International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 02 | Feb 2020 www.irjet.net p-ISSN: 2395-0072

Survey on Named Entity Recognition using Syntactic Parsing for Hindi Language

Prem Thamarakshan[#1], Raj Paliwal[#2], Amit Shukla[#3], Shubhangi Chavan[#4]

[#1,2,3]*Student, Department of Information Technology, University of Mumbai, Pillai College of Engineering, New Panvel, Maharashtra, India*
[#4]*Professor, Department of Information Technology, University of Mumbai, Pillai College of Engineering, New Panvel, Maharashtra, India*

Abstract - *NLP is a branch of artificial intelligence that deals with analyzing, understanding and generating the languages that humans use naturally so as to interpret with computers using natural human languages instead of computer languages. NER is the task of identifying named entities in a given text and distinguishing them based on their entity type. This paper discusses various techniques and models that have been discovered and are used for this process. It also provides analyses on how effective are these techniques and models in the process of Named Entity Recognition (NER).*

Key Words: **Hidden Markov Model, Rule based Approach and List Look Up Approach, Joint Parsing, NER identification, POS tagging.**

1. INTRODUCTION

Natural language processing is the proficiency of a computer system or program to understand and interpret human language. It is a component of computer science, linguistics and artificial intelligence. The development of NLP applications is challenging because computers traditionally require humans to tell them in terms of programming language that is precise, unambiguous and highly structured, or through a limited number of clearly enunciated voice commands. Human spoken language, however, is not always scrupulous -- it often possesses ambiguity and the linguistic structure depends on many complex variables, including slang, regional dialects and social context.

Named entity recognition (NER) is the primary step towards information extraction in which an algorithm takes input as a string of text (sentence or paragraph) identifies relevant nouns (people, places, and organizations) that are mentioned in that string and classify named entities into pre-defined categories just like the names of persons, organizations, locations, expressions of times, quantities, monetary values, percentages, etc. NER is used in many fields in Natural Language Processing (NLP) and it can help in answering many real-world questions, such as:

- Which companies were mentioned within the news article?
- Were specified products mentioned in complaints or reviews?
- Does the tweet contain the name of a specific person?

Natural Language processing is considered to be a difficult problem in computer science. It's the nature of the human language that makes NLP complex to operate. Comprehensively understanding the human language requires understanding of both the words and how the concepts are connected to deliver the intended message. While humans can easily master a language, the ambiguity and imprecise characteristics of the natural languages are what make NLP difficult for machines to implement.

The paper presents a detail survey of various works in process of NER in the field of NLP. Related work and past literature is discussed in this section. The various techniques used in recognition of name entities, the challenges and the problems encountered are discussed in this paper.

II. LITERATURE SURVEY

The process of Named Entity Recognition consists of these stages Stopword removal, Tokenization, Assign a tag to tokenized word, search for Ambiguous word and Entity Recognition. Disambiguation is completed by analyzing the linguistic feature of the word, its preceding word, its following word, etc. Considerable work is already done for foreign languages if we look at the same scenario for South-Asian languages such as Hindi and Marathi, it find out that not much work has been done. As these languages are morphological rich language and unavailability of annotated corpora.

In 2018, **Prince Rana, Sunil Kumar Gupta, Kamlesh Dutta** [1] proposed an approach for identify the named entity where the processing unit is divided into two parts. First is the pre-processing task and second is the post processing task. Pre-processing of text includes tokenization of text followed by comparing the text with the online Hindi dictionary to check whether token is known or unknown word. If the token is unknown word then post processing will perform action on unknown word. The unknown word is compared with the implemented rules to check its identity otherwise identity

of unknown words is checked based on linked annotated corpus. They have achieved accuracy up to a certain level. The accuracy can be increased by increasing the size of the corpus and by handling the disambiguity.

Shrutika Kale and Sharvari Govilkar [2] proposed a survey on different techniques like Rule based Approach, Machine Learning Approach and Hybrid Approach form the major categories of the NLP NER algorithms. Out of following categories it had been observed that the machine learning based approach are best suited and most popular approach. In machine learning there are many sub categories of techniques such as HMM, CRF, SVM, ME. Based on different evaluation techniques and result analysis and also according to the review of their literature survey of experiments conducted across India by different researchers it had been proved that the Rule based, CRF, HMM are mostly implemented for Hindi, Marathi, Urdu, Punjabi, Bengali, Telugu. It had been observed that HMM, CRF gives the best results considering their limitations.

Shubhangi Rathod, Sharvari Govilkar [3] had presented comparison of various POS Tagging techniques for Indian regional languages had been done elaborately. They said that automatic POS tagging makes errors reason being many high frequency words of part-of-speech are ambiguous. Rule-based tagging assigns a word all possible tags and uses context rules to disambiguate statistical tagging assigns a word its most likely tag, based on the n-set values in a training corpus. Hybrid based tagging combines the two approaches.

Zhanming Jie, Aldrian Obaja Muis, Wei Lu [4] had used Dependency trees, which conveys crucial semantic-level information. In this work, they investigate on the way to better utilize the structured information conveyed by dependency trees to enhance the performance of NER. Unlike the present approaches which only exploits the dependency information for designing local features, that they had shown that certain global structured information of the dependency trees are often exploited while building NER models, where such information can provide guided learning and inference.Through extensive experiments, that they had shown that their proposed novel on dependency guided NER model performs competitively with models that supports conventional semi-Markov conditional random fields, while requiring significantly less period.

Simpal Jain and Nidhi Mishra [5] discusses a hybrid based approach for POS tagging on Hindi corpus. This paper is a review of different Techniques for Part of Speech tagging of Hindi language. The Hindi Word Net may be a rich resource, it's getting used by many Hindi Natural language processing (NLP) applications. Hindi Word Net consists of around 1 lakh unique class category of words like Noun, verb, adjective, and adverb. But still, many words are not tagged, so they had used Rule based approach to assign tags to all words, and use context rules to disambiguate stochastic based approach, they had assigned the most likely tag to a word, based on the on-set values frequency in a corpus. Hybrid based tagging, which is a combination of the two approaches. They had concluded that, Hybrid Approach provides higher accuracy, as compared to an individual rule based POS tagger and stochastic POS tagger.

Deepti Chopra, Nisheeth Joshi, Iti Mathur [6] have designed a NER system for the Hindi language using the Hidden Markov Model and got accuracy of 97%. They have explain the importance of HMM and its advantage. The accuracy is high but the size of corpus is limited and the tagset is small. One more challenging aspect of HMM is it requires lot of data for training.

Yavrajdeep Kaur, Er.Rishamjot Kaur [7] have designed a NER system by using Hybrid approach (combination of Rule Based Approach and List look Approach) for Hindi Language. The system is capable of extracting 10 named entities. Their accuracy is 95.77%. In this system they have added three new name entities that is money value, direction values and animal/bird entities. They concluded that by adding more entities and by increasing the size of the corpus accuracy can be increased.

Sachin Pawar, Nitin Ramrakhiyani, Girish K. Palshikar, Pushpak Bhattacharyy, and Swapnil Hingmire [8] over here they have used Distant Supervision framework, which is used to automatically create a large labeled data for training the sequence labeling model. The framework exploits a set of heuristic rules based on corpus statistics for the automatic labeling. Their approach puts together the benefits of heuristic rules, a large unlabeled corpus as well as supervised learning to model complex underlying characteristics of noun phrase occurrences. In comparison to simple English like chunking baseline and a publicly available Marathi Shallow Parser, their method demonstrates a better performance.

Amir Bashir Malik and Khushboo Bansal [9] have describes the problems of NER in the context of Kashmiri Language and provides relevant solutions by using noun identification algorithm and named entity recognition identification algorithm. Not much research has been done for Kashmiri language, by more research on it can improve the accuracy.

Suvarna G Kanakaraddi, V Ramaswamy [11] they have used a new parser called as fuzzy parsing which is less rigid than traditional parsing and appropriate for NLP. They have developed FLSR grammar for this parsing. The language used was C. This paper is made for English language. The input was English language they have generated permutations and on the basis of permutations FSLR algorithm was applied. Fuzzy min max

was to determine the degree of parsing. This parsing is made for partial sentence and gives partial syntactic correctness. The efficiency of result or percent is not described.

Jenny Rose Finkel and Christopher D. Manning [12] they have build a joint model of named entity recognition and parsing which is based on feature-based constituency parser. Their model produces a consistent output, where the named entity spans do not conflict with the phrasal spans of the parse tree. The joint representations allows the information from each type of annotation to improve performance on the other and they have done experiment on OntoNotes corpus and found improvements of about 1.36% absolute F1 score for parsing, and up to 9.0% F1 score for named entity recognition. They said that they would like to add other levels of annotation available in the OntoNotes corpus to their model, including word sense disambiguation and semantic role labeling.

TABLE I : OBSERVATION

Sr. no	Language	Category	Accuracy
1	Hindi	12	97%
2	Hindi	Not Defined	95.77%
3	Punjabi	Not Defined	86.98%
4	Hindi, Marathi	Not Defined	66.05%
5	Bengali	Not Defined	53.36%
6	Kashmiri	Not Defined	93%

Above table shows the accuracies of different Indian languages for NER. Apart from [6], none of them have mentioned the tagset category and most of the paper are the survey paper and does not describe the techniques used for recognition of named entities. Deepti Chopra, Nisheeth Joshi, Iti Mathur [6] has used a HMM model for NER and have good accuracy but has a limitation of dataset. Yavrajdeep Kaur, Er.Rishamjot Kaur [7] has used Hybrid approach (combination of Rule Based Approach and List look Approach) and had got good accuracy but has the same limitation of dataset for hindi.

III. EXISTING SYSTEM

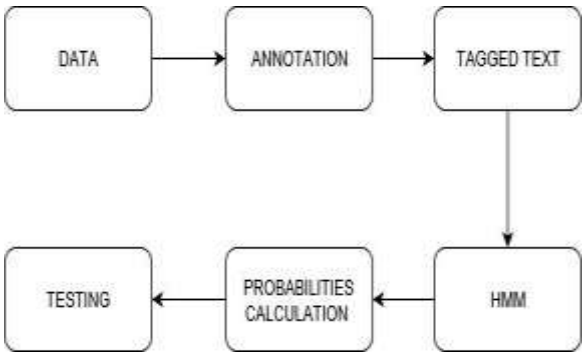

Fig :1 Existing System Architecture

In the above figure of the existing system architecture we can see the process in following steps:

1. **DATA:** The input data is provided to the system.
2. **ANNOTATION:** After preprocessing annotation is done i.e. all tokens are identified. The name entity of person, location, and organization and so on various annotations are done.
3. **TAGGED TEXT:** The annotated text is the tagged text which is used for training the model.
4. **HMM:** The tagged set is given to the Hidden Markov Model (HMM) for training.
5. **PROBABILITIES CALCULATION:** After training three probabilities are calculated (Start probability, Transition Probability and Emission probability) which are used to evaluate the model i.e. to know how good the model is.
6. **TESTING:** After training the model is tested to know how good the model is performing.

IV. PROPOSED SYSTEM

This system helps to find the NER i.e. the Named Entity Recognition of the input sentence. This helps to understand the relationship of the various terms in a sentence for example name, place, etc. This system is proposed for Indian regional languages i.e. Hindi, Marathi.

PROCESSING STEPS:

1. **INPUT:** The input data is provided to the system in the form of a document.

Example:

Input: भारतीय परंपरा के अनुसार आम आदमी

आम खाता है |

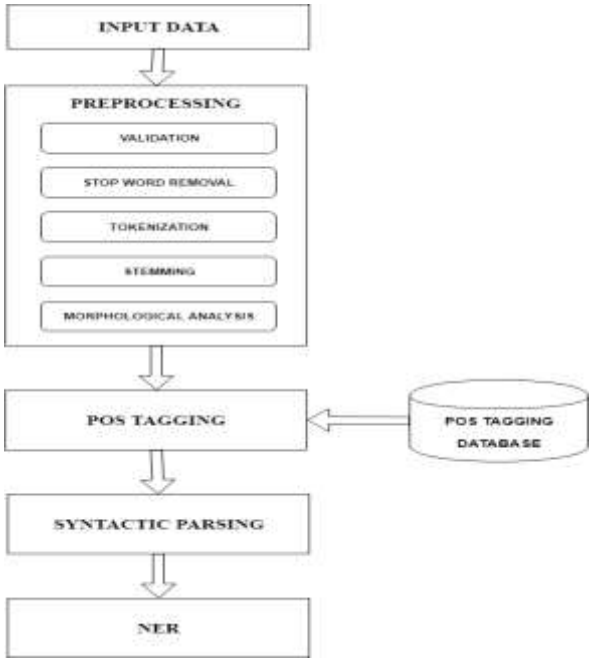

Fig. 2**:** NER using Syntactic Parsing for Hindi

2. **PRE-PROCESSING:** The input data is first processed by removing the stop words in it and carrying out tokenization, stemming and a root word is generated by the morphological analyzer.

- **VALIDATION:** Validation is to check whether the given input text is in language for which the system is implemented. It also checks whether the input is syntactically correct, but does not check the semantic correctness.

  Comparing each input document character with UTF-8.If the character is present in UTF-8 then the input is a valid script(Hindi) for the language i.e. Hindi for which the system is implemented

  Example:

  Input: भारतीय परंपरा के अनुसार आम आदमी आम खाता है |

  According to Indian tradition common man eats mango.(UTF-8,Font=Mangal)

  Output: भारतीय परंपरा के अनुसार आम आदमी आम खाता है

- **STOPWORD REMOVAL:** In stop word removal, a word that occur very frequently and does not contribute much to the context and content, and also have any impact on their existence are removed. Removing unnecessary words in the sentence such as से, को, इस, कि, जो, तो, ही, या, हो, etc. From a predefined list of stop words in the language.

  Example:

  Input: भारतीय परंपरा के अनुसार आम आदमी आम खाता है

  Output: भारतीय परंपरा अनुसार आम आदमी आम खाता

- **TOKENIZATION**: The aim of the tokenization is the exploration of the words in a Sentence where every word, symbol, special character in the sentence is considered as a token.

  After removing stop words each character left is separated as a token.

  Example:

  Input: भारतीय परंपरा के अनुसार आम आदमी आम खाता है

  Output: भारतीय | परंपरा | अनुसार | आम | आदमी | आम | खाता

- **STEMMING:** Trimming or cutting out the extraneous words to the stem is called stemming. Here inflections are removed using stemming algorithms. Here the suffixes and prefixes added to the root word is removed.

  Example:

  Input: भारतीय | परंपरा | अनुसार | आम | आदमी | आम | खाता

  Output: भारती | परंपरा | अनुसार | आम | आदमी | आम | खाता

- **MORPHOLOGICAL ANALYSIS:** Morph analysis is the procedure to find out the root word. It recognizes the inner structure of the word.

  After stemming there is a possibility of not obtaining the exact root word in such cases the morphological analysis is important. The Morphological Analysis takes place with the predefined inflection rules in the system.

  Example:

  Input: भारतीय | परंपरा | अनुसार | आम | आदमी | आम | खाता

  Output: भारत | परंपरा | अनुसार | आम | आदमी | आम | खाता

3. **POS TAGGING:** The parts of speech in the input data is identified and assigned to the word. It also removes the ambiguity in the sentence using word sense disambiguation.

   WSD will be used to remove the ambiguity in the sentences. Each token after stemming and morphing will be assigned with its POS in the sentence.

   Example:

   Input: भारत | परंपरा | अनुसार | आम | आदमी | आम | खाता

   Output: भारत→NNP (Proper Noun)

   परंपरा→ ABN (Abstract Noun)

   अनुसार →CC (Conjunction)

   आम→ JJ (Adjective)

   आदमी→ NN (Common Noun)

   आम→ NN (Common Noun)

   खाता→ VM (Verb)

4. **SYNTACTIC PARSING:** Syntactic parsing is the task of recognizing a sentence and assigning a syntactic structure to it. In this case the input data will be parsed and a parse tree based on the Entity and relationships will be generated. A parse tree will be generated after POS tagging in which we can see the entity with its relation in the sentence.

   Example:

<Sentence id="1">

1	((	NP	<fs af='परंपरा,n,f,sg,3,d,0_का_अनुसार,0' vpos="vib2_3_4" head="परंपरा">
1.1	भारतीय	JJ	<fs af='भारतीय,adj,any,any,,any,,'>
1.2	परंपरा	NN	<fs af='परंपरा,n,f,sg,3,d,0,0' name="परंपरा">
	))		
2	((	NP	<fs af='आदमी,n,m,sg,3,d,0,0' head="आदमी">
2.1	आम	JJ	<fs af='आम,adj,any,any,,any,,'>
2.2	आदमी	NN	<fs af='आदमी,n,m,sg,3,d,0,0' name="आदमी">
	))		
3	((	JJP	<fs af='आम,adj,any,any,,any,,' head="आम">
3.1	आम	JJ	<fs af='आम,adj,any,any,,any,,' name="आम">
	))		
4	((	VGF	<fs af='खा,v,m,sg,2,,ता_है,wA' vpos="tam1_2" head="खाता">
4.1	खाता	VM	<fs af='खा,v,m,sg,any,,ता,wA' name="खाता">
	))		

</Sentence>

5. **NER:** This parse tree constructed will be used to classify these entities which consist of proper nouns like person name, location names, temporal entities, etc. Clusters of different types of entities based on Name, Place, etc will be formed.

Example:

भारतीय परंपरा के अनुसार आम आदमी आम खाता है |

भारत → देश(Country)

आम → फल(Fruit)

V. CONCLUSION

The models discussed in this paper is sufficient for handling named entities in text but still more accuracy is required. The previous models have achieved accuracy up to a certain level; these can be increased by increasing the size of the corpus and also the rules used for finding named entities in the text. Accuracy can also be increased by handling ambiguous entities in the present text. Also considering the ambiguous nature of Indian regional languages the complete automation of named entities is still a very difficult task to achieve is what these references conclude.

ACKNOWLEDGMENT

We would like to show our gratitude to each and every one part of the project and the project guide for sharing their proficiency with us during the course of this research, and we thank other faculty for their so-called insights. We are also immensely grateful to the authors of the referred references for their comments on an earlier version of the manuscript, although any errors are our own and should not tarnish the reputations of these esteemed persons. We would also like to thank the head of the Information Technology department and to the principal of Pillai College of Engineering, New Panvel for extending their support in this course of research.

REFERENCES

[1]. Named Entity Recognition (NER) for Hindi Prince Rana, Sunil Kumar Gupta,Kamlesh Dutta International Journal of Computer Sciences and Engineering Vol.-6, Issue-7, E-ISSN: 2347-2693 July 2018

[2]. Survey of Named Entity Recognition Techniques for various Indian Regional Languages Shrutika Kale and Sharvari Govilkar International Journal of Computer Applications (0975 – 8887) Volume 164 – No 4, April 2017

[3]. Shubhangi Rathod et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 6 (3) , 2015, 2525-2529 Survey of various POS tagging techniques for Indian regional languages Shubhangi Rathod, Sharvari Govilkar 2017

[4]. Efficient Dependency-Guided Named Entity Recognition Zhanming Jie, Aldrian Obaja Muis, Wei Lu Singapore University of Technology and Design 2017.

[5]. Insight of various POS tagging techniques for Hindi Language Simpal Jain and Nidhi Mishra Sept 2017

[6]. Named Entity Recognition in Hindi Using Hidden Markov Model Deepti Chopra, Nisheeth Joshi, Iti Mathur Second International Conference on Computational Intelligence & Communication Technology 2016

[7]. Named Entity Recognition (NER) System for Hindi Language Using Combination of Rule Based Approach and List Look Up Approach Yavrajdeep Kaur, Er.Rishamjot Kaur International Journal of scientific research and management (IJSRM) 2015

[8]. Noun Phrase Chunking for Marathi using Distant Supervision Sachin Pawar Nitin Ramrakhiyani Girish K. Palshikar Pushpak Bhattacharyya Swapnil Hingmire August 2015.

[9]. Named Entity Recognition for Kashmiri Language using Noun Identification and NER Identification Algorithm Amir Bashir Malik and Khushboo Bansal International Journal of Computer Sciences and Engineering Volume-3, Issue-9 E-ISSN: 2347-2693 2015.

[10]. Study of Named Entity Recognition for Indian Languages Hinal Shah, Prachi Bhandari, Krunal Mistry, Shivani Thakor, Mishika Patel and Kamini Ahir 2015

[11]. Natural Language Parsing using Fuzzy Simple LR (FSLR) Parser Suvarna G Kanakaraddi, V Ramaswamy IEEE International Advance Computing Conference (IACC) 2014.

[12]. Joint Parsing and Named Entity Recognition Jenny Rose Finkel and Christopher D. Manning Computer Science Department Stanford University 2009