IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-6, Issue-7, July 2018 E-ISSN: 2347-2693

Named Entity Recognition (NER) for Hindi

[1*]Prince Rana, [2]Sunil Kumar Gupta, [3]Kamlesh Dutta

[1]Research Scholar, IKGPTU, Kapurthala, India
[2]Department of Computer Science & Engineering, BCET, Gurdaspur, India
[3]Department of Computer Science & Engineering, NIT, Hamirpur, India

[*]*Corresponding Author: erprincerana15@gmail.com*

Available online at: www.ijcseonline.org

Accepted: 12/Jul/2018, Published: 31/Jul/2018

***Abstract*:** In this paper, we present the issues and technique for recognition of named entities present in Hindi language text. Here, we discussed the categorization of Unknown word as named entity. Name entities includes person names, city names, email, web addresses etc. The main problem while identifying these words is that its meaning is not present in the dictionary. Our focus revolve around a hybrid approach consists of two sub-approaches such as corpus based and rule based hybrid approach. Experimental results have been shown to measure the accuracy of the system.

***Keywords*:** Named Entity Recognition, Unknown Words, Annotated Corpus, Tokenization

I. INTRODUCTION

Task of Natural Language Processing is to process the natural language of a human being so that it can be beneficial for those who don't or have a little knowledge of the particular language. It processes the language in such a manner that there will be minimal errors. Otherwise it will be of no use for the person who wants to study that language. Major application of natural language processing is machine translation, parsing, sentiment analysis, speech recognition and information retrieval system.

A paragraph consists of sentences and sentences consist of words. All the words in a sentence have their own structure and syntax. They can be noun, pronoun, verb, adverb, adjective, preposition etc. This distribution is same for maximum languages. Few distributions from the above are causes ambiguity during processing. E.g. while translating a Hindi text into English, Names (city names, person names, organization names, dates, email id, website name, and locations) causes' ambiguity, because these names are not present in the Hindi Lexicon. Other reason for ambiguity is that there is no small case or upper case letters in Hindi. Every line starts with a word which does not shows that it is either a lower case or upper case. In English it is somewhat easy to identify these named entities because names are starting from capital letters. Recognition of all types of words is important while processing a language. Named entity recognition plays an important role in natural language processing. For achieving better accuracy these named entities has to be handles properly. Our emphasis is to discuss the solution to increase the efficiency of the system by handling these named entities.

In this paper section 2 we have talk about the related work carries for handling named entities. Section 3 tells about the proposed work to be done for handling named entities in Hindi text. Section 4 describes the evaluation and results. Section 5 shows the conclusion and future scope.

II. RELATED WORK

Information extraction is important part of language processing. These extracted pieces are relevant to specific applications.

In this paper author discusses the hybrid approach for identifying the named entities from the Hindi language text. They discussed look up and rule base approach and achieved 96% of accuracy. [12] In this author discusses literature survey on machine translation and also discuss the various part of speech tags. They also discuss why the proper information extraction is important. [9] In this author discuss the work done to identify named entities for Indian languages. They also show a comparative analysis of named entity recognition. [2] In this paper author presents a step wise mining for detecting multi word expressions from the Hindi text. According to them many of multiword are used in our daily life but they are considering that it is frequently used in formal communication. [17] Author presents a study of kinds and structure of multi word expressions. They have also highlighted the methodologies to extract multi word expressions from parallel English Hindi corpus. [11] In this paper author classify words into the predefined categories

such as location, person-name, organization, date, time etc. They have used machine learning and rule based approach to classify the words. They have also experiment the same with the Conditional Random Fields and Maximum Entropy. Voting method was implemented to improve the performance of the NER system. [10] In this paper author discussed about talked named MANWAI. They have reported 0.792 error rate in Hindi-English and English-Hindi language pair. Their main innovations are named entity recognizer and multi word extractor. Their system has improved the accuracy rate of machine translation system. [16] In this author develop a system to categorize the unknown words. Multi component architecture was developed to create a decision tree for each class of unknown words. They have also showed the previous experimental result on same data. [15] In this paper author present a survey on various approaches (like rule based, statistics based or a combination of both) used identification of Names Entities (NE) in Indian Languages. They have discussed various application areas where named entity recognition is very important. [7] In this paper author applied hybrid approach for Named Entity Recognition (NER) in Manipuri language. Statistical approach (Conditional Random Field, CRF) and Rule-based approach is used and they have achieved a F-score of 92.26%, 94.27% and 93.3% respectively. [4] In this paper author has developed a tool for extracting multi word expressions from Sanskrit text. They have made a tool named "Sequence feature extractor" for this. They also tried to extract multi word groups by word sequence rules and morphological analyzer. [5]

III. THE PROPOSED MODEL

The proposed system work on recognition of unknown words present in a given Hindi text. The system works in two phases. 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. Post processing is divided into two parts.

Firstly 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. The corpus consists of city names, names of persons, organizations names etc. A rule consists of names such as well known surnames like kaur, Singh, Rana etc.

Every language has different semantics based on needs. Different approaches are used for different languages because each and every language is semantically different. Every language gives different meaning when they translate the words. Every language has also its own way of representing and writing of these words. We have used hybrid approach for handling these named entities.

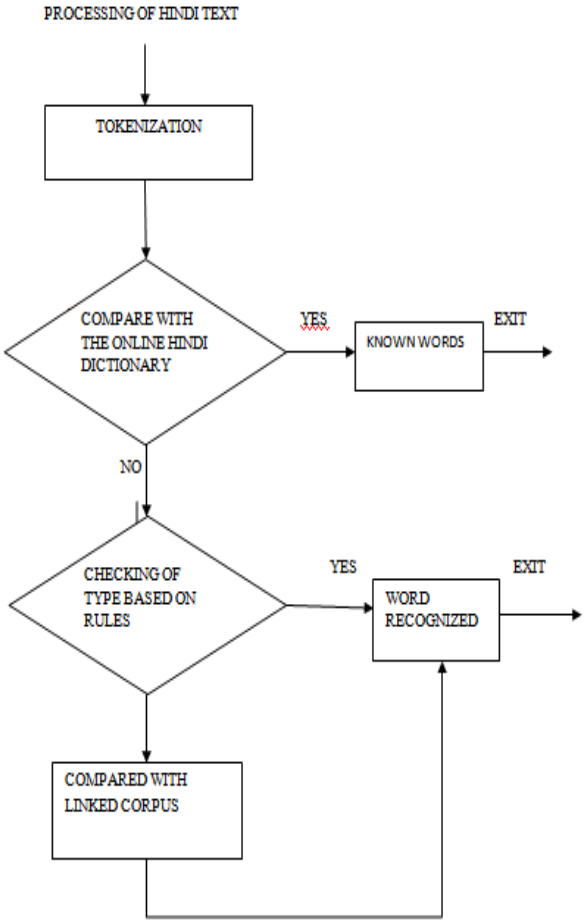

Fig 1: Processing Model

Block Diagram of Model: Block diagram of our proposed model is shown below.

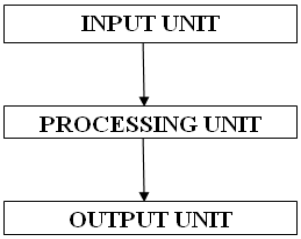

Fig 2 Block Diagram of Proposed Model

Input Unit: In this unit input is to be taken either with the help of keyboard or the file can also be imported. We have also given an on screen keyboard of Hindi text for the ease of users. We have made our input unit a user friendly. It is very easy to use for the new users. We have also used backspace, reset buttons on the model.

Processing Unit: Processing of the input is start by converting the given or input text into tokens. The process to

convert input text into tokens is called tokenization. For this we use online Hindi dictionary to check the known words. Means those words that have a meaning in dictionary. Others are considered as unknown words.
From these unknown words our model identifies the named entities according to the rules. If fail to identify with the rules then annotated corpus is used to match the word.
Output Unit: This unit basically displays the result of the processed input. We can also rest the input and output unit.

IV. EVALUATION AND RESULTS

Evaluation of the model is effective only when following parameters are satisfied. These parameters are size of the corpus, performance of a model and the algorithm used. If the size of the corpus is not so huge then accuracy of the system will be reduced. If the rules for finding named entities are not properly made it also affects the system.
We have tested our proposed model on written text taken from various novels, books, newspapers and some text from online sources also. News papers like Punjab Kesari, Dainik Jagran etc. In each test we get different score because the words used in all the sources different. Total five dataset has been taken for evaluation of our tool. Each dataset consists of different number of words with different frequency of named entities.

Table 1: Experimental Result

S. No.	Number of words in a text	Named Entities in them	Found Accurate words among them	Accuracy in %age
Test 1	600	85	71	83.52
Test 2	700	70	57	82.85
Test 3	500	60	51	85
Test 4	800	50	38	84
Test 5	900	63	48	80.95

The overall average accuracy provided by our model is 83%.

V. CONCLUSION AND FUTURE WORK

Attempt has been made to maximize the performance of named entity recognizer. The model discussed in this paper is sufficient for handling named entities in a text but still more accuracy is required. We have achieved accuracy up to a certain level. We can increase this accuracy of the system 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. Complete automation of named entities is still a very difficult task to achieve.

REFERENCES

[1]. Deepti Chopra, Sudha Morwal, "Named Entity Recognition in Punjabi using Hidden Markov Model", International Journal of Computer Science & Engineering Technology, Vol. 3,issue 12,pp 616-620 ,2012

[2]. Hinal Shah, Prachi Bhandari, Krunal Mistry, Shivani Thakor, Mishika Patel and Kamini Ahir "Study Of Named Entity Recognition For Indian Languages" International Journal of Information Sciences and Techniques vol 6, issue 1/2, pp.11-25,2016.

[3]. Kamaldeep Kaur , Vishal Gupta, "Name Entity Recognition for Punjabi Language", International Journal of Computer Science and Information Technology & Security, Vol. 2, No.3, pp 16051-16055, 2012.

[4]. L. Jimmy, Darvinder Kaur, "Named Entity Recognition in Manipuri: A Hybrid Approach", Springer International Publishing AG, Part of Springer Science Business Media, Vol. 8105, pp 104- 110.

[5]. Murali Nandi, Ramasree R.J. " Rule-based Extraction of Multi-Word Expressions for Elementary Sanskrit Texts" International Journal of Advanced Research in Computer Science and Software Engineering, vol 3, issue 11, pp.661-667, 2013.

[6]. Navneet Kaur Aulakh, Yadwinder Kaur, "Optimized name entity recognition of machine translation", International Journal for Research In applied science and Engineering Technolo gy, Vol. 2, issue 6, pp24-30,2014.

[7]. Prakash Hiremath, Shambhavi B. R "Approaches to Named Entity Recognition in Indian Languages: A Study" International Journal of Engineering and Advanced Technology, vol 3, issue 6, pp.191-194, 2014.

[8]. Rakhi Joon, Archana Singhal "Analysis Of MWE In Hindi Text Using Nltk" International Journal on Natural Language Computing, vol 6, issue 1, pp.13-22,2017.

[9]. Shachi Mall, Umesh Chandra Jaiswal " Survey: Machine Translation for Indian Language" International Journal of Applied Engineering Research, vol 13, issue 1, pp.202-209,2018

[10]. Shilpi Srivastava, Mukund Sanglikar, D.C Kothari "Named Entity Recognition System for Hindi Language: A Hybrid Approach" International Journal of Computational Linguistics vol 2, issue 1, pp.10-23, 2011.

[11]. Vivek Dubey , Pankaj Raghuwanshi , Sapna Vyas "Impact of Multiword Expression in English Hindi Language" International Journal of Emerging Trends & Technology in Computer Science, vol 4, issue 3, pp.101-105,2015

[12]. Yavrajdeep Kaur, Rishamjot Kaur "Named Entity Recognition (NER) system for Hindi Language Using combination of Rule Based Approach and List Look up Approach" International journal of scientific research and management, vol 3, issue 3, pp.2300-2306,2015.

[13]. Amit Goyal "Named Entity Recognition for South Asian Languages" Proceedings of the IJCNLP-08 Workshop on NER for South and South East Asian Languages pp.89-96, 2008.

[14]. Anil Kumar Singh" Extraction and Translation of Multi Word Number Expressions" Proceedings of the 3rd Indian International Conference on Artificial Intelligence, 2007

[15]. Janine Toole "Categorizing unknown words: using decision tree to identify Names and Misspellings" Annual conference of North America,2000

[16]. Liling Tan, Santanu Pal" Manawi: Using Multi-Word Expressions and Named Entities to Improve Machine Translation" Proceedings of the Ninth Workshop on Statistical Machine Translation pp.201-206, 2014.

[17]. Rai Mahesh Kumar Sinha, "Stepwise Mining of Multi-Word Expressions in Hindi" Proceedings of the Workshop on

Multiword Expressions: from Parsing and Generation to the Real World (MWE 2011), pp 110–115, 2011.

[18]. Rajesh Sharma & Vishal Goyal, "Name Entity Recognition Systems for Hindi using CRF approach", International Conference on Information Systems for Indian Languages , pp 31-35 2011.

AUTHORS PROFILE

Prince Rana is a research scholar of IKGPTU. He did his B. Tech and M. Tech from Punjab technical University. He has total 5 international and one national publication.

Dr. Sunil Kumar Gupta did B.E. in Computer Science from Gorakhpur University, Gorakhpur, India in 1988, and M.S. in 1991 and completed Ph.D. in Computer Science from Kurukshetra University, Kurukshetra, India. He possesses 28 years of teaching experience

He has worked as teaching faculty in many reputed institutions in India including N.I.T., Hamirpur (HP). Presently, he is working as Associate Professor in Computer Science & Engg. Department at Beant College of Engineering and Technology, Gurdaspur (India).He has more than 40 research publications. His work is published and cited in highly reputed journals of Elsevier, Springer, Taylor and Francis and IEEE. His areas of interest include database management systems, distributed systems, cloud computing and mobile computing and security.

Dr. (Mrs.) Kamlesh Dutta, Associate Professor, Computer Science & Engineering Department, National Institute of Technology, Hamirpur (HP), India. Coordinator videoconferencing and communication, Member SUGC Member SUGC (curriculum), Member APEC, Member Committee for the revision of UG curriculum, Member, Community Service Cell, Member, Campus video networking committee, Member, Automation Committee Convener DUGC, CSE Department, Coordinator Project, CSE Department

Faculty-In charge CSE , SMDP-II programme, NIT Hamirpur (HP), Member Committee - Sexual Harassment of Women at Work Place, NIT Hamirpur (HP).

