ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering
(An ISO 3297: 2007 Certified Organization)
Website: www.ijircce.com
Vol. 5, Issue 4, April2017

Tweet Segmentation and Summarization Using Named Entity Recognition and Event Detection

Shinma NP
M.Tech Student, Dept. of Information Technology, Rajagiri School of Engineering and Technology, Ernakulam, India

ABSTRACT: Twitter is an emerging, social media where people share their messages and interact with other users messages. Many real-time news and events were posted on the Twitter so that peoples get up to date information. To get the actual essence of the story the tweet summarization is required. A system named SocialSegNER introduced to generate the tweet summary. This model first detects the named entity and event phrases of the tweet, then make summarization by using that.

KEYWORDS: Named entity recognition, part-of-speech, tweet segmentation

I. INTRODUCTION

Twitter is a large social media where each user is an individual news media where each user share their opinions and also absorb information. The fast interactions between users in Twitter help timely detect the events. Event detection helps the users to take favourable users opinion, and a company can quickly make a fast response for their products changes. Tweet summarization of an events or corporations product reviews will get the overall idea about the event and products. Each tweet can send up to 140 characters. Due to this short nature, many noises will be there in a tweet such as misspellings and simple abbreviations. So event detection for existing natural language processing would be the difficult task.

So for detecting the event and recognising the named entity segmentation from a batch of tweets are done. For segmentation, the tweet split into n consecutive segments. A segment may contain a word or more than one. Each segment indicates an event phrases or named entity, and a function is used to calculate the stickiness grade of the segment, i.e., more split the segment would break the meaning of the segment. Then identify whether the segment is valid or not by comparing it to the global word and local word. Global word is the word that presents in the web pages or Wikipedia, that the word contains the universal meaning. Local word identified using local linguistic features and local collocation. Then the named entity is recognised by using two algorithms random walk NER and part-of-speech NER. Event phrases are identified using part-of-speech. After the named entity recognition (NER) and event phrases, pre-constructed the summary for the events which is valid for all events. Which then mixed with the data's from NER and a correct sentence generated.

II. RELATED WORK

Existing techniques mainly designed for implementing the formal text. These methods are not accurate for local linguistic features like POS tags, word capitalization, etc. This features can be done using some supervised learning algorithms, but it can't be applied in tweets because tweets are informal in nature. It contains many short forms of text and local linguistics. To improve POS tagging on tweets, Gimple et al. incorporate tweet specific features including at-mentions, hashtags, URLs, and emotions. In their approach, they measure the confidence of capitalized words and apply phonetic normalization for ill-formed words to address possible peculiar writings in tweets. Normalization of ill-formed words in tweets has established itself as an important research problem. A supervised approach is employed

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering
(An ISO 3297: 2007 Certified Organization)
Website: www.ijircce.com
Vol. 5, Issue 4, April2017

into first identify the ill-formed words. Then, the correct normalization of the ill-formed word is selected based on some lexical similarity measures.[1][3].

III. PROPOSED ALGORITHM

A. *Design Considerations:*

- Initially, the tweet split into m consecutive segments
- Identified whether the segment is valid or not by comparing with the global word and local word
- Named entity is identified using random walk NER and POS
- Event phrases is identified using POS
- Tweet summarization is done using named entity and event phrases.

B. *Description of the Proposed Algorithm:*

Aim of the proposed algorithm is to generate the tweet summarization. The detailed system architecture is given below.

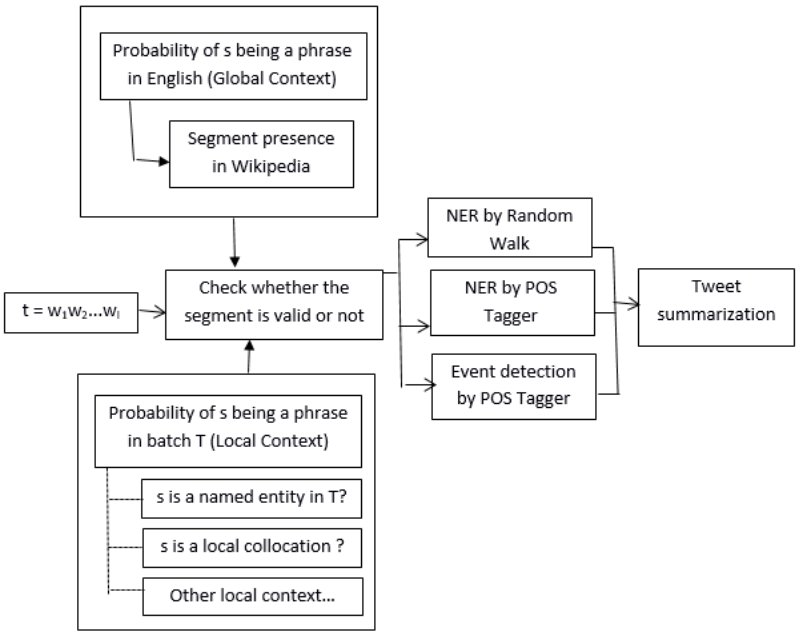

Figure 1: System Architecture

The proposed algorithm is consists of four main steps.

Step 1: Tweet Segmentation:

The SocialSegNER model first perform the tweet segmentation. Consider a tweet t contains l words and the tweet segmentation method split the tweet into m following segments, i.e., t = s1,; s2…sm where s one or more than one words. In this method, the sum of the closeness of segments is a major problem. An excellent closeness grade indicates that the segment s is a phrase and more it split the correct word collocation and meaning of segment would break. For example, “Keep calm” is a segment and together it has a meaning. If the segment is further divided “Keep” and”calm” would be two words and it is meaningless. So during the tweet is splitting the segment is validated. The tweet splits in the order of ngram,… bigram, unigram, where the length of ngram will be a length of the tweet. [1][3].

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering
(An ISO 3297: 2007 Certified Organization)
Website: www.ijircce.com
Vol. 5, Issue 4, April2017

Step 2: Identify Segment is Valid or Not:
The validation was done in four stages. First, check whether the segment is a global word or not by counting the number of appearances of a segment to a significant corpus like Microsoft Web N-Gram Corpus. Second check whether the segment is local word or not by applying multiple off the shelf NERs trained on formal texts to detect named entities in a collection of tweets T. Third, word collocation is a sequence of words or terms that come with more often than would be expected by chance so, measure the segment's sub ngrams of a ngram and check they associated with one another, so as to evaluate the possibility of the ngram being a valid segment. [1]

Step 3: Named Entity and Event Phrases Detection:
By using two algorithms, the named entity identified. They are NER based on the random walk and NER based on POS Tagger. The first NER algorithm based on the information that a named entity often co-exist with other named entities in a collection of tweets. Based on this knowledge a segment graph is drawn. By implementing random walk through the graph the named entity is identified. The second algorithm figures out the part of speech tags in the tweet for the identification of the named entity by seeing noun phrases as named entity using segment rather a word as a unit. A segment may appear in different tweets, and its constituent words may be attached different POS tags in these tweets. Then evaluate the possibility of a segment being a noun phrase by considering the POS tags of its constituent words of all appearances. Then by using natural language tool event is detect in the tweet. [6]

Step 4: Tweet Summarization:
By usingthe named entities and event phrases the tweet summarization is doing. After identifying the named entity and event phrases a weighted undirected graph is made. In the graph, nodes represented the event phrases and named entity differently. If two nodes connected by an undirected edge and if they coexisted in t tweets, and the weight of that edge is t. The PageRank-like algorithm used in TextRank (Mihalcea and Tarau, 2004) used in automatic summarization.[2]

IV. SIMULATION RESULTS

The simulation involves, the tweet data sets for the experiments was extracted from the Twitter by matching some predefined keywords and hashtags for some event using the Twitter API. Then the tweet segmentation is done using the Natural Language Toolkit. The global context is compared with the Wikipedia Dump which is released on 30 January, 2010. This dump contains 3,246,821 articles and there are 4,342,732 distinct entities appeared as anchor texts in these articles. Named Entity and Event phrases are identified using the Natural Language Toolkit.

The simulation result is given below,

Tweet	Named Entity	Event Phrases	Summary
VIDEO:Amazing...Motorcycling daredevils parade on India's 67th Republic Day in New Delhi Credit:PressTV https://t.co/yeFIbdnqbB	India Republic Day New Delhi	parade	parade at New Delhi in Republic Day

Table1. Simulation result

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering
(An ISO 3297: 2007 Certified Organization)
Website: www.ijircce.com
Vol. 5, Issue 4, April2017

V. CONCLUSION

In this paper, summarization of tweet performed. The SocialSegNER model which first segments tweets into meaningful phrases called segment using both global word and local word. Then show the identification of named entity and event phrases using two algorithms. After the named entity and event phrases identification, the summary constructed from the data's that revived from the named entity recognition. The summary created in such a way that the received data construct the summary.

REFERENCES

1. Li, Chenliang, et al. "Tweet segmentation and its application to named entity recognition." *IEEE Transactions on Knowledge and Data Engineering* 27.2, pp. 558-570, 2015.
2. Xu, Wei, et al. "A preliminary study of tweet summarization using information extraction." *NAACL 2013*: pp. 20, 2013.
3. C. Li, J. Weng, Q. He, Y. Yao, A. Datta, A. Sun, and B.S. Lee, "Twiner: Named entity recognition in targeted twitter stream," in Proc. 35th Int.ACM SIGIR Conf. Res. Develop. Inf. Retrieval, pp. 721730, 2012.
4. C. Li, A. Sun, J. Weng, and Q. He, "Exploiting hybrid contexts for tweet segmentation," in Proc. 36th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, pp. 523532, 2013.
5. C. Li, A. Sun, and A. Datta, "Twevent: segment based event detection from tweets," in Proc. 21st ACM Int. Conf. Inf. Knowl. Manage., pp. 155164.35,2012
6. Gimpel, Kevin, et al. "Part-of-speech tagging for twitter: Annotation, features, and experiments." *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*. Association for Computational Linguistics, pp. 42-47, 2011.