Design and Implementation of Named Entity Recognition Technique

R. Subba Rao, Rajanikanta Mohanty, Mohini Prasad Mishra
Dept. of Computer Science and Engineering
Gandhi Engineering College, Bhubaneswar, India

ABSTRACT:

Twitter has attracted millions of users to share and disseminate most up-to-date information, resulting in large volumes of data produced everyday. 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. 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, we show that local linguistic features are more reliable for learning local context compared with term-dependency. As an application, we show that high accuracy is achieved in named entity recognition by applying segment-based part-of-speech (POS) tagging.

1.1 Introduction to project

MICROBLOGGING sites such as Twitter have reshaped the way people find, share,
and disseminate timely information. Many organizations have been reported to create and monitor targeted Twitter streams to collect and understand users’ opinions. 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) event detection and summarization opinion mining 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 its 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 January 2013. In this paper, we focus on the task of tweet segmentation. The goal of this task is to split a tweet into a sequence of consecutive n-grams (n ≥ 1), each of which is called a segment. A segment can be a named entity (e.g., a movie title “finding nemo”), a semantically meaningful information unit (e.g., “officially released”), or any other types of phrases which appear “more than by chance” Fig. 1 gives an example. In this example, a tweet “They said to spare no effort to increase traffic throughput on circle line.” is split into eight segments. Semantically meaningful segments “spare no effort”, “traffic throughput” and “circle line” are preserved. Because these segments preserve semantic meaning of the tweet more precisely than each of its constituent words does, the topic of this tweet can be better captured in the
subsequent processing of this tweet. For instance, this segment-based representation could be used to enhance the extraction of geographical location from tweets because of the segment “circle line”. In fact, segment-based representation has shown its effectiveness over word-based representation in the tasks of named entity recognition and event detection.

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 Hybrid-Seg1 on two tweet data sets 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 addition 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. 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. 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.

1.2. Objective of the Project

Twitter has attracted millions of users to share and disseminate most up-to-date information, resulting in large volumes of data produced everyday. 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. 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, we show that local linguistic features are more reliable for learning local context compared with term-dependency. As an application, we show that high accuracy is achieved in named entity recognition by applying segment-based part-of-speech (POS) tagging.

1.3. Existing System

- 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. However, these techniques experience severe performance deterioration on tweets because of the noisy and short nature of the latter.

- In Existing System, to improve POS tagging on tweets, Ritter et al. train a POS tagger by using CRF model with conventional and tweet-specific features. Brown clustering is applied in their work to deal with the ill-formed words.
DISADVANTAGES OF EXISTING SYSTEM:

- 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.

1.4. Proposed System

- In this paper, we focus on the task of tweet segmentation. The goal of this task is to split a tweet into a sequence of consecutive n-grams, each of which is called a segment. A segment can be a named entity (e.g., a movie title “finding nemo”), a semantically meaningful information unit (e.g., “officially released”), or any other types of phrases which appear “more than by chance”
- To achieve high quality tweet segmentation, we propose 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.
- 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.

ADVANTAGES OF PROPOSED SYSTEM:

- 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.
- 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.

☑ Helps in preserving Semantic meaning of tweets.

System Architecture

\[ I = W_1 W_2 \ldots W_n \]

Maximize the stickiness of segments \( <s_1, s_2, \ldots, s_m> \)

\[ I = S_1 S_2 \ldots S_m \]

OUTPUT SCREENS
CONCLUSIONS

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. Through experiments, we show that segment-based named entity recognition methods achieve much better accuracy than the word-based alternative. We identify two
directions for our future research. One is to further improve the segmentation quality by considering more local factors. The other is to explore the effectiveness of the segmentation-based representation for tasks like tweets summarization, search, hashtag recommendation, etc.

**BIBLIOGRAPHY**


