ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 5, Issue 3 , March 2018

# Analysis of Two Different Approaches for Named Entity Recognition Based on Natural Language Processing.

**Samriddhi Jain, Vijeta Shah, Prof. Sarita Rathod**

U.G. Student, Department of Information Technology, K.J. Somaiya Institute of Engineering and Information Technology, Mumbai, Maharashtra, India
U.G. Student, Department of Information Technology, K.J. Somaiya Institute of Engineering and Information Technology, Mumbai, Maharashtra, India
Professor, Department of Information Technology, K.J. Somaiya Institute of Engineering and Information Technology, Mumbai, Maharashtra, India

**ABSTRACT**: Recognition of the Named Entities[NER] refers to a task of data extraction which is responsible for storing, sorting and searching textual content into pre-defined categories such as the name of a person, organizations, locations, expression of time, quantities, monetary values, and percentages. Recognition of the Named Entities can be implemented using two different approaches such as Rule Based Approach and Statistical Based Approach.

In this Project we do a comparative study of these two approaches on various types of inputs given by us on the named entities of name of person, organization, and location and it will analyze the outcome on the basis of parameters such of Recall, Precision, and F-Measure and determines whether the Rule Based Approach or the Statistical Based Approach should be implemented for better performance and efficiency in Named Entity Recognition.

**KEYWORDS**: Named Entity Recognition, Rule Based, Statistical Based, Recall, Precision, F-Measure, Person, Location, Organization.

## I.INTRODUCTION

### A. Natural language processing:

Processing of Natural Language [NLP] comes under domain of computer science, artificial intelligence and computational linguistics which deals with the interactions between computer and human (natural language), and, in particular, concerned with programming computers to fruitfully process large corpora.
The ultimate task of the natural language processing is to build software that will analyze, understand, and generate human languages naturally, enabling communication with a computer as if it were a human itself [5].

### B. Named entity recognition:

Since we will be using the term "Named Entity", let us define it. It was first introduced by Grishman and Sundheim and is widely used in Natural Language Processing.
Recognition of Named Entity is a sub-task of extracting information that seeks to search and classify named entities (recognizing proper nouns) in text into pre-defined categories such as names of person, organization, location, expression of times, monetary values, percentages, and quantities, etc. Few years ago, the researchers were focusing on extracting structured information from the unstructured text like newspaper articles.
Recognition of Named Entities also plays an important role in reference resolution, different types of disambiguation, and meaning representation in other natural language processing applications. Semantic parsers, part of speech taggers, and thematic meaning representations could all be extended with this type of tagging to provide better results [3].

ISSN: 2350-0328
International Journal of Advanced Research in Science, Engineering and Technology
Vol. 5, Issue 3 , March 2018

## II. APPLICATIONS

Recognition of Named Entity and technique of Information Extraction is required to solve most problems in specific research areas such as Question Answering and Summarization Systems, Retrieval of Information, Machine Translation, Video Annotation, Ontology Learning, Semantic Web Search and Bio-Informatics.

Recognition of Named Entity involves two defined tasks:

[1] The identification of proper nouns in text, and

[2] The classification of these entities into set of pre-defined categories of interests, such as person names, organizations (companies, government organizations, committees, etc.), locations (cities, countries, rivers, etc.), date and time expressions, etc [5].

## III. NAMED ENTITY

The term "Named Entity" as introduced above, was introduced in the sixth Message Understanding Conference (MUC-6). It has provided the benchmark/standard for named entity systems that perform various information extraction tasks.
In MUC-6, Named Entities (NEs) were categorized into three types of labels, each of which uses specific attribute for a particular entity type.

Entities and their labels were defined as follows:

1. ENAMEX: Person, Location, Organization.

2. TIMEX: Date, Time.

3. NUMEX: Money, Percentage, Quantity.

**Example:**
For example, in sentence "Samriddhi and Vijeta lives in Mumbai and work at Accenture."
"Samriddhi" and "Vijeta" are names of Person, "Mumbai" is a Location, and "Accenture" is an Organization, which is recognized by the Named Entity Recognition Systems.

## IV. CHALLENGES FACED IN NAMED ENTITY RECOGNITION

For us humans, it is fairly simple to recognise Entities because mostly all the entities are proper nouns and they have initial character as a capital letter and it is easy to recognise them, but for a machine (Computer) it is hard to grasp that. Some might think that why not use the dictionaries for classifying these entities because most of them are proper nouns, but this is a wrong opinion since as time passes new proper nouns are created continuously so it is hard to keep a static dictionary for the same. [2]

Therefore, it is impossible to add all those proper nouns to a dictionary. Even though some named entities are registered in the dictionary, it is not easy to decide their senses. Most problems in named entity recognition are that they have semantic (sense) ambiguity; on the other hand, a proper noun has different senses according to the context.

For example, "Samriddhi visited Obama at White house", here Samriddhi and Obama are name of Person and White House is a Location, but in "White House announced the list of minister candidates", White House is an Organization. Similarly, Ambiguity arises in the entity "April", it can be a name of a Person and it can also be a name of Month.

ISSN: 2350-0328
International Journal of Advanced Research in Science, Engineering and Technology
Vol. 5, Issue 3 , March 2018

## V. LEARNING METHODS OF NAMED ENTITY RECOGNITION

There are three main method of learning Named Entity:

1. Supervised Learning (SL).

2. Semi-Supervised Learning (SSL).

3. Un-Supervised Learning (UL).

The main concern of Supervised Learning is the requirement of a large annotated corpus which is its limitation. The unavailability of such resources and the prohibitive cost of creating them lead to two other alternative Learning Methods.

### A. Supervised learning:

The concept of Supervised Learning is to gain insight on the features of positive and negative examples of named entity over a large collection of elucidated documents and design rules that capture instances of a given type. The ongoing dominant technique for addressing the recognition of named entity problem is supervised learning.

Supervised methods are combined class of algorithm/pseudo code that understands a model by learning from the annotated training examples. Among all the methods of supervised learning for recognition of named entity, significant work has been done using Hidden Markov Model [1] (HMM), Decision Tress, Maximum Entropy Model (MEM), and Support Vector Machine (SVM).

Typically, supervised methods either understand disambiguation rules based on discriminative features or they try to understand the parameter of assumed distribution that maximizes the likelihood of training data.
The performance of the system depends on the baseline to be transferred to the vocabulary, with the percentage of words that appear without repetition, both in training and test corpus.

### B. Semi-supervised learning:

The term "semi-supervision" (weak supervision) is still rather young. The main technique of Semi-Supervision Learning is called "bootstrapping" which includes a small measure of control, like a row of seeds, for the initialization of the learning process.
Semi supervised learning algorithms use both labelled and unlabelled corpus to create their own hypothesis. Algorithms typically start with small amount of seed data set and create more hypotheses' using large amount of unlabelled corpus.

Semi-supervised learning (SSL) is halfway between supervised and unsupervised learning. In addition to unlabelled data, the algorithm is provided with some supervision information – but not necessarily for all examples. Often, this information standard setting will be the targets associated with some of the examples.
Among the learning methods of semi supervision for recognition of named entity, a lot of work has been done using bootstrapping method.

### C. Un-supervised learning:

Unsupervised learning studies how systems can learn to represent particular input patterns in a way that reflects the statistical structure of the overall collection of input patterns.

The best example of unsupervised learning is clustering. For example, one can try to collect names from clustered groups based on the similarity of context. There are other methods also unattended. Basically, the techniques based on lexical resources (e.g. WordNet), calculated on lexical patterns and statistics on a large unannotated corpus.

ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 5, Issue 3 , March 2018

A major limitation of supervised setting is requirement of specifying large number of features. For learning a good model, a robust set of features and large elucidated corpus is needed. Many languages don't have large elucidated corpus available at their disposal. To deal with lack of annotated text across domains and languages, unsupervised techniques for Recognition of Named Entity have been proposed such as Know-It-all.

## VI.METHODOLOGIES

### A. Rule based approach:

A Rule-Based Recognition of Named Entity algorithm determines the named entity by using a set of rules and a list of dictionaries that are manually pre-defined by human. The rule-based Recognition of Named Entity algorithm implement a set of rules in order to extract pattern and these rules are based on pattern base for location names, pattern base for organization name and etc [2].

The patterns are mostly formed by grammatical, syntactic and orthographic features. Along with this, a list of dictionaries is used to speed up the recognition process. However, the types of dictionaries affect the performance of the NER systems and these dictionaries normally include the list of countries, major cities, companies, common first names and titles.

The Rule based approach requires some set of linguistic rules and the gazetteer lists to develop NER system. Linguistic rules have been developed by linguistic experts and we maintain gazette lists [3].

Disadvantage:

- Lot of human effort is required to main gazetteer list and linguistic rules.
- It is a cost effective and time-consuming process.
- It is highly language dependant and has very low performance.

Rule based approach can be broadly classified as Linguistic Approach and List look up Approach.

### B. Machine/Statistical based approach:

Machine-learning Named Entity Recognition algorithm normally involves the usage of machine learning (ML) techniques and a list of dictionaries. There are two types of Machine Learning model for the Named Entity Recognition algorithms; supervised and unsupervised machine learning model. Unsupervised Named Entity Recognition does not require any training data. The objective of such method is to create the possible annotation from the data. This learning method is not popular among the Machine Learning methods as this unsupervised learning method does not produce good results without any supervised methods. Machine Learning methods are applicable for different domain-specific Named Entity Recognition systems but it requires a large collection of annotated data.
Hence, this might require high time-complexity to pre-process the annotate data [2].

Statistical based approach is also called as machine learning based approach. Machine learning is a way to automatically learn to recognize patterns from the given data and apply them at all the provided situations. For this, a central and vast training set of data is built which is an essential input to learning based approach. This data often takes the form of annotations that are labelled instances of named entities, created by domain experts in a document annotation process. In machine learning, such annotated data are often called labelled data, which are often used to train an extraction model; on the other hand, the data without annotations are called test data [3].

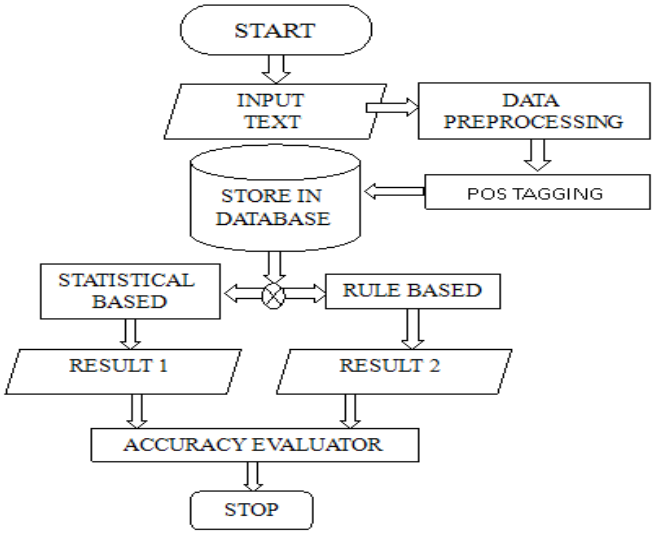
In Named Entity Recognition, the target text objects are tokens (e.g., words) or sequences of tokens for identification and classification. Features are used to create a multidimensional representation of the text objects, which can then be used by learning algorithms for generalization in order to derive patterns that can extract similar data and distinguish

ISSN: 2350-0328
International Journal of Advanced Research in Science, Engineering and Technology
Vol. 5, Issue 3 , March 2018

positive from negative examples. For this analysis, Message Understanding Conference (MUC) data set is used as training dataset.

## VII. DESIGN DETAILS

We will be finding which approach of 'Named Entity Recognition' is more accurate. For this we first take a sentence, a couple of sentences or a paragraph in English. Then we apply POS Tagger on it so that it recognizes all parts of speech and then focuses only on the noun of the sentences. These sentences with their POS tags are stored. Both the algorithms are then applied on it so that each algorithm identifies the entities and store them separately. Since we are using supervised learning method, the application already knows the correct entities and compare it with the entities recognized by the algorithms. By applying formulas of 'Precision', 'Recall' and 'F-measure'; we analyze which of the two approaches is better than the other.



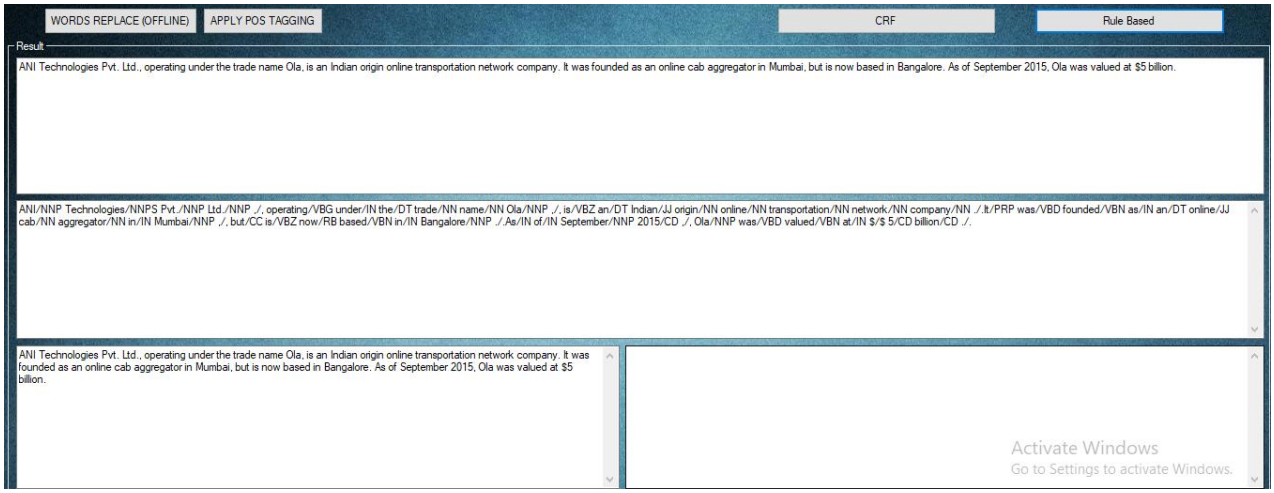

### A. PARTS OF SPEECH TAGGING:

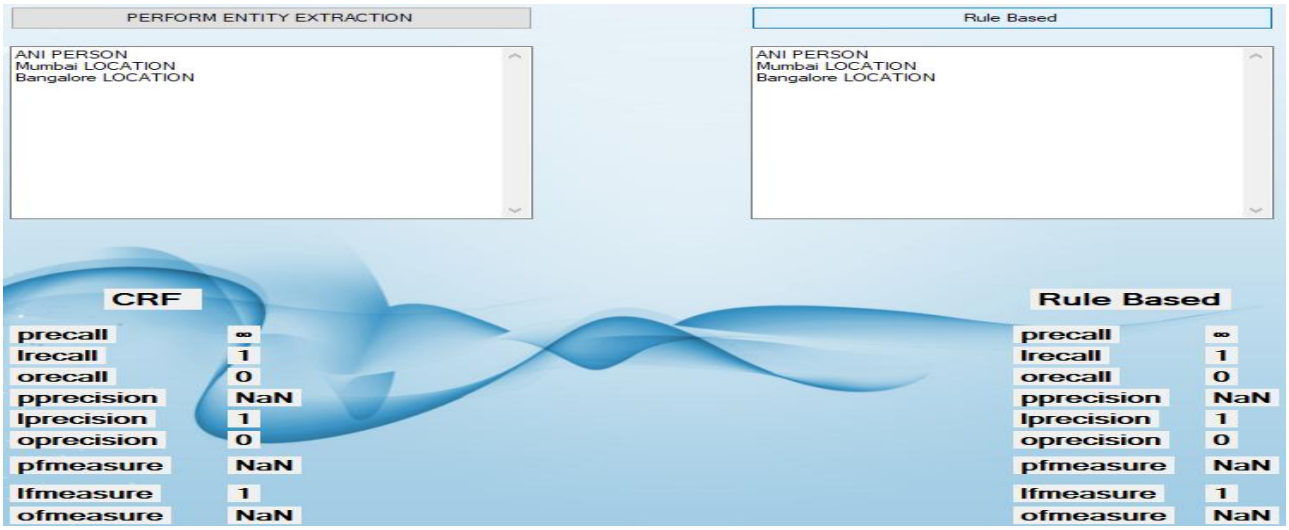
Part-of-speech tagging , in corpus linguistics, can be defined as grammatical tagging or word-category disambiguation, which is the process of denoting a word in a text (a large corpora) as corresponding to a specific part of speech, based upon both its definition and its context, its with adjacent in a phrase, sentence, or paragraph. We learnt the simplified form of this in school, in the identification of words as nouns, verbs, adjectives, adverbs, etc. In part-of-speech tagging by computer, it is difficult to distinguish from 50 to 150 separate parts of speech for English. For example, NN for singular common nouns, NNS for plural common nouns, NP for singular proper nouns.

ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 5, Issue 3 , March 2018




## VIII. EVALUATION PARAMETERS

### A. Precision:

Precision can be defined as the fraction of retrieved instances that are relevant. It is the measure of how much of the information the system returned is correct (accurate) [3] [4].

Precision (P) = No. of correct answers given by system
Total No. of answers given by system

### B. Recall:

Recall can be defined as the fraction of relevant instances that are retrieved. It is the measure of how much relevant information the system has extracted (coverage of system) [3] [4].

Recall (R) = No. of correct answers given by system
Total No. of possible correct answers in text

ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 5, Issue 3 , March 2018

### C. F-measure:

F-Measure can be defined as the harmonic average of precision and recall [3] [4]. These two measures of performance combine to form one measure of performance, the F-measure, which is calculated by the uniformly weighted harmonic mean of precision and recall:

$$\text{F-Measure (F)} = \frac{2*R*P}{(R+P)}$$

## IX. CONCLUSION

In this paper, we analysed different approaches that are used for named entity recognition such as rule-based approach and statistical based approach as well as their accuracy based on the evaluation parameters such as recall, precision, f-measure in supervised learning in natural language processing. We also studied the applications of named entity recognition and few challenges faced by it.

We can conclude from our analysis that Statistical Approach is much accurate and precise than the Rule Based Approach.

## REFERENCES

1. Nita V. Patil, Ajay S. Patil, B.V. Pawar, "HMM based Named Entity Recognition for Inflectional Language", in 2017 International Conference on computer, communication and electronics, July 01-02, 2017.
2. Dr. M. Humera Khanam, Md. A. Khudus, Prof M.S. Prasad Babu, "Named Entity Recognition using machine learning techniques for telgu language", 2016.
3. Gowri Prasad, Fousiya KK, "Named Entity Recognition Approaches", in International Conference on circuit, power, and computing technologies, 2015.
4. K.U. Senevirathne, N.S. Attanayake, "Conditional Random Fields based named entity recognition for sinhala", in IEEE 10th International conference on Industrial and information systems, ICIIS 2015, Dec. 1820, 2015, Sri Lanka.
5. Siham Boulaknadel, Meryem Talha, Driss Aboutajdine, "Amazighe named entity recognition using a rule-based approach", 2014.