Stochastic Modelling and Computational Sciences

COMPARATIVE ANALYSIS OF NAMED ENTITY RECOGNITION IN INDIAN LANGUAGE TEXTS: MONOLINGUAL AND CODE-MIXED LANGUAGES

Anasooya Dutta* **and Sharvari Govilkar**
Department of Computer Engineering, University of Mumbai, PCE, New Panvel, India
anasooya21me@student.mes.ac.in and sgovilkar@mes.ac.in

ABSTRACT
Following the start of the internet, we can see the rise in the use of many social media platforms where people propagate their thoughts in their own language.Also, we can see mixing of two different languages while conveying messages on the web.Thus, making it difficult to perform name entity tasks. This is one of the main challenges of name entity extraction. The complexity of linguistic mixing in code-mixed text poses an additional challenge for Named Entity Recognition.Employing NER techniques can assist in determining the linguistic origins of code-mixed text. In social media posts and comments, code mixing is a common occurrence where individuals use both their regional language and English, which further complicates identifying the language base.The aim of named entity recognition (ner) is to identify and label the words such as person, place,location,etc.This survey paper discusses and compares algorithms used by past researchers on name entity extraction for pure and code mixed Indian languages by considering the most relevant tools, and methods like Convolutional Neural Networks (CNNs), Bidirectional Long Short Term Memory, Transfer learning approaches, and numerous other methods being used lately.

Keywords: NLP, Entity extraction, techniques and methods, challenges, lexical resources, features, machine learning, lexicon based.

1. INTRODUCTION
Natural language processing(NLP) is the indispensable task of data in the present scenario because it considers many sources to get the information, then helps translate human language and thus perform useful tasks.Named entity recognition (NER) is one of the sub-task of NLP.NER helps to recognise and identify the name entities from unstructured data.In recent years, people have started using social media platform to express their concern ,opinions and thoughts.Most of the times, people tend to type in their regional language or mixing two different language which is known as code-mixing thus hinders NER to perform its task.The linguistic variation is a big challenge for NER task.By the definition, Code mixing or code switching can be seen when Grammatical features and lexical items occurs in one sentence.Besides this, there are other challenges like capitalisation,lot of use of abbreviations, emojis,etc.This information helps to form the meaning of the sentences or intention or context behind it.There are application other than what we discussed above in various domain.A few of them are information retrieval which retrieves important points from textual documents.

Past Researchers have applied various technologies in order to get accurate results, curtailing the errors.Regarding Named Entity Recognition (NER) methodologies, various techniques are available. Among the most prevalent are the Rule-Based approach, the Supervised approach, and the Unsupervised approach. This paper undertakes an in-depth analysis of multiple research papers to comprehensively delineate the process of name entity extraction and recognition, as well as to pinpoint established technologies employed for this task.

The survey article is meticulously structured as Section 2, to provide a cohesive and organized exploration of the subject matter.,literature survey of past research paper , Sect. 3, contain the background that is features for entity extraction,Lexicon resources for entity recognition and Baseline Algorithm Sect. 4 contain entity extraction algorithms and its Summary, divided into two classification of NER techniques: machine learning-based emotion detection and deep learning-based entity extraction, and a comparison of both pure and code mixed indian language text.Sect. 6, concludes the paper.

Copyrights @ Roman Science Publications Ins.

2. LITERATURE SURVEY

K.P. Pallavi, L. Sobha and M.M. Ramya[1], The utilization of gazetteers in conjunction with the Conditional Random Fields (CRF) technique has been put forth as a novel approach to devise a named-entity recognition system tailored for the Kannada language. This innovative system demonstrates its efficacy by amalgamating the strengths of gazetteers and CRF. When evaluated on a dedicated test set, the system exhibits an impressive F1-score of 90.41%, showcasing its robust performance in accurately identifying named entities within the Kannada text. This achievement underscores the potential of combining gazetteers and CRF to enhance named-entity recognition tasks in linguistically diverse contexts.

Parth Patil, Aparna B,Maithili S, Onkar L, and Raviraj J.[2], Used the L3Cube-MahaNER dataset to train and evaluate different BERT models, including BERT-base, BERT-large, and multilingual BERT, for Marathi NER.The BERT models were fine-tuned on the L3Cube-MahaNER dataset using a sequence tagging approach.

AjeesPa, Sumam Mary[3], In the context of sequence tagging-based Named Entity Recognition (NER), an innovative approach was implemented. This involved the utilization of a bidirectional long short-term memory (BiLSTM) network, a cutting-edge neural architecture renowned for its proficiency in capturing sequential dependencies. To further augment the performance of the NER system, a conditional random field (CRF) layer was integrated into the model. This CRF layer played a pivotal role in enhancing the overall coherence and structure of the predicted entity labels, resulting in improved accuracy and contextual relevance.

Furthermore, the model's input features were enriched by the inclusion of both word embeddings and character-level embeddings. This dual-input strategy was meticulously chosen to encapsulate a holistic range of linguistic attributes. Word embeddings, derived from pre-trained language representations, enabled the model to grasp intricate semantic nuances and contextual meanings embedded within the text. Meanwhile, character-level embeddings provided a granular insight into morphological aspects, enabling the system to decipher fine-grained details such as prefixes, suffixes, and word composition.

By synergistically integrating these features into the BiLSTM framework, the NER model achieved a comprehensive understanding of the underlying language structure. This enabled it to excel in recognizing and classifying named entities across diverse contexts and linguistic variations. The meticulous orchestration of BiLSTM, CRF, word embeddings, and character-level embeddings thus exemplified a robust and multifaceted approach to NER, showcasing a sophisticated interplay between various techniques to enhance performance and accuracy.

Silja C K and Dr. T K Bijimol[4], Employing a combination of a rule-based approach and a Maximum Entropy classifier, Named Entity Recognition (NER) was executed within the context of the Malayalam language. The outcome of this approach resulted in the development of a system that demonstrated remarkable performance metrics, with precision reaching 92.05%, recall achieving 90.95%, and an impressive F1-score of 91.50%.

Arti Jain, Divakar Yadav, Anuja Arora, Devendra K. Tayal[5] , encompassed a range of variants such as BERT-base, BERT-large, and the versatile multilingual BERT, all of which were meticulously employed to cater to the specific demands of Marathi Named Entity Recognition (NER).Subsequent to the selection and preparation of these BERT models, an intricate and strategic fine-tuning process was embarked upon. This optimization endeavor was carried out exclusively on the L3Cube-MahaNER dataset, meticulously engineered to facilitate the enhancement of Marathi NER performance. The fine-tuning process was executed through a sequence tagging approach, effectively harnessing the inherent contextual capabilities of the dataset to bolster the BERT models' prowess in recognizing and classifying named entities within Marathi text. This dual-pronged approach, incorporating the prowess of BERT models and the sequence tagging methodology, underscores the meticulous and multifaceted nature of the research, geared towards advancing the field of Marathi NER.

Vinay Singh, Deepanshu Vijay, Syed S. Akhtar, Manish Shrivastava[6],Employed were Decision tree, Long Short-Term Memory (LSTM), and Conditional Random Field (CRF) methodologies. The researchers designed a

set of experiments involving diverse machine learning classification algorithms, each integrated with word, character, and lexical features for comprehensive analysis.

Kushagra Singh,Indira Sen,Ponnurangam Kumaraguru[7], Utilized in the study were both Conditional Random Fields (CRF) and Long Short-Term Memory Recurrent Neural Networks (LSTM RNNs). The focus of the research centered around the creation of a token-level language identification system specifically tailored for the context of Hindi-English (Hi-En) code-mixed tweets. This intricate task involved a meticulous amalgamation of CRF and LSTM RNNs to devise a robust framework for accurately identifying and distinguishing the languages present within the code-mixed content of the tweets.

Ruba Priyadharshini,Bharathi Raja Chakravarthi,Mani Vegupatti, John P. McCrae[8] ,Utilized FastText word embeddings extracted from Common Crawl and Wikipedia datasets, encompassing English and Hindi languages, with the latter utilizing the native Hindi-Devanagari script. Furthermore, to cater to the specific nature of NER data derived from Twitter, English Twitter GloVe word embeddings were also integrated into the approach.

Rudra Murthy, Pallab Bhattacharjee, Rahul Sharnagat,Jyotsana Khatri, Diptesh Kanojia, Pushpak Bhattacharyya [8], Various language models were harnessed to undertake the sequence labeling task, focusing on Named Entity Recognition (NER). This served to demonstrate the effectiveness of the dataset through a comprehensive comparative analysis vis-à-vis models trained on a distinct dataset tailored for the specific Hindi NER task. In order to simplify and enhance the annotation workflow, an innovative online tool was developed. This tool was rooted in the PaCMan framework and incorporated a tailored modification to the PaCMan architecture. This adaptation facilitated the seamless integration of untagged NER data, providing a more holistic approach to the NER task.

Suman Dowlagar,Radhika Mamidi [9], the foundational benchmarks employed in this study encompass Conditional Random Field (CRF), MultiCoNER baseline, and Pre-trained multilingual BERT. The neural network framework of choice is PyTorch, complemented by the utilization of two pre-trained models: the multilingual BERT model (bert-base-multilingual-cased) and the XLM-ROBERTa base model (xlm-roberta-base).

Vamshi Krishna Srirangam, Appidi Abhinav Reddy, Vinay Singh, Manish Shrivastava[10],Employed within this research were machine learning models including Conditional Random Fields (CRFs), Decision Trees, and Bidirectional LSTMs. The focus of this study was the introduction of an annotated code-mixed Telugu-English corpus, serving as a pioneering contribution in this field. Notably, this corpus marks the first of its kind. The meticulous annotation process for named entity tags within the corpus was conducted by two linguistically proficient individuals with a strong command over both Telugu and English languages.

Sumukh S, Manish Shrivastava[11],Various machine learning classification models, such as CRF, Bi-LSTM, and Bi-LSTM-CRF, were executed to a corpus with word, character, and lexical features. The primary objective was to assess the impact of each feature and model parameters through a series of experiments. These experiments involved intertwining diverse sets of features together and also aggregating all features simultaneously. Furthermore, the model parameters, such as decision tree criterion ('Information gain' and 'gini') and maximum depth of the tree, were transmuted for the decision tree model. For the CRF model, regularization parameters ('L2 regularization') and optimization algorithms ('Avg. Perceptron' and 'Passive Aggressive') were assorted. The aim of these experiments was to comprehend the effects of each feature and model parameter on the overall performance of the classification models.

3. BACKGROUND

A. Features for Entity Extraction

Name Entity Recognition depends on vivid features to recognize and categorize entities in a text.These features gives important information to the NER model for prediction of the text.Some of the important features are word-based features, Contextual features, Lexical features, Rule-based features, and Sequence labeling features.

Word based features focuses on individual words in a text which include word embedding distributed representations of words that encode semantic and syntactic information, part-of-speech (POS) tags which Labels assigned the words based on their grammatical categories, and Capitalization which checks whether a word starts with an uppercase or lowercase letter.Lexical features which recognise linguistic patterns and information from external resources like Gazetteers,Ontologies and knowledge bases, and Word dictionaries.

Sequence labeling features take the sequential nature of text and relationship between adjacent words into consideration such as N-grams and window based features.

As there is more use of deep learning and contextual word embeddings, features like word embeddings and contextual embeddings have become more prevalent.

B. Lexicon Resources for Entity Recognition

Lexicon resources play a crucial role in Name Entity Recognition (NER) by providing lists of words or phrases that correspond to specific named entity categories like Gazetteers,Ontologies and Knowledge Bases, Domain-Specific Word Lists, WordNet and Word Embeddings Clusters.

Gazetteers are collections of words or phrases linked with specific named entity types. Ontologies and knowledge bases, such as DBpedia, Wikidata, or Freebase, provide well curated information about entities and their relationships.Word embedding models can be used to merge similar words together. By merging words based on their distributional similarity, these clusters can be treated as lexicon resources..When we mix lexicon resources with other techniques, such as machine learning models and rule based, we can see improvement in accuracy and overall coverage of name entity recognition.

C. Baseline Algorithm

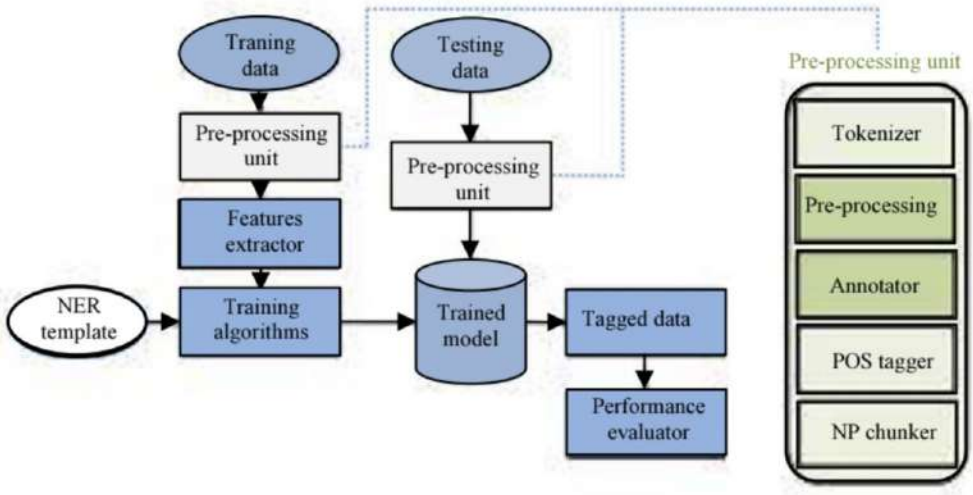

Fig 1: Architecture for kannada NER system(K.P. Pallavi et al. (2018))

Baseline algorithm for name entity analysis is a combination of tokenization, feature extraction using classifiers such as SVM, Naive Base, etc. A common baseline algorithm for Named Entity Recognition (NER) is the rule-based approach. Rule-based methods rely on hand-crafted rules or patterns to search and classify named entities in text. Here's a broad overview of the steps involved in a rule-based NER system.Preprocessing: The text data is preprocessed by tokenizing it into words or subword units (e.g., using whitespace or more advanced tokenization techniques like Byte-Pair Encoding).Post-processing and Filtering: The identified potential named entities are further filtered and refined to remove false positives. This step may involve applying additional heuristics or constraints to improve the precision of the NER system.Pattern Matching:Predefined patterns or regular

expressions to the tokenized text to identify named entities.Labeling and Categorization: Finally, the recognized named entities are labeled and categorized into predefined entity types such as person names, organization names, locations, or dates.

4. ENTITY EXTRACTION TECHNIQUES

CRF Model (Kannada [1], Hinglish [6] [7] [8] [12])
Conditional Random Fields (CRFs) represent a form of probabilistic graphical model adept at addressing named-entity recognition (NER) undertakings. Specifically within NER, CRFs serve to effectively model the conditional likelihood of a named entity sequence within the context of a given input sentence.

In a study by K.P. Pallavi et al. (2018), the CRF model was harnessed to discern and categorize named entities (NEs). These categorized NEs were subsequently stored within tagged data. Following an error analysis, a refinement process involved the adjustment of certain classes present in the input file. For experimental purposes, the researchers sourced publicly available data from the Kannada Wikipedia.

Bidirectional LSTM (Long Short-Term Memory) network (Hinglish [6] ,[10],Malayalam [3])
A bidirectional Long Short-Term Memory (LSTM) network to model the sequence of words in the input text and predict the corresponding entity labels.The authors also used word embeddings, character embeddings, and part-of-speech (POS) tags as input features to the LSTM network. The word embeddings were obtained using a pre-trained word2vec model, while the character embeddings were learned during training. Ajees Pa et al.(2018)Introduce a technique for Named Entity Recognition (NER) employing Neural Networks. The experimental setup involves the utilization of the CUSAT POS tagged corpus, wherein approximately 205,000 words from the said corpus have been meticulously annotated with named entity tags.

BERT-Base and BERT-Multilingual (Marathi [2], HINDI [9])
BERT is a bidirectional transformer-based model that learns contextualized word embeddings by considering the entire input sentence.RoBERTa is an optimized version of BERT that uses larger batch sizes and more training data to achieve better performance on NLP tasks.To use RoBERTa for NER, the model is first pre-trained on a large corpus of text using an unsupervised learning task, such as masked language modeling or next sentence prediction. During pre-training, the model learns to generate high-quality representations of words and their contexts that capture their semantic and syntactic relationships.

The research by Rudra Murthy and colleagues (2022) employed the mBERTbase−cased variant of the multilingual BERT (mBERT) model, featuring 12 layers with 768 hidden layers and a total of 110 million parameters. This variant supports 104 languages. The authors selected a subset of 9,989 sentences from the ILCI tourism domain and combined it with sentences from the news domain, resulting in a dataset encompassing a total of 108,608 sentences.

The authors conducted meticulous hyper-parameter tuning for each model, selecting the hyper-parameters that yielded the highest F-Score on the development set. In another study by Parth Patil et al., the MahaNER BERT4 model, fine-tuned on L3Cube-MahaNER, is denoted as the MahaBERT model. This model has been made publicly available on the model hub. Furthermore, the MahaRoBERTa model showcased superior performance for IOB notations, while MahaBERT demonstrated exceptional outcomes for non-IOB notations.

PURE LANGUAGES:

Table 2.1 Summary of papers and techniques used for named entity extraction

Sr no	Techniques used	Language used in	Authors	Advantages	Disadvantages
1	**Conditional random fields**	**Kannada [1]**	K.P. Pallavi et al. (2018)	This algorithm Showed robust performance even	CRFs optimize the sequence labeling task extensively

	(CRF)	**Malayalam [3]**	Ajees Pa et al. (2018)	with limited annotated data and can easily integrate external source of information such as gazetteers.	during training, they do not explicitly capture extensive context information apart from the sentence level.
		Hindi[6]	Vinay Singh et al. (2022)		
2	**Maximum entropy mode**	**Malayalam [4]**	Sriha CK et al. (2022)	Maximum Entropy models are useful for limited annotated data, as they can effectively handle sparse feature spaces and make predictions even with a small amount of labeled data and allows easy incorporation of various types of features, such as word-level, character-level, or context-based features.	Training Maximum Entropy models can be computationally overpriced,especially when dealing with a large number of features.
		Hindi[5]	Arti Jain et al. (2022)		
3	**BiLSTM**	**Hindi[6]**	Vinay Singh et al. (2022)	Can handle variable-length sequences, which is essential in NER as entities can have different lengths.	BiLSTM takes word embedding as input so when it sees a out of vocab word it cannot predict the output
		Malayalam [3]	Ajees Pa et al. (2018)		
4	**BERT**	**Hindi[6]**	Vinay Singh et al. (2022)	By fine-tuning the pre-trained BERT model on a specific NER task, the model can effectively capture the relevant information for identifying named entities in the text.	Collection and annotation for large dataset (in hindi) can be time-consuming.
		Marathi[2]	Parth et al. (2022)		

CODE MIXED LANGUAGES:

Sr no	Techniques used	Language used in	Authors	Advantages	Disadvantages
1	**Conditional random fields (CRF)**	**Hinglish[7]**	Kushagra Singh et al.(2018)	CRF's global inference capabilities are advantageous for identifying and recognizing multi-word named entities that span multiple tokens.	It depends on a fixed-size context window to capture dependencies among words. Selecting an appropriate window size is vital because a relatively small window may not capture long-range dependencies, while a large window can introduce computational challenges.
		Hinglish [8]	Ruba Priyadharshini et al. (2018)		
		Hinglish [9]	Rudra Murthy et al. (2020)		

2	**rule-based approach**	**Telugu-english[10]**	Vamshi Krishna et al. (2019)	Rule-based NER does not need a large amount of annotated data for training, which is especially beneficial when labeled data is scarce or difficult to obtain and also are defined explicitly, the system designer can choose what types of entities the system recognizes and what it ignores.	If there are errors in the rule definitions then that can lead to incorrect entity recognition, and since rules often work independently, one error in the process can propagate to subsequent steps.
		Kannada-English [11]	Sumukh S et al. (2022)		
3 3	**BERT**	**Hinglish [9]**	Rudra Murthy et al. (2020)	BERT's bidirectional functionality handles ambiguous entities effectively by taking into account both preceding and succeeding context, which can be crucial in resolving ambiguous mentions in NER.	BERT is pre-trained on various tasks, but it doesn't contain specific knowledge about named entities of a particular domain.

5. CHALLENGES IN NAMED ENTITY EXTRACTION

Challenges faced during entity extraction are Lack of meaning of the word, Contextual Ambiguity, Overlapping entities,Noisy data and Cross Lingual challenges.Due to lack of vocabulary, if the training data set does not contain a particular word then it becomes difficult for identification of that word which may lead to incorrect label.Overlapping Entities Text may contain named entities that overlap with each other, making it difficult for NER models to correctly identify and classify them.Text data text data in real-world applications can be noisy, containing misspellings, abbreviations, or incomplete information.Proper segmentation of a sentence is crucial so as to avoid duplicate identification of entities.Moreover, recognising the exact label in such scenarios can be challenging.

6. CONCLUSION

The underlying goal of this work is to explore and analyze all the approaches with their strengths and weaknesses in named entity extraction of pure and code-mixed Indian Languages. To initiate, various tiers of named entity tools were discussed, subsequently accompanied by a brief summary of significant algorithms.In pure Indian languages, BERT along with BiLSTM algorithm yielded the best results for Hindi language, BiLSTM showed best results for Marathi language.Considering code-mixed languages, BERT used for Hinglish language gave best results.

As part of the future work, integration of entity extraction with other NLP tasks can be a solution for code-mixed language,the corpus could be enhanced by additionally providing the corresponding POS tags for each token.For getting robust training dataset, collection and analysis of larger dataset is crucial so that while annotating there are less error.Introducing more mult-bert language model so that while tagging it identifies it accurately.Compared to Pure Indian language text ,less work has been done on codemixed languages,Beyond entity extraction, named entity linking, which involves connecting entity mentions to specific entries in a knowledge base, can be an interesting direction for research.Overall, entity extraction in code-mixed and Indian languages has a bright future, and further study and development in this field will be crucial to the creation of inclusive and successful NLP solutions for a variety of linguistic contexts.

ACKNOWLEDGEMENTS

Visualizing any project without the guidance of experts who have already treated this path before, is difficult. I would take this opportunity to thank my mentor and **HOD Dr. Sharvari Govilkar** and M.E. Coordinator **Dr. Prashant Nitnaware** for their guidance in this project and also for providing me all the details for the presentation. I am also grateful to our principal **Dr. Sandeep Joshi** for extending his help directly and indirectly through various channels in my project work.

REFERENCES

[1] K.P. Pallavi, L. Sobha and M.M. Ramya (2018), ‘‘Named Entity Recognition for Kannada using Gazetteers list with Conditional Random Fields”.

[2] Parth Patil,Aparna B,Maithili S, Onkar L, and Raviraj J.(2022), ‘‘L3Cube-MahaNER: A Marathi Named Entity Recognition Dataset and BERT models’’,20

[3] Ajees Pa,Sumam Mary, ‘‘A Named Entity Recognition System for Malayalam using Neural Networks’’,2018

[4] Sriha CK,Dr TK Bujimol, ‘‘Named Entity Recognition for Malayalam’’,2022

[5] Arti Jain Divakar Yadav Anuja Arora Devendra K. Tayal , ‘‘NER for hindi language using context pattern based maximum entropy.’’,2022

[6] Vinay Singh, Deepanshu Vijay, Syed S. Akhtar, Manish Shrivastava, ‘‘Named Entity Recognition for Hindi-English Code-Mixed Social Media Text.’’,2018

[7] Kushagra Singh, Indira Sen,Ponnurangam Kumaraguru , ‘‘Language Identification And name entity recognition in hinglish code mixed tweets’’,2018

[8] Ruba Priyadharshini,Bharathi Raja Chakravarthi,Mani Vegupatti,John P. McCrae, ‘‘Named Entity Recognition for Code-Mixed Indian Corpus using Meta Embedding.’’,2020

[9] Rudra Murthy, Pallab Bhattacharjee, Rahul Sharnagat,Jyotsana Khatri,Diptesh Kanojia, Pushpak Bhattacharyya, ‘‘HiNER: A Large Hindi Named Entity Recognition Dataset’’,2022

[10] Vamshi Krishna Srirangam, Appidi Abhinav Reddy, Vinay Singh, Manish Shrivastava, ‘‘Corpus Creation and Analysis for Named Entity Recognition in Telugu-English Code-Mixed Social Media Data.’’, 2019

[11] Sumukh S, Manish Shrivastava, ‘‘"Kanglish alli names!" Named Entity Recognition for Kannada-English Code-Mixed Social Media Data.’’,2022