iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 5 **Issue:** XII **Month of publication:** December 2017

DOI:

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor :6.887
Volume 5 Issue XII December 2017- Available at www.ijraset.com

Research Paper on Named Entity Recognition System for Punjabi Language Text

Shavi Juneja[1]
[1]*Research Scholar Punjab, India*

Abstract: ***Natural Language Processing is an area of research and application which deals with computers and explores how computers can be used to understand and manipulate natural language text to do useful things. Natural Language Processing applications are characterized to make complex interdependent decisions which require large amounts of prior knowledge. NER is a sub problem of Natural Language Processing (NLP). In the expression "Named Entity", the word "Named" means to any name which can be belong to the person, place, location, dates , city, state, country etc. Not much work has been done in NER for Indian languages in general and Punjabi in particular. Adequate corpora are not yet available in Punjabi to find the named entity. Hence it is required to develop such a tool that can help to find the named entity from a text. In this paper we are presenting a review that how to create a named entity tool .A number of language independent and dependent various features are extracted from given paragraph. The different NER features have been reviewed to identify and classify the various named entities.Named Entity Recognition (NER) system is an important area of Natural Language Processing (NLP). In the present scenario named entity recognition plays a vital role. Area of NER has been paid more and more attention in recent years as a result of the dramatic and fast increase in the volume of data base. The development of Internet not only causes an explosively growing volume of data base, but also provides people more ways to get data base and to develop more rules. The importance of an efficient technique in NER is rule based approach from the huge collection cannot be overemphasized. One approach for NER is no name entity technique. This technique is used to develop rules and to increase the accuracy in existing systems.***

Keywords: NER, Rule based Approach, List look up approach, Precision, Recall, F-measure.

I. INTRODUCTION

Named entity recognition (NER) is a technology used to recognize proper nouns or entities in text and associate them with the appropriate types. A number of various languages independent and various language dependent features are extracted for NER. Common types in NER systems are location, person name, date, address, designation etc. It is a precursor for many natural languages processing tasks and now established as a key technology to understanding low- level texts. NER is a sub application of Natural Language processing, and it is a sub problem of Information Extraction (IE) and less complex than Information Extraction. A number of various languages independent and various language dependent features are extracted for NER. Common types in NER systems are location, person name, date, address, designation etc. It is a precursor for many natural languages processing tasks and now established as a key technology to understanding low- level texts. Some NER systems are incorporated into Parts-of-Speech (POS) taggers, though there are also many stand-alone applications whereas most of NER systems are based on analyzing patterns of POS tags, they also make use of lists of typed entities like list of possible names means it involves the identification of many named entities such as person names, location names, names of organizations, monetary expressions, dates, numerical expressions or of regular expressions for particular types like address patterns. This type of NER task also known as proper name classification that involves the classification of so called named entities as people, places, products, companies, or monetary amounts.

II. RELATED WORK

In paper (1) the authors discuss about the 'Hybrid Approach'. The hybrid approach is a combination of the rule based approach and list look up approach. In rule based approach, the number of language based rules is formed and various gazetteer lists are prepared in look up approach. In list look up approach, the NER system uses gazetteer to classify words and suitable lists are created. This approach is simple, fast and language independent. It is also easy to retarget as only lists are to be created.
Certain rules are developed which doesn't give the accurate results and hence these rules need modification to achieve better results. Overall accuracy of the proposed system is 85% which can be further improved. In paper (2) the author presents a classifier-combination experimental framework for named entity recognition in which four diverse classifiers (robust linear classifier, maximum entropy, transformation-based learning, and hidden Markov model) are combined under different conditions. When no

gazetteer or other additional training resources are used, the combined system attains a performance of 91.6F on the English development data; integrating name, location and person gazetteers, and named entity systems trained on additional, more general, data reduces the F-measure error by a factor of 15 to 21% on the English data.In paper (3) author represents a review on Named Entity Recognition system. Author describes that the Named entities are phrases that represent person, location, number, time, measure, organization. According to this paper Named Entity Recognition is the task of identifying and classifying named entities into some predefine categories. This paper gives a brief introduction to Named Entity Recognition. It also summarizes various approaches for Named Entity Recognition like Hidden Markov Model, Maximum Entropy Markov Models, Conditional Random Field, Support Vector Machine, Decision Trees and Hybrid approaches. Named Entity Tag sets defined for MUC-6, CoNLL 2002 and 2003 and IJCNLP-2008 shared tasks are also discussed. Different NER features in context to identification and classification of named entities have also been reviewed.In paper (4) author discuss two named-entity recognition models which use characters and character n-grams either exclusively or as an important part of their data representation. The first model is a character-level HMM with minimal context information, and the second model is a maximum-entropy conditional markov model with substantially richer context features. Author's best model achieves an overall accuracy of 86.07% on the English test data (92.31% on the development data). This number represents a 25% error reduction over the same model without word-internal substring features.In paper (5) the author presents a system that improves the accuracy of one NLP technique, Named Entity Recognition or NER, on Twitter data i.e. done by training a recognizer specifically for this type of data. NER is the process of automatically recognizing the words are names of people, places, organizations, locations which would be very beneficial to build a system.

III.PROPOSED WORK

A. *Techniques*

The named entity recognition system describes the various types of features and rules. The accuracy of the existing system is low. The aim of the proposed work is to create various rules to improve the overall accuracy of the system and to increase the corpus size and to increase the number of rules like animal/bird name rule, direction name rule, measurement named rule, transport or vehicle named rule, monetary named rule. For the improvement of accuracy we have to improve the existing rules by adding no name entity technique in which various rules are improved. We will use the hybrid approach to implement the name entity recognition system. This hybrid approach is a combination of "List look up approach", "Rule based approach", "No name entity technique" and use linguistic various features of the Punjabi language. These approaches can be explained in brief as follows:

1) *List lookup approach:* In this approach a corpus of for the names entities of Punjabi language is formed. In this corpus various types of names , for example names of males, females, names of places , locations , rivers , various departments and posts etc. The document from which names are to be extracted is compared with the database created and names entities are found.
2) *Rule based Approach:* Handcrafted systems rely for a great deal on the human intuition of their designers who constructs a large number of rules that capture the intuitive notions that come to mind when contemplating a simple approach for recognizing named entities. For instance, in many languages it is quite common for person names to be preceded by some kind of title.
3) *No name entity technique :* No name entity technique is used which improves or modifies the existing rules. This technique analyze the various NER Systems and results are compared with the existing approaches.

B. *Performance Measurements*

The performance of NER system is calculated by using following three parameters:

1) *Precision (P):*The precision is defined as, the precision parameter is used to measure the number of correct named entities (NEs) obtained by NER system, over the total number of named entities (NEs) extracted by NER system. The precision is represented by P. The following formulae describe how precision can be calculated:P = no. of correct names generated by our system /no. of total names generated by our system
2) *Recall (R):*The recall is defined as; the recall parameter measures the no. of correct named entities obtained by NER system over the total no. of named entities in a text. Thus, recall (R) can be calculated as,R= no. of correct names generated by our system /Total no. of names present in a paragraph
3) *F-measure (F):* The F-measure represented by F and defined as; the f-measure is used to represents the harmonic mean of precision and recall i.e.,F=2RP/R+P

IV. RESULTS

The results can be evaluated by comparing between the existing system and proposed system with parameters as precision, recall and f-measure values of the existing and proposed system.The earlier NER system calculate the values of the precision, recall and f-measure value of NE class such as person, location, organisation, designation, date/ time.The total precision value of NE class is 89.98%, total recall value is 84.55 % and total f- measure value is 85.88 % which are shown following in tabular form:

NE Class	Precision (P %)	Recall (R %)	F-measure (F %)
Person	74.52	62.86	65.67
Location	91.52	92.89	91.25
Organisation	90.27	90.10	88.77
Designation	98.84	87.09	91.98
Date/Time	94.79	89.79	91.75
Total	89.98	84.55	85.88

Table1. Results of existing NEs

Above table describes the results of existing NEs. Now the proposed system generates the more accuracy as compared to existing system and develops the NE Class with more accurate values. The objective of proposed system is to increase the corpus size and to develop the corpus of various Punjabi named entities which include the names of males, females, countries, locations, states, rivers, places etc.The proposed system achieves the accuracy up to 90% and precision, recall value, f- measure of NE class achieves 92.78%, 88.42%, 90.47% accuracy as shown in following table:

NE Class	Precision (P %)	Recall (R %)	F-measure (F %)
Person	81.52	70.50	75.61
Location	93.34	94.80	94.06
Organisation	93.39	92.50	92.94
Designation	99.12	90.89	94.82
Date/Time	96.54	93.45	94.96
Total	92.78	88.42	90.47

Table2. Results of proposed NEs

A. Precision value

The precision parameter is used to measure the number of correct named entities (NEs) obtained by NER system, over the total number of named entities (NEs) extracted by NER system. The results of precision values are expressed following in the graphical form. The graph shows an existing system has less precision value and an increased precision value shows the proposed system. The following graph represents the comparison between the existing and proposed system:

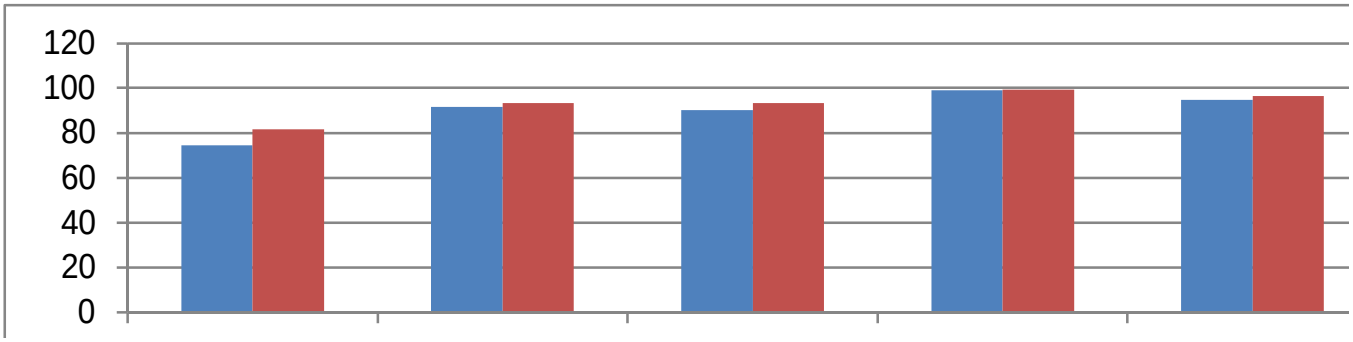


Fig 1. Graph of precision value existing system v/s proposed system

B. Recall value

The recall parameter measures the no. of correct named entities obtained by NER system over the total no. of named entities in a text and denoted as (R).The results of recall values are expressed following in the graphical form. The graph shows an existing system has less recall value and the proposed system shows the more recall value. The following graph represents the comparison between the existing and proposed system:

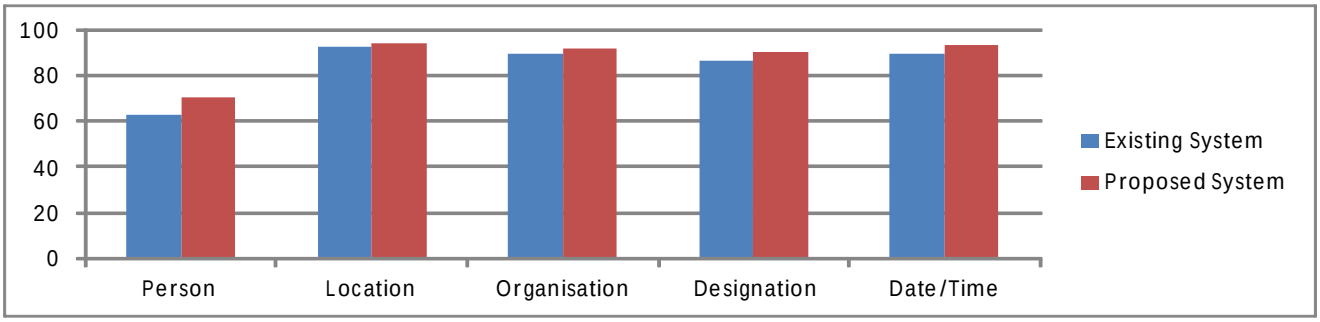


Fig.2 Graph of recall value existing system v/s proposed system

C. F- measure

The results of recall values are expressed the graphical form. The graph shows an existing system has less recall value and the proposed system shows the more recall value. The following graph represents the comparison between the existing and proposed system:

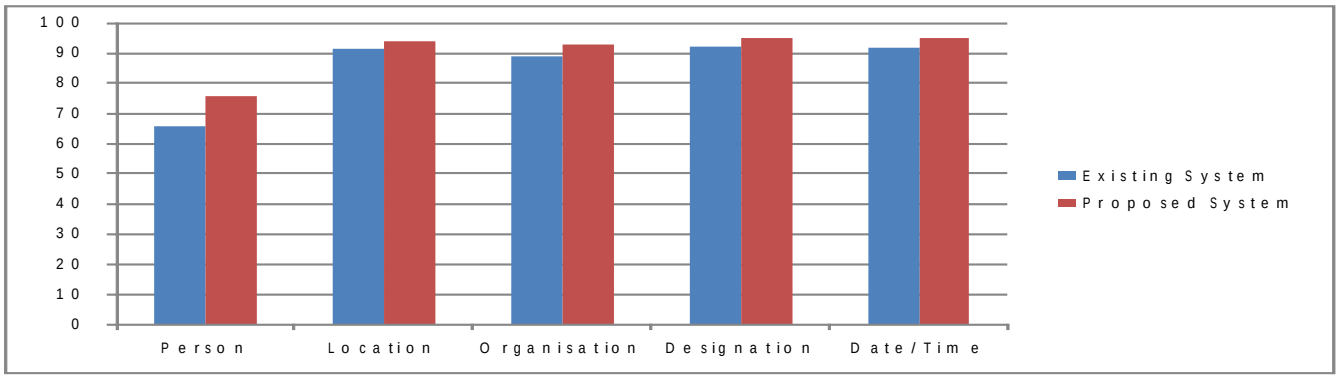


Fig. 3,Graph of f-measure value existing system v/s proposed system

D. New features

The current NER system introduces new features and adds new rules by adding no name entity technique. Through this technique, the proposed system works on new named entities as directions, monetary expressions, animals, birds, transport, measurement expressions. The following table shows the current NER system with new features:

NE Class	Existing System	Proposed System			
			Precision(P %)	Recall (R %)	F-Measure(F %)
Directions	No	Yes	86.23	77.54	81.65
Monetary Expressions	No	Yes	74.65	73.24	73.93
Vehicles	No	Yes	77.74	74.60	76.13
Measurement	No	Yes	84.38	82.62	83.49

expressions					
Animals/Birds	No	Yes	88.37	86.78	87.56

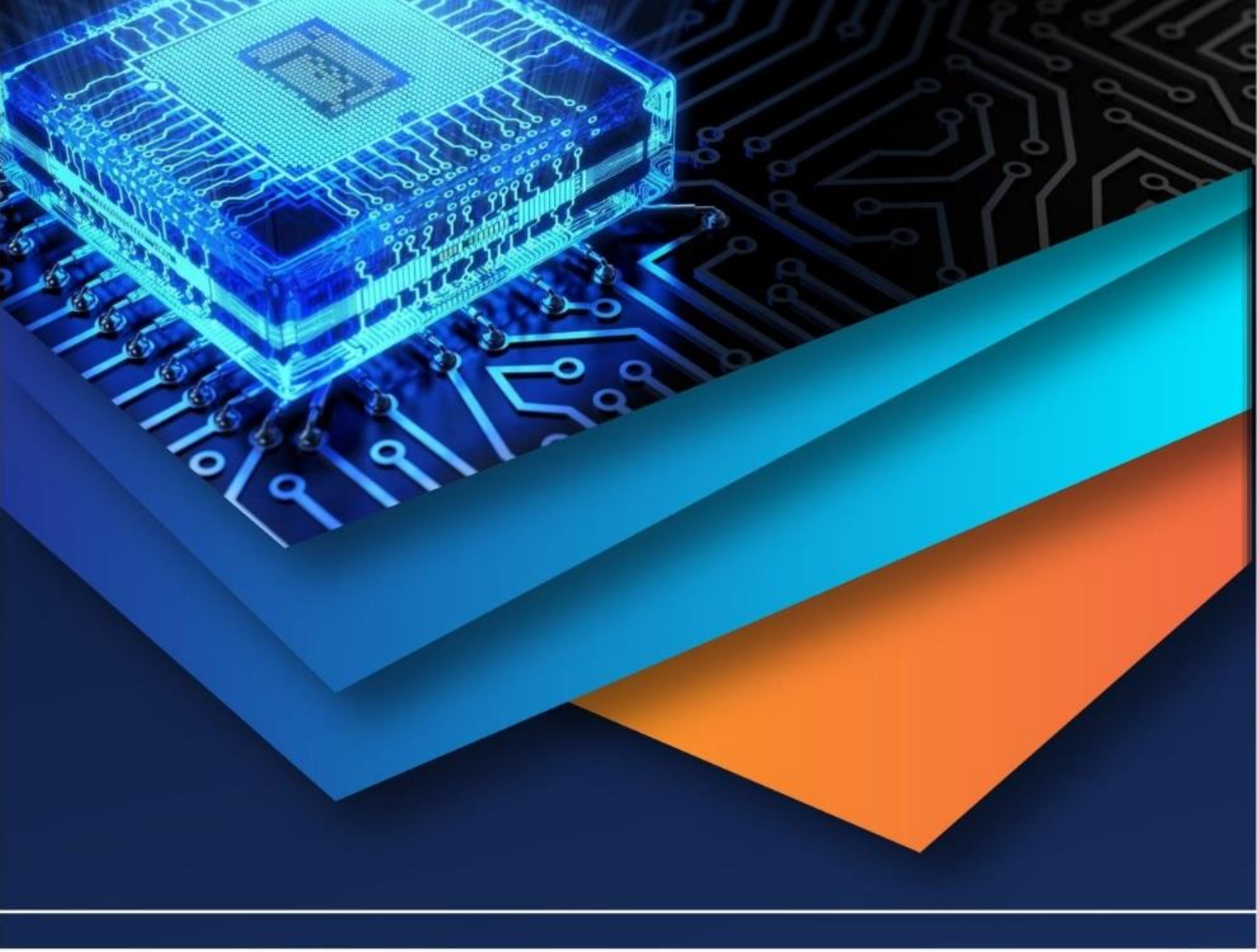
Table 3 Results of new NEs

V. CONCLUSIONS

The proposed NER system works on more new named entities but the existing system does not work in those entities as directions, monetary expressions, animals, birds, transport, measurement expressions and various rules are used to implement them. The NER system is tested against various inputs. The proposed system shown better accuracy compared to the existing systems and it also fetch some new named entities as Transport name rule, Bird/ Animal name rule, Measurement expression name rule, Direction named rule, Monetary Named rule. But the challenges faced during named entity recognition need to be solved for which more detailed study of various natural languages is required. Improved Name entity recognition is most important part of natural language. Future work can be extended to get further more accuracy and more new rules can be developed but there needs to be developing a system with efficient methods which can give more accurate result.The proposed system can't work on those documents which are extracted from multi language like English, Hindi and Punjabi. Corpus for multi languages is also required to be developed separately.

REFERENCES

[1] Kamaldeep Kaur, Vishal Gupta, "Name Entity Recognition for Punjabi Language", IRACST - International Journal of Computer Science and Information Technology & Security (IJCSITS), ISSN: 2249-9555Vol. 2, No.3, June 2012
[2] Radu Florian and Abe Ittycheriah and Hongyan Jing and Tong Zhang , " Named Entity Recognition through Classifier Combination" IBM T.J. Watson Research Center 1101 Kitchawan Rd, Yorktown Heights, NY 10598, USA.
[3] Arshdeep Singh ,Jyoti Rani ,Kuljot Singh , " Named Entity Recognition" : A Review , International Journal of Computer Science and Communication Engineering IJCSCE Special issue on "Emerging Trends in Engineering & Management" ICETE 2013.
[4] Dan Klein, Joseph Smarr, Huy Nguyen, Christopher D. Manning, "Named Entity Recognition with Character-Level Models" Computer Science Dept., Stanford University.
[5] Dr. Timothy W. Finin, William Murnane "Improving Accuracy of Named Entity Recognition on Social Media Data" Master of Science, 2010.
[6] Andrew O. Arnold "Exploiting domain and task regularities for robust named entity recognition" August 2009 CMU-ML-09-109.
[7] Artem Boldyrev, Prof. Dr. Gerhard Weikum, "Dictionary-Based Named Entity Recognition" Universitat de Saarlandes Max-Planck-Institut fur Informatik Databases and Information Systems, December 2013.
[8] Toine Bogers, "Dutch Named Entity Recognition: Optimizing Features, Algorithms, and Output" Sept 2004.
[9] Andrew Borthwick, " A Maximum Entropy Approach to Named Entity Recognition" New York University ,September, 1999.
[10] Jiafeng Guo , Gu Xu , Xueqi Cheng, Hang Li, " Named Entity Recognition in Query" Institute of Computing Technology, CAS.
[11] Lev Ratinov and Dan Roth "Design Challenges and Misconceptions in Named Entity Recognition" Computer Science Department University of Illinois Urbana, IL 61801 USA.
[12] Jamie Callan and Teruko Mitamura, "Knowledge-Based Extraction of Named Entities" School of Computer Science Carnegie Mellon University Pittsburgh, PA 15213-8213, USA.
[13] Zhenzhen Kou, William W. Cohen,(2005) "High-Recall Protein Entity Recognition Using a Dictionary", in 13th Annual International Conference on Intelligent Systems for Molecular Biology.
[14] Sujan Kumar Saha, Partha Sarathi Ghosh, Sudeshna Sarkar, "Named Entity Recognition in Hindi using Maximum Entropy and Transliteration" , Polibits, 38, pp. 33-42, 2008, Indian Institute of Technology ,Kharagpur, India.
[15] Sudeshna Sarkar, Pabitra Mitra, Sujan Kumar Saha , "A Hybrid Approach for Named Entity Recognition in Indian Languages" Proceedings of the IJCNLP-08 Workshop on NER for South and South East Asian Languages, pages 17–24,Hyderabad, India, January 2008.
[16] Arvind Neelakantan and Michael Collins , "Learning Dictionaries for Named Entity Recognition using Minimal Supervision" Department of Computer Science University of Massachusetts, Amherst Amherst, MA, 01003.
[17] Navneet Kaur Aulakh and Er.Yadwinder Kaur, " Review Paper on Name Entity Recognition of Machine Translation", International Journal of Advanced Research in Computer Science and Software Engineering , Volume 4, Issue 4, April 2014 ISSN: 2277 128X.
[18] Karthik Gali, Harshit Surana, Ashwini Vaidya, Praneeth Shishtla and Dipti Misra Sharma. 2008."Aggregating Machine Learning and Rule Based Heuristics for Named Entity Recognition" in the proceedings of the IJCNLP-08 Workshop on NER for South and South East Asian Languages, pages 25-32, Hyderabad, India.
[19] Kumar N. and Bhattacharyya Pushpak. 2006. "Named Entity Recognition in Hindi using MEMM" in the proceedings of Technical Report, IIT Bombay, India.
[20] Mandeep Singh Gill, Gurpreet Singh Lehal and Shiv Sharma Joshi, 2009. "Parts-of-Speech Tagging for Grammar Checking of Punjabi" in the Linguistics Journal Volume 4 Issue 1, pages 6-22.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY