Volume 3, Issue 11, November 2013 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Named Entity Recognition for Indian Languages: A Survey

Pallavi*
Research Scholar
Department of Computer Applications
Hindustan University Chennai, India

Dr. Anitha S Pillai
Professor and HOD
Department of Computer Applications
Hindustan University Chennai, India

Dr. Sobha L
Scientist
AUKBC Research Center MIT Campus Chennai, India

Abstract: ***Named Entity Recognition (NER) is a sub task of Information Extraction (IE) used to identify and classify the names in any given data. Earlier studies were mostly based on hand written rules where as now-a-days Machine Learning models such as Hidden Markov Model (HMM), Maximum Entropy (MaxEnt), Maximum Entropy Markov model (MEMM), Support Vector Machine (SVM), Conditional Random Fields (CRFs) are used to develop NER systems. In this paper we are presenting a survey of NER's for Indian Languages which has been developed using the above learning algorithms. Work on NER has been done only for few Indian Languages like Bengali, Hindi, Tamil, Telugu, Oriya, Punjabi, Urdu and Kannada among 22 official languages. In this paper, we discuss the strategies adopted and performance of these NER'S with respect to recall, precision and F-measure.***

Key Words: Named Entity Recognition (NER), Information Extraction (IE), Hidden Markov Model (HMM), Maximum Entropy (MaxEnt), Maximum Entropy Markov model (MEMM), Support Vector Machine (SVM), Conditional Random Fields (CRFs).

I. INTRODUCTION

Language is one of the fundamental aspects of human communication. The 'Natural Language Processing (NLP)' helps in enabling the human-machine communication [1]. The main goal of NLP is to build specific models that approach human performance in the linguistic tasks of reading, writing, hearing and speaking [4]. Information Extraction is one of the applications of NLP, which refers to the automatic extraction of structured information such as entities, relationships between entities and attributes describing entities from unstructured sources [12]. The Named Entity was introduced in the Sixth Message Understanding Conference (MUC-6) (R. Grishman & Sundheim 1996) [3]. NER is a sub task of IE used to identify and classify the names in any given data. In most cases hand-written rules are used to identify NE's. For classifications Naive Bayes, MaxEnt models and sequence models like HMM, MEMM, SVM, CRFs are used.

The role of NER systems is to locate and classify words in a text into predefined categories such as the names of persons, organizations, locations, expressions of times, quantities etc.

Eg: Ram went to Delhi

Ram and Delhi are named entities where NER need to classify Ram as <Name of a person> and Delhi as <Name of a place>

The NE's can be identified in two ways

1. The given sentence is compared with gazetteer's lists to identify the NE's, where the gazetteer's lists for person names, location names and organization names are created manually.
2. Writing linguistic rules eg: nouns always start with capital letter.

It is challenging to implement NER for Indian languages due to non-capitalization forms for nouns and Indian languages have rich morphology compared to English. Preparing a gazetteer's list for all nouns is impossible. A lot of work has been done in English and other foreign languages like Chinese, Japanese, Korean, Arabic, and Spanish with high accuracy. But not much work has been reported for NER in Indian languages due to insufficient resources and also morphological nature. It is highly challenging to solve linguistic problems for languages with rich morphology. The second section presents the Related work done; third section gives the performance evaluation of those systems with reference to precision, recall and f-measure; and fourth is the Conclusion.

II. RELATED WORK

Saha et al., 2008, [10] developed '*A Hybrid Feature Set based Maximum Entropy (MaxEnt) Hindi Named Entity Recognition*'. The four NE's identified by this system are Person names (P), Location names (L), Organization names (O) and Date (D). Each NE is divided in to 4 classes as begin, continue, end and unique. Collectively 16 NE classes are

created for those NE's and another non-entity class is added to identify the entities which do not belong to any of the above.

MaxEnt, a supervised machine learning technique is applied to solve linguistic problems with the help of orthographic, collocation features and Gazetteers lists. Initially they transliterated already existing English gazetteer's lists to Hindi using the 2-phase transliteration module. This module contains 34 intermediate characters to transliterate English names to Hindi. Similarly Hindi names are transliterated to English. By comparing both Hindi and English transliterated names, the eight entity groups are constructed: month name, days of the week, organization, end word list, person prefix words list, list of common locations, location names list, first names list, middle names list, and surname.

For classification, a semi-automatic induction of context patterns is derived. For each individual NE class, seed entities are collected. Finally the patterns between the seed entities and collected gazetteer's lists are compared. The patterns having low precision are discarded. Other patterns with high precision are generalized by dropping one or more tokens to increase coverage.

A 75.6 f-measure is achieved as baseline result and 81.52 f-measure is achieved after adding gazetteer lists and context patterns into MaxEnt based NER system.

Sujan Kumar Saha et al., 2008, [11] '*A Hybrid Approach for Named Entity Recognition in Indian Languages'*, is implemented for 5 Indian languages - Hindi, Bengali, Oriya, Telugu and Urdu.

Two approaches used here are the Linguistic approach where the typical rules written by linguists and the Machine Learning (ML) approach – Where the system is trained using tags.

Features identified in this NER system are: 1. Static word feature, 2. Context list, 3. Dynamic NE tag, 4. First word, 5. Contains digits, 6. Numerical word, 7. Word suffix, 8. Word prefix, 9. Root information of a word and Parts of speech information. Parts Of Speech(POS) tagging used the Coarse-grained tag set with only three tags - nominal (Nom), postposition (PSP) and other (O). They also worked on Nested entities and Nominal Postpositions (NomSPS).

This paper reported that they received poor accuracy for Oriya, Telugu and Urdu languages compared to the other two languages due to lack of POS information, morphological information, language specific rules and gazetteers lists. Finally, the system was 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.

Ekbal et al., 2008, [2] has worked on **'***Named Entity Recognition in Bengali: A Conditional Random Field Approach'.* CRF is introduced by Lafferty et al., 2001 [5] to build a probabilistic model to segment and label the sequence data. It includes the ability to prove strong independence assumptions over HMM and stochastic grammars. To develop NER for Bengali, they used rule based with CRF approach. They defined conditional probability of a state sequence and named it as the OpenNLP CRF++ to classify the NEs. It was implemented using C++. They achieved f-measure of 90.7 using CRF, more than the HMM result.

Karthik Gail et Al., 2008, [7] NERSSEAL-2008 promulgated a paper on '*Aggregating Machine Learning and Rule Based Heuristics for Named Entity Recognition'*, screened for Indian languages. This paper discusses the ambiguities for Indian languages that deals with the linguistic issues like

- Agglutinative nature and absence of capitalization
- Same meaning for common name and proper name
- Low parts of speech tagging accuracy for nouns
- Spelling variation
- Patterns and suffixes

In this paper, they used CRF's to

- Perform statistical tagging
- Resolve capitalization issue
- Find five preceding words
- Collect lists of suffixes for NE-Persons and NE-Locations
- Collect prefixes such as noun inflections
- Find actual NE-Number
- Find presence of digits and presence of four digits that is year

The Rule Based Heuristics post processing is used to determine the second best tag resulted by the CRF model and deals with nested entities like NE-Number and NE-Measure with the help of Gazetteers list.

They achieved *f - measures* 40.63, 50.06, 39.04, 40.94, and 43.46 for Bengali, Hindi, Oriya, Telugu and Urdu respectively.

Kashif Riaz, 2011, [8] worked on '*Rule-based Named Entity Recognition in Urdu'*. It proves through few examples that the Hindi NER cannot be applied for Urdu. This is due to the differences in the vocabulary and writing styles of Hindi

and Urdu. Since the same NER could not be applied, they derived hand crafted rule based algorithms to identify Named Entities for Urdu.

The rules are constructed to identify the entities similar to Hindi NER [10] with the addition of two more entities – Numbers and Person of influence. The rules are derived for all NE's using below mentioned rule patterns: 1. Punctuation marks, 2.Title of news, 3.Stemming and Suffix rules for locations and organizations, 4. String of names without any prefix or suffix cues, 5.Multiple spellings for same NE, 6. Transliteration problems, 7.Anchor texts based corpuses, 8. Patterns and heuristic grammars.

The Becker-Riaz corpus of 2,262 documents has been cleaned by applying n-gram model to remove XML tags and unwanted contents for readability. A Gazetteers list for locations names is created similar to Hindi NER [10]. The classification has been done by comparing identified NE's with those of Gazetteers lists. An algorithm was implemented using the rules and rule pattern to find NE's.

They ran rule set on the 36,000 token Urdu data provided for IJCNLP 2008 NER Workshop. Without modifying the database, the $f1 - measure$ was 72.4% and after adding a few rules to train the database, the $f1 - measure$ increased to 81.6% on the test set.

Amarappa and Sathyanarayana, 2012, [1] came up with a paper on '*Named Entity Recognition and Classification (NERC) in Kannada language'*, that built a SEMI-Automatic Statistical Machine Learning NLP models based on noun taggers using HMM.

The challenges and issues faced for Kannada language are listed by them are

1. No capitalization
2. High phonetic characteristic of Brahmi script.
3. Non-availability of large gazetteer lists
4. Lack of standardization and spelling
5. Number of frequently used words (common nouns).

Their proposed NERC system for Kannada receives the unannotated text file containing the Kannada document, recognizes the NE's and generates an annotated text document file. Further the output of NERC system is subjected to a suitable cryptographic algorithm to secure the structured corpus.

They came up with 13 noun taggers for NER like person name (NNP), location name (NNL), organization name (NNO), etc. Hidden Markov Model (HMM) is a supervised learning technique and a statistical model with generalized learning method. It is used to develop a NER with symbolic, statistical, connectionist and hybrid natures. NE's and NE Tags are defined with examples in this paper. These tag sets are used to tag each word in the sentence. Python and python Natural Language ToolKit (NLTK) are used to build the NERC models. NLTK has built in functions for performing the basic tasks on the raw text, such as sentence segmentation, tokenization etc. They gained good accuracy for Kannada NERC.

Malarkodi, et al., 2012, [9] experimented their NER on Tamil database, coping with real time challenges using CRF's, which can apply for most of the Indian languages. Labeling task will help to identify the proper nouns in a database. Each word in a database is tagged using POS tag set. Naturally all NER system will help to improve the performance of NLP applications like Machine Translation (MT), IE, Cross Lingual information access system, etc.

They used a standard tagset. Here totally 106 tags are used to tag entities of 3 different categories. Such as entity names, numerical and time expressions. There are 11 entity names, 4 numerical expressions and 7 time expressions. The NE features like suffix; prefix differs from language to language. Dictionary patterns and gazetteer lists vary for every language and it contains only the frequent named entities. POS tag identifies the proper and common nouns, cardinal numbers and also relationship between the current preceding and succeeding words. The named entities are identified in a sentence using phrase chunking.

Tourism and general corpus is collected and divided into training and testing sets in a ratio of 70%:30%. POS and chunking is done using automated tools and that are trained using CRF's on already specified features.

NE's of this system are identified and the results are evaluated manually. Overall they achieved an *f-measure* of 60.36. They have applied the same features to classify NE using SVM model, to compare the results with the CRF's and they have achieved *f-measure* of 58.34%.

Kaur and Vishal Gupta, 2012, [6] built a 'NER for Punjabi' using rule based and list look up approaches. As mentioned earlier, Punjabi is also a language with high clung and inflections, which leads to linguistic problems.

The rule based approach trained the system to identify NEs by writing rules manually for all NE features. The most common words are removed from the database, and then a list look up approach is used with the Gazetteer's lists to classify the identified NEs. Their system resulted with 85.88% *f-measure*.

III. PERFORMANCE EVALUATION TECHNIQUES

All the above papers used precision, recall and f-measure to measure the accuracy of results.

- Recall = Named Entities identified by the system / Total Number of Named entities
- Precision = Named entities identified correctly / Total number of named entities identified
- F-Measure = Harmonic mean of precision and recall, the f-measure.

TABLE 1

SL.NO	LANGUAGE	NO OF TAG SETS USED	TESTED DATA IN NO OF WORDS		F-MEASURE	MODELS USED
			TRAINING SET	TESTING SET		
1	HINDI	17	243K	25K	81.52	MAXENT
2	BENGALI	17	130K	20K	90.7	CRF
3	URDU		36k	36K	81.6	RULE BASED
4	TELUGU	NOUN IDENTIFIERS	13.4K	6.2K	80-97	CRF
5	TAMIL	22	65022	28270	60.36 58.34	CRF SVM
6	PUNJABI	NO TAGS IDENTIFIED			85.88	RULE BASED AND LIST LOOK UP APPROACH
7	KANNADA	13			HIGH ACCURACY	HMM

IV. CONCLUSION

This paper, gives a survey of NER systems that has been developed for some of the Indian Languages like Hindi, Oriya, Punjabi, Marathi, Urdu, Telugu, Tamil, and Kannada. There are several rule based NER systems, containing mainly lexicalized grammar, gazetteer lists, and list of tagged words [9]. Among those Telugu achieved a good *f-measure* using CRF approach. Hybrid approach results good accuracy for bilingual NERs [11].

A hybrid approach that combines rule based and machine learning approaches will give a better result for NER system.

REFERENCES:

[1] Amarappa; Dr. S V Sathyanarayana. 2012. "*Named Entity Recognition and Classification in Kannada Language. International Journal of Electronics and Computer Science Engineering".*

[2] Asif Ekbal; Rejwanul Haque; Sivaji Bandyopadhyay. 2008. "*Named Entity Recognition in Bengali: A Conditional Random Field Approach".*

[3] Grishman, Beth Sundheim. 1996. *Message Understanding Conference-6: "A Brief History". In the proceedings of the 16th International Conference on Computational Linguistics (COLING), pages 466-471, Center for Sprogteknologi, Copenhagen, Denmark*

[4] James Allen. 1995. "Natural Language Understanding". *Pearson Education, Inc.*

[5] John Lafferty; Andrew McCallum; Fernando Pereira. *2001 "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data"*

[6] Kamaldeep Kaur; Vishal Gupta. "*Name Entity Recognition for Punjabi Language". International Journal of Computer Science and Information Technology & Security (IJCSITS), ISSN: 2249-9555 Vol. 2, No.3, June 2012.*

[7] Karthik Gali; Harshit Surana; Ashwini Vaidya; Praneeth Shishtla and Dipti Misra Sharma. "*Aggregating Machine Learning and Rule Based Heuristics for Named Entity Recognition", In Proceedings of IJCNLP-08 workshop IIIT Hyderabad, India, January 2008, pp. 25-31.*

[8] Kashif Riaz. 2010. "*Rule-based Named Entity Recognition in Urdu". Proceedings of the 2010 Named Entities Workshop, ACL 2010, pages 126–135, Uppsala, Sweden.*

[9] Malarkodi, C S; Pattabhi; RK Rao and Sobha; Lalitha Devi.*2012 "Tamil NER – Coping with Real Time Challenges". Proceedings of the Workshop on Machine Translation and Parsing in Indian Languages (MTPIL-2012), pages 23–38, COLING 2012, Mumbai, December 2012.*

[10] S.K.Saha; S. Sarkar; P. Mitra. 2008. "*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.*

[11] Sujan Kumar Saha; Sanjay Chatterji; Sandipan Dandapat; Sudeshna Sarkar; Pabitra Mitra**. "***A Hybrid Approach for Named Entity Recognition in Indian Languages", In Proceedings of IJCNLP-08 workshop IIIT Hyderabad, India, January 2008, pp. 17-24.*

[12] Sunita Sarawagi. 2008. *"Information Extraction".* "*Indian Institute of Technology, CSE, Mumbai 400076, India.*

