RESEARCH ARTICLE OPEN ACCESS

TwiSeg NER-Tweet Segmentation Using Named Entity Recognition

Vijay Choure[1], Kavita Mahajan[2],Nikhil Patil[3], Aishwarya Naik[4],
Mrs.Suvarna Potdukhe[5]
1,2,3,4,5(Department of Information Technology, RMD Sinhgad School of Engineering,SavitrabaiPhule Pune University,PUNE)

Abstract:

Now a days Twitter has provided a way to collect and understand user's opinions about many private or public organizations. All these organizations are reported for the sake to create and monitor the targeted Twitter streams to understand user's views about the organization. Usually a user-defined selection criteria is used to filter and construct the Targeted Twitter stream. There must be an application to detect early crisis and response with such target stream, that require a require a good Named Entity Recognition (NER) system for Twitter, which is able to automatically discover emerging named entities that is potentially linked to the crisis. However, many applications suffer severely from short nature of tweets and noise. We present a framework called *HybridSeg*, which easily extracts and well preserves the linguistic meaning or context information by first splitting the tweets into meaningful segments. The optimal segmentation of a tweet is found after the sum of stickiness score of its candidate segment is maximized.This computed stickiness score considers the probability of segment whether belongs to global context(i.e., being a English phrase) or belongs to local context(i.e., being within a batch of tweets).The framework learns from both contexts.It also has the ability to learn from pseudo feedback. Also from the result of semantic analysis the proposed system provides with sentiment analysis.

Keywords — **NER,Hybrid Segmentation,Natural Language Processing**

I. INTRODUCTION

Sites like twitter have found new ways so that people can find, share and spreadtimely information many organization have been given an account to make a monitortarget twitter stream to understand user opinion and called them to twitter streamis constructed by filtering the tweets with predefined selection criteria. It is essentialto understand tweet language for a huge frame of downstream application due to itscrucial business value of timely information. The tweet length is limited but thereare no restrictions on its writing styles. The tweet often contain misspelling, informalabbreviations and grammatical errors.

The word level language models often prove to be less reliable for error proneand short nature of tweets. For example, given a tweet "I call he, but no answer. He phone in the bag, she dancin," there is no clue to guess its true theme by disregarding word order (i.e., bag-of-word model). On the other hand, the semantic phrases ornamed entities can be used in order to well preserve the semantic information of tweetsdespite of its noisy nature. So we focus on thetweet segmentation task, that splitsa tweet into a sequence of segments. These segments preserve the semantic meaningof tweets more error-free than each of its component words from this semantic analysisresult, sentiment analysis is done that might be good or bad, positive or negative.

II. GOALS & OBJECTIVES

The goal or objective is to find the optimal segmentation of tweet form its candidate segments by maximizing their stickiness scores.the stickiness score takes into account the probability of the segment ,that is,is the segment a English phrase(global context) or is a phrase withinthe batch of tweets(local context).for latter,we evaluate two models for deriving local context and term-dependency in batch of tweets.Instead of using the global context alone more segmentation can be improved by using both global and local context.We want to show that by applying the segment based parts -of -speech tagging in NER higher accuracy can be achieved.

III. EXISTING SYSTEM

- `Predefined criteria' is used to filter tweets.
- Tweet stream is constructed based on factors such as geographical location, time span and predefined keywords.
- Word level models are used for semantic analysis.

The system fails to analyze error prone and short tweets. It loses the semantic meaning in case of grammatically wrong tweets. System fails for SMS/NET Lingo. Precision accuracy is less.

IV. SYSTEM ARCHITECTURE

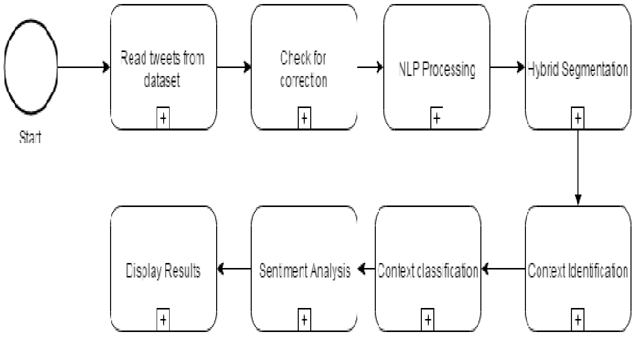

Fig1. System Architecture

V. METHODOLOGY

NLP (Natural Language Processing)

Natural language processing (NLP) is the ability or technique of a computer program to unsderstandhuman speech. NLP is a element artificial intelligence.

Sentiment Analysis

Sentiment is equal to feelings like attitude, subjective impressions,opinions,emotions and not facts.Generally, a binary opposition in opinions is assumed for/against, like/dislike, good/bad, etc. There can be some sentiment analysis jargonlike Semantic Orientation, Polarity.Semantic Analysis can be done using statistics, NLP or machine learning methodsto identify ,extractor otherwise characterize the sentiment content .
Sometimes it is also referred to as opinion mining, although the importance is on Extraction.

Hybrid Segmentation

- Here we are proposing a framework named HybridSeg.
- HybridSeg is further divided into 4 components as shown in fig. 2

In the proposed HybridSeg framework,the tweets are segmented in batch mode. Using a fixed time interval (e.g. a day)tweets are assembled into batches by their publication time.Each batch of tweets are then segmented by HybridSeg collectively.

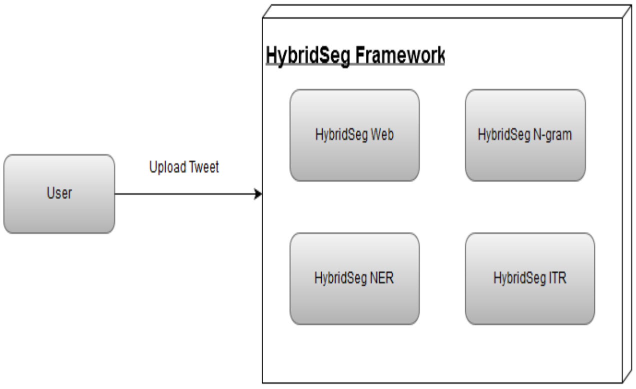

Fig2. Hybrid Segmentation Block Diagram

Named-Entity Recognition (NER)

There are predefined categories such as names of organizations, persons, quantities, monetary values, expressions of times, persons, percentages, etc.The task of NER is to extract information that locates and elements are classified into these predefined categories.

N-gram

N-grams are extensively used in text mining and natural language processing.N-grams are set of co-occuring words and while computing the N-grams we move one word forward.For example,the sequence "to be or not to be"then n-grams areto be,
be or,
or not,
not to,
to be.

here we have five n-grams.Notice that here we moved from to->be to be->or to or->not to not->to to to->be essentially moving on word forward.If n-grams would be :

to be or,
be or not,
or not to,
not to be ,

here we have 4 n-grams.

If N=1 referred to as Unigram,
If N=2 referred to as Bigram,
If N=3 referred to as Trigrams,
If N> called as four grams or five grams and so on.

HybridSegWeb learns from global contextonly,HybridSegIter learns from pseudo feedback
iteratively on top of HybridSegNER.

VI. PROCESS DESCRIPTION

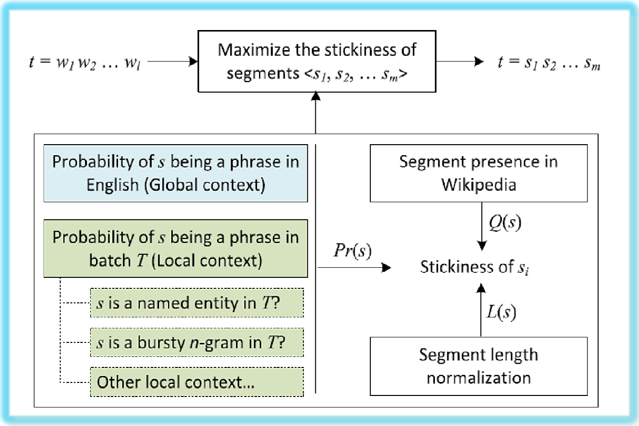
$t = w_1 w_2 \ldots w_l$
Maximize the stickiness of segments $<s_1, s_2, \ldots s_m>$
$t = s_1 s_2 \ldots s_m$
Probability of s being a phrase in English (Global context)
Probability of s being a phrase in batch T (Local context)
s is a named entity in T?
s is a bursty n-gram in T?
Other local context...
$Pr(s)$
Segment presence in Wikipedia
$Q(s)$
Stickiness of s_i
$L(s)$
Segment length normalization

Fig3.Process Description

Tweet Segmentation

Given a tweet t from batch T the problem of tweet segmentation is to split the l word in $t = w_1, w_2, w_3 \ldots w_l$ into $m <= l$ consecutive segments, $t = s_1, s_2, s_3, \ldots s_m$, where each segment s_i contains more than one words.The tweet segmentation problem is formulated as an optimization problem that maximizes the sum of *stickiness* scores of the m segments, shown in above fig. A high stickiness score of segment s indicates that the phrase appears "more than by chance", and further splitting of the phrase could change its meaning or word collocation.Formally, let $Ć(s)$ denote the stickiness function of segment s. The optimal segmentation is defined in the following:

$$\arg \max_{s_1, \ldots, s_m} \sum_{i=1}^{m} \mathcal{C}(s_i).$$

Dynamic programming is used to derive the the optimal segmentation with $O(l)$ time complexity.

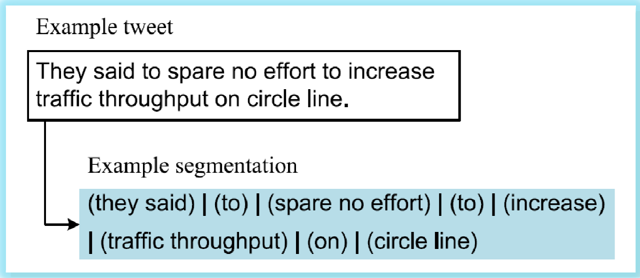
Example tweet
They said to spare no effort to increase traffic throughput on circle line.
Example segmentation
(they said) | (to) | (spare no effort) | (to) | (increase) | (traffic throughput) | (on) | (circle line)

Fig. 2, the segment in a stickiness function takes in 3 factors:

(i) Length normalization *L(s),* (ii) the segment's presence in Wikipedia *Q(s)*, and segment's phraseness *Pr(s)* , or the probability of s being a phrase of global and local contexts. The stickiness of s, *Ć(s)*, is formally defined in Eq. below, which captures the three factors:

$$\mathcal{C}(s) = \mathcal{L}(s) \cdot e^{Q(s)} \cdot \frac{2}{1 + e^{-SCP(s)}}.$$

Length normalization

Tweet segmentation is to extract meaningful phrases, longer segments are preferred for topically specific meanings

Presence in Wikipedia

In our framework,for valid names or phrases Wikipedia serves as an external dictionary.With lots of informal abbreviations and grammatical errors tweets are considered noisy. However, tweets are posted mainly for sharing of information and communication for many purposes.

VII. PROBLEM TYPES

There are two types of problems:

Polynomial (P)

The system accepts input, and we get the output in fixed polynomial time, the inputlarge or small, or simple or complex. Applications of polynomial type are rare. Onesuch example is hash table. A hashtable finds index for a data to be inserted in fixedamount of time because its uses hash function to find index. So for finding index 1 or 100 the time is fixed, which is not in case of sequential search for index. Ourapplication is not of type P because if does not give result in fixed polynomial time.

Non-Polynomial(NP)

There are two sub-types of NP Problems:

[A]NP-Hard:

The system accepts input, but there is no guarantee that we will get the output. Suchsystems do not exist because no one will use the system if there is no guarantee thesystem works for any inputs. Hence our application is again not of NP-Hard typebecause we want to build a system that never fails and guarantees output.

Ex: Turing Machine Halting problem. (You can search for it over internet)

[B] NP-Complete:

Our applicationis NP-Complete.The system accepts input, and we get the output invariable non-polynomial time. Almost all or maximum systems are of NP-Completetype. Even our application is of NP complete type because it guarantees output butnot in fixed amount of time. Now our output time varies because of the input. In ourapplication the core input of system is tweet dataset. When tweets are uploaded byuser, the system (hybridSeg) starts batch processing and determines tweets semanticmeaning. Thus our system produces con_rmed output on tweet dataset as input, alsoour systems performance and output is heavily dependent on no. of tweetsuploadedby the user. Hence our application is NP-complete.

VIII. PROPSOSED SYSTEM

- A novel framework HybridSeg is proposed for tweet segmentation.
- Uses Named Entity Recognition (NER).
- Word level models are replaced with Segmentation model.
- Segmentation is done in batch mode.
- 'HybridSeg' finds the optimal segmentation of tweet by maximizing sum of stickiness scores
- Local linguistic features are more reliable than local context with term dependency
- High Accuracy due to 'NER'
- Despite of noisy tweets, semantic meaning is preserved

IX. CONCLUSION

We have been studying the hybridseg framework which segments tweets into meaningful phrases called segments using both global and local context. Through this framework, we shall demonstrate that the local linguistic features are more reliable than term-dependency in guiding the segmentation process. On the Studies based on Tweet segmentation we found that it helps to preserve the semantic meaning o tweets, which subsequently bene_ts many downstream applications, e.g., named entity-recognition. By comparing papers, we may come to a conclusion that a segment-based named entity recognition method achieves much better accuracy than the word-based alternative, for segmentation purpose hybridseg framework will be used which outputs semantic analysis result. This result then will further perform sentiment analysis results in graphical format. And further user will be provided with both results of semantic analysis and sentiment analysis.

X. FUTURE WORK

1. We plan on adding support for ‘hash tags’.
2. Live connectivity to twitter data can be achieved.
3. Use of Big-Data for scalability.

ACKNOWLEDGMENT

We take this opportunity to thank our project guide Mrs. Suvarna Potdukhe and Head of Department Mrs. Sweta Kale for their valuable guidance and immense support in providing all the necessary facilities,moral support to encourage us which were indispensable in the completion of this project report. We are also thankful to all the staff members of the Department of Information Technology of RMD Sinhgad School of Engineering, Warje for their valuable time, support, comments, suggestion and persuasion. We would also like to thank tha institute for providing the required facilities, Internet access and important books.

REFERENCES

1. Ritter, S. Clark, Mausam, and O. Etzioni, Named entity recognition in tweets: An experimental study, in Proc. Conf. Empirical Methods Natural LanguageProcess., 2011, pp. 15241534.
2. X. Liu, S. Zhang, F. Wei, and M. Zhou, Recognizing named entities in tweets, inProc. 49th Annu. Meeting Assoc. Comput. Linguistics: Human Language Technol.,2011, pp. 359367.
3. L. Ratinov and D. Roth, Design challenges and misconceptions in named entityrecognition, in Proc. 13th Conf. Comput. Natural Language Learn., 2009, pp.147155.
4. J. R. Finkel, T. Grenager, and C. Manning, Incorporating nonlocal informationinto information extraction systems by Gibbs sampling, in Proc. 43rd Annu.Meeting Assoc. Comput. Linguistics, 2005, pp. 363370.
5. G. Zhou and J. Su, Named entity recognition using an hmmbased chunk tagger,in Proc. 40th Annu. Meeting Assoc. Comput. Linguistics, 2002, pp. 473480.
6. K. Gimpel, N. Schneider, B. OConnor, D. Das, D. Mills, J. Eisenstein, M.Heilman, D. Yogatama, J. Flanigan, and N. A. Smith, Part-of-speech tagging for twitter: annotation, features, and experiments, in Proc. 49th Annu. Meeting. Assoc.Comput. Linguistics: Human Language Technol., 2011, pp. 4247.
7. B. Han and T. Baldwin, Lexical normalisation of short text messages: Maknsensa twitter, in Proc. 49th Annu. Meeting. Assoc. Comput. Linguistics: HumanLanguage Technol., 2011, pp. 368378.
8. F. C. T. Chua, W. W. Cohen, J. Betteridge, and E.-P. Lim, Community-basedclassi_cation of noun phrases in twitter, in Proc. 21st ACM Int. Conf. Inf. Knowl.Manage., 2012, pp. 17021706.
9. W.Jiang,M.Sun, Y.Lu,Y.Yang, andQ. Liu,Discriminativelearningwith naturalannotations:Wordsegmentation as a case study, in Proc.Annu.MeetingAssoc.Comput.Linguistics, 2013, pp.761769.