A Proposed Model for Ontology based Development of Sanskrit Named Entity Recognition

Vini Gujarati
Temporary Assistant Professor
Veer Narmad South Gujarat University
Surat

Veena Jokhakar, PhD
Assistant Professor
Veer Narmad South Gujarat University
Surat

ABSTRACT
Named Entity Recognition (NER) is the task of identifying the entities in the text document and categorize them into pre-defined categories such as Person, Location, Organization, etc. It is an important step in the processing of natural text. This paper propose an NER system for Sanskrit language using ontology. In contrast to modern languages, the Sanskrit language has rich morphology, complex compounds and vast use of epithets (descriptive titles, alternative names) which makes entity identification more difficult. To address this problem, we proposed a Model that combines linguistic preprocessing and ontology-aware entity linking to ensure robust recognition of relationship between NEs in Sanskrit text.

General Terms
Information Extraction, Natural Language Processing, Artificial Intelligence, Named Entity Recognition, Ontology

Keywords
Entity Identification, Entity Chunking, Machine Translation, Information Retrieval, Knowledge Graph, Text Summarization

1. INTRODUCTION
In this technological era, the information content over the internet is rapidly increasing. When a user is requesting for information from the vast collection of data from the internet, the response from the internet is presented in an unstructured format. The form of unstructured data may be text, images, audio-video clips. It is difficult for humans to refer all these data and find suitable data. Information Extraction (IE) helps to extract the required data from huge unstructured collection of information. In artificial intelligence, information extraction (IE) is a branch that transforms natural language text into a structured form to perform information processing.

Sanskrit is an ancient and classical language of India in which ever first book of the world Rigveda was compiled. Panini (500 B.C.) was a great landmark in the development of Sanskrit language. He, established about 10 grammar schools prevalent during his time, penned the master book of grammar named Ashtadhyayi which served as inspiration for the later period. Literary Sanskrit and spoken Sanskrit both follow Panini's system of language. It is believed that Sanskrit is a member of the Indo-Aryan or Indo-Germanic language family, which also contains Latin, Greek, and other similar languages. When William Jones, who had previously studied Greek and Latin, encountered Sanskrit, he said that it was more flawless than Greek, more abundant than Latin, and more sophisticated than both. It is significant that despite being ancient and classical, scholars in India and other parts of the world, such as America and Germany, continue to employ Sanskrit as a vehicle of expression. The eighth schedule of the Indian Constitution lists Sanskrit as one of the contemporary Indian languages.

1.1 Natural Language Processing:
Natural Language Processing (NLP) is a field that combines computer science, artificial intelligence, and language studies. It allows computers to understand, interpret, and generate human language, both written and spoken. It uses machine learning and computational linguistics to process and analyze human language data, enabling applications like chatbots, voice assistants, search engines, and translation tools. NLP plays a crucial role in efficiently and comprehensively analyzing text and speech data. This is where NLP performs its phase as a subset of artificial intelligence (AI) to construct system that can recognize the language. NLP can be used in chatbots, text classification, sentiment analysis, machine translation, virtual assistants, speech recognition, text summarization, named entity recognition , and question answering.

1.2 Named Entity Recognition(NER):
Named entity recognition (NER) is a subtask of information extraction (IE). IE apply Natural Language Processing (NLP) for structured information retrieval. Named entity recognition is the process and the method to explore the named entities that are proper nouns and classifying them into pre-defined categories. Named entity recognition is also known as entity identification, entity chucking. Identification of named entities is playing a crucial role in NLP application such as machine translation, Information retrieval, question-answering system, automatic text summarization. NER or entity identification system identifies named entities such as name of person, location, organization, number, date, etc. shown in figure 1. Different approaches are taken for NER in various languages like Hindi, Urdu, Marathi, Punjabi, Tamil, Kannada etc. These approaches are rule-based, machine learning (HMM, MEMM, CRF, SVM) and hybrid approach.

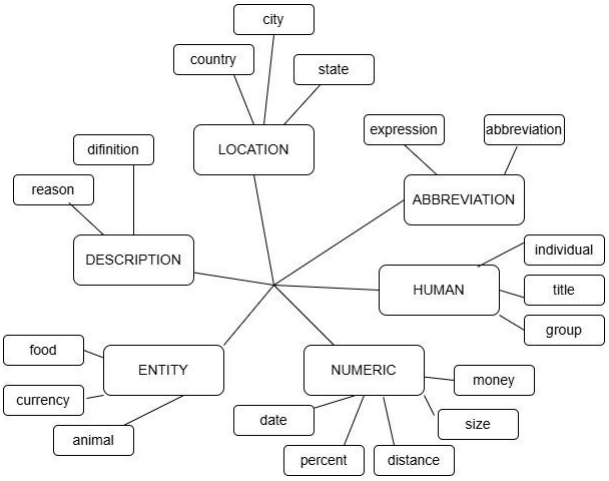

Figure 1. A single Named Entity split into more specific Named Entities [1]

1.3 Ontology

An ontology is a structured framework that defines explicit representation of knowledge within a specific domain, including concepts, properties, and relationships between them. It functions as a logical structure for data, enabling artificial intelligence (AI) systems to understand and share information effectively. Ontologies are used to build machine readable knowledge bases and natural language processing (NLP) for automating reasoning and decision-making.

1.4 Knowledge Graph

A knowledge graph is a structural representation of a network that organizes information about real world entities such as people, places, or concepts and their interrelationships. It uses nodes to represent entities and edges to represent the relationships between them and creates a model that both humans and machines can understand. Knowledge graphs allow for contextual understanding of data, powering intelligent applications and enhancing search engines and virtual assistants.

2. LITERATURE REVIEW

In[1], author has present a survey for different NERC techniques for different Indian languages : Hindi, Marathi, Bengali, Punjabi, Malayalam, Kannada, Telugu, Tamil, Urdu, Oriya. Various techniques are discussed to implement NER such as Rule based approach, machine learning approach (HMM, MEMM, CRF, SVM).

In [2], Survey represents that NER can be implemented using Rule-based or ML based approach. Rule-based use rules written by experts, and ML methods are based on arithmetical models to identify Named Entity(NE) in given text. HMM, CRF, Artificial Neural Network(ANN), Support Vector Machine(SVM), Fuzzy Support Vector Machine(F-SVM) techniques are used to recognize named entity. Survey depicts that 'NER in Malayalam through F-SVM' has high accuracy which is 92.8%, through ANN 88%, SVM 84.248%, CRF 74%, HMM 62.7% and at the conclusion that NER in Sanskrit can be accomplish using F-SVM.

In [3], author has presented a different method for named entity identification in Indian languages. Rule based approach and list lookup for Punjabi, Urdu. Hidden Markov model used for Kannada language. NERC system for Kannada language uses the unannotated text file as input, identifies the Named entities and produce an annotated text file. Conditional random field method for Telugu, Hindi, Bengali, and Tamil languages. Support vector machine, for Bengali language, Bengali news corpus is used to generate training set consists of 150k words. Maximum entropy used for Hindi and Telugu language. For implementation gazetteer list was prepared manually and semi-automatically from the dataset. Hybrid approach for the Hindi language: a combination of rule-based CRF and maximum entropy. NER for Telugu, A CRF with rule-based method.

In [4], describes various NER Tools such as Stanford NER, Lingpipe, Yamcha, Sanchay, CRF++ and Mallet. Survey for NER on various non-indian and indian languages with algorithm and F-Score : English, Swedish, Korean, Thai, Spanish, etc., and Hindi, Bengali, Telugu, Oriya, Punjabi, etc. in conclusion, rule based method and language independent rule and gazetteer list combined together generates more accurate result. For Indian languages, the combination of linguistics and statistical approach is the best combination.

In [5], the author has worked on the NER system for english and hindi languages. English language use mixes case text so that it gives clues like initial capitalized letter that indicates the availability of named entities such as name, place, etc. On the other hand, for hindi , no such hints presents, so it is difficult to recognize name entities in hindi. To overcome this problem they have stored the corresponding hindi text of english words in their database. To evaluate their NER tool, precision, recall and F-measure are used. In conclusion, NER for bilingual has been developed. Database and POS tagger are required to produce the correct output. Since Hindi POS tagger is not available to use, they can't do so much work for Hindi.

In [6], the Rule based approach used to recognize the named entity and develop the various rules to retrieve the named entities in the given input text. For analyzing the system, they have develop two data sets. one dataset for political news, articles and short stories uses 12032 tokens, another dataset for news related to science and business uses 150243 tokens.

In [7], Aim of the paper is sanskrit manuscripts are accessible to people, through translating sanskrit text into english language and to store the document in digital format. This paper provides an approach to recognize sanskrit text with machine learning translation. OCR technique plays an important role to identifies these documents. Authors proposed methodology using convolutional neural networks to identifies sanskrit characters, which provides accurate result to recognize sanskrit character.

In [8], paper presents Mahanama, a large annotated literary dataset for entity linking and named entity coreference for the Sanskrit language, which is low in resource and morphologically rich. The dataset faces challenges because of lexical variation and long range entity references.

In [9], represents the various techniques and features that are implement to identified name entity from various language. Different languages may have diverse morhologies and thus each NER procedures are not similar. Ex : An Arabic NER system can't be used for malayalam text because of different morphological features. well-known methos like rule based, machine learning, and hybrid methods are discussed with its features. Rule based has several features such as Gazetteer, Blacklist, Trigger Words. ML has common features such as Word Length, special markers, word suffixes/prefixes, Stop word, Part-of-speech (POS), etc.

In [10], an NER system based on the phonetic information included in name in the Sanskrit language. Different methods such as exhaustive search, Syntactic pattern recognition method, the rule-based method, Phonetic-based name recognizer are discussed. In phonetic-based, theoretical foundation and phonetic recognizer's design philosophy approach are used.

In [11], the paper presents the NER system for punjabi language using the hybrid approach. Combination of a rule-based approach and machine learning approach(HMM) is used to recognize named entity. Manually tagged dataset is used for creation of training and testing dataset.NER with hybrid approach is able to achieve the precision of 72.92%, recall of 76.27%, F-measure of 74.56%, and precision, recall, and F-measure of 48.27%, respectively, using HMM. In conclusion, the author observed that the proposed NER system performs better.

In [12], the paper depicts the study for NER in filipino text using support vector machine and performance is evaluated compared to an existing named entity recognizer for the same language using a rule-based approach. NER using support

vector machine gives the best result to identify the named entity class with date 95.52% f-measure and overall measure 84.97%.

In [13], the author describes the identification of named entities by using different features, gazetteer list using language dependent features and rule based approaches for telugu text. They perform the whole process in two phase, the first describe the noun identification using telugu dictionaries, noun morphological stemmers, and noun suffix, the second phase recognize the named entities using gazetteer lists related to named entity tags, various suffix features and morphological features.

In [14], the paper presents the NER system for hindi language by hybrid approach(by aggregating both rule based heuristics and hidden markov model). The paper also discusses about NER, different approaches of NER, Performance metrics and challenges such as Lack of capitalization, rich morphologies of indian languages and lack of resources. In conclusion, they have obtained accuracy of about 94.61% by using hybrid approach. If they apply only rule based method, then the accuracy obtained by this method was 47.5%, and using HMM, the accuracy was 89.78%. This shows that hybrid approach gives the very good result in a named entity recognition system.

In [15], the authors have reviewed about named entity recognition and different approaches to identifies named entities. They have identified that CRF approach is best for indian language to identify named entity. HMM is not much used to recognize named entity for indian language . If CRF is used with POS(Part of speech) and the prefix and suffix feature then the performance of NER can be improve.

In [16], the author has proposed a rule based approach to identifying kannada named entities. suffix, prefix list and proper noun dictionary of 5000 words has been prepared to recognize named entity. In kannada proper nouns are indistinguishable and this ambiguity makes Kannada NER challenging. Kannada newspaper Prajavani corpus is used to perform experiments. In conclusion, using rule based method for recognition of Kannada named entities has good precision around 86% and results can be improved by improving the gazetteer list with a proper noun dictionary, prefix, and suffix.

In [17], the author deploys an annotation tool for the Sanskrit NER dataset. The corpus they have selected was Srimad-Bhagavatam. The suggestions are pre-annotated. The heuristic works based on string-matching algorithms. The key idea is to compare Sanskrit text with an English translation and identify words of similar forms, likely named entities (NE). The gazetteer is applied to identify the NEs.

In [18], paper presents a survey on machine translator for sanskrit to english language. machine translation is a part of Natural Language Processing/Computational Linguistics (NLP/CL), which deals with understanding, developing, and computing theories for natural or human languages. The author has discusses different techniques like rule based, corpus based, statistical and interlingual machine translation approach.

In [19], paper presents NER system for Bengali, English, Hindi, Marathi, Punjabi, Tamil and Telugu laguage. Hidden Markov Model based model has been used to developed the system. The system has been trained and tested on NLP TOOLS CONTEST : ICON 2013 dataset. system has obtained F-measure value for bengali-0.8559, English-0.7704, Hindi-0.7520, Marathi-0.4289, Punjabi-0.5455, Tamil-0.4466 and Telugu-0.4003

In [20], an NER system in the Malayalam language using a neural network. Neural networks are impressive tools to study the representation of data with many levels of abstraction. The system uses different features such as POS information of the word, embedded representation of words and suffixes, etc. They have used a corpus of 20615 sentences for training and testing.

The literature review performed above in section 2 discusses the various approaches for NER in Sanskrit, such as rule-based, machine learning (HMM, MEMM, SVM, CRF, F-SVM, and ANN,) and hybrid approaches, but very minimal research has been performed on ontology based Sanskrit NER depicted with knowledge graph.

It has been surveyed that Named Entity Recognition (NER) for different Indian languages and non-Indian languages has been done, but Named Entity Recognition (NER) in Sanskrit has a critical challenge to recognize and classify entities and the relationships between the entities. In contrast to modern languages, the Sanskrit language has rich morphology, complex compounds and sandhi rules, and vast use of epithets (descriptive titles, alternative names), which makes entity identification more difficult. A single proper noun may show up in multiple forms or be substituted by synonymous epithets, for example,

Arjuna may be referred to as *Pārtha, Kaunteya,* or *Dhanañjaya*

Additionally, ambiguity come when a term use as a common noun and a proper noun, for example,

Common noun VS proper noun (e.g., ‘Suraj’ (Person name, sun))

Common noun VS proper noun (e.g., ‘*Gaṅgā*’ (Female Person name, Holy river))

These complication makes conventional NER model difficult to recognize and categorize entities and the relationships between the entities in Sanskrit text.

3. PROPOSED MODEL

The proposed model for NER will use ontology over the Sanskrit corpus and present it in the form of a knowledge graph. Figure (2) shows the whole workflow of the methodology.

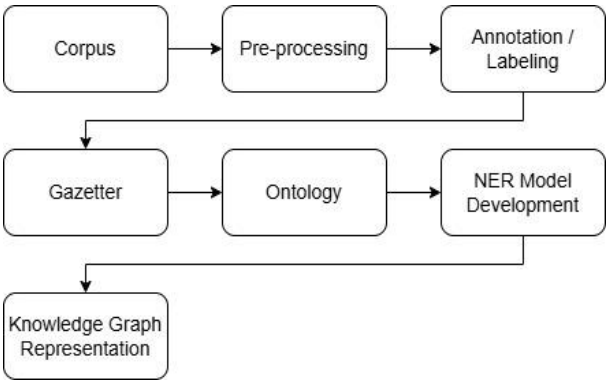

Figure 2. Workflow of Proposed Model

1. Corpus Collection: Collection of Sanskrit texts from Ayodhya Kanda a Classical literature, Mahabharata, Bhagavadgita,. These data are in the raw form in Sanskrit language. These data provides a vast distribution of named entities such as persons, places, organizations, etc.

2. Pre-processing: Normalization will clean the text for further text processing, Sentence segmentation & tokenization will identify text boundaries and split text into tokens., Sandhi

Splitting will bring back original word boundaries, Samāsa Analysis ,Morphological Analysis will identify the root form and provide entity classification.

3. Annotation / Labeling:

- Manually annotate dataset with entity tags such as PERSON, PLACE, ORGANIZATION, DATE, etc.
- Build a standard annotated corpus that validate corpus for Sanskrit NER.

4. Gazetteer & Ontology**:**

- Prepare gazetteer lists (deities, places, rivers, dynasties, characters) that helps to enhance entity recognition and reduce ambiguity.
- Handle epithets (Arjuna = Pārtha, Kaunteya, Dhanañjaya) that map all epithets into a single entity ID to maintain entity consistency.
- Link entities by mapping to classes and connect them with relevant relations(e.g. son of, king of,)

5. NER Model Development: Building and Training NER model using Classical ML or transformers.

6. Knowledge Graph construction: Structural representation of entities and relationship between entities.

4. CONCLUSION

Unlike other modern languages, Sanskirt has rich morphology, sandhi rules and vast use of epithets and alternative names, making a task of NER more complex. This paper concludes with a proposed model that combines linguistic preprocessing (sandhi and samasa handling), epithet handling and ontology based entity linking to ensure robust recognition of relationship between NEs in Sanskrit text and representing the entities and relationships between them using knowledge graph construction. Future work includes enlarging dataset, enhance ontology to add other domains such as mythological roles, etc and development of application like Sanskrit QA system.

5. REFERENCES

[1] Kale, S., & Govilkar, S. (2017). Survey of named entity recognition techniques for various indian regional languages. *International Journal of Computer Applications*, *164*(4), 37-43.

[2] Joseph Mickole James, Sangeetha Jamal : Named Entity Recognition in Sanskrit : A Survey.

[3] Prakash Hiremath, S. B. (2014). Approaches to named entity recognition in indian languages: A study. *International Journal of Engineering and Advanced Technology (IJEAT), ISSN*, *2249*, 8958.

[4] Patil, N., Patil, A. S., & Pawar, B. V. (2016). Survey of named entity recognition systems with respect to Indian and foreign languages. *International Journal of Computer Applications*, *134*(16).

[5] Nanda, M. (2014). The named entity recognizer framework. *International Journal of Innovative Research in Advanced Engineering*, *1*(4), 104-108.

[6] Singh, U., Goyal, V., & Lehal, G. S. (2012, December). Named entity recognition system for Urdu. In *Proceedings of COLING 2012* (pp. 2507-2518).

[7] Kulkarni, I., Tikkal, S., Chaware, S., Kharate, P., & Pandit, A. (2022, February). Proposed Design to Recognize Ancient Sanskrit Manuscripts with Translation Using Machine Learning. In *Proceedings of the International Conference on Innovative Computing & Communication (ICICC)*.

[8] Anonymous (2024). Mahanama : An Epic Literary Dataset For Entity Linking and Named Entity Coreference

[9] Syafiq, M. I., Talib, M. S., Salim, N., Haron, H., & Alwee, R. (2019, August). A concise review of named entity recognition system: Methods and features. In *IOP Conference Series: Materials Science and Engineering* (Vol. 551, No. 1, p. 012052). IOP Publishing.

[10] Srivastava, A., & Rajaraman, V. (2002). Computer recognition of Sanskrit-based Indian names. *IEEE transactions on systems, man, and cybernetics*, *21*(1), 287-290.

[11] Bajwa, K. S., & Kaur, A. (2015). Hybrid approach for named entity recognition. *International Journal of Computer Applications*, *118*(1).

[12] Castillo, J. M., Mateo, M. A. L., Paras, A. D., Sagum, R. A., & Santos, V. D. F. (2013). Named entity recognition using support vector machine for Filipino text documents. *International Journal of Future Computer and Communication*, *2*(5), 530.

[13] Sasidhar, B., Yohan, P. M., Babu, A. V., & Govardhan, A. (2011). Named entity recognition in telugu language using language dependent features and rule based approach. *International Journal of Computer Applications*, *22*(8), 30-34.

[14] Chopra, D., Jahan, N., & Morwal, S. (2012). Hindi named entity recognition by aggregating rule based heuristics and hidden markov model. *International Journal of Information*, *2*(6), 43-52.

[15] Shah, H., Bhandari, P., Mistry, K., Thakor, S., Patel, M., & Ahir, K. (2016). Study of named entity recognition for indian languages. *Int. J. Inf*, *6*(1), 11-25.

[16] Melinamath, B. C. (2014). Rule based methodology for recognition of Kannada named entities. *Int. J. Latest Trends Eng. Technol.(IJLTET)*, *3*, 50-58.

[17] Sujoy, S., Krishna, A., & Goyal, P. (2023, January). Pre-annotation based approach for development of a Sanskrit named entity recognition dataset. In *Proceedings of the Computational Sanskrit & Digital Humanities: Selected papers presented at the 18th World Sanskrit Conference* (pp. 59-70).

[18] V N Jokhakar, A survey on machine translation for Sanskrit Language to English Language,IDEES - International Multidisciplinary Research Journal 6 (2), 184-191.

[19] Gayen, V., & Sarkar, K. (2014). An HMM based named entity recognition system for indian languages: the JU system at ICON 2013. *arXiv preprint arXiv:1405.7397*.

[20] Ajees, A. P., & Idicula, S. M. (2018). A named entity recognition system for Malayalam using neural networks. *Procedia computer science*, *143*, 962-969.