DEVELOPING CLASSIFICATION-BASED NAMED ENTITY RECOGNIZERS (NER) FOR SAMBALPURI AND ODIA APPLYING SUPPORT VECTOR MACHINES (SVM)

Pitambar Behera and Sharmin Muzaffar

This paper demonstrates the development of named Entity Recognizers (NER) applying Support Vector Machines (SVM) for Sambalpuri and Odia. The Sambalpuri corpus amounts to 112k word tokens out of which 5,887 are named entities. On the contrary, 250k ILCI corpus has been applied for Odia out of which 18,447 tokens are named entities. The former accurately recognizes 96.72% whereas the latter provides 98.10% accuracy.

Keywords: *NER, Sambalpuri, NLP, Odia, SVM, Machine Learning, Indo-Aryan languages, Information Retrieval, Natural Language Processing.*

1 Overview

Named entity recognition (NER) is one of the applications of Natural Language Processing and it is considered as the subtask of information retrieval. NER is the process of detecting Named Entities (NEs) in a document and to categorize them into certain named entity classes such as the names of organization, person, location, sport, river, city, country, quantity etc. In English, we have accomplished a lot of work pertaining to recognizing named entities. On the contrary, we have not achieved remarkable accomplishment with regard to detecting NER in Indian languages. India is an abode for 22 official languages along with endangered, lesser-known and less-studied languages. NER is still considered to be an emerging area of research in the field of NLP in the context of Indian languages.

There are various applications of NER such as Information Extraction, Question Answering, Information Retrieval, Automatic Summarization, Machine Translation, etc. The Named Entities can be made known to us by performing computation on a given natural language through rule-based or statistical approaches. The task of identification, extraction and retrieving necessary information can be made faster, if we are already acquainted with the nature, type and functions of named entities. Therefore, NER is the process of detecting, classifying and extracting Named Entities in a document into their corresponding classes with the application of any of the NER based approaches.

1.1 Approaches to named entity recognition

There are basically two broad approaches that are employed in the recognition of named entities (Nayan et al., 2008; Sasidhar et al., 2011; Saha, 2008). These include: Rule-based approach and Machine learning based approach (Kaur and Gupta, 2012; Kaur and Gupta, 2010; Srivastava et al., 2011).

1.1.1 Rule-based approach

Under this section, there are list lookup approach and linguistic approach. So far as the former is concerned, gazetteers are exploited that comprise of different lists of named entity classes and a simple look up or search operation is conducted in order to detect whether a word belongs to a named entity class or not. If a particular word belongs to a named entity class, a named entity label, as specified in the annotation schema, is allotted to that word on the basis of the named entity class which it originally belongs to. On the other hand, in linguistic approach, a linguist is entrusted with the work of formulating heuristic linguistic rules, so that the named entities can be identified as well as classified and extracted easily (Ekbal and Bandyopadyay, 2010; Gupta and Lehal, 2011). The formulated rules are language dependent and cannot be applied in order to identify named entities in any other given language (Kaur and Gupta, 2012). Therefore, data-driven statistical approach became indispensable.

1.1.2 Statistical approach

This approach is motivated by the machine learning theories and algorithms, for instance, Hidden Markov Models (HMM), Maximum Entropy Markov Model (MEMM), Conditional Random Field (CRF), Support Vector Machines (SVMs), Decision Tree and so on.

2 Review of literature

For English, we have developed rule-based NER systems of which the F-measure accuracy ranges from 88-92% (Grishman, 1995; Wakao et al., 1996). These rule-based systems are inoperable and they demand huge amount of linguistic knowledge of the given language. Machine Learning based techniques are operable in nature and does not require huge linguistic knowledge in that language.

Hidden Markov (Bikel et al., 1997), Maximum Entropy (Borthwick), CRF (Li and Mccallum, 2004) models have been commonly applied to different languages for NER identification purposes. A hybrid system combining ME, HMM and rule-based methods has been developed by Srihari et al. (2000).

Gali et al (2008) have reported lexical F-Score accuracy of 40.63%, 39.04%, 40.94%, 43.46% and 50.06% for Bengali, Oriya, Telugu, Urdu and Hindi respectively. For Indian languages, Ekbal and Bandopadhyay (2007) have described an approach to lexical pattern learning. For Hindi and four other European languages, Cucerzan and Yarowsky (1999) have developed NER applying morphological and contextual features. They registered an f-measure of 41.70%. Li and Mccallum (2004) have developed an NER for Hindi applying CRFs and registered an f-measure of 71.50. Kumar and Bhattacharya (2006) have achieved an accuracy of 79.7% f-measure on a maximum entropy markov model for Hindi. Krishnarao et al (2007) have demonstrated a comparative analysis of CRF and SVM for the purpose of recognizing named entities in Hindi. Chopra et al (2012) have prepared an NER for Hindi combining rule-based heuristics and HMM where they have achieved 94.61% accuracy.

Shishtala et al. (2003) have developed an NER for Telugu applying CRF and registered a reported F-value of 44.91%. Ekbal and Bandyopadhyay (2008) have developed an NER for Bengali applying SVM and reported an F-Score of 91.8%. Kaur and others have developed an NER for Punjabi language applying CRF. Balabantaray and others (2013) have developed an NER in Odia having parameterized CRF++ tool in various ways. In doing so, they have applied the gazetteer and parts of speech tags so as to extract various feature sets. So far as the corpus distribution is concerned, they have applied 45k of word tokens data out of which only 1k are named entities. They have achieved an overall precision value of 92%.

3 Support vector machines

Support vector machines are a group of supervised learning models developed by Vapnik (Joachims, 1999). They are associated with learning algorithms that are responsible for analyzing data and recognizing patterns that are further applied to the purpose of classification and regression analysis tasks. When you provide a finite set of training data by marking each as belonging to either of the categories, the SVM training algorithm develops a model. It provides labels to new examples learned from the training examples, making it a non-probabilistic binary linear classifier.

The SVM learns a linear hyperplane when given a particular series of N training examples $\{(x1, y1),\ldots, (x_N, y_N)\}$ where an example x_i stands for a vector R^N and the class annotation label is $y_i \in \{-1,+1\}$. The hyperplane distincly separates positive instances from the set of negative ones with an optimum margin. That maximal margin is defined as the distance of the hyperplane to the nearest of the positive instance and negative set of examples (Gimenenez and Marquez, 2006).

A non-linear classifier decides f (x)=sign(g(x)) for an input vector where f(x)=+1 (x is a member of a given class), f(x)=-1 (x is not a member of a given class). g(x) is proportionate to m, z_i s are support vectors and

$$g(x) = \sum_{i=1}^{m} w_i K(x, z_i) + b$$

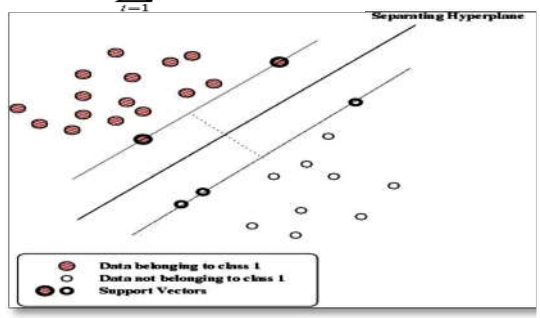

Fig 1: SVM example of hard margin classifying negative and positive examples

K stands for kernel.

Given some training data $\mathcal{D}$, a set of n points of the form

$$\mathcal{D} = \{(\mathbf{x}_i, y_i) \mid \mathbf{x}_i \in \mathbb{R}^p,\ y_i \in \{-1, 1\}\}_{i=1}^{n} \quad (4)$$

4 Methodology

4.1 Corpus annotation schema

Concepts are categorized in three broad categories: Location, Organization, and Person. But we have categorized those into seven general categories so as to incorporate all of them. Entities of person are the names of persons living or dead, of deities, of fictional characters etc. For instance, Hari, Gangadhar Meher, Kalapahada. Organization entities are restricted to institutions, corporations and government agencies such as Sambalpur University, DRDO, etc. Similarly, the entities pertaining to location are countries, streets, mountains, airports, monuments such as Sambalpur, Jharsuguda Airport, Bir Surendra Sai Statue. Temporal entities are the names for months, weeks, special days as for example September, Independence Day. Date Entities are the entities suggesting the dates of a particular month. Entities referring to currency names are Rouble, Yen, Rupee, Yuan etc. Names referring to percentage and fractional numbers are clubbed under this category.

The annotation schema (see fig 2) consists of seven labels such as person (NNP_PER), organization (NNP_ORG), location (NNP_LOC), time (NN_TIM), date (NN_DAT), currency (CD_VAL), and percentage (CD_PER). We have developed our own annotation schema considering various lexical feature sets of named entities in Sambalpuri and Odia.

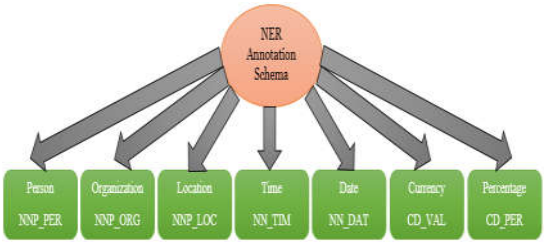

Fig 2: NER annotation schema

The figure demonstrated below explains the examples of named entities pertaining to the broad category of location.

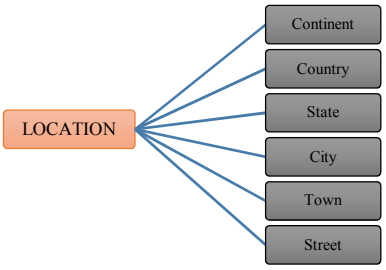

Fig 3: Examples of location named entity

Table 1 below contains the annotation samples of Sambalpuri NER data as according to their respective categories.

Table 1: Sambalpuri annotation sample

Labels	**Sambalpuri Annotation Sample**
NNP_LOC	bʰɛʈɛn NNP_LOC
NNP_NAM	səməlpurɪ NNP_NAM
NNP_ORG	səməlpur NNP_ORG ɪnbʰɑ:rsɪʈɪ NNP_ORG
NNP_PER	pəndɻt NNP_PER josi: NNP_PER
NN_TIM	sɛpʈembər NN_TIM
NN_DAT	12.09.2018 NN_DAT
CD_VAL	ɖɔlɑ:r CD_VAL
CD_PER	12% CD_PER

4.2 Corpus distribution

For the current study, we have taken 112k and 250k data for Sambalpuri and Odia respectively. Out of those aforementioned data, 5887 and 18, 447 have been labelled as NERs in Sambalpuri and Odia respectively. For the purpose of testing, we have applied 2503 word tokens for Sambalpuri and 3003 tokens for Odia.

The corpora for Sambalpuri have been adapted from (Behera et al., 2018); created for POS annotation work. On the other hand, 250k ILCI corpora have been applied for Odia out of which 18,447 tokens are named entities. The POS-annotated corpora for Odia have been adapted from Behera (2016).

Table 2: Corpus distribution

Languages	Training Corpus	Testing Corpus
Sambalpuri	5887/112k	2503
Odia	18, 447/250k	3003

4.3 Feature extraction

The features for a classification-based NER have been selected considering the word tokens, POS, ambiguity and maybe's. The tri-gram feature file has been applied. The configuration file applied in the learning phase encapsulates medium verbose (-V 2) and the directions of automatic learning and annotation have been set to the left-right-left (LRL) mode. All the miscellaneous features have been set to the default mode.

```
# SVMT configuration fileNAME = /home/sanskrit/svmtool/models/odi/ODI
TRAINSET =/home/sanskrit/svmtool/odia.trainSVMDIR
=/home/sanskrit/svmtool/svmlight/W = 5 2 F = 5 10000 X = 7 Dratio = 0.005
REMOVE_FILES = 1do M0 LRL#do M1 LRL#do M2 LRL#do M4 LRL#
--------------------------------------------------------------------------
--------#ambiguous-right [default]A0 = w(-3) w(-2) w(-1) w(0) w(1) w(2) w
(3) w(-2,-1) w(-1,0) w(0,1) w(-1,1) w(1,2) w(-2,-1,0) w(-2,-1,1) w(-1,0,1)
w(-1,1,2) w(0,1,2) p(-3) p(-2) p(-1) p(-2,-1) p(-1,1) p(1,2) p(-2,-1,1) p
(-1,1,2) a(0) a(1) a(2) a(3) m(0) m(1) m(2) m(3) z(2) z(3) z(4) ca(1) cz
(1)A0unk = w(-3) w(-2) w(-1) w(0) w(1) w(2) w(3) w(-2,-1) w(-1,0) w(0,1) w
(-1,1) w(1,2) w(-2,-1,0) w(-2,-1,1) w(-1,0,1) w(-1,1,2) w(0,1,2) p(-3) p(-
2) p(-1) p(-2,-1) p(-1,1) p(1,2) p(-2,-1,1) p(-1,1,2) k(0) k(1) k(2) k(3)
m(0) m(1) m(2) m(3) a(2) a(3) a(4) z(2) z(3) z(4) ca(1) cz(1) L SA AA SN CA
CAA CP CC CN MW#
```

Fig 4: The configurations file for Sambalpuri and Odia NERs

5 Evaluation

The statistical representation in the below tabulated data (see Table 3 and Table 4) demonstrates the evaluation of the NER systems for Sambalpuri and Odia category-wise on three measures: perecntage, precision and recall. Both in Table 3 and Table 4, the highest accuracy is figured in the categories of proper names, person, location, and organization which is indicative of the fact that at the level of POS category of proper nouns, the Sambalpuri (Behera and Dash, 2017) POS tagger (Behera et al., 2018) and Odia (Jha et al., 2014) POS tagger (Behera, 2016, Behera, 2017, Ojha et al., 2015) have also registered a fair amount of accuracy. On all the measures, cent percent accuracy has been achieved in the cattegory of named enitity of time. In both these systems, named entity of persons has registered almost the same amount of accuracy as explained in the following comparative data. The zero figured in Table 3 and Table 4 refers to the fact that during evaluation the test file does not contain any word for the respective categories.

Table 3: Evaluation of Sambalpuri NER

Labels	(%)	precision	recall
NNP_LOC	92.15	96.79	92.15
NNP_NAM	97.60	100	100
NNP_ORG	91.94	93.71	94.17
NNP_PER	99.09	98.60	99.09
NN_TIM	100	100	100
NN_DAT	0	0	0
CD_VAL	100	100	100
CD_PER	0	0	0
Total	96.72	72.62	73.17

Table 4: Evaluation of Odia NER

Labels	(%)	Precision	Recall
NNP_LOC	99.04	98.57	99.04
NNP_NAM	89.83	100	100
NNP_ORG	97.13	96.73	97.13
NNP_PER	99.03	97.73	99.03
NN_TIM	100	100	100
NN_DAT	100	100	100
CD_VAL	83.33	50	100
CD_PER	0	0	0
Total	98.10	80.37	86.90

6 Computational framework

This section highlights the computational architecture of the user interface (see Fig. 5). At the first stage, raw text in Sambalpuri and Odia is provided. Thereafter, the system pre-processes the junk characters and avoids unwanted elements in the text. At the third stage, the texts are tokenized. Then, the text goes to the SVM tool where it is

POS-tagged after being pre-processed. The POS-tagged text is again annotated with NER annotation labels. The text gets detokenized at the detokenization level. Finally the system provides the NER-tagged output after detokenization.

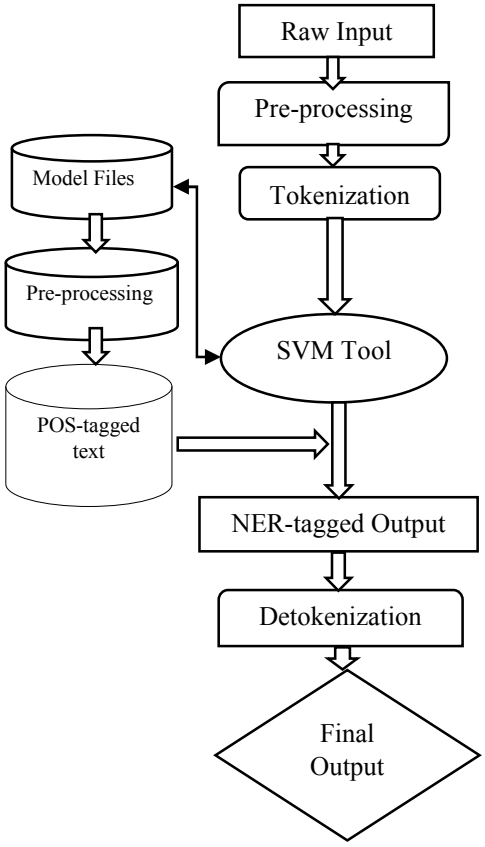

Fig 5: User interface architecture

7 Issues and challenges

One of the major issues faced by Indian languages written in their respective scripts is that there is no capitalization applied for the proper named entities at the initial alphabet unlike English. Therefore, it becomes quite challenging in order to detect the entities properly. The second challenging issue is that they have both inflectional and agglutinating properties, enriched morphological features and relatively free word order. Another inherent linguistic issue to be pondered upon is that they use several common nouns as the proper nouns for example, Pawan, Chandni, Dharti. In addition, dictionaries applied for the recognition task of named entities also contain the above issue that makes the detection process even more complicated. Furthermore, another issue related to the linguistic challenge is the lack of standardization in spelling, especially for the non-scheduled languages such as Sambalpuri, Bhojpuri, Magahi, Awadhi and many others.

The web has been dominated by the English and Mandarin languages. Above all, we need to accept the fact that even after years of endeavor by the CIIL, TDIL, IIIT Hyderabad and other institutions, the resource scarcity situation has not changed drastically so far. As a result, we don't have properly labelled data for NER with standaridization. Owing to the scarcity in resource, we do not have name dictionaries, POS taggers, morphological analyzers, gazetteer lists and so on that can be utilized for the development of NER systems.

So, the need of the hour is to make the Indian languages resource rich by making them available on the digital space. Secondly, we need to develop unambiguous dictionaries clearly distinguishing the proper noun and common noun labels for the same word as exemplified aforementioned. To address this issue, we need to develop more frequency dictionaries which can predict the frequency of the co-occurrence of the proper with the common nouns. The phonetic characteristic of the Brahmi script of the Indian languages can be exploited by the NER systems in order to get the desired output.

8 Conclusion

The rationale for dealing with Sambalpuri and Odia simultaneously is that they are considered to be closely related independent languages. We have presented two classification-based SVM NERs for Odia and Sambalpuri in this study. In addition, we have also applied the POS taggers for respective languages during detection process of named entities. We have achieved 96.72% and 98.10% accuracy in Sambalpuri and Odia respectively. We have evaluated Sambalpuri 72.62/73.17 and Odia 80.37/86.90 in precision and recall methods respectively. Our system performs better than already reported Odia NER. The NER system for Sambalpuri is the first

attempt at successfully recognizing named entities for the language.

This research work is limited to only the identification and extraction of named entities without nested structure. NERs will be fruitful for Information Extraction, Question Answering, Information Retrieval, Automatic Summarization, Machine Translation etc. in Sambalpuri and Odia for further future research.

Acknowledgements

We are indebted to the reviewers and participants of the 38th Linguistics Society of Nepal Annual Conference held in 2017 for their suggestions for qualitative improvement. We also acknowledge the ILCI Project for providing us data for Odia applied in this current research.

References

Balabantaray, Rakesh Chandra; Das, Suprava; and Mishra, Kshirabdhi Tanya. 2013. Case study of named entity recognition in Odia using CRF++ tool. *IJACSA*, 4 (6), 213-216.

Behera, Pitambar. 2016. Evaluation of SVM-based automatic parts of speech tagger for Odia. In *Proceedings of WILDRE-3 (LREC-2016), ,* Portoroz, Slovenia, 32-38.

Behera, Pitambar. 2017. An experimentation with the CRF++ parts of speech tagger for Odia. *Language in India*. Vol 17:1, Jan, (1930-2940), Bloomington, USA.

Behera, Pitambar; Ojha, Atul Kumar; & Jha, Girish Nath. 2015. Issues and challenges in developing statistical pos taggers for Sambalpuri. Human Language Technology: Challenges for Computer Science and Linguistics, ed. By Zygmut Vetulani et al. 393-406. LNAI 10930, Springer International Publishing.

Behera, Pitambar. & Dash, Biswanandan. 2017. Documenting Sambalpuri-Kosli: The Case of a Less-resourced Language. *Indian Journal of Applied Linguistics (IJOAL)*. Bahri Publications, 43(1-2), 128-144.

Chopra, Deepti; Jahan, Nusrat; & Morwal, Sudha. 2012. Hindi named entity recognition by aggregating rule based heuristics and Hidden Markov Model. International Journal of Information, 2(6).

Ekbal, Asif & Bandyopadhyay, Sivaji. 2010. Named entity recognition using Support Vector Machine : A language independent approach. International Journal of Electrical and Electronics Engineering, 4:2.

Gimenez, Jesus &Marquez, Lluis. 2006. SVMTool technical manual v1.3, Barcelona: TALP Research Center, LSI Department, Univeritat Plotecnica de Cataluniya.

Gupta, Vishal & Lehal, Gurpreet Singh. 2011. Named entity recognition for Punjabi language text summarization. International Journal of Computer Applications, 33:3.

Joachims, Thorsten. (1999). Making large scale SVM learning practical. In: Schölkopf B, Burges C, Smola A (eds) Advances in kernel methods - support vector learning. MIT Press, Cambridge.

Jha, Girish Nath; Hellan, Lars; Beermann, Dorothee; Singh, Srishti; Behera, Pitambar; & Banerjee, Esha. 2014. Indian languages on the TypeCraft platform–The case of Hindi and Odia. In *Proceedings of WILDRE-2014 Reykjavik, Iceland*, 84-90.

Kaur, Darvinder & Gupta, Vishal. 2010. A survey of named entity recognition in english and other indian languages. International Journal of Computer Science, 7:6.

Kaur, Darvinder & Gupta, Vishal. 2012. Name entity recognition for Punjabi language. International Journal of Computer Science and Information Technology & Security 2:3.

Nayan, Animesh; Rao, B. Ravi Kiran; Singh, Pawandeep; Sanyal, Sudip; & Sanyal, Ratna. 2008. Named entity recognition for Indian languages. In IJCNLP, 97-104.

Ojha, Atul Kumar; Behera, Pitambar; Singh, Srishti; & Jha, Girish Nath. 2015. Training & evaluation of POS taggers in Indo-Aryan languages: A case of Hindi, Odia and Bhojpuri. In *Proceedings of LTC-2015, Poland*, 524-529.

Saha, Sujan Kumar; Chatterji, Sanjay; Dandapat, Sandipan; Sarkar, Sudeshna; & Mitra, Pabitra. 2008. 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 languages, 17-24.

Saha, Sujan Kumar; Sarkar, Sudeshna; & Mitra, Pabitra. 2008. Gazetteer preparation for named entity recognition in Indian languages. IJCNLP. 9-16.

Sasidhar, B.; Yohan, P. M.; Vinaya Babu, A.; & Govardhan, A. 2011. A survey on Named Entity Recognition in Indian languages with particular reference to Telugu. IJCSI International Journal of Computer Science Issues, 8:2.

Srivastava, S.; Sanglikar, M.; & Kothari, D.C. 2011. Named Entity Recognition system for Hindi language: A hybrid approach. International Journal of Computational Linguistics (IJCL), 2:1.

Kaur, Amandeep; Josan, Gurpreet S; & Kaur, Jagroop. Named entity recognition for Punjabi: A Conditional Random Field approach. In the proceedings of 7th International Conference on Natural Language Processing, Macmillan Publishers, India.

Shishtla, Praneeth M; Gali, Karthik; Pingali, Prasad; & Varma, Vasudeva. 2008. Experiments in Telugu NER: A Conditional Random Field approach. In the proceedings of the IJCNLP-08 Workshop on NER for South and South East Asian Languages, 105-110.

Bikel, Daniel M.; Miller, Scott; Schwartz, Richard; &Weischedel, Ralph. 1997. Nymble: A high performance learning name-finder. In Proceedings of the Fifth Conference on Applied Natural Language Processing, 194–201.

Borthwick, Andrew. 1999. A Maximum Entropy Approach to named entity recognition. Computer Science Department, New York University, Ph.D. thesis.

Cucerzan, Silviu & Yarowsky, David. 1999. Language independent named entity recognition combining morphological and contextual evidence. In Proceedings of the Joint SIGDAT Conference on EMNLP and VLC 1999, 90–99.

Krishnarao, Awaghad Ashish; Gahlot, Himanshu; Srinet, Amit; & Kushwaha, D. S. 2009. A comparison of performance of sequential learning algorithms on the task of Named Entity Recognition for Indian languages. In the proceedings of 9th International Conference on Computer Science. Baton Rouge, LA, USA, 123-132.

Gali, Karthik; Surana, Harshit; Vaidya, Ashwini; Shishtla, Praneeth; & Sharma, Dipti Misra. 2008. Aggregating machine learning and rule based heuristics for named entity recognition. In the proceedings of the IJCNLP-08 Workshop on NER for South and South East Asian Languages, Hyderabad, India, 25-32.

Ekbal, Asif & Bandyopadhyay, Sivaji. 2007. Lexical pattern learning from corpus data for named entity recognition. In Proceedings of International Conference on Natural Language Processing (ICON).

Ekbal, Asif & Bandyopadhyay, Sivaji. Bengali Named Entity Recognition using Support Vector Machine. In the proceedings of the IJCNLP-08 workshop on NER for South and South East Asian Languages, Hyderabad, India, 51-58.

Grishman, Ralph. 1995. The New York University system muc-6 or where's the syntax? In Proceedings of the Sixth Message Understanding Conference.

Kumar N. & Bhattacharyya Pushpak. 2006. Named Entity Recognition in Hindi using MEMM. In Technical Report, IIT Bombay, India.

Li, Wei & Mccallum, Andrew. 2004. Rapid development of Hindi named entity recognition using Conditional Random Fields and feature induction. In ACM Transactions on Computational Logic.

Srihari, R.; Niu, C.; & Li, Wei. 2000. A Hybrid Approach for Named Entity and Sub-Type Tagging. In Proceedings of the sixth conference on Applied natural language processing.

Wakao T.; Gaizauskas, R; & Wilks, Y. 1996. Evaluation of an algorithm for the recognition and classification of proper names. In Proceedings of COLING-96.