IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

Named Entity Recognition in Indian Languages: A Survey

Nusrat Jahan[*1], Sudha Morwal[2]

[*1,2]Department of computer science, Banasthali University, Jaipur-302001, Rajasthan, India

nusratkota@gmail.com

Abstract

Named Entity Recognition (NER) is the process of determining and identifying all proper nouns into pre-defined classes such as persons, places, organization and others. Lots of work has been done in Western languages but for Indian languages it is a difficult and challenging task and also limited due to lack of resources, but it has started to appear recently. In this paper we present a brief summary of NER and its issues in the Indian languages. We also explain the different techniques used in NER and literary work review on different languages done by different scientists. Named Entity Recognition (NER) play big part in various Natural language processing (NLP) task like machine translation, text to speech synthesis, natural language understanding, Information Extraction, Information retrieval, question answering etc. Named Entity recognition comes under the domain of "information extraction", which extracts specific kinds of data from records.

Keywords: Named Entity Recognition (NER), Support Vector Machine (SVM), Maximum Entropy Markov Model (MEMM), Decision Tree (DT), Conditional Random Field (CRF).R

Introduction

Natural Language Processing (NLP) is the computerized approach for analyzing text that is based on both a set of theories and a set of technologies. Named Entity Recognition is a subtask of Information extraction where we locate and classify proper names in text into predefined categories. The named entities may be classified as follows:

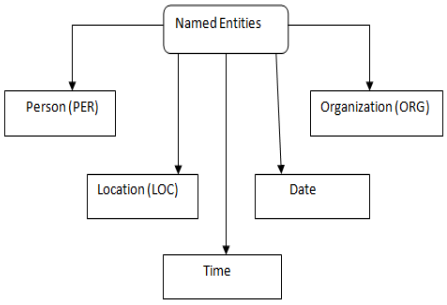

Fig 1: Named entity classification

NER is a precursor for many natural languages processing tasks such as machine translation, more accurate internet

Search engines, automatic indexing of documents, automatic question-answering, information retrieval etc. An accurate NER system is needed for these applications.

In the last decades, substantial efforts have been made and impressive achievements have been obtained in the area of Named Entity recognition (NER) for text documents. Generally NER can be handled as a two-step process - identification of appropriate nouns and its classification. The first thing is the recognition of appropriate nouns from the text and the second phase is the classification of these proper nouns into any one of the classes like person name, organization name, location name and other classes. The main problem of NER is how to tag the words and what tag is assigned to the entities like person, location etc. Sometimes ambiguities are available in the document and we have to take care of them to be able to allocate the appropriate tag.

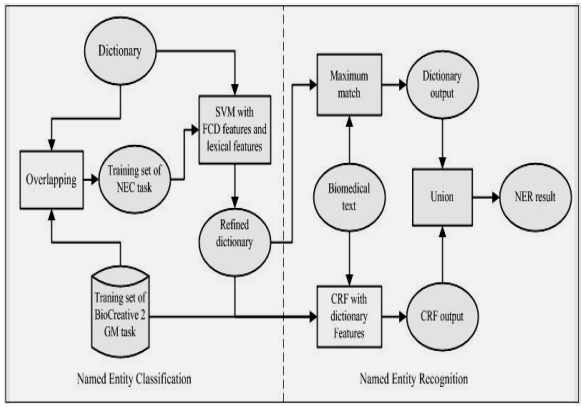

Fig2. System architecture

Example- Consider a Hindi sentence as follows:

" आगरा शहर ताजमहल के लिए जगत प्रसिध्द है जो कि मुगल बादशाह शाहजहाँ द्वारा अपनी प्रिय पत्नी की याद में बनाया गया था। ".

In the above sentence, the NER based system first identifies the Named Entities and then categorize them into different Named Entity classes like:

"आगरा/CITY शहर/OTHER ताजमहल/LOC के/OTHER लिए/OTHER जगत/OTHER प्रसिध्द/OTHER है/OTHER जो/OTHER कि/OTHER मुगल/OTHER बादशाह/OTHER शाहजहाँ/PER द्वारा/OTHER अपनी/OTHER प्रिय/OTHER पत्नी/OTHER की/OTHER याद/OTHER में/OTHER बनाया/OTHER गया/OTHER था/OTHER ।/OTHER

In this sentence, first word आगरा refers to the city name, so it is allotted 'CITY' tag.

The word ताजमहल refers to the name of location. So, it is allotted 'LOC' tag. The word शाहजहाँ refers to the name of a person. So it is assigned the 'PER' tag and rest words are assigned the 'OTHER' tag means not a Named Entity tag.

Similarly consider the example of English language:
"Ram eats apple"
"Ram/PER eats/OTHER apple/FRUIT" Here Ram is the 'PER', eats is assigned 'OTHER' tag and apple is FRUIT.

Consider another example of Bengali language like:-
রোহন একটি ছেলে.
রোহন/PER একটি/OTHER ছেলে/OTHER ./OTHER

Approaches for NER

There are basically three approaches that are employed in Named Entity Recognition. These approaches are:-

A. *The Rule based / Handcrafted Approach*
- *List Lookup Approach*
- *Linguistic Approach*

B. *MMachine Learning Based Approach /Automated Approach*
- *Hidden Markov Models (HMMs)*
- *Maximum Entropy Markov Model*
- Conditional Random Field (CRF)
- Support Vector Machine (SVM)
- Decision Tree (DT)

C. *Hybrid Approach*

A. The Rule based / Handcrafted Approach:

The rule based or handcrafted approach consists of set of handcrafted rules which are written by language expert of particular language domain.

1) LList Lookup Approach:

The list lookup approach has following advantage and disadvantages.

Advantage:-
- NER system uses gazetteer to classify words.
- We just have to create a suitable list in the gazetteer.
- It is simple, fast and language independent.
- It is also easy to retarget as we just have to create lists.

Disadvantage:-
- This approach only works for lists in the gazetteer.
- We have to collect and maintain the gazetteer.
- This approach cannot resolve ambiguity.

1) *Linguistic Approach:*
- NER system uses some language based rules and other heuristic to classify words.
- It needs rich and expressive rules and gives good results.
- It requires an advanced knowledge of grammar and other language related rules.
- These calls for a thorough knowledge and advanced skills related to the Language under consideration are needed to come up with good rules and heuristic.

2) Machine Learning Based Approach /Automated Approach:

Machine-learning (ML) approach Learn rules from annotated corpora. Now a day's machine learning approach is commonly used for NER because training is easy, same ML system can be used for different domains and languages and their maintenance is also less expensive. The various machine learning approaches used are:

1) Hidden Markov Models (HMMs):

It is a generative model. The model assigns a joint probability to paired observation and label sequence. Then the parameters are trained to maximize the joint likelihood of training sets. It uses forward-backward algorithm, Viterbi Algorithm and Estimation-Modification method for modelling. Its basic theory is elegant and easy to understand. Hence it is easier to implement and analyze. In order to define joint probability over observation and label sequence HMM needs to enumerate all possible observation sequence. Hence it makes various assumptions about data like Markovian assumption

i.e. current label depends only on the previous label. Also it is not practical to represent multiple overlapping features and long term dependencies. Number of parameter to be evaluated is huge. So it needs a large data set for training.

2) *Maximum Entropy Markov Models (MEMMs)*:
Advantage of MEMM:-

- It is a conditional probabilistic sequence model.
- It can represent multiple features of a word and can also handle long term dependency.
- It is based on the principle of maximum entropy which states that the least biased model which considers all know facts is the one which maximizes entropy.
- It solves the problem of multiple feature representation and long term dependency issue faced by HMM.
- It has generally increased recall and greater precision than HMM.

Disadvantage of MEMM:-

- It has Label Bias Problem.
- The probability transition leaving any given state must sum to one.
- So it is biased towards states with lower outgoing transitions. The state with single outgoing state transition will ignore all observations.
- To handle Label Bias Problem we can change the state-transition structure or we can start with fully connected model and let the training procedure decide a good structure.

3) Conditional Random Field (CRF):

CRF stands for Conditional Random Field. It is a type of discriminative probabilistic model. It has all the advantages of MEMMs without the label bias problem. CRFs are an undirected graphical model (also known as random field) which is used to calculate the conditional probability of values on assigned output nodes given the values assigned to other assigned input nodes.

4) Support Vector Machine (SVM):

SVM is one of the famous supervised machine learning algorithms for binary classification in all various data set and it gives the best results where the data set is a few. To solve a classification task by a supervised machine learning model like SVM, the task usually involves with training and testing data, which consists of some data instances. The goal of a supervised SVM classifier method is to produce a model which predicts target value of the attributes. For each SVM, there are two data set namely, training and testing, where the SVM used the training set to make a classifier model and classify testing data set based on this model with use of their features[9].

5) Decision Tree (DT):

DT is a powerful and popular tool for classification and prediction. The attractiveness of DT is due to the fact that in contrast to neural network, it presents rules. Rules can readily be expressed so that human can understand them or even directly use them in a database access language like SQL so that records failing into a particular category may be tree. Decision Tree is a classifier in the form of a tree structure where each node is either a leaf node-indicates the value of the target attributes (class) of expressions, or a decision node that specifies some text to be carried out on a single attribute value with one branch and sub-tree for each possible outcome of the text. It is an inductive approach to acquire knowledge on classification [10].

3) Hybrid Model Approach:

In this approach Rule Based approach and Machine Learning approaches are mixed for more accuracy to identify NERs. Here several combinations are used as follows:

- HMM approach and Rule Based approach.
- CRF approach and Rule Based approach.
- MEMM approach and Rule Based approach.
- SVM approach and Rule Based approach.

Issues with NER

In technology for Indian languages NER has an essential need. NER in Indian Languages is a more challenging problem as compared to European languages due to absence of capitalization, resources etc. Because of these difficulties an English NER system cannot be used directly for Indian languages.
The main problem of NER is how to tag the words and what tag is assigned to the entities like person, organization and location etc. Sometimes ambiguities exist in the document and we have to resolve them in order to assign the correct tag. Some of the major issues related with the recognition of named entities in Indian Languages are as follows:

- Unlike English and most of the European languages, Indian languages lack the capitalization information that plays a very important role to identify NEs in those languages.
- Hindi names are ambiguous and this issue makes the recognition a very difficult task.
- Hindi, like other Indian languages, is also a resource poor language. Annotated corpora, name dictionaries, good morphological

analyzers, POS taggers etc. are not yet available in the
required quantity and quality.
- Lack of standardization and spelling.
- Web sources for name lists are available in English, but such lists are not available in Indian languages.
- Although Indian languages have a very old and rich literary history still technology development are recent.
- Non-availability of large gazetteer.
- Named entity recognition systems built in the context of one domain do not usually work well in other domains.
- Indian languages are relatively free-order languages.

Applications of NER

- NER finds application in most of the NLP applications. The following list mentions few of its applications.
- NER is very useful for search engines. NER helps in structuring textual information, and structured information helps in efficient indexing and retrieval of documents for search.
- NER finds application in machine translation, as well. Usually, entities identified as Named Entities are transliterated as opposed to getting translated.
- Before reading an article, if the reader could be shown the named entities, the user would be able to get a fair idea about the contents of the article.
- Automatic indexing of Books: Most of the words indexed in the back index of a book are Named Entities.
- Useful in Biomedical domain to identify Proteins, medicines, diseases, etc.

NE Tagger is usually a sub-task in most of the information extraction tasks because it adds structure to raw information.

Previous Work in NER by Using Different Methods

Author	Year	method	Language	Trained data	Tested data	Precision	Recall	F-measure
[1]	2007	HMM	Bengali	-	-	79.48%	90.02%	84.5%
			Hindi	-	-	74.6%	82.5%	78.35%
[2]	2008	CRF	Bengali	122,467	30505	50.75%	62.92%	58.74%
			Hindi	502,974	38708	75.90%	22.87%	35.08%
			Telugu	64026	6356	34.90%	14.49%	20.45%
			Oriya	93173	24640	4.17%	1.47%	2.16%
			Urdu	35,447	3782	50.84%	22.82%	31.47%
[3]	2008	ME	Hindi	234K	25K	-	-	81.52%
[4]	2008	CRF	Hindi	-	-	-	-	58.85%
[5]	2008	SVM	Bengali	150K	-	89.40%	94.30%	91.80%
[6]	2008	CRF	Telugu	13425	6223	-	-	91.95%
[8]	2009	ME	Bengali	-	-	82.63%	88.01%	85.22%
			Hindi	-	-	79.23%	86.40%	82.66%
[7]	2010	SVM	Hindi	502974	60K	74.34%	80.23%	77.17%
			Bengali	122467	35K	80.12%	88.61%	84.15%

Performance evaluation metric

The performance of the system can be measured by calculating following parameters. These parameters are precision, recall and F Measure.

- **Precision (P):** Precision is the fraction of the documents retrieved that are relevant to the User's information need.

$$\text{Precision (P)} = \frac{\text{correct answers}}{\text{answers produced}}$$

- **Recall (R):** Recall is the fraction of the documents that are relevant to the query that are successfully retrieved.

$$\text{Recall(R)} = \frac{\text{correct answers}}{\text{total possible correct answers}}$$

- **F-Measure:** The weighted harmonic mean of precision and recall, the traditional F-measure or balanced F-score is:

$$F1 - Measure = \frac{2PR}{P+R}$$

Conclusion

In this survey we have studied the different techniques employed for NER, and have identified the various problems in the task particularly for ILs. Named Entity Recognition is one of the current topics of research in the field of NLP. In English and in many European Languages, a lot of work has been done in the field of Named Entity Recognition. F-measure of 93.3% has been achieved till now in NER in English. This fact motivates us not only to perform NER in the Indian Languages but also find ways to improve the performance metrics of a Named Entity Recognition based system (Precision, Recall, F-Measure).Apart from all this, there are also many fascinating approaches that drives us to perform Named Entity Recognition in Indian Languages.

Acknowledgement

We would like to thank all those who helped us in accomplishing this task.

References

[1] A. Ekbal and S. Bandyopadhyay, "A Hidden Markov Model Based Named Entity Recognition System: Bengali and Hindi as Case Studies," in *Proceedings of 2nd International conference in Pattern Recognition and Machine Intelligence*, Kolkata, India, 2007, pp. 545–552.

[2] Asif Ekbal, Rejwanul Haque, Amitava Das, Venkateswarlu Poka and Sivaji Bandyopadhyay "Language Independent Named Entity Recognition in Indian Languages "in *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 available at* : http://ltrc.iiit.ac.in/ner-ssea-08.

[3] S. K. Saha, S. Sarkar, and P. Mitra, "A Hybrid Feature Set based Maximum Entropy Hindi Named Entity Recognition," in *Proceedings of the 3rd International Joint Conference on NLP*, Hyderabad, India, January 2008, pp. 343–349.

[4] A. Goyal, "Named Entity Recognition for South Asian Languages," in *Proceedings of the IJCNLP-08 Workshop on NER for South and South- East Asian Languages*, Hyderabad, India, Jan 2008, pp. 89–96.

[5] A. Ekbal and S. Bandyopadhyay, "Bengali Named Entity Recognition using Support Vector Machine," in *Proceedings of the IJCNLP-08 Workshop on NER for South and South East Asian languages*, Hyderabad, India, January 2008, pp. 51–58.

[6] P.Srikanth and K. N. Murthy, "Named Entity Recognition for Telegu," in *Proceedings of the IJCNLP-08 Workshop on NER for South and South East Asian languages*, Hyderabad, India, Jan 2008, pp. 41–50.

[7] A. Ekbal and S. Bandyopadhyay, "Named Entity Recognition using Support Vector Machine: A Language Independent Approach," *International Journal of Computer, Systems Sciences and Engg(IJCSSE)*, vol. 4, pp.155–170, 2008.

[8] M. Hasanuzzaman, A. Ekbal, and S. Bandyopadhyay, "Maximum Entropy Approach for Named Entity Recognition in Bengali and Hindi," *International Journal of Recent Trends in Engineering*, vol. 1, May 2009.

[9] B. Sasidhar#1, P. M. Yohan*2, Dr. A. Vinaya Babu3, Dr. A. Govardhan4," A Survey on Named Entity Recognition in Indian Languages with particular reference to Telugu**", available at** http://www.ijcsi.org/papers/IJCSI-8-2-438-443.pdf

[10] Asif Ekbal and Sivaji Bandyopadhyay 2008 " Bengali Named Entity Recognition using Support Vector Machine "Proceedings of the IJNLP-08 Workshop on NER for South and South East Asian Languages Hyderabad, India.

[11] 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.Availableat: http://cscjournals.org/csc/manuscript/Journals/IJCL/volume2/Issue1/IJCL-19.pdf

[12]"Padmaja Sharma, Utpal Sharma, Jugal Kalita"Named Entity Recognition: A Survey for the Indian Languages" (Language in India www.languageinindia.com 11:5 May 2011 Special Volume: Problems of Parsing in Indian Languages.) Available at http://www.languageinind