Volume 4, Issue 6, June 2014 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Review Paper on Named Entity Recognition System for Punjabi Language Text

Shavi Juneja
M..tech Student
Department of Computer Science
GZS PTU Campus Bathinda
Punjab, India ,

Er. Jyoti Rani
Assistant Professor
Department of Computer Science
GZS PTU Campus Bathinda
Punjab, India

Abstract— Natural Language Processing applications are characterized to make complex interdependent decisions which require large amounts of prior knowledge. 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.

Keywords— NER, Rule based Approach, List look up approach, Linguistic approach.

I. INTRODUCTION

Named Entity Recognition (NER) is a sub problem of information extraction (IE) and is slightly less complex than IE. Named entity recognition is a technology used to recognize proper nouns or entities in text and associating them with the appropriate types. 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. But most of the NER systems are based on the analyzing patterns of POS tags and they also make the use of lists of typed entities like list of possible person names or we can also use the regular expressions for particular types like address patterns or location, target or actor of any terrorist event to be extracted, these named entities first need to be recognized means firstly they are recognized. This 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. 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 foundation of NLP lies in a number of disciplines for example computer and information sciences, linguistics, mathematics, electrical and electronic engineering, artificial intelligence and robotics, psychology, etc. In natural language processing approaches are divided into four categories namely symbolic, statistical, connectionist, and hybrid approach. 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.

II. APPLICATIONS

Major tasks in NLP include are as follows:

- Automatic summarization
- Foreign language reading aid
- Foreign language writing aid
- Information extraction

Information retrieval (IR) – Information retrieval plays one of the most important task in Natural language processing. IR is concerned with storing, searching and retrieving information. It is a separate field within computer science and applications for many natural languages processing . It involves the identification of named entities such as person names, location names, names of organizations, monetary expressions, dates, numerical expressions etc.

In the taxonomy of Computational Linguistics, the NER falls in the category of Information Extraction which deals with the extraction of specific information from the given documents. The main role of NER is to identify expressions such as date and time as well as names of people, places, and organizations.

III. RELATED WORK

Arshdeep Singh ,Jyoti Rani ,Kuljot Singh 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 [1].

Dan Klein, Joseph Smarr, Huy Nguyen, Christopher D. Manning, 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 [2].

Kamaldeep Kaur, vishal Gupta, "Name Entity Recognition for Punjabi Language" 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 [3]

IV. PROBLEM DEFINITION

Accuracy of the NER systems for Indian Languages is very low because neither proper corpus is available nor Proper rule based system is available. Performance of the existing systems is comes out to be 50%-70% which is very low. Hybrid approach will be used in this proposed work which consist of two approaches which are "table look up technique" and "Rule based Approach" to achieve better results.

V. EXISTING TECHNIQUES

a) Rule based Approach :

In this approach a large number of rules are created by considering the features of the Punjabi language. The handcrafted systems follows the 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 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. Thus for many cases it is necessary that person name must follow some kind of title.

b) 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. The NER system uses gazetteer to classify words and suitable lists are created. It is simple, fast and language independent and it is also easy to retarget as only lists are to be created. But its disadvantage of is to maintain the gazetteer list and it cannot resolve ambiguity.

c) Linguistic approach:

The NER system uses some language based rules which are manually written by linguists and other heuristic to classify words. It needs rich and expressive rules and gives good results. The main disadvantage is that these require huge experience and grammatical knowledge of the particular language or domain and also not easily portable and has high acquisition cost. It is very specific to the target data.

d) Machine Learning based approach:

i. Supervised approach: This approach uses the program that learns to classify the given set of labelled examples that are made up of the same number of features. Each example is thus represented with respect to the different feature spaces. This approach requires preparing labelled training data to construct a statistical model, but it cannot achieve a good performance without a large amount of training data, because of data sparseness problem.

ii. Unsupervised approach: In this approach an unsupervised model learns without any feedback. In this learning, the goal of the program is to build representations from the data. These representations can then be used for data compression, classifying, decision making, and other purposes. It is not a very popular approach for NER and the systems that uses the unsupervised learning are usually not completely unsupervised.

VI. CONCLUSION

In this paper authors represent the review on name entity recognition system from a text written in Punjabi language. Literature of various papers is presented along with the techniques used by them. Authors conclude that a lot of work has left for this problem domain. With this there is a need of various new rules to be formed to improve the existing results.

And many first names in Punjabi are also common nouns, this limitation lowers the down the performance of the system which demands to produce the NER system for Punjabi language which generate the more promising results then that of existing ones with improved rule based techniques.

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] Sujeet Kumar, (2008), " Named Entity Recognition for Hindi", Indian Institute of Technology, Kanpur

[5] 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

[6] Gobinda G. Chowdhury ,Dept. of Computer and Information Sciences ,University of Strathclyde, Glasgow G1 1XH, UK

[7] 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

[8] ZhenzhenKou, William W. Cohen,(2005) "High-Recall Protein Entity Recognition Using a Dictionary", in 13th Annual International Conference on Intelligent Systems for Molecular Biology

[9] Mohammad Hasanuzzaman, Asif Ekbal and Sivaji Bandyopadhay, (2009), " Maximum Entropy Approach for Named Entity Recognition in Bengali and Hindi", Academy Publisher, International Journal of Recent Trends in Engineering, Vol. 1, No. 1.

[10] R. Grishman, Sundheim,(1996), " Message Understanding Conference6:A Brief History", Proceedings of International Conference on Computational Linguistics.

[11] Sujan Kumar Saha, Sudeshna Sarkar, Pabitra Mitra, (2008), "Gazetteer Preparation for Named Entity Recognition in Indian Languages", the 6th workshop on Asian Language Resources