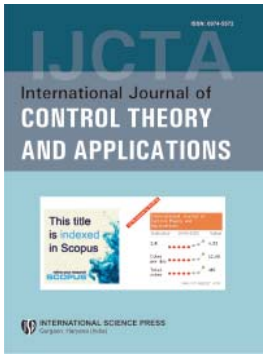

International Journal of Control Theory and Applications

ISSN : 0974–5572

Volume 9 • Number 42 • 2016

CRF Approaches to Kokborok Named Entity Recognition, a Low Resource Language

Abhijit Debbarma[a] Paritosh Bhattacharya[b] and Bipul Syam Purkayastha[c]

[a]*Research Scholar, NIT Agartala, Computer Sc and Engineering, Jirania, Tripura, India*

E-mail: db.abi@mail.com

[b]*Associate Professor, Department of Mathematics, NIT Agartala, Jirania, Tripura, India*

E-mail: pari76@rediffmail.com[2],

[c]*Professor, Computer Science, Assam University, Silchar, Assam, India*

E-mail: bipul_sh@hotmail.com[3]

Abstract: Named entity recognition is the process of identifying the given entity from a text input corpus. The named entity may be name of place, organization, person etc. The problem of named entity recognition has gain much research highlight for major languages like English, but for low resource language the research is still yet to achieve its acceptance. The research on Named Entity Recognition is not a new field anymore. Named entity recognition (NER) is widely used in the application of information retrieval and other application of natural language processing. It is also used widely to enhance the application of Question Answering system. We have tried to study the problem of named entity recognition for Kokborok language, a low resource language. Kokborok is the official language of the Indian state of Tripura. It is spoken in the state of Tripura in India and in the Chitagong Hill Tract region of Bangladesh. The problem of NER can be solved through rule based method or by machine learning approach. This paper tries to study the scope of machine learning approach to named entity recognition for a low resource language. We have applied the conditional random field (CRF) approach for our work. The CRF approach when combined with post processing technique gives better NER result. An initial study has shown encouraging result. We were able to obtain an F1 score of 71.59 for our work.

Keywords: *CRF, Named entity recognition, Machine learning.*

1. INTRODUCTION

Named entity recognition (NER) is the processes of identifying the desired entity from a given text. The entity can be name of a place (LOC), person (PER), organization (ORG) or others (O). Traditionally NER looks into the proper noun in finding the named entity. Named entity recognition is considered one of the most important features in the application of data mining. Named entity recognition (NER) is the process to identify the named entity in a given text. The entities are basically the Proper Noun. NER system tries to recognize the NAME, LOCATION, ORGANISATION, QUANTITY, TIME etc. The NER system is widely used in many applications of Natural Language Processing (NLP) areas. It is widely used in the areas of Information

Retrieval (IR), Question Answer System (QA), Machine Translation (MT) etc. The research in NER is not new to the English language. Researches on Indian languages have not been well developed like the English language. Recently many Indian languages too were seen keenly working in the field of NER system. Indian languages like Hindi, Sanskrit, Telugu, Tamil, Bengali etc were being studied in several research institute in India. 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.

2. PREVIOUS WORK

Several researchers have studied the issues pertaining to the NLP and NER. [1]Lec Ratinov and Dan Roth in their paper "*Design Challenges and Misconceptions in Named Entity Recognition*" gave 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.

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. 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. [3]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*[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 introduced the concept of support vector machine on a concept of linear hyperplane[5]. McNamee and Mayfield used the SVM algorithm as a binary decision problem[6]. Munro and Manning[7] in their paper titled "*Accurate unsupervised joint named-entity extraction from unaligned parallel text*", 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 *et al*[8] discuseses 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 are language specific. A rule based Urdu NER system was discussed by Riaz[9]. The paper studies the general steps required in building an 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. Miran *et al*[12] has used the word embedding technique to sole the named entity recognition problem. 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.

3. KOKBOROK

Kokborok language belongs to the Tibeto-Burman language family. It is widely spoken in the state of Tripura and adjoining areas in the state of Assam, Mizoram and also in the Chitagong Hill region of Bangladesh. Kokborok has a flexible word order. This feature like for many Indian languages makes the NER system difficult to recognize the correct entity. A Debbarma *et al*[13] has done some work on Kokborok NER based on frequency statistics. They have also discussed on the characteristics of Kokborok language in respect to the implementation of it in the areas of named entity recognition. There are many challenges of Kokborok language to implement named entity recognition. Some of the challenges are discussed by the authors. The digitization of the Kokborok language is not found much and those found are also not named entity recognition ready. This makes the issue more difficult to get the text corpus to train the NER system. The problems facing the named entity recognition for developing for Kokborok language are briefly discussed below:

1. Obtaining tag data is one of the major task of named entity recognition. As a low resource language we don't have any NER tagged data available for the language. This becomes a very challenging issue for developing an NER system.
2. Kokborok like many Indian languages is a word order free language. This makes the system to predict falsely for a named entity in the machine learning approaches which follows pattern to predict correct entity.
    a) Manik [B-PER] Sarkar [I-PER] Agartala [B-LOC] o miya Takarjala [B-LOC] ni sokphaikha.
    b) Miya Takarjala [B-LOC] ni Agartala [B-LOC] o Manik [B-PER] Sarkar [I-PER] sokphaikha.
    c) Agartala [B-LOC] o Manik [B-PER] Sarkar [I-PER] miya Takarjala [B-LOC] ni sokphaikha
    d) Takarjala [B-LOC] ni Agartala [B-LOC] o miya Manik [B-PER] Sarkar [I-PER] sokphaikha
    e) Agartala [B-LOC] o sokphaikha miya Manik [B-PER] Sarkar [I-PER] Takarjala [B-LOC] ni.
    f) All the five sentences above are of the same meaning which means *"Manik Sarkar arrived at Agartala from Takarjala"*.
3. As Kokborok is a low resource language and developing language. The standardization of the language is a major issue. The spelling variation by various authors makes the spelling of Kokborok word looks different.
4. Ambiguity is another concern for named entity recognition system. The Kokborok language is also facing this problem of ambiguity. ***Khumulung****[Place] ni* ***khumulung****[Flower Garden] o phaidi.*

4. CONDITIONAL RANDOM FIELD

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 [3]Lafferty et al. It is model used mainly in the area of pattern recognition and machine learning. [4]McCullum et al 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. [3]Laferty has defined Conditional random field as:

Let G = (V, E) be a graph such that $Y = (Y_v)v \in V$, so that Y is indexed by the vertices of G. Then (X,Y) is a conditional random field in case, when conditioned on X, the random variables Y_v obey the Markov property with respect to the graph:

$$p(Y_v \mid X, Y_w, w \neq v) = p(Y_v \mid X, Y_w, w \sim v),$$

where $w \sim v$ means that *w* and *v* are neighbors in G.

Applying this technique in named entity recognition we predict 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.

5. KOKBOROK NER

The Kokborok named entity recognition is being developed using the CRF based approach. The data we have been using for our work is collected from *the Naikol Kokpin*, a Kokborok newspaper from Tripura. As there are no training data available we have to tag manually the data. The tagging of data for the named entity recognition is a very laborious task and time consuming. We have tagged the corpus based on the BIO (Beginning, Inside and Outside) format for NER. The named entity tag we consider for our work are LOC for location, ORG for organization, PER for person name, DAY for day name, NUM for numbers, DAT for date and month and O for others. As no ready tagged data are available we have to prepare our own named entity tagged data. As named entity recognition study for the language is relatively new we have to create our own data. As no standard rule has been found in the person name we have seen most of the name of a person is found to be name of a place or some words in Kokborok.

Table 1
Example of Kokborok Named entity

NER Tag	*Example*
B-PER	James B-PER
B-PER I-PER	Aisrang B-PER Debbarma I-PER
B-LOC	Khumulwng –B-LOC
B-LOC I-LOC	Longtrai – B-LOC Valley – I-LOC
B-ORG	TTAADC
B-ORG I-ORG	Tripura B-ORG University I-ORG
B-NUM	3849
B-DAY	Koktisal
O	Non Named Entity

We have tagged 3467 words for our experiments. The tag set are in BIO format as shown in the above table.

6. EXPERIMENT

We have used the Stanford CRF NER system[14] or CRFClassifier available at Stanford University website[15]. The CRFClassifier is used for our work in developing the Kokborok named entity recognition system. The feature forms the basic in getting the accuracy for named entity recognition. The features we have used are the Ngrams of the sequence the presence word and previous word. Standford CRF doesn't need parts-of-speech tagging (POS) to work, hence we have not used POS as a feature for our work. The algorithm works without taking the POS as one of its features which is important for a low resource language.

The default properties used with the CRF tagger are as follows:

no Mid N Grams	=	true
use Disjunctive	=	true
max N Gram Leng	=	6
use Prev	=	true
use Next	=	true
use Sequences	=	true
use Prev Sequences	=	true
max Left	=	1 # the next 4 deal with word shape features
use Type Seqs	=	true
use Type Seqs 2	=	true
use Typey Sequences	=	true
word Shape	=	chris 2 use LC

To initiate the work we first tokenize the sentences. The tokenize words are then run through the Kokborok stemmer. We have used the Kokborok stemmer as developed by A Debbarma *et al*[16]. Kokborok being a highly inflected language has a complex form of suffix.

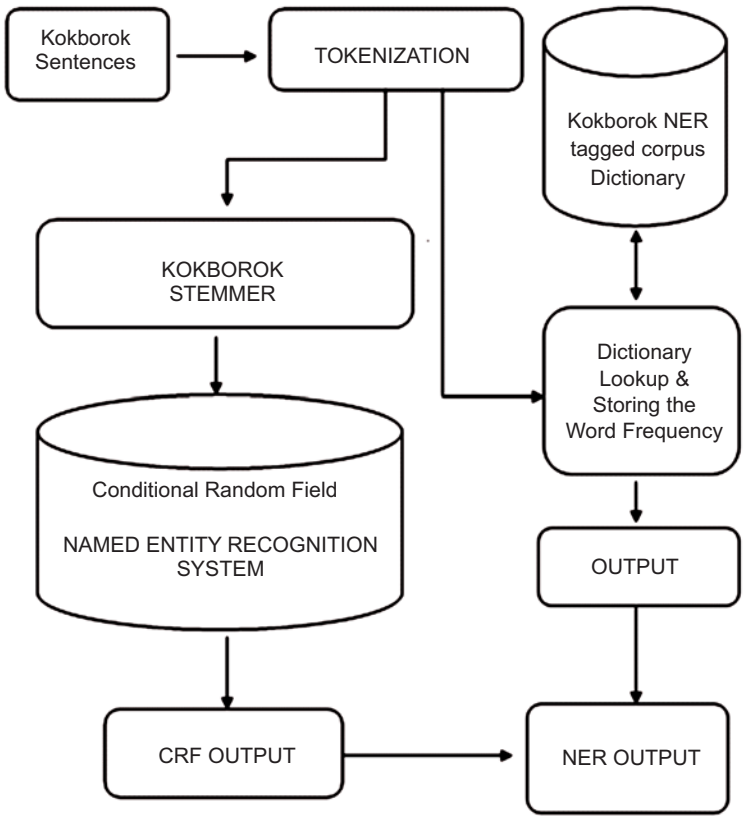

Figure 1: Diagram of CRF based Kokborok NER

The output from the stemmer is then process through the CRF named entity recognition engine. The output of the CRF NER system is then validated by the output that we obtain from the dictionary lookup approach. The final result is then obtained.

The final output words are then tag as per the BIO tagset notation for NER as stated above in Table 1. We have conducted the experiments with our tag data with the use of stemmer and without the use of stemmer.

7. RESULT

Result of named entity recognition experiment can be evaluated by using the precision, recall and F1 score of the experiments. Precision is the percentage of selected tag word that are found correct. The recall is the percentage of correct tag word that are selected by the system. The F1 measure is the harmonic of precision and recall.

$$F1 = 2PR/(P + R).$$

The results of the experiment that we conducted consist of 3467 tag words. The first experiment that we conducted was using the Kokborok stemmer to create our NER model. The experiment provides us an F1 score of 71.59. Whereas the experiment conducted without the use of Kokborok stemmer yield us result of F1 score 69.78.

This result slightly improves when we combine the statistical approach of CRF with the dictionary based method[13]. The false negative is decreases slightly and the F1 score increases.

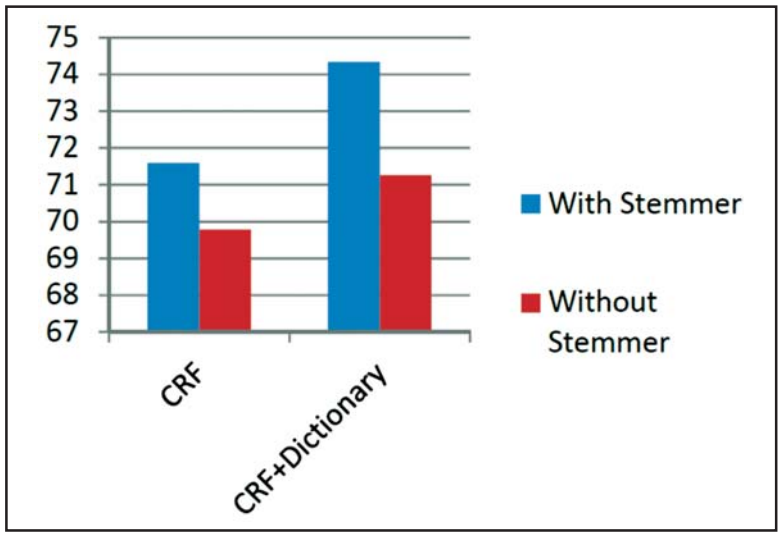

Figure 2: Graphical representation of F1 score

8. CONCLUSIONS

Machine learning approaching is being implemented to solve the problem of named entity recognition for Kokborok language. The CRF method of NER has been used for many Indian languages. The sequential graphical model when used for Indian language context works better in what is called hybrid model. In this work the word that are not found in the training data are tend to wrongly tag. The number of named entity in our training data is lesser. This affected the system to give rise to false negative in the output.

The statistical approach has a drawback in our work as the size of training data is lesser. The sequential approach is affected as a result of it. The hybrid approach is found to be working for the Indian language. We would like to experiment the Kokborok NER in hybrid approach in our future endeavor. We have used the CRF based approach as developed by Stanford University. The machine learning approach gives a good and faster result then the rule based approach. We will be trying to develop a rule based approach using the information gathered during our machine learning approach.

REFERENCES

[1] Lec Ratinov and Dan Roth, "Design Challenges and Misconceptions in Named Entity Recognition" CoNLL'09.

[2] Daniel M. Bikel, Richard Schwartz, and Ralph M. Weischedel. "An algorithm that learns whats in a name". *Machine Learning*, 34(1-3):211231, Feb 1999

[3] John D. Laferty, Andrew McCallum, and Fernando C. N. Pereira. "Conditional random fields: Probabilistic models for segmenting and labeling sequence data". In Proceedings of the Eighteenth International Conference on Machine Learning, ICML '01, pages 282{289, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc. ISBN 1-55860-778-1.

[4] Andrew McCallum and Wei Li. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003 - Volume 4, CONLL '03, pages 188{191, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.

[5] Corinna Cortes and Vladimir Vapnik. Support-vector networks. In Machine Learning, pages 273-297,1995.

[6] Paul McNamee and James Mayfield. Entity extraction without language-specific resources. In proceedings of the 6th conference on Natural language learning - Volume 20, COLING-02, pages 1-4, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.

[7] Robert Munro and Christopher D. Manning. Accurate unsupervised joint named-entity extraction from unaligned parallel text. In Proceedings of the 4th Named Entity Workshop, NEWS '12, pages 21-29, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.

[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", Named Entity Workshop, ACL 2010, pages 126-135, Uppsala, Sweden.

[10] Umrinder Pal Singh, Vishal Goyal, Gurpreet Singh Lehal, "Named Entity Recognition System for Urdu", COLING 2012.

[11] Asif Ekbal, Rejwanul Haque, Amitava Das, Venkateswarlu Poka and Sivaji Bandyopadhyay, "Language Independent Named Entity Recognition in Indian Languages", Proceedings of the IJCNLP-08 Workshop on NER for South and South East Asian Languages, pages 33–40, Hyderabad, India, January 2008. 2008 Asian Federation of Natural Language Processing.

[12] Miran Seok, Hye-Jeong Song, Chan-Young Park, Jong-Dae Kim, Yu-seop Kim, 2016.International Journal of Software Engineering and Its Applications Vol. 10, No. 2 (2016), pp. 93-104

[13] A. Debbarma, P. Bhattacharya, B. S. Purkayastha, "Frequency based named entity recognition system for under resource language", Proc. International Conference on Control Instrumentation Communication and Computational Technologies (ICCICCT) IEEE, pp. 847-849, 2014.

[14] Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. *Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005)*, pp. 363-370.

[15] http://nlp.stanford.edu/software/CRF-NER.shtml, Accessed on 11/12/2016.

[16] A Debbarma, BS Purkayastha., Paritosh Bhattacharya., "Stemmer for Resource Scarce Language using String Similarity Measure", International Conference on Reliability, Optimization and Information Technology (ICROIT-2014)" Gaziabad., Feb 6-8, 2014.