International Journal of scientific research and management (IJSRM)
\||Volume||3||Issue||3||Pages|| 2300-2306||2015||
Website: www.ijsrm.in ISSN (e): 2321-3418

Named Entity Recognition (NER) System for Hindi Language Using Combination of Rule Based Approach and List Look Up Approach

Yavrajdeep Kaur[1], Er.Rishamjot Kaur[2]

[1]Department of Computer Science & Engineering,
Baba Farid College of Engineering and Techonolgy, Bathinda, India
ssskaur10@gmail.com

[2]Department of Information Technology,
Baba Farid College of Engineering and Techonolgy, Bathinda, India
jotrishamsran@gmail.com

Abstract: *Named Entity Recognition is a discipline for recognition of named entities in a document associating them with the proper types. Information Extraction is the main requirement of Named Entity Recognition (NER) system. The main functionality of NER is to identify Named Entities. Named Entities can be name of a Person, Location name, Organization name, Time, Date, etc. The NER is a major application of Natural Language Processing (NLP). In proposed system, a NER system is designed and implemented by using Hybrid Approach for Hindi Language. Hybrid Approach is the combination of Rule Based Approach and List look Approach. Named entities have been identifying by using some existing rules and adding new rules. Named entities have been stored in corpus; from here named entities have been identified from corpus. The accuracy of the proposed system comes out to be approximately 96%.*

Keywords: *Named Entity Recognition, Natural Language Processing, Hybrid Approach, Rule Based Approach, List look Approach.*

1. Introduction

1.1 NLP (Natural Language Processing)

Natural Language Processing (NLP) is a field of computer science and linguistics concerned with the interactions between computers and human (natural) languages. In theory, natural-language processing is a very attractive method of human-computer interaction. Natural-language understanding is sometimes referred to as an AI-complete problem, because natural-language recognition seems to require extensive knowledge about the outside world and the ability to manipulate it.

The history of NLP generally starts in the 1950s, although work can be found from earlier periods. Although NLP may encompass both text and speech, work on speech processing has evolved into a separate field. Natural language generation systems convert information from computer databases into readable human language.

NLP is an area of research and application that explores how computers can be used to understand and manipulate natural language text or speech to do useful things. NLP researchers aim to gather knowledge on how human beings understand and use language so that appropriate tools and techniques can be developed to make computer systems understand and manipulate natural languages to perform the desired tasks. The foundations of NLP lie in a number of disciplines, viz. computer and information sciences, linguistics, mathematics, electrical and electronic engineering, artificial intelligence and robotics, psychology, etc.

Natural language processing approaches fall roughly into four categories: symbolic, statistical, connectionist, and hybrid. Symbolic and statistical approaches have coexisted since the early days of this field. Connectionist NLP work first appeared in the 1960's. For a long time, symbolic approaches dominated the field. In the 1980's, statistical approaches regained popularity as a result of the availability of critical computational resources and the need to deal with broad, real world contexts. Connectionist approaches also recovered from earlier criticism by demonstrating the utility of neural networks in NLP.

Various sub problems in NLP include speech segmentation, text segmentation, part of speech tagging, word sense disambiguation, syntactic ambiguity.ent and are identified in italic type, within parentheses, following the example. Some components, such as multi-leveled equations, graphics, and tables are not prescribed, although the various table text styles are provided. The formatter will need to create these components, incorporating the applicable criteria that follow [4].

Applications of NLP: NLP used in various ways such as:

- Information Extraction
- Automatic summarization
- Machine translation
- Named entity recognition (NER)

- Natural language generation
- Natural language understanding
- Optical character recognition (OCR)
- Question answering
- Part-of-speech tagging
- Speech segmentation
- Speech recognition
- Information extraction (IE) etc.

1.2 Named Entity Recognition (NER)

Named Entity Recognition (NER) is basically concerns with Natural Language Processing (NLP). Named Entity Recognition (NER) is the application of NLP. The main goal of Named Entity Recognition (NER) is to identify and then classify named entities into some categories. The task of Named Entity Recognition (NER) is to identify all named entities from given document or paragraph and after that classify all named entities such as

- Name of the person (person can be male or female)
- Location name (location can be city, state, country etc)
- Organization name
- Date
- Vehicle name etc.

Applications of Named Entity Recognition (NER): NER has many applications .It is used in various ways like:

- Information retrieval
- Question answer generating system
- Machine Translation
- Transliteration
- Spell Checker
- Information Extraction

The architecture of Named Entity Recognition system shown as:

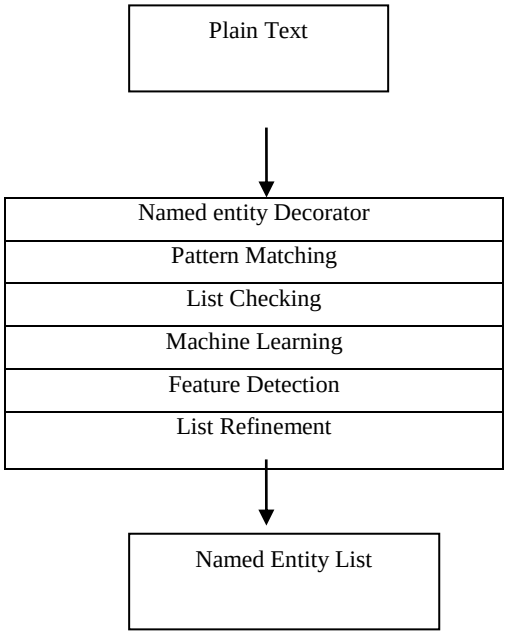


Figure 1.2: Architecture of NER system [4]

Further example is given to explain Named Entity Recognition (NER) that is:

भगत सिंह का जन्म 28 सितंबर 1907 में हूआ था। उनके पिता का नाम सरदार किशन सिंह और माता का नाम विद्यावती कौर था। यह एक सिख परिवार था जिसने आर्य समाज के विचार को अपना लिया था। उनके परिवार पर आर्य समाज व महर्षि दयानन्द की विचारधारा का गहरा प्रभाव था ! अधिकांश क्रांतिकारियों को देश प्रेम की प्रेरणा महर्षि दयानन्द के साहित्य व आर्य समाज से मिली ! अमृतसर में 13 अप्रैल 1919 को हूए जलियाँवाला बाग हत्याकाण्ड ने भगत सिंह की सोच पर गहरा प्रभाव डाला था।

First task is to identify all named entities from given Hindi paragraph.

भगत सिंह, 28 सितंबर 1907, किशन सिंह, विद्यावती कौर, सिख, आर्य समाज , दयानन्द, अमृतसर, 13 अप्रैल 1919

Now next task is to categorize them:

Name of the person: भगत सिंह, किशन सिंह, विद्यावती कौर, दयानन्द

Date: 28 सितंबर 1907, 13 अप्रैल 1919

Location: अमृतसर

Organization name: सिख, आर्य समाज

1.3. Existing Approaches of NER

The named entity recognition systems mainly can be divided into

1.3.1. Rule Based System or Hand crafted System

1.3.2. Machine Learning System

1.3.1. Rule Based System/Hand Crafted System

This approach was used in earlier time for NER that is based on pattern matching. Hand crafted methods can be further divided into:

(a)List Look up Method

(b)Linguistic approach

The drawbacks of these rule based techniques are:

They need huge experience grammatical knowledge on the particular language. The development is generally takes too much time. Changes in the system are not easy to maintain. These systems are not transferable, which means that one rule-based NER system made for a particular language or domain cannot be used for other languages or domains.

1.3.2. Machine learning based approach

This approach is used in now days. These techniques produce output in a very short time. . Machine learning based approach can be further divided into:

(a)Supervised approach

(b)Unsupervised approach

Various machine learning methods are:

- Hidden Markov Models (HMM)
- Maximum Entropy Markov Models (MEMMs)

Disadvantages: A drawback of MEMMs is that they potentially suffer from the Label Bias problem.

- Conditional Random Field (CRF)

2. Problem Definition

2.1 Indian language has lack in research work for NER system. Mostly researches have been done in foreign languages, for example English, Chinese and Spanish etc. This is major drawback of NER. Main cause is non-availability of data for corpus. Information on internet and on web pages is mostly in English language. For storing database entries in corpus, either translation or transliteration has been done.

2.2 NER system has highly needed to enhance new rules, like, existing system for NER do not use 'No Name Entity Rule'. This rule has a big impact on the performance of NER system. Rules need to be redefined to achieve better results.

2.3 Its need to resolve ambiguities, which is also a problem in NER in Hindi.

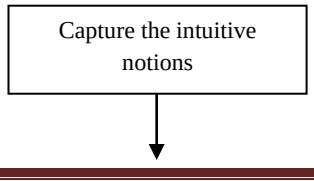
3. Proposed Work

Proposed system works on Hybrid Approach for Named Entity Recognition (NER) in Hindi. The word 'Hybrid' means combination of two or more than two approaches. Two approaches are Rule Based approach and List Look Up approach. The proposed NER system is capable of extracting 10 named entities that are, person names, location names, city/state/country names organization names, dates, money value, measurement value, direction value , transport and bird/animal values.

3.1 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.

The proposed system works on new rule that is 'no name entity rule'. No name entity rule is used which improves or modifies the existing rules. This rule analyzes the various NER Systems and results are compared with the existing approaches.



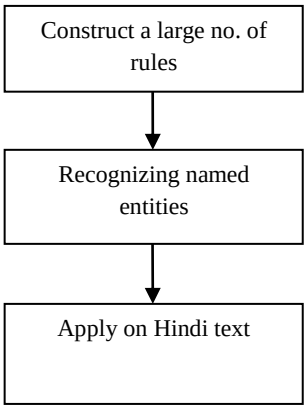


Figure 3.1: Rule Based Approach

3.2 List Lookup Approach

In this approach a corpus is created of the names entities for Hindi language. In this corpus various types of tables are created, for example table for names of persons, , location name , suffix table , prefix table , no name entity table , Organization table etc , document from which names are to be extracted is compared with the database created and names entities are identified.

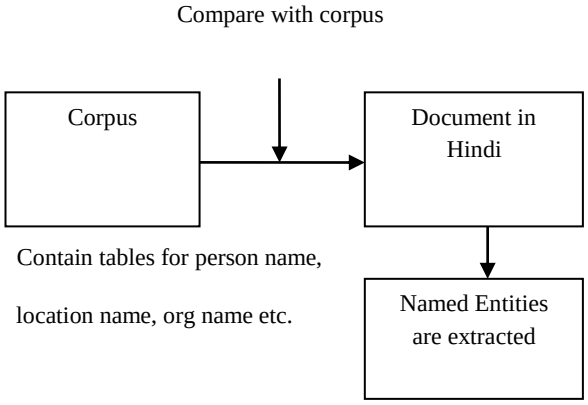


Figure 3.2: List Look up Method

Named entities have been stored in corpus from Hindi newspapers on web. Some named entities have been either translated or transliteration because contents on web are mostly in English language.

4. Performance Measurement

The performance of the NER system is measured using three parameters i.e. Precision (P), Recall(R) and F-measure.

The precision measures the number of correct NEs, obtained by NER system, over the total number of NEs extracted by NER system.

PRECISION (P) = No. of correct NE's / Total no. of NE's given by system

The recall measures the number of correct NEs, obtained by NER system over the total number of NEs in a text that has been used for testing.
RECALL (R) = No. of correct NE's / Total no. of NE's in document.

The F-measure represents harmonic mean of precision and recall.
F-MEASURE = 2RP / PR

The F-measure (also F-score or F-measure) is a score to measure the accuracy of the system. It considers both the precision p and the recall r of the system to calculate average score.

5. Results

The proposed NER system has been implemented using vb.net platform and gazetteers lists are stored as tables in the database. The system need input documents which contain name entities such as person names, location names, organization names, date, money value , animal/birds entities, direction entities etc. These test documents are taken from e-copies of Hindi newspapers such as Danik Jagran; Punjab kesari etc.
The proposed system uses 62 word files for testing. Some input data has been created according to name entities. Some of the links of newspaper which are used for gathering input are:

http://navbharattimes.indiatimes.com/

http://www.livehindustan.com/location/rajwarkhabre/39-8-localnews-30.html

With these 62 inputs, the results were calculated using three parameters precision, recall and f-measure and corresponding graphical representation were drawn.

Results for proposed NER system is depicted in table 5.1. This result is calculated for 9 name entities that are Person, Location, Organization, Date, Measurement, Direction, Transport and Animal/Bird. Graph 5.1 has been drawn corresponding to table 5.1.

Table 1: Results of Proposed NER System for Hindi Language

ENTITY	*TOTAL NE's IN DOCUMENT*	*TOTAL NE's GIVEN BY SYSTEM*	*CORRECT NE's*	*PRECISION P*	*RECALL R*	*F-MEASURE*
Person	276	270	265	0.9814	0.9601	97.06
Location	285	280	275	0.9821	0.9649	97.34
Organization	159	143	139	0.972	0.8742	92.05
Date	88	84	79	0.9404	0.8977	91.86
Money	37	36	36	1	0.9729	98.63
Measurement	98	97	95	0.9793	0.9693	97.43
Direction	27	27	25	0.9259	0.9259	92.59
Transport	28	27	27	1	0.9642	98.18
Animal	32	30	30	1	0.9375	96.77
Total Accuracy of Proposed NER System						95.77%

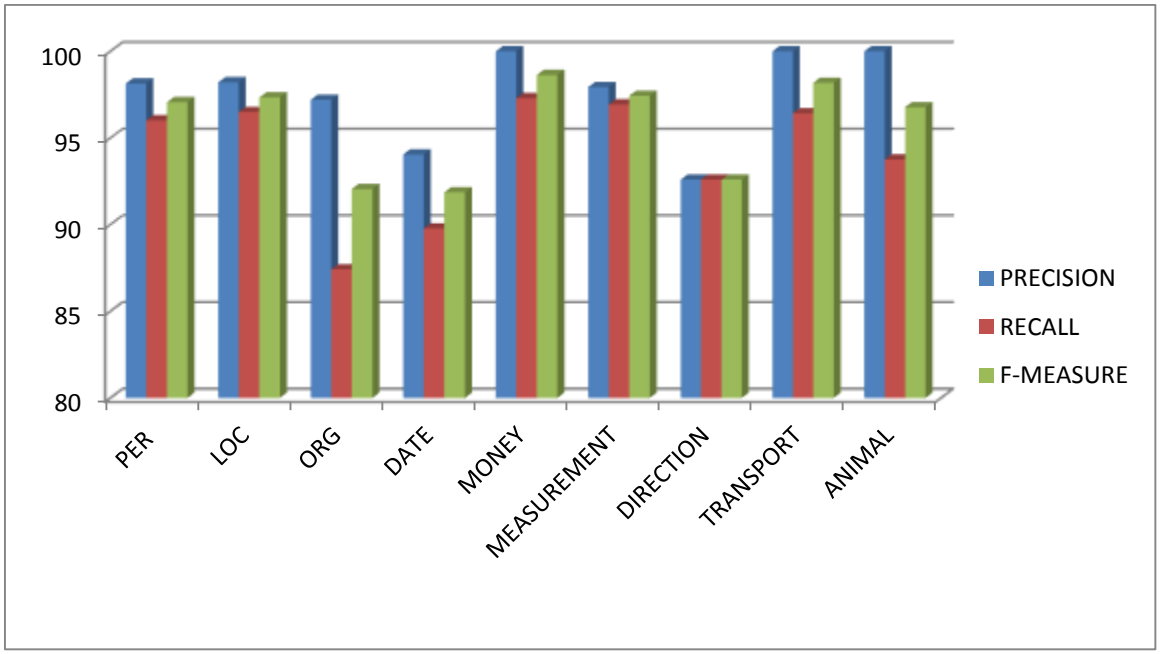


Figure 5.1: Graphical Representation for Precision, Recall and F-Measure

Conclusion

Very less work has been done in NER for Hindi and other Indian languages. This paper depicts Named Entity Recognition (NER) system for Hindi language. The proposed NER system for Hindi language used Hybrid Approach. The hybrid approach is the combination of two approaches; rule based approach and list look up approach. The proposed system identifies three new name entities that is money value, direction values and animal/bird entities; and adding new rule 'no name entity rule' to improve the overall accuracy of the system. 'No name entity rule' improved or modified the existing rules. Different tables have been created in database for Hindi language and named entities have been extracted from these tables in list look up approach. Three parameters have been used to calculate the accuracy of the proposed system. These three parameters are: Precision (P), Recall(R) and F-measure (F). The accuracy of the proposed system is 95.77%. The accuracy of the proposed system depends upon the named entity stored in the database and handcrafted rules to identify the named entity.

6. Future Scope

In future further work can be done for NER in Hindi. In future a new technique can be developed to enhance the performance of NER in Hindi. New rules can be created to improve the performance of NER system. In future the size of corpus can be increased by adding more name entities in database. The accuracy of the overall system and rules can be improved. More name entities such as title and English abbreviations (e.g. BSNL, CBI etc.) can be extracted. More corpus size, more accuracy can be achieved.

References

[1] Deepti Chopra, Nusrat Jahan, Sudha Morwal, "Hindi Named Entity Recognition By Aggregating Rule Based Heuristics and Hidden Markov Model", International Journal of Information Sciences and Techniques (IJIST) Vol.2, No.6, November 2012.

[2] Darvinder kaur, Vishal Gupta, "A survey of Named Entity Recognition in English and other Indian Languages", IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 6, November 2010.

[3] Yunita Sari, Mohd Fadzil Hassan , Norshuhani Zamin , " Rule-based Pattern Extractor and Named Entity Recognition: A Hybrid Approach" , 978-1-4244-6716-711 0/$26.00 ©20 1 0 IEEE .

[4] 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.

[5] Radu Florian and Abe Ittycheriah and Hongyan Jing and Tong Zhang, "Named Entity Recognition through Classifier Combination".

[6] 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.

[7] Dan Klein, Joseph Smarr, Huy Nguyen, Christopher D. Manning, "Named Entity Recognition with Character-Level Models".

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

[9] Andrew McCallum and Wei Li, "Early Results for Named Entity Recognition with Conditional Random Fields, Feature Induction and Web-Enhanced Lexicons", in 7th Conference on Natural Language Learning(CoNLL).

[10] 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.

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

[12] 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.

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

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

[15] Sujan Kumar Saha, Partha Sarathi Ghosh, Sudeshna Sarkar, and Pabitra Mitra, "Named Entity Recognition in Hindi using Maximum Entropy and Transliteration" , October 22, 2008.

[16] Sujan Kumar Saha, Sanjay Chatterji, Sandipan Dandapat, Sudeshna Sarkar ,Pabitra Mitra, "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. C 2008 Asian Federation of Natural Language Processing.

[17] Animesh Nayan, B. Ravi Kiran Rao, Pawandeep Singh, Sudip Sanyal and Ratna Sanya, "Named Entity Recognition for Indian Languages" , Proceedings of the IJCNLP-08 Workshop on NER for South and South East Asian Languages, pages

[18] 97–104, Hyderabad, India, January 2008. C 2008 Asian Federation of Natural Language Processing.

[19] Shilpi Srivastava, Mukund Sanglikar, D.C Kothari, "Named Entity Recognition System for Hindi Language: A Hybrid Approach", International Journal

of Computational Linguistics (IJCL), Volume (2): Issue (1): 2011.

[20] David Nadeau, Satoshi Sekine, "A survey of named entity recognition and classification".

[21] Lev Ratinov ,Dan Roth , "Design Challenges and Misconceptions in Named Entity Recognition" ,Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL), pages 147–155, Boulder, Colorado, June 2009. C 2009 Association for Computational Linguistics.

[22] Jiafeng Guo, GuXu, Xueqi Cheng,Hang Li, "Named Entity Recognition in Query".

[23] Asif Ekbal, Rejwanul Haque, Amitava Das, Venkateswarlu Poka , Sivaji Bandyopadhyay, "Language Independent Named Entity Recognition in Indian Languages" , Proceedings of the IJCNLP-08 Workshop on NER for South and South East Asian Languages, pages 33–40, Hyderabad, India, January 2008. C 2008 Asian Federation of Natural Language Processing.

[24] Dinesh Kumar, Gurpreet Singh Josan, "Part of Speech Taggers for Morphologically Rich Indian Languages: A Survey", International Journal of Computer Applications (0975 – 8887) Volume 6– No.5, September 2010.

Author Profile

Yavrajdeep Kaur is a Student of M.Tech (computer science Engg.) at Baba Farid College of engineering and technology, Bathinda. She has received her B.Tech in Computer Science from Punjabi University Neighbourhood Campus, Rampura Phul, in 2012. She is perusing her M.Tech Thesis in the area of Natural Language Processing.

Er.Rishamjot Kaur received M.Tech degrees in Computer Science from Punjabi University, Patiala in 2012. She is working as Assistant Professor in Department of Information Technology at Baba Farid College of Engineering and Technology, Bathinda, India