or not. [7]Explored, automatic detachment of sarcastic messages from linguistic and pragmatic features of tweets.

3. PROPOSED METHODS

We recommend and assess two segment-based NER algorithms. Both are taking tweets as input and are not training based. First algorithm is applying random walk method based on co-occurrence of named entities. It assumes that named entities are more likely to co-occur together in tweet streams. The second algorithm is utilizes Part-of-Speech method. This method is giving importance to POS tags of element words in segments. Here namedentities are segments that are likely to be a noun phrase. The other algorithm utilizes Part-of-Speech (POS) tags of the constituent words in segments. The segments that are likely to be a noun phrase are considered as named entities.

3.1. Disadvantages in the existing system

1) Twitter is a place where people share their perspectives on different issues occurring around them, where there is a barrier in which only certain characters will be used for uploading, thus make the user a mini irrational.
2) The most problem is when updating a status , if a spelling mistakes appears it couldn't be resolved automatically, where it has to be removed manually that arise lot of in suitable to the particular user

3.2. Advantages in the proposed system;

1) To overcome the First issue for the registered users of these DLL(DYNAMIC LINK LIBRARY) supportive file can able to upload to their maximum character as per their wish, thus provide a wider view of Sharing their view in a social site.

2) To overcome the second disadvantage for the registered users of these DLL can provide an automatic spelling check for one or two words before uploading that helps the user to provide a safe updating where it doesn't create problems.

To rectify the misspellings occur while tweeting in the tweet application, we chose the DLL file which is used in the MS word.

Also we included random walk and parts of speech methods. Since the complete implementation of the algorithm is not feasible, regarding our base paper , we took the concept alone and also added a comparison in tweet spellingcorrection .

4. METHODS AND METHODOLOGIES

4.1. DLL

A DLL library consists of code and data that can be reused by more than one program at a time. It is used as a supportive file for other application and it does not contain an entry point which means it does not contain a Main Function) ,so it cannot run individually.Os does not create a separate process for any DLL rather DLL will run in the same process created for execution. A DLL file can be reused by other application.

4.2NER Algorithm

For information sharing and communication, so many common phrases are used in tweets and it contains lots of errors in spellings and grammar.So we put forward two segment-based NER algorithms –Random Walk (RW) and Part Of Speech (POS) and both of these algorithms are unsupervised and the input is taken as tweet segments.

4.2.1. Random walk

Random Walk consists of sequence of random steps and it is applied to the segment. First this model reads the entire text and return back to the beginning (i.e., position 0) Then it jumps to the neighboring position or next word (i.e., position 1) and rectify the errors. And this process continues until it reaches the last word of the sentence.

4.2.2. Part of speech

Rather than reading the entire text in RW, POS compares the adjacent and related words in phrase and rectify the errors in the beginning itself.

We thoroughly examined random walk and parts -of -speech and organized the concept alone in a comparison manner .The exact algorithm is not implemented as it is not possible to do so.

5. EXPERIMENTAL RESULTS

For the experiment we selected different tweets and analyzed using Random walk and Parts of speech methods. From the experiment we got that Parts of Speech method is better than Random walk.

Graphical representation of the performance of two algorithms Random Walk and Part-Of-Speech.

The figure below shows the output, and from this we proved that by comparing both the algorithms POS shows better and faster performance compared to random walk.

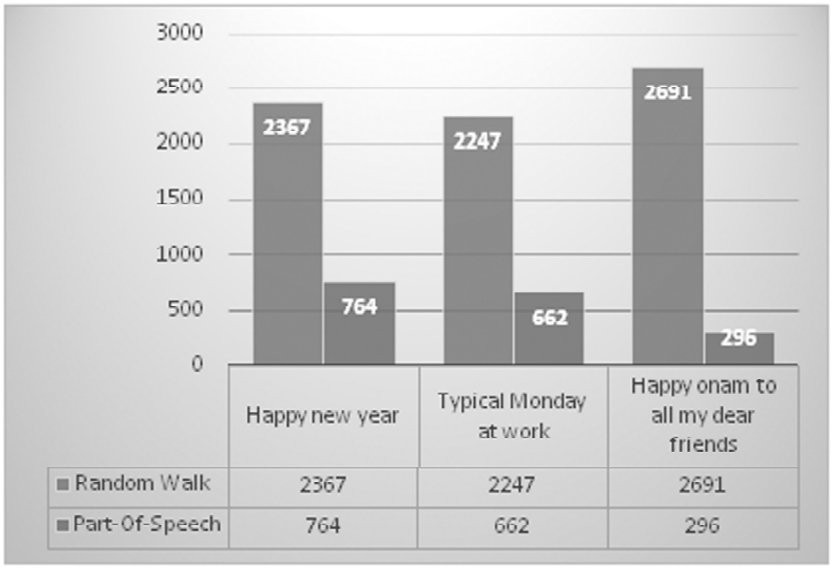

Figure 1: Random Walk method

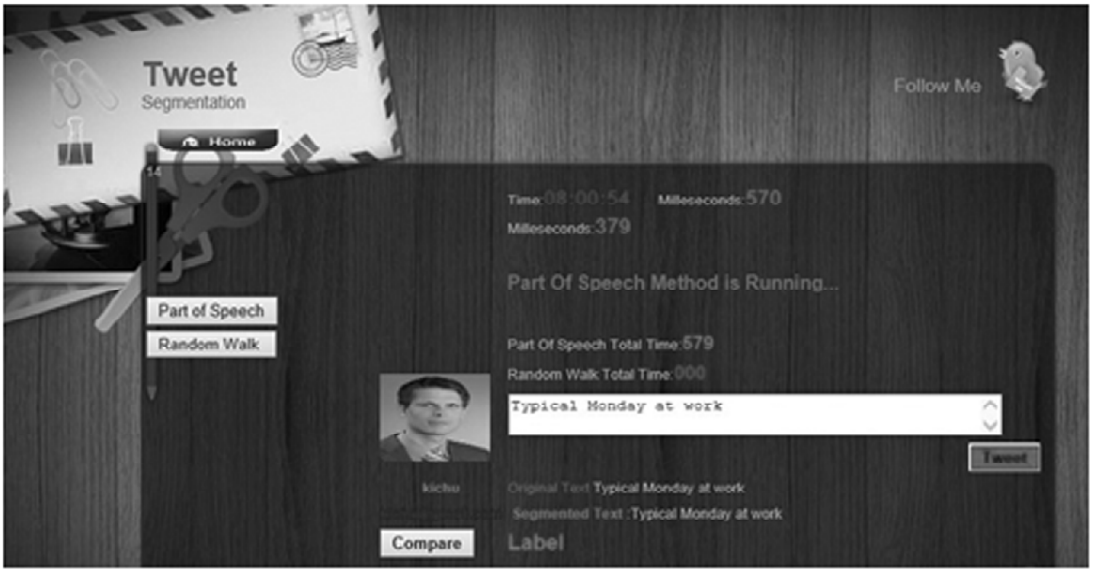

Figure 2: Part-Of-Speech method

Figure 3: Comparison of RW and POS methods

6. CONCLUSIONS

In this paper, we chose DLL file is used to rectify the misspellings occur while tweeting in the tweet application. Experimental results show that Part-of-speech method has got a betteraccuracy and performance when compared to the Random Walk method. The proposed methodology proves to be efficient for larger data-sets to get faster results.

REFERENCES

[1] Li, Chenliang, et al. "Twiner: named entity recognition in targeted twitter stream." *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2012.

[2] KARATAY, DENIZ. "TWEET RECOMMENDATION UNDER USER INTEREST MODELING WITH NAMED ENTITY RECOGNITION." PhD diss., MIDDLE EAST TECHNICAL UNIVERSITY, 2014.

[3] Li, Chenliang, Aixin Sun, Jianshu Weng, and Qi He. "Exploiting hybrid contexts for tweet segmentation." In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pp. 523-532. ACM, 2013.

[4] Karatay, Deniz, and Pinar Karagoz. "User Interest Modeling in Twitter with Named Entity Recognition." *Making Sense of Microposts (# Microposts2015)* (2015).

[5] Chavan, Mr. Chetan, and Ranjeetsingh Suryawanshi. "Tweet Segmentation and Named Entity Recognition."

[6] Rajadesingan, Ashwin, Reza Zafarani, and Huan Liu. "Sarcasm detection on twitter: A behavioral modeling approach." In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, pp. 97-106. ACM, 2015.

[7] Gonzalez Ibanez, Roberto, Smaranda Muresan, and Nina Wacholder. "Identifying sarcasm in Twitter: a closer look." In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pp. 581-586. Association for Computational Linguistics, 2011.

[8] Ritter, Alan, Sam Clark, and Oren Etzioni. "Named entity recognition in tweets: an experimental study." In Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 1524-1534. Association for Computational Linguistics, 2011.

[9] Mukherjee, Triparna, and Asoke Nath. "International Journal of Advanced Research in Computer Science and Software Engineering." International Journal 5, no. 6 (2015).

[10] Wang, Chaoyue, and Guohong Fu. "Chinese Tweets Segmentation based on Morphemes." CLP 2012 (2012): 106.