Research Journal of Applied Sciences, Engineering and Technology 8(10): 1272-1278, 2014
DOI:10.19026/rjaset.8.1095
ISSN: 2040-7459; e-ISSN: 2040-7467
© 2014 Maxwell Scientific Publication Corp.
Submitted: July 14, 2014 Accepted: September 13, 2014 Published: September 15, 2014

Research Article

Challenges of Urdu Named Entity Recognition: A Scarce Resourced Language

[1,3]Saeeda Naz, [1]Arif Iqbal Umar, [1]Syed Hamad Shirazi, [2]Sajjad Ahmad Khan,
[2]Imtiaz Ahmed and [2]Akbar Ali Khan
[1]Department of Information Technology, Hazara University, Mansehra, Pakistan
[2]COMSATS Institute of Information Technology, Abbottabad, KPK, Pakistan
[3]Higher Education Department, GGPGC NO.1 Abbottabad, KPK, Pakistan

Abstract: In this study, we present a brief overview of Named Entity Recognition (NER) system, various approaches followed for NER systems and finally NER systems for Urdu language. Urdu language raises several challenges to Natural Language Processing (NLP) largely due to its rich morphology. Research against NER systems in Urdu language is at infancy stage therefore the focus of this study is on challenges and peculiarities of Urdu NER system. In this study we also explore the previous work done on NER systems for South and South East Asian Languages (SSEAL). Finally, we conclude the existing work in Urdu NER which is a scarce resourced and morphologically rich language and other SSEAL which have similar features to Urdu language.

Keywords: CRF, ME, SSEAL, Urdu named entity recognition

INTRODUCTION

Named Entity Recognition (NER) is a process of searching the text to detect entities ('atomic elements') in a text and to classify them into predefined classes such as the names of persons, organizations, locations, expressions of times, quantities, etc. For example consider the following sentence:

> "Microsoft launched its first retail version of Microsoft Windows on November 20, 1985"

An accurate NER system would extract two NEs from the above sentence:

- "Microsoft" as an organization
- "November 20, 1985" as a date

NER is a basic tool for all application areas of Natural Language Processing (NLP) such as Automatic Summarization, Machine Translation, Information Extraction, Information Retrieval, Question Answering, Text Mining and Genetics etc. Performance of all these applications depends on NER system. These applications can perform well if the named entities are recognized and grouped accurately.

The "Named Entity" word was used and promoted in the sixth and seventh "Machine Understanding Conferences" (MUC). The Message Understanding Conference (MUC) was initiated in 1987 by DARPA (Defense Advanced Research Projects Agency) to foster the development of enhanced algorithms for information extraction. For the 6th MUC, one of the evaluation tasks was "Named Entity Recognition", which brought this study field into limelight. According to Ekbal and Bandyopadhyay (2008a) these conferences defined the milestones for English Named Entity Recognition (NER) systems. The concept of MUC-6 and MUC-7 was also adapted by "Multilingual Entity Task" (MET-2) for Japanese NER. The Conference on Computational Natural Language Learning (CONLL-2002) focused Dutch and Spanish languages and CONLL-2003 was organized for German. In CoNLL 2002 and CoNLL 2003 concerned on language-independent NER. According to Sang (2002) in CoNLL 2002 the participants evaluated their systems on Spanish and Dutch corpora and on English and German data in 2003. IOB2 annotation was used for tagging the data in both evaluations. IOB2 scheme is a variant of the IOB scheme introduced by Ramshaw and Marcus (1995). This tagging scheme rules are discussed below:

- Words which are Outside NEs are tagged as "O-TYPE"
- "B-TYPE" tag is used for the first word (Beginning) of an NE of class TYPE
- Words which are part of an NE of class TYPE but are not the first word are tagged as "I-TYPE" (Inside)

The MUC-6, CoNLL 2002 and 2003 competitions have proven to be a valuable resource for further work on NER systems. Variety of techniques has been discovered in the proceedings of these competitions.

Corresponding Author: Saeeda Naz, Department of Information Technology, Hazara University, Mansehra, Pakistan
This work is licensed under a Creative Commons Attribution 4.0 International License (URL: http://creativecommons.org/licenses/by/4.0/).

These competitions have given an idea about the most competent Machine Learning (ML) techniques for the NER task such as Hidden Markov Model (HMM), Maximum Entropy (ME), Support Vector Machine (SVM) and Conditional Random Fields (CRF). Furthermore most competent systems have used one of the following approaches:

- Two-phase method in which the first phase detects the boundaries of the NEs and the second phase classifies
- Combined different machine learning techniques in order to improve their results by means of the advantages of different modeling techniques

NER systems for English, European languages and some Asian languages (Chinese, Japanese etc.) have reached to their maturity level and have yielded result with very high accuracies. These languages become rich resourced languages. But development of NER system is a challenging and more complicated task in the South and South East Asian Languages (SSEALs) due to poor resources and some features such as lack of capitalization and spelling variations etc. Some work has been made on some Indian languages NER system recently while very little computational research work has been done in the area of NER for Urdu language. There is a need to go through the existing work on Urdu NER system and their comparison for getting attention of researchers that construction of accurate Urdu NER is very crucial and important on the Internet because Urdu has got very much political importance by reason of its close relation with Muslim world.

NER APPROACHES

In literature, three main approaches have been used for the development of NER systems.

Rule based approach is also called Handcrafted Approach. It is based on seeking named entities in the text by using linguistic or handcrafted rules manually written by linguists along with gazetteer lists. Saha *et al*. (2008a) have main disadvantages of rule-based techniques. According to them, huge experience and grammatical knowledge of the particular language or domain is required. The techniques developed for the rule based systems of a language cannot be applied for other languages or domains. The rule based NER systems makes use of gazetteer lists and dictionaries. Chaudhuri and Bhattacharya (2008) have discussed that the rule-based systems cannot tackle ambiguous situations very well.

Statistical approach is based on ML models like HMM, ME, CRF and SVM etc. These methods need a large sized Named Entity tagged corpus for training. The NE tagged corpus is used to train the statistical models so that they can acquire high level language knowledge. The training data in case of statistical model must be annotated with all of the concerned entities and their types. Furthermore the training data should match the data on which the system will be run. Gazetteer lists and dictionaries are also used to classify words for achieving better results in the statistical approach. Statistical approach is not domain specific therefore; it is easily applicable and trainable for other languages or domains. Maintenance of ML based NER systems is also very easy and cheaper than the rule based NER system.

Hybrid NER systems use ML approaches along with hand crafted-rules. Gazetteer lists are also used in hybrid systems. The hybrid systems are mostly used for morphologically rich languages because of their complex nature. These systems yield result with high accuracy but have the same problem of being non-portable to other languages or domains due to linguistic rules. Current trend in NER is to make use of machine learning or statistical approaches because of their adoptable and trainable nature, such systems are easy to maintain and are cheaper as compared to rule based systems. Srikanth and Murthy (2008) have discussed that machine learning techniques are relatively independent of language and domain and no expert knowledge is needed.

LITERATURE REVIEW

A number of different techniques have been used for the development of NER systems for different languages since 1991. A surfeit of algorithms has been developed for NER of English and other European languages and has achieved high recognition rates. Comparatively very few NER algorithms have been developed for South and South East Asian languages. The following sections discuss different earlier research carried out to develop NER systems.

Rule based approaches: Among the earlier research papers in the field of NER area, Rau and Jacobs (1991) has presented a rule based NER system for identification and classification of different company names. The accuracy of system is over 95%. Cucerzan and Yarowsky (1999) have developed a language independent NER system for Hindi language by using contextual and morphological evidences for five languages such as English, Greek, Romanian, Turkish and Hindi. The performance of Hindi NER system is very low and has f-measure of 41.70 with very low 27.84% recall and nearly 85% precision.

Statistical approaches: Borthwick (1999) has presented a NER system based on Maximum Entropy (ME) for English language and has achieved F-measure of 84.22%. Li and McCallum (2003) have presented a Conditional Random Field (CRF) for the development of NER system for Hindi language. The system has 71.50% accuracy. The authors provided large array of lexical test and used feature induction for constructing the features automatically. These both helped in discovering the relevant features. The early stopping

and Gaussian prior have been used for reducing over fitting.

Nadeau *et al*. (2006) have presented semi-supervised approach for the development of an English NER system by classifying 100 named entities. The System has achieved F-measure value in the range 78-87%. Saha *et al*. (2008b) have used Maximum Entropy based NER system for Hindi language. The system has achieved F-value of 80.01% by using word selection and word clustering based feature reduction techniques. Ekbal *et al*. (2008) have developed statistical Conditional Random Field (CRF) model for the development of NER system for South and South East Asian languages, particularly for Bengali, Hindi, Telugu, Oriya and Urdu. Different contextual information and variety of features have been used for seeking and recognizing 12 classes of Named Entities in the system. The language independent features for all the languages have been used except Bengali and Hindi languages. The rules for identifying nested NEs for all the five languages have been used. The gazetteer lists for Bengali and Hindi languages have also been used. The system has achieved F-measure of 59.39% for Bengali, 33.12% for Hindi, 28.71% for Oriya4.749% for Telugu and 35.52% for Urdu. Goyal (2008) has developed CRF based NER system for Hindi language. This machine learning algorithm has been trained using NLPAI Machine Learning Contest 2007 data. The comparison on Hindi data and English data of CoNLL shared task of 2003 has also been discussed. The proposed system has been divided into three sub tasks. The first module called NER module recognizes NE in the text, second module called NEC module classifies the recognized Named Entities according to their types and third module called NNE module identifies the Nested Named Entities (NNE). The tags used for this system are: person, organization, location names, measure, time, number, domain specific terms, abbreviation, title and designation. IOB model is used in NER module. The author divided the test data into two sets called test set 1 and 2. The method has been evaluated on test set 1 and 2 and achieved nested F1-measure around 50.1% and maximal F1-measure around 49.2% for test set 1 and nested F1-measure around 43.70% and maximal F1 measure around 44.97 for test set 2 and F1-measure of 58.85% on development set. Ekbal and Bandyopadhyay (2008b) and Rau and Jacobs (1991) have presented NER system based on Support Vector Machine (SVM) for Bengali language. Different contextual information of the words along with a variety of features has been used to predict different NE classes. The training set for experiment has partially NE tagged corpus collected from online Bengali newspapers. Results of various experiments has showed overall average recall value of 94.3%, precision value of 89.4% and F-measure value of 91.8% of the system. VijayKrishna and Sobha (2008) have developed CRF based Tamil NER system for tourism purposes. A hierarchical tag set consisting of 106 tags have been used to handle morphological inflection and nested Named Entities. A corpus of size of 94 k has been manually tagged for POS, NP chunking and NE annotations. The corpus has been divided into training data and the test data. The system has F-measure of 80.44%. Gali *et al*. (2008) have developed CRF based NER system for Telugu. The language dependent and independent features have been used for the experiments. The system has F-value of 44.91%. The authors have observed that the use of suffix and prefix information helps a lot in seeking the category. Gupta and Arora (2009) have presented a CRF based NER system for Hindi. The data collected from the tourism domain has been used as a training data for model and manually tagged in the IOB format. The maximum f-measure achieved by system is up to 66.7% for Person, 69.5% for Location and 58% for organization. Raju *et al*. (2010) have developed ME based NER system for Telugu. The data of corpus has been collected from the Telugu Wikipedia and newspapers. The system has been evaluated with the manually tagged test data, different contextual information of the words and Gazetteer list. Gazetteer list has been prepared manually or semi-automatically from the corpus. The System has achieved an F-measure of 72.07% for person, 6.76, 68.40 and 45.28% for organization, location and others respectively. Ekbal and Saha (2011) have developed a multi-objective simulated annealing based classifier ensemble NER system for three scarce resourced languages like Hindi, Bengali and Telugu. The recall, precision and F-measure values are 93.95, 95.15 and 94.55%, for Bengali, 93.35, 92.25 and 92.80%, for Hindi and 84.02, 96.56 and 89.85%, respectively for Telugu, respectively. Conditional Random Field (CRF), Maximum Entropy (ME) and Support Vector Machine (SVM) have been used to construct different models using language independent features. An ensemble system has been used to find appropriate weight of vote for each output class in each classifier.

Hybrid approaches: Bikel *et al*. (1997) have developed IdentiFinder using HMM for English and Spanish languages to extract proper names and to make four categories including names, times, dates and numerical quantities. The system has achieved F-measure of 90.44%. Biswas *et al*. (2010) have presented a hybrid system for Oriya NER based on ME, HMM and some handcrafted rules to recognize NEs. The IOB annotated data has been used. The system has an F-measure from 75 to 90%. Saha *et al*. (2008b) have presented NER system using Maximum Entropy approach for Hindi, Bengali, Telugu, Oriya and Urdu. Linguistic rules and gazetteer lists have also been used to achieve better performance of NER for Hindi and Bengali languages. The NER system has F-measures of 65.13, 65.96, 44.65, 18.74 and 35.47% for Hindi,

Table 1: Different approaches used for SSEA languages

Author	Languages	Approaches	F-measures (%)
Cucerzan and Yarowsky (1999)	Hindi, English, Greek, Romanian, Turkish	Language independent features	41.70
Li and McCallum (2003)	Hindi	CRF	71.50
Saha *et al*. (2008a)	Hindi	ME	80.01
Saha *et al*. (2008b)	Hindi, Bengali, Telugu, Oriya, Urdu	ME	65.13, 65.96, 44.65, 18.74, 35.47
Gali *et al*. (2008)	Hindi, Bengali, Telugu, Oriya, Urdu	CRF	40.63, 50.06, 39.04, 40.94, 43.46
Ekbal *et al*. (2008)	Bengali, Hindi, Telugu, Oriya, Urdu	CRF	59.39, 33.12, 47.49, 28.71, 35.52
Ekbal and Bandyopadhyay (2008b)	Bengali	SVM	91.80
Chaudhuri and Bhattacharya (2008)	Bangla	N-gram+dictionary+rules	89.51
VijayKrishna and Sobha (2008)	Tamil	CRF	80.44
Srikanth and Murthy (2008)	Telugu	Rules then CRF	80-97
Goyal (2008)	Hindi	CRF	58.85
Kumar and Kiran (2008)	Bengali, Hindi, Oriya, Telugu, Urdu	CRF, HHM, rules	38.25, 44.73
Gupta and Arora (2009)	Hindi	CRF	66.7 (for person), 69.5 (for location), 58 (for organization)
Raju *et al*. (2010)	Telugu	ME	48.12
Ekbal and Saha (2011)	Hindi, Bengali, Telugu	Ensemble	94.55, 92.80, 89.85
Srivastava *et al*. (2011)	Hindi	CRF, ME, rules, voting	46.43, 39.99, 91.25, 82.95

Bengali, Oriya, Telugu and Urdu respectively. Gali *et al*. (2008) have developed a CRF based NER system for five languages including Hindi, Bengali, Telugu, Oriya and Urdu. The machine learning approach and hand written rules or heuristics have been used. The NER system has been trained for Hindi and Telugu languages. The system has an accuracy of 40.63, 50.06, 39.04, 40.94 and 43.46 F-values for Bengali, Hindi, Oriya, Telugu and Urdu, respectively without sufficient linguistic resources. Kumar and Kiran (2008) have presented NER system for five languages including Urdu using CRF, HMM and rules. The system has 39.77, 46.84, 45.84, 46.58, 44.73 F-measures for Bengali, Hindi, Oriya, Telugu and Urdu using rules with HMM and 35.71, 40.49, 36.76, 45.62 and 38.25% F-measures for Bengali, Hindi, Oriya, Telugu and Urdu using hybrid CRF model, respectively. Hybrid HMM model has showed better performance than hybrid CRF model for all the languages. Chaudhuri and Bhattacharya (2008) have developed NER system for Indian script Bangla. Three-stage approach for automated identification Named Entities has been used. Dictionary based, rules based and left-right co-occurrences statistics (n-gram) have been used for Named Entity. A popular corpus named AnandabazarPatrika has been used for system experiments. The system has 85.50% recall, 94.24% precision and 89.51% F-measure.

Srikanth and Murthy (2008) have used CRF based Noun Tagger for Telugu language using 13,425 words manually tagged data for training and 6,223 words as test data. The system has F-value of Noun Tagger up to 92%. The rules based NER system has been developed for identifying names of person, place and organization. Using this rule based tagger through bootstrapping; a manually checked Named Entity tagged corpus of 72, 157 words has been developed. Afterward CRF based NER system has been developed for Telugu. The overall F-measures of the system ranging from 80 to 97%. Srivastava *et al*. (2011) have presented hybrid approach for Hindi NER system. Rules have been formulated over Conditional Random Field (CRF) model and Maximum Entropy (ME) model using features of POS and orthography for overcoming limitations of machine learning models for complex morphological languages like Hindi. The voting method has also been used to improve the performance of the NER system. Based on comparisons, CRF achieves better result than ME and rule based result. Sharma *et al*. (2011) has reported a survey for NER systems for Indian languages including clear explanation of NER and challenges related to NER. The authors have discussed three approaches and existing work with the used methodology for NER system in five Indian languages such as Urdu, Bengali, Telugu, Hindi and Oriya. In addition, results in terms of F-measure for different Indian languages using various approaches have been discussed. Summary of different approaches used for SSEA languages is given in Table 1.

EXISTING WORK ON URDU NER

Earlier research on NER for digital Urdu text has been carried out by Becker and Riaz (2002). Issues pertaining to Urdu language have been discussed and a corpus of 2200 Urdu documents has been developed. A comprehensive contribution has been made by NER workshop publications of IJCNLP in 2008 at IIT Hyderabad on Bengali, Hindi, Oriya, Telugu and Urdu but no study has been performed exclusively for Urdu

Table 2: Different approaches used for developing Urdu NER system

Author	Approaches	F-measures (%)	Corpus
Saha *et al*. (2008a, b)	ME	35.47	36,000 tokens
Gali *et al*. (2008)	CRF	43.46	36,000 tokens
Ekbal *et al*. (2008)	CRF	35.52	36,000 tokens
Kumar and Kiran (2008)	HHM+rules, CRF+rules	44.73, 38.25	36,000 tokens
Mukund *et al*. (2010)	ME, CRF	55.30, 68.90	55,000 tokens
Riaz (2010)	Hand crafted rules	91.10, 81.60	2,262 documents, 36,000 tokens
Singh *et al*. (2012)	Rules based	74.09	1,62,275 tokens

language. Corpus of 36000 words has been provided by IJCNLP 2008 workshop for the researchers to produce their results.

Mukund *et al*. (2010) has developed an information extraction system for Urdu language. The sub module of NER has been developed for information extraction system by using two models; ME and CRF based NER for Urdu. The results of ME have 55.3% F-measures and CRF based module for NER F-measure value of 68.9%.

Riaz (2010) has presented a rule based approach for Urdu NER system. Different rules have been formulated from 200 documents of Becker-Riaz corpus and have extracted 600 documents out of 2,262 documents for better evaluation during experimentation. The system has F-measure of 91.1% with 90.7% recall and 91.5% precision. This rule based NER has been tested on 36000 Urdu words' corpus of NER Workshop IJCNLP 2008 and has achieved F-measures of 72.4% without any change in the rule set. The results have been later improved by developing new rules after analyzing the training set. The developed rule-based approach for Urdu NER shows encouraging results.

Singh *et al*. (2012) presented rules based NER in Urdu languages for thirteen NEs and evaluated the proposed system on two different sets which were collected from different news sources. The overall accuracy rate is 74.09%.

The following Table 2 summarizes different approaches used for developing Urdu NER system.

CHALLENGES IN URDU NER

The large number of ambiguities of NE and the problems related to the Urdu language makes NER a challenging task for Urdu language. The construction of a robust Urdu NER is a complicated task because of the following limitations.

No capitalization: In English orthography capitalization of the initial letter is a specific signal that indicates that a word or sequence of words is a NE. Urdu has no such special signal that makes the detection of NEs more challenging. Thus, in Urdu language there is no difference between a NE and the other word from lexical point of view.

Scarce resources: A standard and huge corpus is the basic requirement for NLP related tasks but unfortunately there is no standard NE tagged Urdu corpus available. The available Urdu NE tagged corpora are:

- EMILLE (2003) corpus
- Becker-Riaz (2002) corpus
- IJCNLP workshop (2008) corpus
- CRL Annotated Corpus

The EMILLE corpus contains long running articles that do not have a lot of Named Entities (Riaz, 2010). Becker-Riaz corpus contains 2,262 short news articles and has a very rich content for Named Entity Recognition. NER workshop of IJCNLP in 2008 provided a big corpus that contained 36,000 Urdu tokens. Computing Research Laboratory (CRL) has an annotated corpus of 55,000 NE in Urdu for the machine translation task. As per (Mukund *et al*., 2010) the data of the CRL is written in the "news writing" style and follows. All of these contain very limited number of tokens as compare to English corpus that has millions of tokens or words.

Agglutinative nature feature: Some additional features can be added to the word to have more complex meaning. Agglutinative languages form sentences by adding a suffix to the root forms of the word, e.g., پا کستا ن (Pakistan is location) to پا کستا نی (Pakistani is also location) but it is difficult for NER system to detect as a NE and classify as location.

Free word order: In Urdu Language SOV (Subject Object Verb) word order is used but usually the writers do not follow the same word order e.g., an English sentence "Ahmad closed the bag of books" can be written in Urdu "کتابوں کا بستہ احمد نے بند کیا" ("Kitabo ka basta Ahmad ne band kia") and " ا حمد نے کتابوں کا بستہ بند کیا" ("Ahmad ne kitabo ka basta band kia"). The use of such different word orders makes the NE identification more challenging.

Complexity of spelling variations: Different writers can write same NEs in various forms using different spellings in different situations even for native names e.g., نعمان، نومان، نمان.

Borrow words: Some words are taken from other languages e.g., (Palwasha) پلو شہ is taken from Pashto language, (Zeemal) ز یمل is taken from Balochi

language and (Toyota) ٹویوٹا is taken from English Language.

Nested named entities: A nested Named Entities are made up of more than one proper name which is nested or overlap with one another. The individual token may need more than one label for nested Named Entity which makes the classification task difficult. For example: عبدالولی خان یونیورسٹی is a NE of the type organization but it also consists of NE of type person (عبدالولی خان). Now consider another nested NE پشاور یونیورسٹی (organization name) but it also contains location name i.e., پشاور. To handle nested NEs in Urdu is very challenging task and still it needs attention of researchers.

Compound named entities: A compound Named Entity is composed of multiple words. This brings more challenges to accurately detect the beginning and the ending of a multi-word NE. To extract such NEs like محمد علی جناح (person name) as single NE is difficult.

Conjunction ambiguity: Some entities are made up by using conjunction word such as اور e.g., علی اوربلال سی این جی (organization name) is a conjunct NE which cannot be recognized as a single NE by the NER system.

Ambiguous nature of NEs: A Named Entity can be used as a person name or organization name or as a word other than nouns e.g., نور is a name of person and also equivalent to the English word "light".

Ambiguity in acronyms: English systems easily recognize acronyms due to the capitalization rule, but in Urdu it is quite difficult to recognize acronyms. For example (UNO) یو-این-او ، (BBC) بی بی سی in Urdu cannot be recognized as NEs.

CONCLUSION

It can be concluded that Urdu NER task has not been thoroughly investigated or experimented with due to scarce resources and the inherent complex features. Urdu is rich morphological language due to the fact that it has inherited various features from many languages like Sanskrit, Arabic, Persian, English and Turkish etc. It lies in the category of right to left languages therefore for processing; Unicode encoding and proper font usage is required. The published research has identified that Urdu language demands detailed investigation regarding the application of different existing techniques employed for NE in different languages. Moreover it emphasizes to explore new techniques and to upgrade the existing ones to tackle all the inherent problems of Urdu language.

REFERENCES

Becker, D. and K. Riaz, 2002. A study in Urdu corpus construction. Proceeding of the 3rd Workshop on Asian Language Resources and International Standardization at the 19th International Conference on Computational Linguistics. August, 2002.

Bikel, D.M., S. Miller, R. Schwartz and R. Weischedel, 1997. Nymble: A high-performance learning name-finder. Proceeding of the 5th Conference on Applied Natural Language Processing. Association for Computational Linguistics, 1997.

Biswas, S., S.P. Mishra, S. Acharya and S. Mohanty, 2010. A hybrid oriya named entity recognition system: Harnessing the power of rule. Int. J. Artif. Intell. Expert Syst. (IJAE), 1(1): 1-6.

Borthwick, A., 1999. A maximum entropy approach to named entity recognition. Ph.D. Thesis, Computer Science Department, New York University, New York.

Chaudhuri, B.B. and S. Bhattacharya, 2008. An experiment on automatic detection of named entities in Bangla. Proceeding of the IJCNLP-08 Workshop on NER for South and South East Asian Languages, pp: 75-82.

Cucerzan, S. and D. Yarowsky, 1999. Language independent named entity recognition combining morphological and contextual evidence. Proceeding of the Joint SIGDAT Conference on EMNLP and VLC, pp: 90-99.

Ekbal, A. and S. Bandyopadhyay, 2008a. Named entity recognition using support vector machine: A language independent approach. Int. J. Comput. Syst. Sci. Eng. (IJCSSE), 4: 155-170.

Ekbal, A. and S. Bandyopadhyay, 2008b. Bengali named entity recognition using support vector machine. Proceeding of the IJCNLP-Workshop on NER for South and South East Asian Languages. Hyderabad, India.

Ekbal, A. and S. Saha, 2011. A multiobjective simulated annealing approach for classifier ensemble: Named entity recognition in Indian languages as case studies. Expert Syst. Appl., 38(12): 14760-14772.

Ekbal, A., R. Haque and S. Bandyopadhyay, 2008. Named entity recogntion in Bengali: A conditional random field approach. Proceeding of IJCNLP. India, pp: 589-594.

Gali, K., H. Surana, A. Vaidya, P. Shishtla and D.M. Sharma, 2008. Aggregating machine learning and rule based heuristic for named entity recognition. Proceeding of the IJCNLP-08 Workshop on NER for South and South East Asian Languages. Hyderabad, India, pp: 25-32.

Goyal, V., 2008. Named entity recognition for south Asian languages. Proceeding of the IJCNLP-08 Workshop on NER for South and South-East Asian Languages. Hyderabad, India.

Gupta, P.K. and S. Arora, 2009. An approach for named entity recognition system for Hindi: An experimental study. Proceeding of ASCNT-CDAC. Noida, India, pp: 103-108.

Kumar, P.P. and V.R. Kiran, 2008. A hybrid named entity recognition system for south Asian languages. Proceeding of the IJCNLP-08 Workshop on NER for South and South East Asian Languages, pp: 83-88.

Li, W. and A. McCallum, 2003. Rapid development of hindi named entity recognition using conditional random fields and feature induction. ACM T. Asian Lang. Inform. Process. (TALIP), 2(3): 290-294.

Mukund, S., R. Srihari and E. Peterson, 2010. An information-extraction system for Urdu-a resource-poor language. ACM T. Asian Lang. Inform. Process., 9(4).

Nadeau, D., Peter D. Turney and S. Matwin, 2006. Unsupervised named-entity recognition: Generating gazetteers and resolving ambiguity. Proceeding of 19th Conference of the Canadian Society for Computational Studies of Intelligence, (AI'06), pp: 266-277.

Raju, B.S., D.S.V. Raju and K. Kumar, 2010. Named entity recognition for Telegu using maximum entropy model. J. Theor. Appl. Inform. Technol., 3: 125-130.

Ramshaw, L.A. and M.P. Marcus, 1995. Text chunking using transformation-based learning. Proceeding of the 3d ACL Workshop on Very Large Corpora, pp: 82-94.

Rau, L.F. and P.S. Jacobs, 1991. Creating segmented databases from free text for text retrieval. Proceeding of the 14th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '91), pp: 337-346.

Riaz, K., 2010. Rule-based named entity recognition in Urdu. Proceeding of the 2010 Named Entities Workshop (NEWS, 2010), pp: 126-135.

Saha, S.K., S. Sarkar and P. Mitra, 2008a. A hybrid feature set based maximum entropy Hindi named entity recognition. Proceeding of the 3rd International Joint Conference on Natural Language Processing. Hyderabad, India.

Saha, S.K., P.S. Ghosh, S. Sarkar and P. Mitra, 2008b. Named entity recognition in Hindi using maximum entropy and transliteration. Polibits, 38: 33-42.

Sang, E.F.T.K., 2002. Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition. Proceeding of the 6th Conference on Natural Language Learning (COLING-02), pp: 1-4.

Sharma, P., U. Sharma and J. Kalita, 2011. Named entity recognition: A survey for the Indian languages. Parsing in Indian Languages, pp: 35-39.

Singh, U., V. Goyal and G.S. Lehal, 2012. Named entity recognition system for Urdu. Proceeding of COLING, pp: 2507-2518.

Srikanth, P. and K.N. Murthy, 2008. Named entity recognition for Telegu. Proceeding of the IJCNLP-Workshop on NER for South and South East Asian Languages. Hyderabad, India, pp: 41-52.

Srivastava, S., M Sanglikar and D.C. Kothari, 2011. Named entity recognition system for Hindi language: A hybrid approach. Int. J. Comput. Linguist. (IJCL), 2(1).

VijayKrishna, R. and L. Sobha, 2008. Domain focused Named Entity Recognizer for Tamil using conditional random fields. Proceedings of the IJCNLP-08 Wokshop on NER for South and South East Asian Languages. Hyderabad, India, pp: 59-66.