Named Entity Recognition for the Agricultural Domain

Malarkodi C. S.[1], Elisabeth Lex[2], Sobha Lalitha Devi[1]

[1]AU-KBC Research Center, MIT Campus of Anna University, Chromepet, Chennai, India
[2]Knowledge Technologies Institute, Graz University of Technology, Austria
{csmalarkodi,sobha}@au-kbc.org, elex@know-center.at

Abstract. Agricultural data have a major role in the planning and success of rural development activities. Agriculturalists, planners, policy makers, government officials, farmers and researchers require relevant information to trigger decision making processes. This paper presents our approach towards extracting named entities from real-world agricultural data from different areas of agriculture using Conditional Random Fields (CRFs). Specifically, we have created a Named Entity tagset consisting of 19 fine grained tags. To the best of our knowledge, there is no specific tag set and annotated corpus available for the agricultural domain. We have performed several experiments using different combination of features and obtained encouraging results. Most of the issues observed in an error analysis have been addressed by post-processing heuristic rules, which resulted in a significant improvement of our system's accuracy.

Keywords: Named Entity Recognition, Text mining, Information Extraction, Natural Language Processing, Agricultural data

1 Introduction

Named Entity Recognition (NER) has a fundamental role in Information Extraction (IE) and text mining applications like Question/Answer (Q&A) systems, event/product monitoring or customer-product relation extraction systems. For instance, in Q&A systems, named entities are answer strings to 'WH' questions. Text mining attempts to find the knowledgeable information from unstructured data lying on the web [24]. IE is considered as an important component of Text mining, since it aims to represent the structured information from the unstructured web data. The structured information extracted from IE systems, are mainly named entities. Since the identification of named entities discovers the hidden knowledge from data, it is considered as the main component in various activities related to text mining. The main goal of this work is to develop NER system for agriculture domain which supports the Information Systems and Text mining activities concerning agriculture sector. Agriculture is a main source of livelihood. Information about agricultural sciences are essential to improve agricultural productivity, research & development efforts. Food and agriculture organization emphasize that agricultural information is the cru-

cial factor in rural development activities [21]. For example, stake-holders within the agricultural domain, such as livestock & food processing industries, pesticide industries, researchers, or policy-makers need to retrieve entities such as crop names, fertilizers and price factors and how they are reflected for sound decision making regarding agricultural products. Though the extensive amount of work has been done on NER across various domains, very few have been found in the agricultural domain. NER from agricultural data enlightens the following information; 1) crop production and its location 2) policies or schemes benefiting farmers 3) natural disasters affecting crop cultivation 4) pesticides to control the pests 5) diseases affecting plant growth. The present work focused on the developmet of NER system for agriculture domain. We have developed the fine-grained NE tagset with 19 tags which covers the main key-terms in the respective domain. The paper is organized as follows: section 1 presents the overview of the state-of-art systems about NER, Section 2 describes the tagset design, corpus collection and NE annotation, Section 3 explains the various features used for system development, Experiments, results and error analysis are explained in Section 4. Finally the paper concludes with section 5.

1.1 Related Work

NER aims at the identification and classification of proper nouns into predefined categories like person, location, organization, etc.. Initially, NER was defined in MUC 6 as part of Information Extraction (IE). A survey of fifteen years of research on NER has been conducted by Nadeau [17]. They presented various features and methods used for NE identification from 1996 to 2006. A high performance named entity system for English and Spanish had been built using the standard version of a Hidden Markov Model (HMM) and obtained 90% accuracy [3]. Another NER system was developed to extract person names from e-mail[16]. Over a couple of decades, lot of research has been carried out on NER across languages. For example, Rössler et al. developed a NE engine for German based on Support Vector Machines (SVM) [19]. A NER system has been built for Dutch based on a genetic algorithm approach[5], an another for English using a combination of four classifier [8], while Federico et al. proposed a NER for Italian that is based on a boot-strapping process [7].

A more general overview on approaches used for NER across laguages is given in the survey of Kaur et al. [9]. They listed NER methods that follow rule based approaches, machine learning techniques such as decision trees, Naïve Bayes, Hidden Markov Model (HMM), Maximum Entropy Model (MEMM), Conditional Random Fields (CRF) and hybrid approaches, i.e. the combination of both rule based and machine learning approaches.

Vijayakrishna et al. worked on Tamil NER for the Tourism domain using CRF. Their system handles nested tagging of named entities with a hierarchical tag set containing 106 tags [22]. Malarkodi et al. proposed a NER model with language dependent and independent features for English, Tamil, Telugu, Bengali, Punjabi and Marathi using CRF [15]. Ekbal et al. developed NER systems for the two leading Indian

languages, namely Bengali and Hindi using CRF [6]. Bindu et al. developed the Malayalam NER engine and discussed how their work support Q&A systems [4].

The work presented in [18] is also very much related to ours as they also perform NER in the domain of agriculture. In their work, they constructed auto-matic NE gazetteers using unsupervised learning, more specifically, a variant of Multiword Expression Distance. They have used three NE tags, namely crop, disease and chemical_treatment. For each NE type, the gazetteers were generated automatically and the effectiveness of the dictionary was compared with Wikipedia articles related to agriculture.

1.2 Our Contributions

The contributions presented in this work are threefold: 1) In order to extract a wide range of information from the agriculture domain, we have designed a fine-grained tag-set comprising of 19 entity types. 2) The NE annotated corpus has been created for agriculture with 1L word forms 3) We have constructed the baseline system with basic minimal features such as word, POS and chunking information 4) Based on the analysis, we also incorporated rich set of linguistic features for the system development 5) Post processing heuristics is applied further to fine-tune the system.

2 Tagset Design and Corpus Preparation

2.1 Our NER Tagset

We have paid special attention towards developing fine-grained tagsets for our work. The proposed tagset consisting of 19 tags for named entities can help to understand the semantic classes of entities for the agricultural domain. Table 1 explains the rationale behind the tags along examples.

Table 1. NE Tagset With Example

NE Tags	Description	Example
Person	Names of person and nationals	John Smith, Indian
Location	Names of cities, continents, water bodies	Europe, Chennai
Organization	Names of institutions, companies, industries	Common organization Of agricultural market
Chemicals	Fertilizers, pesticides, fungicides	Nitrogen, nitrate
Crop	Names of fruits, vegetables, cereals, grains	Apple, carrot, wheat
Organism	Names of animals and micro-organisms	Sheep, Escherichia coli,
Policy	Agricultural aids or policies	Common agricultural Policy
Climate	Denotes the climatic conditions	Summer, winter

Food items	Plant/animal products	Cheese, milk, bread
Diseases	Diseases affecting plant growth	Late Blight, brown rot
Natural Disaster	Disasters affecting crop production.	Famine, earthquake, Flood
Events	Conferences, workshops, meetings and exhibition	National conference on agricultural & Food Security
Nutrients	Fats, minerals, vitamins and carbohydrates	Vitamin A, calcium
Count	Number of items	350 people
Distance	Distance measures such as feet, meter, km., etc.	15 inches, 250 acres
Quantity	Quantity measures such as litter, tonnes, grams	8 tonnes, 10 kg
Money	Currency value such as the euro, rupee, dollar etc.	$90, 100 euro, Rs. 1000
Temperature	Numerical measure of climatic condition such as degree, Celsius	70C
Year_Month_Date	Denotes the year, month, day and date	May 26, 29/10/2013

2.2 Corpus Collection and Annotation

Corpus Collection. As the system's performance depends on training data for the CRF, the corpus collection is a crucial factor. The dataset used in the proposed work has been collected from Wikipedia articles that are related to agriculture as well as from reputed websites in the European Union pertained to agriculture. The corpus has been collected in such a manner that it would cover major aspects of farming from crop cultivation to agricultural productivity. Therefore for data collection we focused on the sub-domains like crop cultivation and management, food processing industries and research institutions, subsidiaries, organisms and diseases, food products, natural disasters and risk management, agribusiness and marketing.

NE Annotation. The agricultural corpus comprises of 100k word forms. The raw text is tokenized and pre-processed with part of speech and chunking information. Named entities are manually annotated and represented in BIO standard mentioned in CONLL 2003 shared task. Issues encountered during the annotation process are as follows: 1) ambiguity between NEs 2) ambiguity between NE and non-NE 3) boundary limitation 4) abbreviations. Firstly, there exists an ambiguity between named entities of one type and another; for example, Jersey is a location name in some instances and Jersey is a cattle name in some cases. Purcari is a wine sector name in some examples and the same instance denotes the wine name in other examples. Secondly, ambiguity aroused in the case of whether to consider certain instances as NE or not. For example, “common organization of agricultural markets” is an organization name but ambiguity arisen whether to consider words like fast-food market, agricultural market, as NE or non-NE. As those instances are important in this domain, we con-

sidered them as NE. Thirdly, due to the definite descriptions it is difficult to determine the NE boundaries. Fourthly, abbreviations of organization, policy and event names are hard to identify without contextual knowledge. The corpus statistics are shown in the table 2. The corpus is randomly divided into 80:20 for training and test data.

Table 2. Corpus Statistics

Corpus	Total No of Words	Total No of NEs
Train Corpus	80k tokens	8,924
Test Corpus	20k tokens	2,117

3 Our Approach

CRF is a probabilistic model used to segment and label the sequential data [12][23]. It selects the label sequence y which maximizes the conditional probability of p(y|x) to the observation sequence x. The probability of a label sequence y given an observation sequence x is given below:

$$P(y|x,\lambda) = \frac{1}{z(x)} \exp \sum_{i=1}^{n} \sum_{j} \lambda_j f_j(y_{i-1}, y_i, x, i)$$

$$z(x) = \sum_{i=1}^{n} \sum_{j} \lambda_j f_j(y_{i-1}, y_i, x, i)$$

Where z is normalization factor, fj (Y_{i-1}, Y_i, x,i) is a transition feature function of an observation sequence and the labels at position i and i − 1. For example, consider the task of assigning the label y to the word x named 'Netherland', then the transition function fj(y_{i-1},y_i,x,i) = 1 if y_i="LOCATION" and the suffix of i_{th} word is "land"; otherwise 0; If the weight λ_j associated with the above feature is large and positive, then the words ending with the suffix "land" is labelled as NE type "LOCATION".

We have developed the base NER engine, with minimal basic features. Later, we added heuristic rules to improve the system's performance. Our method is a hybrid approach, as we are using both machine learning technique and linguistic rules. We came up with the linguistically inspired features which are explained below.

3.1 Syntactic Level Features

POS and Chunk Information . POS play a significant role in NE identification, as they contain information about the linguistic category of words. We have considered the POS tags occurring in a window of five. Noun phrase chunking helps identify the NE boundary.

Proper Noun. As most of the Named Entities belong to the grammatical category proper noun, we gave importance to the POS tag of the proper nouns.

Frequent POS patterns of NE. The most frequent POS tags preceding the NE occurring in w-1(word preceding the NE) and w-2 positions are considered as a feature.

3.2 Lexico-Syntactic Features

Cue Phrases and POS. Cue phrases are the key words occurring as part of an entity. For instance, organization names are following or preceding by cue phrases like university, consortium, ltd. The key-terms like blight, rots are occurring as part of plant disease.

Occurrence of proper noun after preposition 'in'. We have examined the proper nouns following the preposition 'in'. Our analysis revealed that in most of the cases proper noun (NNP) that comes immediately after the preposition 'in' are location names.

Numerical Feature. Digit patterns exhibit useful information in predicting numerical entities. For instance, numerical values of length four tend to identify year names. POS of the numerical value CD (Cardinal Number) preceded and succeeded by quantity measures and distance measure are considered as a feature.

4 Experiments and Results

The agriculture corpus is randomly divided in the ratio of 80:20 for training and test data. We measured the performance of our system in terms of precision, recall and f-measure. In order to find the best feature set, we have conducted several experiments using various combinations of features. We also have performed 10 fold cross validation. Feature-wise performances are shown in tTable 3.

Table 3. Feature-wise Results

S.NO	Feature Combination	Precision	Recall	F-M
1	Word, POS and chunk information	80.96	71.75	76.35
2	(1), capitalization	81.59	72.63	77.11
3	(1), noun phrase	81.69	72.63	77.16
4	(1),(2),(3), Occurrence of NP after 'IN'	82.67	74.15	78.41
5	(1),(2),(3),(4), key words	82.83	76.01	79.42
6	(1),(2),(3),(4),(5), affixes	83.50	77.58	80.54
7	(1),(2),(3),(4),(5),(6), Numeric features	83.61	78.81	81.21
8	(1),(2),(3),(4),(5),(6),(7),POS patterns of NE	84.45	79.67	82.06

4.1 Contribution of Various Features

Initially, we have applied the combination of word, POS and chunking information to determine the performance of our system when using the minimal feature set. We obtained an accuracy of F-M: 76.35% for the base NE system. In comparison with

results shown in a row (1), including capitalization as feature improves the accuracy by 1% and including proper nouns leads to the increase of 1% in precision and 1% in recall value. The occurrence of proper noun after the preposition "in" enables a positive effect on f-measure by 1%. Especially this feature increased the performace of NE type "Location", since most of the proper nouns followed by the preposition "in" are location names. The key word feature provides the improvement of 2% recall while the affix feature provides a slight improvement on f-measure. Inclusion of keyword feature contributes to raise an accuracy of NE types such as "organization, policy, diseases". The affix feature leads to the improvement of chemical and location entities. Finally, the incorporation of numerical features boost the identification of numerical entities such as "Unit, Money and Year_Month_Date"and POS patterns succeeding & preceding NE seems to be effective for all NE types. The highest accuracy obtained from the CRF approach is 82.06% (F-Measure). Tag-wise, F-Measure (F-M) values for entity expressions are depicted in Figure 1.

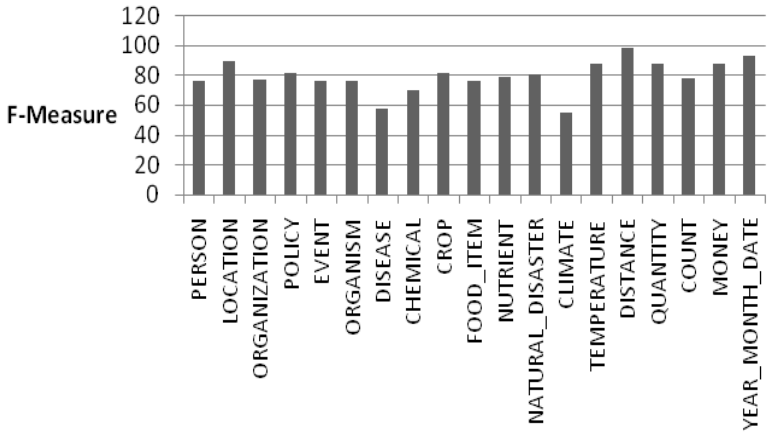

Fig. 1. Tag-wise Results from CRF

4.2 Cross Validation

Cross validation is used to assess how our results can be generalized to an independent dataset. Specifically, we performed 10 fold-cross validation and report the average accuracy is shown in table 4. During the analysis on classification results, we have observed that 70 percent of Organization names in our corpus consist of more than six token length and 80% of such instances are correctly identified by the system. In entity expressions, location tag obtained the highest score of 90.83% precision and 88.19% recall. Figure 1. shows that other than the entities DISEASE and CLIMATE, all entity expressions scored above 75% accuracy. Except NE type COUNT, we achieved more than 80% for all numeric entities. Crop name was the second highest one.

Table 4. 10-Fold Cross-Validation Results

S. No.	Precision	Recall	F-M
1	83.06	77.48	80.27
2	84.20	78.25	81.22
3	84.60	79.13	81.86
4	84.73	81.35	83.04
5	82.05	79.06	80.55
6	85.55	82.22	83.88
7	82.30	78.66	80.48
8	86.95	82.05	84.50
9	81.81	77.21	79.51
10	87.32	80.84	84.08
Avg	84.25	79.62	82.23

4.3 Error Analysis

From our observation, we found that the errors in our system are due to inconsistencies and ambiguities between entities. Abbreviation of policy names is misclassified as organization in some cases. Such instances can be handled using the NE type of the respective expansion. In some cases, abbreviations will occur right after the expansion. In the example of the Common Agricultural Policy (CAP), if the system has tagged the expansion correctly and fails to identify the abbreviation 'CAP', we can tag the NE type for 'CAP' as 'POLICY' by using the NE tag of the expansion which precedes the abbreviation. Parts of organization names may be misclassified as Location and food_item as crop type in a few instances. This problem arises when one type of NE occurs as part of another type of NE (entity within entity, i.e. a so-called nested entity). It can be handled using the cue phrases of the respective types with the combination of POS and orthographic features.

4.4 Impact of Post-processing

In order to improve the accuracy of our system and to remove inconsistencies, we applied linguistic and heuristic rules following the CRF output. Some of the rules implemented in the post-processing are discussed below.

The rule 1 given in table 5 describes that if the B-tag (Beginning Tag) of same NE type occurs in two consecutive positions within the same phrase, the second B-tag should be replaced as I-tag (Inside Tag). The rule2 illustrates that if the I-tag start without B-tag, then the I-tag in the beginning of an entity should be changed to B-tag. In row3, instead of tagging “Sugar Cane Juice“ as food_item, the system has tagged part of the NE “Sugar Cane“ as “CROP“. This ambiguity has occurred due to being a nested entity. As we considered the maximal entity, we handled this ambiguity using linguistics rules based on the POS, orthographic features and key words. Thus, we have handled partial tagging and ambiguities that exists between NEs. Tag wise results obtained after post-processing is given in Table 6.

Table 5. Post Processing Results

Post Processing Rule	Example	Comments
-2 NNP+B-NP B-TAG1	-2 John B-PER	-2 John B-PER
-1 NNP+I-NP B-TAG1	-1 Mathew B-PER	-1 Mathew I-PER
0 NNP+I-NP I-TAG1	0 Zen I-PER	0 Zen I-PER
-2 NNP I-TAG1	2 Rural I-POLICY	-2 Rural B-POLICY
-1 NNP I-TAG1	-1 Agri I-POLICY	-1 Agri I-POLICY
0 NNP I-TAG1	0 Policy I-POLICY	0 Policy I-POLICY
0 NNP B-TAG1	0 Sugar B-CROP	0 Sugar B-FOOD_ITEM
1 NNP I-TAG1	1 Cane I-CROP	1 Cane I-FOOD_ITEM
2 Keyword+NN|NNP	2 Juice O	2 Juice I-FOOD_ITEM

Table 6. Tag-wise results after Post-Processing

NE Type	Precision	Recall	F-measure
PER_LOC_ORG	82.59	83.59	83.09
CROP	82.31	76.7	79.50
ORGANISM	84.28	74.68	79.48
DISEASE	62.5	98.5	80.50
CHEMICAL	86.27	72.13	79.20
FOOD_ITEM	80.32	78.01	79.16
POLICY	82.17	81.1	81.63
OTHERS	85.93	83.26	84.59
\NUMERICAL	91.50	90.43	90.96
TIME	92.54	92.95	92.74
AVERAGE	83.24	83.13	83.18

Where PER_LOC_ORG- Person, Location and Organization, OTHERS include the NE type Natural Disaster, Climate, Nutrient and Event. As the post processing rules reduced the number of false positives and increased the number of true positives, there was a slight decrement in precision and 4% increment in the recall. Figures 2 and 3 show the results before and after post-processing.

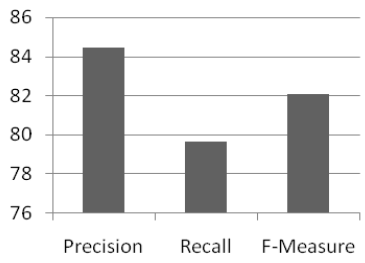

Fig. 2. Results before Post-Processing

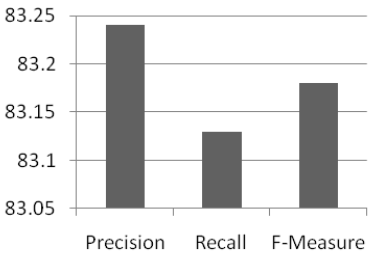

Fig. 3. Results after Post-Processing

4.5 Comparison with CreateGazetteMED

Patil et al. worked on the agriculture domain with 3 NE tags[18]. The highest precision obtained by the CreateGazetteMED algorithm is 66.2% for crop, 92.8% for disease and 88.6% for chemical. We have achieved a precision of 82.31% for crop, 62.5% for disease and 86.27% for chemical. The precision we scored for the NE category crop is higher and chemical is quite closer to CreateGazetteMED results. There was a precision drop for disease tag, due to the false positives.

5 Conclusion

In this paper, we have presented a NER system for the agriculture domain. To the best of our knowledge, this work is the first attempt in generating a NE annotated corpus and NE system with a major tagset for agriculture. We have collected data from various sub-domains of agriculture starting from cultivation to marketing and we designed the NE tag-set with 19 fine grained tags so that it could cover prominent entities in the agricultural field. Our system exploits both linguistically enriched as well as domain independent features. With our system, we achieved 76% accuracy with minimal features such as word, POS and chunk information. Based on a detailed corpus analysis, more features are incorporated and our results were improved by 7%. We have implemented post-processing heuristics to overcome tag ambiguities which resulted in an improvement of the overall precision of our system to 83.24% and a recall of 83.13%. Our results show that our system is comparable with existing NER models. In future we plan to extend this work by developing a robust relation extraction system for the agricultural domain.

Acknowledgement

This work has been funded by the European Commission as part of the WIQ-EI project (project no. 269180) within the FP7 People Programme. It is the result of the collaboration between AU-KBC Research Centre, Chennai, India and Know Center, Graz, Austria. The Know-Center is funded within the Austrian COMET program – Competence Centers for Excellent Technologies - under the auspices of the Austrian Federal Ministry of Transport, Innovation and Technology, the Austrian Federal Ministry of Economy, Family and Youth, and the State of Styria. COMET is managed by the Austrian Research Promotion Agency FFG.

References

1. Palanisamy, A., Lalitha Devi, S.: HMM based POS Tagger for a Relatively Free Word Order Language. Research in Computing Science, vol. 18, pp. 37-48. (2006)

2. Benajiba, Y., Rosso, P.: Arabic named entity recognition using conditional random fields. In: Proceedings of the workshop on HLT & NLP within the Arabic World, LREC, vol. 8, pp. 143-153. (2008)
3. Bikel, D,M., Miller, S., Schwartz, R., Weischedel R.: Nymble:A high-performance learning name-finder. In: Proceedings of Fifth Conference on Applied Natural Language Processing, pp. 194-201. (1997)
4. Bindu, M.S., Sumam, M.I.: Article: Named Entity Recognizer employing Multiclass Support Vector Machines for the Development of Question Answering System. International Journal of Computer Applications, vol. 25(10), pp. 40-46. (2011)
5. Desmet, B., Véronique, H.: Dutch named entity recognition using ensemble classifiers. In: Proceedings of 20th Meeting of Computational Linguistics in the Netherlands (CLIN), Landelijke Onderzoeksschool Taalwetenschap (LOT), (2010)
6. Ekbal, A., Bandyopadhyay, S.: A Conditional Random Field Approach for Named Entity Recognition in Bengali and Hindi. Linguistic Issues in Language Technology, vol. 2(1), pp. 1-44, (2009)
7. Federico, M., Bertoldi, N., Sandrini V.: Bootstrapping named entity recognition for Italian broadcast news. In: Proceedings of ACL-02 conference on Empirical methods in natural language processing, vol.10, pp. 96-303. (2002)
8. Florian, R., Ittycheriah, A., Jing, H., Zhang, T.: Named entity recognition through classifier combination. In: Proceedings of seventh conference on Natural language learning at HLT-NAACL 2003 Association for Computational Linguistics, pp. 168-171. (2003)
9. Kaur, D., Gupta, V.: A survey of named entity recognition in English and other Indian languages. In: Proceedings of the IJCSI, pp. 239-245. (2010)
10. Kudo, T.: CRF++, an open source toolkit for CRF. http://crfpp.sourceforge.net, 2005.
11. Kumar, S., Jha, G.N., Lalitha Devi, S.: Challenges in Developing Named Entity Recognition System for Sanskrit, In: Proceedings of Workshop on Indian Language and Data: Resources and Evaluation Workshop Programme, pp. 70-75. (2012)
12. Lafferty, J., McCallum, A., Pereira, F.: Conditional Random Fields for segmenting and labelling sequence data. In: Proceedings of ICML-01, pp. 282-289. (2001)
13. Lalitha Devi, S., Vijay, S.R.: Noun Phrase Chunker for Tamil. In: Proceedings of Symposium on Modeling and Shallow Parsing of Indian Languages, Indian Institute of Technology, Mumbai, pp. 194-198. (2006)
14. Malarkodi, C.S., Pattabhi, R.K., Lalitha Devi, S.: Tamil NER–Coping with Real Time Challenges, In: Proceedings of Workshop on Machine Translation and Parsing in Indian Languages(MTPIL-2012), COLING, pp. 23-38 (2012)
15. Malarkodi, C.S., Lalitha Devi, S.: A Deeper Look into Features for NE Resolution in Indian Languages. In: Proceedings of workshop on Indian Language Data: Resources and Evaluation, LREC, Istanbul, (2012)
16. Minkov, E., Richard, C.W., William, W.C.: Extracting personal names from email: applying named entity recognition to informal text. In: Proceedings of conference on Human Language Technology and Empirical Methods in Natural Language Processing, Association for Computational Linguistics, pp. 443-450. (2005)
17. Nadeau, D., Sekine S.: A survey of named entity recognition and classification, Linguisticae Investigationes, vol. 30(1), pp.3-26. (2007)
18. Patil, A., Sachin, P., Girish K.P.: Named Entity Extraction Using Information Distance. In: Proceedings of Sixth International Joint Conference on Natural Language Processing, pp. 1264-1270. (2013)

19. Rössler, M.: Corpus-based Learning of Lexical Resources for German Named Entity Recognition, In: Proceedings of Fourth International Conference on Language Resources and Evaluation (LREC'04), (2004)
20. Saha, S.K., Chatterji, S., Dandapat, S., Sarkar, S. Mitra P.: A hybrid approach for named entity recognition in Indian languages. In: Proceedings of IJCNLP-08 Workshop on NER for South and South East Asian Languages, pp. 17-24 (2008)
21. Vidanapathirana, N.P.: Agricultural information systems and their applications for development of agriculture and rural community a review study. In: Proceedings of 35th Information Systems Research Seminar in Scandinavia, (2012)
22. Vijayakrishna, R. Lalitha Devi, S.: Domain focused Named Entity for Tamil using Conditional Random Fields, In: proceedings of IJNLP-08 workshop on NER for South and South East Asian Languages, Hyderabad, India, pp. 59-66. (2008)
23. Wallach, H.M.: Conditional random fields: An introduction. Technical Reports (CIS), MSCIS-04-21, (2004)
24. Witten, I. H.: Text mining, Practical handbook of Internet computing. vol. 4(1), pp. 56-59. (2005)