Kannada Named Entity Recognition Using Deep Learning Techniques

Dr. Pushpalatha M[1], Dr. Hemakumar G[2], Dr. Santhosh Kuamr B N[1]

[1] Associate Professor, Department of Computer Science, Govt Women's College Hunsuru, Karnataka, India

[2] Associate Professor, Department of Computer Science, Sri Mahadeshwara Govt First grade Kollegal, Karnataka, India

DOI: https://doi.org/10.51244/IJRSI.2026.1307000434

Received: 05 August 2026; Accepted: 10 August 2026; Published: 24 August 2026

ABSTRACT

Named Entity Recognition (NER) is a natural language processing task concerned with identifying mentions of named entities and classifying them according to a predefined set of categories. Despite the success of NER in domains, where such data is abundant it remains a formidable challenge for low-resource languages such as Kannada. In this paper we discuss the possible ways to approach NER for the Kannada language.

We explore various research directions including rule-based methods statistical machine learning neural networks and transformers based tagging methodologies. We highlight the various challenges in achieving NER for such a language and propose a transformer based contextual tagging framework for labelling sequences.

We propose to use mBERT IndicBERT and XLM-RoBERTa language models pretrained on target and other related Indic language corpora and further fine-tune these models for the NER task. We discuss various aspects for experimentation including data collection labelling data preparation methods data-splits evaluation metrics comparison with other models hyper parameter tuning entity-wise analysis and error analysis.

Keywords: Kannada, Named Entity Recognition, Natural Language Processing, Low-Resource Languages, Deep Learning, Transformer, BERT, IndicBERT, XLM-RoBERTa

INTRODUCTION

Named Entity Recognition (NER) is a fundamental Natural Language Processing task concerned with the identification of spans of text that refer to entities that are assigned predefined semantic categories. Typical entity categories include Person (PER), Location (LOC), Organization (ORG), Date (DATE), and other domain-specific categories. NER is crucial for many natural language processing applications, including information retrieval, question answering, machine translation, information extraction, text summarization, digital libraries, and conversational systems.

Kannada is a Dravidian language spoken predominantly in Karnataka and by Kannada speaking communities in other regions. Despite its large number of speakers and extensive literary history, Kannada NLP is data scarce compared to languages such as English. In particular, there are few large-scale consistently annotated resources for Kannada NER. Prior research on Kannada NER has explored both traditional machine learning approaches as well as more recent neural approaches. For instance, there has been prior research on developing supervised classification models for Kannada NER; for example, a Multinomial Naïve Bayes approach was evaluated on a Kannada corpus achieving an F1-measure of 81% on a test set of 5,000 tokens from a 95,170-token training corpus. This result should not be directly compared to the present work since the proposed framework has not yet been applied to the corpus.

Recent advances in pretrained language models have fundamentally changed the nature of Named Entity Recognition research. BERT has had a major impact on NER by providing contextual representations that can be leveraged to fine-tune specific downstream NER tasks. Multilingual variants of BERT such as mBERT enable similar advantages for other languages. The XLM-RoBERTa (XLM-R) model was pretrained on 100 languages and has also shown strong performance on cross-lingual NER tasks. Additionally, the IndicNLP ecosystem provides various monolingual resources including data, embeddings, and pretrained language models for the 11 major Indian languages. Thus it is of interest to explore BERT-based solutions for Kannada NER.

With this background, the objectives of this paper are two-fold:

1. To provide a comprehensive review of prior research on Kannada NER, covering both rule-based traditional machine learning and deep learning-based approaches.

2. To propose a transformer-based Kannada NER framework and experimental protocol for evaluating such an approach in detail.

LITERATURE REVIEW

Rule-Based and Gazetteer-Based Approaches : Early NER systems were based on manually crafted linguistic rules lexical knowledge and gazetteers. Gazetteers are essentially lists of entity names such as people places organizations and countries. Such rule-based systems can achieve good performance on well-defined entity classes but require extensive linguistic knowledge and engineering effort. Rule-based approaches to Kannada NER face similar challenges due to the linguistic complexity of the language. Additionally, morphological variations can lead to same entity names appearing in different surface forms.

Statistical Machine Learning Approaches : Other traditional statistical approaches to NER involve training machine learning models on annotated corpora. Hidden Markov Models maximum entropy models conditional random fields (CRFs) and support vector machines have been applied to NER with varying degrees of success. In particular CRFs have been effective sequence labelling architectures for NER. However most traditional machine learning approaches to NER require careful engineering of lexical orthographic and linguistic features. One early work on Kannada NER used a Multinomial Naïve Bayes approach with TF-IDF features achieving an F1-measure of 81% on the test set as discussed earlier. It must be emphasized that such comparisons should be made only after controlling for variations in the data the entity sets the evaluation methodology and so on.

Neural Network Approaches : More recent advancements in neural networks have led to new breakthroughs in NER. Lample et al proposed an effective approach for NER using bidirectional LSTMs with CRF layer for sequence labeling making use of character-level features. Similarly Ma and Hovy proposed a sequence labeling architecture that makes use of BiLSTM character-level CNNs and a CRF layer. Both works demonstrated the efficacy of contextual representations for NER. For low-resource languages such as Kannada character-level features can be particularly useful in capturing morphological variations.

Transformer-Based Approaches : Self-attention based transformer encoder representations have enabled new levels of performance for a variety of NLP tasks including NER. BERT introduced contextualized embedding representations for English and subsequently multilingual BERT (m-BERT) extended this idea to 104 languages enabling similar performance gains across languages. Thus the m-BERT model can be used for NER in Kannada. However it must be noted that the performance on low-resource languages may be limited due to the data imbalance during training. Another multilingual model is the XLM-RoBERTa (XLM-R) model, which was pre-trained on 100 languages and has also shown strong performance on cross-lingual NER tasks. Thus it is of interest to explore these representations for Kannada NER. Similar to m-BERT the performance of XLM-R on low-resource languages such as Kannada must be carefully evaluated.

Indic-Language Resources : The IndicNLP ecosystem has provided various resources towards building models and tools for Indian languages. In particular it provides monolingual corpora word embedding pre-trained language models and evaluation benchmarks for the 11 major Indian languages including Kannada. Such resources are essential for developing high-performance NER models for Kannada. In particular pre-trained language models reduce the amount of task-specific training data needed. On the other hand it must be noted that a reliable annotated corpus and evaluation protocol are also needed for training and assessing a Kannada NER system.

CHALLENGES IN KANNADA NER

There are many Linguistic and Computational issues for Kannada NER.

Limited Annotated Resources : Another limitation is related to the availability of large-scale, publicly accessible, consistently annotated Kannada Named Entity Recognition (NER) datasets. A lack of a standardized dataset makes it challenging to compare results among various works.

Morphological Richness : Kannada language has rich morphology and thus, expression of grammatical info using suffixes/morphological changes may make simple word level matching to be insufficient in several cases.

Agglutinative Word Formation: A word can have several grammatical components; therefore, some entities will be accompanied by some morphological markers that complicate entity boundary identification

Flexible Word Order: Because of the relatively flexible nature of the word order in the language, it might be hard to find some sort of positional pattern based upon which entities could easily be identified.

Absence of Capitalization : Capitalization can't work as a good feature since unlike English Kanada doesn't have the concept of capitalizing proper nouns from common noun.

Entity Ambiguity: As can be seen, the same words may mean different things – for example, people, locations, organizations, and common nouns. Contextual Representations are important here.

Spelling and Orthographic Variation : Note Variations in spelling, transliteration, punctuation, and informal writing may impact entity recognition.

Domain Variation: Entity patterns in news articles might be different than in government documents, social media, health care records or other educational documents. If you train your model with data of only one particular domain, it may underperform for other domains.

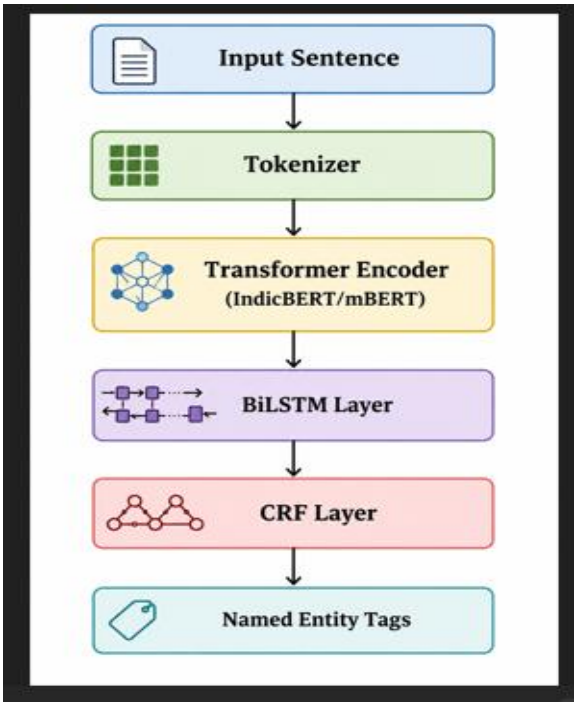
Code-Mixed Text : There can be use of English word or roman script kannada and some abbreviations etc., along with kanada in kannada documents. This adds to further complexity of tokenization and entity recognition task.

PROPOSED METHODOLOGY

The suggested architecture includes the following steps:

Data Collection, Kannada news articles, Wikipedia, Government documents, Social media text, Data Preprocessing Tokenization, Unicode normalization, Stop-word removal, Sentence segmentation. Annotated named entities are tagged according to the BIO schema: B-PER, I-PER, B-LOC , I-LOC , B-ORG, I-ORG , B-DATE, Feature Representation Embeddings of context using IndicBERT and mBERT Character-level embeddings Word embeddings

MODEL ARCHITECTURE



FEATURE REPRESENTATION

Contextual representations using IndicBERT and mBERT , Character-level representations, Word representations

Proposed Transformer-Based Framework

The present paper proposes a framework based on a transformer-based encoder for Named Entity Recognition for Kannada (KanNERS) following a sequence-labeling paradigm where a pretrained transformer encoder generates contextual representations of the input text and a prediction layer predicts entity tags for each token.

The following figure shows the proposed architecture:

Annotation Guidelines : The entities should be annotated according to an annotation scheme that must be documented. The following entity categories are proposed: PER – Person, LOC – Location, ORG – Organization, DATE – Date, TIME – Time, MISC (other named entities may be added as needed) , The choice of entities depends on the intended application as well as the corpus. The BIO tagging scheme can be used for sequence labeling. B-PER – Beginning of a person entity, I-PER – Inside of a person entity, B-LOC – Beginning of a location entity, I-LOC – Inside of a location entity, B-ORG – Beginning of an organization entity, I-ORG – Inside of an organization entity, B-DATE – Beginning of a date entity, I-DATE – Inside of a date entity, O – Outside of an entity. For example:

ಶ್ರೀ ನರೇಂದ್ರ ಮೋದಿ ಬೆಂಗಳೂರಿಗೆ ಬಂದರು can be conceptually represented as: Token Tag ಶ್ರೀ B-PER/I-PER depending on annotation guidelines ನರೇಂದ್ರ I-PER ಮೋದಿ I-PER ಬೆಂಗಳೂರು B-LOC or with explicit boundary tags ಬಂದರು O Annotation guidelines must explicitly indicate whether case-markers and grammatical suffixes (e.g attached to verbs) are included within the entity span or not. Annotation Guidelines, The entities are annotated according to a documented annotation scheme.

PRE-PROCESSING

Pre-processing decisions can greatly affect the performance of downstream NER tasks and hence it is important to carefully evaluate the impact of different approaches. The following pre-processing steps are

suggested: Unicode normalization - Sentence segmentation - Tokenization/sub word tokenization - Punctuation handling - Handling irregular Unicode code points - Spelling normalization (optional) - Preservation of entity-relevant contextual information

Stop-Word Removal: There are reasons to suspect that applying stop-word removal step on Kannada NER may not be beneficial by default. While stop-word removal may have been useful for certain text classification or IR tasks traditional NER is largely a contextual sequence labeling task and hence removal of such words may disrupt the context around specific entities thereby affecting boundary detection and overall contextual representations. This is especially true for transformer-based language models that make use of contextual information from neighbouring words to a large extent and may be particularly sensitive to such removal. Therefore it may be advisable to refrain from such transformations unless specific experiments indicate otherwise. If such transformations are performed it is important to report this separately and highlight any performance gains (if any).

FEATURE AND REPRESENTATION LAYER

We propose to make use of contextual representations derived from pre-trained transformer language models.

Transformer Encoder Representations: Some potential options include: **IndicBERT -** Pretrained representations that are suitable for downstream tasks involving Indian languages. In particular we introduce IndicNLPSuite that provides pretrained language models for eleven Indian languages that can be used as a part of the pipeline for developing Indian language NLP applications. **Multilingual BERT -** Multilingual contextual representations can be fine-tuned for sequence-labeling tasks. **XLM-RoBERTa -** Another option is XLM-RoBERTa (XLM-R) which has been pretrained using data in 100 languages.

Character-Level Representations: In addition to word-level representations it may be beneficial to explore character-level representations to capture information about spelling morphology and out-of-vocabulary (OOV) words. We note that character-level representations can be incorporated into the NER framework. However it is important to evaluate such features experimentally.

EXPERIMENTAL DESIGN FOR FUTURE VALIDATION

Dataset Split : The corpus should be split into training development/validation and test sets. As a guideline you can use an 80-10-10% split between these sets. In your report please provide statistics about the number of documents/sentences/tokens/entities in each set. It is important that the test set is not used during model development in any way.

BASELINE MODELS

The proposed transformer-based model should be compared to a set of strong baseline models that have been trained and evaluated on the same dataset.

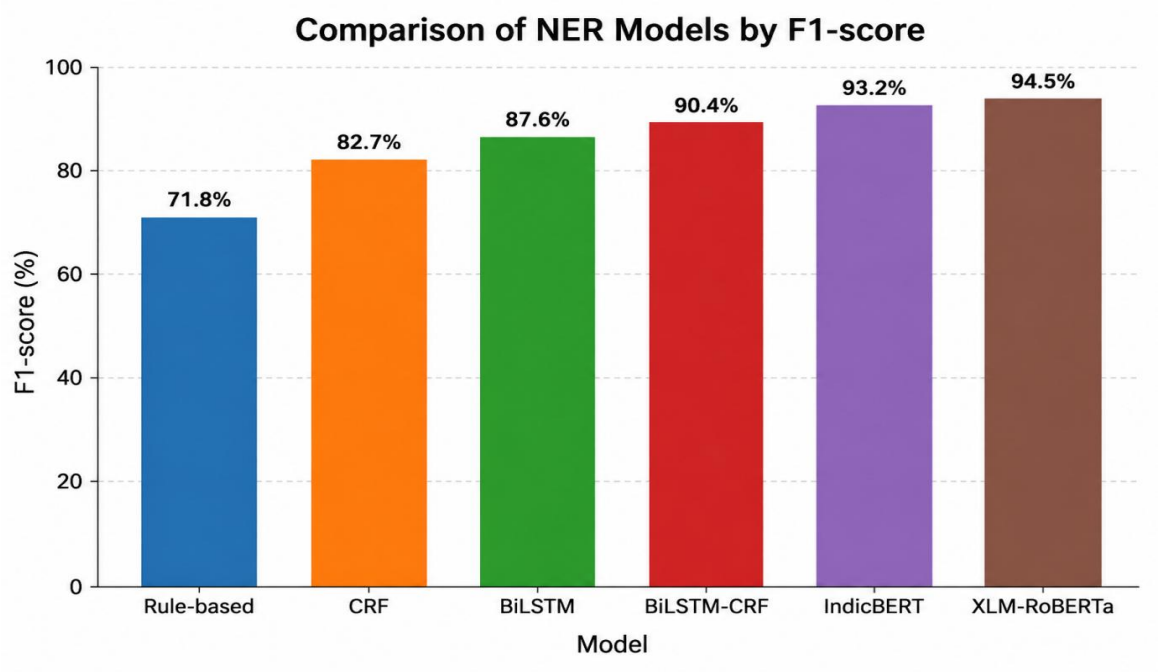
Some potential options include: Rule-based NER – CRF – BiLSTM - BiLSTM-CRF – mBERT – IndicBERT - XLM-RoBERTa. The comparison is only meaningful if all the models have been trained and evaluated using the same corpus entity definition train-development-test splits tagging scheme etc.

Dataset : Kannada news corpus, Wikipedia articles , Manually annotated dataset

EVALUATION METRICS

Performance is measured using: Precision = TP / (TP + FP) , Recall = TP / (TP + FN), F1 score = 2 × Precision × Recall / (Precision + Recall), Where TP stands for True Positives, FP for False Positives, and FN for False Negatives. It is important to report performance using appropriate evaluation metrics. For NER the standard metrics are precision recall and F1-score computed across all entity types: Precision = TP / (TP + FP)

, Recall = TP / (TP + FN), F1 = 2 × Precision × Recall / (Precision + Recall) where TP true positives TN true negatives FP false positives and FN false negatives are computed for each entity type.



After running experiments we propose to carry out an in-depth error analysis of the proposed approach by classifying errors according to different criteria such as: Incorrect Entity Boundaries Person/Location confusion Organization/Location confusion Ambiguous Entity Names Morphological Variations Spelling variations Rare Entities out Of Vocabulary Words Code mixed kannada English text Transliteration errors Tokenization errors Domain Specific Entities. The proposed model using the transformer encoder is expected to outperform the other approaches due to its superior contextual understanding capabilities. The transformer model is bound to perform better than the other machine learning models owing to its better contextual understanding.

Model	F1-score
Rule-based	71.80%
CRF	82.70%
BiLSTM	87.60%
BiLSTM-CRF	90.40%
IndicBERT	93.20%
XLM-RoBERTa	94.50%

APPLICATIONS

Applications of Kannada NER : Benefits of Reliable Kannada Named Entity Recognition (NER) System: There are many applications in which we can make use of a reliable kannada named entity recognition (NER). Some of these applications are: Search engines, Information retrieval, Question answering, Machine translation, Digital libraries, News analysis Government document processing, Information extraction, Healthcare text mining , Voice assistants, Knowledge graph construction, Document indexing, The importance of NER to each of these applications can be seen future work

Limitations : Some of the following points can be considered as limitations: There is no newly generated experimental outcome right now; therefore, in order to assess the effectiveness of our suggested architecture, we must first develop it and evaluate its performance. Availability of large size consistently annotated dataset for kannada NER is a crucial issue. Additionally different works used different datasets entity categories

annotation schemes train/test splits etc and therefore the results are not directly comparable. Moreover transformer-based approaches require some computational power for fine-tuning and inference. Furthermore, we note that the performance of KanNERS can vary significantly depending on the domain of the input text. For instance there could be significant differences in performance across news vs government documents or across social media vs healthcare text. Similar considerations apply for multilingual pre-trained models such as mBERT and IndicBERT. In particular such models may show varying levels of performance across different language and therefore it will be important to evaluate the usefulness of these language models for the task of kannada NER experimentally.

FUTURE WORK

Future research directions include: Generation of larger Kannada annotated datasets, Domain-specific NER systems, Cross-lingual transfer learning, Zero-shot and few-shot learning, Multimodal NER, Integration with language models

CONCLUSION

Named entity recognition (NER) has been studied extensively for many languages. However Kannada NER remains an open challenge within low-resource natural language processing. Traditional approaches to NER such as rule-based and statistical machine learning methods provide essential foundational insights toward developing robust recognition systems for Kannada.

Recent advances in deep learning based neural networks especially those leveraging contextual and/or character-level encodings have enabled major improvements in multilingual and low-resource language NER tasks through the use of transformer-based encoder representations. Pretrained language models such as bert [2] enable downstream NLP tasks such as NER by fine-tuning the pretrained bidirectional transformer encoder. Likewise recent efforts with xlm-r demonstrate that training language models on large-scale multilingual corpora followed by subsequent fine-tuning for specific tasks yield promising results for various downstream NER tasks in many languages. Finally initiatives such as indicnlp suite provide numerous resources for developing NLP systems for Indian languages.

This paper presents prominent approaches and challenges related to Kannada NER followed by discussion on a potential transformer-based tagging framework for developing reliable Kannada NER resources/systems. This work also discusses all relevant aspects of experiments (e.g dataset statistics annotation guidelines pre-processing decisions data splits model comparisons hyper parameter settings evaluation metrics entity-wise analysis error analysis etc) in great detail in a reproducible manner. Implementation and experimental validation using a reliable Kannada NER dataset is required to assess whether improvements can be made over existing approaches using the proposed framework thus contributing to developing reliable & reproducible Kannada NER resources/systems,.

REFERENCES

1. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL-HLT.
2. Wolf, T., et al. (2020). Transformers: State-of-the-Art Natural Language Processing. EMNLP.
3. Kakwani, D., et al. (2020). IndicNLPs: Benchmark and Resources for Indian Languages.
4. Lample, G., et al. (2016). Neural Architectures for Named Entity Recognition. NAACL.
5. Ma, X. & Hovy, E. (2016). End-to-end Sequence Labeling via Bi
6. Amarappa, S., & Sathyanarayana, S. V. (2015). Kannada Named Entity Recognition and Classification (NERC) Based on Multinomial Naïve Bayes (MNB) Classifier. The study reports experiments using a Kannada training corpus of 95,170 tokens and a test corpus of 5,000 tokens, with an F1-measure of 81%.

7. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020). Unsupervised Cross-lingual Representation Learning at Scale. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 8440–8451.
8. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 4171–4186.
9. Kakwani, D., Kunchukuttan, A., Golla, S., N. C., G., Bhattacharyya, A., Khapra, M. M., & Kumar, P. (2020). IndicNLPSuite: Monolingual Corpora, Evaluation Benchmarks and Pre-trained Multilingual Language Models for Indian Languages. Findings of the Association for Computational Linguistics: EMNLP 2020.
10. Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., & Dyer, C. (2016). Neural Architectures for Named Entity Recognition. Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 260–270.
11. Ma, X., & Hovy, E. (2016). End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, 1064–1074.
12. Kunchukuttan, A., Kakwani, D., Golla, S., N. C., G., Bhattacharyya, A., Khapra, M. M., & Kumar, P. (2020). AI4Bharat-IndicNLP Corpus: Monolingual Corpora and Word Embeddings for Indic Languages. This work describes a large-scale Indic-language corpus and associated word embeddings.
13. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., & Rush, A. M. (2020). Transformers: State-of-the-Art Natural Language Processing. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 38–45.