IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 2, February 2017

# Named Entity Recognition Approaches and Challenges

**Sanjana Kamath[1], Rupali Wagh[2]**

Student, Department of Computer Science, Christ University, Bengaluru, India [1]

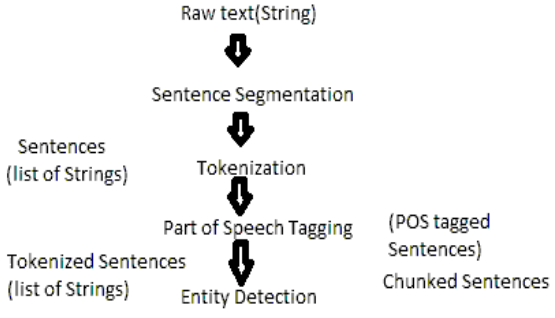
Associate Professor, Department of Computer Science, Christ University, Bengaluru, India [2]

**Abstract:** With the increase in availability of data, extraction of useful information from this data has become most important activity across all domains. When the data is available as documents written in natural language, information extraction becomes more challenging. Named Entity recognition (NER) is a technique used extensively for automatic extraction of useful information from unstructured natural language document collections. Used both, for web applications as well as stand-alone systems, NER is considered as one of the major step in Natural Language Processing (NLP) for analysis of text. This paper discusses basics of NER, various algorithms used for NER and major applications and challenges in the field of NER.

**Keywords**: Named Entity Recognition, Information Extraction, Natural Language Processing, Part of Speech.

## I. INTRODUCTION

Named Entity Recognition is one of the major tasks in Natural Language processing (NLP).NER is an active area of research for past many years. Named Entity is a word or a phrase of sentence that clearly identifies an item. Automatic data Extraction is one of the prime applications of Natural Language Processing. Most Information Extraction systems follow rigid approaches of extracting information. These approaches are generally specific to a document or a particular language. NER frameworks are used to find and arrange words in content into predefined classifications for example, the names of people, associations, areas, articulations of times, amounts and so forth. E.g.: kejriwal went to Delhi. Kejriwal and Delhi is named substances where a NER system needs to identify kejriwal as <Name of a person> and Delhi as <Name of a place>. Named-element acknowledgement is a subtask of data extraction that looks to find named elements. In many domains, data related to NER may be confidential and a technique is required to which it will process any kind of documents without any kind of pre requisite knowledge. One of the major roles of NLP is to generate models that will be useful for human to machine type communication.FIG 1 IDENTIFYING NAMED ENTITY




The Way of Identifying an Entity from a raw text and sorting these texts into sub groups is called Entity Recognition. Basic steps of NER are shown in the above diagram.

This paper presents a review of basic NER process, current approaches and techniques used in NER, NER applications and challenges.The paper is organized as follows. The second section of this paper presents related work in the field of NER. The third section gives,an overview of main tasks and applications of Named entity recognition and, the fourth section discusses about the challenges.

## II. RELATED WORK

The paper titled "**Named Entity Recognition: A Literature Survey" [1]**explains supervised techniques, semi-supervised and unsupervised techniques used for NER. Named entity Recognition is a major task in Natural Language processing. Main focus of this paper is to improve the NER system mainly for Indian languages.The paper has used statistical methods for identification. The identification of entities is done by identifying entity names i.e. place names, temporal expressions and numerical expressions. Entities can be identified using supervised learning algorithms. Hidden Markov models, decision trees and support vector machines are highlighted as most popularly used supervised learning algorithms for NER .Semi-supervised algorithms use both labelled and unlabelled data sets. These algorithms usually identify small seed data and move to large amount of un annotated data. The paper has also discusses Named entity Recognition tasks, its applications and motivations for pursuing research in this area for future.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 2, February 2017

The paper titled "**NERD: Evaluating Named Entity Recognition Tools in the Web of Data**" **[2]** highlights the important role of NER in extracting structured information by identification important features from texts called as Entities and linking these entities to web Resources by typed interpretations. The paper also proposes algorithms for extracting of texts like identifying name of people, location, organization names, time and quantity related texts. These algorithms are used for improvising search operations, and finding a way of getting meaningful relation between the extracted entities. The algorithmscan be applied for extracting semantic data sets to identify person names, organization names and are further divided into categories. The paper also focuses on improvement of content searching, and finding useful meanings from the extracted entities. The paper demonstrates experimental evaluation of extracting entities with a human driven technique like algorithms. Two main methods are demonstrated in the paper -Controlled Experiments and uncontrolled Experiments (controlled experiments results in asking few people to evaluate the output received from the same news article and rating the same and uncontrolled experiments refers to giving the same task to people to rate in any article without any restrictions on the article) .The evaluation of the same was performed and focused on different precisions such as classification of information into different units.

The Paper titled "**Design Challenges and Misconceptions in Named Entity Recognition"[3]**presents a typical way of representing named Entity Recognition that uses various features to identify a new art of named Entity Recognition. Four fundamental features namely text chunks representation, inference algorithm, using non-local features and external knowledge were used for identification purpose. It also states that Named Entity Recognition is an intensive task that requires prior knowledge across several different domains.

The paper titled **"Named Entity Recognition from Diverse Text Types"[4]** focuses on Named Entity recognition system which mainly aims to reduce the need for costly and time-consuming systems to new applications. This demonstrates the automation of this process as to which resource to use, the implementation or the outcome of the resource mainly depends on the performance and different data representation of the algorithm.

The Paper Titled "**Named Entity Recognition in Tweets: An Experimental Study"**[**5]**addresses the approaches and challenges related to the style of unintentional language being used in social media's these days.The paper alsohighlights the issues in classifying named entities in tweets. Two major concerns mentioned are varied named entity types like brand names, company names movies, no restriction on length of the texts or named entitiesetc and lack of background knowledge .And to solve this issue the proposal was related to distantly supervised approach to accumulate large amounts of data and related dictionaries. To address the existing challenges tools have been built that train on both labelled and unlabelled data.

The Paper titled **"Named Entity Recognition for Question Answering" [6] proposes a different approach** wherenamed entity recognition is done is in a question answer format. As compared to named entity in the traditional way this paper has a better way of getting results. The methodology gives a better chance to find questions to all answers. The future work of this paper includes multiple labels on Entity Recognition on higher Question Answering systems.

The paper titled **"Named Entity Recognition: Exploring Features**" **[7]** focuses on complete features in identifying supervised NER, and various combinations of features and their result on recognizing performance. This paper mainly focuses on variety of features,which are mined from a word being labelled and analysis is done on the same and the effectiveness of a supervised NER system, various individual features and combinations on the effectiveness of named entity recognition is also focused in the paper. The paper aims to extend their work on clustering features and their effective combinations of Named entity recognition.

The Paper titled "**A Survey of Named Entity Recognition and Classification"[8]**aims at improving the named entity recognition for Indian languages using both supervised and non-supervised methods, and various statistical measures are used for the study of the same and also study of neural network approaches for named entity recognition. The paper made observations like language factor, domain factor; entity factor etc.The paper emphasizes on the suitability of neural networks the field of NER.

The paper titled "**Named Entity Recognition for Indian Languages:" A Survey [9]** deals with how languages play a vital role in NER, Language being the fundamental goal for communication and helps in enabling machine type communication. This paper tells about how language plays a vigorous role in hearing, talking, speaking etc. The major part of NER is to identify and categories different words in a text format into its subsequent categories like person name, place names, quantities etc. E.g.: Sonia is from Gujarat. This can be identified in two ways Sonia is a person name and Gujarat a place name .The paper came up with 13 noun taggers for entity recognition like person names, location names and organization names ,also used Hidden markov model in supervised learning technique and statistical models with generalized learning method in this paper. The major challenge of this paper is that all Indian languages do not have capitalized forms of nouns and Indian languages are varied when compared to other languages. The paper deals with languages such as Oriya, Punjabi and related Indian languages.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 2, February 2017

The paper titled **"Named Entity: History and Future" [10]** discusses about the history of NER future of the same, the problems faced and how these problems could be solved. The paper describes how drastic changes are taking place in this field, changes from tagging of only proper names to tagging a wide variety of words and expressions which humans call it information. The paper demonstrates results after three types of study namely weakly supervised, active learning and unsupervised learning. The weakly learning focuses on extracting relations of entities such as book titles, author names etc, active learning have better outcomes that could be achieved by annotating each data that has been tagged. Unsupervised learning refers to data without label. Entity Recognition plays a vital role as a technology for applications in the field of natural language processing.

The paper titled **"Named Entity Recognition using Machine Learning and pattern Selection Rules"[11]**discusses about significance of machine learning in NER. The paper has proposed methods and rules namely hybrid method and maximum entropy model. The data used are extracted from tagged data sets. This model focuses on unidentified words. The paper mainly focuses on entropy model and this work can be shifted to any other domain as the data is trained data set.

The paper titled **"A survey of Named Entity Recognition in English and other Indian Languages"[12]**presents overview done on different methodologies of named entity recognition in different Indian languages. Paper highlights language specific aspects of named entity recognition with respect to Spanish, Chinese. Also a study is done with respect to the work done in other Indian languages like Oriya and Punjabi.

The paper titled **"Constructing Dictionaries for Named Entity Recognition on Specific Domains from the Web" [13]** addresse show automatic dictionaries can be constructed for Named Entity Recognition on specific domains such as maps, restaurant guides and so on. It also explains about how NER is the first step towards Information Extraction and developing dictionaries plays a very important role. NER deals with wide variety of domains on the Web. This paper mainly focuses on improving the performance of NER by creating dictionaries that are created using HTML documents .The major aim of this paper is dictionary creation on specific domains using for Entity Identification and also emphasizes on applications of such automated dictionaries on web.

The paper titled **"Named Entity Discovery Using Comparable News Articles" [14]** Discusses the importance of NER in the field of News articles. In this field, data sparseness (if no information exists in a document) is a major concern for NER. The paper overcomes this by observing that NER appears in a regular pattern in news articles and the way common nouns are used. The paper describes the techniques used and is classified using time series distribution of entities. With the help of newspapers that appear in a regular pattern. This research worked with newspaper as sample based on 365 days newspaper. A method called normalization was applied by dividing the number of article's containing each word by total number of articles on that day. The research was specific to this field as newspaper data depends on date and time.

The paper titled**" A Simple Semi-supervised Algorithm for Named Entity Recognition" [15]**focuses on a simple semi supervised algorithm to identify unlabelled data .The data collected are domain independent. The accuracy of algorithm was achieved when the data that has been used are for different types of domains. It also states that compared to other semi supervised learning, the algorithm proposed in this paper has better performance.

## III. APPLICATIONS OF NER

Major Applications of Named Entity Recognition are in different fields. Amount of data available on the web has increased exponentially in last few decades. NER is used primarily for automatic extraction of data from unstructured documents. Major applications of Named Entity Recognition are as mentioned below: NER is used effectively for automatic identification of events like disasters and crimes. The articles can be extracted using web based news aggregation method (is software or an application which extracts web substances). A study is done on different languages. The application of gathering news articles is not easy due to the complex language used in the articles. Information given in the newspaper are generally scattered all around in several sentences and different documents which is a major Task to identify NER. Another Important observation that NER is seen synchronously in lot of news articles, Synchronicity of names i.e. how often an Entity appears in a news article. Nouns, proper nouns etc. can also identified and can be used for populating databases.

Named Entity Recognition is popularly used for identification of patient names, patient address etc. from Electronic Medical Records. An automated tool which uses NER is used to process patient information. The Extracted Information can be useful for professionals in medical field for further study. The tool extracts patient's information and cla*ssifies them accordingly into different categories. Study of NER in the field of Medical Domain is very significant. Medical Entity Recognition refers to a sub task of information extraction that locates entities in medical field (E.g. Medicine names, Medical tools etc.). Major challenge in this field is the complicated terms (medicine names) used and hence makes it difficult for identifying.

Social media analysis requires NLP applications.NLP tool are used extensively for analysing tweets in twitter and

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 2, February 2017

other Social Media is through using parts of speech tagging, and through chunking. NER is used in such domains for recognizing different Entities. Messages which are posted on Facebook and other social media may be informal in nature. Hence classifying these named entities in social media is a difficult task .The text varies from company names, movie names, brand names etc. Another Application in this field is identifying real life events from documents that are available on the web that is gaining lot of fame .The method integrates NER, topic clustering and event detection. NER is gaining lot of popularity in social media analysis as social media like twitter and Facebook are used by people for their opinions which form the basis of many business processes.

## IV. CHALLENGES

NER, though considered to be a basic NLP function, is challenged by various complexities that are inherent in any natural language. Few of the challenges are described below:

**Ambiguity and Abbreviations** -One of the major challenges in identifying named entities is language. Recognizing words which can have multiple meanings or words that can be a part of different sentences. Another major challenge is classifying similar words from texts.

Multiple words or sentences can be written in different forms .Words can be abbreviated for ease of writing and understanding .Same words can be written in long forms. Words which will sometimes require some label for identification is another major challenge.

**Spelling Variations**-The vowels (a, e, i, o, u) in English language plays a very important role. Words which do not make a major difference in phonetics but make a major difference in the way of writing and its spelling.

**Foreign Words**-Words which are not used very frequently these days, or words that are not heard by a lot of people, is another major challenge in this field. Words like person names, location names etc.

## V. CONCLUSION

Named Entity Recognition, a sub process of natural language processing, plays very important role in automated information extraction. In today's web world where huge amount of information is available as natural language documents, NER has gained lot of significance. Basic tasks, algorithms used for NER and major application domains are discussed in the paper. This Paper also discusses major challenges in NER.

## REFERENCES

[1] Sharnagat, Rahul. "Named Entity Recognition: A Literature Survey." (2014).

[2] Rizzo, Giuseppe, and Raphaël Troncy. "Nerd: evaluating named entity recognition tools in the web of data." (2011): 1-16.

[3] Ratinov, Lev, and Dan Roth. "Design challenges and misconceptions in named entity recognition." Proceedings of the Thirteenth Conference on Computational Natural Language Learning. Association for Computational Linguistics, 2009.

[4] Maynard, Diana, et al. "Named entity recognition from diverse text types." Recent Advances in Natural Language Processing 2001 Conference. 2001.

[5] Ritter, Alan, Sam Clark, and Oren Etzioni. "Named entity recognition in tweets: an experimental study." Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2011.

[6] Mollá, Diego, Menno Van Zaanen, and Steve Cassidy. "Named entity recognition in question answering of speech data." Proceedings of the Australasian Language Technology Workshop. 2007.

[7] Tkachenko, Maksim, and Andrey Simanovsky. "Named entity recognition: Exploring features." KONVENS. 2012.

[8] Nadeau, David, and Satoshi Sekine. "A survey of named entity recognition and classification." Lingvisticae Investigationes 30.1 (2007): 3-26.

[9] Pillai, Anitha S., and L. Sobha. "Named entity recognition for indian languages: A survey." International Journal 3.11 (2013).

[10] Sekine, Satoshi. "Named entity: History and future." Project notes, New York University (2004): 4.

[11] Seon, Choong-Nyoung, et al. "Named Entity Recognition using Machine Learning Methods and Pattern-Selection Rules." NLPRS. 2001.

[12] Kaur, Darvinder, and Vishal Gupta. "A survey of named entity recognition in english and other Indian languages." IJCSI International Journal of Computer Science Issues 7.6 (2010): 1694-0814.

[13] Shinzato, Keiji, et al. "Constructing dictionaries for named entity recognition on specific domains from the Web." Web Content Mining with Human Language Technologies Workshop on the 5th International Semantic Web. 2006.

[14] Shinyama, Yusuke, and Satoshi Sekine."Named entity discovery using comparable news articles." Proceedings of the 20th international conference on Computational Linguistics. Association for Computational Linguistics, 2004.

[15] Liao, Wenhui, and Sriharsha Veeramachaneni. "A simple semi-supervised algorithm for named entity recognition." Proceedings of the NAACL HLT 2009 Workshop on Semi-Supervised Learning for Natural Language Processing. Association for Computational Linguistics, 2009.