Natural Language Processing in Low-Resource Languages: Progress and Prospects

Ritul Phukan[1], Monalisa Daimari[2], Anupam Kharghoria[3], Biman Basumatary[4]
[1,2,3,4]Department of Computer Science and Engineering, Assam Down Town University, Guwahati, India

Abstract

Low-resource languages—languages with limited annotated corpora, lexicons, and digital resources—pose major challenges for modern natural language processing (NLP). Recent progress in transfer learning, multilingual pretraining, parameter-efficient adaptation, data augmentation, and community-driven dataset creation has substantially improved capabilities for many such languages, yet large performance gaps remain compared to high-resource languages. This article surveys the technical advances that enable NLP for low-resource languages (including unsupervised and weakly supervised methods, multilingual and massively multilingual models, few-shot and in-context learning with large language models, and adapter/LoRA-style parameter-efficient fine-tuning). We examine practical pipelines for tasks such as machine translation, speech recognition, OCR, and information extraction; describe prominent dataset and community projects; summarize typical evaluation strategies and their pitfalls; and outline promising research directions (community data collection, privacy-preserving methods, on-device adaptation, and ethics-aware deployments). The review highlights approaches that balance performance, compute cost, and data-efficiency, and recommends research and deployment practices to accelerate inclusive language technology.

Keywords

Low-resource languages, transfer learning, multilingual pretraining, few-shot learning, LoRA / adapters, data augmentation, machine translation, speech datasets, Masakhane, Common Voice

1. Introduction

Natural Language Processing (NLP) has become one of the most transformative domains in artificial intelligence, with applications ranging from machine translation and speech recognition to information retrieval and conversational systems. However, the benefits of these advancements have largely accrued to a small set of high-resource languages such as English, Chinese, French, and Spanish, which dominate digital resources and computational investments. In contrast, a vast majority of the world's 7,000+ languages are categorized as *low-resource*—languages with little or no annotated corpora, minimal presence on the internet, and insufficient linguistic tools. This disparity creates a digital divide in which speakers of underrepresented languages are systematically excluded from technological progress.

Low-resource languages are often spoken by marginalized or indigenous communities, making their inclusion not just a technical issue but also a matter of cultural preservation, social equity, and linguistic diversity. For example, many African and South Asian languages lack large-scale text corpora or speech datasets, and several indigenous languages remain primarily oral traditions with no standardized orthography. The rapid decline of linguistic diversity—UNESCO warns that nearly half of the world's languages may become extinct within the next century—further underscores the urgency of developing NLP solutions that work for these languages.

While traditional supervised learning approaches struggle in such contexts due to the scarcity of annotated data, recent progress in transfer learning, multilingual pretraining, unsupervised machine translation, and community-driven dataset creation has shown encouraging results. These advances not only improve the accessibility of digital technologies but also foster inclusivity by enabling speakers of low-resource languages to interact with technology in their native tongue. Nevertheless, critical challenges remain, particularly in the domains of linguistic representation, computational cost, and evaluation reliability. The following sections detail these challenges and provide a foundation for understanding the current state and future prospects of NLP in low-resource contexts.

2. Background and Core Challenges

Low-resource languages pose a unique set of challenges that distinguish them from their high-resource counterparts. The most significant issue is the **scarcity of data**, both in terms of raw text and annotated corpora. Most modern NLP models thrive on large volumes of high-quality data for training, but many low-resource languages have little to no representation in the digital space. While some may exist in oral traditions or localized publications, the absence of digitized text corpora makes it extremely difficult to build foundational resources. In particular, the lack of parallel corpora for translation tasks and the absence of labeled speech datasets hinder the progress of machine translation, automatic speech recognition, and text classification in these languages.

Another major challenge is the **linguistic complexity and variability** inherent to many of these languages. Low-resource languages often exhibit rich morphological structures, highly inflectional word forms, and diverse syntactic patterns that differ greatly from those of high-resource languages. This richness, while linguistically fascinating, increases data sparsity, since a single lemma can generate dozens or even hundreds of inflected forms. Furthermore, dialectal variation, code-switching practices, and the lack of standardized orthography introduce inconsistencies into the data, complicating both tokenization and model training.

Adding to these difficulties is the **lack of computational and institutional resources** in regions where such languages are spoken. Research on high-resource languages benefits from centralized infrastructure, large research teams, and substantial funding, while low-resource contexts often rely on small community-driven efforts with limited access to high-performance computing. This creates an asymmetry where the same advanced deep learning models that push the boundaries of English or Chinese NLP cannot be easily deployed or fine-tuned for less-represented languages.

Finally, **ethical and sociolinguistic issues** must be carefully considered. Many communities are concerned about how their data is collected and used. Extracting cultural texts or oral narratives without consent can lead to exploitation and mistrust. Sustainable progress in this area therefore requires methods that not only address technical gaps but also involve native speaker communities, respect cultural sensitivities, and ensure that the benefits of language technologies are shared with the people who contribute to their development.

3. Current Methods and Approaches

Despite these challenges, recent years have seen encouraging progress in building NLP systems for low-resource languages. One of the most effective strategies has been **transfer learning through multilingual pretraining**. Large language models such as mBERT, XLM-R, and mT5 have been trained on hundreds of languages simultaneously, allowing knowledge learned from high-resource languages to be transferred to related low-resource ones. For example, training on English and Hindi can indirectly benefit models for Nepali or Assamese due to shared vocabulary and grammatical features. This cross-lingual sharing has been a major step forward, but the gains vary significantly depending on how well the target language is represented in the pretraining corpus.

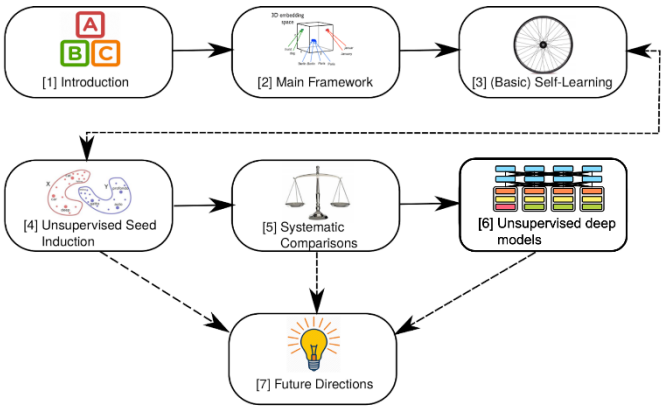

Figure 1: Cross-Lingual Transfer in Multilingual Models

Another important approach is **parameter-efficient adaptation**, which has become increasingly popular due to its cost-effectiveness. Instead of fine-tuning the entire model—a process that requires vast amounts of data and computing power—techniques such as adapters, LoRA (Low-Rank Adaptation), and prefix tuning allow researchers to adjust only a small subset of model parameters. This makes it possible to adapt large pretrained models to new languages with limited datasets and modest computing resources, a key requirement for teams working in resource-constrained environments.

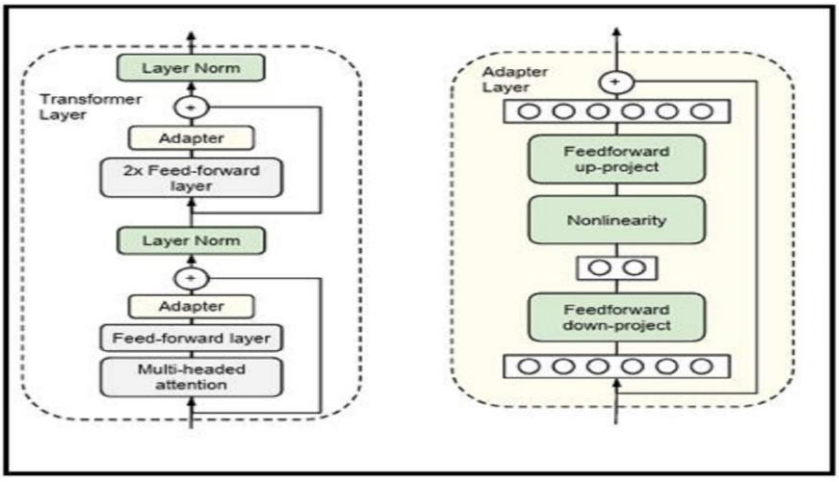

Figure 2: Parameter-Efficient Adaptation Techniques

Beyond model adaptation, data augmentation and synthetic data generation play an essential role in compensating for the lack of natural training material. Methods such as back-translation, morphological transformations, paraphrasing, and transliteration-based expansion create additional examples that enhance model robustness. While these techniques cannot fully substitute for genuine data, they help bridge gaps and improve generalization when used carefully. The rise of large-scale community projects has also reshaped the research landscape. Initiatives like Masakhane in Africa and Mozilla's Common Voice have demonstrated that grassroots collaboration can yield significant datasets for translation and speech recognition. By engaging native speakers and volunteers, these projects generate resources that would otherwise remain unavailable, while also fostering local ownership of the technology. Finally, few-shot and in-context learning with large language models is emerging as a promising direction. Recent studies show that even when a language is underrepresented in pretraining data, carefully designed prompts with a handful of examples can enable the model to perform basic translation, classification, or summarization tasks. While this capability is not yet reliable enough for widespread deployment, it indicates that future foundation models may offer scalable solutions for hundreds of currently neglected languages.

4. Datasets, Community Projects, and Infrastructure

The availability of high-quality datasets is the cornerstone of any successful NLP system, and for low-resource languages, this remains one of the most significant bottlenecks. While commercial and academic institutions have historically concentrated on collecting data for globally dominant languages, much of the progress for underrepresented languages has emerged from community-driven initiatives. These efforts not only fill critical gaps but also ensure that language technology development respects local cultural and ethical norms.

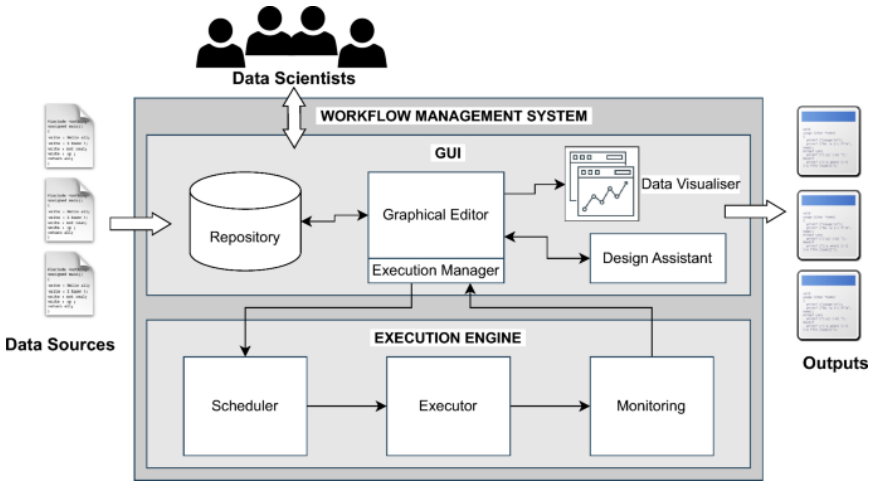

Figure 3: Community-Driven Data Collection Pipeline for Low-Resource Languages

One of the most influential grassroots movements in this domain is the Masakhane project, which brings together researchers and volunteers across Africa to build machine translation systems for African languages. Masakhane demonstrates the power of distributed collaboration, where native speakers contribute to dataset creation,

annotation, and validation. The project has already produced several parallel corpora and translation benchmarks, significantly increasing the visibility of African languages in global NLP research.

Similarly, Mozilla's Common Voice initiative provides a large-scale, crowd-sourced speech corpus covering dozens of languages, including many that were previously absent from mainstream datasets. Contributors donate voice samples through a web platform, and these recordings are validated by the community to ensure quality. This open-source approach has proven invaluable for building automatic speech recognition (ASR) systems in languages with little prior digital presence.

In addition to large-scale projects, regional universities and independent organizations are playing a vital role in data digitization. Many initiatives focus on digitizing folk literature, local newspapers, and government records, converting them into machine-readable corpora. Such efforts are crucial for tasks like part-of-speech tagging, named entity recognition, and sentiment analysis, which require domain-specific training examples.

Beyond datasets, infrastructure for sharing and standardization is also advancing. Platforms like Hugging Face Datasets and GitHub repositories enable open access to language resources, while standardized annotation formats and benchmarks improve reproducibility across studies. However, challenges remain in ensuring data consistency, handling dialectal diversity, and maintaining ethical data governance.

Overall, the progress in dataset creation and community collaboration underscores the fact that technological advances in low-resource NLP cannot be achieved through algorithms alone. They depend equally on the social infrastructure of collaborative data sharing, linguistic expertise, and community participation. By continuing to invest in grassroots projects and fostering open platforms, the NLP community can accelerate the inclusion of marginalized languages into the digital ecosystem.

5. Evaluation Practices and Case Studies

Evaluating NLP systems for low-resource languages is one of the most difficult aspects of research in this domain. While building models is already a challenge due to limited data, designing reliable and fair evaluation methods is equally complex. Standard benchmarks and metrics, which work reasonably well for high-resource languages, often fail to capture the true performance and usability of systems for underrepresented languages. One major limitation is the scarcity of representative test sets. For languages with only a few hundred sentences available for evaluation, results can be unstable and highly sensitive to dataset composition. A model that performs well on such a test set may fail completely in real-world applications, especially when exposed to dialectal or domain variation. Moreover, test sets are often domain-specific, such as newswire texts, which do not reflect everyday language use, limiting the generalizability of evaluation results. Another difficulty arises from the limitations of automatic metrics. For example, BLEU scores are widely used for machine translation but may not capture semantic adequacy or fluency in morphologically rich languages. Similarly, Word Error Rate (WER) for speech recognition can be misleading when orthography is inconsistent or when there is no agreed-upon spelling standard. In these cases, human evaluation—where native speakers assess outputs for fluency, adequacy, or usability—remains the gold standard, though it is costly and time-consuming.

Case studies from recent community-driven efforts illustrate both progress and limitations. In the Masakhane project, machine translation systems for several African languages were evaluated using BLEU and chrF scores, but the project also emphasized human evaluations to better understand the cultural and linguistic nuances of the outputs. In another example, the Common Voice corpus enabled the training of automatic speech recognition systems for underrepresented languages. While early results showed promising accuracy levels, community-led testing revealed significant performance drops in spontaneous, conversational speech compared to scripted recordings, highlighting the gap between benchmark results and practical usability. These examples underscore the importance of adopting multi-dimensional evaluation practices. A robust strategy often combines automatic metrics with human evaluation, supplemented by error analysis that highlights where systems succeed or fail. For low-resource languages, even small-scale case studies are invaluable, as they reveal practical issues such as domain mismatch, dialectal coverage, or cultural appropriateness that would otherwise go unnoticed in purely quantitative evaluations.

In summary, evaluation in low-resource NLP is not just a technical necessity but also a community-centered process. By involving native speakers in both designing evaluation protocols and interpreting results, researchers can achieve assessments that are both scientifically rigorous and socially meaningful.

6. Conclusion

Natural Language Processing for low-resource languages is both a pressing challenge and an opportunity for impactful innovation. While remarkable progress has been made in developing techniques such as transfer

learning, multilingual pretraining, and data augmentation, these methods are not a substitute for the creation of high-quality linguistic resources and active community participation. The scarcity of annotated corpora, orthographic variation, and lack of standardized benchmarks continue to slow down progress, but collaborative and community-driven initiatives have shown that these barriers can be gradually overcome.

This article highlighted the causes of resource scarcity, the range of methodologies applied to address them, and the importance of carefully designed evaluation frameworks. The role of community involvement was emphasized throughout, as sustainable progress depends on empowering native speakers to contribute to resource creation, annotation, and system evaluation. Moreover, case studies such as Masakhane and Common Voice demonstrate that grassroots collaboration can yield systems that are both linguistically robust and culturally appropriate.

Looking ahead, the future of NLP in low-resource languages lies in **inclusive research practices** that integrate computational advances with linguistic and cultural expertise. By promoting open-source resources, shared benchmarks, and participatory evaluation, researchers and communities can together ensure that technological development benefits speakers of all languages, not just those with abundant data. In doing so, NLP research can move closer to its goal of universal accessibility, ensuring digital equality and preserving linguistic diversity for generations to come.

References

[1] A. Conneau et al., “Unsupervised cross-lingual representation learning at scale,” in Proc. ACL, 2020, pp. 8440–8451.

[2] S. Ruder, I. Vulić, and A. Søgaard, “A survey of cross-lingual word embedding models,” J. Artif. Intell. Res., vol. 65, pp. 569–631, 2019.

[3] J. Tiedemann, “Parallel data, tools and interfaces in OPUS,” in Proc. LREC, 2012, pp. 2214–2218.

[4] M. Nekoto et al., “Participatory research for low-resourced machine translation: A case study in African languages,” Findings of ACL: EMNLP 2020, pp. 2144–2160.

[5] K. Heffernan, A. Salesky, and A. Post, “Bitext mining using distant supervision for low-resource languages,” in Proc. NAACL-HLT, 2021, pp. 3617–3629.

[6] J. Schneider et al., “Common Voice: A massively-multilingual speech corpus,” in Proc. LREC, 2020, pp. 4218–4226.