Named entity recognition in Odia language: a rule-based approach

Amrita Anandika, Sujata Chakravarty and Bijay Kumar Paikaray*

Department of CSE,
Centurion University of Technology and Management,
Odisha, India
Email: amrita.anandika@gmail.com
Email: chakravartys69@gmail.com
Email: bijaypaikaray87@gmail.com
*Corresponding author

Abstract: NLP can be defined as a process of automatic handling of human spoken languages by computers. NLP is a vast research field linked to multiple areas like artificial intelligence, machine translation, information retrieval etc. NER is an information extraction (IE) process concerned with extracting named entities (NE) from raw data and categorising these NEs into some predefined classes. The process of recognising and extracting NEs from unstructured corpus or data is an essential task for solving different complications in various research fields such as, question answering, summarisation system, information extraction, machine learning, semantic web search, bio-informatics, video annotation and many more. For this research work a classical methodology, i.e., rule-based approach is used to construct a system for automatic identification of NEs from tourism domain data. The system considers words with their repetition and without their repetition and acquires 83% and 71% accuracy respectively.

Keywords: named entity recognition; NER; named entity extraction; information retrieval; machine translation; natural language processing; NLP.

Reference to this paper should be made as follows: Anandika, A., Chakravarty, S. and Paikaray, B.K. (2023) 'Named entity recognition in Odia language: a rule-based approach', *Int. J. Reasoning-based Intelligent Systems*, Vol. 15, No. 1, pp.15–21.

Biographical notes: Amrita Anandika is currently pursuing her PhD in the Department of Computer Science and Engineering, Centurion University of Technology and Management, Odisha, India. Her areas of research include natural language processing and machine learning.

Sujata Chakravarty is currently working as a Professor in the Department of CSE, Centurion University of Technology and Management, Bhubaneswar, India. Her research areas include information security, computational intelligence and evolutionary computing, bio-medical data classification, smart agriculture, natural language processing, intrusion detection system in computer-network, and analysis and prediction of different financial time series data.

Bijay Kumar Paikaray is currently pursuing his PhD in the Department of Computer Science and Engineering, Centurion University of Technology and Management, Odisha, India. His areas of research include high-performance computing, information security and IoT.

1 Introduction

The term NE is currently used in various IE applications. In 1996, it was introduced by Grishman and Sundheim in the Sixth Message Understanding Conference (MUC-6). Named entity recognition (NER) can be defined as a process of information extraction that is concerned with recognising and classifying NEs in a given sentence or text. It is a subpart of NLP (Mansouri et al., 2008; Anandika and Mishra, 2019). Natural language can be defined as the communication approach between humans specifically, speech and text. NLP is the method of automatic manipulation of human spoken language, i.e., natural language by software.

NER is being an essential task in the field of NLP. It is a part of information extraction process which handles structured as well as unstructured text and identifies the terminologies that indicates name of people, places, organisations and companies. Generally NER has two steps. First step identifies the appropriate names from the given text, and the second step, categorises these names into set of groups or classes such as person names, organisations like committees, companies, government organisations, locations such as cities, countries, rivers, date and time expressions, etc. (Dhariya et al., 2017). These groups are previously defined. NE hierarchy is classified into three types where the first one is entity (ENAMEX). It consists of

person, organisation and location name. Second one is time expression (TIMEX). It consists of date and time. Third one is numeric expression (NUMEX). It consists of money, percentage values.

Different types of methodologies exist in research literature for NER extraction, such as: terminology-driven NER, rule-based NER, machine learning-based NER, hybrid NER. Presently, machine-learning (ML) approaches are widely used for NER due to the nature of easily trainable and adoptable to other domains and languages. Moreover, in ML approaches maintenance of trained data is less expensive. However such methodologies demand rich corpuses which are not available as on date for many languages where very poor research has been done like Odia language. Hence, researchers need to look for more feasible alternative like rule-based method.

In the following manner rest of the paper is arranged. Section 2 presents literature review. In Section 3 NER in Odia language is discussed. Section 4 represents different challenges associated with Indian languages. Section 5 gives brief description of the proposed system. In Section 6, experimental result of the proposed system is given and finally, Section 7 concludes the research work and focuses on future directions.

2 Literature review

Biswas et al. in 2009 have designed a hybrid model by using maximum entropy model (MEM) and hidden Markov model (HMM) along with some linguistic rules to identify NEs (NE) in Odia language. Firstly, they have used MEM to identify NEs in Odia corpus, and then they tagged them temporarily as reference. After this to train the HMM they have used the tagged corpus of MEM as training process. This method can reach to high precision and high recall, if it is being supplied with sufficient data for training and a proper error correction mechanism. They have used some grammatical features such as orthography feature, word suffix and word prefix, part-of-speech feature, morphological information and information regarding surrounding words as well as their corresponding tags for developing an Odia NER system based on MEM and HMM. Gazetteers are used to identify the designation and title of a person name. They have also developed some linguistic rules in Odia language to identify time, number etc.

Abdallah et al. in 2012 have developed a NER system for Arabic language (NERA) by integrating a machine learning classifier with a rule-based system which is previously developed. This system is a combination of rule-based NERA system, and ML classification system. In this system, two sets of features are extracted for each word from unstructured text where the first set is based on rule-based features and it contains the name entity tags that are identified by rule-based approach by considering the word in question and window of surrounding words and the features of second set are general features which are based on the experience of a developer. A parser, gazetteer and filtration mechanism is required by this system. This recognition system consists of following two steps: lookup procedure also called as whitelist/gazetteer which contain lists of known NEs. It performs the identification based on the lists of NEs. A parser, that contains sets of different grammatical rules. These rules are expressed as regular expressions which are derived from local lexical content analysis. These whitelists are fixed static dictionaries of NEs which are matched with the target text. The words of target text that exactly match with the whitelist entries are NEs.

Petasis et al. in 2012 have presented a system which supports both rule-based NER approach and classification approach. This system follows an inventive approach for the use of learning in NERC. This system does not use ML in the construction of NERC system, rather it is used autonomously. The system that is created by using ML is used to monitor the performance of existing rule-based NERC system. Feedback regarding the under control rule-based approach like, whether the rule-based approach is out of date and it requires an update is provided by the new system. This system does not need any manual tagging for training data which is the main advantage. An iterative process is used in this approach. First to train the classifier few labelled examples are supplied, then to test the classifier some unlabeled examples are given. They have implemented this approach in Greek and French language.

Day et al. in 2014 have developed a semi-hybrid approach for the NER task. In development of such system they have focused on development of stemming tool, POS tagging tool and a NER detection tool by combining HMM with look-up algorithm and rule-based approach. In the proposed system most of the work is done by HMM (Dey et al., 2014). The look-up algorithm and some rules are used for handling ambiguous words. This system can also be used for solving word sense disambiguation problem.

Mathur and Saxena in 2014 have developed a system which transliterates NEs from English language to Hindi language. The system contains two modules. To transliterate the NEs, both the modules use phoneme-based approach. A CMU pronouncing dictionary that has a collection of 133,270 words along with their respective pronunciation is used by the module.1. Suppose a word which is to be transliterated and it is not found in CMU pronouncing dictionary then, module.2 will be used. Module.2 uses 5-gram approach where maximum of five letters are used for generating transliterated target letter.

Bajwa and Kaur in 2015 developed a combination of rule-based and supervised learning (HMM)-based approach. Two different interpretations are used in this paper. First interpretation is HMM-based and second one is based on the combination of rule-based and HMM. But the disadvantage with this approach is that proper nouns are not automatically tagged which are directed to the generation of training and testing dataset as no dataset is accessible.

Ahmad and Satyaraj in 2015 have implemented MEM for retrieval of NEs from the database. To train the system they have used gazetteer list so that the system will retrieve those words that have the maximum entropy amongst all

others. It is also proved to be the fastest method in retrieving and classifying the entity sets from database. Advantage of this method is, it has increased the freedom of choosing features to represent observations and sequence tagging. It is also observed that when the MEM is used to retrieve the information from the gazetteer list it gives better result.

Chopra et al. in 2016 developed a HMM for the NER in Hindi language. Advantage of this approach is that, difficulty of managing the list of open class words while adopting the gazetteers is handled effectively. But this methodology cannot handle unknown names in a precised way.

Boros et al. in 2017 have developed an optimised decision tree computation algorithm that follows the guidelines of ID3 algorithm. This algorithm calculates entropy and information gain by using single pass over to the training data. They have also developed a tree-pruning algorithm to solve the issue of over fitting of the training set. They implemented a result caching method in order to increase the speed of the system. They have seen that the tree-pruning achieves satisfactory accuracy.

Wang et al. in 2008 have designed and implemented classifiers ensemble approaches for biomedical NER by using four different learning algorithms. They are generalised winnow, conditional random field, support vector machine and maximum entropy. They also compared the performances of three different classifiers ensemble strategies, i.e., arbitration rules, stacked generalisation and cascade generalisation. The stacked generalisation involves class-stacking and class-attribute-stacking. They have also discovered different features for biomedical NER like local features, full text features and external resource features.

3 Named entity recognition in Odia

Odia is known as the official language of Odisha state and the first language of more than 35 million of people. It is also used as second and third language by many people of India. The internal linguistic composition of state Odisha has many tribal groups that linguistically belong to two different individual language family that are Dravidian and Munda (Swain and Pati, 2013). Through borrowing in Odia the NEs in these languages are nativised.

Odia language is recognised as a classical language because it is rich in literature and also has a history of more than thousand and twenty years. Odia script consists of 64 letters (14 vowels, 50 consonants) and 10 digits. In Odia language there is no upper case or lowercase letters. Odia language is a free word order language meaning that in Odia one single sentence can be written in different ways (Balabantaray et al., 2013), e.g.:

- ଭୁବନେଶ୍ୱର ଓଡିଶାର ରାଜଧାନୀ। (Bhubaneswar Odishara rajadhanee)

This sentence can also be written as,

- ଓଡିଶାର ରାଜଧାନୀ ଭୁବନେଶ୍ୱର। (Odishara rajadhanee Bhubaneswar)

One major problem in recognising NE in Odia is that in this language no morphological and punctuation marks are present which can help in identifying the NEs. It can be concluded that, in Odia NEs are based on the semantic features of this language. NEs of Odia language have similarity with the NEs of other languages to some extent and it is also different from NEs of other languages (Abdallah et al., 2012; Biswas et al., 2010). It is different according to the area that is associated with tourism, history and culture of the state, geographical diversity.

Although lots of work has been done in NER for many Indian as well as non-Indian languages but, a very little work has been done for Odia language. Hence, enough resources and appropriate corpus are not available for Odia language. In this study a rule-based model has been adopted for identifying NEs from any document. After identification of NEs it has been categorised into person, location and organisation names as per rules. To implement this system an Odia corpus from tourism domain has been adopted. Hence, detection of NEs from a less annotated corpus is the main challenge.

4 Challenges for Indian languages

For many European languages NER system works correctly especially for English language, whereas, many difficulties are still there for Indian languages (Anandika and Mishra, 2019; Shah et al., 2016). Some of the difficulties are discussed here:

- *No capitalisation:* in English language capitalisation plays vital role for recognising proper nouns. But, in Indian languages capitalisation cannot be applicable.
- *Morphologically rich:* Indian languages are morphologically very rich that's why it is too difficult to identify the root word, e.g., in Hindi, words Pyasa and Pyasi both represents same meaning that is Thristy. But the original word is Pyas (Thrist).
- *Ambiguity between common noun and proper noun:* for Indian languages the most common issue is ambiguity between common noun and proper noun because most of the names of the people are dictionary words (like Gagan, Aakash) and without capitalisation unlike western names.
  - a *Company vs. fruit:* some name entity which represents name of an organisation can be used as name of a fruit, e.g., Apple.
  - b *Person vs. month:* sometimes a month name is used as name of any person, e.g., June.
  - c *Date vs. time:* in some cases date expression represents both date and time expression.

d *Person vs. location:* some name entity which indicates a location name can also be used as any person's name, e.g., Let us consider the common word Rose. It means Rose is a flower but at the same time it can also be name of a person which creates ambiguity between common noun and proper noun.

- *Ambiguity in suffixes:* Indian languages may have a number of post positions attached to the root word to form a single word. For example: Manipur refers to name of a location but Manipuree means anyone who lives in Manipur. When 'ree' is added to the word meaning of the root word gets changed.
- *Standardisation deficiency and spell variations:* one of the biggest problems in many Indian languages is the spelling variation, i.e., same word is spelled differently by different people, e.g., a word 'you' in Hindi it is spelled as 'aap', 'tum', 'tu'. In Odia it is spelled 'apana', 'tume'. In Bengali, it is spelled as 'apani', 'tumi'. In Assamese, it is spelled as 'aponar', 'tomar'. In Telugu it is spelled as 'miru'.
- *Less resources and less labelled data:* to work in NER for any Indian languages a person faces many difficulties like, very less availability of resources as well as a proper annotated data because very less amount of work has done in NER for different Indian languages. The tools required for preprocessing the data such as chunking and speech tagging gives very poor performance (Chopra et al., 2016). In Indian languages it is too difficult to find large amount of corpus or training data. Basic resources like good morphological analyser, part of speech (POS) tagger, name lists are not available for many Indian languages or are at research stage. Whereas a large amount of resources are available in English.
- *Agglutinative nature:* when some additional features are added to a word in order to make its meaning more complex then it is called as agglutinative nature which is very common for Indian languages, e.g., in Assamese, Guwahati refers to a name entity of type location but Guwahatiya is not a name entity. It refers anyone who stays in Guwahati. Hence, the name entity can be found which may have appeared as compound word or with any suffix. In this case the root word of the name entity has to be identified.
- *Foreign words:* some NEs are often language specific like State Bank of India. When such entity is used in any other language text, either it has to be translated to that language or transliterated. Transliteration is defined as a process of writing source language expression in target language characters based on phonetic similarity, e.g., 'Aapnaar naam kee?' (What is your name?) And 'Morbhal' (I am fine) are the transliterations of Bengali and Assamese language in Roman. NE phrases are most difficult to translate because they are domain specific and not found in bilingual dictionaries. Translation of some NEs does not make any sense like All India Radio, Air India etc. Hence transliteration of NEs is more important even when they can be translated.

Moreover Odia language is highly inflectional and morphologically rich in nature. Odia language has relatively free word order meaning that a name entity can appear for both subject and object positions. At the same time there is no subject-verb agreement and due to free word order, named entities can appear at any position. In English language, web sources for name lists are available. But there are no such lists available for Odia language (Swain and Pati, 2013). Lack of labelled data. Larger gazetteers are not available in Odia language. To overcome above stated problems associated with Odia language a rule-based system is proposed for NER.

5 Proposed rule-based approach

Initially, NER systems were constructed on the basis of some hand-crafted rules. Basically, the rules that are created by humans form the background of a rule-based NER system. A rule-based approach is a method which uses human made rules for storing, sorting and manipulating data. It is a classical approach of NER used by many systems (Dhariya et al., 2017). A rule-based system needs a set of information or data source along with a set of rules to manipulate that data. Most of the time, these rules are referred to be as if statements, i.e., if A happens then do B. Generally, it focuses on retrieving names on the basis of some rules (Wakao et al., 1996). Such as: grammatical (part or speech), syntactic (word precedence), orthographic feature (capitalisation), dictionary. These rules are manually written by a linguistic expert. Normally, name identification system executes the text in three different phases (Gali et al., 2008). They are: recognising phrases, recognising patterns, and merging.

5.1 Reasons behind using rule-based system

In rule-based system, the main advantage is, it does not require any training data. Precision is high in rule-based system. The system is cost efficient. This system provides better accuracy. The system can be available to the user easily. The system operates at high speed. Due to predefined rules, error rate is very less in rule-based system. Hence, the system possesses high accuracy. There is reduced amount of risk in terms of accuracy. Output of the system is stable as they are generated according to the predefined rules. Hence the output cannot be indeterminate. The rule-based system gives result in the same way as a human does.

In this work a tagged Odia corpus is used which is collected from tourism sector. The corpus is based on description of different tourist places, present in different states of India. The dataset consists of 1,000 lines. Each line starts with a line header and each word in the lines is attached with its respective tag by the symbol '\'. Tag

defines how a particular word is used in the sentence, i.e., either it is used as noun or proper noun or verb etc. The below example describes about the structure of a line from the dataset.

- E.g., htd9001 ମେରିନ\N_NNP ରାଷ୍ଟ୍ରୀୟ\JJ ଉଦ୍ୟାନ\N_NN ଆଣ୍ଡାମାନ\N_NNP ଜିଲ୍ଲାରେ\N_NN ଅବସ୍ଥିତ\JJ ।\RD_PUNC

5.2 Features considered in the proposed model

- Context word feature: the term context represents the immediate linguistic environment of a word. However it is not always explicit. Sometime it may be hidden within the neighbouring words of the keyword that are used in the same piece of text. Two types of context are there. Such as: local context and topical context. Local context represents the surrounding members of the key word, i.e., one or two immediately before and after words of the key word. Topical context represents the topic of the text where the key word is utilised. We have considered local context in our research work. For example: If a key word has Shrijukta, Shriman, Shrimati, Kumari as its previous word then the key word will be a NE of person name. Similarly if the key word has Sahara, Nagara, Grama as it is after word then it will be a location name.
- Word suffix: word suffix refers to the end letters of a word up to a fixed length, i.e., the last two or last three letters of the word. The feature of a fixed length word suffix can be applied to the current word or to any surrounding word. This feature is considered as one of the most powerful and helpful approach for identifying a NE. For example: If a key word has Natha, Kanta as suffix then that word is a NE of person type. If a key word has Gada, Pur as suffix then that word is a NE of location type.
- Word prefix: prefix information of a word is also needed for NE detection. Prefix refers to the starting letter of a key word up to some fixed length like starting two or three letters of the key word. A fixed length word prefix feature can be applied to the current word or to the surrounding words of the keyword.
- Part of speech (POS) tag: the POS tag of a key word or its surrounding words can also be considered as a useful feature for identifying NE.

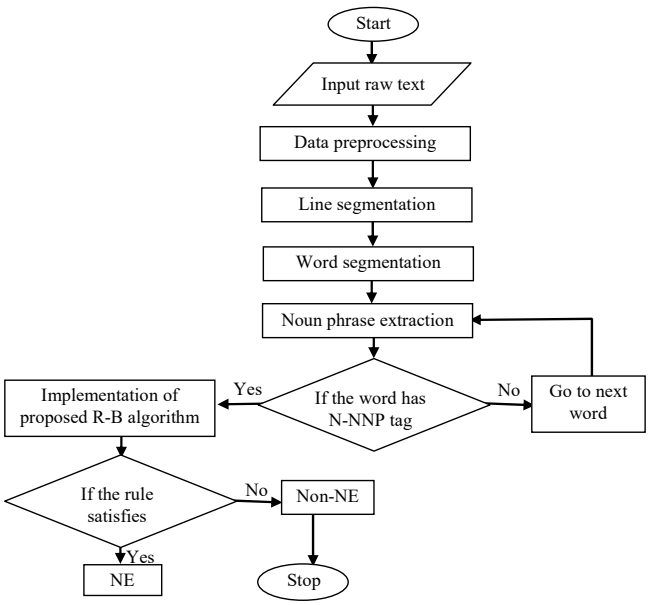
E.g., in most of the cases a word which is tagged as NNP is a named entity apart from cases like:

- ମୌଲିଙ୍ଗ ରାଷ୍ଟ୍ରୀୟ ଉଦ୍ୟାନ ଅରୁଣାଞ୍ଚଳପ୍ରଦେଶର ପୂର୍ବସିୟାଙ୍ଗ ଜିଲ୍ଲାରେ ଅବସ୍ଥିତ ।

In the above sentence ମୌଲିଙ୍ଗ is tagged as NNP but it is not a NE. The proposed rule-based NER system for Odia language is implemented in this work by using python. The platform used for the implementation of the language is Natural Language Tool Kit (NLTK). NLTK is an open source library or platform for developing Python programs to work with human language data for applying NLP. Figure 1 shows the flow diagram of the proposed system.

Figure 1 shows the flowchart for the proposed system. In the first step of the flow chart, raw text data is given as input. Here the input raw text data is a tagged Odia dataset. Next step represents data pre-processing where line segmentation is done first followed by word segmentation. After the word segmentation process, noun phrase extraction has been carried out. In the next step a single word is considered at a time to check whether the word has an N-NNP tag or not. N-NNP tag represents that the associated word is used as a proper noun. Generally proper nouns are name specific like people, places, things etc. If the word does not have an N-NNP tag then, go to the next word and again check for the N-NNP tag. If the word has N-NNP tag then implement the proposed rule-based algorithm, which results in identifying whether the word is a named entity or not.

Figure 1 Flow diagram of rule-based approach for NER



6 Experimental result

The proposed rule-based approach has been applied on a dataset which is based on tourism and it consists of 1,000 lines. In order to calculate the performance of the proposed system, confusion matrix has been considered.

- true positive (TP): observation is true and predicted as true
- true negative (TN): observation is false and predicted as false.
- false positive (FP): observation is false and predicted as true.
- false negative (FN): observation is true and predicted as false.

- class 0: non-N-NNP tagged word
- class 1: N-NNP tagged word.

Accuracy of the proposed system is calculated by considering the words present in the dataset with their repetition and without their repetition (Kaur and Gupta, 2010). Hence two different confusion matrices and two different accuracies are obtained (Guo et al., 2019; Parai et al., 2009; Das et al., 2020). Performance of the proposed system is calculated in terms of precision, recall, F1-score and accuracy. Equations for calculation of these measures are given below:

- *Precision:* precision is calculated as the ratio of total number of correctly classified true values to the total number of predicted true values. High precision points to low FP. The formula for precision is given in equation (1).

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

- *Recall:* recall is calculated by dividing the total number accurately classified true values with the total number of true values. High recall points to low FN. Equation (2) shows the formula for recall.

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

- *F1-score:* F1-score is calculated by using Harmonic mean in place of Arithmetic mean. The formula for F1-score is shown in equation (3).

$$F1\text{-}score = \frac{2 * Recall * Precision}{Recall + Precsion} \tag{3}$$

- *Accuracy:* below equation represents the formula for calculating accuracy.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

As discussed in Table 3, the proposed system got an accuracy of 83% with word repetition and 71% of accuracy without word repetition. Word repetition represents considering multiple occurrence of same word in different places in the dataset. Without word repetition represents considering only one occurrence of any word in the whole dataset. From this research work it can be concluded that, if only one occurrence of any word is considered then the proposed system gives an average performance. Whereas if any word considered with its multiple occurrences (as many times it is being used in different places in the dataset) then, the proposed system gives better performance. Hence, word repetition can be used as a key to enhance the performance of any rule-based system.

Table 1 Confusion matrix with word repetition

n = 942	*Predicted false*	*Predicted true*
Actual false	TN = 1,526	FP = 210
Actual true	FN = 169	TP = 303

Table 2 Confusion matrix without word repetition

n = 942	*Predicted false*	*Predicted true*
Actual false	TN = 625	FP = 210
Actual true	FN = 169	TP = 303

Table 3 Performance measures with word repetition and without word repetition

Performance measures with word repetition			
Precision	*Recall*	*F1-score*	*Accuracy*
0.59	0.64	0.61	0.83
Performance measures without word repetition			
Precision	*Recall*	*F1-score*	*Accuracy*
0.59	0.64	0.61	0.71

7 Conclusions and future directions

In this study, a rule-based NER system for Odia language has been presented. The proposed system is implemented on tourism dataset and got accuracy of 83% and 71% by considering words with their repetition and without their repetition respectively. In this system, different features like context word feature, word suffix, word prefix and POS tag are considered. Precision, recall and F1-score have been considered to measure the efficacy of the system. Although rule-based approach gives promising results, but it has one major challenge. It is highly domain dependant. If the research domain is changed then changes has to be made in the system structure.

In future research, more features can be added to rule-based system in order to enhance the system performance. This system can also be implemented in other domain as well as in other languages that are similar to Odia language like Bengali and Asami. Moreover, different high level rules or mechanism can also be incorporated in order to reduce the ambiguity between the named entities present in Odia language. A hybrid model can also be developed by combining the proposed rule-based system with any machine learning approaches.

References

Abdallah, S., Shaalan, K. and Shoaib, M. (2012) 'Integrating rule-based system with classification for Arabic named entity recognition', *International Conference on Intelligent Text Processing and Computational Linguistics*, Springer, pp.311–322.

Ahmed, I. and Sathyaraj, R. (2015) 'Named entity recognition by using maximum entropy', *International Journal of Database Theory and Application*, Vol. 8, No. 2, pp.43–50.

Anandika, A. and Mishra, S.P. (2019) 'A study on machine learning approaches for named entity recognition', *International Conference on Applied Machine Learning (ICAML)*, pp.153–159.

Bajwa, K.S. and Kaur, A. (2015) 'Hybrid approach for named entity recognition', *International Journal of Computer Applications*, Vol. 118, No. 1, p.3641.

Balabantaray, R.C., Lenka, S.K. and Sahoo, D. (2013) 'Name entity recognizer for Odia using conditional random fields', *Indian Journal of Science and Technology*, Vol. 6, No. 4, pp.4290–4293.

Biswas, S., Mishra, S.P., Acharya, S. and Mohanty, S. (2010) 'A hybrid Oriya name entity recognizer: harnessing the power of rule', *International Journal of Artificial Intelligence and Expert Systems (IJEA)*, Vol. 1, No. 1, pp.1–6, ISSN: 2180-1282.

Biswas, S., Mohanty, S. and Mishra, S.P. (2009) 'A hybrid Oriya named entity recognition system: integrating HMM with MaxEnt', *Second International Conference on Emerging Trends in Engineering & Technology*, IEEE, pp.639–643.

Boros, T., Dumitrescu, S.D. and Pipa, S. (2017) 'Fast and accurate decision trees for natural language processing tasks', *Proceedings of Recent Advances in Natural Language Processing*, pp.103–110.

Chopra, D., Joshi, N. and Mathur, I. (2016) 'Named entity recognition in Hindi using hidden Markov model', *IEEE Second International Conference on Computational Intelligence Communication Technology*, pp.581–586.

Chopra, D., Joshi, N. and Mathur, I. (2016) 'Named entity recognition in Hindi using hidden Markov model', *IEEE Second International Conference on Computational Intelligence Communication Technology*, pp.581–586.

Das, D., Singh, M., Mohanty, S.S. and Chakravarty, S. (2020) 'Leaf disease detection using support vector machine', in *2020 International Conference on Communication and Signal Processing (ICCSP)*, IEEE, July, pp.1036–1040.

Dey, A., Paul, A. and Purkayastha, B. (2014) 'Named entity recognition for Nepali language: a semi hybrid approach', *International Journal of Engineering and Innovative Technology*, Vol. 3, No. 8, pp.21–25.

Dhariya, O., Malviya, S. and Tiwary, U.S. (2017) 'A hybrid approach for Hindi- English machine translation', *International Conference on Information Networking (ICOIN)*, IEEE, pp.389–399.

Gali, K., Sharma, H., Vaidya, A., Shisthla, P. and Sharma, D.M. (2008) 'Aggregating machine learning and rule-based heuristics for named entity recognition', *IJCNLP-08. Workshop on NER for South and South East Asian Languages*, pp.25–32.

Guo, J., Han, Y. and Ke, Y. (2019) 'A neural-based re-ranking model for Chinese named entity recognition', *International Journal of Reasoning-based Intelligent Systems (IJRIS)*, Vol. 11, No. 3, pp.265–272.

Kaur, D. and Gupta, V. (2010) 'A survey of named entity recognition in English and other Indian languages', *International Journal of Computer Science Issues (IJCSI)*, Vol. 7, No. 6, pp.239–245.

Mansouri, A., Affendey, L.S. and Mamat, A. (2008) 'Named entity recognition approaches', *International Journal of Computer Science and Network Security*, Vol. 8, No. 2, pp.339–344.

Mathur, S. and Saxena, V.P. (2014) 'Hybrid approach to English-Hindi name entity Transliteration', *Proceedings of IEEE Students Conference on Electrical, Electronics and Computer Science*, pp.1–5.

Parai, G.K., Tenneti, T., Borah, P.K., Shah, S. and Sanyal, S. (2009) 'Document summarisation using combination and reduction of extracted sentences', *International Journal of Reasoning-based Intelligent Systems*, Vol. 1, Nos. 3–4, pp.191–199.

Petasis, G., Vichot, F., Wolinski, F., Paliouras, G., Karkaletsis, V. and Spyropoulos, C.D. (2012) 'Using machine learning to maintain rule-based named-entity recognition and classification systems', *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, pp.426–433.

Shah, H., Bhandari, P., Mistry, K., Thakor, S., Patel, M. and Ahir, K. (2016) 'Study of named entity recognition on indian languages', *International Journal of Information Science and Techniques (IJIST)*, Vol. 6, Nos. 1–2, pp.11–25.

Swain, D. and Pati, C. (2013) 'Named entity disambiguation in Odia', *International Journal on Advanced Computer Theory and Engineering (IJACTE)*, Vol. 2, No. 4, pp.137–143.

Wakao, T., Gaizauskas, R. and Wilks, Y. (1996) 'Evaluation of an algorithm for recognition and classification of proper names', *Proceedings of COLING-96*.

Wang, H., Zhao, T., Tan, H. and Zhang, S. (2008) 'Biomedical named entity recognition based on classifiers ensemble', *International Journal of Computer Science and Applications*, Vol. 5, pp.1–11.