International Journal of Engineering Research-Online
A Peer Reviewed International Journal
Email:editorijoer@gmail.com http://www.ijoer.in

Vol.4., Issue.2., 2016
(Mar-Apr)

RESEARCH ARTICLE

ISSN: 2321-7758

NAMED ENTITY RECOGNITION BY TWEET SEGMENTATION

SWATHY.K.SHAJI, Dr.SOBHANA N.V
Department of Computer Science and Engineering,
Rajiv Gandhi Institute of Technology, Kottayam

ABSTRACT
Named entity recognition is a task of identifying the named entities such as person, organizations, locations, expressions of times, quantities etc. from a huge corpus[1]. It is one of the major tasks in natural language processing. The paper proposes a novel framework for tweet segmentation in batch mode. By splitting the tweets into meaningful segments the semantic or context information is well preserved. Here evaluates two NER algorithms and the results shows that the identification of named entities can be significantly improved by using the proposed framework along with POS tagging
Keywords: Twitter stream, Tweet segmentation, Natural language processing, Wikipedia, Named Entity Recognition, Precision, Recall, F-measure

INTRODUCTION

Twitter has attracted millions of users to share their information creating huge volume of data produced every day. It is a very difficult and time consuming task to handle this huge amount of data. Thus the segmentation of tweets and identifying the named entities is considered to be a tedious one. Given that the tweet has limited length (140 characters), and there is no restrictions in expressing one's opinion. The short and error prone nature of tweets makes it unreliable for downstream applications.

This paper mainly focuses on the task of tweet segmentation along with its application to named entity recognition. Here propose a novel Hybrid framework [1] for segmenting the tweets. Hybrid segment the tweets based on batch mode. For example the tweets under a particular time period are grouped into batches and thus continue the segmentation. Hybrid framework learns from both local and global context. Local context considers the probability that a segment be a meaningful phrase within a batch of tweets. Global context on the other hand considers the probability that the segment being a phrase in English, it is actually done by identifying the meaningful segments from Wikipedia, Microsoft Web N Gram [1] etc.

Tweet segmentation is done by splitting the tweets into consecutive n-grams which is called a segment. The segment can be a named entity, a semantically meaningful information unit or any other type of phrases that appears more than once in a group of tweets. To improve the segmentation quality, here propose a Hybrid framework that learns from both local and global context. Global context has been used for tweet segmentation, but the accuracy of segmentation is low when compared with segmentation using both local context and

global context. Since the tweets are posted for the purpose of information sharing analyze tweets on the basis of content from Wikipedia is not at all good for obtaining the desired accuracy. The global context derived from web pages such as Microsoft Web N Gram [1] corpus or from Wikipedia thus helps in identifying the meaningful segments in tweets. Local context deals with analyzing the tweets under acertain publication time (e.g. tweets published in 1 day). Some of the local phrases cannot be easily identified by using the global context alone. For example consider the case of a song 'she dancin' which cannot be identified when considering the external knowledge bases only. This phrase can be easily identified when considering the tweets from a particular day. The word that is occurring more than once can be treated as a phrase and hence should be preserved. Next task for segmentation is based on learning from pseudofeedback. The segments recognized based on local context having high confidence normally serve as a feedback for the extraction of more meaningful segments. The method using pseudo feedback has been conducted iteratively and the iterative learning mechanism is termed as HybridSegiter.

As an application of tweet segmentation to named entity recognition here evaluate two NER algorithms. Both algorithms are unsupervised in nature. One of the algorithm that exploits the co-occurrence of named entities in tweets by applying the Random walk model. The random walk model builds a segment graph, in the graph the nodes represent the segments identified by the Hybrid framework. An edge exists between the nodes if and only if the segments co-occur in some tweets. The random walk model is then applied to the segment graph for identifying the named entities. Second algorithm exploits part-of-speech tags for the identification of named entities as noun phrases. A segment may appear in multiple tweets. The constituent words of that particular phrase are assigned with different POS tags. Then estimate the likelihood of segment being a phrase by considering the POS of the individual words in a segment.

RELATED WORKS

Ritter et al [9] suggested an NLP pipeline which begins with part-of-speech tagging and ends with named entity recognition. Named entity recognition in tweets is a difficult task because of the noisy nature of tweets, also tweet contain distinct entity types which cannot be easily identified. Most of the NLP task including named entity recognition and Information extraction applies the part-of-speech tagging. Ritter et al [9] compares the T-POS with Stanford POS tagger. The misclassification made by Stanford POS tagger can be significantly reduced by using the T-POS model.

Shallow parsing[9] is applied for the identification of noun phrases, verb phrases, propositional phrases etc. The method can be effectively used for Information Extraction and named entity recognition. Capitalization of words is another way of identifying the named entities. For better identification of named entities it is important to determine whether the capitalization is informative or uninformative. Ritter et al [9] build a capitalization classifier named T-CAP[9] to effectively determine whether capitalization is informative or not. It also suggests the method for segmenting and classifying the named entities.The strengths of the suggested work are: (i) evaluated many existing tools for POS tagging[2], chunking, and named entity recognition. Found that T-POS outperforms the Stanford POS tagger by 41%.(ii) Presented a distantly supervised approach based on labeled LDA which significantly improves the F1 score by 25%.(iii) Exploited large dictionaries of entities from Freebase and thus able to identify more type of entities.

Named entity recognition (NER) (also known as entity identification and entity extraction) is a subtask of informationGimpel et al [2] develop a tag set, annotate data and develop features. The tagging result brings 90% accuracy. They build an English POS tagger especially for twitter data. The major contributions of their work were: (i) Developed a POS tag set for twitter (ii) They manually tagged 1.827 tweets (iii) Develop features for twitter POS tagging and conducted experiments

in order to evaluate the features.

Annotation proceeds through three stages. In stage 0 they developed a set of 20 coarse grained tags, and then pre-tagged the tweets using a WSJ-trained POS tagger[2]. In stage 1 2217 tweets were distributed, of them 390 were found to be non-English and then removed. In stage 2 Two annotators reviewed all the English tweets from stage1. The major advantage of their work is that:(i) Developed a POS tagger[2] for twitter and made it available for research community.(ii) The approach can be applied to other linguistic analysis in social media.(iii) The annotated data can be used for semi supervised learning.

Normalization of ill formed words in tweets has been studied by Bo Han et al [3]. Tweets usually are noisy thus make it unsuitable for NLP[3]. The method which proposes a mechanism to detect the ill formed words in tweets then generates the correction words based on morphophonemic similarity. This is usually done by analyzing word similarity and the context in which they are tweeting.The ill formed words usually include all the individual instance of typos, adhocabbreviations[3], phonetic substitutions and unconventional spellings. The paper proposes methods for conventional spell checking and recovering mechanism for the commonly used short hand abbreviations.

Xiaohua Liu et al[4] proposed a method to combine the K-Nearest Neighbors classifier with a linear conditional random field (CRF) model in order to tackle the challenges of the short and noisy nature of the tweets. KNN based classifier conduct pre-labeling for collecting global coarse evidence and CRF model[4] uses sequential labeling to obtain the fine grained information. The proposed approach usually uses a KNN classifier[4] along with CRF model[4]. KNN classifier[4] is adopted for word level classification. Following the classification a two stage prediction aggregation method is used. The pre-labeled result are then fed into linear CRF model, it actually conduct a fine grained NER on tweets.

The experiment analysis of tweet dataset shows that proposed approach considerably has high performance compared with the already existing one. Semi supervised learning is used in NER, it learns from both labeled and unlabeled data. It is normally being used when the labeled data is scarce and the unlabeled data is abundant.

Freddy chong tat chua et al [5] proposes a method to classify the noun phrases to some specific categories like politics, sports etc. Due to the conversational nature of tweets, it is difficult to find out the known keywords. Thus a classification mechanism is proposed, that uses a feature vector. Feature vector is defined based on the user's behavior and his social activities. Part of Speech (POS) Information[5]: The POS of the current and/or the surrounding word(s) can be used as features.A sentence level semantic category recognizer (SentReg)[5] has been used for classifying NP. Here use NP+LDA (latent dirichlet allocation)[5] for finding topic distribution of authors. By using the technique frequent topics on which the author is tweeting can be easily found out.

SilviuCucerzan [6] proposes a large scale system for identifying named entities based on information extracted from a large knowledge base. The paper discusses both named entity recognition and disambiguation. Disambiguation [6]employs a vast amount of contextual and category information extracted fromWikipedia. Information extraction from Wikipedia usually has to deal with four types of articles namely entity pages, redirection pages, disambiguation pages and list pages.

JianfangGao [7] addresses issues in Chinese natural language processing. The proposed approach has three unique components namely taxonomy of Chinese words, unified approach of word breaking and unknown word detection and finally customizable display of word segmentation. The paper defines a taxonomy system that categorize Chinese words [7] into five types namely lexicon words, morphologically derived words, factoids, named entities and new words. Wenbin Jiang [8] et al utilizes internet as an external knowledge base which has massive natural annotations. The annotations may include font, layout, color and hyperlink. They propose a

character classification model that must factorize the whole prediction into atomic predictions. The paper discusses Word segmentation[8], which splits a sequence into subsequence each of which represents a meaningful word. They build a perceptron training algorithm to train the classifier for the character classification problem. The annotation difference between the outputs of constraint decoding and normal decoding is used to train the classifier.

IITWEET SEGMENTATION

Given a tweet t from a batch T, the task is to split the tweets into meaningful segments. The word in tweet t can be represented as t = $w_1,w_2,\ldots\ldots,w_l$ the segmented tweets after applying the segmenting mechanism is represented as t = $s_1, s_2,\ldots\ldots,s_m$ .Stickness score is another important parameter that is to be applied to evaluate the presence of a segment that is being a phrase. High stickness score means that the phrase is appearing more than by chance. Stickness score has to deal with three parameters: (i) length normalization, Normalized length has been computed for determining whether the segment is longer or short. The segment having longer length should be considered of having some special meaning and should be preserved as such. The normalized segment length is given by L(s) = 1 if |s| = 1, thus L(s) can be computed by using the equation (1)[1]:

$$L(s) = \frac{|s|-1}{|s|} \quad (1)$$

(ii) Presence in Wikipedia: Wikipedia serve as an external data source for identifying whether the segments are meaningful phrase in English.

(iii) Segment phraseness: The last component that must estimate the probability that a segment being a valid phrase in English.

The diagrammatic illustration of the framework used for tweet segmentation is indicated in the Figure.1.

The diagrammatic illustration shows that the tweet may contain a number of words represented by t = $w_1,w_2,\ldots\ldots,w_l$corresponding words from tweets are analyzed for finding the segments having the highest stickness score.

The stickness score of segments has being found out by considering the probability of segment being phrase in a batch of tweets i.e. local context and also by considering whether segment is a meaningful phrase in English i.e. global context. This probability value along with the segment presence in Wikipedia and segment length Normalization is used for finding the stickness score of individual segments. Next step is the evaluation of segments having high stickness score. A segment having high stickness score means that it is a phrase that appears more than once and should be preserved as such. Thus the segment having high stickness score should be preserved as such. The list of segments thus obtained should be represented ast = $s_1, s_2,\ldots\ldots,s_m$. Segment thus obtained can be a word or can also be a phrase having more than one word or a group of words.

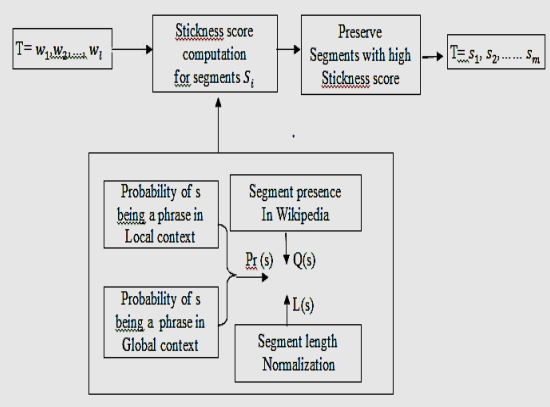

Figure.1. Hybrid Framework

Observations for Segmentation

Tweets are considered to be noisy in nature due to the presence of informal abbreviations and grammatical errors. Tweets are being posted mainly for the purpose of information sharing and communication among a group of individuals. Several observations should be made for segmenting the tweets. Observations show the difficulties in identifying the named entities from tweets.

Observation 1: Tweets may often contain the phrases that appear most frequently, the common phrases in English should be preserved as such and should not be further segmented. The phrases in English can be identified from the tweets by considering its presence from Microsoft Web N Gram corpus. This corpus provides a good estimate of the commonly used phrases in English. The tweet

segmentation framework proposed here segment the tweets by preserving the named entities and commonly used phrases in English.

Observation 2: Local context [1] should be considered for the identification of some words or phrase that cannot be identified when considering the global context (Wikipedia) alone. Local context evaluation should be made by grouping the tweets under a certain publication time (say one day). The phrases that appear in more than one tweet should be preserved as such. The linguistic features in tweets often helps in the identification of named entities with relatively high accuracy.

Observation 3: Global context [1] evaluation should be done by evaluating the segments presence in Wikipedia. Wikipedia can be considered as a huge corpus having a lot of articles. Most of the relevant segments can be found out by using this huge corpus.

III EXPERIMENTAL DESIGN FRAMEWORK

The design framework that we are using for splitting the tweets is the Hybrid framework that learns from both local and global context. Local context is one that split the tweets by considering the segment presence within a batch of tweets. The tweet under a certain publication period is being grouped into a batch and then segment the tweet based on its presence within the batch of tweets. Global context [1] on the other hand segment the tweets by considering its presence in the Wikipedia and Microsoft Web N Gram[1]. Wikipedia is a huge corpus that contain a number of articles and the Microsoft Web N Gram that provides a good estimate of the commonly used phrases in English. The design framework for segmenting the tweets is being indicated in the Figure.2.Figure.2. shows the major processing steps for the tweet segmentation and the identification of named entities. First we have the tweet dataset. Our task is to group the tweets into a batch under a certain publication time (say one day). The tweets thusobtained should be given to a Hybrid segmentationframework that learns from both global and local context. The segments thus obtained should be evaluated for identifying the named entities. Here evaluate two NER algorithms Random walk model and POS Tagging. Random walk model identifies the named entities by learning from a segment graph and POS Tagging is being used for the identification of noun phrases among the tweets.

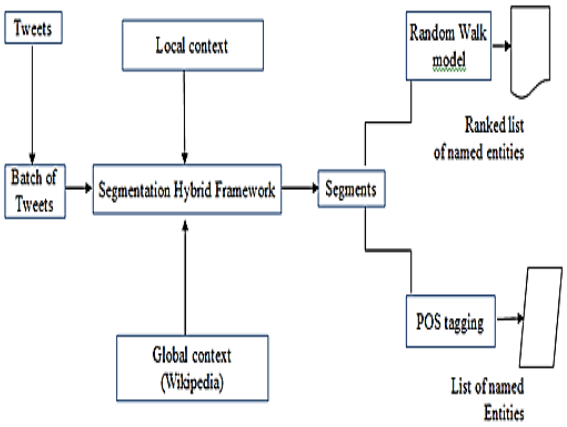

Figure.2. Processing steps for the identification of Named entities

Learning from Global Context [1]

The tweets are being posted for the purpose of information sharing and communication among a group of individuals. Due to the noisy nature of tweets it is a very difficult task to segment the tweets and thus identifying the named entities and semantically meaningful segments. Global context [1] actually deals with the segment comparison based on the Wikipedia content and Microsoft Web N Gram content. Global context derived from web pages or Wikipedia thus helps in identifying the meaningful segments in tweets and thereby result in the identification of named entities.

Many text mining and natural language processing tasks such as text categorization, topic detection, information extraction etc has to deal with the use of Wikipedia content. Entity linking [9] is applied here for finding the segments from the Wikipedia. Entity linking needs an external knowledge base for deriving the entity mentions, the commonly used knowledge base for entity extraction, linking, named entity disambiguation is the Wikipedia.

Learning from Local Context [2]

Local context that described in the paper has to deal with the phrase comparison within a batch of tweets. The tweets from the tweet dataset should be grouped into a batch. The batch normally contains the tweets under a certain publication time (say one day). Tweets are highly time sensitive in nature and therefore finding some phrases by considering the global context alone is not an efficient way of identification. Some phrases like 'She Dancin' a music album cannot be found out when looking on the global context alone. But the phrase can be easily identified when looking on the group of tweets related with the particular phrase. Tweets posted during a certain time period mostly use the phrase and therefore the identification of phrase within the batch of tweets is not a very difficult task. Local context along with the comparison using global context definitely improves the accuracy of tweet segmentation.

Learning from Local Collocation [1]

The word Collocation deals with the arbitrary recurrence of word combination in a group of documents. Consider a segment$w_1w_2w_3$ having three words represented by then the possible combination of these words such as w_1w_2, w_2w_3, $w_3w_2w_1$etc are considered to be positively correlated with each other. This shows the way that the sub n gram of a meaningful phrase correlates with each other. Hence this also provides a way to find the meaningful phrases from tweets by learning from the local collocations. A segment can be considered to be a valid one if it co-occur together in a group of tweets. If the sample size is larger, then the word collocations can be considered to be a suitable measure to find the valid segments. If the sample size is smaller, then global context is the better way of finding the meaningful segments.

IV SEGMENT BASED NAMED ENTITY RECOGNITION

Named entity recognition is the task of recognizing the proper nouns or entities in text and relating them to a predefined set of categories. Most of the NER systems are based on analyzing the patterns of POS tags. The categories to which the entities belong may be location name, organization, date, expression, percentage, person name etc. Named entity recognition [4] has a number of applications in the field of natural language processing. It has been used in information extraction, parsing, machine translation, question answering etc. NER systems have been mostly used in the field of bioinformatics and molecular biology for extracting the entities. Most of the NER systems are word based, here we employ the segmentation method for the identification of named entities. Named entity recognition task has been performed for evaluating the performance of two algorithms: Random walk model and POS Tagging. Both algorithms are exploited for the identification of named entities.

NER by Random Walk model

The random walk model algorithm exploits the co-occurrence of named entities in a group of tweets. The random walk model algorithm builds a segment graph by identifying the segments as nodes in the graph. The segments can be joined by an edge if they co-occur together in a group of tweets. The algorithm which forms a graph based structure.The co-occurrence of these segments can be thus used for the identification of named entities. The weight of an edge is being evaluated by using the Jaccard coefficient [1] between the corresponding segments. Let p(s) denotes the stationary probability of a segment s; the segment is thus weighted by [1],

$$Y(s) = e^{Q(s)}.P(s) \quad (2)$$

It indicates that the segment that mostly appears as an anchor text in Wikipedia is identified as a named entity.

NER by POS Tagger [5]

The NER algorithm using the POS tagger exploits the part of speech tags in tweets. Here the constituent words in a segment are assigned with different POS tags. We estimate the likelihood of a segment being a named entity by analyzing the POS tags of its constituent words in all its appearances. The table 1 shows some of the commonly used tags, its descriptions and examples.

Table.1. POS Tags and its description

Tag	Description	Example
N	Common Noun (NN,NNS)	books someone
M	Proper noun + Verbal	Mark'll
Z	Proper noun + Possessive	India's
S	Nominal + Possessive pronoun	book's someone's
$	numeral	2010; four; 9.25
A	adjective	beautiful
X	Existential there	both
#	Hash tag	#acl
@	At mention	@mark
U	URL or Email address	http://xyz.com
E	emoticon	:-), (:

The word in the segment has been assigned with different POS tags. The named entities thus identified can be assigned to different categories which are predefined one.

V EXPERIMENTAL RESULTS

The experiment has been conducted by using the tweet dataset having 5000 tweets. Experiment has been conducted to evaluate the accuracy of two NER algorithms: Random walk model [1] and POS Tagger [1]. We evaluate the performance of the system by making comparison with the already developed systems using global context alone. Experimental results should be made by comparing different methods: (i) Segmentation accuracy of Hybrid framework[1] using global context (ii) Segmentation framework using local and global context [1](iii) Named entity identification by random walk model (iv) Named entity recognition by POS Tagger.

Experimental Settings

Experiment has been conducted by using the tweet dataset having 5000 tweets. The tweets should be such that some phrases should be repeating in more than one tweet. Experiment has been done to evaluate the segmentation accuracy of two algorithms: Random walk model and POS Tagger. We used the Wikipedia dump for evaluation. This dump may contain 32,46,821 articles and there are about 4,342,732 distinct entities appeared as anchor text in these articles.

The evaluation of both algorithms shows that the segmentation accuracy can be significantly improved by using the Hybrid framework along with POS Tagging. Here the tweet dataset having 5000 tweets were evaluated, the dataset may also include some additional fields like author name, date published and the tweets. The date published field should be used for grouping the tweets under a certain publication time. The phrases that appear more than once in a group of tweets should be identified as segment and should not be further segmented. It thus helps in preserving the semantic meaning of the tweets. The evaluation of the algorithms has been done by taking the time as a parameter.

Evaluation

The task of tweets segmentation is to split the tweets into meaningful segments. In ideal situation the tweet segmentation method should be evaluated by comparing its segmentation result with the manually segmented tweets. It is reasonably a big task to manually segment the big sized tweets. We choose to evaluate the method by determining whether the segments are correctly detected. Here we use the Recall measure denoted by Re, which denotes the percentage of manually annotated tweets that iscorrectly split as segments. We evaluate two segmentation methods in the experiments:

(i) HybridWeb learns from global context only.

(ii) HybridNER learns from global context and local context.

Then we evaluate the time it takes for segmenting the tweets in both the methods. First method has been implemented by using the random walk model and thereby splitting the segments. This has been done by using the URL for searching the key word (http://dumps.wikimedia.org/enwiki/). It has been

performed by comparing the words or phrase with the Wikipedia content. Second method is based on learning from both local and global context. Learning from local context usually have to deal with the grouping of tweets under a certain publication day (say 1 day), then segment the tweets based on both Random walk model and by using POS tagging [2].

Performance Measure

The execution time for both algorithm can be evaluated by considering the system time at the time of execution. Since both systems use the same dataset for evaluation the variation in time shows the efficiency of one system over the other. The tweets are segmented and the identification of phrases can be better understood by the resulting output. As an application of this segmentation, here describes the task of named entity recognition. The segmentation result is then passed to two NER algorithms: Random Walk model and POS tagging[2][4][5]. The result obtained from this evaluation is that the phrases having more than one word can be easily identified by using the technique.

Combination of words can also be identified and which can-not be further segmented. Here the segmentation is actually done by two ways: First do the segmentation only by comparing the occurrence of word in the Wikipedia. Second the segmentation done using both local context and global context. Local context comparison is made by grouping the tweets of a certain day and finding the similar words in the tweets. If a combination of words occur in more than one tweets the it should be identified as a named entity and should be listed.

The execution time difference of two method (Random walk model and POS tagging) when identifying phrases having varying length is indicated as in the form of a table format. Table 2 shows the comparison of Random walk model and POS tagger for varying length phrases.

Table.2. Execution time of two NER Algorithms

Number of words in phrase	Execution Time(ms) (Random walk model)	Execution Time(ms) (POS tagger)
ONE word	64	35
TWO word	74	40
THREE word	89	45
FOUR word	120	90

Thus the table shows the exact time difference between the two implemented algorithms. The time variation is due to the difference in the way they are segmenting the tweets. The POS tagger can be used for segmentation but it cannot list the named entities if it doesn't have a meaning in the context of tweets. Random walk model thus segment the tweets and identify all the phrases that appears in the tweets as named entities but in some cases the phrases aren't named entities but it will also be listed as a named entity.

As an application of tweet segmentation, here propose and evaluate two segment-based NER algorithms. Both algorithms are unsupervised in nature and take tweet segments as input. One algorithm exploits co-occurrence of named entities in targeted Twitter streams by applying random walk (RW) with the assumption that named entities are more likely to co-occur together.The other algorithm utilizes Part-of-Speech (POS) tags [2] of the constituent words in segments. The segments that are likely to be a noun phrase (NP) are considered as named entities. Our experimental results show that (i) the quality of tweet segmentation significantly affects the accuracy of NER, and (ii) POS-based NER method outperforms RW-based method on both data sets.

Another method of comparison is also needed to identify the difference in the way of identifying the named entities. Here implement two methods of identifying named entities by using global context alone and by using both local and global context. The way of identification using global context is done by comparing the tweet keywords with the Wikipedia content. Local context

comparison is made by grouping the tweets under a certain publication time and thus identifying the named entities.

Table 3 below shows some keywords identification using global context alone and by using local context.

Table.3. Identification of some keywords by global context and local context.

Keywords	Global Context	Local context
Traffic_throughput	Not recognized	Recognized
She_Dancin (music band)	Not recognized	Recognized
Circle line	Not recognized	Recognized
Circle	Recognized(search count:128982)	Recognized
Coloumbia_pictures	Not recognized	Recognized
Sailing competition	Not recognized	Recognized

The table thus shows that most of the phrases that we are using in tweets are left unrecognized when using the global context alone. If we are using both local and global context the unrecognized phrases can easily be identified thus improving the efficiency. This might be used, if we want to find the reviews of people about a particular lm. Suppose the lm is released in that particular day, then it is impossible to find the lm name when we are looking at the global context only. It can be easily be identified when we are using both local context and global context. At that particular day most of the tweets are related with the lm reviews so that the phrase that appears more than once can easily be identified by the two algorithms.

To draw conclusions about the performance of the tweet segmentation it should be compared with the other method for the segmentation.

Segmentation accuracy

We evaluate three segmentation methods in the experiments: (i) HybridSegWeb [1] learns from global context only, (ii) HybridSegNER [1] learns from global context and local context. HybridSegWeb [1] that learns from global context alone, that means the segmentation is done by comparing the keyword with the Wikipedia content. Table.4. reports the segmentation accuracy (recall measure) achieved by the two methods on the same data set.

Table.4. Recall value of Hybrid framework that learns from global context (HybridSegWeb) and that learns from both local and global context (HybridSegNER)

Method	Recall Value
HybridSegWeb	0.742
HybridSegNER	0.806

Three observations can be made from the results.

(i) HybridNER achieves significantly better segmentation accuracy than HybridWeb. It shows that local context does help to improve tweet segmentation quality largely.

(ii)Learning local context is more effective than learning from local word collocation in improving segmentation accuracy, thus HybridNER outperforms HybridWeb on tweet data sets.

Method analysis and Comparison

To evaluate the method and compare it with the already developed methods should deal with three types of segments:

(i)Fully detected segments (FS)[1]: all occurrences of the segments are detected from the batch of tweets

(ii)Missed segments (MS)[1]: not a single occurrence of the segment is detected from the previous iteration.

(iii)Partially detected segments (PS)[1]: some but not all occurrences of the segments are detected.

Table 5 shows the number of detected, partially detected and missed segments.

Table.5. Number of fully detected, partially detected and missed segments of two methods

Method	Fully Detected	Partially Detected	Missed segment
HybridSegWEB	3150	620	1230
HybridSegNER	4200	560	240

In information retrieval contexts, precision and

recall are defined in terms of a set of retrieved documents and a set of relevant documents. In the field of information retrieval, precision is the fraction of retrieved documents that are relevant to the query:

$$P = \frac{\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}}{\{\text{retrieved documents}\}} \quad (3)$$

Precision is also used with recall, the percent of all relevant documents that is returned by the search. The two measures are sometimes used together in the F1 Score (or f-measure) to provide a single measurement for a system. Recall in information retrieval is the fraction of the documents that are relevant and are successfully retrieved.

$$R = \frac{\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}}{\{\text{relevant documents}\}} (4)$$

In a set of documents recall is the number of correct results divided by the number of results that should have been returned. The F measure that combines precision and recall and should be evaluated by,

$$F = 2.\frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (5)$$

Evaluation Metric

The accuracy of NER is evaluated by Precision (P), Recall (R) and F1. P is the percentage of the recognized named entities that are truly named entities; R is the percentage of the named entities that are correctly recognized. The type of the named entity (e.g., person, location, and organization) is ignored. Similar to the segmentation recall measure, each occurrence of a named entity in a specific position of a tweet is considered as one instance.

Table 6 evaluate variations of the two methods namely GlobalSeg [1], HybridSegRW [1], HybridSegPOS [1] Here denotes HybridSegWeb [1] uses global context, and HybridSegPOS[1] is the best method that uses both global and local context along with POS Tagging. However, with better segmentation results, HybridSegPOS is much better than HybridSegRW [1]. By F1 measure, HybridPOS achieves the best NER result. We also observe that both the segment-based approaches HybridSegPOS [1]and HybridSegRW[1] favor the popular named entities.

Table.6. precision, recall and F measure values for three different segmentation tasks

Method	P	R	F
GlobalSeg	0.416	0.312	0.356
HybridSegRW	0.510	0.333	0.398
HybrisSegPOS	0.636	0.388	0.482

The graph which is being plotted by considering Recall, precision and F measure in Y-axis and Method of segmentation in X-axis. It is being indicated in the following Figure .3.

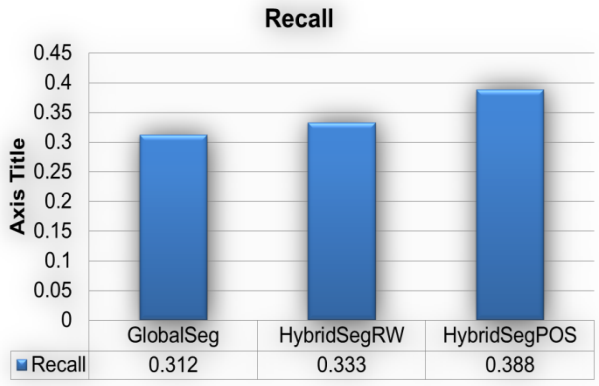

Figure.3. Recall value for three segmentation methods

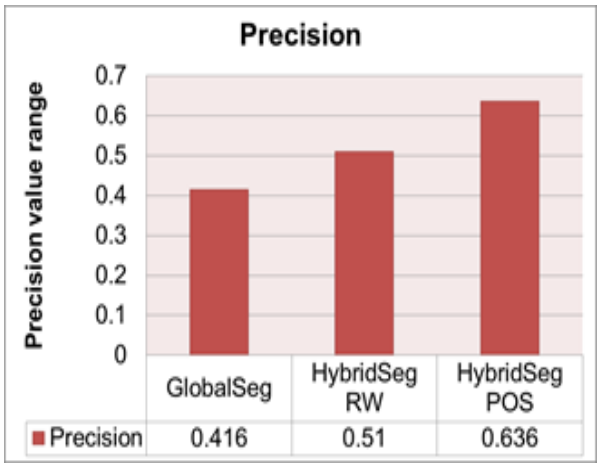

Figure.4. Precision value for three segmentation methods

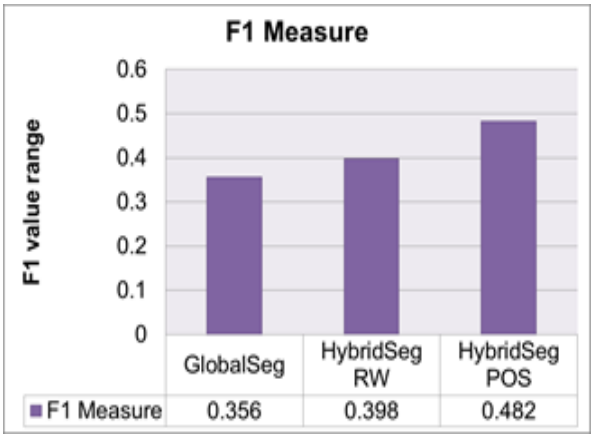

Figure.5. F measure value for three segmentation methods

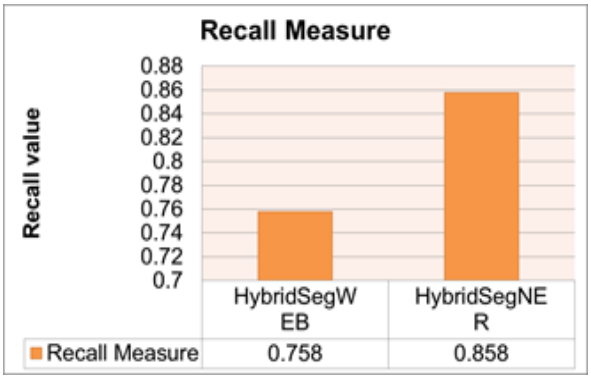

Figure.6. Recall value for two general segmentation methods

By considering the concept of fully detected, partially detected and missed segments we can plot a graph with method adopted on the X-axis and the segment category on the Y-axis. It can be indicated by using the Figure.7. below:

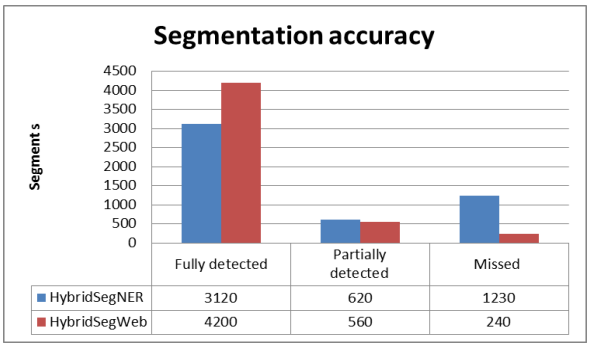

Figure.7. Graph showing the number of Fully detected, Partially detected and Missed segments in two Segmentation approaches

CONCLUSION

In this paper, we have made the task of tweet segmentation and thereby the identification of named entities. It is a difficult task to identify the named entities from a large corpus. Tweet segmentation based on HybridSeg [1] framework definitely improves the accuracy of identifying the named entities. Here evaluated two NER algorithms random walk model and POS Tagger, the results shows that the performance of POS Tagger is better as compared with the other algorithm. The work can be extended in such a way as to categorize the named entities detected by the proposed framework. This can be done byusing techniques of entity linking, summarization, entity relationship extraction etc.

REFERENCES

[1]. 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, 2015

[2]. K. Gimpel, N. Schneider, B. O'Connor, 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. 42–47.

[3]. B. Han and T. Baldwin, "Lexical normalisation of short text messages: Maknsens a #twitter," in Proc. 49th Annu. Meeting. Assoc. Comput. Linguistics: Human Language Technol., 2011, pp. 368–378.

[4]. 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.

[5]. F. C. T. Chua, W. W. Cohen, J. Betteridge, and E.-P.Lim, "Community-based classification of noun phrases in twitter," in Proc. 21st ACM Int. Conf. Inf.

Knowl.Manage., 2012, pp. 1702–1706.

[6]. [6] S. Cucerzan, "Large-scale named entity disambiguation based on wikipedia data," in Proc. Joint Conf. Empirical Methods Natural Language Process. Comput. Natural Language Learn., 2007, pp. 708–716.

[7]. J. Gao, M. Li, C. Huang, and A. Wu, "Chinese word segmentation and named entity recognition: A pragmatic approach," in Comput. Linguist., vol. 31, pp. 531–574, 2005.

[8]. W.Jiang,M.Sun, Y.L€u,Y.Yang, andQ. Liu,"Discriminativelearningwith natural annotations:Wordsegmentation as a casestudy,inProc.Annu.MeetingAssoc.Comput.Linguistics, 2013, pp.761–769.

[9]. A. Ritter, S. Clark, Mausam, and O. Etzioni, "Named entity recognition in tweets: An experimental study," in Proc. Conf. Empirical Methods Natural Language Process., 2011, pp. 1524–1534.