Approaches to Named Entity Recognition in Indian Languages: A Study

Prakash Hiremath, Shambhavi B. R

***Abstract*— Named Entity Recognition (NER) is subtask of information extraction that seeks to locate and classify the elements in some text into pre-defined categories. NER finds its application in Natural Language Processing tasks like machine translation, question-answering systems and automatic summarization. The approaches to NER are rule based, statistics based or a combination of both. In this paper, we present a survey of these various approaches for identification of Names Entities (NE) in Indian Languages.**

***Index Terms*— Named Entity Recognition (NER), Natural Language Processing, Machine Learning**

I. INTRODUCTION

Named Entities are atomic elements in text belonging to pre-defined category such as name of a person, organization, location etc. Named Entity Recognition is the task of identifying such Named Entities. According to the Message Understanding Conference (MUC) initiated by DARPA [1], the NER generally work on common entity types like organization, person, location, person, date, time, measurement and number. For example consider a sentence: Rajesh joined BMSCE as a PG student in Bangalore on 14th April 2014. Here 'Rajesh' is PERSON entity, 'BMSCE' is ORGANIZATION entity, 'Bangalore' is a LOCATION entity and '14th April 2014' is DATE entity. Identification of Named Entities is a very important in NLP applications like question-answering systems and automatic summarization, machine translation system and information extraction system. For the last few years the Indian language content on different media types have raised by many folds due to the availability and usability of internet at fingertip. The content growth is driven by people from non-metros, small and semi-urban cities. The automatic processing of these huge data plays major role for companies in understanding the public view on their product and processes. This requires Natural Language Processing (NLP) systems which identify the entities and relation between them. Hence NER is necessary. The paper is organized as follows. The next section presents an overview of NER. Section 3 surveys on NER work done in English. In English, Capitalization is the major key for identification of NEs and lot of work has been done in this field. A survey on NER for Indian languages based on diverse approaches and the comparative study of the F-measure is presented in section 4.

Manuscript published on 30 August 2014.
***** Correspondence Author (s)

Prakash Hiremath, M.Tech Student, Department of Information Science and Engineering, B. M. S. College of Engineering, Bangalore, India.
Dr. Shambhavi B. R, Assoc. Prof., Department of Information Science and Engineering, B. M. S. College of Engineering, Bangalore, India.

II. OVERVIEW OF NER

The approaches to NER are namely Rule Based Learning, Machine Based Learning or a hybrid of both. Rule-based systems typically obtain better precision, but at the cost of lower recall and months of work by experienced computational linguists. Machine Learning includes Hidden Markov Model [2], Conditional Random Fields [3], Maximum Entropy [4], Decision Tree, and Support Vector Machines [5]. Modern systems often follow machine learning approaches.

Performance Evaluation Metrics:

- Precision (P): Precision is the fraction of the correct tags generated by the NER to the total number of tags generated.

Precision (P) =Correct answers/answers produced

- Recall(R): Recall is the fraction of the correct tags generated by the NER to the total number of correct tags.

Recall (R) = correct answers/ total possible correct answers.

- F-Score: F-score is the weighted harmonic mean of precision and recall.

F-Measure = $(\beta^2 +1)PR/(\beta^2R+P)$.

β is the weighting between precision and recall and typically β=1.

III. PREVIOUS WORK FOR ENGLISH

Lot of work has been done on NER for English employing machine leaning techniques, using both supervised learning and unsupervised learning. In 1995 Ralph Grishman developed a rule based NER using some specialized name dictionaries including names of all countries , names of major cities, companies, common first names[6] with a result of 86 % recall, 90% precision and 88.19 % F-measure. In 1999, Bothwick developed a ML based system [7] i.e. MaxEnt. It used 8 dictionaries. In 2001 Fleiscman [8] proposed a method for categorization of location names using Baysian and Decision tree and the accuracy is 80%. In 2003 Hsinhsi Chen [9] proposed an algorithm for Named Entity for information retrieval, various types of information from various levels of text are employed, including character conditions, static information. The recall and precision rate for the extraction of names was (87.33%, 82.33%), for organization name it was (76.67%, 79.33%) and for location name it was (77%, 82%). In 2002 Collin proposed ranking algorithm for Named Entity Extraction using Maximum entropy and reports 84% precision, 86% recall and 85% F-measure.

IV. NER FOR INDIAN LANGUAGES

NLP research around the world has taken major turn with the advent of effective machine learning algorithms and the creation of large annotated corpora. Not much work has been done in NER for Indian Languages because annotated corpora and other lexical resources are not available. Due to absence of capitalization and lack of large labeled dataset, standardization and spelling variation, English NER cannot be directly used for Indian languages. Also the ambiguities for Indian languages that deals with the linguistic issues like

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

Shared tasks have been organized by NLP associations with an intention to get researchers and developers to work on a particular problem and come up with the best systems. These contest like events foster advancing the state of art in the area. The first workshop in NER was conducted for five Indic languages (Hindi, Bengali, Oriya, Telugu and Urdu) by IJCNLP 2008. Recently shared task for NER for Indian Languages was held in conjunction with FIRE 2013. The registered teams submitted runs for English, Hindi, Tamil, Malayalam and Bengali.

A. RULE BASED APPROACH

Kaur and Vishal Gupta built a NER for Punjabi [10] using Rule based and list look up approaches with a result of 85.88% F-measure. Riaz proposed a rule based approach for Urdu using small scale gazetteers [11] and reports 90% recall, F-measure 93.14% and 96.4 precision. Bhuvaneshwari C Melinamath has proposed Rule based approach of NER for Kannada [12]. The designed system takes the raw corpus which is converted into transliterated corpus. Then the text file is divided into sequence of tokens and removes the delimiters like [.,?]. Also they have used rules to identify abbreviations in the text. Each token is searched in dictionary. The methodology follow the sequence of translate, tokenize, search dictionary and identify entity. Famous Kannada newspaper Prajavani corpus tool is used to carry out experiments. The report shows 86% precision and 90% recall.

B. HIDDEN MARKOV MODEL

Hidden Markov Model is a statistical model in which the system being modeled is assumed to be a Markov process with unobserved state. Each state has a probability distribution over the possible output tokens. Therefore the sequence of tokens generated by an HMM gives some information about the sequence of states. Amarappa and Satyanarayana published a paper “Named Entity Recognition and Classification in Kannada language” [13], which built a SEMI-Automatic Statistical Machine Learning NLP models based on noun taggers using Hidden Markov Model (HMM). This proposed NERC system for Kannada takes the unannotated Kannada text file as an input, recognizes the NE’s and generates an annotated text document file. The structured corpus is made secure by subjecting the output of NERC to a suitable cryptographic algorithm. They came up with 13 noun taggers. In this paper NE’s and NE tags are defined with examples. The tag sets are used to tag each word in the sentence. NERC is built using Python and Python NLTK. NLTK has built in functions for performing the basic operation on the text such as sentence segmentation, tokenization etc. This Kannada NERC has attained good accuracy.

C. CONDITIONAL RANDOM FIELD

Conditional Random Field is a probabilistic framework for labeling and segmenting structured data, such as sequences, trees and lattices. The underlying idea is that of defining a conditional probability distribution over label sequences given a particular observation sequence, rather than a joint distribution over both label and observation sequences. The primary advantage of CRFs over hidden Markov models is their conditional nature, resulting in the relaxation of the independence assumptions required by HMMs in order to ensure tractable inference. Praneeth, et. al[14] has developed an NER system for Telugu using CRF. The evaluation has reported F-score of 44.91 %. Goyal has proposed NER for Hindi [15] using CRF in 2008. Evaluation has reported F-measure of 58.85% on development test. Ekbal proposed NER for Bengali [16] using CRF the evaluation has reported f-score of 90.7 %. Amandeep, Gurpreet, Jagroop proposed an NER using for Punjabi [17]. Evaluation has reported f-score of 80.92%. Malarkodi, et al., 2012, [18] experimented their NER on Tamil database, coping with real time challenges using CRF. They achieved an f-measure of 60.36 %. VijayKrishna and Sobha [19] developed a domain specific Tamil NER for tourism. It handles morphological inflection and nested tagging of named entities with a hierarchical tag set consisting of 106 tags. The system comes out with an F-measure of 80.44%.

D. SUPPORT VECTOR MACHINE

Support Vector Machine is supervised learning model with associated learning algorithms that analyze data and recognize patterns, used for classification and regression analysis. Given a set of training examples, each marked as belonging to one of two categories, an SVM training algorithm builds a model that assigns new examples into one category or the other, making it a non-probabilistic binary linear classifier. In 2008, Ekbal and Bandopadhyay [20] again developed another NER system for Bengali using SVM approach. A partially NE tagged Bengali news corpus has been used to create the training set for the experiment and the training set consists of 150K word forms that is manually tagged with 17 tags. The evaluation reported an F-measure of 91.8%.

E. MAXIMUM ENTROPY

Maximum Entropy Markov Model 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. Each source state has a exponential model that takes the observation feature as input and output a distribution over possible next state. Output labels are associated with states. In 2006 Kumar and Bhattacharyya proposed an NER system for Hindi [21]. They achieved an F-score of 79.7% using the Maximum Entropy Markov Model.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

Raju et.al [22] have developed a Telugu NER system by using ME approach. The system makes use of the different contextual information of the words and Gazetteer list was also prepared manually or semi-automatically from the corpus and came out with a an F-measure of 72.07%, 6.76%, 68.40% and 45.28% for person, organization, location and others respectively.

F. HYBRID APPROACH

In **Hybrid approach** both Machine Learning and Rule based systems are combined to achieve high accuracy. Shilpi Srivatsava proposed a hybrid approach [23] for Hindi, a combination of rule based CRF and Maximum Entropy for named entity recognition system for Hindi with a result of 96% precision, 86.96 recall and 91% F-measure. Saha et al., [24] developed "A Hybrid Feature Set based Maximum Entropy (MaxEnt) Hindi Named Entity Recognition". The four NEs identified by this system are Person names (P), Location names (L), Organization names (O) and Date (D). 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 ME based NER system. Pandian et.al [25] presented a hybrid three-stage approach for Tamil NER. The E-M(HMM) algorithm is used to identify the best sequence for the first two phases and then modified to resolve the free-word order problem. Both NER tags and POS tags are used as the hidden variables in the algorithm. Finally the system comes out with an F-measure of about 72.72% for various entity types. P Srikanth and K. Narayana Murthy in 2008 proposed "NER for Telugu" [26] a CRF approach with rule based NER System. Named Entity tagged corpus of 72,157 words has been developed using the rule based tagger through bootstrapping and features applied on three NE classes person, place and organization. The resulted f-measure lies between 80 and 97%. Table.1 shows the NER work done on Indian languages using different models. NER systems which are built using Rule based and Hybrid approach show better accuracy than other models.

TABLE I. RESULTS OF NER MODELS FOR DIFFERENT INDIAN LANGAUGES

Author	**Models used**	**Language**	**F-measure (%)**
[14]	CRF	Telugu	44.91%
[15]		Hindi	58.85%
[16]		Bengali	90.7%
[17]		Punjabi	80.92%
[18]		Tamil	60.36%
[19]		Tamil	80.44%
[20]	SVM	Bengali	91.8%
[13]	HMM	Kannada	High accuracy
[21]	MEMM	Hindi	79.17%
[10]	Rule based	Punjabi	85.88%
[11]	Rule based	Kannada	86%
[12]	Rule based	Urdu	93.14%
[23]	CRF + ME + Rule	Hindi	91
[25]	Rule + HMM	Tamil	72.72
[26]	Rule + CRM	Telugu	80-97

V. CONCLUSION

Applications of Natural Language Processing are many like machine translation, text processing, information retrieval, speech recognition and so on. Named Entity Recognition is a critical task in all of these NLP applications. Literature survey reveals that hybrid models which combine both rules and a machine learning algorithm perform better for Indian languages. It can be also be seen that the work already done or being done in NER for Indian languages is very small compared to what all needs to be done.

REFERENCES

[1] Charles L. Wayne. 1991., "A snapshot of two DARPA speech and Natural Language Programs" in the proceedings of workshop on Speech and Natural Languages, pages 103-404, Pacific Grove, California. Association for Computational Linguistics.
[2] B. D. M, M. Scott, S. Richard, and W. Ralph, "A High Performance Learning Name-finder," in Proceedings of the fifth Conference on Applied Natural language Processing, 1997, pp. 194–201.
[3] J. Lafferty, A.McCallum, and F. Pereira, "Probabilistic Models for Segmenting and Labelling Sequence Data, "in Proceedings of the Eighteenth International Conference on Machine Learning (ICML-2001), 2001.
[4] B. Andrew, "A Maximum Entropy Approach to NER," Ph.D. dissertation, 1999.
[5] Cortes and Vapnik, "Support Vector Network, Machine Learning," 1995, pp. 273–297.
[6] R. Grishman. 1995. "The NYU system for MUC-6 or Where's the Syntax" in the proceedings of Sixth Message Understanding Conference (MUC-6) , pages 167-195, Fairfax, Virginia.
[7] Andrew Borthwick. 1999. "Maximum Entropy Approach to Named Entity Recognition" Ph.D. thesis, New York University.
[8] Michael Fleischman, "Automated sub categorization of named entities". Proc. Conference of the European Chapter of Association for Computational Linguistic, pp 25–30, 2001.
[9] Yungwei ding hsinhsi Chen and Shihchung Tsal, "Named entity extraction for information retrieval". Proc. of HLT-NAACL.
[10] 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.
[11] Riaz K. Rule-based named entity recognition in Urdu. In Proceedings of the Named Entities Workshop. pages 126-135.2010
[12] Bhuvaneshwari C Melinamath." Rule based Methodology for Recognition of Kannada Named Entities", (IJLTET) Vol. 3 ISSN: 2278-621. March 2014.
[13] Amarappa, Dr. S V Sathyanarayana. 2012. "Named Entity Recognition and Classification in Kannada Language". International Journal of Electronics and Computer Science Engineering.
[14] Praneeth M Shishtla, Prasad Pingali, and Vasudeva Varma 2008 "ACharacter n-gram Based Approach for Improved Recall in Indian Language NER s" Proceedings of the IJNLP-08 Workshop on NER for South and South East Asian Languages Hyderabad, India.
[15] A. Goyal, "Named Entity Recognition for South Asian Languages," in Proceedings ofthe IJCNLP-08 Workshop on NER for South andSouth- East Asian Languages, Hyderabad, India, Jan 2008, pp. 89–96.
[16] Ekbal and S. Bandyopadhyay, "Named entity recognition in Bengali: A Conditional random field". Proc. ICON, pp 123–128, 2008.
[17] Amandeep Kaur, Gurpreet Singh Josan and Jagroop Kaur. 2009 Named Entity Recognition for Punjabi: A Conditional Random Field Approach. In Proceedings of 7th international conference on Natural Language ProcessingICON-09. Macmillan Publishers, India.

[18] 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.

[19] Vijayakrishna R and Sobha L, “Domain focused Named Entity Recognizer for Tamil using Conditional Random Fields,” in Proceedings of the IJCNLP-08 Workshop on NER for South and South East Asian languages, Hyderabad, India, 2008, pp. 59–66.

[20] Asif Ekbal, Sivaji Bandyopadhyay. “Bengali Named Entity Recognition using Support Vector Machine” in the proceedings of the IJCNLP-08 workshop on NER for South and South East Asian Languages, pages 51-58, Hyderabad, India.

[21] Kumar N. and Bhattacharyya Pushpak. 2006. “Named Entity Recognition in Hindi using MEMM” in the proceedings of Technical Report, IIT Bombay, India.

[22] G. Raju, B.Srinivasu, D. S. V. Raju, and K. Kumar, “Named Entity Recognition for Telegu using Maximum Entropy Model,” Journal of Theoretical and Applied Information Technology, vol. 3, pp. 125–130, 2010.

[23] Mukund Sangalikar, Shilpi Srivatsava and D.C. Kothari. ”Named entity recognition System for Hindi language”. International journal of Computational Linguistics Volume (2), pp 10–23.

[24] Sujan Kumar Saha, Sanjay Chatterji, Sandipan Dantapat,Sudeshna Sarkar and Pabitra Mitra 2008 “A Hybrid Approach for Named Entity Recognition in Indian Languages” Proceedings of the IJNLP-08 Sorkshop on Ner for South and South East Asian Languages Hyderabad, India.

[25] S.Pandian, K.A.Pavithra and T.Geetha, “Hybrid Three-stage Named Entity Recognizer for Tamil,” INFOS2008, March 2008.

[26] P Srikanth and Kavi Narayana Murthy 2008 “Named Entity Recognition for Telugu” Proceedings of the IJNLP-08 Workshop on NER for South and South East Asian Languages Hyderabad, India.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)