Chapter 1

1 INTRODUCTION

The term “ Named Entity ” (NE) is in current use in Machine Translation(Information Extraction or Information Retrieval) applications. It was introduced at the sixth Message Understanding Conference (MUC-6)(Grishman and Sundheim 1996), which motivated Information Retrieval(IR) or Information Extraction (IE) researches in the 1990s. At the time, MUC was focusing on Information Retrieval(IR) or Information Extraction(IE) tasks wherein structured information on company and defense-related activities are extracted from unstructured text, such as newspaper articles. It is essential to recognize information units such as names of person, organization, location, numeric expressions, including time, date, money and percentage for defining Information Retrieval(IR) or Information Extraction(IE). This process of identifying named entity from text or document which also an important sub-task of Information Retrieval(IR) or Information Extraction (IE) is called " Named Entity Recognition " (NER). Before the NER field was recognized in 1996, significant research was conducted by extracting proper names from texts. A paper published in 1991 by Lisa F. Rau (1991) is often cited as the root of the field.

For more than fifteen years, a dynamic research community advanced the fundamental knowledge and the engineered solutions to create an NER system. In its canonical form, the input of an NER system is a text and the output is information on boundaries and types of NEs found in the text. The vast majority of proposed systems fall in two categories: the handmade rule-based systems; and the supervised learning-based systems. In both approaches, large collections of documents are analysed by hand to obtain sufficient knowledge for designing rules or for feeding machine learning algorithms. Expert linguists must execute this important amount of work, which in turn limits the building and maintenance of large-scale NER systems.

Named Entity Recognition (NER) involves processing structured and unstructured documents and identifying expressions that refers to people, places, organizations and companies and so forth. For humans, Named Entity Recognition is intuitively simple. Many named entities are proper nouns and have initial capital letters and can easily be recognized that way. But in case of Indian languages it is very difficult. There is no capitalization on leading character words or proper nouns in the sentence.

Named Entity Recognition (NER) can be used for different tasks. It can be used as a self – standing tool for full text searching and filtering. Also it can be used as a pre-processing tool for other Natural Language Processing (NLP) tasks. These tasks can take advantage of marked named entity (NE) and handle them separately, which often results in better performance. Some of these tasks are Machine Translation, Question Answering, Text Summarization, Language Modeling and Sentiment Analysis.

1.1 Different approaches of NER

There are different methodologies for Named Entity Recognition(NER) detection. These approaches are :
A. Rule based or Linguistic approach.
B. Machine learning (ML) based approach.
C. Hybrid approach.

A. Linguistic or Rule based approach.

In rule based or linguistic approaches linguists uses their hand written rules for Named Entity Recognition (NER) detection . These rules are language specific. So different rule based Named Entity Recognition (NER) system are:
a. Lexicalized grammar.
b. Gazetteer lists.
c. List of trigger words.

B. Machine learning (ML) based approach

The machine learning methods used for accurate Named Entity Recognition (NER) detection are as follows:
a. Hidden Markov Models (HMM).
b. Decision Trees.
c. Maximum Entropy Models (ME).
d. Support Vector Machines (SVM).

e. Conditional Random Fields (CRF).

All of the above machine learning approaches have their own advantages and disadvantages. Maximum entropy model cannot solve the label biasing problem. Markov Models can solve the sequence labeling problem very efficiently. The conditional probabilistic characteristic of Conditional Random Fields and Maximum Entropy Models are very useful for development of Named Entity Recognition (NER) system. Conditional Random Fields is flexible to capture many correlated features, including non-independent and overlapping features [P. K. Gupta and S. Arora 2009].

C. Hybrid approach

We can uses both rule based and machine learning methods in Hybrid Approach. So we can combine any of the two methods in Hybrid Approach in order to improve the performance of the Named Entity Recognition (NER) system. So the hybrid approach may be combination of Hidden Markov Model and Conditional Random Field (CRF) model or Conditional Random Field (CRF) and Maximum Entropy Models (ME) approach. In our thesis we consider the Hybrid Approach i.e. Gazetteer method and Hidden Markov Model to increase the accuracy of the Named Entity Recognition (NER) System.

Table 1: Comparison of Machine learning approach and Rule based

Rule Based Approach	Machine Learning Approach
This approach contains set of hand written rules. Rules are written by the language experts so for this approach human experts are required.	Developers do not need language expertise.
Require only small amount of training data.	Require large amounts of annotated training data.
These systems are not transferable to other languages or domains.	Once we build the machine learning based system may be used other language or domains.
Development can be very time consuming.	It requires less human effort.
Some changes may be hard to accommodate.	Some changes may require re-annotation of the entire training corpus.

1.2 Current status in Named Entity Recognition (NER) for Indian Language(IL)

Research in the field of Named Entity Recognition (NER) in Indian languages is still in the initial stage as compared to other foreign languages such as English, Spanish, and Chinese etc. For European Languages especially for English and for East Asian language Accurate Named Entity Recognition (NER) systems are now available. The problem of Named Entity Recognition (NER) is still far from being solved for south and South East Asian languages. For Indian languages there are many issues which make the nature of the problem different.

For example:- The number of frequently used words (common nouns) which can also be used as names (Proper nouns) is very large for European language where a large proportion of the first names are not used as common words.

1.3 Issues with Indian Language

Lots of Named Entity Recognition (NER) system has been built for English Languages. But for Indian Languages we cannot use such Named Entity Recognition (NER) system because of the following reason [Padmaja Sharma, Utpal Sharma, Jugal Kalita 2011]:

a. Indian languages lack the capitalization information that plays a very important role to identify Named Entities in those languages, Unlike English and most of the European languages.

b. In Named Entity Recognition (NER) the detection of Indian names are very difficult task because of the ambiguity problem.

c. Indian languages are resource poor language. Annotated corpora, name dictionaries, good morphological analysers, POS taggers etc. are not yet available in the required quantity and quality [Suleiman H. Mustafa and Qasem A. Al-Radaideh 2004].

d. Lack of standardization and spelling [Suleiman H. Mustafa and Qasem A. Al-Radaideh 2004].

e. Web sources for name lists are available in English, but such lists are not available in Indian languages.

f. Although Indian languages have a very old and rich literary history still technology development are recent [Arindam Dey, Dr.Bipul Syam Purkayastha 2013].

g. Non-availability of large gazetteer.

i. Named entity recognition systems built in the context of one domain do not usually work well in other domains.

j. Indian languages are relatively free-order languages [Sujan Kumar Saha, Sudeshna Sarkar, Pabitra Mitra].