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 using Hybrid Approach

Shavi Juneja

Research Scholar, Dept. of CSE, GZS PTU Campus, Bathinda, Punjab, India

ABSTRACT*:* Named Entity Recognition is an important area of Natural Language Processing. 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 importance of an efficient technique in NER is rule based approach from the huge collection cannot be overemphasized. The hybrid approach is used 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. In list look up approach a corpus of for the various named entities of Punjabi language is formed which consist of Names of Persons (male and female), location names, country names, cities names, state names, organization names. 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*:* Rule base approach, List look up approach, No name entity Technique, Natural language processing, 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. Named entity recognition is simple for humans and indeed building a NER system with performance is easy. Many named entities are proper names and have initial capital letters and can easily be recognized that way. Expressions are difficult to extract in NER using traditional natural language processing (NLP) because they belong to the open class of expressions i.e. there is an infinite variety and new expressions are constantly being created. Performance can be further improved by using so-called gazetteers or seed lists i.e. lists of people, places and companies. A sentence can be tagged quite easily using these simple rules and lists. However, attaining human performance levels of NER is still difficult due to the large amount of ambiguity present in natural language.

The main purpose of ner system is to generate the handcrafted rules by which named entities such as person names, location names , name of places etc and 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 is based on I) To improve the existing rule by adding no name entity technique. ii) To analyze various NER Systems and to increase the no. of rules like animal/bird name rule, direction name rule, measurement named rule, transport or vehicle named rule, monetary named rule.

II. Related work

In [2] 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 [3] 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 [5] authors represent 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 [7] 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. A trained model is developed and it is compared to baseline entity detection rate with an off-the-shelf NER system. They built a platform to manage the large volume and high flow rate of social media data, including database design, building a sufficiently fast language detection algorithm, and extract the tools to quickly retrieve a subset of tweet collection.

III. PROPOSED SYSTEM

A. *Design Considerations:*

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. All the work has been implemented in ASP.NET visual studio 2010; SQL Server data base 2008 and framework used 4.0.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:
i) 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.
ii) 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.

B. *Description of the Proposed System:*

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. In existing rules the various problems are found, due to these problems their accuracy is low and they need further some improvements. No name entity technique sorts out these problems and develops some more efficient and accurate rules. More appropriate rules are thus needed to be created in existing system to improve the accuracy of the overall system. Improved features by No name entity Technique:

Problems with Existing Systems:

i)Corpus size is small means large corpus is not available.
ii) It can't recognize names consist of multiple words

- Can't recognize Monetary Expressions Ex. 1000/-
- It can't recognize that these are 1000 rupees
-

IV. PERFORMANCE EVALUATION

The performance of NER system is calculated by using following three parameters:
i) Precision
ii) Recall
iii) F- measure

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

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

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

V. RESULTS

The results are explained through the comparison between the existing system and proposed system and it is is done by comparing the 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:

DOI: 10.15680/IJIRCCE.2017. 0510068

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

F- measure

The f-measure values of the proposed system of NE class are calculated and their results are compared with the existing system. The following table shows the f-measure value of existing system and proposed system of NE class:

NE Class	Existing System F measure (F %)	Proposed System F measure (F %)
Person	65.67	75.61
Location	91.25	94.06
Organisation	88.77	92.94
Designation	91.98	94.82
Date/Time	91.75	94.96

Table3. Results of f-measure value

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:

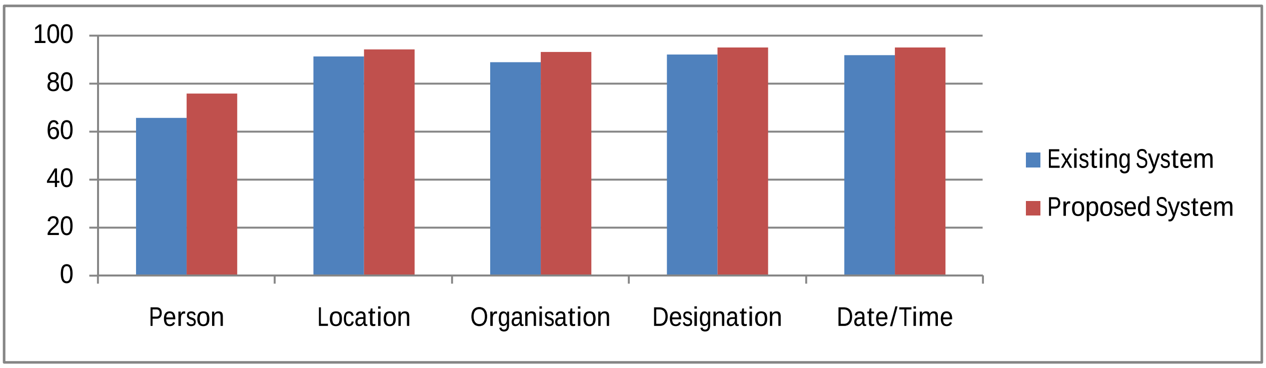

Fig 1. Graph of f-measure value existing system v/s proposed system

VI. Conclusion and Future Work

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 work shows 90% accuracy with input is fetched from database. The proposed system shown better accuracy compared to the existing systems and it also fetch some new named entities. 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. 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. 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. Dan Klein, Joseph Smarr, Huy Nguyen, Christopher D. Manning, "Named Entity Recognition with Character-Level Models" Computer Science Dept., Stanford University.
2. 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.
3. 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.
4. Andrew O. Arnold "Exploiting domain and task regularities for robust named entity recognition" August 2009 CMU-ML-09-109.
5. 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.
6. Artem Boldyrev, Prof. Dr. Gerhard Weikum, "Dictionary-Based Named Entity Recognition" Universitat des Saarlandes Max-Planck-Institut fur Informatik Databases and Information Systems, December 2013.
7. Dr. Timothy W. Finin, William Murnane "Improving Accuracy of Named Entity Recognition on Social Media Data" Master of Science, 2010.
8. 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.
9. 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.
10. 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.
11. 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.
12. 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.
13. Kumar N. and Bhattacharyya Pushpak. 2006. "Named Entity Recognition in Hindi using MEMM" in the proceedings of Technical Report, IIT Bombay, India.
14. 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.

DOI: 10.15680/IJIRCCE.2017. 0510068