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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 4, April 2016

Named Entity Recognition with the Use of Tweet Segmentation

Akshay Shinde, Sachin Yelmar, Rahul Yelwante, Nilesh Chavan

Student, Dept. of Information Technology, Dhole Patil COE., Wagholi, Pune, Maharashtra, India

ABSTRACT: Twitter is attracting as lot of user using twitter. Many type of data will be stored and share on twitter hence large amount data produced every day.As using tweets the context data will be spitted accurately into segment. For the latter, propose and evaluate the models to derive local context by considering the linguistic features and the term dependency in a batch of tweets, respectively. HybridSeg is designed to iteratively learn from the confident segments as a pseudo feedback. Experiments on two tweet data sets show that tweet segmentation quality is improved by learning both the global and local contexts compared with using global context alone. Through the analysis, identified that local linguistic features are reliable for learning local context compared with term-dependency. As an application, tried to show that high accuracy is achieved in named entity recognition by applying segment-based part-of-speech tagging.

KEYWORDS: Tweet Segmentation, Named Entity Recognition.

I. INTRODUCTION

MICROBLOGGING sites like as the Twitter have reshaped the way people find, share, and disseminate timely information. Many organizations have reported to create and monitor targeted Twitter streams to collect and understand users 'opinions. With the predefine selection criteria targeted tweeter stream will be filtering. Due to the invaluable business value of timely information from these tweets it is imperative i.e. important to understand tweets 'language for the large body of downstream applications like as named entity recognition (NER) [4] event detection and summarization [5] [6] opinion mining sentiment analysis and many others.

Given the length of a tweet is limited (i.e., 140 characters) and no restrictions on its writing styles, tweets often contain misspellings, grammatical errors, and informal abbreviations? The error-prone and the short nature of tweets often make the word-level language models for tweets are less reliable. For example, given a tweet _I call her, no answer, and her phone in the bag, she dancing, there is no clue or no idea to guess its true theme by disregarding word order (i.e., bag-of word model). The situation is further described and exacerbated with the limited context provided by the tweet. That is, more than one explanation for this tweet is derived by different readers if the tweet is considered in isolation. On the other hand, due to the noisy nature of tweets, the core semantic information is well preserved in tweets in the form of semantic phrases or named entities. For example, the emerging phrase she dancing in the related tweets indicates that its a key concept it classifies the tweet into family of tweets talking about the song She Dancing‖ a trend topic in Bay Area.

Overview-First, to obtain tweets on the target event precisely, apply semantic analysis of a tweet. For example, users make tweets which can be such as _Now it is shaking' or _Earthquake' for which shaking or earthquake are considered as keywords, but users may also make tweets such as _Someone is shaking hands with their friends' Or _I am attending an Earthquake Conference.' After this then try and prepare the training data and then devise a classifier using a Support Vector Machine (SVM) based on features such as the number of words, keywords in a tweet, and the context of target-event words. After doing this then achieve a probabilistic spatiotemporal model of an event, then make a crucial assumption that each Twitter user is regarded and considered as a sensor and each tweet as sensory information.

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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 4, April 2016

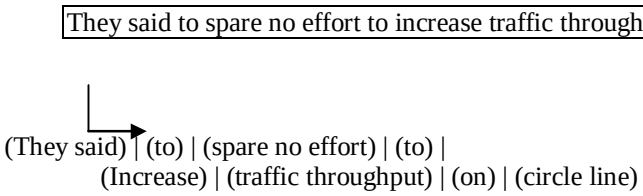

Fig. 1. Example of tweet segmentation.

II. PROJECT IDEA

User The main focus is given on the task of tweet segmentation. The goal of this task is to split a tweet into a sequence of consecutive n-grams (n > 1), each of which is called a segment. A segment can be a named entity (e.g., a movie title "finding nemo"), a semantically meaningful information unit (e.g., "officially released'), or any other types of phrases which appear "more than by chance". Fig. 1 gives an example. In this example, a tweet "They said to spare no effort to increase traffic throughput on circle line." is split into eight segments. Semantically meaningful segments "spare no effort", "traffic throughput" and "circle line" are preserved. These segments help for the semantic meaning of the twits more effectively than similar word does. Segment based tweet can improve geographical location from the tweet as using segments circle line. In fact, segment-based representation has shown its effectiveness over word-based representation in the tasks of named entity recognition and event detection. Note that, a named entity is valid segment; but a segment may not necessarily be a named entity. In the segment "Korea versus Greece" is detected for the event related to the world cup match between Korea and Greece. Then identified two directions for the research. One is to further improve the segmentation quality by considering more local factors. The other is to explore the effectiveness of the segmentation-based representation for tasks like tweets summarization, search, hashtag recommendation, etc.

III. MOTIVATION

To achieve high quality tweet segmentation, used a generic tweet segmentation framework, named HybridSeg. HybridSeg learns from both global and local contexts, and has the ability of learning from pseudo feedback. The HybridSeg framework which segments tweets into meaningful phrases called segments using both global and local context. Through the framework, it is demonstrated that local linguistic features are more reliable than termdependency in guiding the segmentation process. This finding opens opportunities for tools developed for formal text to be applied to tweets which are believed to be much noisier than formal text. Tweet segmentation helps to preserve the semantic meaning of tweets, which subsequently benefits many downstream applications, e.g., named entity recognition. Through experiments, it is shown that a segment-based named entity recognition method achieves much better accuracy than the word-based alternative.

IV. LITERATURE SURVEY

In the work [2] Author said, Social events are events that occur between people where at least one person is aware of the other and of the event taking place. Extracting social events can play an important role in a wide range of applications, such as the construction of social network. Here author has introduced the task of social event extraction for tweets, an important source of fresh events. One main challenge is the lack of information in a single tweet, which is rooted in the short and noise-prone nature of tweets. Author propose to collectively extract social events from multiple similar tweets using a novel factor graph, to harvest the redundancy in tweets, i.e., the repeated occurrences of a social event in several tweets. Then evaluate the method on a human annotated data set, and show that it outperforms all baselines, with an absolute gain.

In the work [3] Author said, the challenges of Named Entities Recognition (NER) for tweets lie in the insufficient information in a tweet and the unavailability of training data. Author proposed to combine a K-Nearest Neighbors classifier with a linear Conditional Random Fields model under a semi-supervised learning framework to tackle these challenges. The K-Nearest Neighbors based classifier conducts pre-labeling to collect global coarse evidence across tweets while the Conditional Random Fields model conducts sequential labeling to capture fine-grained information

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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 4, April 2016

encoded in a tweet. The semi-supervised learning plus the gazetteers alleviate the lack of training data. Extensive experiments show the advantages of the method over the baselines as well as the effectiveness of KNN and semi supervised learning.

In the work [4] Author said, Event detection from tweets is an important task to understand the current events/topics attracting a large number of common users. However, the unique characteristics of tweets (e.g., short and noisy content, diverse and fast changing topics, and large data volume) make event detection a challenging task. Most existing techniques proposed for well written documents (e.g., news articles) cannot be directly adopted. Author proposed a segment-based event detection system for tweets, called Tweet event. Tweet event first detects busty tweet segments as event segments and then clusters the event segments into events considering both their frequency distribution and content similarity. More specifically, each tweet is split into nonoverlapping segments (i.e., phrases possibly refer to named entities or semantically meaningful information units). The busty segments are identified within a fixed time window based on their frequency patterns, and each busty segment is described by the set of tweets containing the segment published within that time window. The similarity between a pair of busty segments is computed using their associated tweets. After clustering busty segments into candidate events, Wikipedia is exploited to identify the realistic events and to derive the most newsworthy segments to describe the identified events. We evaluate Tweet event and compare it with the state-of-the-art method using 4.3 million tweets published by Singapore-based users in June 2010. In experiments, Tweet event outperforms the state-of-the-art method by a large margin in terms of both precision and recall. More importantly, the events detected by Tweet event can be easily interpreted with little background knowledge because of the newsworthy segments. Author also shows that Tweet event is efficient and scalable, leading to a desirable solution for event detection from tweets.

V. IMPLEMENTATION STRATEGIES

A. ARCHITECTURE

The general overview of the system architecture is shown below. The method used in the study segments the tweets and generates named entity candidates. These candidates have to be validated so that they can be used as an indicator of the user‘s interest. In this step, Wikipedia is chosen as a reference for a segment to be a named entity, or not. Since our Tweet collection is in Turkish, Turkish Wikipedia dump published by Wikipedia is obtained. For named entities to be extracted successfully, the informal writing style in tweets has to be handled. Generally named entities are assumed as words written in uppercase or mixed case phrases where uppercased letters are at the beginning and ending, and almost all of the studies bases on this assumption. However, capitalization is not a strong indicator in tweet-like informal texts, sometimes even misleading. To extract named entities in tweets, the effect of the informality of the tweets has to be minimized as possible. The preprocessing tasks applied can be divided into two logical group. Pre-segmenting, and Correcting. Removal of links, hashtags, mentions, conjunctives, stop words, vocatives, slang words and elimination of punctuation are considered as pre segmentation. It is assumed that parts in the texts before and after a redundant word, or a punctuation mark cannot form a named entity together, therefore every removal of a word is considered as it segments the tweet as well as punctuation does it naturally. Removal of repeating characters that are used to express a feeling such as exaggerating, or yelling, handling mistyping and asciification related problems are considered as correcting and can be thought of conversion of tweets from informal to formal. In the following subsections, described the NER and user profile modelling and recommendation steps in more detail.

The main goal is to reduce Twitter users‘ effort to access to the tweet carrying the information of interest. To this aim, a tweet recommendation system under a user interest model generated via named entities is presented. The system mainly involves six phases; data gathering, knowledge base construction, data preprocessing, named entity recognition, user interest model generation based on named entities and finally recommendation. General information on the phases is as follows:

- Data Gathering is the process of collecting a Twitter user‘s data, including user‘s friends‘ posts as well as user‘s ownposts. In this phase, user-friend relationship is also extracted and friends‘ relative ranking is generated as an output.
- Knowledge Base Construction is the process of generating a graph-based knowledge base of TurkishWikipedia article titles and their links to each other, in order to validate named entity candidates generated as an output of Named Entity Recognition phase. Keeping this knowledge base up to date is also

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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 4, April 2016

included in this phase. Although other phases iteratively follow each other and on e‘s output is the other‘s input, this phase is independent and conducted in parallel.

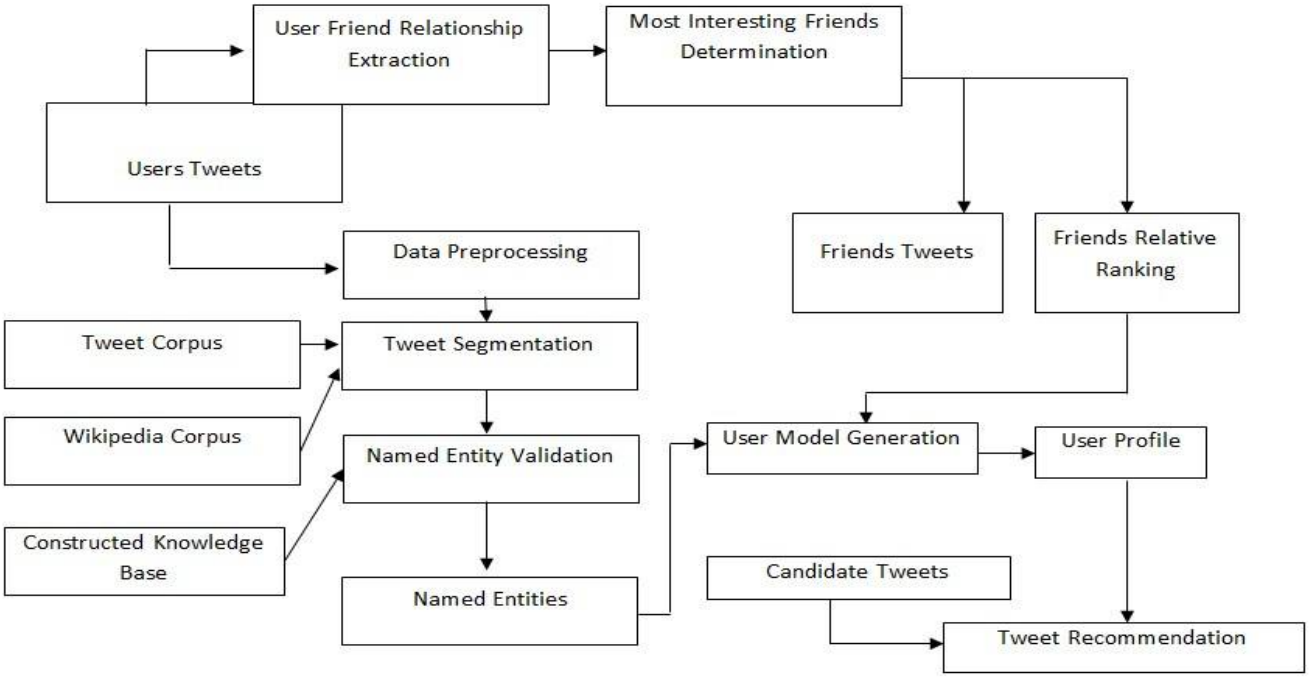

Fig 2System Architecture.

- Data Preprocessing includes removing unnecessary parts of tweet texts such as mentions, hashtags, smileys, vocatives, links etc. Since informal writing style is commonly adopted in tweets, this phase is also responsible from normalizing the tweet text such as getting rid of unnecessarily repeated characters, slang words, correcting asciification related problems.
- Named Entity Recognition is the next phase of data pre-processing phase. In this phase, tweet segmentation on preprocessed tweets is carried out by means of global context and segments as candidate named entities are generated. Then, these candidates are validated as named entities or ignored by usage of previously constructed knowledge base of Turkish Wikipedia article titles.
- User Interest Model Generation phase is a must. In this phase, using named entities extracted from user‘s and user‘s friends ‘tweets and user-friend relationships, a user interest model is generated. In other words, a Twitter user is represented via weighted named entities.
- Tweet Recommendation is the last phase, where two kinds of recommendation applications applied by comparing candidate tweets with the generated user interest model. Tweet classification which is the task of deciding whether a candidate tweet is interesting for the user or not, and tweet ranking which aims to sort tweets from the most recommendable to the least recommendable are performed in this phase.

B. HybridSeg Framework

The proposed HybridSeg framework segments tweets in batch mode. Tweets from a targeted Twitter stream are grouped into batches by their publication time using a fixed time interval (e.g., a day). Each batch of tweets are then segmented by HybridSeg collectively.

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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 4, April 2016

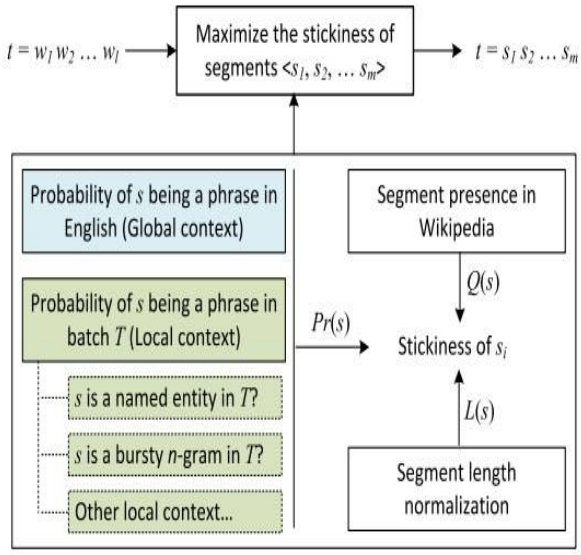

Fig 3 HybridSeg framework.

C. Tweet Segmentation

Tweets are considered noisy with lots of informal abbreviations and grammatical errors. However, tweets are posted mainly for information sharing and communication among many purposes.

Given an individual tweet t ∈ Ti, the problem of tweet segmentation is to split t into m consecutive segments, t = s1,s2...sm; each segment contains one or more words. A high stickiness score of segment s indicates that it is not suitable to further split segment s, as it breaks the correct word collocation. In other words, a high stickiness value indicates that a segment cannot be further split at any internal position. If the word length of tweet t is L, there exist 2L−1 possible segmentations.

Observations for Tweet Segmentation:

Observation 1. Word collocations of named entities and common phrases in English are well preserved in Tweets. Many named entities and common phrases are preserved in tweets for information sharing and dissemination. In this sense, Pr(s) can be estimated by counting a segment‘s appearances in a very large English corpus (i.e., global context). In our implementation, we turn to Microsoft Web NGram corpus. This N-Gram corpus is derived from all web documents indexed by Microsoft Bing in the EN-US market. It provides a good estimate of the statistics of commonly used phrases in English.

Observation 2. Many tweets contain useful linguistic features. Although many tweets contain unreliable linguistic features like misspellings and unreliable capitalizations, there exist tweets composed in proper English. For example, tweets published by official accounts of news agencies, organizations, and advertisers are often well written. The linguistic features in these tweets enable named entity recognition with relatively high accuracy**.**

Observation 3. Tweets in a targeted stream are not topically independent to each other within a time window. Many tweets published within a short time period talk about the same theme. These similar tweets largely share the same segments. For example, similar tweets have been grouped together to collectively detect events, and an event can be represented by the common discriminative segments across tweets.

The latter two observations essentially reveal the same phenomenon: local context in a batch of tweets complements global context in segmenting tweets. For example, person names emerging from bursty events may not be recorded in Wikipedia. However, if the names are reported in tweets by news agencies or mentioned in many tweets, there is a good chance to segment these names correctly based on local linguistic features or local word collocation from the batch of tweets.

D. Segment-Based Named Entity Recognition

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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 4, April 2016

Here select named entity recognition as a downstream application to demonstrate the benefit of tweet segmentation. Investigate two segment-based NER algorithms. The first one identifies named entities froma pool of segments (extracted by HybridSeg) by exploiting the co-occurrences of named entities. The second one does so based on the POS tags of the constituent words of the segments.

NER by Random Walk:
The first NER algorithm is based on the observation that a named entity often co-occurs with other named entities in a batch of tweets (i.e., the gregarious property). Based on this observation, build a segment graph. A node in this graph is a segment identified by HybridSeg. An edge exists between two nodes if they co-occur in some tweets; and the weight of the edge is measured by Jaccard Coefficient between the two corresponding segments. A random walk model is then applied to the segment graph.

NER by POS Tagger:
Due to the short nature of tweets, the gregarious property may be weak. The second algorithm then explores the partofspeech tags in tweets for NER by considering noun phrases as named entities using segment instead of word as a unit. A segment may appear in different tweets and its constituent words may be assigned different POS tags in these tweets. Then estimate the likelihood of a segment being a noun phrase by considering the POS tags of its constituent words of all appearance

VIII. RESULTS

The following figures shows the result of implementations.

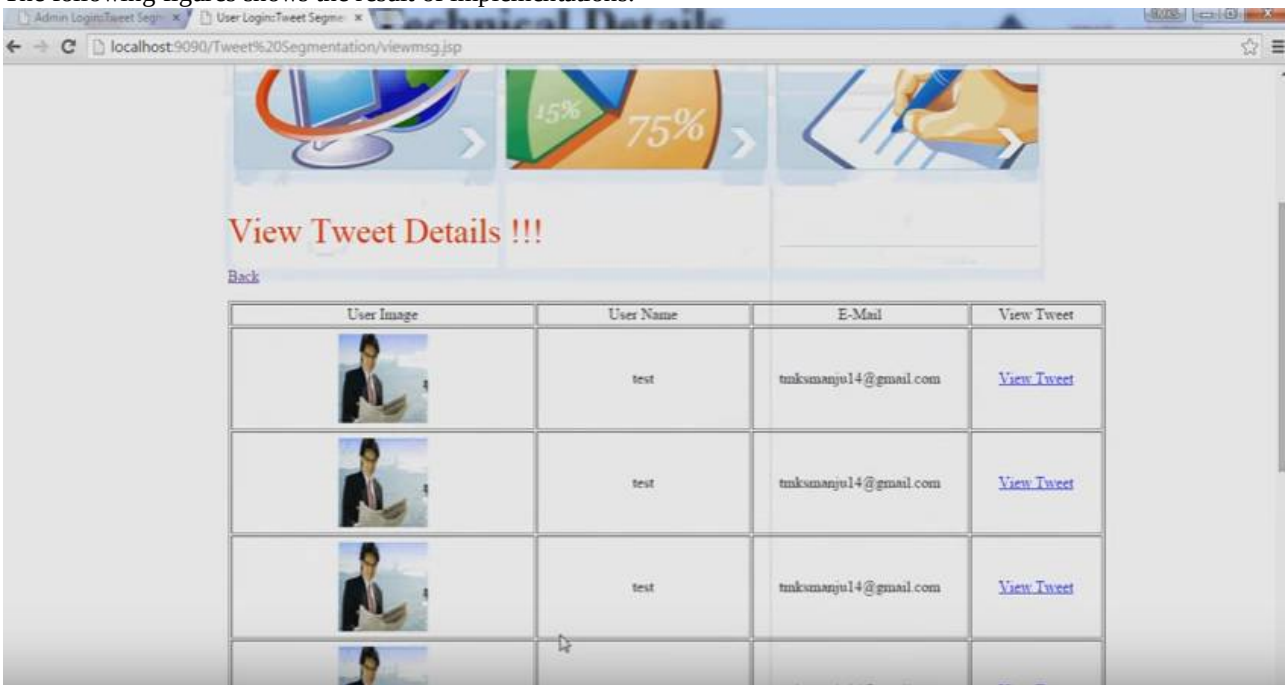

Fig 4. Admin view

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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 4, April 2016

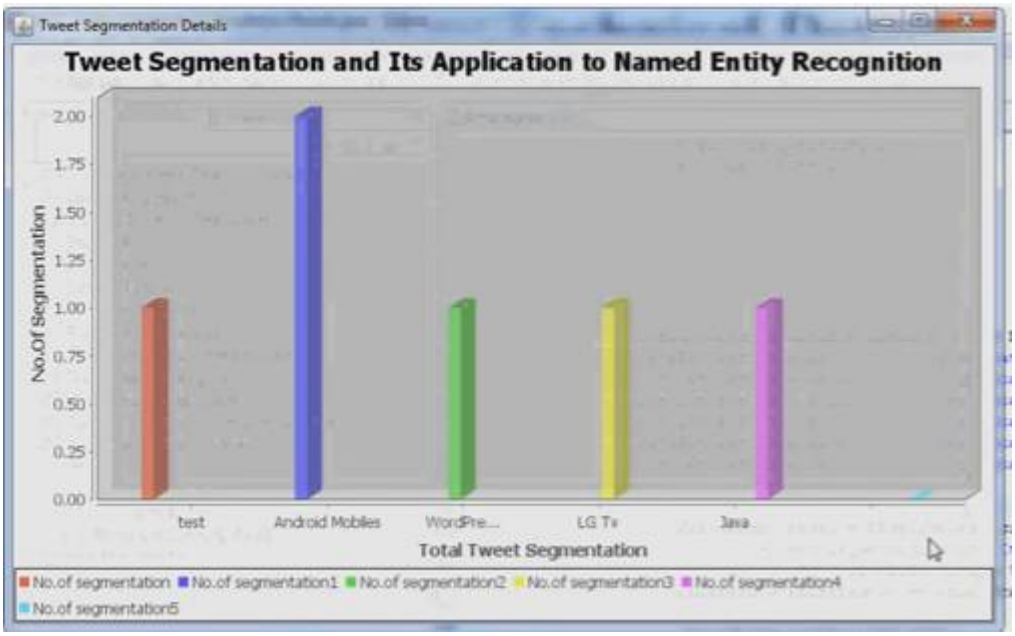

Fig. 5 Tweet segmentation

IX. CONCLUSION

Twitter, as a new type of social media, is gaining importance and has attracted great interests and from both industry and academia. Many private and public organizations have been reported to monitor Twitter stream to collect and understand users 'opinions about the organizations. Nevertheless, it is practically infeasible and unnecessary to monitor and listen the whole Twitter stream, due to it extremely large volume. Therefore, targeted Twitter streams are usually monitored instead. The HybridSeg framework which segments tweets into meaningful phrases called segments using both global and local context. Through the framework, it is demonstrated that local linguistic features are more reliable than term-dependency in guiding the segmentation process. This finding then opens opportunities for tools developed for formal text to be applied to tweets which are believed to be much noisy than formal text. Tweet segmentation helps to preserve the semantic meaning of tweets, which benefits many downstream applications like e.g., named entity recognition. And a segment-based named entity recognition method achieves much better accuracy than the word-based alternative.

FUTURE WORK

The future extension of our work is to further improve the segmentation quality by considering more local factors. And second improvement is to explore the effectiveness of the segmentation-based representation for tasks like tweets summarization, search, hashtag recommendation, etc. Also in future we will add large dataset for the segmentation purpose.

REFERENCES

1. M.Chenliang Li, Aixin Sun, JianshuWeng, and Qi He, ―Tweet Segmentation and Its Application to Named Entity Recognition‖ IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 27, NO. 2, FEBRUARY 2015
2. X. Liu, X. Zhou, Z. Fu, F. Wei, and M. Zhou, ―Exacting social events for tweets using a factor graph,‖inProc. AAAI Conf. Artif. Intell., 2012.
3. X. Liu, S. Zhang, F. Wei, and M. Zhou, ―Recognizing named entities in tweets,‖ in Proc. 49th Annu. Meeting Assoc. Comput. Linguistics: Human Language Technol., 2011.pp. 359-367
4. C. Li, A. Sun, and A. Datta, ―Twevent: segment-based event detection from tweets,‖ in Proc. 21st ACM Int. Conf. Inf. Knowl. Manage., 2012, pp. 155–164.

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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 4, April 2016

5. 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, 2013
6. A. Cui, M. Zhang, Y. Liu, S. Ma, and K. Zhang, ―Discover breaking events with popular hashtags in twitter,‖ in Proc. 21st ACM Int. Conf. Inf. Knowl. Manage., 2012
7. H. Ney, U. Essen, and R. Kneser, ―On structuring probabilistic dependences in stochastic language modelling,‖Comput. Speech Language, vol. 8
8. K. Nishida, T. Hoshide, and K. Fujimura, ―Improving tweet stream classification by detecting changes in word probability,‖ in Proc. 35th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 2012.
9. X. Wang, A. McCallum, and X. Wei, ―Topical n-grams: Phrase and topic discovery, with an application to information retrieval,‖ in Proc. IEEE 7th Int. Conf. Data Mining, 2007.
10. X. Wang, F. Wei, X. Liu, M. Zhou, and M. Zhang, ―Topic sentiment analysis in twitter: a graph-based hashtag sentiment classification approach,‖ in Proc. 20th ACM Int. Conf. Inf. Knowl. Manage., 2011, pp. 1031–1040.
11. L. Ratinov and D. Roth, ―Design challenges and misconceptions in named entity recognition,‖ in Proc. 13th Conf. Comput. Natural Language Learn., 2009, pp. 147–155
12. 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, 2012