Named Entity Recognition: A Survey for the Indian Languages

Padmaja Sharma
Dept. of CSE
Tezpur University
Assam, India 784028
psharma@tezu.ernet.in

Utpal Sharma
Dept.of CSE
Tezpur University
Assam, India 784028
utpal@tezu.ernet.in

Jugal Kalita
Dept.of CS
University of Colorado at Colorado Springs
Colorado, USA 80918
kalita@eas.uccs.edu

***Abstract*—Named Entity Recognition(NER) is the process of identifying and classifying all proper noun into pre-defined classes such as persons, locations, organization and others. Work on NER in Indian languages is a difficult and challenging task and also limited due to scarcity of resources, but it has started to appear recently. In this paper we present a brief overview of NER and its issues in the Indian languages. We also describe the different approaches used in NER and also the work in NER in different Indian languages like Bengali, Telugu, Hindi, Oriya and Urdu along with the methodologies used. Lastly we presented the results obtained for the different Indian languages in terms of F-measure.**

I. Introduction

Natural Language Processing (NLP) is the computerized approach for analyzing text that is based on both a set of theories and a set of technologies. Named Entity Recognition (NER) is an important task in almost all NLP areas such as Machine Translation (MT), Question Answering (QA), Automatic Summarization (AS), Information Retrieval(IR), Information Extraction(IE), etc.

NER can be defined as a two stage problem - Identification of the proper noun and the classification of these proper noun into a set of classes such as person names, location names (cities, countries etc), organization names (companies, government organizations, committees, etc.), miscellaneous names (date, time, number, percentage, monetary expressions, number expressions and measurement expressions). Thus NER can be said as the process of identifying and classifying the tokens into the above predefined classes.

II. Basic Problems in Named Entity Recognition

The basic problems of NER are-

1) Common noun Vs proper noun- Common noun sometimes occurs as a person name such as “Suraj” which means sun, thus creating ambiguities between common noun and proper noun.
2) Organization Vs person name- “Amulya” as a person name as well as an organization, that creates ambiguity between proper noun and group indicative noun.
3) Organization Vs place name- “Tezpur” which act both as an organization and place name.
4) Person name Vs place name- When is the word “Kashi” being used as a person name and when as the name of a place.

Two broadly used approaches in NER are:

1) Rule-based NER
2) Statistics-based NER

Statistical methods such as Hidden Markov Model (HMM) [1], Conditional Random Field (CRF) [2], Support Vector Machine (SVM) [3], Maximum Entropy (ME) [4], Decision Tree (DT) [5] are the most widely used approaches. Besides the above two approaches, NER also make use of the Hybrid model which combines the strongest point from both the Rule based and statistical methods. This method is particularly used when data is less and complex Named Entities (NE) classes are used. Sirhari et.al [6] introduce a Hybrid system by combination of HMM, ME and handcrafted grammatical rules to build an NER system.

III. Problem faced in Indian Languages(ILs)

While significant work has been done in English NER, with a good level of accuracy, work in IL has started to appear only very recently. Some issues faced in Indian languages-

1) There is no concept of capitalization of leading characters of names in Indian Languages unlike English and other European languages which plays an important role in identifying NE’s.
2) Indian languages are relatively free-order languages.
3) Unavailability of resources such as Parts of speech (POS) tagger, good morphological analyzer, etc for ILs. Name lists are found available in web which are in English but no such lists for Indian Languages can be seen.
4) Some of the Indian languages like Assamese, Telugu are agglutinative in nature.
5) Indian languages are highly inflectional and morphologically rich in nature.

IV. Methodologies/Approaches

NER system can either be Rule-based or Statistics based. Machine Learning techniques(MLT)/Statistics based methods described below are successfully used for NER .

A. Hidden Markov Model (HMM):

HMM is a statistical model in which the system being modeled is assumed to be a Markov process with unobserved state. In this approach the state is not directly visible, but output depends on the state and is visible. Instead of single independent decisions, the model considers a sequence of decisions. Following are the assumptions of HMM-

- Each state depends on its immediate predecessor.
- Each observation value depends on the current state.
- Need to enumerate all observations.

.

The equation for HMM is given as-
$P(X) = \sum \prod_{i=0}^{n} P(y_i(y_{i-1})p(x_i|y_i))$

where,

$$X = (x_1, ..., x_n)$$

$$Y = (y_1, ..., y_n)$$

B. Conditional Random Field (CRF):

CRF are undirected graphical models a special case of which corresponds to conditionally trained finite state machines. They can incorporate a large number of arbitrary, non independent features and is used to calculate the conditional probability of values on designated output nodes given values on other designated input nodes. The conditional probability of a state sequence $S = (s1, s2..sT)$ given an observation sequence $O = (o1, o2, o3...ot)$ is calculated as

$$P(s|o) = \frac{1}{Z_o} \exp(\sum_{t=1}^{T} \sum_{k} \lambda_k f_k(S_{t-1}, S_t, o, t)$$

Where Z_o is a normalization factor overall state sequence.

$$Z_o = \sum \exp(\sum_{t=1}^{T} \sum_{k} \lambda_k f_k(S_{t-1}, S_t, o, t)$$

and $f_k(S_{t-1}, S_t, o, t)$ is a feature function whose weight λ_k is to be learned via training.

C. Support Vector Machine(SVM):

SVM first introduced by Vapnik are relatively new machine learning approaches for solving two-class pattern recognition problem. In the field of NLP, SVM is applied to text categorization and are reported to have high accuracy. It is a supervised machine learning algorithm for binary classification.

D. Maximum Entropy (ME):

The Maximum Entropy framework estimates probabilities based on the principle of making as few assumptions as possible other than the constraints imposed. Such constraints are derived from training data, expressing some relationship between features and outcomes. The probability distribution that satisfies the above property is the one with the highest entropy and has the exponential form

$P(o|h) = \frac{1}{z(h)} \prod_{j=1}^{k} \propto_j f_j(h, o)$

where o refers to the outcome, h the history(or context) and z(h) is a normalization function. In addition each feature function $f_j(h, o)$ is a binary function. The parameter $\propto_j$ are estimated by a procedure called Generalized Iterative Scaling(GIS) [7]. This is an iterative method that improves the estimation of the parameter at each iteration.

E. Decision Tree (DT):

DT is a powerful and popular tool for classification and prediction. The attractiveness of DT is due to the fact that in contrast to neural network, it represents rules. Rules can readily be expressed so that human can understand them or even directly use them in a database access language like SQL so that records falling into a particular category may be tree.

Decision Tree is a classifier in the form of a tree structure where each node is either a leaf node-indicates the value of the target attributes(class)of expressions, or a decision node that specifies some test to be carried out on a single attribute value with one branch and sub-tree for each possible outcome of the text. It is an inductive approach to acquire knowledge on classification.

V. Existing work on different Indian languages in NER

A. Hindi

Saha et.al(2008) [8] describes the development of Hindi NER using ME approach. The training data consists about 234 k words,collected from the newspaper "Dainik Jagaran" and is manually tagged with 17 classes including one class for not name and consists of 16,482 NEs. The paper also reports the development of a module for semi-automatic learning of context pattern. The system was evaluated using a blind test corpus of 25K words having 4 classes and achieved an F-measure of 81.52%.

Goyal(2008) [9] focuses on building a NER for Hindi using CRF. This method was evaluated on test set1 and test set 2 and attains a maximum F1-measure around 49.2% and nested F1-measure around 50.1% for test set1 maximum F1-measure around 44.97% and nested F1-measure around 43.70% for test set2 and F-measure of 58.85% on development set.

Saha et.al(2008) [10] has identified suitable features for Hindi NER task that are used to develop an ME based Hindi NER system. Two-phase transliteration methodology was used to make the English lists useful in the Hindi NER task. The system showed a considerable performance after using the transliteration based gazetteer lists. This transliteration approach is also applied to Bengali besides Hindi NER task and is seen to be effective. The highest F-measure achieved by ME based system is 75.89% which is then inceased 81.2% by using the transliteration based gazetteer list.

Li and McCallum(2004) [11] describes the application of CRF with feature induction to a Hindi NER. They discover

relevant features by providing a large array of lexical test and using feature induction to construct the features that increases the conditional likelihood. Combination of Gaussian prior and early-stopping based on the results of 10-fold cross validation is used to reduce over fitting.

Gupta and Arora(2009) [12] describes the observation made from the experiment conducted on CRF model for developing Hindi NER. It shows some features which makes the development of NER system complex. It also describes the different approaches for NER. The data used for the training of the model was taken from Tourism domain and it is manually tagged in IOB format.

B. Bengali

It is the seventh popular language in the world, second in India and the national language of Bangladesh. Ekbal and Bandyopadhyay(2009) [13] reports about the development of NER in Bengali by combining the output of the classifier like ME, CRF and SVM. The training set consists of 150k word form to detect the four Named Entity tags namely person, location, organization and miscellaneous names. Lexical context pattern generated from an unlabeled Bengali corpus containing 3 million wordform have been used to improve the performance of the classifier. Evaluation results of 30K wordforms have found the overall recall, precision and f-score values of 87.11%, 83.61% and 85.32%, which shows an improvement of 4.66% in f-score over the best performing SVM based system and an improvement of 9.5% in f-score over the least performing ME based system.

On the other hand work by Ekbal et.al [14] shows the development of Bengali NER system using the statistical CRF. The system make use of different contextual information of the words along with the variety of features for identifying Named Entity classes. The training set comprises of 150k wordform which is manually annotated with 17 tags. Experimental results of the 10-fold cross validation test shows the effectiveness of proposed CRF based NER system with an overall average recall, precision and f-score values of 93.8%, 87.8% and 90.7%.

Ekbal and Bandyopadhyay(2010) [15] developed NER system for Hindi and Bengali using SVM. An annotated corpora of 122,467 tokens of Bengali and 502,974 tokens of Hindi has been used tagged with 12 NE classes. The NER system has been tested with the gold standard test sets of 35K, and 60K tokens for Bengali and Hindi. Evaluation results have demonstrated the recall, precision and f-score of 88.61%, 80.12% and 84.15% for Bengali whereas 80.23%, 74.34% and 77.17% for Hindi.

Hasan et.al(2009) [16] presented a learning-based named entity recognizer for Bengali that donot rely on manually-constructed gazetteers in which they developed two architectures for the NER system. The corpus consisting of 77942 words is tagged with one of 26 tags in the tagset defined by IIT Hyderabad where they used CRF++ to train the POS tagging model. Evaluation results shows that the recognizer achieved an improvement of 7.5% in F-measure over a baseline recognizer.

Chaudhuri and Bhattacharya(2008) [17] has made an experiment on automatic detection of Named Entities in Bangla. Three-stage approach has been used namely-dictionary based for named entity, rules for named entity and left-right co-occurrences statistics. Corpus of Anandabazar Patrika has been used from the year 2001-2004. The manual tagging was done by the linguistic based on the global knowledge. Experimental results has shown the average recall, precision and f-measure to be 85.50%,94.24% and 89.51%.

Ekbal and Bandyopadhyay(2008) [18] developed NER system for Bengali using SVM. The system makes use of the different contextual information of the words along with the variety of features that are helpful in predicting the Named entities. A partially NE tagged Bengali news corpus has been used to create the training set for the experiment and the training set consists of 150K wordforms that is manually tagged with 17 tags. Experimental results of the 10 fold cross validation test shows the effectiveness of the proposed SVM based NER system with the overall average recall, precision and F-score values of 94.3%, 89.4% and 91.8%.

Ekbal and Bandyopadhyay(2008) [19] reports about the development of Bengali news corpus from the web consisting of 34 million wordforms. A part of this corpus of 150K wordforms is manually tagged with 16 NE and one non NE tag and additionally 30 K wordforms is tagged with a tagset of 12 NE tags defined for the IJCNLP-08 NER shared task for SSEAL. A tag conversion routine has been developed to convert the 16 NE tagged corpus of 150 K wordforms to the corpus tagged with IJCNLP-08 12 NE tags where the former has been used to develop the Bengali NER system using HMM, ME,CRF, SVM. Evaluation results of the 10 fold cross validation tests gives the F-score of 84.5% for HMM, 87.4% for ME, 90.7% for CRF and 91.8% for SVM.

Ekbal and Bandyopadhyay(2008) [20]describes the development of a web-based Bengali news corpus consisting of 34 million wordforms.The performance of the system is compared for two system- one is by using the lexical contextual patterns and the other using linguistic features along with the same set of lexical contextual pattern and came with the conclusion that the use of linguistic knowledge yields an highest F-value of 75.40%, 72.30%, 71.37% and 70.13% for person, location, organization and miscellaneous names.

Ekbal and Bandyopadhyay(2009) [21] describes a voted NER system by using Appropriate Unlabeled Data. This method is based on supervised classifier namely ME, SVM, CRF where SVM uses two different system known as forward parsing and backward parsing. The system has been tested for Bengali containing 35,143 news document and 10 million wordfroms and makes use of language independent features along with different contextual information of the words. Finally the models have been combined together into a final system by a weighted voting technique and the experimental

results show the effectiveness of the proposed approach with the overall recall precision and f-score values of 93.81%, 92.18% and 92.98%.

Ekbal and Bandyopadhyay(2008) [22] reports about the development of NER system in Bengali by combining the outputs of the classifier like ME, CRF, SVM. The corpus consisting of 250K wordforms is manually tagged with four NE namely person, location, organization and miscellaneous. The system makes use of different contextual information of the words along with the variety of features that helps in identifying the NEs. Experimental results shows the effectiveness of the proposed approach with the overall average recall, precision and f-score values of 90.78%, 87.35% and 89.03% respectively. This shows an improvement of 11.8% in f-score over the best performing SVM based baseline system and an improvement of 15.11% in f-score over the least performing ME based baseline system.

Hasanuzzaman et.al(2009) [23] describes the development of NER system in Bengali and Hindi using ME framework with 12 NE tags. A tag conversion routine has been developed in order to convert the fine-grained NE tagset of 12 tags to a coarse-grained NE tagset of 4 tags namely person name, location name, organization name and miscellaneous name. The system makes use of different contextual information of the words along with the variety of orthographic word - level features that helps in predicting the four NE classes. Ten fold cross validation test results the average recall, precision and f-measure of 88.01%, 82.63%, 85.22% for Bengali and 86.4%, 79.23% and 82.66% for Hindi.

Ekbal and Bandyopadhyay(2007) [24] reported the development of HMM based NER system. For Bengali it was tested manually over a corpus containing 34 million wordforms developed from the online Bengali newspaper. A portion of the tagged news corpus containing 150,000 wordforms is used to train the NER system through HMM-based parts of speech tagger with 26 different POS tags and the training set thus obtained is a corpus tagged with 16 NE tags and one non NE tag and the experimental results of the 10-fold cross validation yields an average Recall, Precision and F-score values of 90.2%, 79.48% and 84.5% respectively. After this the HMM-based NER system is also trained and tested with Hindi data to show the effectiveness for the language independent features. Tthe results for Hindi NER shows an average Recall, Precision and F-score values of 82.5%, 74.6% and 78.35% respectively.

C. Telugu

Telugu being a language of the Dravidian family, is the third most spoken language in India and official language of Andhra Pradesh.

Srikanth and Murthy (2008) [25] have used part of the LERC-UoH Telugu corpus where CRF based Noun Tagger is built using 13,425 words manually tagged data and tested on a test data set of 6,223 words and came out with an F-measure of 91.95%. Then they develop a rule-based NER system consisting of 72,152 words including 6,268 Named Entities where they identified some issues related to Telegu NER and later develop a CRF based NER system for telegu and obtained an overall F-measures between 80% and 97% in various experiments.

Shishtla et.al(2008) [26] conducted an experiment on the development data released as a part of NER for South and South East Asian Languages (NERSSEAL) Competition. The Corpus consisting of 64026 tokens was tagged using the IOB format (Ramshaw and Marcus, 1995). The author have showed experiments with various features for Telugu. The best performing model gave an F-1 measure of 44.91%.

Raju et.al [27] have developed a Telugu NER system by using ME approach. The corpus was collected from the iinaaDu, vaarta news papers and Telugu Wikipedia. Manually tagged test data is prepared to evaluate the system. The system makes use of the different contextual information of the words and Gazetteer list was also prepared manually or semi-automatically from the corpus and came out with a an F-measure of 72.07% for person, 6.76%, 68.40% and 45.28% for organization, location and others respectively.

D. Tamil

VijayKrishna and Sobha(2008) [28] developed a domain specific Tamil NER for tourism by using CRF. It handles morphological inflection and nested tagging of named entities with a heirarchial tageset consisting of 106 tags. A corpus of 94k is manually tagged for POS, NP chunking, and NE annotations. The corpus is divided into training data and the test data where CRF is trained with the former one and CRF models for each of the levels in the hierarchy are obtained. The system comes out with a F-measure of 80.44%.

Pandian et.al(2008) [29] presented a hybrid three-stage approach for Tamil NER. The E-M(HMM) algorithm is used to identify the best sequence for the first two phases and then modified to resolve the free-word order problem. Both NER tags and POS tags are used as the hidden variables in the algorithm. Finally the system comes out with an F-measure of about 72.72% for various entity types.

E. Oriya

Biswas et.al [30] presented a hybrid system for Oriya NER that applies both ME and HMM and some handcrafted rules to recognize NEs. Firstly the ME model is used to identify the named entities from the corpus and then this tagged corpus is regarded as training data for HMM which is used for the final tagging. Different features have been considered and linguistic rules help a lot for identification of named entities. The annotated data used in the system is in IOB format. Finally the system comes with an F-measure between 75% to 90%.

VI. Analysis

From the above survey we have seen that though the work in NER in IL is limited, still considerable work has been done for the Bengali language. The level of accuracy obtained for these languages are described in the (Table 1, 2) along with the approaches used. We can see that CRF is the most widely used approach which shows an effective results for the Indian Languages in comparison to the other approaches. Our survey reveals that Ekbal and Bandyopadhyay [18] achieved highest accuracy using CRF 90.7%, using SVM 91.8, using ME 87.4% and using HMM 84.5% for Bengali.

VII. Conclusion and Future Work

In this survey we have studied the different techniques employed for NER, and have identified the various problems in the task particularly for ILs. In addition to these approaches researchers can also try using other approaches like DT, Genetic algorithm, Artificial and Neural Network etc that which already showed an excellent performance in the other languages like English, Germany etc. Also NER should be attempted for other IL in which no such work has been attempted so far.

TABLE I
Comparison of the approaches with their accuracy for the different Indian Languages. Fm : Maximal F-measure, Fn : Nested F-Measure, Fl: Lexical F-measure, BIA : Baseline Induced Affixes, BIAW : Baseline Induced Affixes Wiki: Classifier- outputs of ME, CRF,SVM.

Language	Author	Approach	Accuracy(%)
Telugu	[25]	CRF	80-97
	[26]	CRF	44.91
	[27]	ME	P-72.07
			O-60.76
			L-68.40
			Others-45.28
Tamil	[28]	CRF	80.44
	[29]	HMM	72.72
Hindi	[10]	ME	75.89
	[8]	ME	81.52
	[9]	CRF	58.85
Bengali	[18]	SVM	91.8
	[17]	n-gram	89.51
	[14]	CRF	90.7
	[13]	Classifiers	85.32
	[19]	MLT	HMM- 84.5
			ME -87.4
			CRF -90.7
			SVM -91.8
	[21]	Classifier	92.98
	[22]	Classifier	89.03
	[20]	MLT	P-75.40
			L-72.30
			O-71.37
			Others-70.13
	[16]	CRF	Baseline-65.57
			BIA-69.32
			BIAW-71.99
Bengali+Hindi	[15]	SVM	Bengali-84.15
			Hindi-77.17
Bengali+Hindi	[23]	ME	Bengali-85.22
			Hindi-82.66
Bengali+Hindi	[24]	HMM	Bengali-84.5
			Hindi-78.35

TABLE II
Comparison of the approaches with their accuracy for South and South East Asian Languages

Author	Approach	Language	Fm	Fn	Fl	F measure
[31]	CRF	Bengali	53.36	53.46	59.39	-
[32]	ME	Hindi	-	-	-	65.13
		Bengali	-	-	-	65.96
		Oriya		-	-	44.65
		Telugu	-	-	-	18.74
		Urdu		-	-	35.47
[33]	CRF	Bengali	35.65	33.94	40.63	-
		Hindi	48.71	50.47	50.06	-
		Oriya	29.29	26.06	39.04	-
		Telugu	8.19	43.19	40.94	-
		Urdu	39.86	39.01	43.46-	-
[34]	ME	Bengali	12.50	11.97	12.30	-
		Hindi	29.24	28.48	25.68	-
		Oriya	13.94	11.91	19.44	-
		Telugu	00.32	01.08	08.75	-
		Urdu	26.41	24.39	27.73-	-
[35]	CRF	Bengali	31.48	30.79	35.71	-
		Hindi	42.27	41.56	40.49	-
		Oriya	25.66	22.82	36.76	-
		Telugu	21.56	17.02	45.62	-
		Urdu	33.17	31.78	38.25-	-
	HMM	Bengali	33.50	32.83	39.77	-
		Hindi	48.30	47.16	46.84	-
		Oriya	28.24	25.86	45.84	-
		Telugu	13.33	32.37	46.58	-
		Urdu	34.48	36.83	44.73-	-
[36]	N-gram	Telugu	-	-	-	49.62
		Hindi	-	-	-	45.07

References

[1] B. D. M, M. Scott, S. Richard, and W. Ralph, "A High Performance Learning Name-finder," in *Proceedings of the fifth Conference on Applied Natural language Processing*, 1997, pp. 194–201.

[2] J. Lafferty, A. McCallum, and F. Pereira, "Probabilistic Models forSegmenting and Labelling Sequence Data," in *Proceedings of the Eighteenth International Conference on Machine Learning(ICML-2001)*, 2001.

[3] Cortes and Vapnik, "Support Vector Network ,MachineLearning," 1995, pp. 273–297.

[4] B. Andrew, "A Maximum Entropy Approach to NER," Ph.D. dissertation, 1999.

[5] F.Bechet, A.Nasr, and F.Genet, "Tagging Unknown Proper Names using Decision Trees," in *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistic*, 2000.

[6] R.Sirhari, C.Nui, and W.Li, "A Hybrid Approach for Named Entity and Sub-Type Tagging," in *Proceedings of the sixth conference on Applied natural language processing, Acm Pp*, 2000, pp. 247–254.

[7] J. Darroch and D.Ratcliff, "Generalized iterative scaling for log-linear models," *The Annals of Mathematical Statistics*, vol. 43(5), 1972.

[8] S. K. Saha, S. Sarkar, and P. Mitra, "A Hybrid Feature Set based Maximum Entropy Hindi Named Entity Recognition," in *Proceedings of the 3rd International Joint Conference on NLP*, Hyderabad,India, January 2008, pp. 343–349.

[9] A. Goyal, "Named Entity Recognition for South Asian Languages," in *Proceedings of the IJCNLP-08 Workshop on NER for South and South-East Asian Languages*, Hyderabad, India, Jan 2008, pp. 89–96.

[10] S. K. Saha, P. S. Ghosh, S. Sarkar, and P. Mitra, "Named Entity Recognition in Hindi using Maximum Entropy and Transliteration," *Research journal on Computer Science and Computer Engineering with Applications*, pp. 33–41, 2008.

[11] W. Li and A. McCallum, "Rapid Development of Hindi Named Entity Recognition using Conditional Random Fields and Feature Induction (Short Paper)," *ACM Transactions on Computational Logic*, pp. 290–294, Sept 2003.

[12] P. K. Gupta and S. Arora, "An Approach for Named Entity Recognition System for Hindi: An Experimental Study," in *Proceedings of ASCNT-2009*, CDAC, Noida, India, pp. 103–108.

[13] A. Ekbal and S. Bandyopadhyay, "Bengali Named Entity Recognition using Classifier Combination," in *Proceedings of 2009 Seventh International Conference on Advances in Pattern Recognition*, pp. 259–262.
[14] A. Ekbal, R. Haque, and S. Bandyopadhyay, "Named Entity Recogntion in Bengali: A Conditional Random Field," in *Proceedings of ICON*, India, pp. 123–128.
[15] A. Ekbal and S. Bandyopadhyay, "Named Entity Recognition using Support Vector Machine: A Language Independent Approach," *International Journal of Computer, Systems Sciences and Engg(IJCSSE)*, vol. 4, pp. 155–170, 2008.
[16] K. S. Hasan, M. ur Rahman, and V. Ng, "Learning -Based Named Entity Recognition for Morphologically-Rich Resource-Scare Languages," in *Proceedings of the 12th Conference of the European Chapter of the ACL*, Athens, Greece, 2009, pp. 354–362.
[17] B. B. Chaudhuri and S. Bhattacharya, "An Experiment on Automatic Detection of Named Entities in Bangla," in *Proceedings of the IJCNLP-08 Workshop on NER for South and South East Asian laanguages*, Hyderabad, India, January 2008, pp. 75–82.
[18] A. Ekbal and S. Bandyopadhyay, "Bengali Named Entity Rcognition using Support Vector Machine," in *Proceedings of the IJCNLP-08 Workshop on NER for South and South East Asian laanguages*, Hyderabad, India, January 2008, pp. 51–58.
[19] ——, "Development of Bengali Named Entity Tagged Corpus and its Use in NER System," in *Proceedings of the 6th Workshop on Asian Language Resources*, 2008.
[20] ——, "A web-based Bengali news corpus for named entity recognition," *Language Resources & Evaluation*, vol. 42, pp. 173–182, 2008.
[21] ——, "Voted NER System using Appropriate Unlabelled Data," in *Proceedings of the 2009 Named Entities Workshop, ACL-IJCNLP 2009*, Suntec, Singapore, August 2009, pp. 202–210.
[22] ——, "Improving the Performance of a NER System by Post-processing andVoting," in *Proceedings of 2008 Joint IAPR International Workshop on Structural Syntactic and Statistical Pattern Recognition*, Orlando, Florida, 2008, pp. 831–841.
[23] M. Hasanuzzaman, A. Ekbal, and S. Bandyopadhyay, "Maximum Entropy Approach for Named Entity Recognition in Bengali and Hindi," *International Journal of Recent Trends in Engineering*, vol. 1, May 2009.
[24] A. Ekbal and S. Bandyopadhyay, "AHidden Markov Model Based Named Entity Recognition System:Bengali and Hindi as Case Studies," in *Proceedings of 2nd International conference in Pattern Recognition and Machine Intelligence*, Kolkata, India, 2007, pp. 545–552.
[25] P.Srikanth and K. N. Murthy, "Named Entity Recognition for Telegu," in *Proceedings of the IJCNLP-08 Wokshop on NER for South and South East Asian languages*, Hyderabad, India, Jan 2008, pp. 41–50.
[26] P. M. Shishtla, K. Gali, P. Pingali, and V. Varma, "Experiments in Telegu NER: A Conditional Random Field Approach," in *Proceedings of the IJCNLP-08 Workshop on NER for South and South East Asian laanguages*, Hyderabad, India, January 2008, pp. 105–110.
[27] G. Raju, B.Srinivasu, D. S. V. Raju, and K. Kumar, "Named Entity Recognition for Telegu using Maximum Entropy Model," *Journal of Theoretical and Applied Information Technology*, vol. 3, pp. 125–130, 2010.
[28] V. R and S. L, "Domain focussed Named Entity Recognizer for Tamil using Conditional Random Fields," in *Proceedings of the IJCNLP-08 Wokshop on NER for South and South East Asian languages*, Hyderabad, India, 2008, pp. 59–66.
[29] S. Pandian, K. A. Pavithra, and T. Geetha, "Hybrid Three-stage Named Entity Recognizer for Tamil," *INFOS2008*, March 2008.
[30] S.Biswas, S.P.Mohanty, S.Acharya, and S.Mohanty, "A Hybrid Oriya Named Entity Recogntion system," in *Proceedings of the CoNLL*, Edmonton, Canada, 2003.
[31] A. Ekbal, R. Haque, A. Das, V. Poka, and S. Bandyopadhyay, "Language Independent Named Entity Recognition in Indian Languages," in *Proceedings of the IJCNLP-08 Wokshop on NER for South and South East Asian languages*, Hyderabad, India, 2008, pp. 33–40.
[32] S. K. Saha, S. Chatterji, and S. Dandapat, "A Hybrid Approach for Named Entity Recognition in Indian Languages," in *Proceedings of the IJCNLP-08 Workshop on NER for South and South East Asian laanguages*, Hyderabad, India, January 2008, pp. 17–24.
[33] K. Gali, H. Surana, A. Vaidya, P. Shishtla, and D. M. Sharma, "Aggregating Machine Learning and Rule Based Heuristic for Named Entity Recognition," in *Proceedings of the IJCNLP-08 Workshop on NER for South and South East Asian laanguages*, Hyderabad, India, January 2008, pp. 25–32.
[34] A. K. Singh, "Named Entity Rcognition for South and South East Asian Languages," in *Proceedings of the IJCNLP-08 Workshop on NER for South and South East Asian laanguages*, Hyderabad, India, January 2008, pp. 5–16.
[35] P. K. P and R. K. V, "A Hybrid Named Entity Rcognition System for South Asian Languages," in *Proceedings of the IJCNLP-08 Workshop on NER for South and South East Asian laanguages*, Hyderabad, India, January 2008, pp. 83–88.
[36] P. M. Shishtla, P. Pingali, and vasudeva Varma, "A Character n-gram Approach for Improved Recall in Indian Language NER," in *Proceedings of the IJCNLP-08 Workshop on NER for South and South East Asian laanguages*, Hyderabad, India, January 2008, pp. 67–74.