© 2016 IJSRSET | Volume 2 | Issue 2 | Print ISSN : 2395-1990 | Online ISSN : 2394-4099
Themed Section: Engineering and Technology

A Novel Framework for Tweet segmentation and its Application to Named Entity Recognition

Anuja A. Thete [1], J. S. Karnewar[2]
Department of Computer Science and Engineering, Jagadambha College of Engineering and Tehnology,Yavatmal, Sant Gadge Baba Amravati University, Amravati, Maharashtra

ABSTRACT

Twitter has become one of the most important communication channels with its ability providing the most up-to-date and newsworthy information. Considering wide use of twitter as the source of information, reaching an interesting tweet for user among a bunch of tweets is challenging. A huge amount of tweets sent per day by hundred millions of users, information overload is inevitable. For extracting information in large volume of tweets, Named Entity Recognition (NER), methods on formal texts. However, many applications in Information Retrieval (IR) and Natural Language Processing (NLP) suffer severely from the noisy and short nature of tweets. In this paper, we propose a novel framework for tweet segmentation in a batch mode, called HybridSeg by splitting tweets into meaningful segments, the semantic or context information is well preserved and easily extracted by the downstream applications. HybridSeg finds the optimal segmentation of a tweet by maximizing the sum of the stickiness scores of its candidate segments. The stickiness score considers the probability of a segment being a phrase in English (i.e., global context) and the probability of a segment being a phrase within the batch of tweets (i.e., local context). For the latter, we propose and evaluate two models to derive local context by considering the linguistic features and term-dependency in a batch of tweets, respectively. HybridSeg is also designed to iteratively learn from confident segments as pseudo feedback. As an application, we show that high accuracy is achieved in named entity recognition by applying segment-based part-of-speech (POS) tagging.

Keywords : Twitter Stream, Tweet Segmentation, Named Entity Recognition, Linguistic Processing

I. INTRODUCTION

Twitter, as a new type of social media, has seen tremendous growth in recent years. It has attracted great interests from both industry and academia. Many private and/or public organizations have been reported to monitor Twitter stream to collect and understand user's opinions about the organizations. Nevertheless, due to the extremely large volume of tweets published every day, it is practically infeasible and unnecessary to listen and monitor the whole Twitter stream. Therefore, targeted Twitter streams are usually monitored instead; each such stream contains tweets that potentially satisfy some information needs of the monitoring organization. Targeted Twitter stream is usually constructed by filtering tweets with user-defined selection criteria depends on the information needs. Targeted Twitter stream is usually constructed by filtering tweets with predefined selection criteria (e.g., tweets published by users from a geographical region, tweets that match one or more predefined keywords). Due to its invaluable business value of timely information from these tweets, it is imperative to understand tweets' language for a large body of downstream applications, such as named entity recognition (NER) [1], [3], [4], event detection and summarization [5], [6], [7], opinion mining [8], [9], sentiment analysis and many others.

Given the limited length of a tweet (i.e., 140 characters) and no restrictions on its writing styles, tweets often

contain grammatical errors, misspellings, and informal abbreviations. The error-prone and short nature of tweets often make the word-level language models for tweets less reliable. For example, given a tweet "I call her, no answer. Her phone in the bag, she dancin.", there is no clue to guess it's true theme by disregarding word order (i.e., bag-of-word model).

The situation is further exacerbated with the limited context provided by the tweet. That is, more than one explanation for this tweet could be derived by different readers if the tweet is considered in isolation. On the other hand, despite the noisy nature of tweets, the core semantic information is well preserved in tweets in the form of named entities or semantic phrases. For example, the emerging phrase "she dancin" in the related tweets indicates that it is a key concept – it classifies this tweet into the family of tweets talking about the song "She Dancin", a trend topic in Bay Area in Jan, 2013.

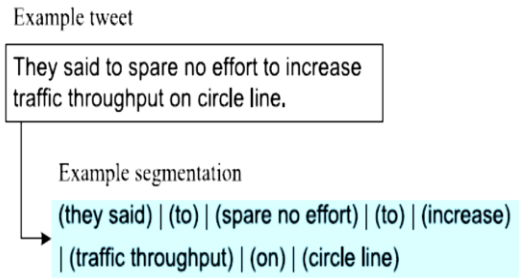

Figure 1. Example of tweet Segmentation

II. METHODS AND MATERIAL

A. Literature Review

Both tweet segmentation and named entity recognition are considered important subtasks in NLP. Many existing NLP techniques heavily rely on linguistic features, such as POS tags of the surrounding words, word capitalization, trigger words (e.g., Mr., Dr.), and gazetteers. These linguistic features, together with effective supervised learning algorithms (e.g., hidden markov model (HMM) and conditional random field (CRF)), achieve very good performance on formal text corpus [14], [15], [16]. However, these techniques experience severe performance deterioration on tweets because of the noisy and short nature of the latter. There have been a lot of attempts to incorporate tweet's unique characteristics into the conventional NLP techniques. To improve POS tagging on tweets, Ritter et al. train a POS tagger by using CRF model with conventional and tweet-specific features [3]. Brown clustering is applied in their work to deal with the ill-formed words. Gimple et al. incorporate tweet-specific features including at-mentions, hashtags, URLs, and emotions [5] with the help of a new labeling scheme. In their approach, they measure the confidence of capitalized words and apply phonetic normalization to ill-formed words to address possible peculiar writings in tweets. It was reported to outperform the state-of-the-art Stanford POS tagger on tweets. Normalization of ill-formed words in tweets has established itself as an important research problem. A supervised approach is employed in to first identify the ill-formed words. Then, the correct normalization of the ill-formed word is selected based on a number of lexical similarity measures. Both supervised and unsupervised approaches have been proposed for named entity recognition in tweets. T-NER, a part of the tweet-specific NLP framework in [3], first segments named entities using a CRF model with orthographic, contextual, dictionary and tweet-specific features. It then labels the named entities by applying Labeled-LDA with the external knowledge base Freebase.2 The NER solution proposed in [4] is also based on a CRF model. It is a two-stage prediction aggregation model. In the first stage, a KNN-based classifier is used to conduct wordlevel classification, leveraging the similar and recently labeled tweets. In the second stage, those predictions, along with other linguistic features, are fed into a CRF model for finer-grained classification. Chua et al. propose to extract noun phrases from tweets using an unsupervised approach which is mainly based on POS tagging. Each extracted noun phrase is a candidate named entity.

B. Analysis of Problem

✓ **Existing System**

Nowadays we have so many social networking sites. But we are using all sites for update states and sharing photos, videos and so on. There is no alert for bad weather situation, earthquakes and so on. That's why we will go to do develop those options in existing system.

Disadvantage

1. Time loss.
2. Quality of message reduced.

✓ **Problem Statement**

This paper presents a real-time nature of Twitter that is designed to ascertain whether we can extract valid information from it. An event notification system that monitors tweets and delivers notification promptly using knowledge from the investigation. In this, we take three steps: first, we crawl numerous tweets related to target events; second, we propose probabilistic models to extract events from those tweets and estimate locations of events; finally, we developed an alerting reporting system that extracts earthquakes from Twitter and sends a message to registered users. Here, we explain our methods using an earthquake as a target event.

✓ **Scope**

First, to obtain tweets on the target event precisely, we apply semantic analysis of a tweet. For example, users might make tweets such as "Earthquake!" or "Now it is shaking," for which earthquake or shaking could be keywords, but users might also make tweets such as "I am attending an Earthquake Conference," or "Someone is shaking hands with my boss." We prepare the training data and devise a classifier using a Support Vector Machine (SVM) based on features such as keywords in a tweet, the number of words, and the context of target-event words. After doing so, we obtain a probabilistic spatiotemporal model of an event. We then make a crucial assumption: each Twitter user is regarded as a sensor and each tweet as sensory information.

✓ **Applications of Proposed System**

[1]. Our work is also related to entity linking (EL). EL is to identify the mention of a named entity and link it to an entry in a knowledge base like Wikipedia.

[2]. Through our framework, we demonstrate that local linguistic features are more reliable than term-dependency 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.

[3]. Helps in preserving Semantic meaning of tweets.

✓ **Objective**

- HybridSeg finds the optimal segmentation of a tweet by maximizing the sum of the stickiness scores of its candidate segments.
- The stickiness score considers the probability of a segment being a phrase in English (i.e., global context) and the probability of a segment being a phrase within the batch of tweets (i.e., local context).
- Evaluate two models to derive local context by considering the linguistic features and term-dependency in a batch of tweets, respectively.
- Experiments on two tweet data sets
- Analysis and comparison of results.

✓ **Algorithm Explanation**

- As an application of tweet segmentation, we 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 of the constituent words in segments

C. System Architecture

To achieve high quality tweet segmentation, we proposed a generic tweet segmentation framework, named HybridSeg. HybridSeg learns from both global and local contexts, and has the ability of learning from pseudo feedback.

✓ **Global context**

Tweets are posted for information sharing and communication. The named entities and semantic phrases are well preserved in tweets. The global context derived from Web pages (e.g., Microsoft Web N-Gram corpus) or Wikipedia therefore helps identifying the meaningful segments in tweets. The method realizing the proposed framework that solely relies on global context is denoted by HybridSegWeb.

✓ **Local context.**

Tweets are highly time-sensitive so that many emerging phrases like “She Dancin” cannot be found in external knowledge bases. However, considering a large number of tweets published within a short time period (e.g., a day) containing the phrase, it is not difficult to recognize “She Dancin” as a valid and meaningful segment. We therefore investigate two local contexts, namely local linguistic features and local collocation. Observe that tweets from many official accounts of news agencies, organizations, and advertisers are likely well written. The well preserved linguistic features in these tweets facilitate named entity recognition with high accuracy. Each named entity is a valid segment. The method utilizing local linguistic features is denoted by HybridSegNER. It obtains confident segments based on the voting results of multiple off-the-shelf NER tools. Another method utilizing local collocation knowledge, denoted by HybridSegNGram, is proposed based on the observation that many tweets published within a short time period are about the same topic. HybridSegNGram segments tweets by estimating the term-dependency within a batch of tweets.

✓ **Pseudo feedback**

The segments recognized based on local context with high confidence serve as good feedback to extract more meaningful segments. The learning from pseudo feedback is conducted iteratively and the method implementing the iterative learning is named HybridSegIter. We conduct extensive experimental analysis on HybridSeg1 on two tweet datasets and evaluate the quality of tweet segmentation against manually annotated tweets. Our experimental results show that HybridSegNER and HybridSegNGram, the two methods incorporating local context in additional to global context, achieve significant improvement in segmentation quality over HybridSegWeb, the method use global context alone. Between the former two methods, HybridSegNER is less sensitive to parameter settings than HybridSegNGram and achieves better segmentation quality. With iterative learning from pseudo feedback, HybridSegIter further improves the segmentation quality.

D. Modules

- Admin Module
- User Module

✓ **Admin Module**

In this module admin can register and he can login. After login he will do users list and delete users. He can see user’s tweets and he can delete also. He gets messages from users about earthquakes. Admin will confirm and he will forward to the all users in this site.

✓ **User Module**

In this module user can register. And he can login. User can sent tweets to others and he will receive tweets from others. User can replay for others tweets. He can send only 140 characters tweets. He can search people in this site and he will follow other users and he will also followed by other users. He can send earthquakes reports to admin. And all users get earthquake alerts from admin. Whenever admin send earthquake alerts that time users are moving to another places. User can update his profile information and he can upload his photos.

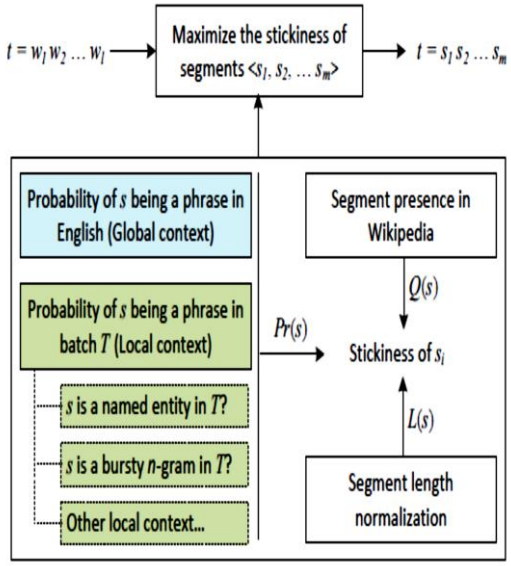

Figure 2. Hybridseg Framework without learning from pseudo feedback

III. RESULTS AND DISCUSSION

A. 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.

✓ **Tweet Segmentation**

Given a tweet t from batch T , the problem of tweet segmentation is to split the ` words in t = w1w2 : : :w` into m _ ` consecutive segments, t = s1s2:::sm, where each segment si contains one or more words. We formulate the tweet segmentation problem as an optimization problem to maximize the sum of stickiness scores of the m segments, shown in Figure 3. A high stickiness score of segment s indicates that it is a phrase which appears "more than by chance", and further splitting it could break the correct word collocation or the semantic meaning of the phrase. Formally, let C(s) denote the stickiness function of segment s.

✓ **Segment based Named Entity Recognition**

In this paper, we select named entity recognition as a downstream application to demonstrate the benefit of tweet segmentation. We investigate two segment based NER algorithms. The first one identifies named entities from a 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, we 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.

Table 1. Three POS tags as the indicator of segment being a noun phrase

Tag	Definition	Examples
N	common noun (NN, NNS)	books; someone
^	proper noun (NNP, NNPS)	lebron; usa; iPad
$	numeral (CD)	2010; four; 9:30

✓ **NER By POS Tagger**

Due to the short nature of tweets, the gregarious Property may be weak. The second algorithm then explores the part-of-speech tags in tweets for NER by considering noun phrases as named entities using segment [20] 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. We estimate the likelihood of a segment being a noun phrase (NP) by considering the POS tags of its constituent words of all appearances. Table 1 lists three POS tags that are considered as the indicators of a segment being a noun phrase.

B. Learning From Local Context

Illustrated in Figure 3, the segment phraseness Pr(s) is computed based on both global and local contexts. Based on Observation 1, Pr(s) is estimated using the n-gram probability provided by Microsoft Web NGram service, derived from English Web pages. We now detail the estimation of Pr(s) by learning from local context based[19]. Specifically, we propose learning Pr(s) from the results of using off-the-shelf Named Entity Recognizers (NERs), and learning Pr(s) from local word collocation in a batch of tweets. The two corresponding methods utilizing the local context are denoted by HybridSegNER and HybridSegNGram respectively.

✓ **Learning from Weak NERs**

To leverage the local linguistic features of well-written tweets, we apply multiple off-the-shelf NERs trained on formal texts to detect named entities in a batch of tweets T by voting. Voting by multiple NERs partially all eviates the errors due to noise in tweets. Because these NERs are not specifically trained on tweets, we also call them weak NERs. Recall that each named entity is a valid segment, the detected named entities are valid segments.

✓ **Learning from Local Collocation**

Collocation is defined as an arbitrary and recurrent word combination. Let w1w2w3 be a valid segment, it is expected that sub-n-grams fw1; w2; w3; w 1w2 ;w 2w 3g are positively correlated with one another. Thus, we need a measure that captures the extent to which the sub-n-grams of a n-gram are correlated with one

another, so as to estimate the probability of the n-gram being a valid segment.

✓ Absolute Discounting Smoothing

At first glance, it seems that applying maximum likelihood estimation is straightforward. However, because Pr(w1) is set to 1, then P^rNGram(w1 : : :wn) = fw1:::wn=fw1 . More importantly, due to the informal writing style and limited length of tweets, people often use a sub-ngram to refer to a n-gram. For example, either first name or last name is often used in tweets to refer to the same person instead of her full name. We therefore adopt absolute discounting smoothing method [15] to boost up the likelihood of a valid segment.

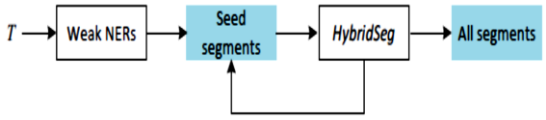

Figure 3. The Iterative process of $\text{Hybridseg}_{\text{Iter}}$

✓ Right-to-left Smoothing

Like most n-gram models, the model in Eq. 8 follows the writing order of leftto- right. However, it is reported that the latter words in a n-gram often carry more information [18]. For example, ”justin bieber” is a bursty segment in some days of tweets data in our pilot study. Since”justin” is far more prominent than word”bieber”, the ngram probability of the segment is relative small. However, we observe that”justin” almost always precedes”bieber” when the latter occurs. Given this, we introduce a right-to-left smoothing (RLS) method mainly for name detection.

IV. CONCLUSION

In this paper, we present the HybridSeg framework which segments tweets into meaningful phrases called segments using both global and local context. Through our framework, we demonstrate that local linguistic features are more reliable than term dependency 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 more noisy than formal text. Tweet segmentation helps to preserve the semantic meaning of tweets, which subsequently benefits many downstream applications, e.g. named entity recognition. We identify from this paper to improve segment quality by considering more local factors.

V. REFERENCES

[1] 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 SIGIR,2012, pp. 721–730.

[2] C. Li, A. Sun, J. Weng, and Q. He, “*Exploiting hybrid contexts for tweet segmentation*,” in SIGIR, Volume No. 3 , 2013, pp. 523–532.

[3] A. Ritter, S. Clark, Mausam, and O. Etzioni, “*Named entity recognition in tweets: An experimental study*,” in EMNLP,2011, pp. 1524–1534.

[4] X. Liu, S. Zhang, F. Wei, and M. Zhou, “*Recognizing named entities in tweets*,” in ACL, 2011, pp. 359–367.

[5] X. Liu, X. Zhou, Z. Fu, F. Wei, and M. Zhou, “*Exacting social events for tweets using a factor graph*,” in AAAI, Volume No. 2 , 2012.

[6] A. Cui, M. Zhang, Y. Liu, S. Ma, and K. Zhang, “*Discover breaking events with popular hashtags in twitter*,” in CIKM, 2012, pp. 1794–1798.

[7] A. Ritter, Mausam, O. Etzioni, and S. Clark, “*Open domain event extraction from twitter*,” in KDD, 2012, pp. 1104–1112.

[8] X. Meng, F. Wei, X. Liu, M. Zhou, S. Li, and H. Wang, “*Entitycentric topic-oriented opinion summarization in twitter*,” in KDD, 2012, pp. 379–387.

[9] Z. Luo, M. Osborne, and T. Wang, “*Opinion retrieval in twitter*,” in ICWSM , 2012, pp. 202- 215

[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 CIKM, 2011, pp. 1031–1040.

[11] K.-L. Liu, W.-J. Li, and M. Guo, “*Emoticon smoothed language models for twitter sentiment analysis*,” in AAAI, 2012.

[12] J. Weng, C. Li, A. Sun, Q. He, “*Tweet Segmentation and its Application to Named Entity Recognition*,” in IEEE Transactions, 2015, pp. 1–15.

[13] C. Li, A. Sun, J. Weng, and Q. He, “*Exploiting hybrid contexts for tweet segmentation*,” in SIGIR, 2013, pp. 523–532.

[14] L. Ratinov and D. Roth, *“Design challenges and misconceptions in named entity recognition*,” in CoNLL, 2009, pp. 147– 155.

[15] 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 ACL-HLT , 2011, pp. 42–47.

[16] B. Han and T. Baldwin, “*Lexical normalisation of short text messages: Makn sens a #twitter*,” in ACL, 2011, pp. 368–378.

[17] F. C. T. Chua, W. W. Cohen, J. Betteridge, and E.-P. Lim, “*Community-based classification of noun phrases in twitter*,” in CIKM, 2012, pp. 1702–1706

[18] G. Zhou and J. Su, “*Named entity recognition using an hmmbased chunk tagger*,” in ACL, 2002, pp. 473–480.

[19] S. Cucerzan, “*Large-scale named entity disambiguation based on wikipedia data*,” in EMNLP-CoNLL, 2007, pp. 708–716.

[20] Chenliang Li, Aixin Sun, Jianshu Weng, and Qi He,”*Exploiting hybrid contexts for tweet segmentation*” In Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval, SIGIR ’13,pages 523–532, New York, NY, USA, 2013.