IOSR Journal of Engineering (IOSRJEN) **www.iosrjen.org**
ISSN (e): 2250-3021, ISSN (p): 2278-8719
Vol. 08, Issue 6 (June. 2018), ||V (III) || PP 71-75

A Survey on Named Entity Recognition of the Indian Languages with special reference to Manipuri

Priyadarshini Lamabam[1]
[1]*(Computer Science Department ,Kendriya Vidhyalaya No.1 Imphal,India)*

Abstract: - Named Entity Recognition(NER) is a subdivision of information extraction that helps to identify the various elements in text into certain categories. It plays an important role in the applications of Natural Language Processing like Machine Translation, Part of Speech tagging, Information Retrieval, Question Answering etc. There has been numerous works on NER of English and other foreign languages. Compared to them, there has been lesser works on NER of the Indian languages and the work is more limited when it comes to a regional language like Manipuri. After a survey on the various papers, the type of approaches used in NER and how they have been adopted in the Indian languages with special reference to Manipuri has been discussed in this paper.

Keywords: - Name entity recognition, natural language processing, rule-based, machine learning, hybrid approach

Date of Submission: 01-06-2018 Date of acceptance: 16-06-2018

I. INTRODUCTION

Human-machine interaction is one of the most important form of communication that helps in the automation of the various tasks. The Natural Language Processing (NLP,)an application of computer science helps in enabling it.There has been growing interest in this field of research since the early 1990s. Building of specific models that approach human performance in the linguistic tasks of reading, writing, hearing and speaking [1]has been described. Named Entity Recognition refers to the automatic extraction of structured information such as names of persons, organizations, locations, expressions of times, quantities, monetary values, percentages from unstructured sources like newspapers,articles,social media.
Most NER system work by taking a block of text, eg:

Ronaldo has been working in Delhi University since 1990
And an annotated block of text is produced by the system highlighting the names of entities:
[Ronaldo] Person has been working in [Delhi University]Organisation since [1990]Time

Early in the 1990s NER systems aimed primarily at extraction from formal documents like journalistic articles. At the later stages name entity extraction from several types of informal text , such as weblogs and text transcripts from conversational telephone speech conversations began.

Now,coming to the research work on NER,there are various approaches for implementing it. They are namely- Rule Based that uses hand-written rules, Machine Learning based which includes HMM(Hidden Markov Model), Maximum Entropy, Decision Tree, Support Vector Machines and Conditional Random Fields and Hybrid Approach(that uses various form of combination of machine learning algorithms).

Handcrafted rule-based systems usually give good results, but they need months for the development of high number of rules by experienced linguists so it may fail in those situations where a full-fledge NLP based system is needed in a short time. So the researchers are now working on statistical and hybrid approaches. Machine learning approaches require a large annotated corpora to train the classifier models for the various Name entity classes.

A lot of work for NER has been done for English,most of the European languages and some of the Asian languages like Chinese, Japanese and Korean with high accuracy. This is because of the presence of capitalization features and large labelled corpus. But not much work has been reported for NER in Indian languages due to various reasons like rich morphology,lack of capitalization,lack of standardization and spelling variation, scarcity of tools and resources.English NER cannot be directly applied to the Indian languages so it becomes highly challenging to solve linguistic problems for languages like this.

Manipuri, one of the scheduled languages of the Indian Constitution,is a Tibeto-Burman and highly agglutinative, monosyllabic(where a single letter is meaningful) and compounding (where words are formed easily) . Considering all these, there are some challenges occurred during the processing of Manipuri Language like the other Indian languages.

- Lack of capitalization features which is quite an important feature of NER.
- A lot of NEs in Manipuri can appear in the dictionary with some other specific meanings.
- Manipuri is a highly inflectional language providing one of the richest and most challenging sets of linguistic and statistical features resulting in long and complex wordforms.
- Manipuri is a relatively free word order language. Thus NEs can appear in subject and object positions making the NER task more difficult compared to others and in order to make a language machine readable, there should be enough tools, operating system or applications such as processor, translator, compiler, and encoding, decoding support in order to make the particular language computerized.
- Manipuri is a resource-constrained language. Resources like annotated corpus, name dictionaries, sophisticated morphological analyzers, POS taggers etc. are not yet available.
- The word categories in Manipuri language are not well defined. Grammatically, in some cases there is a difference between structural and contextual meaning of a particular word.
-

II. RELATED WORK

In paper [2], the authors have discussed and represented the various approaches used for NER of Indian languages.After a comprehensive study they concluded that the hybrid models which combines both rule-based technique and machine learning algorithms perform better for Indian languages due to their vast difference from foreign languages.

In paper [3], authors have investigated the various features and machine learning algorithms for NER in Inflectional Language. They made a corpus collection of 1000 sentences from the different domains of an inflectional language. One third of 1000 sentences were used to train the different NER systems and one third for testing. The different types of features including bag- o-characters, word segmentation, part-of-speech, and section information, and different machine learning algorithms including conditional random fields(CRF), support vector machines(SVM), maximum entropy(ME), and structural SVM(SSVM) were used. All classifiers were trained on the training dataset and precision, recall, and F-measure were evaluated.

The authors developed an NER for Hindi language [4] that identifies the named entities of Person names (P), Location names (L), Organization names (O) and Date (D). Each Named entity(NE) has a subdivision of classes like begin, continue, end and unique resulting in 16 NE classes and another non-entity class to identify the entities not belonging to any of above. They transliterated the already existing English gazetteer’s lists to Hindi using the 2-phase transliteration module and Hindi names are also transliterated to English. The eight entity groups are constructed namely,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. Seed entities are collected for each individual NE class. They made a comparison between the seed entities and collected gazetteer’s lists . 81.52 f-measure is achieved after the addition of gazetteer lists and context patterns into MaxEnt based NER system, a supervised machine learning technique as compared to 75.6 f-measure (baseline result).

Description about NER and its various approaches have been discussed by the authors in [5].There is lack of proper corpus for Hindi because of which they decided to design a hybrid approach for NER in hindi to improve the accuracy.They have designed it using the combination of rule based approach and list look up approach in [6].They have identified money value,direction value and animal/bird entities as new name entities.No name entity rule is used in their proposed system and their accuracy comes to be approximately 96%.

In [7] they have implemented a hybrid approach consisting of linguistic approach with a number of linguists rules and Machine Learning approach where tagset is used to train the models for the named entity recognition of 5 Indian languages - Hindi, Bengali, Oriya, Telugu and Urdu. The various features used are Static word feature, Context list, Dynamic NE tag, First word, Contains digits, Numerical word, Word suffix,Word prefix, Root information of a word and Parts of speech information. Parts Of Speech(POS) used the three tags - nominal (Nom), postposition (PSP) and other (O). Their system could recognize 12 classes of NEs with 65.13% f-value (Hindi), 65.96% f-value(Bengali) and 44.65%(oriya), 18.74%(Telugu), and 35.47%(Urdu).

In [8] NERSSEAL-2008 discusses about the ambiguities of the various Indian languages.They have 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 helps to determine the second best tag from the CRF model and deals with nested entities like NE-Number and NE-Measure with the help of Gazetteers list. Their system gave the performance of F–scores of 40.63,39.04, 50.06, 40.94, and 43.46 for Bengali, Oriya ,Hindi, Telugu and Urdu respectively.

The authors in [9] proposed a method that comprises a combination of Maximum Entropy (MaxEnt), Conditional Random Field (CRF) and Support Vector Machine (SVM) for NER in Bengali. After taking approximately 272k word forms of training set for testing, they have developed semi-supervised learning technique that uses the unlabeled data. According to them, the use of large corpora is not enough but the system should be able to automatically select effective documents and sentences from the unlabeled data. So finally, they used an approach that consists of a weighted voting approach to combine all the models. Their system can secure recall, precision, and F-score values of 93.79%, 91.34% and 92.55% respectively.

A rule based with CRF approach for NER of Bengali was used in [10].They defined conditional probability of a state sequence and OpenNLP CRF++ was the name given to it to classify the NEs. It was implemented using C++. Such approach helps to achieve an f-measure of 90.7

The authors in [11] have worked for one of the resource scarce Indian language i.e, Urdu. Its vocabulary and writing style is quite unique and different from Hindi so Hindi NER cannot be applied for Urdu. Hand crafted rule based algorithms were designed for Urdu. The pattern of the rules used are 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 clues, 5.Multiple spellings for same NE, 6. Transliteration problems, 7.Anchor texts based corpuses, 8. Patterns and heuristic grammars. They have used Becker-Riaz corpus consisting of 2,262 documents after cleaning by applying n-gram model to remove XML tags and some unwanted contents. Their system was tested on the 36,000 token Urdu data provided for IJCNLP 2008 NER Workshop. The F1 – measure was 72.4% and rose to 81.6% after addition of few rules.

In paper [12], authors have developed an NER for Urdu, a resource scarce language with two n-gram models,namely unigram and bigram. They made use of gazetteer lists with both techniques as well as some smoothing techniques with bigram NER tagger. This system could categorize 5 classes of NEs i.e. Person, Location, Organization, Date and Time.The unigram tagger trained with training data and combined with gazetteers can produce 65.21% precision, 88.63% recall and 75.14% f-measure while the bigram NER tagger could produce 66.20% precision, 88.18% recall and 75.83% f-measure.

The authors in [13] have developed NERC (Name entity recognition and classification) system for kannada with the help of SEMI-Automatic Statistical Machine Learning NLP models based on noun taggers using HMM. After the NER of a text document,the output is secured by crptographic algorithm.Hidden Markov Model (HMM) is a supervised learning technique and a statistical model with generalized learning method. They have used Python and python Natural Language ToolKit (NLTK) for building the NERC models as it has built in functions for performing the NLP basic tasks owing to which they have gained good accuracy.

Authors have use Support Vector Machine approach [14] for NER for Nepali text. They have created the corpus manually as nepali language is a free order language and so the NER on Nepali was very complex but they have made a system which has efficient feature extraction and comprehensive recognition techniques. Their system can learn well from the small set of training data and NE recognition is increased when the size of training data is increased.

In [15] the authors have experimented their NER on Tamil database. A Standard tagset is used with total of 106 tags 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 POS tag helps to identify the proper and common nouns, cardinal numbers and also the relationship between the current preceding and succeeding words. Phrase chunking is used to identify the named entities in a sentence. Automated tools are used to perform POS and chunking and are trained using CRF with the various features.F-measure of 60.36 is achieved by them. When Support Vector Machine (SVM) is trained with the same features, they achieved a lesser accuracy of 58.34%

An NER for Punjabi is built using rule based and list look up approaches in [16] . The rules are written manually for the system to identify the NEs. After the removal of most common words from the database, a list look up approach is used with the Gazetteer's lists. Their system could accurately identify the NEs with an
f-measure of 85.88%

In paper [17], authors have described about a Translation model that check the accuracy of target sentences provided the source sentence and decoder that maximizes the probability of translated text of target language. They found that there are some character exists in English which have double meaning like “you” and “u” and the poor word selection led to major inaccuracies in the transliteration.They proposed a framework for machine learning where an NER model is used to find proper entities and thereby enhancing the capability of machine translation and the language model performs efficiently.

Authors have proposed a named-entity recognition system [18] that combines named entity extraction with a simple form of named-entity disambiguation. Their unsupervised system is compared with that of the supervised system using the Message Understanding Conference MUC7 NER corpus and showed that their technique can be applied to other named-entity types, like the brand name of the cars, the bridge names. They have even included an experiment with car brands and they planned to create a system that can recognize named entities in a given document without training.

Above are the various NER works done on Indian languages. Now coming to the special reference of this paper which is about Manipuri it is already described about the unique nature of Manipuri language. Owing to that only two NER systems have been designed for it till now.The first one [19]is by kishorjit where they have designed two models,one is based on active learning technique using lexical context patterns from an unlabeled news corpus(http://www.thesangaiexpress.com/) of 174,921 wordforms while the other is based on well known machine learning algorithm i.e.Support vector machine.They manually annotated the corpus with four NE tags comprising of person name,location name,organisation name and miscellaneous name.The SVM classifier uses different contextual information of the words,orthographic word-level features and the lexical context patterns from the unlabelled corpus as features for prediction of the NE classes.After trying out different features they found that Prefixes and suffixes of length upto three characters of the current word, dynamic NE tags of the previous two words, POS tags of the previous two and next two words, Digit information, Length of the word, Infrequent word could give best results of Recall, Precision and F-Score values of 93.91%, 95.32% and 94.59% respectively after training and testing with 28,629 and 4,763 wordforms.

Another system has been designed for Manipuri NER by the authors in [20] using Conditional based approach(CRF).For the data, a Manipuri text document is used and is filtered to rectify the spelling mistakes and syntax of a sentence by a linguist expert from Linguistic Department, Manipur University. In the corpus words written in English,are rewritten in Manipuri to get free from erroneous output. Out of 55,000 tokens which is the Gold standard data, a total of 50,000 words have been used for training and testing is done on the rest 5000 words.

Manipuri being a complex and resource-scarce language ,the researchers face a lot of problems in the experiments. So they involve a lot of features namely Current word,Surrounding Stem words as feature,Acceptable suffixes,Acceptable prefixes as feature,Binary notation if a suffix(es) is present,Number of acceptable suffixes as feature,Binary notation if a prefix(es) is present, Digit features, Binary Notation of general salutations/preceding word of Name Entity,Length of the word,Word frequency and Surrounding POS tag. Their system could give a Recall of 81.12, Precision of 85.67 and F-Score of 83.33. The features are quite useful in performing the NE identification and therefore can be tried in other Indian languages.

.

III. CONCLUSION

A lot of work has been done for English,European and Asian languages but not considerable amount of work could be done for all the Indian languages due to the various reasons discussed in the paper. The work becomes more complex when it is confined to a regional language like Manipuri. And coming to the recent trend of social media text,which is contrary to formal documents many researchers are attracted towards it. But it is found that no NER has been developed for Manipuri posts from social media like Facebook and Twitter.So as a future scope , more approaches can be explored to develop NER for Manipuri if they could perform better and also deviate the experiments towards the social media text and see their performance. Such approaches will led to the development of the Manipuri language and bring it up to the global platform .

REFERENCES

[1] James Allen.Natural Language Understanding. *Pearson Education, Inc*.

[2] P. Hiremath & B. R. Shambhavi.Approaches to Named Entity Recognition in Indian Languages: A Study. *International Journal of Engineering and Advanced Technology.(2014) Vol. 3 pp191-194*.

[3] A. Dey, J. Abedinand & B. Purkayastha.A Comprehensive Study Of Named Entity Recognition On Inflectional Languages. *International Journal of Advanced Research in Computer Science and Software Engineering. (2014) Vol. 4, pp696-701.*

[4] S.K.Saha, S. Sarkar, 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.*

[5] Y.Kaur & R.Kaur. A review Name Entity Recognition in Hindi. *International Journal of Computer Engineering and Application.(2014) Vol. 7, pp1-8.*

[6] Y. Kaur & R. Kaur.Named Entity Recognition system for Hindi Language using combination of rule based approach and list look up approach. *International Journal of scientific research and management.2015 Vol. 3, pp 2300-2306.*

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

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

[9] A. Ekbal & S. Bandyopadhyay.Named Entity Recognition Using Appropriate Unlabeled Data, Post-Processing and Voting.2009

[10] Asif Ekbal, Rejwanul Haque, Sivaji Bandyopadhyay. Named Entity Recognition in Bengali: A Conditional Random Field Approach.2008

[11] kashif Riaz.Rule-based Named Entity Recognition in Urdu. 2010 *Named Entities Workshop, ACL 2010, pages 126–135*, Uppsala, Sweden.

[12] F. Jahangir, W. Anwar, U. Bajwa1 & X. Wang.N-Gram and gazetteer list based Named Entity Recognition for Urdu: A scarce resourced Language. unpublished.

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

[14] S. Bam & T. B. Shahi.Named Entity Recognition for Nepali text using Support Vector Machine. *Intelligent Information Management, (2014) pp21-29.*

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

[16] 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.*

[17] R. Sharnagat,.Named Entity Recognition: A Literature Survey. unpublished.

[18] S. Kulkarni.A survey on Named Entity Recognition for South Indian Languages.*National Conference on Indian Language Computing.2014*

[19] Thoudam Doren Singh, Kishorjit Nongmeikakpam, Asif Ekbal, Sivaji Bandyopadhyay.Name Entity Recognition for Manipuri Using SVM. *Pacific Asia Conference on Language, Information and Computation", Hong Kong, (2009)*

[20] Kishorjit Nongmeikakpam, Leisram Newton Singh, Tontang Shangkhunem, Bishworjit Salam, Chanu, Sivaji Bandyopadhyay.CRF Based Name Entity Recognition in Manipuri: A Highly Agglutinative Indian Language. *8th International Conference on Natural Language, IIT Kharagpur, India, (2011)*

Priyadarshini Lamabam "A Survey on Named Entity Recognition of the Indian Languages with special reference to Manipuri". IOSR Journal of Engineering (IOSRJEN), vol. 08, no. 6, 2018, pp. 71-75.