International Journal of Engineering Research-Online
A Peer Reviewed International Journal
Articles available online http://www.ijoer.in; editorijoer@gmail.com

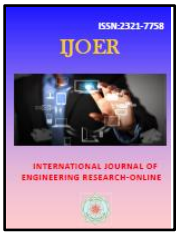
Vol.6., Issue.5, 2018
Sept-Oct.,

RESEARCH ARTICLE

ISSN: 2321-7758

NAMED ENTITY RECOGNITION (NER): A STUDY IN TWEETS

SWATHY.K.SHAJI
M G University



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 paper, we study the basic techniques that is used for the extraction of named entities from a group of tweets. By splitting tweets into meaningful segments, the semantic or context information is well preserved and easily extracted by the downstream applications. Here we analyse the two basic methods used for the identification of named entities: Shallow parsing and POS Tagging. Also the paper describes an orthographic feature for recognizing the named entities. In tweets capitalization is much reliable than in text. The linguistic features in these tweets enable named entity recognition with relatively high accuracy.

1. INTRODUCTION

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) [1], [2], [3], [4], event detection and summarization [5], [6], [7], opinion mining [8], [9],sentiment analysis [10], [11], 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. 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" [12]. Figure 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 on effort", "traffic throughput" and "circleline" 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. Note that, a named entity is valid segment; but a segment may not necessarily be a named entity. In the segment "korea vs greece" is detected for the event related to the world cup match between Korea and Greece. 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. Status Messages posted on Social Media websites such as Facebook and Twitter present a new and challenging style of text for language technology due to their noisy and informal nature. Like SMS (Kobus et al., 2008), tweets are particularly terse and difficult. Yet tweets provide a unique compilation of information that is more up to-date and inclusive than news articles, due to the low-barrier to tweeting, and the proliferation of mobile devices.1 The corpus of tweets already exceeds the size of the Library of Congress (Hachman, 2011)and is growing far more rapidly. Due to the volume of tweets, it is natural to consider named-entity recognition, information extraction, and text mining over tweets. We find that classifying named entities in tweets is a difficult task for two reasons. First, tweets contain a plethora of distinctive named entity types (Companies, Products, Bands, Movies, and more). Almost all these types (except for People and Locations) are relatively infrequent, so even a large sample of manually annotated tweets will contain few training examples. Secondly, due to Twitter's 140 character limit, tweets often lack sufficient context to determine an entity's type without the aid of background knowledge. To address these issues we propose a distantly supervised approach which applies Labeled LDA (Ramage et al., 2009) to leverage large amounts of unlabeled data in addition to large dictionaries of entities gathered from Freebase, and combines information about an entity's context across its mentions.

2. RELATED WORK

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. 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. 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 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, 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. The NER solution proposed in 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 word level 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. 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. Conventionally, EL involves a NER system followed by a linking system. Recently, Sil and Yates propose to combine named entity recognition and linking into a joint model. Similarly, Guo et al. propose a structural SVM solution to simultaneously recognize mention and resolve the linking. While entity linking aims to identify the boundary of a named entity and resolve its meaning based on an external knowledge base, a typical NER. The system identifies entity mentions only, like the work presented here. It is difficult to make a fair comparison between these two techniques. Tweet segmentation is conceptually similar to Chinese word segmentation (CSW). Text in Chinese is a continuous sequence of characters. Segmenting the sequence into meaningful words is the first step in most applications. State-of-the-art CSW methods are mostly developed using supervised learning techniques like perceptron learning and CRF model. Both linguistic and lexicon features are used in the supervised learning in CSW. Tweets are extremely noisy with misspellings, informal abbreviations, and grammatical errors. These adverse properties lead to a huge number of training samples for applying a supervised learning technique. Here, we exploit the semantic information of external knowledge bases and local contexts to recognize meaningful segments like named entities and semantic phrases in Tweets. Very recently, similar idea has also been explored for CSW by Jiang et al. They propose to prune the search space in CSW by exploiting the natural annotations in the Web. Their experimental results show significant improvement by using simple local features.

3. SHALLOW SYNTAX IN TWEETS

We first study two fundamental NLP tasks – POS tagging and noun-phrase chunking.

A. Part of Speech Tagging: Part of speech tagging is applicable to a wide range of NLP tasks including named entity segmentation and information extraction. Prior experiments have suggested that POS tagging has a very strong baseline: assign each word to its most frequent tag and assign each Out of Vocabulary (OOV) word the most common POS tag. This baseline obtained a 0.9 accuracy on the Brown corpus (Charniak et al., 1993). However, the application of a similar baseline on tweets obtains a much weaker 0.76, exposing the challenging nature of Twitter data. A key reason for this drop in accuracy is that Twitter contains far more OOV words than grammatical text. Although NNP is the most frequent tag for OOV words, only about 1/3 are NNPs. The performance of off-the-shelf news-trained POS taggers also suffers on Twitter data. The state of-the-art Stanford POS tagger (Toutanova et al., 2003) improves on the baseline, obtaining an accuracy of 0.8. This performance is impressive given that its training data, the Penn Treebank WSJ (PTB), is so different in style from Twitter, however it is a huge drop from the 97% accuracy reported on the PTB. There are several reasons for this drop in performance.

B. Shallow Parsing: Shallow parsing, or chunking is the task of identifying non-recursive phrases, such as noun phrases, verb phrases, and prepositional phrases in text. Accurate shallow parsing of tweets could benefit several applications such as Information Extraction and Named Entity Recognition.

C. Capitalization: A key orthographic feature for recognizing named entities is capitalization (Florian, 2002; Downey et al., 2007). Unfortunately in tweets, capitalization is much less reliable than in edited texts. In addition, there is a wide variety in the styles of capitalization. In some tweets capitalization is informative, whereas in other cases, non-entity words are capitalized simply for emphasis. Some tweets contain all lowercase words (8%), whereas others are in ALL CAPS (0.6%). To address this issue, it is helpful to incorporate information based on the entire content of the message to determine whether or not its capitalization is informative. To this end, we build a capitalization classifier, T-CAP, which predicts whether or not a tweet is informatively capitalized. Its output is used as a feature for Named Entity Recognition. We manually labeled our 800 tweet corpus as having either “informative” or “uninformative” capitalization. The criteria we use for labeling is as follows: if a tweet contains any non-entity words which are capitalized, but do not begin

a sentence, or it contains any entities which are not capitalized, then its capitalization is "uninformative", otherwise it is "informative".

4. NAMED ENTITY RECOGNITION

We now discuss our approach to named entity recognition on Twitter data. As with POS tagging and shallow parsing, off the shelf named-entity recognizers perform poorly on tweets. For example, applying the Stanford Named Entity Recognizer results in the following output: [Yess]ORG! [Yess]ORG! Its official [Nintendo] LOC announced today that they Will release the [Nintendo]ORG 3DS in north [America]LOC march 27 for $250.The OOV word 'Yess' is mistaken as a named entity. In addition, although the first occurrence of 'Nintendo' is correctly segmented, it is misclassified, whereas the second occurrence is improperly segmented – it should be the product "Nintendo 3DS". Finally "north America" should be segmented as a LOCATION, rather than just 'America'. In general, news-trained Named Entity Recognizers seem to rely heavily on capitalization, which we know to be unreliable in tweets. Following Collins and Singer (1999) Downey et al. (2007) and Elsner et al. (2009), we treat classification and segmentation of named entities as separate tasks. This allows us to more easily apply techniques better suited towards each task. For example, we are able to use discriminative methods for named entity segmentation and distantly supervised approaches for classification. While it might be beneficial to jointly model segmentation and (distantly supervised) classification using a joint sequence labeling and topic model similar to that proposed by Sauper et al. (2010), we leave this for potential future work. Because most words found in tweets are not part of an entity, we need a larger annotated dataset to effectively learn a model of named entities. We therefore use a randomly sampled set of 2,400 tweets for NER. All experiments (Tables 6, 8-10) report results using 4-fold cross validation.

A. Segmenting Named Entities: Because capitalization in Twitter is less informative than news, in-domain data is needed to train models which rely less heavily on capitalization, and also are able to utilize features provided by T-CAP. We exhaustively annotated our set of 2,400 tweets (34K tokens) with named entities.8 A convention on Twitter is to refer to other users using the @ symbol followed by their unique username. We deliberately choose not to annotate @usernames as entities in our data set because they are both unambiguous, and trivial to identify with 100% accuracy using a simple regular expression, and would only serve to inflate our performance statistics. While there is ambiguity as to the type of @usernames (for example, they can refer to people or companies), we believe they could be more easily classified using features of their associated user's profile than contextual features of the text. T-SEG models Named Entity Segmentation as a sequence-labeling task using IOB encoding for representing segmentations (each word either begins, is inside, or is outside of a named entity), and uses Conditional Random Fields for learning and inference. Again we include orthographic, contextual and dictionary features; our dictionaries included a set of type lists gathered from Freebase. In addition, we use the Brown clusters and outputs of T-POS, T-CHUNK and T-CAP in generating features.

5. OBSERVATIONS FOR TWEET SEGMENTATION

Tweets are considered noisy with lots of informal abbreviations and grammatical errors. However, tweets are posted mainly for information sharing and communication among many purposes. Observation 1: Word collocations of named entities and common phrases in English are well preserved in Tweets. Many named entities and common phrases are preserved in tweets for information sharing and dissemination. In this sense, Pr(s) can be estimated by counting a segment's appearances in a very large English corpus (i.e., global context). In our implementation, we turn to Microsoft Web N-Gram corpus. This N-Gram corpus is derived from all Web documents indexed by Microsoft Bing in the EN-US market. It provides a good estimate of the statistics of commonly used phrases in English. Observation 2: Many tweets contain useful linguistic features. Although many tweets contain unreliable linguistic features like misspellings and unreliable capitalizations, there exist tweets composed in proper English. For example, tweets published by official accounts of news agencies, organizations,

and advertisers are often well written. The linguistic features in these tweets enable named entity recognition with relatively high accuracy.

Observation 3: Tweets in a targeted stream are not topically independent to each other within a time window. Many tweets published within a short time period talk about the same theme. These similar tweets largely share the same segments. For example, similar tweets have been grouped together to collectively detect events, and an event can be represented by the common discriminative segments across tweets. The latter two observations essentially reveal the same phenomenon: local context in a batch of tweets complements global context in segmenting tweets. For example, person names emerging from bursty events may not be recorded in Wikipedia. However, if the names are reported in tweets by news agencies or mentioned in many tweets, there is a good chance to segment these names correctly based on local linguistic features or local word collocation from the batch of tweets.

6. CONCLUSIONS

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. 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, hash tag recommendation, etc. We have demonstrated that existing tools for POS tagging, Chunking and Named Entity Recognition perform quite poorly when applied to Tweets. Additionally we have shown the benefits of features generated from T-POS and T-CHUNK in segmenting Named Entities.

3. 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, 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, 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.

[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]. Edward Benson, Aria Haghighi, and Regina Barzilay. 2011. Event discovery in social media feeds. In The 49th Annual Meeting of the Association for Computational Linguistics, Portland, Oregon, USA. To appear.