Segmentation of Tweets for Multilingual Named Entity Recognition

Rashmi Rachh[1] & Vanaja H Kulkarni[2]
[1,2]Department of Computer Science and Engineering
Center of PG studies, VTU, Belagavi, Karnataka, India

Abstract: *Now-a-days Twitter is providing a way to collect and recognize user's views about many private and public organizations. However, traditional applications suffer rigorously due to short nature of tweets. A novel framework called HybridSeg is presented which easily extract and well preserves the linguistic meaning or context information by dividing tweets into significant segments and finds the optimal segmentation of tweets based on the amount of stickiness score of individual segments. The stickiness score is calculated based on the likelihood of a segment being a English phrase (i.e. global linguistic context) and likelihood of a segment being phrase within the bunch of tweets(i.e. local linguistic context).This framework will iteratively learn the candidate segments based on pseudo feedback. This model is trained on a tweet data sets, it shows that the quality of tweet segmentation has improved by learning both on global and local context. The experimental results demonstrate that tweet segmentation model significantly benefits the downstream applications***.

1. Introduction

Sites like twitter have set up novel techniques so that people can find share and extent timely information. Twitter has attracted many users to share and disseminate latest information, resulting in large scale of data produced every day. Numerous associations are interested in collecting the user opinions about their organizations, and they are in need of some named entity recognition models where they can gather and monitor the targeted twitter streams. Targeted twitter stream is constructed by straining tweets based on predefined selection criteria Targeted twitter stream is than supervised to collect and empathize user's opinion about the organization. To address the above challenges caused by tweets, this application presents a framework for targeted twitter streams called HybridSeg. HybridSeg model is the combination of both local and global linguistic context and also has the ability to learn from pseudo feedback. This application mainly focused on the task of tweet segmentation. The main goal of tweet segmentation is to split tweet into a sequence of successive n-grams (n>0), each of them are called segments. A segment is a named entity, a semantically significant information unit which appears "more than by chance" [5], [6]. The idea is to segment a tweet into a sequence of sequential phrases. Given a tweet of six words w1w2w3w4w5w6, we segment it as w1w2w3||w4w5w6 rather than w1||w2w3w4w5w6, if C (w1w2w3) +C (w4w5w6) > C (w1) + C (w2w3w4w5w6), where C (.) calculates the probability of a word or phrase being a valid segment. Below figure 1 shows the example of tweet segmentation.

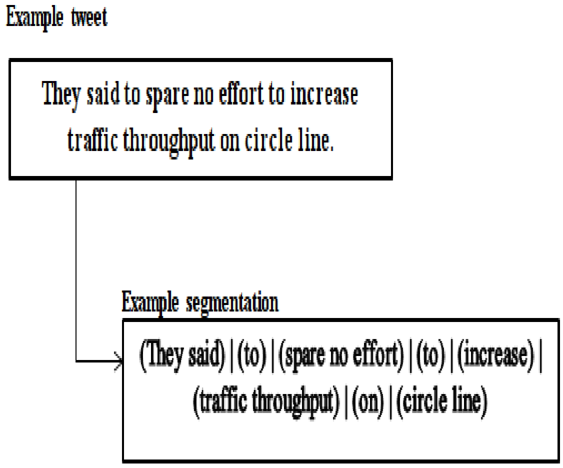
Example tweet
They said to spare no effort to increase traffic throughput on circle line.
Example segmentation
(They said) | (to) | (spare no effort) | (to) | (increase) | (traffic throughput) | (on) | (circle line)

Figure. 1. Example of tweet segmentation *[1]*

In this example, the tweet is split into eight segments. The meaningful segments such as "spare no effort", "traffic throughput", and "circle line" are preserved [1]. Because these segment uphold the semantic meaning of the tweets more accurately, than each of its consecutive words. A named entity is a valid segment, but it is not necessary that a segment should be a named entity.

In order to achieve a high quality segmentation of tweets, this project proposes a model called HybridSeg. HybridSeg model is the combination of both local and global linguistic context and also has the ability to learn from pseudo feedback.

1.1. Global linguistic context

The global linguistic context is solely derived from external knowledge bases known as GlobalSeg [2]. The GlobalSeg is the combination of Microsoft Web N-Gram corpus (i.e. web pages) or Wikipedia key phraseness, which helps to identify the meaningful segments' in tweets. Figure 2 shows the process carried out by global linguistic context.

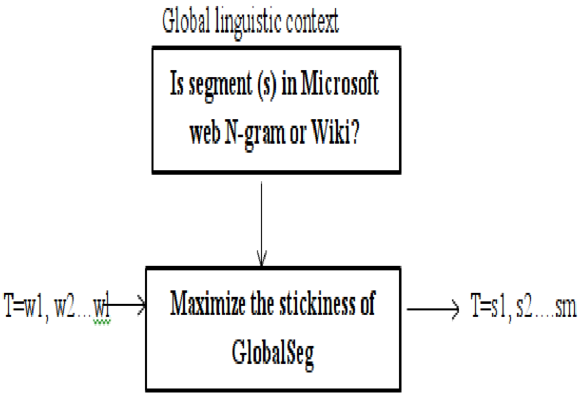
Global linguistic context
Is segment (s) in Microsoft web N-gram or Wiki?
T=w1, w2…wl
Maximize the stickiness of GlobalSeg
T=s1, s2….sm

Figure 2.The framework of global linguistic context [2]

1.2. Local linguistic context

Tweets are highly time sensitive so that many phrases like "he driven" cannot be found in external knowledge base. Purely relying on external knowledge base is not reliable. So along with existing external knowledge base, the local linguistic context is also incorporated and named it as LocalSeg. LocalSeg conducts tweet segmentation in batch mode. Tweets from targeted twitter stream are grouped into batches based on their publication time using determined time interval. Each batch of tweets is then segmented by LocalSeg collectively. Figure 3 shows the process carried out by local linguistic context."

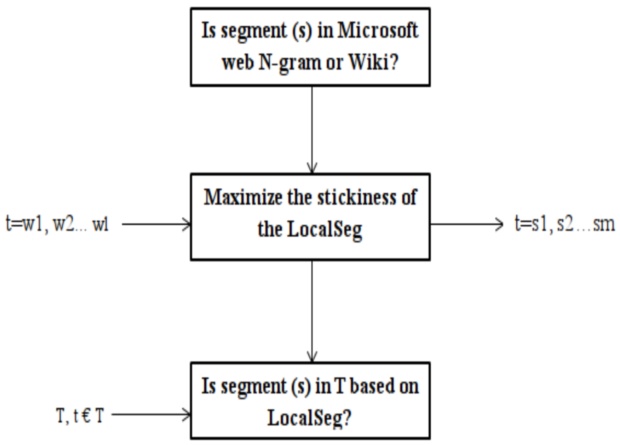
Is segment (s) in Microsoft web N-gram or Wiki?
t=w1, w2… wl
Maximize the stickiness of the LocalSeg
t=s1, s2…sm
T, t € T
Is segment (s) in T based on LocalSeg?

Figure 3. The framework of local linguistic context [2]

1.3. Pseudo feedback

The segments are recognized based on off-the-shelf NER tools can have high precision but low recall rate. In order to have better segmentation, the learning from pseudo feedback is conducted iteratively. Figure.4 shows the iterative process of LocalSeg.

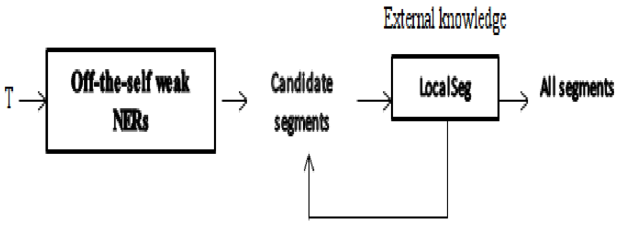

Figure 4. The iterative process of HybridSeg [2]

A brief introduction to NLP, tweet segmentation and named entity recognition has been provided. The existing approaches used for segmentation and named entity recognition are discussed in next section.

In this section, a brief introduction to NLP, tweet-segmentation and NER has been provided. The approach used for segmentation and NER are discussed in next sections.

2. Proposed framework

The below figure 5 shows the system architecture of HybridSeg framework. Initially tweets are downloaded using Twitter API from targeted twitter stream. Each tweet from the batch of tweets is segmented into n-grams. Probability of each segment is calculated based on global linguistic context as well as local linguistic context. The probability of global linguistic context is calculated using Microsoft web n-gram [3] and Wikipedia and the probability of local linguistic context is calculated using three off shelf NERs. Iterative process called pseudo feedback is carried out to improve the probability of the segments.

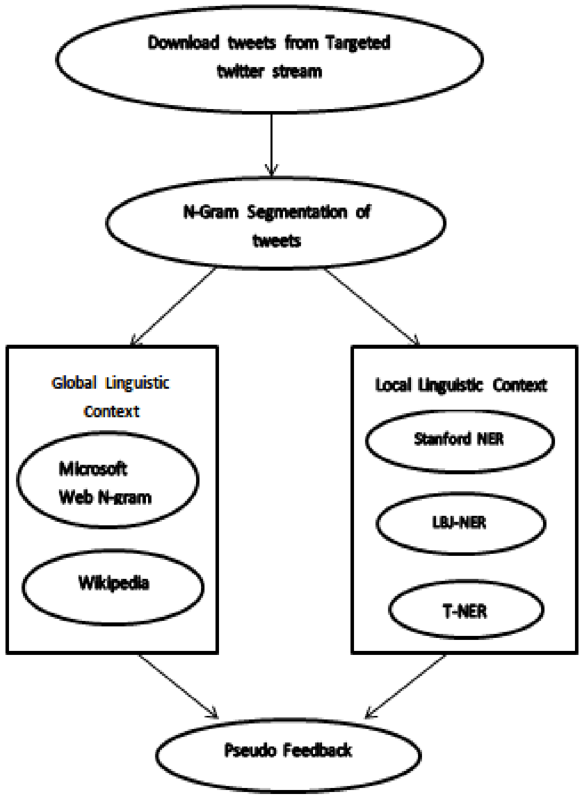
Download tweets from Targeted twitter stream
N-Gram Segmentation of tweets
Global Linguistic Context
Microsoft Web N-gram
Wikipedia
Local Linguistic Context
Stanford NER
LBJ-NER
T-NER
Pseudo Feedback

Figure 5. System architecture of HybridSeg

3. Experiment and result

We led experiment on dataset i.e., the adani group that was utilised to assess the GlobalSeg. Three weak NERs, in particular, LBJ-NER, Stanford-NER, and T-NER were utilized as contribution as a part of HybridSeg for learning local context. We compare segmentation precision of HybridSeg against GlobalSeg. Since GlobalSeg was utilized to identify named entities in [4], we likewise report the precision of named entity recognition utilizing HybridSeg and GlobalSeg individually.

Review that, the assignment of tweet segmentation is to split tweet into semantically significant portions. However, manual segmentation of a sensibly estimated information gathering is extremely costly and requires great comprehension of the tweets. Since each named entity is a substantial segment, the annotated named entity serve as halfway ground truth for the evaluation. We utilize two measures, namely Recall and Frequency-biased Recall.

Recall, denoted by Re, measures the rate of manually clarified named entities that are effectively split as segment. Since a segmentation technique yields precisely conceivable segmentation for every tweet, utilizing clarified named entities as fractional ground truth, recall is the same as precision in this setting.

Frequency-biased Recall, denoted by Re_F, gives the components in the recurrence of the named entities in a cluster of tweets. Since the continuous named entities are regularly identified with intriguing issues or developing occasions in the targeted twitter stream, accurately distinguishing these named entities as segments is basic for downstream applications, similar to opinion mining.

Letf_s^g signifies the no of occurrence of segment s in the manually annotated ground truth G. Let Re(s) is the recall of segment s Є G which is the rate of segment s being effectively extracted from all events of s in the annotated tweets and Z is normalization variable expecting all named entities are redressed segments. The frequency-biased recall is given by Eq. (1):

$$Re_F = \frac{1}{Z}\sum_{s\epsilon G} \log(f_s^g + \varepsilon).Re(s) \qquad (1)$$

3.1. Impact of λ Adaption

We exploit the local context by utilizing linear combination as a part of the estimation of SCP scores. The decision of λ to a great extent influences the execution of the tweet segmentation process. While a little λ may not adequately misuse the local context, an extensive λ could make the local context overwhelm the segmentation procedure which may unfavourably influence the segments with weak local context as a result of their predetermined number of events.

Figure 7 shows the effect of changing λ on HybridSeg regarding Re and Re_F in the 0^{th} cycle (rf Eq.1). For simple demonstration, we plot the standardized score acquired by Eq. 1 with various λ, indicated by score in the figure. Watch that score is emphatically corresponded with the execution measurements Re also, Re_F on the dataset.

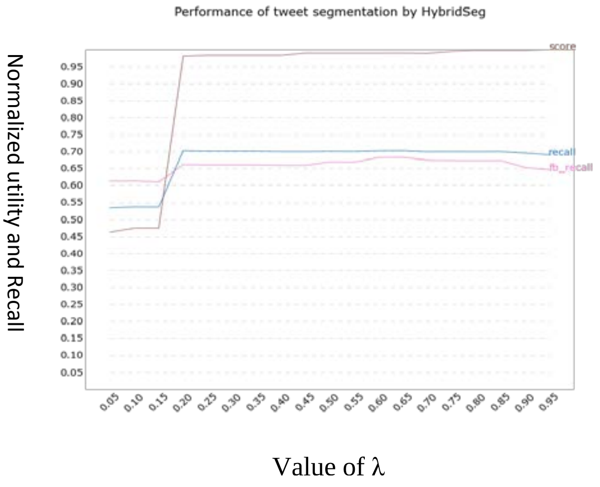

Figure 7 Impact of Re, ReF and score (λ) (%) values of HybridSeg with varying λ in the range of [0, 0.95].

4. 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 Proc. 5th International ACM SIGIR Conference Res. Development Inf. Retrieval, 2012, pp. 721-730.

[2] Kuansan Wang, Christopher Thrasher,Evelyne Viegas, xiaolong Li,Bo-jume(Paul) Hsu,June 2010, “An Overview of Microsoft Web N-gram Corpus and Applications” Proceedings of the NAACL HLT 2010:Demonstration Session, pages 45-48, Los Angeles, California.

[3] X. Liu, S. Zhang, F. Wei, and M. Zhou. Recognizing named entities in tweets. In ACL, pages 359-367, 2011.

[4] A. Ritter, S. Clak, Mausam, and o. Etzioni. Named entity recognition in tweets: An experimental study. In EMNLP, pages 1524-1534, 2011.

[5] J. Kazama and K. Torisawa. Exploiting Wikipedia as external knowledge for named entity recognition. In Proc. of EMNLP-CONLL, 2007.

[6] D. Ramage, D. Hall, R. Nallapati, and C.D. Manning, Labeled lda: a supervised topic model for credit attribution in multi-labeled corpora. In Proc. of EMNLP, 2009.

[7] L.Ratinov and D.Roth.Design challenges and misconceptions in named entity recognition. In Proc. of CONLL, 2009.