International Journal of Computational Intelligence Research
ISSN 0973-1873 Volume 13, Number 5 (2017), pp. 1027-1040

http://www.ripublication.com

Improved Speech Recognition, Classification, Extraction using Conditional Random Field with Kernel Approach

Rohit Kumar[1], Priya Batta[1], Deepshikha Chhabra[1] & Lokesh Pawar[1]

[1]Chandigarh University, Gharuan, Punjab, India.

Abstract

Extracting useful information from the pool of big data gives birth to new domain known as Information Extraction. The domain of Information Extraction has its genesis in Natural Language Processing (NLP). The fundamental drift in this field takes the birth from various competitions that are focused on the recognition and extraction of named entities such as names of people, organizations etc. As the world become more data oriented by advent of Internet, new applications of processing of structured and unstructured data comes in light. Most of the interest is to extract and classify named entities like person, organization and location etc. that is a subtask of Information Extraction known as Entity Extraction and Classification. In this field, number of models, from handcrafted rules to unsupervised learning techniques, were proposed but extracting entities and then classifying them with inference of data present before and after it with great accuracy is still a bottleneck. In this thesis work, a model is proposed that is using Kernel function for reducing the overlapping of information and Conditional Random Field (CRF) for predicting dependency among features. The results will be analyzed by using parameters like precision, recall and accuracy.

Keywords: Conditional Random Field; Hidden Markov Model; Kernel function; Named Entity Extraction and Classification.

I. INTRODUCTION

We have entered in the era where we are awashing in the flood of data. Due to increase in number of embedded sensors, smart phones, personal computers, data is created at unprecedented rate. Generating pertinent information from this data, gives

birth to new field that is emerged as Information Extraction (IE). Having roots in Natural Language Processing, it emerges as a new domain for the automatic recognition and extraction of structured information like entities, relations among entities and features that describe entities from unstructured data.
While extracting the information, the most striking point was that most of the concentration is to recognize proper information units like persons, locations, organizations, dates, etc. Identifying and classifying these entities is one of the important sub-task of IE and known as Named Entity extraction and Classification (NERC). This term was coined for the Sixth Message Understanding Conference (MUC-6) [1]. Recently NERC is receiving more attention as a fundamental building block of Natural Language application. Its usefulness is applicable in numerous domains of medicine, e-retail, entertainment etc. where there is need to recognize pharmaceutical names, product names or pop groups. The desired entities vary from field to field.

Most of the work is focused on the identification of named entities such as names of persons, organizations, locations etc. from the available natural language textual data. The early systems were focused on the handcrafted rules, where all the rules were coded by programmers, that are still one of the best way to generate high quality information. Afterwards, the statistical learning models were started to replace the handcrafted rules based systems that were unable to handle the noisy data. The two famous approaches: Hidden Markov Models and Maximum Entropy Models were deployed in parallel that were latterly superseded by conditional model popularly known as Conditional Random Field. In this race, still there is no particular winner. If one model is focused on one point then it lags behind from the other viewpoint. Extracting entities and then classifying them with inference of data present before and after it with great accuracy is still a bottleneck. In this paper, a model is proposed that is using Kernel function for reducing the overlapping of information and Conditional Random Field (CRF) for predicting dependency among features. The Corpus that is used to implement this model is CONLL-2003. The results will be evaluated by using parameters like accuracy, recall and precision.

II. RELATED WORK

A. Lisa F. Rau, Clifton Park, NY. [1991]

Lisa F. Rau et al. [1] proposed a new way to extract the names of various organizations from the given text by employing a compound based approach. This method is a combination of heuristical methods, exception lists and deep analysis of database. Primarily, this technique find out the suffixes related with the names of organizations and corporations and then seeks to fetch the starting of organization names. This technique was focused not only focused on capitalized text but also yields good results on mixed case textual data. When the name of organization is recognized, then this proficiency goes forward to extract the names of organizations.

B. Daniel M. Bikel, S. Miller, R. Schwartz, and R. Weischedel [1997]

Daniel M. Bikel [2] presents a new statistical technique that is focused on the searching of named entities and also works well on non recursive entities of textual data. This technique was based on a variate of a stochastic probability model that is known as Hidden Markov Model (HMM). It is used to search names and other non-recursive entities in text (as per the MUC-6 definition of the NE task), using a variate of the standard Hidden Markov Model. The system was named as Nymble and it was successful in generating the F-measure that is approximately equal to 90.

C. Sergey Brin [1998]

Sergey Brin [3] explained that Information Extraction systems basically have prerequisite that is two associative array. Firstly, it should be a semantic lexicon and secondly a dictionary of extraction patterns for the related field. In the paper, a multi-bootstrapping method is proposed that creates the semantic lexicon and extraction patterns at the concurrently. Primarily as a source data, this approach needs un-annotated training textual data and a fistful amount of seed words for any predefined category and then this technique bootstrap its extraction into the semantic lexicon, that acts as a foundation of extraction of patterns. To arrive this technique more racy, a 2^{nd} level of bootstrapping this approach more racy, we add a second level of bootstrapping that holds the crucial lexicon entries that are generated by mutual bootstrapping and then resumes the procedure. This multi level bootstrapping approach is assessed on a collaborative web pages and a database of terrorism news articles. Lastly, this system is able to generate rich dictionaries for varied semantic categories.

D. Masayuki Asahara and Yuji Matsumoto [2003]

Masayuki Asahara et al. [4] explained the most striking sub-field Named Entity Extraction as an authoritative subtask of document-processing such as information extraction and question answering. In Named Entity extraction of Japanese text, a distinctive way is applied. With the help of this technique, a shower of morphological analysis, Parts Of Speech tagging and chunking. But, few scenerios exists where segmentation granularity negates the output of this applied method and the fundamental building blocks of Named Entities, therefore the extraction of few Named Entities are about to inconceivable with the help of this approach. To solve this issue, a new way is proposed that is based on character-based chunking. Primarily, the multiple outputs are created by analyzing the input sentence with the help of statistical morphological analyzer. Afterwards, character types are assigned to every character and also Parts Of Speech tags are allocated. Lastly, a chunker that is based on Support Vector Machine takes few parts of input sentence as Named Entities. This technique brings more reliable and suitable data to the chunker than former methods that were focused on single morphological analysis outputs. This technique is employed to IREX NE extraction task. The output that is yielded by this

approach reveals its validation by generating the value of F-measure equals to 87.5.From this value, it is clear that this approach is transcendence and effectual.

E. Richard Evans [2003]

Richard Evans [5] discussed the role of Named Entity Recognition systems in the various domains. The paper presents a new system, named as NERO (Named Entity Recognizing and Classification in Open Domain) that can be implied in a generic way. This system is focused on the recognition of different types of entities that are commonly required while considering the named entity in any domain or in any condition. The first phase, that is recognition phase is executed by finding out the capitalized phrases in a document and then a query is presented to the search engine that works on the super ordinate word of capital sequence. A cluster is obtained to deduce a topology of named entities for any document. These super ordinate word of capitalized phrases are employed to classify the document according to this topology. This technique is evaluated on a small sized dataset and then its classification is tested. Lastly, it is concluded that this method is able to yield accuracy up to 70%.

F. Enrique Alfonseca and Suresh Manandhar [2003]

Enrique Alfonseca et al. [6] arise the point that Knowledge Acquisition is the major chokepoint while developing various types of practical applications like inference engines. In this paper, a methodology is explained that can be used to broaden the ontology with the information that is related with any particular field. The basic vantage about this technique is that it is purely unsupervised. This vantage provides its usefulness while employing this approach for any language and field. The outputs are effective but still its needs some advancement in accuracy and precision so that this technique can be implemented to ontologies that are of large size.

G. Yusuke Shinyama And Satoshi Sekine [2004]

Yusuke Shinyama et al. [7] presented her methodology to improve the process of Named Entity Recognition. According to the method, the Named Entities are find out by using the distribution of words throughout the news articles. It is one of the grandness task for using Named Entity Recognition in the applications of Natural Language Processing. The major issue that comes in the Field of NERC is sparseness of data. The basic concept that is used in designing this technique is that the entity with a proper name comes in news articles repeatedly in news articles but a common noun does not exist repeatedly. By using this approach, the model successfully able to obtain rarely occurred named entities with the accuracy of 90%. This much high precision is achieved just by comparing the time series distributions of 2 articles. The value of recall is not satisfactory however this approach is still able to improve the lexical knowledge of a Named Entity tagger.

H. Ki-Joong Lee, Young-Sook Hwang, Seonho Kim, Hae-Chang Rim [2004]

Ki Joong Lee et al. [8] presented the importance of Named Entity Recognition (NER) in the field of biomedical. This paper reveals the importance of information acquisition in the biomedical science and regarded it as primal task that gives lots of useful information. The author nominates a very fantastic Named Entity Recognizer that consists of two phases and developed on the basis of Support Vector Machine. The two phases are boundary identification phase and semantic classification phase. In previous work, when one phase support Vector Machine (SVM) was implemented then issues of multi-class and unbalanced class becomes severe regarding cost of training and performance. To remove formerly stated issues, this approach differentiates the tasks of Named Entity into two parts. Firstly, suitable Support Vector Machine classifiers are employed to fetch appropriate features and afterwards a hierarchical classification technique is employed that is purely based on ontology. The outputs of this experiment, that was employed on the GENIA corpus, reveals that the proposed approach is efficient in ameliorating the performance as well as bring down the calculating cost. This system is able to achieve the F- measure equal to 74.8 for the identification of boundary values and F-measure equal to 66.7 for the semantic classification.

I. Ulf Leser and Jorg Hakenberg [2005]

Ulf Leser et al. [9] presented an extensive survey on the work done on Named Entity Recognizer systems in the domain of biomedical science. Starting with the advantages offered by NER systems in this domain, the author states it as a fundamental technique for automatic or semi-automatic analysis of text data. Advanced NER tools are prerequisites for many applications like Information Extraction (IE), Information Retrieval (IR) and classification of documents. Taking into account the back years, NER systems have achieved appreciable concern in many fields especially in Bioinformatics. Many systems have been excogitated and implemented. Several systems and algorithms have been excogitated and implemented. In this paper, the issues related with NER systems are exploited and also the fundamental algorithm based systems are adumbrated.

J. Robert Leaman and Graciela Gonzalez [2008]

Robert Leaman et al. [10] presented the enhancing measure of research related to named entity in the biomedical field for the recognition and extraction of named entities related with medicines, proteins etc. In the globe, different research teams try to yield the substantial development in this field. Due to it, there is a requirement to develop a system that is free of cost and open source. To satisfy this need, a new system named as BANNER, is proposed in this paper. This system is betrothed to assist as a benchmark for the biomedical field. This machine learning system is carried out at Java platform and based on the Conditional Random Field. Basically, this system was developed as a domain independent system. In this system, only

unannealed semantic features are not employed. This system is able to gain best performance as compare to other existing systems.

K. Pallavi and Dr. Anitha S Pillai [2013]

Pallavi et al. [11] presented the Named Entity Recognition and classification (NERC) as a subtask of Information Extraction that is employed to find out and then assigning classes to the proper names in available Corpus. Primarily, the Named Entity and Recognition systems were developed by focusing on handcrafted rules. However, later on this approach is modified by using the Machine learning techniques like Hidden Markov Model, variant forms of Hidden Markov Models like Maximum Entropy Markov model, Support Vector machine, Conditional Random Field .The poster presents an intensive survey of Named Entity Recognition and Classification systems that were developed for different languages used in India. While concluding the paper, it can be sited that there are only few Indian languages like Hindi, Punjabi, Urdu etc. that have good NERC systems. Lastly, the different performance measures are compared on the basis of Recall, Precision and F- value. Only limiting to available systems, the NERC based on Conditional Random Field succeeds in achieving higher rank.

L. Joel Nothmana, Nicky Ringlanda, Will Radford, Tara Murphy [2013]

Joel Nothman et al. [12] presented a new approach towards recognition, extraction and classification of Named Entities. With the help of this approach, tremendous and multilingual training annotations are automatically created for the recognition of named entities. By just tapping the textual data and structure of available Wikipedia. Most of the existing Named Entity recognizing systems are developed on the basis of statistical models that annotate the data, so that names of person, organizations and locations are identified and classified. Most NER systems trust on statistical models of annotated data to identify and classify names of people, locations and organizations in text. The chokepoint of these systems is their dependence on costly annotations. This approach firstly classify the articles of Wikipedia into types of named entities , then a system is trained and then the results are evaluated on seven thousand and two hundred Wikipedia articles, that are in nine languages, that are manually labeled . This approach is capable to gains the accuracy near about 95 percent.

M. Surya Bahadur Bam, Tej Bahadur Shahi [2014]

Surya Bahadur Bam et al. [13] discussed the basic goal of Named Entity Recognizer systems to find out and then classify inflexible designators found in textual data like proper names, names of biological species and expressions related with time in already defined categories. A springing up interest of researchers is noticed in this domain from the early 19^{th} century. The recognition of Named Entity has crucial purpose in different areas of natural language processing like Machine Translation,

Extraction of Information and other arenas. A new system is proposed for the recognition and classification of Nepali textual data that is founded on the basis of Support Vector Machine that is well known technique of Machine learning employed for classification. Primarily, a set of features are fetched from the given data set of training. Then, the proficiency and precision of Support Vector machine classifiers are evaluated by applying it on three different sized training datasets. The Named Entity Recognizer system is evaluated with 10 available datasets of Nepali textual information. The intensity of this approach can be seen in the effective feature extraction and its way to include all recognition techniques. However, this system is bounded to only small set of features and also it takes in account an dictionary that directly influence its performance. The ability of system to learn is carefully noticed and it is concluded that this model is able to grasp sharply from very small sized training dataset. The most noticeable feature of this system is that its ability to learn things is sharply accelerated with increase in the training data set.

N. Prakash Hiremath, Shambhavi B. R [2014]

Prakash Hiremath et al.[14] discussed the Named Entity Recognition and Classification as one of the prominent subtask of Information Extraction that tries to identify, fetch and then classify the elements of given textual information in the predefined classes. The most striking applications of Named Entity Recognition (NER) systems are applies in the domain of Natural Language Processing such as machine translation, systems based on question answering and creating of summary automatically. A brief information is presented on the techniques used in Named Entity Recognizer systems like handcrafted, statistical and clustering of both of this. An extensive survey is executed on the pros and cons of existing techniques.

O. Andrea K. Thomer and Nicholas M. Weber [2014]

Andrea K. Thomer et al. [15] discussed the importance of Named Entity Recognizer as a heuristic way to optimize the classification of documents manually. The proposed approach is formulated while working on a project related with cooperative work through the acknowledgement. This technique was developed as part of a project studying cooperative work via the acknowledgement commands that comes in light while considering the corpus of officially published articles in journals, found in a corpus of officially published journal articles. It also presents the incertitude in the outputs generated by text mining tools in fast and rough manner. The results verifies the validity of this approach by providing the accuracy near about 80 percent.

III. METHODOLOGY

There are number of milestones that need to be achieved in order to reach the goal, so following are the sequential steps that will lead to attain the goal.

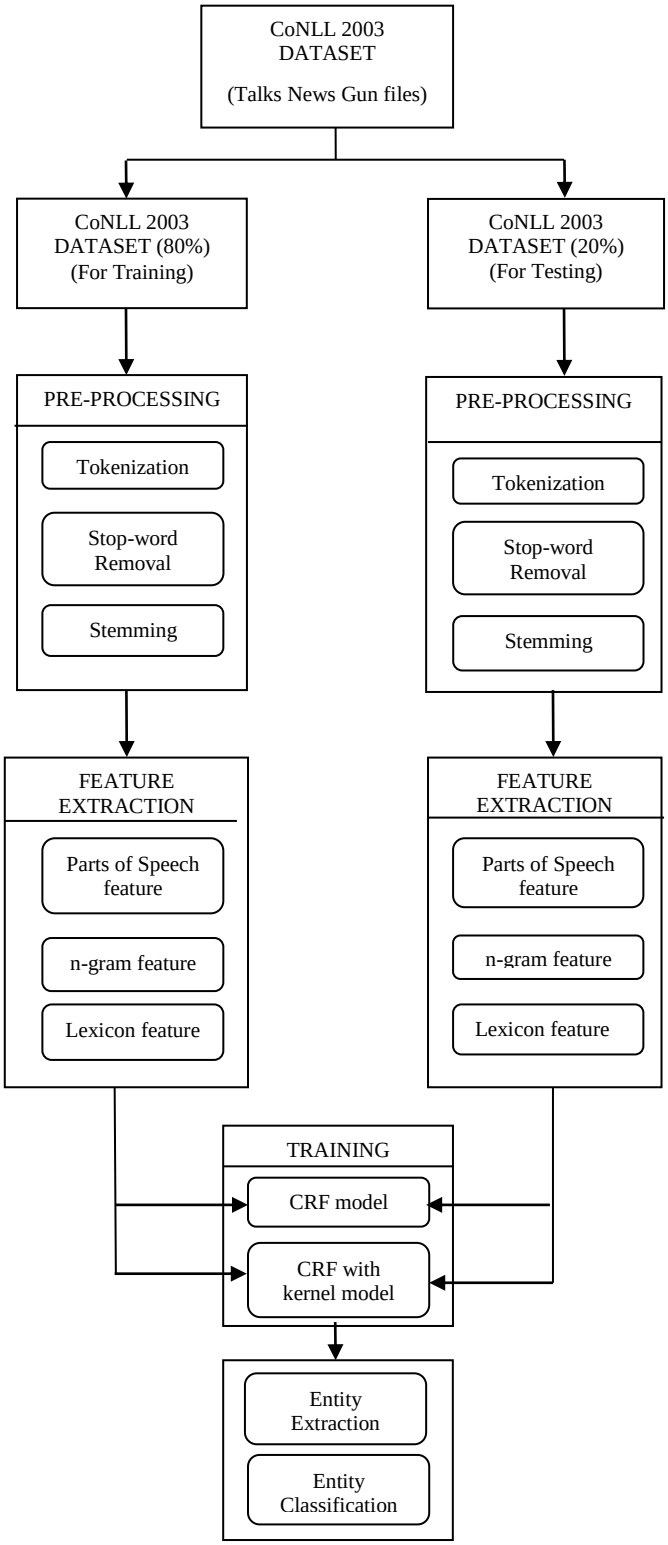

Fig. 1. Proposed Methodology for Entity Extraction and Classification

1. Database CoNLL 2003: It is dataset of about 200 GB. It consists of many types of data files. From these files, Talks News Gun file is selected to make learn and test the proposed model. This dataset is split up in the ratio of 80 by 20. The 80% files data is utilize to make the system learn through both CRF and CRF with kernel function models. The rest 20% files data is used for test the performance of both models. This ratio can be vary according to the requirements.

2. Pre-Processing of Unstructured Data: Initially, the unstructured data is preprocessed to remove the undesired words or features in order to create only useful word set.

   Tokenization: It is the preliminary step of pre-processing of unstructured data. It is the procedure to divide the given stream of textual data into words, symbols or useful elements that are known as tokens. Three most noticeable features of tokenization are:

   a) The alphabetic characters that are closely related in a string are considered as a part of one token.

   b) Whitespaces and punctuation marks are used to identify tokens.

   c) The output after tokenization may or may not carry whitespaces.

   Stop-Word removal: In this step, the stop words that do not contribute to semantics of the textual data like a, an, the, and, are, as, at, be, etc. are removed. These words are predefined for the English language but these can be added if desired

   Stemming: It is a heuristic procedure that chopped off the end of the words or the affixes of the derived words in order to retrieve the base form of word

   e.g. playing $\longrightarrow$ play

3. Feature Extraction: We are extracting features like n-gram, parts of speech and lexicon. These features are extracted from CONLL 2003 news data after tokenization and removal of stop words. Once these features are extracted separately, then they are combined.

   Parts of Speech: It is the procedure in which each word of the text is annotated with the suitable Part of Speech tag on the basis of context of their appearance. In this process, words are placed in the categories according to their role in the sentence. The basic categories are: Article, Adjective, Noun, Number, Preposition etc.

   n-gram feature: It is a connected series of n-items of textual data. It facilitates to predict the type of upcoming word.

   N-gram with size one is called unigram.

   N-gram with size two is called bigram.

   N-gram with size three is called trigram.

4. Training: Train systems for Conditional Random Field and Conditional Random Field with kernel function with the help of extracted features.
5. Entity Extraction and Classification: With the help of these models, the entities are extracted and classified according to the four categories. These categories are: person, location, organization and miscellaneous.

A. Proposed Algorithm

1) Proposed Algorithm (Algorithm for entity extraction and classification)

Input: Unstructured text of Talks News Gun file
Output: Extract the entities like person, location, organization and miscellaneous

```
1. For I= length (doc)
2.   For I=0 to length (doc)
3.       For doc: split in sentences
             X tokenization (sentence)
             Y stemming
             Z stop word removal
                Extract Part of speech(Z)
                N-gram(Z)
                Lexicon features(Z)
         End.
     End.
 4.  Combine all features (n-gram+ Lexicon+ parts of speech)
 5.  For I= 0 to Len (doc)
         Train CRF with kernel
     End.
6.   Model of CRF with Kernel.
7.   For (I= 0 to I < length (doctest. sentences))
         {
             Extract features according to Step 3
           Input in CRF with Kernel model.
         }
     End.
8.   Entity extraction and classified.
```

IV. IMPLEMENTATIONS AND RESULTS

The entity extraction process of CRF with kernel function model is compared with CRF model on the basis of three parameters: accuracy, precision and recall. The Talks News Gun files of CoNLL 2003 dataset are used for testing these models.

It is clearly seen that the proposed model greatly improved these three parameters. The precision is improved between the range of 8 to 10 %. The accuracy is improved

between the range of 6 to 15 % and finally recall is improved between the range of 6 to 14%.

TABLE I. COMPARISON OF ENTITY EXTRACTION BASED ON PRECISION, ACCURACY ANDD RECALL ACHIEVED WHEN CRF AND CRF WITH KERNEL FUNCTION MODELS ARE TESTED

M O D E L	**DATASET**	**ACCURACY**	**PRECISION**	**RECALL**
CRF	CoNLL 2003	77.09%	78.01%	76.19%
CRF WITH KERNEL FUNCTION	CoNLL 2003	92.99%	93.28%	92.71%

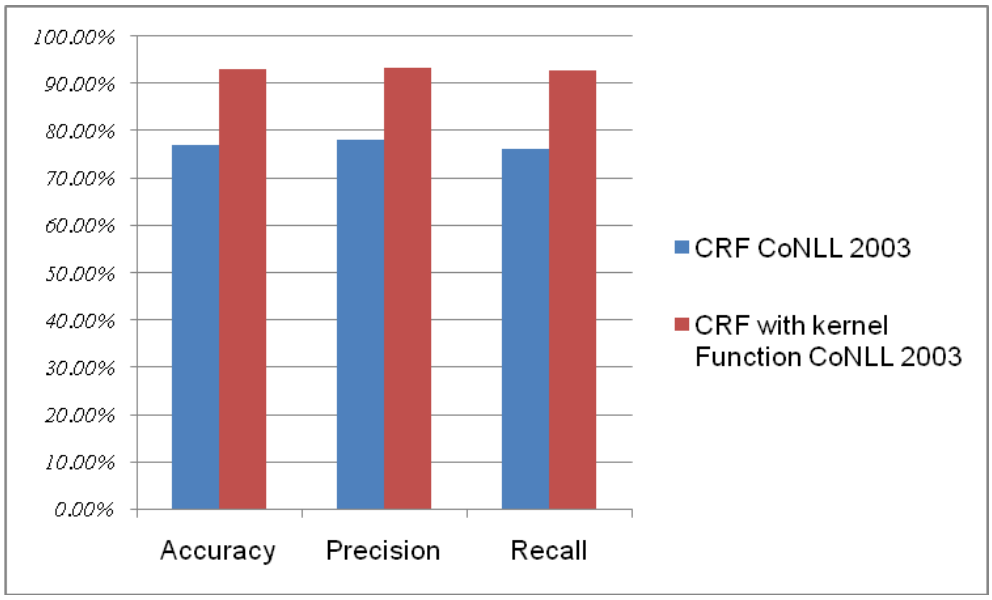

Fig. 2. Diagrammatically representation of comparison between CRF and CRF with kernel function models for Entity Extraction

After extraction process, these extracted entities are classified according to four predefined categories that are person, location, organization and miscellaneous. The results are compared on three parameters i.e. accuracy, precision and recall. The Corpus is same i.e. CoNLL 2003.

TABLE II. COMPARISON OF ENTITY CLASSIFICATION BASED ON PRECISION, ACCURACY ANDD RECALL ACHIEVED WHEN CRF AND CRF WITH KERNEL FUNCTION MODELS ARE TESTED

MODEL	CATEGORY	ACCURACY	PRECISION	RECALL
CRF	PERSON	85.97%	84.83%	87.14%
	LOCATION	86.66 %	86.64%	86.69%
	ORGANISATION	76.50 %	79.19%	73.99%
	MISCELLENEOUS	75.31 %	76.13%	74.50%
CRF with kernel Function	PERSON	91.73%	91.61%	91.86%
	LOCATION	92.51%	92.54%	92.49%
	ORGANISATION	82.83%	84.21%	81.51%
	MISCELLENEOUS	85.44%	89.06%	82.10%

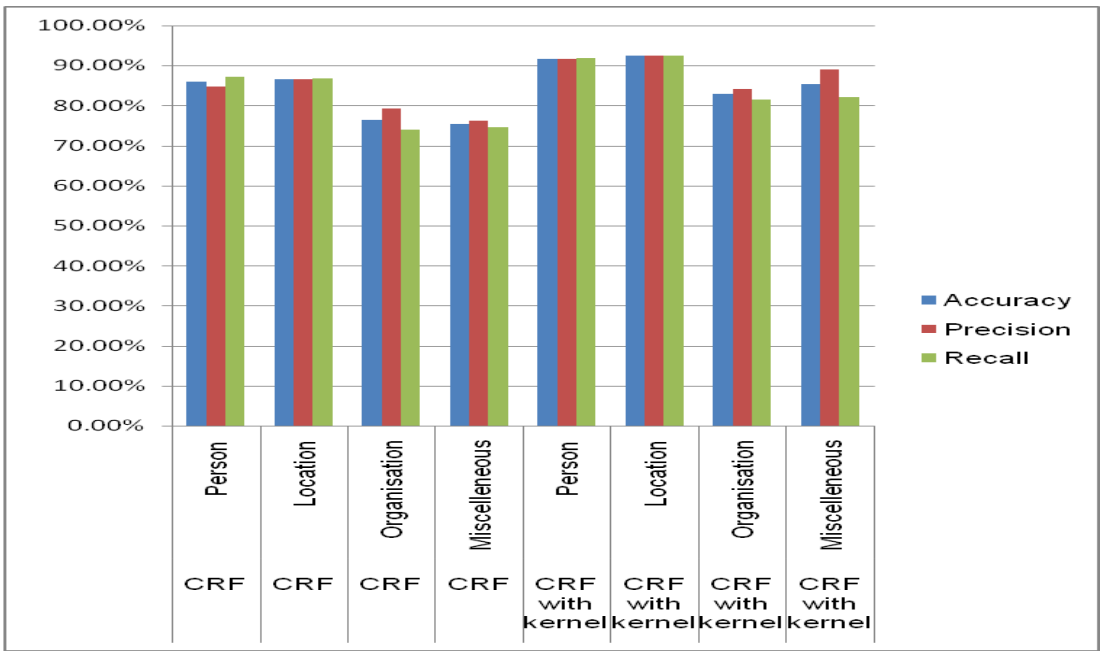

Fig. 3. Diagrammatically representation of comparison between CRF and CRF with kernel function models for Entity Classification

V. CONCLUSION

NERC is one of the most striking research area. In NERC, these days lots of research is contributed for different languages like Punjabi, Nepali, Telugu etc. We have implemented new model based on CRF with Kernel function and compared it with CRF model using the dataset CoNLL 2003. The model is successful to improve accuracy, precision and recall as compare to CRF model that was the most successful model till now. After devotion of many years of researchers to NERC, it still offers number of research options like improvement in Parts of Speech tagger so that it become able to tag proper noun as its own type rather than common noun. There is still a need of NERC model that is generic. The other issue that should be handled is word sense disambiguation.

REFERENCES

[1] Rau F. Lisa, "Extracting Company Names from Text," In Proc. Conference on Artificial Intelligence Applications of IEEE , Year 1991.

[2] Bikel Daniel M.; Miller, S. Schwartz, R. Weischedel, "Nymble: a High-Performance Learning Name-finder," In Proc. Conference on Applied Natural Language Processing, Year 1997.

[3] Brin Sergey, "Extracting Patterns and Relations from the World Wide Web," In Proc. Conference of Extending Database Technology, Workshop on the Web and Databases, Year 1998.

[4] Asahara, Masayuki, Matsumoto, Y. "Japanese Named Entity Extraction with Redundant Morphological Analysis," In Proc. Human Language Technology Conference - North American chapter of the Association for Computational Linguistics, Year 2003.

[5] Richard Evans, "A Framework for Named Entity Recognition in the Open Domain," Year 2003.

[6] Enrique Alfonseca and Suresh Manandhar, " An Unsupervised Method for General Named Entity Recognition and Automated Concept Discovery," In Proc. International Conference on General WordNet, Year 2003.

[7] Yusuke Shinyama And Satoshi Sekine, "Named Entity Discovery Using Comparable News Articles," Year2004.

[8] Ki-Joong Lee, Young-Sook Hwang, Seonho Kim, Hae-Chang Rim, "Biomedical named entity recognition using two-phase model based on SVMs," In Journal of Biomedical Informatics, ISSN: 37 , pp. 436–447, Year 2004.

[9] Ulf Leser, Jorg Hakenberg, "What makes a gene name .Named entity recognition in the biomedical literature," Year 2003.

[10] Robert Leaman, Graciela, "BANNER: An Executable Survey Of Advances In Biomedical Named Entity Recognition", pp. 436–447, Year2008.

[11] Pallavi and Dr. Anitha S Pillai, " Named Entity Recognition for Indian Languages: A Survey," International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 11, ISSN: 2277 128X, Year 2013.

[12] Joel Nothmana, Nicky Ringlanda, Will Radford, Tara Murphy, "Learning multilingual named entity recognition from Wikipedia," Year 2013.

[13] Surya Bahadur Bam, Tej Bahadur Shahi, "Named Entity Recognition for Nepali Text Using Support Vector Machines," Intelligent Information Management, Year 2014.

[14] Prakash Hiremath, Shambhavi B. R, "Approaches to Named Entity Recognition in Indian Languages: A Study", International Journal of Engineering and Advanced Technology (IJEAT), ISSN: 2249 – 8958, Volume-3 Issue-6,, August 2014.

[15] Andrea K. Thomer and Nicholas M. Weber, Year 2014.