International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878, Volume-8 Issue-3, September 2019

Named Entity Recognition for a Low Resource Language

Abhijit Debbarma, Paritosh Bhattacharya, Bipul Shyam Purkayastha

Abstract: ***Kokborok named entity recognition using the rules based approach is being studied in this paper. Named entity recognition is one of the applications of natural language processing. It is considered a subtask for information extraction. Named entity recognition is the means of identifying the named entity for some specific task. We have studied the named entity recognition system for the Kokborok language. Kokborok is the official language of the state of Tripura situated in the north eastern part of India. It is also widely spoken in other part of the north eastern state of India and adjoining areas of Bangladesh. The named entities are like the name of person, organization, location etc. Named entity recognitions are studied using the machine learning approach, rule based approach or the hybrid approach combining the machine learning and rule based approaches. Rule based named entity recognitions are influence by the linguistic knowledge of the language. Machine learning approach requires a large number of training data. Kokborok being a low resource language has very limited number of training data. The rule based approach requires linguistic rules and the results are not depended on the size of data available. We have framed a heuristic rules for identifying the named entity based on linguistic knowledge of the language. An encouraging result is obtained after we test our data with the rule based approach. We also tried to study and frame the rules for the counting system in Kokborok in this paper. The rule based approach to named entity recognition is found suitable for low resource language with limited digital work and absence of named entity tagged data. We have framed a suitable algorithm using the rules for solving the named entity recognition task for obtaining a desirable result.***

Index Terms**:** ***NER, Kokborok, Rule base, NLP***

I. INTRODUCTION

The study of named entity recognition (NER) is to identify a predefined entity from a set of given text inputs. The entities are defined as per the required recognition task. The general entities are defined within the proper noun. The named entity recognition engine normally tries to recognize the PERSON NAME entity, LOCATION entity, ORGANISATION entity, QUANTITY entity, TIME entity etc. There are many applications of the NER system. It is widely used in many areas of natural language processing of Information Retrieval (IR), Question Answer system (QA), Machine Translation (MT) etc. The research on NER is mostly studied in the English language. The research in Indian languages for named entity recognition is not as developed as the English language. Many Indian languages are now studying the named entity recognition research of in various universities and research center. Studies of named entity recognition on various Indian languages like Bengal, Sanskrit Telugu, Hindi, Tamil, etc were being taken up in various Indian universities and institutes.

The study of named entity recognition system has got significance importance after the Message Understanding Conference (MUC6, MUC7). Several researches have been studied to solve the problem of named entity recognition. There are three main ways of solving the named entity recognition task viz. the approach using the linguistic rules, the machine learning approaches and a hybrid approach which is the combination of machine learning approach and the rule based approach

II. PREVIOUS WORK

Named entity recognition is the processes of identifying the named entity. Several researchers have studied the issues pertaining to the NLP and NER. Lec Ratinov and Dan Roth [1] in their research study give an inside of the NER system. It tells us that BILOU tag format is better than the BIO tag format. The paper also compares the performance of different decoding algorithm. The Greedy decoding algorithm is found faster and better then the Viterbi algorithm. It further discusses on non local feature consisting of context aggregation prediction history etc. Importance of unlabeled text to create word cluster technique is also being discussed. The authors come to conclusion that the use of gazetteers in the NER system increases the efficiency of the NER system.

The different approaches to solving named entity recognition problem are the supervised method and the unsupervised method. Supervised methods are those algorithms that learn from the learning pattern from a given training class. After the training set is provided the algorithm learns and annotate as per the learned training set. Some of the popular supervised method used in solving NER problems are Hidden Markov Model (HMM), Conditional Random Field (CRF), Support Vector Machine (SVM), Maximum Entropy Model (MEM), Decision Tree (DT) etc. Bikel et al [2] used the method of Hidden Markov Model to solve the NER problem. They developed an NER system called the Identifinder system, where only a single label can be assigned to a word in context.

Manuscript published on 30 September 2019
* Correspondence Author
Abhijit Debbarma*, PhD Scholar, Department of Computer Sc & Engineering, NIT Agartala, Jirania, Tripura, India.
Dr. Paritosh Bhattacharya, Associate Professor, Department of Computer Sc & Engineering, NIT Agartala, Jirania, Tripura, India.
Prof. Bipul Shyam Purkayastha, Professor, Department of Computer Science, Assam University, Silchar, Assam, India.

The system gives the desired result to the given word or declares it as NOT-A-NAME. The model generated used the popular Vitervi algorithm. Conditional Random Field (CRF) was first introduced by Lafferty et al [3]. It is model used mainly in the area of pattern recognition and machine learning. McCullum et al [4] tried to solve the NER problem using the CRF method. They proposed a feature set for NE. An accuracy of 84% was obtained for the CoNLL shared task for English language. Vapnik and Cortes [5] introduced the concept of support vector machine on a concept of linear hyperplane. McNamee and Mayfield [6] used the SVM algorithm as a binary decision problem. Manning [7] in their paper have proposed a system that generates seed candidates through local, cross-language edit likelihood and then bootstraps to make broad predictions across two languages, optimizing combined contextual, word-shape and alignment models. Animesh Nayan et al. [8] discuses on the issue of solving the NER problems using the language independent approaches. To address the use of soundex algorithm and editex algorithm. The algorithms are used to the find the similarity between the strings. An Indian language Hindi is being transliterated to English. The transliterated strings is then match with the English string based on the Soundex algortihm and Editex algorithm. The similar string are then used for recognition the named entities. The authors compare their work with the Stanford Named Entity Recognition System and found the method they used as better.

The rule based NER system follows a certain linguistic pattern according to the language. The study of several linguistic rules for the Urdu language to solve the named entity recognition system was discussed in [9]. The paper studies the general steps required in building a Urdu NER system. In spite of Hindi being similar to Urdu language, the gazetteers for Hindi language cannot be used in Urdu language NER system to improve the accuracy. It also shows that the ruled based approaches are better than the machine learning approaches like the CRF approach. A rule-based Urdu NER algorithm outperforms the models that use statistical learning like CRF. Singh et al [10] discussed in details about the rule based Urdu named entity recognition system. The combined the ruled based approach with the dictionary lookup approaches. They obtain a good accuracy but the limitation to have large NER corpus degrade the accuracy level of the named entity recognition of Urdu language. Ekbal et al [11] has developed an independent NER system for Indian languages. They have implemented the work using the conditional random field approach.

Thus we have seen a various techniques and methods in implementing the algorithms for solving the named entity recognitions. The Indian languages are reported to have more accuracy using the hybrid approaches while the semi supervised method is more suitable for resource constrained languages.

III. KOKBOROK

Kokborok is spoken in the state of Tripura, situated in the North Eastern part of India. It is also spoken in the state of Mizoram, Assam and Chittagong Hill Tract region in Bangladesh. As per the language family, the Kokborok belongs to the Tibeto-Burman language family [12]. The language is similar to other north eastern languages of the family spoken like Boro, Garo, Dhimasa etc. It is believed that the Kokborok language has its own script called Koloma which was used by the Royal Family of Tripura, but due to its unpopularity with the general public the script died a natural dead. Presently Kokborok is found to both the Bengali script and Romans Script by the speaker and writers of the language of which the later has greater acceptance among the educated Tiprasa intellectuals. Kokborok is highly inflectional language.

IV. ISSUES OF KOKBOROK NER

Kokborok belongs to the Tibeto-Burman languages family. Very less work has been done on the computational study of Kokborok language. Very little research study has been found for Kokborok language. Kokborok stemmer and its application in various field of NLP has been reported in [13-14]. Like many others Indian languages the Kokborok also has many drawbacks in successfully developing a Kokborok named entity recognition system. Kokborok being a low resource language has many constraints like other languages of India. We study here the various issues we found in developing Kokborok named entity recognition.

The various issues limiting the development of Kokborok NER are briefly reported below:

a) Scanty computerized Kokborok work: The language has very limited digital work suitable for research. This unavailability of digital work is a big hindrance to NER research especially for machine learning approach where a large number of training data are required.

b) Absence of annotated data: Availability of annotative data is very important to develop a named entity recognition system. We need a sufficient number of NER tagged data. Kokborok being a low research language have very less digital data and does not have any tagged data.

c) Inconsistent uses Spelling: Having a standard spelling is another challenge for Kokborok like any other Indian languages. An inconsistence use of spelling by various people has been seen amongst the Kokborok writers. Words are being written in different ways. The influences of social media have aggravated the spelling issues in Kokborok. UAtwi is also written as WAtwi.

d) Word Order: The sequence of word in Kokborok doesn't follow a fixed pattern. Kokborok follows a free word order form. The sequence of words can be represented in a different ways for a same given sentence.

e) Ambiguity: The ambiguity observed in the uses of word and named entity in Kokborok is another challenge for Kokborok NER system. Like Indian names which are often ambiguous the Kokborok name too is highly ambiguous and this makes the NER system predict the right entity.

f) NLP resources and tool: Like most of the Indian languages Kokborok being a resource scarce language doesn't have a ready to use NLP tools. There are no developed NLP tools such as POS taggers or Morphological analyzer for Kokborok that are publicly available. NLP tools helps in better recognition of correct entity.

g) Capitalization: For English language the capitalization is quite clear and unambiguous. But Kokborok follows no such rules.

h) Standardization: Kokborok has no standard practice of representing a word. Like the spelling of various words are not standardized and the word suffixes are also found to inconsistence in its uses.

V. KOKBOROK NAMED ENTITY RECOGNITION

Named entity recognition can be solved through a machine learning approach or through a rule based approach. Statistical based approach for Kokborok named entity recognition as been studied in limited domain. The statistical approach uses the statistical information from a training dataset. It creates a model based on the training data. This approach works well when large numbers of training datasets are available. This approach is used widely for resource rich language. Frequency based Kokborok NER system was studied in [15] which consists of a dictionary of annotated Kokborok words. The words are tagged for person (PER), organization (ORG), location (LOC) and others (O) for non entity. At the beginning the words are tokenize and then the words are tagged following the IOB format. As per IOB format the tagged word obtained are B-PER, I-PER, B-LOC, I-LOC, B-ORG, I-ORG and O. The frequency based Kokborok NER system uses the gazetteer for looking up the named entity and also it uses the statistical information to tagged the named entity.

Conditional Random Field (CRF) is a statistical modeling technique where the nearby entities are taken into consideration for sequence labeling. The Conditional Random Field (CRF) was first introduced by [Lafferty et al]. It is model used mainly in the area of pattern recognition and machine learning. McCullum et al tried to solve the NER problem using the CRF method. Applying this technique in named entity recognition a prediction is obtained for the named entity based on the training tag set. CRF are undirected graphical models which are trained conditionally to predict the output. The CRF approach are said to be more suitable for sequence modeling for named entity recognition. The CRF and SVM approach to Kokborok named entity recognition has been studied [16-17]. The tagged data has been seen as a constraint for statistical based approach.

VI. DEVELOPING RULES FOR KOKBOROK NER

We are developing rules for our named entity recognition system based on the statistical feature we obtained.

Linguistics feature are also taken when we frame our rules. The morphological feature of the word is considered important information in framing the rules. Framing rules for named entity recognition requires a lot of linguistic knowledge of the language. The rule based approach follows some specific rules to solve the problem of NER. The drawbacks of the rules based approach are that it requires to be updated regularly. The rules based approaches are also mainly found to be language specific. This approach is found to be suitable for low resource language where the tools and resources for the language are unavailable. The linguistic influenced rules that we have used in our algorithm to study the named entity recognition are as follows:

a) Checking the Prefixes: If the prefixes contains Mr. Mrs., Mg., Mgt., etc the next word is used to identify person entity. Like Mr Maitang Debbarma or Mg. Maitang Debbarma

b) The uses of the word takhuk or bukhuk without any suffixes are an indication that the next word to follow is a person entity.

c) If the next word is amchai or kami, para, nagar etc then it is most likely the present word is name of a location. Makhumai kami or Makhumai amchai

d) Checking the Suffixes. The suffixes analysis using the stemmer helps us understand the entity. Eg. If the word has suffix 'NI' then it is most likely to the named entity of either a PERSON or LOCATION. Khumulung o [LOC], Khumulung ni [LOC], Aisrang ni [PER] Aisrang no [PER].

The common suffix that allow us to identify the Person, Location and Organisation are given below:

Person: Aisrang [name of a person] Aisrang-ni, Aisrang-no, Aisrang bai, Aisrang-khe, Aisrang ya, Aisrang phu, Aisrang-le

Location: Agartala [name of a location] Agartalani, Agartala-no, Agartala bai, Agartala khe, Agartala ya, Agartala phu, Agartala o

Organisation: ADC ni, ADC-no, ADC bai, ADC khe, ADC ya, ADC-phu, ADC-o,

Counting Method in Kokborok follows the cluster method of counting system. Kokborok uses different prefix only for categorizing the counting system. We have tried to study the cluster method of counting in Kokborok. The Kokborok counting system uses the following ways for counting one to ten like sa nwi, tham, brwi, ba, dok, sni, char, chuku and chi. The common prefixes used for counting in Kokborok are discussed below:

a) Khok- is used for counting money, eg khok***sa***
b) Kai- is used for counting things.
c) Kong-is used for counting any cylindrical shape
d) Lai- is used for counting leaves or pages
e) Thai- is used for counting fruits
f) Dek- is used for branches of trees etc
g) Khung- is for roof shape structures
h) Dul- is used for round shape
i) Khorok- is used for head counting (humans)
j) Ma- is used for counting animals
k) Dam- is used for time.
l) Kol is used for counting cylindrical or circle shape
m) Lep – is used for counting flat circular objects.

We design a simple rule based algorithm based on the above model. We also have separate files for different entities. A separate dictionary file is stored for location marker, person marker, time marker, organization marker. A separate list of prefix is stored for identifying the time marker. We have done our experiment on 550 sentences consisting of 8910 token of words. The experiment gives us result of 83.6% on the entity. The result also shows that the location entity and person entity gives a confusing result. The system works better for the known words.

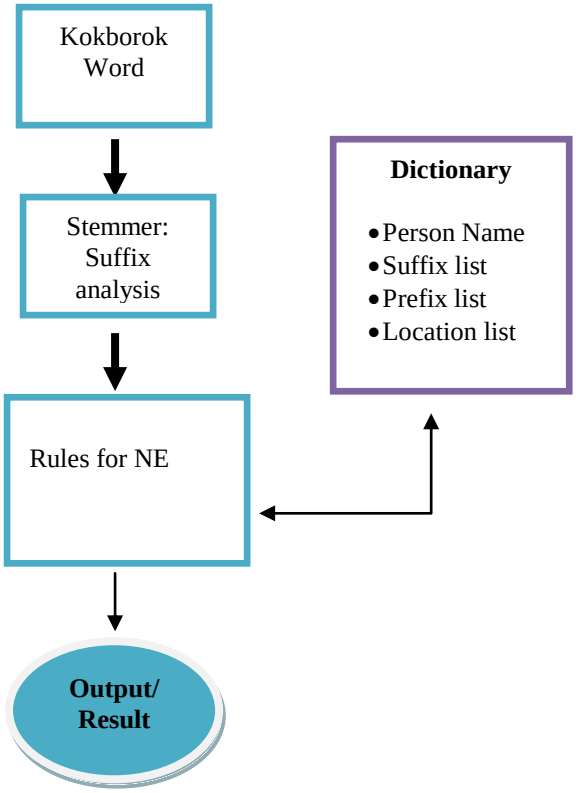

Fig: Rule based Kokborok NER

VII. CONCLUSION AND FUTURE SCOPE

A rule based approach for named entity recognition system has been studied for Kokborok language. We have classified our result based on Person, Location, Organisaiton and Numbers. Unavailability of tag data remains one of the core issues in developing a well tuned named entity. Machine learning approaches rely on large training data inputs. The rule based named entity gives us result to 83.6%. The rules based have a drawback on confusing entity. It is suitable for a low recourse language with minimum data availability. The rule based model relies on the rules but it has to be updated for a new situations. The rules based works for close domain work. As no research on prefix analysis of Kokborok words are done, it is difficult to find the accuracy of the analysis. The prefix used for counting the Kokborok counting system has been studied. A lot more analysis would be done in future work to get better accuracy. The database for the name list will be updated which is very important for rules based to check for NER tag data. However with the limited data and rules we have able to frame rules for analyzing the named entity recognition system. The rule based approach gives us an encouraging the study the linguistic aspect of the language. But the machine learning based would be suitable once the tag data is available. The Hybrid models combining the rules based with the machine learning approach would give us a better result for an open domain named entity recognition.

REFERENCES

1. Ratinov and D Roth, “Design Challenges and Misconceptions in Named Entity Recognition” CoNLL 2009.
2. Daniel M. Bikel, Schwartz, and Ralph M. Weischedel. “An algorithm that learns whats in a name”. Machine Learninng, 1999
3. John D. La_erty, Andrew McCallum, and Fernando C. N. Pereira. “Conditional random fields: Probabilistic models for segmenting and labeling sequence data”. In Proceedings ICML '01, Morgan Kaufmann Publishers Inc. 2001.
4. A McCallum and W Li. “Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons”. In Proceedings of HLT-NAACL 2003 - Volume 4, CONLL '03, ACM, 2003
5. Cortes and Vapnik. “Support-vector networks”. In Machine Learning, pages 273-297,1995.
6. P McNamee and J Mayfield. “Entity extraction without language-specific resources”. In proceedings COLING-02, ACM, 2002.
7. Robert and D. Manning. “Accurate unsupervised joint named-entity extraction from unaligned parallel text”. In Proceedings of NEWS '12, pages 21-29, ACL, 2012.
8. Animesh Nayan, B. Ravi Kiran Rao, Pawandeep Sing, Sudip Sayal and Ratna Sanyal, “Named Entity Recognition for Indian Languages”
9. Riaz, Kashif, “Rule-based Named Entity Recognition in Urdu”, NE Workshop, ACL 2010.
10. Umrinder Pal Singh, V Goyal, GS Lehal, “Named Entity Recognition System for Urdu”, Proceeding of COLING 2012.
11. Asif Ekbal, etal, “Language Independent Named Entity Recognition in Indian Languages”, Proceedings of the Workshop on NER for South and South East Asian Languages, IJCNLP-08 2008
12. Kumud Kundu Chowdhury, “Kokborok the promising language of North East”, Tripura, India.
13. A Debbarma, “A simple Kokborok word Stemmer and its application in spell checking”, Proceeding of the ICFC2012, Bangalore, Narosa Publication, 2012.
14. A Debbarma, BS Purkayastha., Paritosh Bhattacharya., “Stemmer for Resource Scarce Language using String Similarity Measure” , ICROIT-2014, Ghaziabad, Feb 6-8, 2014.
15. A. Debbarma, P. Bhattacharya, B. S. Purkayastha, "Frequency based named entity recognition system for under resource language", Proc. ICCICCT, IEEE, pp. 847-849, 2014
16. Abhijit Debbarma, Paritosh Bhattacharya and Bipul Syam Purkayastha, “SVM approach to named entitiy recognition for low resource language”, Journal of Advanced Research in Dynamical and Control Systems, ISSN: 1943-023X, Issue: 02-Special Issue, Pages: 738-747, Year: 2017
17. Abhijit Debbarma, Paritosh Bhattacharya and Bipul Syam Purkayastha, “CRF Approaches to Kokborok Named Entity Recognition, a Low Resource Language”, International Journal of Control Theory and Applications, ISSN : 0974–5572, Volume 9, No 42, Number 1 • 2016