ISSN No. 0976-5697

Volume 8, No. 5, May – June 2017

International Journal of Advanced Research in Computer Science

REVIEW ARTICLE

Available Online at www.ijarcs.info

Recognizing Named Entities in Text based User Reviews

Richa Chaturvedi
Department of Computer Science & Engineering,
Amity School of Engineering Technology,
Amity University, Lucknow, Uttar Pradesh, India

Dr. Deepak Arora
Department of Computer Science & Engineering,
Amity School of Engineering Technology,
Amity University, Uttar Pradesh, Lucknow, India

Bramah Hazela, Pooja Khanna
Department of Computer Science & Engineering,
Amity School of Engineering Technology,
Amity University, Uttar Pradesh, Lucknow, India

Abstract: Named Entity Recognition is a subtask in Information Extraction. In Named Entity Recognition (NER) we try to identify each of the words provided into some categories predefined by us. These categories or classes can be organization, name, time, place etc. Named Entity Recognition is a part of Natural Language Processing which is aims at making human and computer interactions more meaningful and efficient. NLP is nowadays being effectively being used in the area of automatic text summarization, cross language information retrieval, speech recognition and query named entity recognition. This paper focuses on Named entity recognition, how it is performed and a how apply NER to web search queries so that correct and more user personalized results are displayed by the web search engines..

Keywords: Natural Language Processing (NLP), Named Entity Recognition (NER), Message Understanding Conference (MUC)

I. INTRODUCTION

THIS paper represents the study of Named Entity Recognition system which is a part of Natural Language Processing. Updates posted on various Social Media platforms like Twitter and Facebook has opened for us a new area for research. Tweets are up to date and provide inclusive data which can be great help for companies that want to target a particular type of clients. The numbers of tweets are increasing day by day and hence tools like named-entity recognition are used on these sentences for effective outcome [1]. Typically Named Entity Recognition is a task that extracts entity classes like location, person, time etc. A named entity can be a proper name, or a common name. For example, named entities can be "Nancy Drew," "Trump," or "Cold Play." When a sentence or query is submitted it is analyzed such that words present in the sentence are correctly classified into entities. Basic categories of entities are: Enamex which includes names of person, organization, location etc. Timex- includes date and time and Numex includes numbers such as money and percentages [14].

Search queries usually consist of linguistic units that are submitted by the user to the search engines. Search queries typically consist of words that specify the users need. But these queries are unstructured, ambiguous and also short; techniques to classify named entities are required. Using NER (named entity recognition) search engines can make sense about the contents on web as interconnected entities.

Table I. Example of Named Entity Hierarchy. Question marks means there exists context-dependency

Named Entity	Proper Noun	Rigid Desig.	Unique identifier	Example of domain/purpose	Source and NE type
water	No	Yes	? (sparkling or still)	Chemistry	SH: natur. obj-mineral
$10million	No	?	? (American dollars or Mexican pesos)	Finances	MUC-7: Money
whale	No	?	? (orca or minke)	Biology	SH: sea animal
Bedouins	No	?	? (specific people)	Army/ News	CoNLL03: Miscellanea
Airbus A310	Yes	?	? (specific airplane)	Army/Tech. Watch	CoNLL03: Miscellanea
Batman	Yes	Yes	? (people disguised)	Movies/Costumes	ACE: Person
Cytokine	Yes	?	? (specific proteins)	Biomedicine	GEN: protein family / group
twelve o'clock noon	No	?	? (specific day)	News	MUC-7: Time

Let us take an example and consider a suppose 'Bill Clinton work' is a query and if the user of search engine submits the query following the English spelling rules, like capitalization of nouns, then 'Bill Clinton' will be detected as a name and a noun easily. This is not the case in all the scenarios some users express the information need, queries in cases we don't have the enough contexts about so classifying them as entity is difficult. Sometimes, orthographic rules, such as capitalization, cannot be relied upon or trusted because some users who are not aware of these rules often do not follow the orthographic rules of spelling when entering their search queries into the search engines [7]. Furthermore, search queries tend to be highly ambiguous as are the contextual clues embedded within these queries, and resolving this ambiguity is very important to optimize the query end result. Finally, the Web is now consistently changing and also search queries, so relying completely on supervised named entity

recognition models made or customized for specific search queries may also not be able to give us the desired results.

Search queries are added with snippets that have named entity recognition of the search results displayed at the very top related to the query. Hence, a snippet should always resort to inform the user of the search engine who submitted the query about what the page is about and how to provide context within which the search keywords occur. Furthermore, often enough contextual clues so that named entities are identified and classified are provided. Every query entered is separately processed referring to a proper name or address or time is extracted

Google first patented the technique for re-writing query where named entity recognition was introduced. The following year, Microsoft patented their augmenting 'blocks'. Google's acquired Metaweb, it operates the Freebase directory. Recently, in the year 2013, Microsoft also patented 'entity-based' which enables detection of named entities and it also organizes search results according to the search query.

Different named entity recognition (NER) techniques are being utilized. Search log is now being used for NER. Typically a search log is kind of repository which is maintained by search engines to record its user's activities along with their submitted search queries. The use of extracting named entities from respective search log is that they have mentions of named entities that are written in the form of users perspectives.

The research area confined with the named entity recognition in web search queries have to solve the following questions:- Considering the lack of time or brevity and lack of correct grammatical structure usage how can we identify and classify named entities so that our search queries give an optimized result? How can we identify the distinction between named entities when there is more than one named entity in a search query?

Thus, an approach for NER and classification in queries are required that satisfy the users need for information.

II. LITERATURE SURVEY

The classes of the named entities can be labels, topics, or categories. A named entity may have more than one class or label. Furthermore, it will be based on the query whether or not one or more classes may be more probable classes, and others can become less likely classes. For example, search query "The Iron lady biopic" we can find "lady" as a named entity, and we now can assign "Movie" to it and "Game" or "Book" as less like probable classes, and "Music" as not probable class. If the search query was only "iron lady," then "Person" and "Politician" may be more probable classes..

Approaches for named entity recognition can be supervised or unsupervised, make use of HMM[6] and SVMs. Neural networks and Maximum entropy classifiers can also be used. Most of the systems use approaches that are based on either statistical or are rule based. Rule based approach was developed by Ralph Grishman in the early 1995, in this approach a large dictionary had to be developed which had names of all probable first names, locations, organizations, books, movies etc. The cons of this type of approach is that there is a requirement of huge data and grammatical knowledge and also these system developed for one language or area field cannot be transferred to another language or field on the other hand an unsupervised approach won't require such a large database of named entities[10].

We can evaluate the quality of an output of the NER system three terms: Precision, Recall and F score. Precision is the number of how many positive predictions were actual positive observations. Recall is how many correctly predicted positive observations are exiting to the all observations in actual class. F1 score is mean of both precision and recall. F measure values can be derived by [10]:

$$F1 \text{ score} = (\gamma\ 2 +1)PR/(\ \gamma\ 2R+P) \quad (1)$$

γ is the weighting between precision and recall

Recall and precision are tasks like IR and text categorization, where there typically is one grain size (documents) (Cunningham et all,1997).These measures behave funnily for NER when there are presences of boundary errors (which are common)

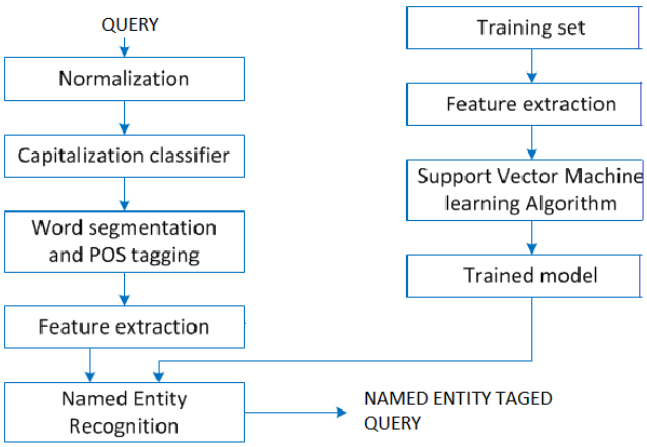

Fig. 1. NER model

Three standard approaches that can be utilized to classify named entity are:

1. Hand-written regular expressions
2. Classifiers - Generative: Naïve Bayes and Discriminative: Maxent models [15].
3. Sequence model- HMMs [6], CMMs and CRFs

For texts that are unstructured or human-written, following can be used efficiently,

1. Part-of-speech (POS) tagging –Here we will underline each word as a proper noun and preposition [3].
2. Syntactic parsing – This will Identify phrases into PP,VP,NP

The importance of classifying named entity or recognition of this named entity is because of the fact that Entity names tend to form the main content of any document. Named entity recognition is often used as an initial step for a chain of processes. The further step of processing can be to find the similarities between two or more Named entities that were recognized at the first step. Further processing on data can help in discovering "what" or "how" present in the given query [5]. Word sense disambiguation is also a problem that can be solved by linking named entities to each other. By

word sense disambiguation we mean to identify which context of the word is used in the given sentence. For example the word 'minute' has two different meanings one is related to time and other means a small amount of something. If using NER we could identify which 'minute' is used in the sentence we can improve the search results. Considerable amount of research work has been done which aims to decrease the ambiguity, increase the robustness and also increase the portability of an NER system. When the corpus of NER tool changes it has been noted that performance of tool decreases [2].

Identifying the named entities present in the search queries submitted can be useful to us because understanding search intents can become better. For example, just by classifying a named entity, it can become easier to provide better search results and thus improvement in the ranking in a relevance search can be obtained. Furthermore, query suggestions that are more suited can be showed by treating named entities and contexts separately. For example, suppose using the search query given by the user "Hercules walkthrough," the following query can be suggested- "Hercules cheats" this can be a more relevant suggestion by the search engine because it has the context in the same class. Also, "Grand theft auto walkthrough" can also be one of the relevant suggestions because it has a named entity in the same class which is gaming [9].

A new probability using approach toward Query Named Entity Recognition uses the log data or click-through data can also be applied. Without the losing much information contained within a query, one can represent named entity can as triple of (e, t, c.), where 'e' means a named entity, t means the context of the named entity, and 'c' denotes the class of the named entity [9]. Here 't' can be kept empty (i.e., no context is present) in search queries having only a named entity, an input query. The goal of one example approach to NERQ is that the finding of the triple (e, t, c) for a given search query q which also has the largest joint probability P(e,t,c). That also points out that the context and the class of a named entity the user most probably intended in the query. A context may be thus being blank, null, or empty if a query contains only a named entity, and no other terms. For example [8] the word "music" can take context as "lyrics" or "mp3", the word "Avatar" can have two classes "Game" or "Movie"

The steps for recognizing the named entities present in a search query has the following steps:

1. First step is to choose a seed point that comprises a seed named entity;
2. Second step consists of assigning a classification to the seed chosen and the named entity based on a predefined taxonomy;
3. Thirdly, training the probabilistic topic model which is based on the seed choose and thus named entity and the classification is done;
4. Then we will receive an input query from a user;
5. Then detecting, by the processor, another named entity in the input query;
6. Next step is representing the input query as one or more triples (e, t, c), where e denotes the detected other named entity, t tells the context of the detected other named entity, and c will represents a particular classification for the detected other named entity that is determined;
7. Now we will calculate a joint probability for individual ones of the one or more triples;
8. Identification of the largest joint probability associated with the respective one of the one or more triples;
9. Prediction is done, by the recognition module, and an assigned classification for the other detected named entity based at least in part on the identified largest joint probability;
   And lastly
10. Presenting the detected other named entity and the assigned classification to the user.

Some application of Named Entity Recognition is to implement filtered search for text query input. NER is used in phrase based auto suggested resolution and also in Question and answering system to detect entities which are not explicitly defined under the particular discussion (Guo et al., 2009). Thus each Question and answering system discussion can become easier to understand. NER is also used to tag contents and listing all over the website and to improve property posting experience of user. We can resort to show preselected fields for which the user is reluctant or lazy to choose from a drop down menu. Many service providers put all projects they deal in to come up in search results but hamper the search relevance. Thus steps can be taken to detect spamminess. Spamminess means keyword stuffing phenomena.

III. CONCLUSION & FUTURE WORK

Since NER is a recent area of research in comparison with other language processing approaches, still we have seen ample research work done in this area and this area will possibly continue to be a major area of research since many uses of NER have been found which will make existing systems to work more efficiently [4]. Many NER tools are available online that are open source and have their own corpus some of the tools are Stanford NLP, Mallet, Natural Language Toolkit, Apache NLP and GATE. Named Entity Recognition is now also being applied on search queries by various researches so that the result of the search is more specific and more closer to what the user wanted. The above literature review specifies the uses of NER and states some of the methodologies used. Some of the steps how named entities are performed on search queries are stated. In future further work can be done for NER in Hindi search queries. In future a new technique or a model can be developed so that the performance of QNER is enhanced in Hindi. Moreover, New rules can be derived that are according to Hindi language and thus improvement in the performance of QNER system which uses Hindi language can be used as compared to English, Hindi does not support any capitalization of proper nouns . In future the size of corpus can be increased by adding more name entities in database as more the corpus size means more is the accuracy that can be achieved.

IV. REFERENCES

[1] Bandyopadhyay, A., Roy, D., Mitra, M., Saha, S.: Named entity recognition from tweets. In Proc of 16th LWA Workshops: KDML, IR and FGWM, Aachen, Germany, September 8-10, 2014, pp. 218–225.

[2] M. Marrero, S. Sánchez-Cuadrado, J. Morato Lara, and Y. Andreadakis, "Evaluation of Named Entity Extraction Systems,"

In proc of International Conference on Intelligent Text Processing and Computational Linguistics, 2009.

[3] Helmut Schmid., Probabilistic part-of-speech tagging using decision trees. In Proc of International Conference on New Methods in Language Processing, 1994

[4] Wohleb, R. "Natural Language Processing: Understanding Its Future," PC/AI, Nov. /Dec., 2001.

[5] Y. Shinyama and S. Sekine, "Named entity discovery using comparable news articles," in International Conference on Computational Linguistics, pp. 848-853, 2004.

[6] Daniel M. Bikel, Scott Miller, Richard Schwartz and Ralph Weischedel. 1997 "Nymble: a highperformance learning name-finder", In proc of International conference on Applied natural language processing, pp. 194-201, San Francisco, CA, USA Morgan Kaufmann Inc.

[7] Downey, D., Broadhead, M., Etzioni, O.: Locating complex named entities in web text. In: IJCAI 2007, Proceedings of the 20th International Joint Conference on Artificial Intelligence, Hyderabad, India, January 6-12, 2007. pp. 2733–2739 (2007)

[8] S. Bergsma and Q. I. Wang. Learning noun phrase query segmentation. In EMNLP-CoNLL '07, pages 819–826, 2007.

[9] Jiafeng Guo, Gu Xu, Xueqi Cheng, Hang Li, Named Entity Recognition in Query Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval,2009

[10] Darvinder kaur, Vishal Gupta, A survey of Named Entity Recognition in English and other Indian Languages, IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 6, November 2010

