Tweet Segmentation and Its Application to Named Entity Recognition

Ramya

[1]Ramya , Address: Thanjavur
[2]Professor: Mrs. Hemalatha, M.Sc., M.Phil., M.C.A., Dept. of Computer Science, Shrimathi Indira Gandhi College, Tamil Nadu, India

---***---

Abstract - *Twitter has attracted millions of users to share and disseminate most up-to-date information, resulting in large volumes of data produced every day. However, many applications in Information Retrieval (IR) and Natural Language Processing (NLP) suffer severely from the noisy and short nature of tweets. In this project, propose a novel framework for tweet segmentation in a batch mode, called Hybrid Segmentation. 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, 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. Experiments on two tweet data sets show that tweet segmentation quality is significantly improved by learning both global and local contexts compared with using global context alone. Through analysis and comparison, show that local linguistic features are more reliable for learning local context compared with term-dependency. As an application, show that high accuracy is achieved in named entity recognition by applying segment-based part-of-speech (POS) tagging.*

Key Words**: Twitter stream, tweet segmentation, named entity recognition, linguistic processing, Wikipedia**

1. INTRODUCTION

Tweets are posted for information sharing and communication. The named entities and semantic phrases are well preserved in tweets. The method realizing the proposed framework that solely relies on global context is denoted by HybridSeg. Tweets are highly time-sensitive. 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 HybridSeg NER. It obtains confident segments based on the voting results of multiple off-the-shelf NER tools. Another method utilizing local collocation knowledge, denoted by HybridSeg N-Gram, is proposed based on the observation that many tweets published within a short time period are about the same topic. HybridSeg N-Gram segments tweets by estimating the term-dependency within a batch of tweets. To improve POS tagging on tweets, Ritter et al. train a POS tagger by using CRF model with conventional and tweet-specific features. Even though short text strings might be a problem, sentiment analysis within micro blogging has shown that Twitter can be seen as a valid online indicator of political sentiment. A lot of advertising and branding is now tied to Twitter. Furthermore, and most importantly, the first place one goes to find any breaking news is to search it on Twitter. The focus of this project is clustering with the TF-IDF weighted mechanism of daily technology news tweets of prominent bloggers and news sites using Apache Mahout and to evaluate the effects of introducing and removing anonymous words on the quality of clustering. This project restricts itself to only tweets in the English language.

2. Existing System

Nowadays people are using so many social networking sites are. But the people 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. The problem is that lot of time consumed to the people for checking the message submitted in the social networking and file a case in cyber crime. After verification cyber crime defense commission will remove the post according to the cyber crime rules in a span of time. This will not solve the problem only to remove the things after the issue arises. Due to ignorance and lack of understanding of Social Media privacy feature, people make many mistakes. Another situation to consider has to do with the availability of information too personal, whether in pictures or text. You cannot give out too much personal information. As throughout the Internet. The User can send their post with some illegal words and letters.

2.1 Survey

A. User tracking using tweet segmentation and word

Many organizations have been reported to create and monitoring targeted Twitter streams to collect a bunch of information and understand according to user's view. Targeted Twitter stream is main usually constructed by filtering tweets and that abused words with predefined selection criteria. Due to its invaluable business value of timely information from these tweets, it's a necessary to understand that abused word's language for a large body of downstream applications, such as named entity recognition (NER), event detecting and summarizing that particular word, opinion mining, sentiment analysis, and etc. In these proposed system application is developed which take tweet is input and search semantic negative or illegal words from database. Generate report of that abusing word and send to cyber crime's site. Then depending on the tweet, track the related information of the person. Track all the tweets of that particular person and track that person through identification databases. Then take a action by cyber crime. And after that prevent the tweet.

B. Measuring the controversy level of Arabic trending topics on Twitter

Social micro-blogging systems like Twitter are used today as a platform that enables its users to write down about different topics. One important aspect of such human interactions is the existence of debate and disagreement. The most heated debates are found on controversial topics. Detecting such topics can be very beneficial in understanding the behavior of online social networks users and the dynamics of their interactions. Such an understanding leads to better ways of handling and predicting how the "online crowds" will act. Several approaches have been proposed for detecting controversy in online communication. Some of them represent the interactions in the form of graphs and study their properties in order to determine whether the topic of interaction is controversial or not. Other approaches rely on the content of the exchanged messages. In this study, we focus on the former approach in identifying the controversy level of the trending topics on Twitter. Unlike many previous works, then do not limit ourselves to a certain domain. Moreover, the main focus on social content written in Arabic about hot events occurring in the Middle East. To the best of our knowledge, ours is the first work to undertake this approach in studying controversy in general topics written in Arabic. Collect a large dataset of tweets on different trending topics from different domains. First apply several approaches for controversy detection and compare their outcomes to determine which one is the most consistent measure.

C. Event Detection and Key Posts Discovering in Social Media Data Streams

Microblogging, as a popular social media service, has become a new information channel for users to receive and exchange the most up-to-date information on current events. Consequently, it is crucial to detect new emerging events and to discover the key posts which have the potential to actively disseminate the events through microblogging. However, traditional event detection models require human intervention for detecting the number of topics to be explored, which significantly reduces the efficiency and accuracy of event detection. Most of the existing methods focus only on event detection and are unable to discover the key posts related to the events, making it challenging to track momentous events in timely manner. To tackle these problems, a HITS (Hypertext Induced Topic Search) based topic decision method, named TD-HITS is proposed. TD-HITS can automatically detect the number of topics as well as discovering associated key posts from a large number of posts. The experimental results are based on a Twitter dataset to demonstrate the effectiveness of our proposed methods for both detecting events and identifying key posts.

D. Tweet modeling with LSTM recurrent neural networks for hash tag recommendation

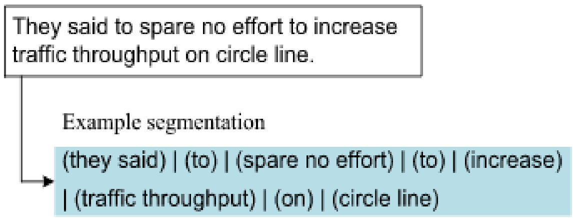
The hash symbol, called a hash tag, is used to mark the keyword or topic in a tweet. It was created organically by users as a way to categorize messages. Hash tags also provide valuable information for many research applications such as sentiment classification and topic analysis. However, only a small number of tweets are manually annotated. Therefore, an automatic hash tag recommendation method is needed to help users tag their new tweets. Previous methods mostly use conventional machine learning classifiers such as SVM or utilize collaborative filtering technique. A bottleneck of these approaches is that they all use the TF-IDF scheme to represent tweets and ignore the semantic information in tweets. In this paper, we also regard hash tag recommendation as a classification task but propose a novel recurrent neural network model to learn vector-based tweet representations to recommend hashtags. More precisely, we use a skip-gram model to generate distributed word representations and then apply a convolution neural network to learn semantic sentence vectors. Afterwards, we make use of the sentence vectors to train a long short-term memory recurrent neural network (LSTM-RNN). We directly use the produced tweet vectors as features to classify hashtags without any feature engineering. Experiments on real world data from Twitter to recommend hashtags show that our proposed LSTM-RNN model outperforms state-of-the-art methods and LSTM unit also obtains the best performance compared to standard RNN and gated recurrent unit (GRU).

E. Exploring human emotion via Twitter

Sentiment analysis or opinion mining on twitter data is an emerging topic in research. In this paper, we have described a system for emotion analysis of tweets using only the core text. Tweets are usually short, more ambiguous and contains a huge amount of noisy data, sometimes it is difficult to understand the user's opinion. The main challenge is to feature extraction for the purpose of classification and feature extraction depends on the perfection of preprocessing of a tweet. The preprocessing is the most difficult task, since it can be done in various ways and the methods or steps applied in preprocessing are not distinct. Most of the researches in this topic, have been focused on binary (positive and negative) and 3-way (positive, negative and neutral) classifications. In this paper, we have focused on emotion classification of tweets as multi-class classification. We have chosen basic human emotions (happiness, sadness, surprise, disgust) and neutral as our emotion classes. According to the experimental results, our approach improved the performance of multi-class classification of twitter data.

3. Proposed System

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 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. Tweets are highly time-sensitive so that many emerging phrases like "She Dancing" cannot be found in external knowledge bases. However, considering a large number of tweets published within a short time period containing the phrase, it is not difficult to recognize "She Dancing" as a valid and meaningful segment. 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.


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)

Fig1- Sample Twitter Segmentation

Text Mining Algorithm

Text mining, also referred to as text data mining, roughly equivalent to text analytics, is the process of deriving high-quality information from text. High-quality information is typically derived through the devising of patterns and trends through means such as statistical pattern learning. Text mining usually involves the process of structuring the input text (usually parsing, along with the addition of some derived linguistic features and the removal of others, and subsequent insertion into a database), deriving patterns within the structured data, and finally evaluation and interpretation of the output. 'High quality' in text mining usually refers to some combination of relevance, novelty, and interestingness. Typical text mining tasks include text categorization, text clustering, concept/entity extraction, production of granular taxonomies, sentiment analysis, document summarization, and entity relation modeling. Text analysis involves information retrieval, lexical analysis to study word frequency distributions, pattern recognition, tagging/annotation, information extraction, data mining techniques including link and association analysis, visualization, and predictive analytics. The overarching goal is, essentially, to turn text into data for analysis, via application of natural language processing and analytical methods.

4. Conclusion

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 noise than formal text. Tweet segmentation helps to preserve the semantic meaning of tweets, which subsequently benefits many downstream applications, e.g. named entity recognition. Through experiments, we show that segment based named entity recognition methods achieves much better accuracy than the word-based alternative.

REFERENCES

1. Twiner: Named entity recognition in targeted twitter stream- C. Li, J. Weng, Q. He, Y. Yao, A. Datta, A. Sun, and B.-S. Lee.\

2. Augmenting naive bayes classifiers with statistical language models- F. Peng, D. Schuurmans, and S. Wang.

3. Topic sentiment analysis in twitter: a graph-based hash tag sentiment classification approach- X. Wang, F. Wei, X. Liu, M. Zhou, and M. Zhang.

4. Twevent: segment-based event detection from tweets- C. Li, A. Sun, and A. Datta-

5. Incorporating nonlocal information into information extraction systems by gibbs sampling- J. R. Finkel, T. Grenager, and C. Manning.