International Journal of Engineering Research and General Science Volume 3, Issue 6, November-December, 2015
ISSN 2091-2730

Survey on Named Entity Recognition System over Twitter Data

Ms. Minal S.Sonmale[1], Prof. Rajaram H.Ambole[2]

[1]Student, M.E., Department of Computer Engineering, VPCOE, Baramati, Pune, Maharashtra, India.
sonmaleminal@gmail.com

[2]Assistant Professor, Department of Computer Engineering, VPCOE, Baramati, Pune, Maharashtra, India.
rajaram.ambole@gmail.com

Abstract— Twitter has allowed millions of users to share and spread most up-to-date information which results into large volume of data generated every day. Due to extremely useful business information obtained from these tweets, it is necessary to understand tweets language for downstream applications, such as Named Entity Recognition (NER). Real time applications like Traffic detection system, Early crisis detection and response with target twitter stream required good NER system, which automatically find emerging named entities that are potentially linked to the crisis and traffic, but tweets are infamous for their error-prone and short nature. This leads to failure of much conventional NER techniques, which heavily depend on local linguistic features, such as capitalization, POS tags of previous words etc. Recently segment-based tweet representation has showed effectiveness in NER.The goal of this survey is to provide a comprehensive review of NER system over twitter data and different NER approaches to improve their effectiveness in named entity recognition applications.

Keywords— Named Entity Recognition (NER), target twitter stream, POS tags, local linguistic features, tweet segmentation.

I. INTRODUCTION

Twitter, as a new type of social media, has seen tremendous growth in recent years. It has attracted great interests from both industry and academia. Millions of users share and spread most up-to-date information on twitter which results into large volume of data generated every day. Many private and/or public organizations have been reported to monitor Twitter stream to collect and understand user's opinions about the organizations. We can get extremely useful business value from these tweets, so it is necessary to understand tweets language for a large body of downstream applications such as NER.

A. Named Entity Recognition Concept

NER is a subtask of information extraction that seeks to locate and classify named entities, named entity is a text element indicating the name of a person, organization and location. As shown in Fig.1, task of NER also related to Entity Linking (EL). EL is used to identify the mention of a named entity and link it to an entry in a knowledge base like Wikipedia.

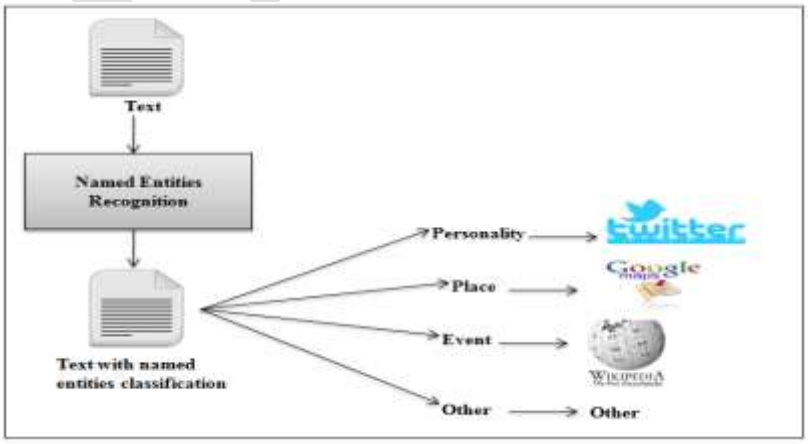

Fig.1 NER System with entity classification & linking

The first paper on NER was presented at the Seventh IEEE Conference on Artificial Intelligence Applications which is presented by Lisa F. Rau (1991). In Rau's paper she describe a NER system that extract and recognize company names, system developed by Lisa based on heuristics and handcrafted rules. At first, English is the most popular langrage factor to research NER, but along with the development of research in these areas, more and more kinds of language have been researched. In this survey, we focus on NER over twitter data. Traffic detection system, early crisis detection and response with target twitter stream are real time applications which required good NER system, which is able to automatically discover emerging named entities that are potentially linked to the crisis and traffic. Many existing NER 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 tweets. Tweets often contain grammatical errors, misspellings, and informal abbreviations because of two reasons –

1. The short length of a tweet (i.e., 140 characters)
2. No restriction on its writing styles.

This leads to failure of much conventional NER techniques, but recently segment-based tweet representation (tweet segmentation) has demonstrated effectiveness in NER. Segment-based tweet representation split tweets into meaningful phrases or segments and checks its validity by using external knowledge bases (e.g., Microsoft Web N-Gram corpus, Wikipedia Dumps) and local context information embedded in the tweets.

1) Named Entity

Named entity is a text element indicating the name of a person, organization and location. For example:

[Shubham]$_{\text{Person}}$ bought 300 shares of Aceme Corp.]$_{\text{Organization}}$ in [2015] $_{\text{Time.}}$

Here, Shubham, Aceme Corp. and 2015 are named entities which is classify under the person, organization and time class respectively. Fig.2 shows named entities based on their pre-defined class.

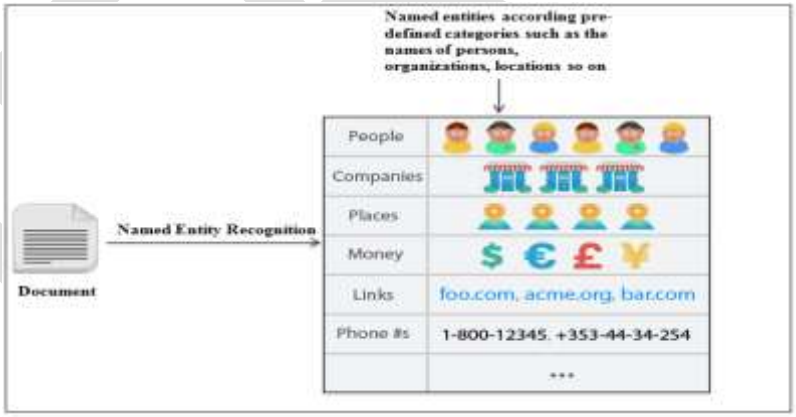

Fig.2 Named entities based on their class.

2) NER Applications

NER is useful in many Natural Language Processing applications such as question answering, information extraction, machine translation, parsing. It also provides person or organization names with their information. Usually, NER systems are used in the areas of entity identification in the bioinformatics, molecular biology and medical natural language processing communities.NER also used in real time applications.

II. APPROCHES TO NER

In this section, some NER approaches are reviewed.

B. Supervised methods

Supervised methods are class of algorithm that learns a model by looking at annotated training examples. Supervised learning algorithms for NER are Hidden Markov Model (HMM), Maximum Entropy Models (ME), Decision Trees, Support Vector Machines (SVM) and Conditional Random Fields (CRF). These all are forms of the supervised learning approach that typically consist of a system that reads a large corpus, memorizes lists of entities, and creates disambiguation rules based on discriminative features.

1) Hidden Markov Model

HMM is the earliest model applied for solving NER problem by Bikel et al. (1999). Bikel proposed a system *IdentiFinder* to identify named entities. In *IdentiFinder* system only single label can be assigned to a word in context. Therefore the model assigns to every word either one of the desired classes or the label NOT-A-NAME which means “none of the desired classes".

2) Maximum Entropy based Model

Maximum entropy model is discriminative model like HMM. In Maximum entropy based Model given a set of features and training data the model directly learns the weight for discriminative features for entity classification. Objective of the model is to maximize the entropy of the data, so as to generalize as much as possible for the training data.

3) Decision Trees

Decision Tree is a tree structure used to make decisions at the nodes and obtain some result at the leaf nodes. A path in the tree represents a sequence of decisions leading to the classification at the leaf node. Decision trees are attractive because the rules can be easily grasps from the tree. It is a well liked tool for prediction and classification.

4) CRF Based Model

Lafferty et al. (2001) proposed Conditional random field model as a statistical modeling tool for pattern recognition and machine learning using structured prediction. McCallum and Li (2003) developed feature induction method for CRF in NE.

5) SVM Based Model

Support Vector Machine was first introduced by Cortes and Vapnik in 1995 which is based on the idea of learning a linear hyperplane that separate the positive examples from negative example by large margin. Large margin suggests that the distance between the hyperplane and the point from either instance is maximum. Support vectors are points closest to hyperplane on either side.

C. Unsupervised methods

Problem with supervised algorithms is it required large number of features. For learning a good model, a robust set of features and large annotated corpus is needed. Many languages don't have large annotated corpus available at their disposal. To deal with lack of annotated text across domains and languages, unsupervised techniques for NER have been proposed.

D. Semi-supervised methods

Semi supervised learning algorithms use both labeled and unlabeled corpus to create their own hypothesis. Algorithms typically start with small amount of seed data set and create more hypotheses using large amount of unlabeled corpus.

III. RELATED WORK

There have been many NER systems developed not only in academia but also in industry some of them are reviewed in this section. GuoDong Zhou and Jian Su[1] proposed a Hidden Markov Model (HMM) and an HMM-based chunk tagger, by using which NER system is built to recognize and classify names, times and numerical quantities. This NER system achieves very good performance on formal text corpus but experience severe performance deterioration on tweets because of the noisy and short nature of the tweets.

The NER solution proposed by L. Ratinov and D. Roth [2] presented a simple model for NER that uses expressive features to achieve new state of the art performance on the NER task. System explored four fundamental design decisions: text chunks representation, inference algorithm, using non-local features and external knowledge entity types. This system can gain consistent performance across several domains, most interestingly in WebPages, where the named entities had fewer contexts and were different in nature from the named entities in the training set but when evaluating the system, it matched against the gold tokenization ignoring punctuation marks.

Chenliang Liy and Aixin Suny [3] presented a novel 2-step unsupervised NER system for targeted Twitter stream, called *TwiNER*. In the first step, dynamic programming algorithm with global context obtained from Wikipedia and Web N-Gram corpus is used for tweet segmentation. Tweet segmentation partition tweets into valid segments (phrases).Each such tweet segment is a candidate named entity. In the second step, *TwiNER* constructs a random walk model to utilize the gregarious property in the local context derived from the Twitter stream. The highly-ranked segments have a higher chance of being true named entities. By aggregating local context and global context *TwiNER* is able to recognize new named entities which may not appear in Wikipedia yet but this system is not able to address the problem of entity type classification.

Chaua et al [5] proposed a noun phrase (NP) classification method that automatically finding noun phrases (NPs) as keywords for event monitoring in Twitter. This system proposed 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. Here first model NP+LDA could classify NPs into political categories more accurately than state of the art tagger (SentReg) based on a large knowledge base, by incorporating community information in NP+RART system can further improve the accuracy of classifier. Finally, system could classify sports NPs in a political dominated Twitter data set.

Silviu Cucerzan[6] proposed large-scale system for the named entity recognition and semantic disambiguation based on information extracted from a large encyclopedic collection and Web search results. Through a process of maximizing the agreement between the contextual information extracted from Wikipedia and the context of a document, as well as the agreement among the category tags associated with the candidate entities, the implemented system shows high disambiguation accuracy on both news stories and Wikipedia articles. This system treat mention detection and entity disambiguation as two different problems. Milne and Witten[7] proposed system that describes how to automatically cross-reference documents with Wikipedia. It explains how machine learning can be used to identify significant terms within unstructured text, and enrich it with links to the appropriate Wikipedia articles.

To rigorously address the Twitter entity linking problem, Guo et al [8] proposed a structural SVM algorithm for entity linking that jointly optimizes mention detection and entity disambiguation as a single end-to-end task. Merging mention detection and entity disambiguation into a single end-to-end entity linking task increase performance of system but this system use structural SVM algorithm which requires NP-hard inference which is computationally expensive.Sil and A. Yates [9] developed a re-ranking model that performs joint named entity recognition and entity linking. The discriminative re-ranking framework allow to introduce features into the model that capture the dependency between entity linking decisions and mention boundary decisions . Furthermore, the model can handle collective classification of entity links, at least for nearby groups of entities. The joint NER and EL model has strong empirical results, outperforming a number of state-of the-art NER and EL systems on several benchmark datasets while remaining computationally inexpensive. System tended to have much lower accuracy on long chains of entities.

TABLE I: SUMMARY OF LITERATURE SURVEY

Sr.No.	Paper Title	Authors	Methods/Techniques used
1	Named Entity Recognition using an HMM-based Chunk Tagger[1]	G.Zhou , J.Su[2002]	HMM-based Chunk Tagger.
2	Design challenges and misconceptions in named entity recognition[2]	L. Ratinov[2009]	Supervised learning algorithms (e.g., HMM , CRF).
3	TwiNER: Named Entity Recognition in Targeted Twitter Stream[3]	Chenliang Li, Weng [2012]	2-step unsupervised NER System, 1st step –tweet segmentation using dynamic programming algorithm.2nd step-use random walk model.
4	Exploiting hybrid contexts for Tweet segmentation[4]	Chenliang Li , Aixin Suny[2013]	Proposed tweet segmentation framework-HybridSeg for NER.
5	Community-Based Classification of Noun Phrases in Twitter[5]	F.C.T. Chua [2012]	Unsupervised approach based on POS tagger. Extract noun phrases as candidate named entity.
6	Large-Scale Named Entity Disambiguation Based on Wikipedia Data[6]	S. Cucerzan [2007]	NER system followed by a linking system.
7	Learning to Link with Wikipedia[7]	D. N. Milne [2008]	Proposed link detector and disambiguator.
8	To Link or Not to Link? A Study on End-to-End Tweet Entity Linking[8]	Guo et al [2013]	SVM Algorithm for entity linking.
9	Re-ranking for Joint Named-Entity Recognition and Linking[9]	A.Sil, A.Yate[2013]	Joint NER and EL Model.

IV. CONCLUSION

The Named Entity Recognition field has been growing for more than fifteen years. Its purpose is to find and classify mentions of rigid designators from text such as proper names and temporal expressions. In this survey, we have shown NER system and their approaches. We found that tweet segmentation has been proven to be effective in the tasks of NER. Tweet segmentation aggregates local context and global context to calculate the probability that segment being named entity. By doing so, we can be able to recognize named entities with high confidence and new named entities which may not appear in Wikipedia yet.

REFERENCES:

[1] G.Zhou and J.Su, "Named entity recognition using an hmm chunk tagger," in proc 40th Annu. Meeting Assoc.Comput.Linguistics, pp.473-480, 2002.

[2] L. Ratinov and D. Roth, "Design challenges and misconceptions in named entity recognition," in Proc. 13th Conf. Comput.Natural Language Learn, pp. 147–155, 2009.

[3] 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. 35th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, pp. 721–730, 2012.

[4] C. Li, A. Sun, J. Weng and Q. He, "Exploiting hybrid contexts for tweet segmentation," in Proc. 36th Int. ACM SIGIR Conf. Res.Develop. Inf. Retrieval, pp. 523–532, 2013.

[5] F. C. T. Chua, W. W. Cohen, J. Betteridge, and E.-P. Lim,"Community-based classification of noun phrases in twitter," in Proc. 21st ACM Int. Conf. Inf. Knowl. Manage, pp. 1702–1706, 2012.

[6] S. Cucerzan, "Large-scale named entity disambiguation based on wikipedia data," in Proc. Joint Conf. Empirical Methods Natural Language Process. Comput. Natural Language Learn, pp. 708–716, 2007.

[7] D. N. Milne and I. H. Witten, "Learning to link with wikipedia,"-in Proc. 17th ACM Int. Conf. Inf. Knowl. Manage , pp. 509–518,2008.

[8] S. Guo, M.-W. Chang, and E. Kiciman, "To link or not to link? A study on end-to-end tweet entity linking," in Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Human Language Technol., pp. 1020–1030, 2013.

[9] A. Sil and A. Yates, "Re-ranking for joint named-entity recognition and linking," in Proc. 22nd ACM Int. Conf. Inf. Knowl. Manage, pp. 2369–2374, 2013.