A Brief Survey on Name Entity Recognition in Natural Language Processing For Indian Languages

N.Vasunthira Devi

Ph.D., Research Scholar in Computer Science,
Mother Teresa Women's University, Kodaikanal, India
vasunthira@gmail.com

Dr.R.Ponnusamy

Professor, Department of Computer Science & Engineering,
Sri Lakshmi Ammal Engineering College, Chennai, Tamilnadu, India.
r_ponnusamy@hotmail.com

***Abstract*— NER is a subsequent task in NLP. Named Entity Recognition (NER) is the mission of handling text to classify and categorize names, which is a vital element in various NLP applications, permitting the extraction of valuable sequence of things from documents. An accurate NER and extraction is mainly resolve maximum difficulties in research part such as Information Retrieval (IR), Video Annotation, Question Answering (QA) and Summarization Systems, Semantic Web Search, Machine Translation (MT) and Bioinformatics. At the present time many researchers use different NER approaches such as Rule-based Approach, Machine Learning-based Approach and Hybrid Approach, to classify names from text. This paper presents a study about Named Entity Recognition research completed for Indian languages. The study and opinions related to approaches, procedures and features essential to implement NER for various languages particularly for Indian languages is reported.**

Keyword - Name Entity Recognition; Natural language processing; Information Retrieval; Question Answering System;

I. INTRODUCTION

The word “Named Entity”, the term Named checks the job to those entities for which one or various inflexible designators views as referent. NER is mostly used in NLP. Language is a most needed thing for human beings. To create machine understand such thoughtful of NLP is used. There has been increasing attention in this part of study since the later 1990s. NER is a programmed method and it is a subtask of Information Extraction (IE) that pursues to detect and categorize features. The pre-defined categories in recorded text such things as the Person’s Name, locations, Organizations, time, percentages, financial values, capacities etc.,

The assignment of Named Entity Recognition (NER) is to classify entirely accurate nouns in a paper into pre-defined classes. NLP is two-step method such as (i) the identification of accurate nouns and (ii) classification. Identification is troubled with marking the incidence of a term/phrase as Named Entity in the particular phrase and classification is for signifying segment of the recognized NE. Even though a much of work has been completed in English and additional foreign languages similar to Greek, Spanish, French, Chinese etc., with great precision but regarding study in Indian languages is at primary step only. Now a review of study prepared in English and Indian languages are presented. Initial structures are creating use of handcrafted through rule-based approach. But current structures most frequently use ML approach. For example, the important key is to a question answering is to classify the questioning point “who, what, when, where, etc., In various circumstances the questioning point matches to a Named Entity.

Ex: “Sachin Tendulkar was born on 24th April 1973 in Bombay”

Now “Sachin Tendulkar” is classified in Person entity, “Bombay” is classified in Location entity and “24th April 1973” is classified in Date entity. The NE system, in biology text records can mechanically extract the predefined classes from documents. While in excess of the years around has been significant work finished for Named Entity Recognition in the interest in Indian language has been fairly low until in recent times. Indian languages have large amount of morphologically equated to other foreign languages. Rule based approach presents well-improved performance and exactness if the set of comments is under control.

The Indian languages are morphologically having a great deal in structures and agglutinative in environment, from now rule-based methods may well be unsuccessful in circumstances where we want full-fledge Natural Language Processing based classification. To make completely developed NLP based classification we want great quantity of rules and language knowledge which is hard hence many researchers are currently employed on geometric and hybrid methodologies. In excess of the previous period of ten years Indian language relaxed on many media categories such as Mesh, E-mail, Blogs, and chats has improved expressively. A state of satisfaction growth is operated by persons as of non-metros and less than normal cities essentially need to perform this extremely large data records automatically. Specifically companies are excite to find something public opinion on their products also routine actions. This involves Natural Language Processing Software Systems (NLPSS) which classify entities, identification of relation among entities. In future an Automatic Named Entity Recognizer (ANER) is required.

II. RELATED FACTORS

As large number of papers related to NER was presented in the message understanding conference (MUC), some basic factors came up, which needed prime emphasis while working in this field.

A. Language Factor

A large amount of research work has been done in English. Most of the domains in English have been explored. But this has confined the work to the particular language only. Language independence and multilingual are the prime problems in this field. Languages like German, Spanish and Dutch have been studied in their own prospect. Chinese, Japanese, French, Italian, Greek, have been studied in abundant of literature. Survey on Bulgarian, Hindi, Danish, Korean, and Turkish are in progress. Even Arabic has started receiving attention. But this works continues to be in it own boundary. Developing a multilingual NER is of prime attention in the acurrent scenario.

B. Domain Factor

Initial work for NER ignored the two main factors of a language, firstly the type of text or textual genre like informal, scientific, technical etc and secondly domain like business, sports, tourism etc. The domain and the type of text collectively form the corpus to learn and test the system. But a single domain may have many corpuses within it. Hence a particular language will have large corpora of text to evaluate. Relating the corpora of one domain with another is a major challenge.

C. Entity and Tagging

The “Named entity” expression in Named entity recognition confines it to identify only those entities which are rigid and have some particular designation. For English these rigid designated entities are the proper nouns.

The most studied entity types are names of “person”, “location” and “organization”. These types were collectively called the “Enamex” in MUC-6. However, many papers describe that these types can be further divided to subtypes like person can be a doctor, celebrity, politician and locations can be country, state, heritage sites etc (M. Fleischman 2001, S. Lee & Geunbae Lee 2005). Also there are other entity types beyond the basic three such as time, money, occasions, accessories etc. These entities are collectively classified as miscellaneous type. All the named entities beyond enamex are put in this type. But the miscellaneous type includes a large number of entities collectively which is not efficient. An NER with large number of identification and tagging will be considered as an efficient one.

III.EXISTING METHODS USING NER

In paper[3], the authors describe a hybrid scheme that relates maximum entropy model (Max- Ent), language specific rules and gazetteers to the task of named entity recognition (NER) in Indian languages designed for the IJCNLP NERSSEAL shared task. Then some language specific rules are added to the system to recognize some specific NE classes. Also they had added some gazetteers and context patterns to the system to increase the performance. After preparing the one-level NER system, he applied a set of rules to identify the nested entities. The system is able to recognize 12 classes of NEs with 65.13% f-value in Hindi, 65.96% f-value in Bengali and 44.65%, 18.74%, and 35.47% f-value in Oriya, Telugu and Urdu respectively.

In paper[10], the authors have evaluate named entity extraction system. The suitability of the algorithms for recognition and classification of entities (NERC) is evaluated through competitions such as MUC, CONLL or ACE. In general, these competitions are limited to the recognition of predefined entity types in certain languages. In this paper a set of NERC tools are assessed. The study complements the information provided by other competitions and aids the choice or the design of more suitable NER tools for a specific project.

In paper[3], the researchers survey some NER approaches and its performance. Nowadays more researchers use different methods of NER, to identify names from text. In this paper, the author review these methods and compare them based on approach and also performance using the Message Understanding Conference (MUC) named entity definition and its standard data set to find their strength and weakness of each these methods. They proposed a robust and novel Machine Learning Based method called Fuzzy support Vector Machine (FSVM) for NER.

In paper[5], the author have briefly studied to exploring features for named entity recognition in Lithuanian text corpus. They attempt to solve a named entity recognition task for Lithuanian, using a supervised machine learning approach and exploring different sets of features in terms of orthographic and grammatical information, different windows, etc. Although the performance is significantly higher when language dependent features based on gazetteer lookup and automatic grammatical tools (part-of-speech tagger, lemmatizer or stemmer) are taken into account; they demonstrate that the performance does not degrade when features based on grammatical tools are replaced with affix information only. The best results (micro-averaged F-score=0.895) were obtained using all available features, but the results decreased by only 0.002 when features based on grammatical tools were omitted.

In paper[14], the authors discussed NER tweets. SMS tweets are particularly terse and difficult. This paper addresses this issue by re-building the NLP pipeline beginning with part-of-speech tagging, through chunking, to named-entity recognition. Their novel T-NER system doubles F1 score compared with the Stanford NER system. T-NER leverages the redundancy inherent in tweets to achieve this performance, using LabeledLDA to exploit Freebase dictionaries as a source of distant supervision. LabeledLDA outperforms cotraining, increasing F1 by 25% over ten common entity types.

IV. NER FOR INDIAN LANGUAGES

NLP investigation about the world takes a main opportunity using the arrival of real ML algorithms and the design of huge explained quantities. Not considerably efforts been prepared in NER for Indian Languages since explained quantities and additional lexical resources are not presented. Owing to deficiency of capitalization and deficiency of huge branded dataset, calibration and meaning variation, English NER cannot be straightly used for Indian languages.

Detection of NEs in raw information is not easy in Indian languages because Indian languages do not have capitalization. Indian languages have highly phonetic characteristics. Resources like gazetteers, dictionaries, POS taggers, morphological analyzers are not easily available. Lot of variations exists in spellings writing style. Work on NER in Indian languages is a difficult and challenging task and also limited due to scarcity of resources, but it has started to appear.

NER on Indian languages is difficult because of different writing methodologies, writing style variations, difficult morphology, little availability of annotated corpora and agglutinative nature like in Telugu. Many researchers have concluded that rule based approach for NER gives satisfactory results with sufficient gazetteers list and language independent rules. Rule based approach is not very easy for NER system development in Indian languages and therefore language independent NER system using hybrid models is needed. Srikanth and Narayana have developed CRF based noun tagger, trained on manually tagged data of 13,425 words and test dataset of size 6,223 words. 92% of F-Score have been given by name tagger. The Telugu NER based on ME by using news articles form Eenadu Vaartha newspaper and data from Telugu Wikipedia using the roman forms of the articles. The system is evaluated with and without using name list with ME and observed that ME using name list performs best for NER. NER system development preferred news articles because news is rich source of NEs of almost all categories. The work mentioned a useful technique to develop tagged Bengali news corpus from web for Bengali NER. Used phonetic matching algorithm Editex and Fuzzy string matching technique Soundex to recognize NEs in Hindi. The system has reported 81%precision. It is observed that large set of annotated data is yet to be available for Indian Languages. A novel NER approach which combines the global distributional characteristics with local context based on MEMM was presented in this research. A hybrid machine learning approach by using MaxEnt and HMM was presented by Biswas et.al. For NER in Oriya. 32 different rules were developed to identify numbers, measures and time. Gazetteers of specialized names were developed by translation into Oriya.

Language	Year	Technique/Algorithm	F1
Hindi	1999	Morphological & contextual clues[38]	41.70
	2003	CRF, Feature Induction[39]	71.50
	2006	MEMM[40]	79.70
	2009	Editex, Soundex[32]	65.69
	2009	MaxEnt[41]	82.66
	2010	CRF[42]	78.29
	2010	CLGIN[33]	82.90
	2010	SVM[43]	77.17
	2011	CRF, MaxEnt, Domain Rules[44]	80.82
	2011	SVM [44]	80.21
Bengali	2007	HMM[45]	83.79
	2008	CRF[46]	90.70
	2009	MaxEnt[41]	85.22
	2009	CRF[42]	81.15
	2010	SVM[43]	84.15
	2011	SVM[45]	83.39
Telugu	2008	CRF + Majority Tag[29]	84.49
	2010	MaxEnt[30]	67.07
	2011	Survey[28]	-
Oriya	2010	MaxEnt + HMM [34]	84.08
Punjabi	2011	condition based list lookup[47]	86.25
	2012	Domain rule, list look up[48]	85.88
Tamil	2008	CRF[49]	80.44
	2008	E-M(HMM)[50]	72.72
Urdu	2010	MaxEnt[51]	74.67
Nepali	2014	HMM + Rule based[25]	85.15

Bhattacharya et.al. developed hand-crafted rule based named entity recognizer for Marathi. The rules were constructed for extracting instances of NE classes using TILDE and WARMR techniques of inductive logic programming. TILDE is extension of traditional c4.5 decision tree learner to first order logic and WARMR is an extension of apriori algorithm to first order logic. Authors have used tagged data of 3,884 sentences in Marathi and 27,748 sentences in Hindi. NER system developed by using GATE (a framework and graphical development environment which enables users to develop and deploy language engineering components and resources in a robust fashion). Vasudev Verma et.al. proposed an approach to identify the NEs present in under resourced languages by utilizing the NEs present in English. Bisecting k-means algorithm is performed for clustering multilingual documents based on the identified NEs. Table 5 shows NER work done for some Indian languages, year of publication, techniques used and F-score obtained. Many NER systems observed here are implemented using more than one technique and evaluated with more than one dataset. F-score value in table 4 and 5 is the best reported performance of that respective system. Some NER systems reported more than one F-score, in case of such systems the average of F-scores is presented in this survey. The performance of NER system is measured by metrics precision, recall and F-measure. Precision measures how many of the tokens tagged are tagged correctly. Recall measures how many of the tokens are tagged are indeed tagged. F-Score is harmonic mean of precision and recall.

V. METHODOLOGIES OF NER

The NER approaches are

A. Rule Based Method

Rule based approach generally apprehensive with physical comments inscribed by the syntax. A lot of rules based NER encloses:

- Lexicalized Grammar
- Gazetteer list
- List of triggered words

In this approach for each individual classification of named entities different rules are given by the linguist. These rules are then implemented when the user enters the text. Whenever the system gets the text it first searches for the named entity and then compares it with the rules that have been used. Once the rule is matched, the system fetches the classification and gives the required output.

B. Statistical Approach or Machine Learning Based- Method

The statistical approach is quite different from the rule based approach. Where in the rule based approach the task of detecting and classifying the named entity solely depended on the rules given by the linguist, in the statistical approach mathematical logic and formulas are used for the same purpose.

Statistical approach differs from the traditional processing in that instead of having a linguist manually construct some model of linguistic phenomenon, the model is instead (semi-) automatically constructed from linguistically annotated resource. Methods for assigning parts of speech tags to words, categories to texts, parse trees to sentences, and so on, are (semi-) automatically acquired using machine learning techniques. In statistical approach a corpus is initially studied and based on the corpus a training module is made where the system is trained to identify the named entities and then on their occurrences in the corpus with particular context and class a probability value is counted. Every time when text is given based on the probability value the result is fetched.

Machine learning based method disturbed with a number of pre-defined approaches. These are:

- → Hidden Markov Models (HMM)
- → Decision Trees
- → Maximum Entropy Models (ME)
- → Support Vector Machines (SVM)
- → Conditional Random Fields (CRF)

C. Hybrid Method

It is a method where further than two methods are used in instruction to increase the presentation of the NER method. So the hybrid method can be a mixture of HMM model and CRF model or CRF and ME method or Gazetteer method with HMM method etc.

The two traditional approaches of Named Entity Recognition (NER) are: rule based approach and statistical or machine learning approach. The rule based approach achieves better accuracy but requires a large amount of labor by linguist and domain experts. Because of this, recent research in NER is concentrated in machine learning technique which requires only manually training set documents. Apart from these traditional approaches, the latest approach is the hybrid approach which combines both machine learning techniques and manually written rules. Hence, it benefits from both the approaches and can outperform manually written rules and machine learning.

VI. CHALLENGES IN NER

It is a technical challenge to build NER systems that performs exactly as that of human because of complex interrelations among various parts of sentence and the variety of languages (e.g. Hindi does not have capitalization clues). The challenges are:

- → Open nature of vocabulary
- → Clues such as capitalization
- → Overlap between NE Types
- → Indirect occurrences of NE
- → Different ways of referring to same entity

The effective NER systems use large amount of commonsense knowledge.

VII. NER FEATURES

Descriptors or characteristic attributes of words designed for algorithmic consumption are referred as features. The features are:

- → Word form and POS tags (if available)
- → Orthographic features: Like capitalization, decimal, digits
- → Word type patterns: Conjunction of types like capitalized, quote, functional etc.
- → Bag of words: Word forms, irrespective of position
- → Trigger words: Like New York City
- → Affixes Like Hyderabad, Rampur, Mehdipatnam, Lingampally
- → Gazetteer features: class in the gazetteer
- → Left and right context
- → Token length: Number of letters in a word
- → Previous history: Classes of preceding Named Entities

VIII. CONCLUSION

Entity Recognition (NER) is a unstable arena in present stage. NER attempts to extract a piece of writing and categorize the inflexible designators since extracted text. A lot of new studies are working on in this arena. In this research, an impression of the two simple methods to improve a NER has been defined. The statistical method contains fine points of certain existing methods which are used. Moreover the many learning methods to invent the system for this method are also talk over. As a final point the new method, the Hybrid approach that can be gained by joining the two simple methods has also been briefly stated.

NER will have a major influence on our community as it will allow us pull out more and more information starting a extracted text by classifying and organizing the NE which may be recognized or unidentified to us.

REFERENCES

[1] Asif Ekbal & Sivaji Bandyopadhyay, "Voted NER System using Appropriate Unlabeled Data", in 2009 Named Entities Workshop, ACL-IJCNLP, pp.202–210, Aug.2009.
[2] B. Sasidhar; P. M. Yohan; Dr. A. Vinaya Babu & Dr. A. Govardhan, "A Survey on Named Entity Recognition in Indian Languages with particular reference to Telugu" in IJCSI Vol.8, Issue.2, pp.438-443, Mar.2011.
[3] Darvinder kaur & Vishal Gupta, "A survey of Named Entity Recognition in English and other Indian Languages", in International Journal of Computer Science, Vol. 7, Issue 6, pp.239-245, Nov.2010.
[4] Faryal Jahangir; Waqas Anwar; Usama Ijaz Bajwa & Xuan Wang, "N-gram and Gazetteer List Based Named Entity Recognition for Urdu: A Scarce Resourced Language", in 10th Workshop on Asian Language Resources, pp. 95–104, Dec.2012.
[5] Harjinder Kaur & Dr. Vijay Laxmi, "A Web Based English to Punjabi MT System for News Headlines", in International Journal of Advanced Research in Computer Science and Software Engineering, Vol.3, Issue.6, pp.1092-1094, June 2013.
[6] Karthik Gali, Harshit Surana, Ashwini Vaidya, Praneeth Shishtla & Dipti Misra Sharma, "Aggregating Machine Learning and Rule Based Heuristics for Named Entity Recognition", pp.25–32, Jan 2008.
[7] Kashif Riaz, "Rule-based Named Entity Recognition in Urdu", in Association for Computational Linguistics, pp. 126–135, July 2010.
[8] Nusrat Jahan; Sudha Morwal & Deepti Chopra, "Named Entity Recognition in Indian Languages Using Gazetteer Method and Hidden Markov Model: A Hybrid Approach", in International Journal of Computer Science & Engineering Technology, Vol.3, Issue.12, pp.621-628, Dec.2012.
[9] Nita Patil; Ajay S. Patil & B. V. Pawar, "Survey of Named Entity Recognition Systems with respect to Indian and Foreign Languages", in International Journal of Computer Applications, Vol.134, Issue.16, pp.21-26, Jan.2016.
[10] N. Abinaya; M. Anand Kumar & K. P. Soman, "Randomized Kernel Approach for Named Entity Recognition in Tamil", in Indian Journal of Science and Technology, Vol.8, Issue.24, Spe.2015.
[11] Sujan Kumar Saha; Sanjay Chatterji; Sandipan Dandapat; Sudeshna Sarkar & Pabitra Mitra, "A Hybrid Approach for Named Entity Recognition in Indian Languages", in IJCNLP-08 Workshop, pp.17-24, Jan.2008.
[12] Sujan Kumar Saha, Sanjay Chatterji Sandipan Dandapat, Sudeshna Sarkar & Pabitra Mitra, "A Hybrid Approach for Named Entity Recognition in Indian Languages", in IJCNLP, pp. 17–24, January 2008.
[13] Thoudam Doren Singh, Kishorjit Nongmeikapam, Asif Ekbal & Sivaji Bandyopadhyay, "Named Entity Recognition for Manipuri Using Support Vector Machine", in 23rd Pacific Asia Conference on Language, Information and Computation, pages 811–818, 2009.
[14] Veerpal Kaur; Amandeep kaur Sarao & Jagtar Singh, "Hybrid Approach for Hindi to English Transliteration System for Proper Nouns", in IJCSIT Vol.5, Issue.5, pp.6361-6366, 2014.