**Tweet Segmentation and Its Application to Named Entity Recognition**

**OVERVIEW**

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.

**DATA**

* 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.

**MODELS TO SOLVE**

* 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.

**Project Team Members and Roles:**

We are three members in the group shared the work equally akhila katkam is doing the testing work and sai karthik napa is doing the back end data collection, Avinash is contributing the material required and Mallikarjuna provides Data Analysis for the projects.

|  |  |
| --- | --- |
| Sai karthik Napa (0270) | Data Collection |
| Akhila Katkam (0736) | Testing |
| Manepalli Mallikarjuna (1620) | Data analysis,coding |
| Avinash Chowdary Koganti (1474) | API development, documentation |