Asian Journal of Convergence in Technology Volume IV Issue I
ISSN NO: 2350-1146 I.F-5.11

Survey of two different Approaches for Named Entity Recognition.

Based on Natural Language Processing.

Prof. Sarita Rathod, Samriddhi Jain, and Vijeta Shah *Department of Information Technoogy, K. J. Somaiya Institute of Engineering and Information Technology, Mumbai, India*
vijeta.shah@somaiya.edu

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

This Project does a comparative study of these two approaches on various types of inputs on the named entities like name of person, organization, and location and analyzes the outcome on the basis of parameters such as 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, ANNIE, CRF, Recall, Precision, F-Measure.

I. INTRODUCTION

A. Natural language processing:

Natural Language Processing [NLP] is a field of computer science, artificial intelligence and computational linguistics concerned with the interactions between computer and human (natural language), and, in particular, concerned with programming computers to fruitfully process large corpora.

The ultimate goal 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:

The term "Named Entity", which was first introduced by Grishman and Sundheim, is widely used in Natural Language Processing.

Named entity recognition is a sub-task of information extraction that seeks to locate 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. The researchers were focusing on extracting structured information from the unstructured text like newspaper articles.

Not only is named entity recognition a subtask of information extraction, but it also plays a vital role in reference resolution, other 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].

II. APPLICATIONS

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

Named Entity Recognition involves two tasks, which is firstly the identification of proper nouns in text, and secondly 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" was introduced in the sixth Message Understanding Conference (MUC-6). In fact, the MUC conferences were the events that have contributed in a decisive way to the research of this area. It has provided the benchmark for named entity systems that perform a variety of 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 humans, Named Entity Recognition is intuitively simple, because many named entities are proper nouns and most of them have initial capital letters and can easily be recognized that way, but for machine (Computer), it is so hard.

One might think the named entities can be classified easily using dictionaries, because most of named entities are proper nouns, but this is a wrong opinion. As time passes, new proper nouns are created continuously [2].

Therefore, it is impossible to add all those entities to a dictionary. Even though 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, "Sam visited Bush at White house", here Sam and Bush 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 "June", it can be a name of a Person and it can also be a name of Month.

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

A. Supervised learning:

The idea of Supervised Learning is to study the features of positive and negative examples of named entity over a large collection of annotated documents and design rules that capture instances of a given type. The current dominant technique for addressing the named entity recognition problem is supervised learning.

Supervised methods are class of algorithm that learns a model by looking at annotated training examples. Among the supervised learning methods for named entity recognition, considerable 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 learn disambiguation rules based on discriminative features or try to learn the parameter of assumed distribution that maximizes the likelihood of training data.

A baseline SL method, which is often proposed, consists of tagging words of a test corpus, if they are annotated as entities in the 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 relatively young. The main Semi-Supervised Learning technology is called "bootstrapping" and includes a small measure of control, like a row of seeds, for the beginning of the learning process.

For example, a system aimed at "disease names" could prompt the user to give a small number of example names. Then the system looks for sentences that contain these names, and tries to identify some clues from the context of five common examples. Then the system tries to other cases of the disease names that appear to be found in similar contexts. The learning curve is then reapplied to the newly found examples, you discover relevant new contexts. By repeating this process, a large number of disease names and a variety of contexts will eventually be obtained. Recent experiments in semi-supervised Named Entity Recognition report that rival performances Baseline monitoring approaches.

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.

Among the semi-supervised learning methods for named entity recognition, considerable work has been done using bootstrapping method.

C. Un-supervised learning:

The typical approach to 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.

A major problem with supervised setting is requirement of specifying 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 Named Entity Recognition have been proposed such as Know-It-all.

VI. METHODOLOGIES

A. Rule based approach:

A Rule-Based Named Entity Recognition algorithm detects the named entity by using a set of rules and a list of dictionaries that are manually pre-defined by human. The rule-based Named Entity Recognition algorithm applies 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 made up from grammatical, syntactic and orthographic features. In addition to that, 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. 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:

A 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 preprocess 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 take the form of annotations that are labeled instances of named entities, created by domain experts in a document annotation process. In machine learning, such annotated data are often called labeled 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 positive from negative examples. For this analysis, Message Understanding Conference (MUC) data set is used as training dataset.

VII. SYSTEMS BASED ON TWO DIFFERENT APPROACHES

A. Gate:

GATE Is an Acronym for General Architecture for Text Engineering, developed by The University of Sheffield, is a framework for text analysis developed in JAVA, available as open-source software.

GATE is an infrastructure for developing and deploying software components that process human language. It is nearly 15 years old and is in active use for all types of computational

task involving human language. GATE excels at text analysis of all shapes and sizes.

GATE is open source free software; users can obtain free support from the user and developer community via GATE.ac.uk or on a commercial basis from our industrial partners [5].

GATE is not only a framework for text engineering, but also is architecture and a development environment. ANNIE, a Nearly-New Information Extraction System that is distributed with an Information Extraction system by GATE. Named Entity Recognition is one of function in ANNIE. These recourses can be used as one unit or used as individual components along with others.

Annie:

Annie, A Nearly New Information System is an entity extraction module incorporated in the GATE framework.

It is open-source and under a GNU license, developed at the University of Sheffield. It is implemented in Java and incorporates in the form of plug-ins and libraries its own or external resources for a variety of aspects related to natural language processing (i.e. Lucene, MinorThird, Google, Weka etc.). It can be used as an API but it also provides its own interface for an independent use.

Annie also offers as a module a set of default resources (i.e. tokenizer, sentence splitter, POS tagging, co-reference resolution, gazetteers, etc.) that can be used in combination for the capture of entities. This set can be substituted by other plugins or even be disabled. The evaluation of the tool has been realized using its default resources, which are adapted for the English language.

B. Conditional random fields:

As Conditional Random Fields (CRFs) are a class of statistical modelling method often applied in pattern recognition and machine learning, they are used for structured prediction. Whereas an ordinary classifier predicts a label for a single sample without regard to “neighboring" samples, a Conditional Random Fields can take context into account [4].

The model uses sequence modelling algorithms which are probabilistic in nature. Sequence labeling is a type of pattern recognition task that involves the algorithmic assignment of a categorical label to each member of a sequence of observed values. A common example of a sequence labeling task is part of speech tagging, which seeks to assign a part of speech to each word in an input sentence or document. Sequence labeling can be treated as a set of independent classification tasks, one per member of the sequence.

Because of the strong ability to integrate any kind of features which plays an important role during training, CRFs becomes one of the key factors affecting the Named Entity Recognition performance.

The features of CNER include not only the internal features from context, such as character information, Parts-Of-Speech and boundary, but also the external features based on the statistical results such as surname that the prefix of family names, the suffix of location and organization and so on. In addition, the feature template is also found to play an important role in Conditional Random Fields based Named Entity Recognition.

VIII. EVALUATION PARAMETERS

A. Precision:

Precision is the fraction of retrieved instances that are relevant.

Measure of how much of the information the system returned is correct (accuracy) [3] [4].

$$\text{Precision (P)} = \frac{\text{No. of correct answers given by system}}{\text{Total No. of answers given by system}}$$

B. Recall:

Recall is the fraction of relevant instances that are retrieved.

Measure of how much relevant information the system has extracted (coverage of system) [3] [4].

$$\text{Recall (R)} = \frac{\text{No. of correct answers given by system}}{\text{Total No. of possible correct answers in text}}$$

C. F-measure:

F-Measure is 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 computed by the uniformly weighted harmonic mean of precision and recall:

$$\text{F-MEASURE (F)} = \frac{R*P}{0.5*(R+P)}$$

R = Recall.

P = Precision.

IX. CONCLUSION

In this paper, we surveyed 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.

ACKNOWLDGEMENT

The work presented in this paper is supported by Prof. Harsh Bhor from K.J. Somaiya Institute of Engineering and Information Technology.

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 18-20, 2015, Sri Lanka.

[5] Siham Boulaknadel, Meryem Talha, Driss Aboutajdine, "Amazighe named entity recognition using a rule based approach", 2014.