Named Entity Recognizer employing Multiclass Support Vector Machines for the Development of Question Answering Systems

Bindu.M.S
M.G University College of Engineering
Muttom, Thodupuzha

Sumam Mary Idicula
Dept. of Computer Science
CUSAT, Cochin

ABSTRACT
Named Entity Recognition (NER) seeks to locate and classify atomic elements in text into predefined categories such as names of person, organization, location, Quantities, Percentage etc. Named entities tell us the roles of each meaning bearing word in a sentence and hence identification of these entities certainly helps us to extract the essence of the text which is very important in Question Answering(QA) , Information Extraction (IE) and Summarization. The system presented here is a Named Entity (NE) Classifier created using Multiclass Support Vector Machines based on linguistic grammar principles. Malayalam NER is a difficult task as each word of named entity has no specific feature such as Capitalization feature in English. NERs in other languages are not suitable for Malayalam language since its morphology, syntax and lexical semantics is different from them. Also there is no tagged corpus available for training. For testing this system, documents from well known Malayalam news papers and magazines containing passages from five different fields such as sports, health, politics, science and agriculture are selected. Experimental results show that the average precision recall and F-measure values are 89.12%, 89.15% and 89.13% respectively.

General Terms
Natural Language Processing, Support Vector Machines

Keywords
Named Entity Recognition, Parts-of-Speech Tagger, Phrase chunker, Compound word splitter

1. INTRODUCTION
Several preprocessing steps are necessary for natural language applications such as Information Retrieval (IR), Data Mining QA systems and Summarization [1]. One major step is named entity recognition. Main task of NER is to identify and classify words or phrases in a document into named entities. Identification is concerned with marking the presence of a word/phrase as NE in the given sentence and classification is for denoting the role of the identified NE.

The term named entity was evolved during the sixth Message Understanding Conference (MUC-6, 1995); people who were focusing on Information Extraction (IE) noticed that it is essential to recognize information units like names including person, location, organization, money and percentage expressions. Identifying references to these entities in a text was recognized as one of the important preprocessing step of IE and was called NER and Classification [2].

Question Answering systems ,an important application of Artificial Intelligence(AI) mostly requires retrieval of nouns or noun phrases as answers to the questions raised by the users. This paper addresses the problem of NER in a Question Answering (QA) system which involves detection of named entities based on the keywords given in the query [1]. To find the answer, questions are analyzed and determine the type of expected answers. Then the answer retrieval module retrieves either a document, passage or a phrase to answer the query. Questions such as 'when', 'where', 'what',' why' etc. mostly requires an answer that contains a noun or a noun phrase. Then the required information is retrieved from the stored document.

NER is an important problem since search queries are often with respect to nouns especially proper nouns. Also some human names are selected from names of common nouns. Sometimes same names are used for organizations, pets, locations etc. Also people are daily making new names for person, organization and location. All these names cannot be maintained in a dictionary and also it is a time consuming task [3].

A lot of work has been done in the field of NER for English and European Languages. In English Capitalization is a major clue for identifying person names. Some efforts have been made for Telugu, Hindi and Bengali. As of now we have no information regarding Malayalam NER work and no tag set is been identified so far.

Malayalam belongs to the Dravidian family of languages and is one of the 4 major languages of this family. It is one of the 22 scheduled languages of India with official language status in the state of Kerala. It is spoken by 35.9 million people. Malayalam is a morphologically rich agglutinative language and relatively of free order. Also Malayalam has a productive morphology that allows the creation of complex words which are often highly ambiguous [4].

Support Vector Machines (SVM) is a supervised machine learning algorithm which analyzes data and recognizes patterns used for statistical classification and regression analysis [5]. SVM is a linear classifier for solving binary classification problem which can be extended to multiclass classification.

Table1. Few Examples of Named Entities

Questions	Answer type	
	Named Entity	**Lexical Category**
Where is your house?	LOCATION	Proper Noun/Noun Phrase
Who played cricket today?	PERSON/ORGANISATION	Proper Noun
When was it happened?	DATE	Number
How was it happened?	DESCRIPTION	Noun Phrase
How much is the distance from here?	DISTANCE	Number

Linguistic features are mapped to SVM feature vectors and setting these feature to 1 if it exists else 0 otherwise. These feature vectors are used to decide whether a word is a NE or not and class of NE in which the word belongs to.

2. RELATED WORKS

NER systems can be categorized into three classes namely Hand-made Rule based NER, Machine learning based NER and Hybrid NER.

Rule based NER is a linguistic approach for extracting names that uses rules prepared by experienced computational linguistics. Such a system may provide better accuracy and precision but at the cost of lower recall. Developing and maintaining rules and dictionaries is a time consuming task. Also these systems are not transferable i.e. systems developed for one domain cannot be used for another language or domain. Several rule based systems containing lexicalized grammar, gazetteer lists and a list of trigger words are available for English and few Indian languages. Some are found in [6] and [7].

Machine learning methods are using either supervised learning or unsupervised learning techniques. Supervised learning can achieve good performance when large amount of high quality training data is available. The training data must be labeled with all of the entities of interest and their types. The training data should match the data on which, the system will run, otherwise performance will degrade. Hidden Markov model (HMM), Maximum Entropy Markov Model (MEMM), Conditional Random Field (CRF) and Decision trees are some of the statistical models proposed based on supervised learning methods. MEMM is one of the most popular statistical model that is been applied in [8] and [9].HMM is another well known model that can be found in [10].

Ekbal and Bandyopadhyay described an approach to lexical pattern learning for Indian Languages in [11].Hindi NER system is developed using CRF with feature induction [12]. An approach to identify nested NEs is presented in [13] where NER problem is treated as a binary classification problem and solved it using SVM.

Hybrid Methods either use combinations of different machine learning methods or combinations of rule based and machine learning methods. [14] Presents a cascaded hybrid model for Chinese NER.

3. NAMED ENTITIES AND FEATURE SELECTION

When we analyze the questions which are asked in a QA system, it is clear that the expected answers are either nouns/noun phrases, more specifically named entities. A list of frequently asked questions and expected answer types are given in table 1.

3.1 Named Entities

We have designed a NE tag set for five domains which includes 95 tags altogether and where 15 tags are exclusively for health domains. The named entity extractor identifies person names, diseases, organizations, fertilizers, locations, dates etc.

Table2: NE tags for a health Question Answering system

PERSON	LOCATION	SYMPTOMS
ORGANISATION	DISEASE	AGE
GEAN	MEDICINE	TIME
CARRIER	MEANS	DATE
JOURNAL	PRECAUTION	DURATION

3.2 Features

NE's are theoretically identified and classified by using features (various abstract entities that combine to specify underlying phonological, morphological, semantic, and syntactic properties of linguistic forms and that act as the targets of linguistic rules and operations).Two kinds of features that have been commonly used are internal and external features; internal features are the ones provided from within the sequence of words that constitute the entity, in contrast, external features are those that can be obtained by the context in which entities appear[15].

In this system clues present as inner word (internal feature) and context word (external features) are used for NE identification and classification. Some of these features are Language independent features while others are language dependant features.

3.2.1 Language Independent feature

3.2.1.1 Word Prefix/Suffix

The term Prefix/suffix means any sequence of first/last characters of word information of current word and surrounding words can be treated as features. Prefix/Suffix information of a current and surrounding words are useful in highly inflected language like Malayalam. Word suffix information is useful to identify the named entities. Variable length suffixes can be matched against the list of suffixes for different classes of NEs.
A list of linguistic suffixes (verbs, adjectives, adverbs, nouns) is prepared which helps to recognize 'Not a Named Entity' cases.
Certain suffixes are helpful in detecting Named Entities such as PERSON, LOCATION etc.

3.2.1.2 Digit Information

This gives word level orthographic information. If a word contains digits or special symbols corresponding binary feature is set to 1 otherwise 0.Three binary valued features are considered depending upon the presence of digits and/or the number of tokens, combination of digits and symbols. These binary valued features are helpful in recognizing DATE expressions, TIME expressions, AGE, YEAR etc.

3.2.1.3 Length of the word

If the length of the word is greater than two, the binary feature 'Length' is set to 1; otherwise it is set to 0. Named Entities, since they are nouns/noun phrases or open class entities they are rarely shorter words.

3.2.1.4 Position

There are two binary valued features; the feature 'First word' is set to 1 if the current word is the first word of the sentence else set to 0 and the feature 'Last word 'is set to 1 if the given word occurs at the end of the sentence.

3.2.1.5 Frequent word list

According to Luhn words with very high frequency and very low frequency are not sense carrying agents, either they are rare words or closed class words.

3.2.1.6 Surrounding words

Previous and next words of a particular word are used as features. This feature is multivalued. Different window sizes were used for different domains in different experiments.

3.2.1.7 NE Information

The NE tags of the previous words are used dynamic feature.

3.2.2 Language Dependant feature

3.2.2.1 Parts of Speech information

POS of the current word and the surrounding words are important to recognize NEs. We have used our own POS tagger developed with Extended CRF. POS also indicates whether the word is a 'standalone word' or part of a phrase.

3.2.2.2 Phrase Chunk Information

Certain NEs are noun phrases which appear either directly or as a part of other phrases such as postpositional phrases. Phrase chunks are labeled using Artificial Immunity System based chunker developed by us.

4. METHODOLOGY

This paper describes the application of SVM to the task of NER. Given a natural language input sequence $W1^N = w_1w_2w_3...w_n$ we choose the NE tag sequence $C1^N = c_1c_2c_3...cn$ with the highest probability among all possible tag sequences.

4.1 Support Vector Machines

Support vector Machines proposed by Vapnik is a set of Machine learning algorithms based on statistical methods. It is known as one of the best supervised learning algorithms and has been successfully applied to natural language tasks such as text categorization, phrase chunking etc.

Assigning a Named Entity Label to a word in the sentence can be taken as a classification task which is common in machine learning. A word may belong to one of the two classes and then the goal is to decide which class the word will be in. SVM constructs hyper planes to classify the words or data points. Any hyper plane can be defined by the following (W. x) +b=0 where w is a normal vector and it is perpendicular to the hyperplane. The parameter $\frac{b}{\|\mathbf{w}\|}$ determines the offset of the hyperplane from the origin along the normal vector $\mathbf{W}$ [13].

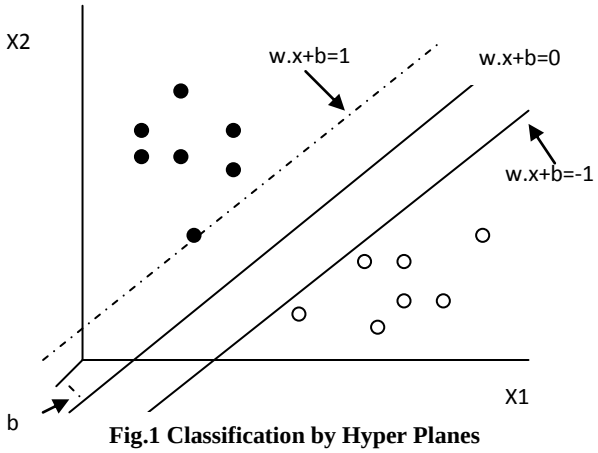

Fig.1 Classification by Hyper Planes

We can select the two hyperplanes of the margin in a way that there are no points between them and then try to maximize their distance.For the best classification we have to choose the best hyperplane that is the one that represents the largest separation, or margin, between the two classes.

In its simplest form training an SVM amounts to finding the hyper plane that separates the +ve training samples from the -ve samples by the largest possible margin. This hyper plane is then used to classify the test vectors, those that on one side of the hyper plane are classified as members of +ve class while the others are classified as members of –ve class.

5. NAMED ENTITY RECOGNIZER FOR MALAYALAM NLP APPLICATIONS

NER involves the identification of named entities such as person, location, dates, Diseases etc. NER emerged as subtask of DARPA sponsored Message Understanding Conference. In the taxonomy of Computational Linguistics, NER falls within the category of IE which deals with the extraction of specific information from the given documents.

The system is trained over a corpus using SVM learning method.200 documents are collected from various journals, dailies and web sites including articles on agriculture, sports, science, health and politics. This collection is then split into training and test sets. From the training data following features were extracted for each kind of named entity. For the test data, features are extracted for each word and encoded them into vectors. Then using the multiclass SVM each word is classified into the most appropriate category of named entity.

5.1 Preprocessing Stage

5.1.1 POS Tagger

Working of this tagger in fig.2 is as follows. The input document is divided into tokens. Then each token is sent to the word analyzer for the detailed analysis. First it checks the token to decide whether it is a compound word or not. If it is a simple word the local information is collected from the lexicon. Else the token is sent to the compound word splitter to find out the morphological details and the constituents of it. The tagger assigns to each token, all possible POS tags with the help of local information provided by the word analyzer. Then to resolve the ambiguity, Extended Conditional Random Field Disambiguator is used with contextual information and eliminates all but one tag.

5.1.2 Phrase Chunker

For the phrase chunker a POS tagged document is the input. Using the phrase grammar the POS string can be analyzed to identify the phrase chunks.

Most of the sentences in a Malayalam documents are compound or complex sentences. Hence to separate the phrase chunks first of all clauses are to be separated. Each clause obtained from the clause identifier (fig.2) is taken one by one and sent to the Phrase separator. All the phrases corresponding to each clause is identified and separated by the phrase separator. Then they are labeled with phrase tags.

5.2 Named Entity Marker

In our application any word may fall into more than two classes. Hence we have used multiclass SVM which assigns suitable label to each word from a finite set of classes'.

Multiclass SVM is implemented by reducing multi classifier problem into multiple binary classifiers. For every pair of distinct classes a binary classifier is constructed, hence all together M (M-1)/2 binary classifiers are constructed where M is the total number of named entity classes [16]. The binary classifier Cij is trained taking the examples from class Wi as positive and the examples from Wj as negative. For a new example x if classifier Cij say x is in class Wi, then the vote for the class Wi is added by one. Otherwise the vote for class Wj is increased by one. After each of the M (M-1)/2 binary classifier makes its vote, MWV (Max- Wins Vote) strategy assigns x to the class with the largest number of votes.

For example in fig3 there are three classes A, B and C. Using MWV strategy, 3*(3-1)/2= 3 binary classifiers are created namely $C_{AB,}$ C_{BC} and $C_{CA.}$ C_{AB} classifies word in the test data as

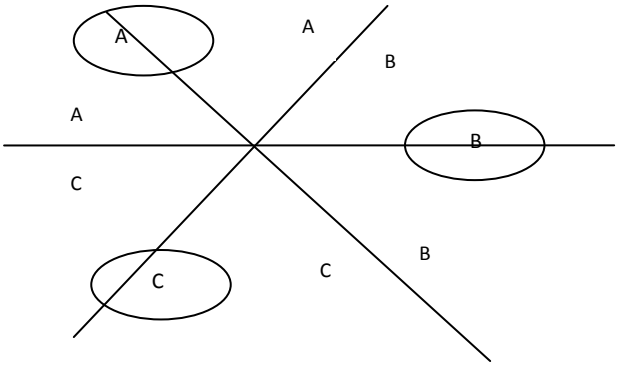

Fig 3. Diagram of Pairwise SVM Decision boundaries on a basic Problem

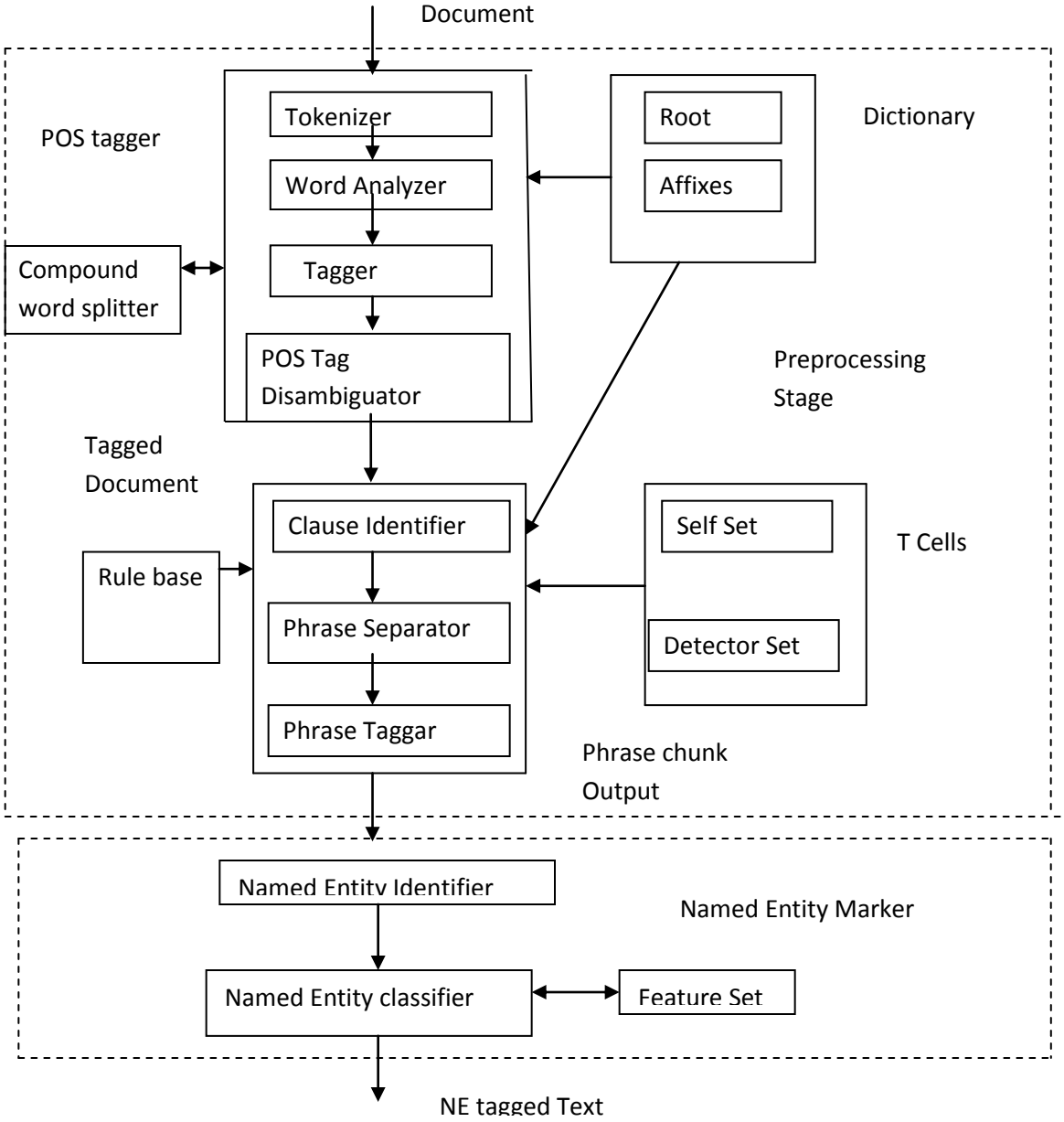

Fig.2 Detailed Architecture of Named Entity Recognizer

of class A or class B by increasing the count corresponding to A or B. C_{BC} and C_{CA} repeats the same process and increases the count corresponding A or B or C. Named entity class with the highest count is taken as the NE tag of the given word.

6. TESTS AND DISCUSSIONS

NER is designed and implemented using J2SDK1.4.2 and MySQL. Its performance is evaluated using standardized techniques precision, recall and F-score where Precision is defined as a ratio of number of correct NER tags to the number of NER tags in the output and recall is the ratio of number of correct NER tags to the number of NER tags in the test data. F-score = 2*recall*precision/ (recall+ precision) [17].

Documents related to five different fields are selected as the corpus. Then we randomly selected 8000 sentences for training and 2000 sentences as test set. Precision, recall and F-score obtained for various types of NE tags are shown in table3.Highest recall is 96.21 for the entity LOCATION and highest precision for the entity JOURNAL Highest F-Score is obtained for the named entity JOURNAL.

We could overcome the following challenges raised by the Malayalam language features by considering the word level and phrase level information i.e. by morphological analysis, POS tagging and phrase chunking.

- Agglutinative Nature

Malayalam is a highly inflectional and agglutinative language. 85% of words in Malayalam text are compound words and hence role of these words can be decided only by knowing its components and their types. Role of an entity depends on the importance of the word which is decided by local and global information. To derive local information, each word is analyzed and collected its component details.

- Word Order

Malayalam sentence is a sequence of words where words may appear in any order and each word can be a combination of any number of stems and affixes. Even though there is no specific order for the words in the sentence, within a chunk word categories are related.

In Malayalam language there is no distinction between uppercase and lowercase. Hence proper techniques are to be adopted to overcome such challenges

7. CONCLUSION

Data classification, Question Answering, Information Retrieval (IR) and Machine Translation (MT) are some of the applications of Natural Language Processing. NER plays an important role in accurate answer retrieval and efficient MT systems. The objective of NER is to categorize "all sense carrying" words in a document into predefined classes like PERSON, LOCATION, DISEASES and miscellaneous.

An SVM based system is prepared for the NER task in Malayalam. Results show that Multiclass SVM is a promising method for Malayalam NE classification. Approach presented here requires linguistic preprocessing of the document text such as Morphological analysis, Part of Speech Tagging and phrase chunking.

Future developments includes a domain independent NER and its application in any QA/IE system.

Table3. NER Performance by named entity type

Named Entity	Recall	Precision	F-Measure
PERSON	95.32	94.28	94.78
ORGANISATION	89.25	88.46	88.85
LOCATION	96.21	95.30	95.75
DISEASE	88.41	89.53	88.97
SYMPTOMS	86.38	84.14	85.25
CAUSE	82.18	83.60	82.88
AGE	87.64	89.43	88.52
TIME	90.40	92.52	91.44
DATE	92.67	91.45	92.05
DURATION	91.20	90.16	90.68
CARRIER	95.47	94.50	94.98
JOURNAL	96.00	97.23	96.61
MEDICINE	92.87	94.75	93.80
MEANS	74.49	76.28	75.37
PRECAUTION	78.34	75.65	76.97

8. REFERENCES

[1] Diego Moll ´a and Menno van Zaanen and Daniel Smith, "Named Entity Recognition for Question Answering", Proceedings of the 2006 Australasian Language Technology Workshop (ALTW2006), pages 51–58

[2] 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.

[3] Lev Ratinov Dan Roth, "Design Challenges and Misconceptions in Named Entity Recognition", Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL), pages 147–155, Boulder, Colorado, June 2009

[4] A .R .Rajarajavarma," Keralapanineeyam", National Book Stall, Kottayam, 2000.

[5] Asif Ekbal and Sivaji Bandyopadhyay," Named Entity Recognition using Support Vector Machine: A Language Independent Approach", International Journal of Electrical and Electronics Engineering 4:2 2010

[6] Kashif Riaz , "Rule-based Named Entity Recognition in Urdu", Proceedings of the 2010 Named Entities Workshop, ACL 2010, pages 126–135, Uppsala, Sweden, 16 July 2010.

[7] B. Sasidhar, P. M. Yohan,Dr. A. Vinaya Babu, Dr. A. Govardhan," Named Entity Recognition in Telugu Language using Language Dependent Features and Rule based Approach", International Journal of Computer Applications (0975 – 8887) Volume 22– No.8, May 2011

[8] Mohammad Hasanuzzaman1, Asif Ekbal2 and Sivaji Bandyopadhyay3," Maximum Entropy Approach for Named Entity Recognition in Bengali and Hindi", International Journal of Recent Trends in Engineering, Vol. 1,No.1, May 2009

[9] Kitoogo Fredrick Edward, Venansius Baryamureeba and Guy De Pauw; Towards Domain Independent Named Entity Recognition, International Journal of Computing and ICT Research, Vol. 2, No. 2, pp. 84 –95

[10] GuoDong Zhou Jian Su ," Named Entity Recognition using an HMM-based Chunk Tagger", "Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, July 2002, pp. 473-480.

[11] Asif Ekbal and Sivaji Bandyopadhyay, "Named Entity Recognition Using Appropriate Unlabeled Data, Post-processing and Voting", Informatica 34 (2010) 55–76

[12] Burr Settles," Biomedical Named Entity Recognition Using Conditional Random Fields and Rich Feature Sets", Proceedings of the COLING 2004 International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA). Geneva, Switzerland. 2004.

[13] Sujan Kumar Saha *, Shashi Narayan, Sudeshna Sarkar, Pabitra Mitra," A composite kernel for named entity recognition", Pattern Recognition Letters (2010)

[14] Xiaofeng Yu,"Chinese Named Entity Recognition with Cascaded Hybrid model",Proceedings of NAACL HLT 2007 Companion Volume, pp 197-200,April 2007

[15] Gjorgji Madzarov, Dejan Gjorgjevikj and Ivan Chorbev," A Multi-class SVM Classifier Utilizing Binary Decision Tree", Informatica 33 (2009) 233-241

[16] Samir AbdelRahman, Mohamed Elarnaoty, Marwa Magdy and Aly Fahmy," Integrated Machine Learning Techniques for Arabic Named Entity Recognition", IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 4, No 3, July 2010

[17] Alireza Mansouri,Lilly Suriani Affendey,Ali Mamat," Named Entity Recognitin Approaches",International Journal of Computer Science and Network Security,Vol.8,No.2,Februart,2008.