==========
Language in India www.languageinindia.com ISSN 1930-2940 Vol. 25:11 November 2025
==========

Evaluating IndicNER for Telugu: Entity-Level Performance and Error Analysis

Vanka Ramesh
Research Scholar
University of Hyderabad
vankaramesh2001@gmail.com

==========

Abstract

This paper presents a comprehensive evaluation of IndicNER, a multilingual transformer-based Named Entity Recognition (NER) system developed by AI4Bharat, on Telugu-language data. A dataset comprising 100 headlines collected manually from the Andhra Jyothi newspaper was annotated for three entity types: Person (PER), Location (LOC), and Organization (ORG). The IndicNER model's output was benchmarked against human annotations using standard metrics—Precision, Recall, and F1-score. While the system achieves satisfactory results for frequently named entities, it exhibits significant performance drops in handling morphologically rich expressions, compound names, and regional or domain-specific entities. Detailed error analysis reveals systematic challenges, including the misclassification of honorifics, code-mixed tokens, and inconsistent tagging of multi-word entities. The findings underscore the need for domain-adapted NER models and more representative training corpora for Telugu and other low-resource Indian languages.

1. Introduction

Named Entity Recognition (NER) is a foundational task in Natural Language Processing (NLP), which involves identifying and classifying entities, such as persons, locations, and organizations, within text. Over the past two decades, NER systems for high-resource languages like English have achieved remarkable accuracy due to the availability of large annotated corpora and advanced deep learning architectures. However, for an Indian language like Telugu, which is

spoken by over 80 million people primarily in southern India, the development and evaluation of robust NER systems remain a significant challenge.

Dravidian languages, including Telugu, present unique linguistic complexities for NER systems, including agglutinative morphology, rich inflectional forms, compound nouns, and frequent use of honorifics. Furthermore, Media language, with some grammatical liberties and shortforms, is often characterized by code-mixing, domain-specific jargon and abbreviations, making entity recognition more difficult. Despite recent advancements in multilingual models, the performance of these systems on Telugu data is underexplored and largely unvalidated.

IndicNER, a transformer-based NER model developed by IIT Madras as part of AI4Bharat, represents one of the first major efforts to bring pre-trained NER capabilities to Indian languages. While IndicNER has demonstrated promising results across several languages in the Indic NLP landscape, its performance on domain-specific and informal Telugu text—such as that found in newspaper headlines—has not been rigorously benchmarked or analyzed.

This paper presents a comprehensive evaluation of IndicNER on Telugu news headlines, focusing on three entity types available with IndicNER: Person (PER), Location (LOC), and Organization (ORG). A curated dataset of 100 headlines was manually annotated and compared against the automatic annotations generated by IndicNER. The evaluation employs standard metrics—Precision, Recall, and F1-score—based on a confusion matrix framework. Beyond quantitative analysis, the study conducts a detailed error analysis to uncover common patterns in misrecognition, such as erroneous tagging of compound entities, misidentification of names versus titles, and missed entities due to tokenization errors or contextual ambiguity.

The findings of this study offer critical insights into the current capabilities and limitations of IndicNER when applied to Telugu. Through both quantitative benchmarking and qualitative error analysis, the paper highlights the system where it fails, revealing consistent patterns of misclassification and omission. This work evaluates a state-of-the-art multilingual NER system on this data, and outlines practical recommendations for enhancing entity recognition performance in morphologically rich and linguistically complex languages like Telugu.

2. Related Works

Named Entity Recognition (NER) has been studied extensively for Indian languages over the last two decades, with early efforts focusing on rule-based and statistical methods. For Telugu, early systems primarily used Conditional Random Fields (CRFs) and language-specific features (Raju & Rao, 2011; Srikanth & Murthy, 2010). These systems often suffered from low recall due to the morphological richness of Telugu and the lack of annotated corpora.

In recent years, transformer-based multilingual models such as BERT (Devlin et al., 2019), IndicBERT (Kakwani et al., 2020), and MuRIL (Khanuja et al., 2021) have shown promise in improving Indian language NER through zero-shot and fine-tuned approaches. Building upon these architectures, AI4Bharat introduced IndicNER, a multilingual NER system trained on synthetic and crowd-sourced annotated corpora across 20 Indian languages, including Telugu. The system is capable of recognizing three core entity types—Person, Location, and Organization—using a unified BIO tagging format (AI4Bharat, 2022).

While IndicNER has reported competitive macro-F1 scores on internal test sets, there is limited independent evaluation of its performance. Previous benchmarking efforts for Indian NER systems have mostly evaluated model performance on curated, domain-specific datasets (Patel et al., 2022), neglecting practical settings like code-mixed content and short forms, which are common among Indian languages.

A few studies have compared automatic and manual annotations in Indian languages (Jain et al., 2020), suggesting that transformer-based models often miss fine-grained linguistic cues However, no comprehensive error analysis has yet been performed for Telugu using real-world annotations, especially to identify systematic patterns of misclassification, such as suffix-triggered errors, partial entity recognition, or handling of honorifics and compound names. This paper addresses this gap by providing an empirical evaluation of IndicNER against manually annotated Telugu headlines, highlighting specific error patterns and discussing the implications for domain-specific and context-aware NER modeling in low-resource settings.

3. Date Collection and Annotation

To evaluate the effectiveness of IndicNER in recognizing named entities in Telugu, a dataset of 100 news headlines was curated from the digital edition of *Andhra Jyothi*. The headlines cover multiple domains such as politics, economy, sports, and science, ensuring linguistic and topical diversity suitable for testing entity recognition. Each headline was tokenized and annotated for three entity types—Person (PER), Location (LOC), and Organization (ORG)—using two parallel methods: First, automatic annotation using IndicNER, which applies the standard BIO (Begin-Inside-Outside) tagging format, and then manual annotation conducted by human annotators, following established Indian language NER guidelines and protocols (Bharati et al., 2009; Saha et al., 2008). The manually annotated dataset serves as the reference gold standard, providing a benchmark for evaluating IndicNER's performance on Telugu text.

4. Methodology

The study adopts a comparative evaluation methodology to measure IndicNER's accuracy and reliability against manual annotation. The evaluation process includes three stages: entity alignment, metric computation, and error analysis. In the entity alignment stage, only tokens labeled as named entities by either IndicNER or the manual annotators were selected. Each token was classified as: True Positive (TP) if IndicNER correctly identified the entity type, False Positive (FP) if IndicNER incorrectly predicted an entity, False Negative (FN) if IndicNER failed to recognize a manually annotated entity, and non-entity tokens were excluded to focus solely on entity recognition accuracy. Next, Precision, Recall, and F1-score were computed using the following formulas:

$$Precision = TP / (TP + FP)$$

$$Recall = TP / (TP + FN)$$

$$F1\text{-}score = 2 \times (Precision \times Recall) / (Precision + Recall)$$

These metrics were calculated both overall and separately for PER, LOC, and ORG categories to identify type-specific trends. Finally, a qualitative error analysis was conducted by reviewing cases of mismatch to identify recurring linguistic patterns and system limitations. This multi-layered methodology ensures a balanced assessment of IndicNER's strengths and weaknesses in handling Telugu-language data.

5. Results

The performance of IndicNER compared with the manually annotated dataset of Telugu news headlines was evaluated using Precision, Recall, and F1-score. These metrics were computed at two levels: (i) an overall evaluation across all entity types, and (ii) specific entity types: Person (PER), Location (LOC), and Organization (ORG).

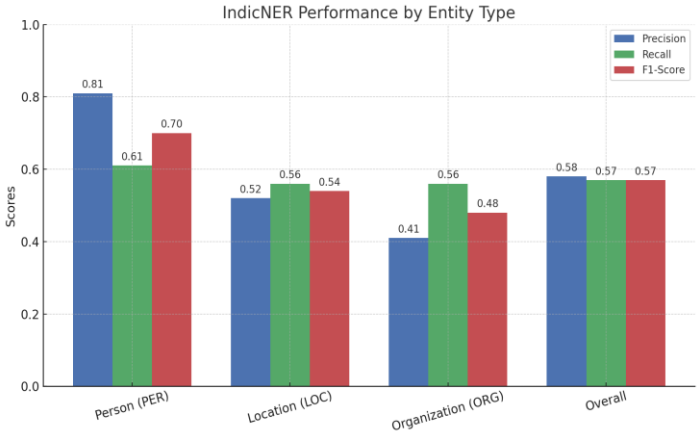
5.1. Overall Evaluation:

At the aggregate level, IndicNER achieved a Precision of 0.58, a Recall of 0.57, and an F1-score of 0.57 across all entity types. These metrics reflect a reasonably balanced but moderate performance in recognizing named entities. Precision of 0.58 indicates a fair level of correctness in the identified entities, the slightly lower recall shows that some entities were predicted incorrectly. While the near parity between precision and recall suggests that the model makes consistent predictions, but neither captures all relevant entities nor filters out false positives effectively. This balance likely results from a generic training setup lacking domain-specific adaptation to Telugu, particularly in the informal and compressed structure of headline text. The overall F1-score of 0.57 underscores the model's potential for foundational applications, but also highlights the need for linguistic refinement and task-specific tuning to support high-stakes use cases in Telugu.

5.2. Entity-Type Specific Evaluation

When analyzed by entity type, IndicNER showed the highest accuracy in recognizing Person (PER) entities, with a Precision of 0.81, a Recall of 0.61, and an F1-score of 0.70. This can be attributed to the higher frequency and structural consistency of person names. Performance on Location (LOC) entities was moderate, with a Precision of 0.52, a Recall of 0.56, and an F1-score

of 0.54. Many errors in these categories were due to compounds and abbreviations, which the model failed to capture fully.



The lowest performance was observed in the Organization (ORG) category, where the model attained a Precision of 0.41, a Recall of 0.56, and an F1-score of 0.48. Most misclassifications occurred due to confusion between organization names and other entity types, especially when honorifics or abbreviations were used. Additionally, the system often failed to capture the full span of multi-word organization names, leading to boundary mismatches.

These results indicate that while IndicNER is effective for certain types, especially person entity types, its performance varies significantly depending on the linguistic structure and context of the entities.

6. Error Analysis

To identify the linguistic and algorithmic limitations of IndicNER in the context of Telugu named entity recognition, a structured and critical error analysis was conducted and identified six prominent error categories. This section presents these error categories with carefully selected examples, accompanied by detailed linguistic interpretation and technical implications.

6.1 Boundary Mismatches and Partial Recognition

IndicNER often fails to recognize the complete span of a multi-token entity, especially with personal names, which are commonly formed using multiple lexical units (e.g., given name + father's name + surname). This failure manifests in the form of entity splitting or truncated recognition, leading to severe issues in entity linking and co-reference tasks.

Example 1:

Headline: padeVlluga rABAruT radrAnu veMTAduTunnArA: rAhuL gAMDI

IndicNER: rAhuL **<O>**, gAMDI **<B-Per>**

The model fails to associate the token "rAhuL" with its surname "gAMDI," and incorrectly treating them as separate or unrelated entities. This reflects an underlying limitation in sequence modeling, particularly for names without typical Western capitalization cues or fixed formats.

Example 2:

Headline: bOnamettiina RS praveeN kumAr..

IndicNER: bOnamettiina **<B-Per>**, ArES **<I-Per>**, praveeN **<I-Per>**, kumAr **<I-Per>**

The entity is tagged fully; IndicNER does not differentiate between *bOnamettiina*, a compound of the noun *bOnam* and verb participle *ettiina*. The uniform tagging of all four tokens as a single person entity overlooks the semantic distinction within the phrase. This suggests that although token-level tagging is accurate, the model lacks structural understanding of Telugu naming conventions, especially when descriptive elements precede names.

6.2 Entity-Type Confusion

Often, the place names present as parts of institutional names (e.g., University of Hyderabad), and vice versa. IndicNER frequently misclassifies the type of an entity, especially in cases where contextual or compound noun interpretation is required. Such mislabeling reduces the semantic value of the entity and may misguide downstream classification or summarization tasks.

Example 1:

Headline: 14 rOjullO 50 hatyalu..eMDIyE pAlita bIHArlO kSIiNincina sAMtibadratalu

IndicNER: eMDIyE **<O>**, bIHAr **<B-Org>**

"bIHAr" is misclassified as an organization instead of a location, and "eMDIyE" is ignored altogether, creating confusion through the abbreviation. This shows IndicNER's tendency to

default to high-frequency labels and inability to resolve geopolitical references in compound structures.

Example 2:

Headline: siddArAmayya kannumUta aMTu meTA anuvAda dOSam

IndicNER: siddArAmayya <O> kannumUta <B-Per>

IndicNER incorrectly tags *kannumUta* 'death' as a person entity, while missing the actual named entity siddArAmayya, a known person. This error reveals the model's reliance on positional heuristics rather than semantic understanding. IndicNER misclassifies this due to its inability to parse such syntactic patterns, mistaking the contextual verb for a named entity and skipping the subject. This kind of type confusion reduces the reliability of entity tagging in obituaries or death reports—an important genre in Telugu news.

6.3 Systemic Omissions and False Negatives

False negatives are critical in evaluating recall. We observed that IndicNER often misses entire entities, especially those that are (i) region-specific, (ii) infrequent, or (iii) linguistically irregular. These systemic omissions lead to data sparsity issues and lower utility for regional language deployment.

Example 1:

Headline: MP mitunreDDi arrestu..!

IndicNER: mitunreDDi **<B-Org>**

IndicNER erroneously classifies *"mitunreDDi"* as an organization rather than a person, despite contextual cues such as the presence of *"eMPi"* (MP) and *"arrestu"* that indicate a person. This misclassification exemplifies a false negative in the person category and an incorrect assignment to the organization type. The error underscores the model's insufficient contextual grounding and a bias toward frequent label patterns rather than semantically coherent tagging.

Example 2:

Headline: krSNamma ku varada hOrU.. SrISailam prAjekTu

IndicNER: SrISailam **<O>**

The model overlooks “SrISailam,” a well-known religious and geographical location, possibly due to morphological complexity or contextual ambiguity. The name of the project may represent location, a famous personality, and culturally related terms, which results in confusion to the system. This severely impacts tasks such as event extraction or religious news monitoring.

6.4 Morphological Suffix Challenges

Telugu frequently attaches postpositions, honorifics, and case markers directly to named entities. When such suffixes are not segmented, NER systems like IndicNER tend to ignore or partially recognize the entity, leading to both false negatives and span inconsistencies.

Example 1:

Headline: ememlYe pallA rAjESvar reDDi nu parAmarSinchina mahES bigAla

IndicNER: reDDi nu**<O>**

The suffix “nu” (accusative case) causes the model to misinterpret or ignore the token “reDDi.” This suggests a lack of morphological preprocessing and suffix-aware training.

Example 2:

Headline: revant reDDi DilhI payaNAlapai SVEtapatram

IndicNER: revant **<B-Per>** reDDi **<O>**

IndicNER fails to incorporate the second half of a common person's name due to its occurrence in a complex predicate context. Morphological disambiguation is essential to avoid such omissions.

6.5 Abbreviation and Acronym Errors

Telugu headlines often include abbreviations of political parties, institutions, or government bodies in Roman or mixed scripts. IndicNER frequently fails to generalize across these tokens, especially when they appear outside the script or word forms seen in training data.

Example 1:

Headline: KCR kiT ku mangaLam.. rU.65 kOTlU kendrAnki vApass

IndicNER: KCR **<B-Org>**, kiT **<I-Org>**

Despite context suggesting a person's name, the model interprets the pair as an organization, possibly due to surface form biases in uppercase acronym patterns.

Example 2:

Headline: RCB de tappu.. cinnaswAmi ghaTanapai karnATaka prabhutva nivEDika

IndicNER: RCB **<B-Loc>**

The RCB (Royal Challengers of Bengaluru) is wrongly classified as a location, suggesting that abbreviations are resolved only based on structural patterns of the individual entity, not on contextual function.

6.6 Code-Mixed and Romanized Entity Errors

Telugu news headlines often integrate English-origin named entities, either in full (e.g., Netflix, OU) or transliterated form. IndicNER shows significant weakness in detecting these entities, indicating insufficient multilingual modeling.

Example 1:

Headline: AkAS praim misail parIksha success

IndicNER: praim **<B-Per>**, misail **<I-Per>**

The initial token “AkAS” is missed entirely, and subsequent tokens are reinterpreted as a new entity. This affects recognition of named weapon systems or scientific objects in code-mixed headlines.

Example 2:

Headline: AIS, FBI, CBI lAnti saMsthAlapai tinmar

IndicNER: AIS**<O>**, FBI**<O>**, CBI**<O>**

These globally known organization names are not recognized, pointing to critical lexical gaps in the model and the need for external gazetteer integration or multilingual pretraining.

7. Discussion

This study presents a comprehensive evaluation of IndicNER, a multilingual Named Entity Recognition system, in the context of Telugu news headlines, a linguistically complex and morphologically rich domain. While the system demonstrates baseline competence—especially in recognizing Person (PER) entities—our findings reveal critical limitations in handling entity boundaries, entity type disambiguation, and morphological variations. These limitations significantly affect the model's precision, recall, and real-world applicability across tasks involving information extraction, event tracking, and media analysis.

The quantitative results show that IndicNER achieves a moderate overall performance with a Precision of 0.58, a Recall of 0.57, and an F1-score of 0.57. When disaggregated by entity type, the system performs best on Person names (F1 = 0.70), but struggles with Location (F1 = 0.54) and Organization (F1 = 0.48) entities. This discrepancy highlights a systemic bias in the model's learning, which favors frequently occurring and syntactically simpler entities, while underperforming on structurally complex or domain-specific expressions.

A qualitative error analysis further illuminates the nuanced challenges faced by the system. We identified six major error types: (i) boundary mismatches, where entity spans are incompletely captured; (ii) entity-type confusion, particularly between locations and organizations; (iii) systemic omissions, where the model fails to detect valid entities altogether; (iv) morphological suffix failures, reflecting the model's inability to segment agglutinative tokens correctly; (v) abbreviation misclassification, especially with Romanized or uppercase short forms; and (vi) code-mixed entity failures, where IndicNER is unable to detect English-script or transliterated named entities embedded in Telugu.

The error patterns suggest that IndicNER lacks contextual grounding and morphological sensitivity, both of which are crucial for high-fidelity NER in Telugu. These observations reinforce that surface-level modeling, even when trained on large multilingual corpora, is insufficient for capturing the complexities of Indian languages. From a practical standpoint, these limitations are

not just academic. Inaccurate or incomplete entity recognition can degrade the performance of downstream tasks such as automated news summarization, sentiment analysis, policy monitoring, and named entity linking in multilingual information retrieval systems. The fact that important organizations, people, or locations are either mislabeled or omitted entirely raises serious concerns about the system's trustworthiness and scalability in real-world applications.

7.1. Consideration and Implications

The findings of this study carry significant implications for the advancement of Named Entity Recognition in Telugu and other Indian languages. The analysis reveals that a one-size-fits-all approach, such as that adopted by IndicNER, is insufficient for languages with agglutinative structures, complex honorific systems, and frequent code-mixing. Therefore, future systems must incorporate morphological preprocessing techniques that can segment case markers and honorific suffixes commonly found in Telugu. Fine-tuning models on domain-specific corpora—covering diverse sectors such as politics, entertainment, and education—will help in adapting to varying linguistic styles and terminologies used in real-world text. Additionally, expanding gazetteers to include regional place names, culturally specific organizations, and Romanized forms is essential for improving both recall and contextual accuracy. The inability to recognize multi-word or compound entities also calls for post-processing mechanisms, such as CRF-based span merging or rule-based refinement, to ensure complete and coherent entity tagging. Finally, the integration of script-adaptive training techniques can significantly improve the recognition of code-mixed or Roman-script entities, which are prevalent in digital Telugu content. Collectively, these enhancements would not only improve the performance of IndicNER but also contribute to more inclusive, reliable, and culturally aligned NLP tools for Telugu and other underrepresented Indian languages.

8. Conclusion

This study critically evaluated the performance of IndicNER on a manually annotated dataset of Telugu news headlines, highlighting both its strengths and limitations. While the system performs reasonably well in identifying Person entities, it struggles with Location and Organization types, particularly in handling multi-word names, suffixes, and code-mixed tokens.

The error analysis revealed systematic issues, including boundary mismatches, entity-type confusion, and failure to recognize morphologically complex or low-frequency entities. These findings underscore the need for language-specific adaptations in NER systems for Telugu. Enhancements such as morphological preprocessing, domain-specific fine-tuning, and expanded regional gazetteers are essential to improve accuracy and coverage. Overall, this work offers a practical evaluation of IndicNER's readiness for Telugu and provides a roadmap for future development of more robust, context-aware NER tools for Indian languages.

References

AI4Bharat. (2022). *IndicNER: Named Entity Recognition models for Indian languages*. Retrieved from https://github.com/AI4Bharat/IndicNER

Bharati, A., Husain, S., Sangal, R., & Sharma, D. M. (2009). ICON-2009 NLP Tools Contest: Named Entity Recognition. *Proceedings of ICON-2009*.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of NAACL-HLT*, 4171–4186.

Ekbal, A., & Bandyopadhyay, S. (2008). Named Entity Recognition using Support Vector Machine: A language independent approach. *International Journal of Computer Systems Science and Engineering*, *4*(1), 35–51.

Jain, A., Mishra, R., & Sankaranarayanan, K. (2020). Evaluating Annotation Consistency for Named Entity Recognition in Indian Languages. Proceedings of LREC 2020, 4556–4562.

Kakwani, D., Chaudhary, Y., Dandapat, S., Bhattacharyya, P., & Kunchukuttan, A. (2020). IndicNLPSuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 4948–4961.

Khanuja, S., Dandapat, S., Kakwani, D., Choudhury, M., & Bhattacharyya, P. (2021). MuRIL: Multilingual representations for Indian languages. *arXiv preprint arXiv:2103.10730*.

Lafferty, J., McCallum, A., & Pereira, F. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *Proceedings of the 18th International Conference on Machine Learning (ICML)*, 282–289.

Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., & Dyer, C. (2016). Neural architectures for named entity recognition. *Proceedings of NAACL-HLT 2016*, 260–270.

Patel, M., Kulkarni, M., & Bhattacharyya, P. (2022). A Comparative Study of Named Entity Recognition Tools on Indian Languages. *Journal of Computational Linguistics and Applications*, *13*(2), 45–63.

Raju, A. B., & Rao, D. M. (2011). Named Entity Recognition for Telugu Using CRFs. *International Journal of Computer Applications*, *27*(11), 21–25.

Saha, S. K., Ekbal, A., & Bandyopadhyay, S. (2008). A hybrid feature set based maximum entropy Hindi named entity recognition. *Proceedings of ICON-2008*.

Srikanth, K., & Murthy, K. N. (2010). Named Entity Recognition for Telugu Using Conditional Random Fields. *Proceedings of the Workshop on South and Southeast Asian NLP (WSSANLP)*, 47–54.