International Journal of Advance and Applied Research
www.ijaar.co.in
ISSN – 2347-7075 Impact Factor – 8.141
Peer Reviewed Bi-Monthly
Vol. 6 No. 38 September - October - 2025

Natural Language Processing for Indian Languages: Challenges, Resources, and Directions

Ms. Vikhe Rupali Sopanrao
Assistant Professor,
Women's College of Home Science and BCA, Loni.
Corresponding Author – Ms. Vikhe Rupali Sopanrao
DOI - 10.5281/zenodo.17312811

Abstract:

India's linguistic diversity presents both a rich opportunity and a significant challenge for Natural Language Processing (NLP). This paper surveys the state of NLP for Indian languages, covering linguistic characteristics, available corpora and benchmarks, recent model advances (including Indic-specific pre-trained models), task-specific progress (machine translation, speech, NER, sentiment analysis), and outstanding challenges such as low-resource settings, script diversity, and code-mixing. We discuss engineering strategies that have shown promise—transfer learning, transliteration-aware training, and multilingual pretraining—and outline research directions including benchmark standardization, inclusive data collection, and application of large language models. A curated list of resources and a recommended research roadmap are provided to help researchers and practitioners plan future work.

Keywords: Indian languages, Indic NLP, multilingual models, low-resource languages, datasets, benchmarks, MuRIL, IndicBERT

Introduction:

Indian languages form a linguistically vast and varied geography multiple families(Indo- Aryan, Dravidian, Austroasiatic, Tibeto-Burman), numerous scripts, agglutinative and inflectional morphologies, and wide law-mixing with English. Despite a large speaker base, numerous Indian languages are under-resourced in NLP terms — limited labeled data, inconsistent orthography, and many standardized marks. At the same time, digital relinquishment and government enterprise have created instigation for erecting usable language technologies in the Indian environment. This paper synthesizes recent progress, registers core coffers, and highlights algorithmic and evaluation requirements.

Linguistic characteristics relevant to NLP:

• Script diversity: A single language can be written in different scripts or transliterated into Latin; multiple languages use distinct scripts with typographic properties that affect tokenization and OCR.

• Morphology: Several Indian languages show rich morphology and compounding, increasing sparsity at the word level and making subword approaches important.

• Free word order: Many Indian languages allow relatively free word order which impacts syntactic parsing and alignment for MT.

• Code-mixing and transliteration: Real-world text, especially on social media, often contains script-switching and English code-mixing; models must handle mixed-script inputs robustly.

• Dialectal variation: Regional dialects and domain-specific registers (e.g., legal, agricultural) create domain-shift challenges.

Resources and Corpora:

Recent community efforts have dramatically improved the resource landscape for Indian languages. Key resources include:

• Indic corpora (AI4Bharat / IndicNLP): Large monolingual corpora covering multiple major Indian languages collected from web crawls, news, and digital archives. These corpora support pretraining and downstream tasks.

• Samanantar: Large parallel corpora for English–Indic language pairs, useful for machine translation and cross-lingual transfer.

• OSCAR and CommonCrawl derivatives: Noisy but massive sources of text for pretraining.

• Task-specific datasets: NER, POS, QA, sentiment datasets produced by academic groups, shared tasks, and industry labs.

• Benchmarks: IndicGLUE and other pan-Indic benchmarks collate several NLU tasks across languages to enable comparative evaluation.

Model Developments:

1. Pre-trained multilingual and Indic-specific models:

• mBERT / XLM-R: General multilingual models that provide strong baselines but underrepresent many Indian languages in training data.

• IndicBERT: A family of models trained specifically on multiple Indic languages to better capture language-specific patterns.

• MuRIL: A multilingual representation model explicitly trained for Indian languages, augmented with transliterated and translated pairs to help cross-script and cross-lingual performance.

These models demonstrate that targeted pretraining on in-language text and transliteration-aware strategies significantly improve downstream performance compared to generic multilingual models.

2. Task-specific methods:

• Transliteration-aware tokenization: Integrating transliteration pipelines or joint modeling of script variants helps reduce noise from Latin-script transliterations.

• Morphology-aware approaches: Morpheme segmentation or subword regularization reduces sparsity for morphologically rich languages.

• Data augmentation and synthetic parallel data: Backtranslation, synthetic transliteration, and translation-based data augmentation boost MT and classification performance in low-resource scenarios.

Major NLP Tasks: Progress & Challenges:

1. Machine Translation (MT):

Neural MT systems trained on parallel corpora (Samanantar, etc.) have enabled usable translation for many language pairs, especially when combined with transfer from high-resource languages and backtranslation. Challenges include domain mismatch and low-quality noisy parallel data for several languages.

2. Automatic Speech Recognition (ASR) and Text-to-Speech (TTS):

Speech datasets and end-to-end modeling have matured for a handful of major languages, but many languages still lack sizeable, high-quality speech corpora. Script support for speech technologies is complicated by orthographic normalization issues.

3. Named Entity Recognition (NER), POS, Parsing:

NER datasets exist for some major languages; however, cross-lingual transfer and annotation standards vary. Language-specific

POS tags and dependency annotation schemes require harmonization to allow multi-lingual parsers.

4. Sentiment and Social Media Analysis:

Code-mixed text dominates social media. Models that explicitly model code-mixing and transliteration show better robustness. Labeled datasets remain limited and skewed toward particular dialects or domains.

5. Question Answering (QA) and Reading Comprehension:

QA datasets have been created for several Indian languages; model performance improves with multilingual pretraining and careful dataset translation, but complex reasoning and culturally specific knowledge remain challenging.

Evaluation and benchmarks:

Benchmarks such as IndicGLUE provide unified NLU evaluation across tasks and languages, while newer efforts (e.g., BharatBench, IndicMMLU variants) aim to broaden task coverage and include industry-relevant use cases. Standardized evaluation is essential to compare approaches fairly and identify where resources should be allocated.

Ethical considerations, bias, and inclusion:

• Representation bias: Most datasets overrepresent certain languages, dialects, or formal registers (news), which biases models toward those varieties.

• Privacy and consent: Data collection must follow ethical norms, especially for speech and user-generated content.

• Accessibility: Building inclusive systems requires datasets and evaluation that reflect real users, including low-literacy and non-standard script users.

Open problems and future directions:

• Scaling to many low-resource languages: Automated dataset creation, weak supervision, and multilingual transfer learning are promising routes.

• Better handling of code-mixing: Joint models that can process mixed-script and mixed-language inputs natively.

• LLMs and instruction-following models for Indic languages: Adapting large language models and aligning them to regional languages and user needs.

• Multimodal and grounding: Combining text, speech, and vision for richer applications (e.g., agricultural advisories in local languages).

Practical roadmap for researchers:

1. Start with resources: Use IndicNLP corpora and catalogs to gather monolingual and parallel data.
2. Choose an approach: For low-resource languages, prioritize transfer learning from related languages and transliteration augmentation.
3. Benchmark: Evaluate on IndicGLUE or task-specific datasets; report language-wise breakdowns.
4. Ethics & release: Ensure documentation, consent (for speech), and clear license terms for dataset/model release.

Conclusion:

NLP for Indian languages has matured substantially in the past few years thanks to community efforts and targeted modeling strategies. However, large gaps remain for many languages and domains. Continued focus on data collection, inclusive benchmarks, and language-specific modeling will be required to build robust, equitable language technologies for India.

References:

1. D. Kakwani et al., "Monolingual Corpora, Evaluation Benchmarks and Pre-…" (IndicNLP resources).
2. Simran Khanuja et al., "MuRIL: Multilingual Representations for Indian Languages" (2021).
3. Simran Khanuja, Sebastian Ruder, Partha Talukdar, "Evaluating the Diversity, Equity and Inclusion of NLP Technology: A Case Study for Indian Languages" (EACL 2023 findings).
4. S Ghosh et al., "IndicFinNLP: Financial Natural Language Processing for Indian Languages" (LREC 2024).
5. B.S. Harish et al., "A comprehensive survey on Indian regional language…" (2020).