ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 5, Issue 10, October 2017

Named Entity Recognition System for Punjabi Language Text

Shavi Juneja

Research Scholar, Dept. of CSE, GZS PTU Campus, Bathinda, 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 that 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

KEYWORDS*:* NER, Rule based Approach, List look up approach, Linguistic approach.

I. Introduction

Natural language processing is a field of computer science and linguistics deals the interactions among computers and human languages and is a very attractive method of human computer attraction . The various NLP researchers aims to gather knowledge on how human beings can understand and use language so that to develop various tools and techniques to make computer systems more understandable in a appropriate way. Further computer manipulate over the natural languages in order to perform the desired tasks.

The foundation of NLP lie in a number fields as computer and information sciences, linguistics, mathematics, electrical and electronic engineering, artificial intelligence and robotics, psychology, etc.Research in natural language processing has been going on for several decades since 1940s. The machine translation was the first computer-based application related to natural language. Natural language processing approaches fall into four categories: symbolic, statistical, connectionist, and hybrid. Symbolic and statistical approaches. The various sub problems in NLP include speech segmentation, text segmentation, part of speech tagging, word sense disambiguation, syntactic ambiguity and these are identified in italic type, within parentheses. The main role of NER is to identify expressions such as date and time, names of people, places, and organizations these expressions are difficult to extract using traditional natural language processing as they follows the open class of expressions, i.e. there is an infinite variety and new expressions are constantly being created. Automatically extracting proper names is useful to many problems such as machine translation, information retrieval, information extraction, question answering and summarization.NER has important significance in the Internet search engines and in many of the Language Engineering applications.

Named Entity Recognition (NER) is a sub problem of information extraction (IE) and is slightly less complex than IE.Named entities consists of any type of word like adverbs, prepositions, adjectives, and even some verbs, but the majority of the named entities are made up of the nouns.. The common types in NER systems are location, person name, date, address, etc. Some NER systems are incorporated into Parts-of-Speech (POS) taggers, even though there are also various stand - alone applications. NER task also called as `proper name classification' that involves the identification as well as the classification of the so-called named entities that describes the expressions that refers to people, places, organizations, products, companies, and even dates, times, or monetary amounts. This in turn means that every word needs to be categorized as belonging to a named entity or not. Or in other words we can say that not only the boundaries of these named entities determined but also the type of their named entity explained.

II. Related work

In [1] author represents a review on Named Entity Recognition system. Author describes that the Named entities 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 [2] 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 [3] author describes an 'Hybrid Approach'. The hybrid approach is an 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.

III. Proposed work

A. *Performance Parameters:*

The performance of NER system is measured by using the following 3 parameters:

- Precision (P)
- Recall (R)
- F-measure (F)

Let the total no. Of names present in paragraph are 12 and our system generates 9 names. In this 9 names , 7 names are correct and 2 are incorrect. Thus , precision , recall and F-measure can be calculated.

The precision 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.

Thus,

P = no. of correct names generated by system / no. of total names generated by system

The recall measures the no. Of correct named entities obtained by NER system over the total no. Of named entities in a text. Thus, recall can be calculated as,

R= no. of correct names generated by system / total no. of names present in a paragraph

The F-measure is used to represents the harmonic mean of precision and recall i.e.,

$$F=2RP/R+P$$

B. *Description of the Proposed System:*

The current NER system generates some more features with some more attributes are added in a proposed system explained as follows. The NE class direction tells about the directions as north, south, west, east, north - west , south - east, west- south, north- east these are extracted through the direction named rule. This rule gives us the information regarding person is travelling in which direction, any company or organization is situated in north direction or either in south direction.

The monetary expressions describes this rule is an important application of information retrieval. The information retrieval is concerned with the storing, searching and retrieving information. It is a separate field within computer science or closer to databases, but IR relies on some NLP methods (for example, stemming). This rule generates the monetary expressions like Dollar, euro, paisa, rupees etc.

IV. REPRESENTATION OF RULES

The rules generated in our proposed system are explained in the form of table. Rules such as Transport name rule, Bird/ Animal name rule, Measurement expression name rule, Direction named rule, Monetary Named rule. These rules build our proposed system.

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 expressions	No	Yes	84.38	82.62	83.49
Animals/Birds	No	Yes	88.37	86.78	87.56

Table1. New Named Entities

V. RESULTS

The precision values of the NE class are calculated and their results are compared with the existing system.
The existing system shows the less precision value and our system improves the accuracy with increase in precision value. The following table shows the precision value of existing system and proposed system of NE class:

NE Class	Existing System Precision value (P %)	Proposed System Precision value (P %)
Person	74.52	81.52
Location	91.52	93.34
Organisation	90.27	93.39
Designation	98.84	99.12
Date/Time	94.79	96.54

Table 2. Results of precision value

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:

DOI: 10.15680/IJIRCCE.2017. 0510066

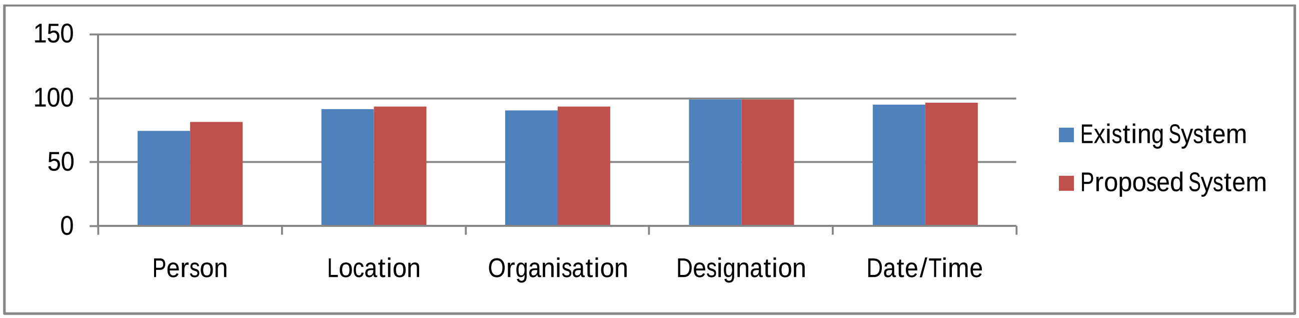

Fig1. Graph of precision value existing system v/s proposed system

The recall values of the NE class of proposed system are calculated and their results are compared with the existing system. The existing system shows the less recall value and our system improves the accuracy with increase in recall value. The following table shows the recall value of existing system and proposed system of NE class:

NE Class	Existing System Recall value (R %)	Proposed System Recall value (R %)
Person	62.86	70.50
Location	92.89	94.80
Organisation	90.10	92.50
Designation	87.09	90.89
Date/Time	89.79	93.45

Table3. Results of recall value

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:

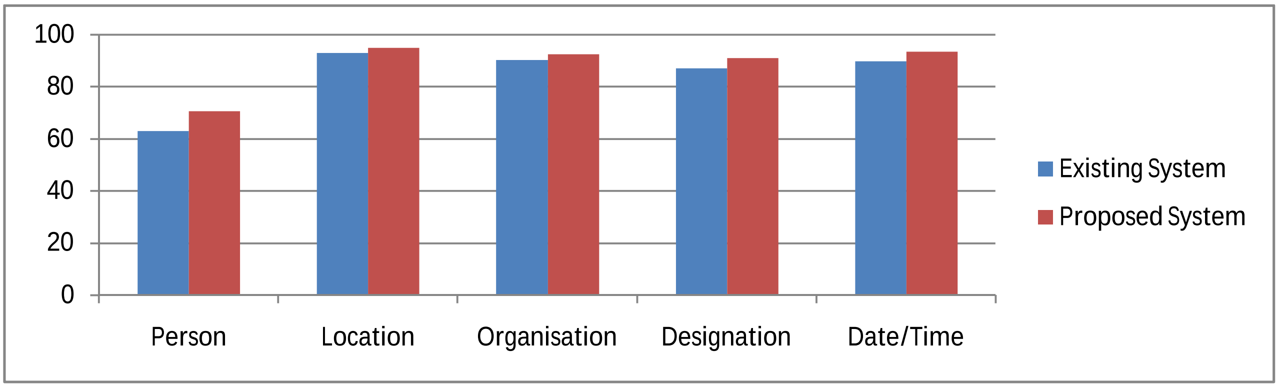

Fig 2. Graph of recall value existing system v/s proposed system

VI. Conclusion and Future Work

The NER system is tested against various inputs. Proposed work shows 90% accuracy with input is fetched from database.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. There are some characters which have double meaning to solve this ambiguity further improvement is required. In future system can be made which will work on various languages like English, Hindi, Punjabi altogether.Corpus for these languages is also required to be developed separately.

REFERENCES

1. 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.
2. Dan Klein, Joseph Smarr, Huy Nguyen, Christopher D. Manning, Named Entity Recognition with Character-Level Models
3. 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
4. Gobinda G. Chowdhury ,Dept. of Computer and Information Sciences ,University of Strathclyde, Glasgow G1 1XH, UK
5. ZhenzhenKou, William W. Cohen,(2005) “High-Recall Protein Entity Recognition Using a Dictionary”, in 13th Annual International Conference on Intelligent Systems for Molecular Biology.
6. Sujan Kumar Saha, Sudeshna Sarkar, Pabitra Mitra, (2008), “Gazetteer Preparation for Named Entity Recognition in Indian Languages”, the 6th workshop on Asian Language Resources.
7. Sujeet Kumar, (2008), “ Named Entity Recognition for Hindi”, Indian Institute of Technology, Kanpur
8. Wei Li and Andrew McCallum, “Rapid Development of Hindi Named Entity Recognition Using Conditional Random Fields and Feature Induction”, in ACM Transactions on Asian language information Processing, 2003
9. Tzonhan Tsai, Shihung Wu, Chengwei Lee, Chengwei Shih, and Wenlian Hsu, “Mencius: A Chinese Named Entity Recognizer using the Maximum Entropy based Hybrid Model”, International Journal of Computational Linguistics of Chinese Language Processing, Vol. 9; Nov. 1, 2004
10. R.Grishman, Sundheim,(1996), “ Message Understanding Conference6:A Brief History”, Proceedings of International Conference on Computational Linguistics