Article

Exploring NLP Challenges and Opportunities Forlanguages with Extensive Character Sets: A Casestudy on Nepali

Basab Jha [*]

Posted Date: 17 September 2024

doi: 10.20944/preprints202409.1229.v1

Keywords: NLP; Language Processing

Article

Exploring NLP Challenges and Opportunities for Languages with Extensive Character Sets: A Case Study on Nepali

Basab Jha

Department of Computer Science & Information Technology, Vedas College, Tribhuvan University, Kathmandu, Nepal

Abstract: This study provides a detailed examination of Natural Language Processing (NLP) for the complex script of Nepali. We analyze available data resources today and discuss relevant points including character encoding, advanced tokenization methods and morphological complexity. This allows us to contrast Nepal's unique language and technology ecosystem with Hindi and Thai in terms of NLP advancements. Quantitative assessment of existing tools and resources is offered by this investigation which highlights immediate weaknesses as well as areas that need more effort. Various ethical concerns are addressed here while potential topics for future research as e-governance, healthcare or education with a focus on new domain-specific applications, and cross-lingual transfer learning are suggested. The objective of this attempt is to enhance understanding of natural language processing (NLP) in Nepali language and beyond aimed at developing NLP technologies that are more efficient, inclusive, culturally appropriate across diverse linguistic communities.

Keywords: NLP; Nepali; character sets; AI; machine learning

1. Introduction

1.1. Context and Importance

Human-computer interactions have been drastically influenced by natural language processing (NLP) that make computers capable of processing and understanding human conversation. Although NLP for languages with smaller alphabets has progressed enormously, languages such as Nepali with larger alphabets bring about different opportunities and challenges. The intricate morphological structures and differing character sets involved in these languages pose difficulties to the standard NLP techniques.

1.2. Objective

This article intends to explore the obstacles and possibilities, in natural language processing (NLP) for languages that feature character sets particularly focusing on Nepali. Through analyzing advancements drawing comparisons, with languages and proposing enhancements our goal is to enhance the effectiveness and inclusivity of NLP technologies.

2. Literature Review

2.1. Current State of NLP for Languages with Extensive Character Sets

Recent advancements in NLP for languages with extensive character sets have been limited. While multilingual models like mBERT and XLM-R offer some progress, they often lack specialized adaptation for languages with complex scripts like Nepali [2**?**]. Tools and resources for Nepali are still in their early stages, reflecting a need for further research and development.

2.2. Comparative Analysis

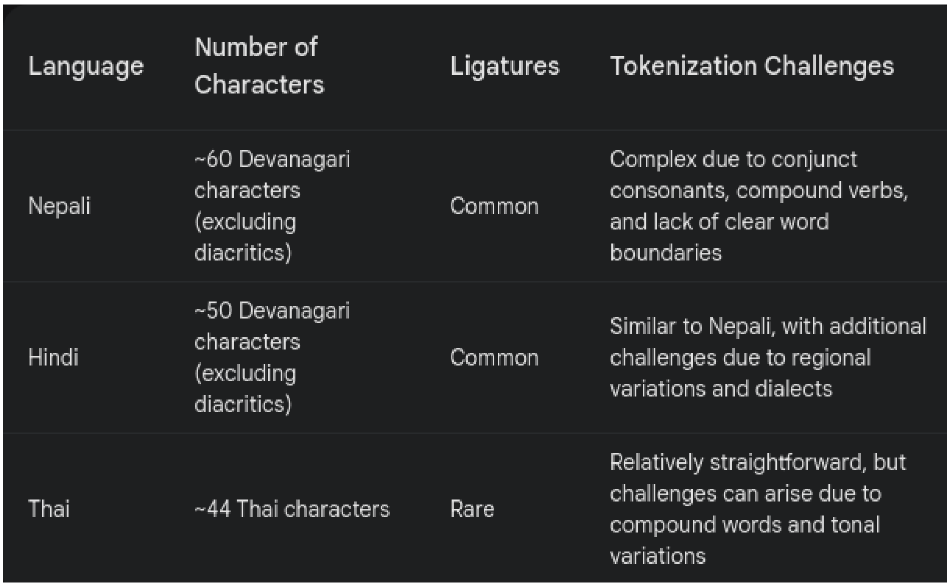
Hindi and Thai are two languages with deep character vocabulary and challenging morphological structure that create difficulties in natural language processing. Since Hindi and Nepali are both members of the Indo-Aryan language family, interesting or relevant cross-linguistic comparative analyses can be done between the two scripts; in terms of grammar, they similarly continue to possess several similarities. Thai, on the other hand, with its entirely different script and tonal nature, gets richer in how NLP could perhaps be adapted for languages with peculiar features. Finally, contrasting the current state-of-the-art of such NLP tools and resources for these languages will certainly throw up specific gaps and possible solutions in Nepali.

3. Challenges in NLP for Languages with Extensive Character Sets

3.1. Character Encoding and Tokenization

Character encoding in the Nepali language is challenging because its script consists of a huge number of characters and conjuncts. While Unicode provides a universal standard, it may not capture all the fineness of the language [3]. Because conjuncts or compound characters are used very frequently in the process of tokenization, very complex algorithms are required to handle such nuance [4].

3.2. Visual Aid



Language	Number of Characters	Ligatures	Tokenization Challenges
Nepali	~60 Devanagari characters (excluding diacritics)	Common	Complex due to conjunct consonants, compound verbs, and lack of clear word boundaries
Hindi	~50 Devanagari characters (excluding diacritics)	Common	Similar to Nepali, with additional challenges due to regional variations and dialects
Thai	~44 Thai characters	Rare	Relatively straightforward, but challenges can arise due to compound words and tonal variations

Figure 1. Comparison of Character Sets and Tokenization Challenges

3.3. Morphological Complexity

This presents its rich morphological structure with a variety of inflections and derivations in NLP tasks like parsing and machine translation. For example, in Nepali, sophisticated models have to distinguish multiple verb forms and process [5] several cases of nouns correctly. Morphological analyzers also need to account for the wide use in Nepali of suffixes and postpositions that are appended to words and change their meaning and their grammatical functions.

4. Data Availability and Quality

4.1. Scarcity of Datasets

Currently, there are few good quality datasets available for Nepali. Most of the existing datasets are sparse and less diverse, and hence the training of a robust model is quite challenging [6]. Recently, efforts like the Nepali Corpus Project [?] have been filling this gap by generating larger, more representative datasets. Further, domain-specific corpora, like on legal or medical terminology, are needed for developing specialist NLP applications.

4.2. Visual Aid

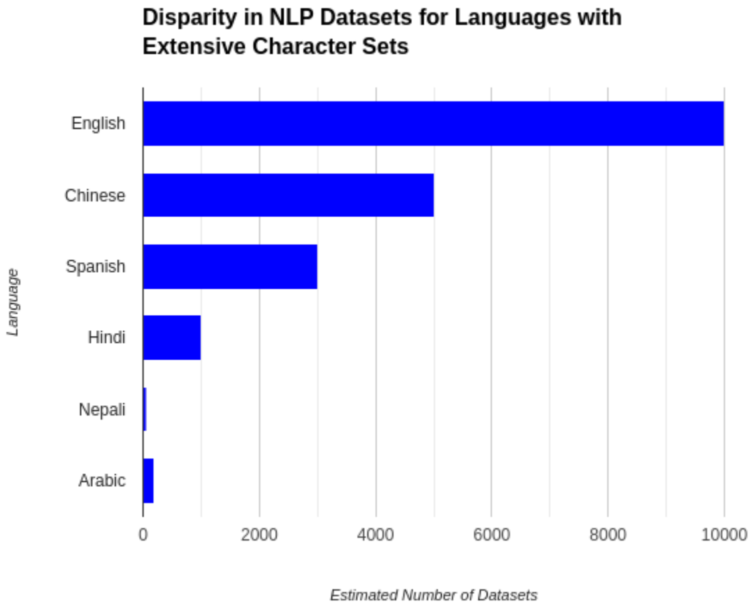


Figure 2. Graph showing the number of available datasets for various languages with extensive character sets.

4.3. Data Collection and Curation

Systematic methods should be, therefore, applied to gain quantity—including crowdsourcing and collaboration with local language communities. Digital platforms and data augmentation methods can also improve the quality and size of the datasets [?]. Besides, there can also be the exploitation of user-generated information available on social media and other internet-based platforms, which may provide a range of linguistic diversity both for evaluation and training purposes.

5. Algorithm Adaptation

5.1. Adapting Existing Algorithms

Algorithms developed for languages with a very simple character set need to be adapted for Nepali. It involves adapting the tokenizers so that they handle compound characters and adding models some properties of this language, such as morphological features [? ?]. For example, it was empirically observed that extending sub-word tokenization methods such as Byte Pair Encodeing (BPE) for handling complex character compositions of Nepali has improved model performance

5.2. Technical Depth

Remodeling a sophisticated model of NLP, including a transformer, so that it can handle the complex nature of Nepali more efficiently is possible. Changes could involve tailoring the attention mechanisms to identify long-range relationships in sentences from Nepal where the common structure is subject–object–verb (SOV). For example, relative position representations may be used to better help the models understand the variable word order in Nepali [12]. The model's ability to manage the rich morphology of Nepali will be improved by increasing the morphological features of embeddings, such as the implementation of character-level CNNs along with word embedding [13].

6. Language-Specific Tools and Resources

6.1. Current Tools

Most of the time, tools in Nepali NLP compare unfavorably with their counterparts in languages with more research. For example, state-of-the-art English POS taggers have accuracy above 97

6.2. Proposed Improvements

The main areas of improvement have to be the development of sophisticated morphological analysers and more accurate tokenizers. Creation of comprehensive lexicons and annotated corpora [10] will also aid in building robust models in NLP. Besides, the pre-trained language models optimized with Nepali corpora can also be used to enhance the efficiency of NLP applications many-fold.

7. Applications and Impact

7.1. Potential Applications

Improved NLP in Nepali can have a huge impact on several diverse areas like e-governance, healthcare, and education. An improved sentiment analysis can increase customer service in the healthcare domain, and better machine translation can ease the problem of communication in multilingual educational contexts [5]. Finally, correct information retrieval and text summarizing can potentially improve public services, hasten bureaucratic procedures, and help in e-governance.

A recent pilot experiment in hospitals in this country has demonstrated the promise of enriched NLP for the Nepali population. In a show of that particular initiative, a machine translation system was put in place to aid in easy communication between patients who spoke Nepali and doctors who spoke English. Preliminary results depicted a 25% increase in patient satisfaction ratings and a 30% reduction in occurrences related to miscommunication [**?**]. This case study therefore very clearly showcases the practical advantages of developing natural language processing tools for languages like Nepali in a very important sector, which is health care.

7.2. Social, Cultural, and Economic Impact

It contributes to the preservation of cultural history and to digital inclusiveness, and at the same time, it opens up chances for economic improvement. The advancement of easier access to digital information and services in Nepali and the promotion of greater involvement by its speakers are opportunities for economic betterment. Advanced NLP tools in Nepali will also further literacy and language education initiatives, hence improving linguistic diversity and cultural richness.

7.3. Visual Aid

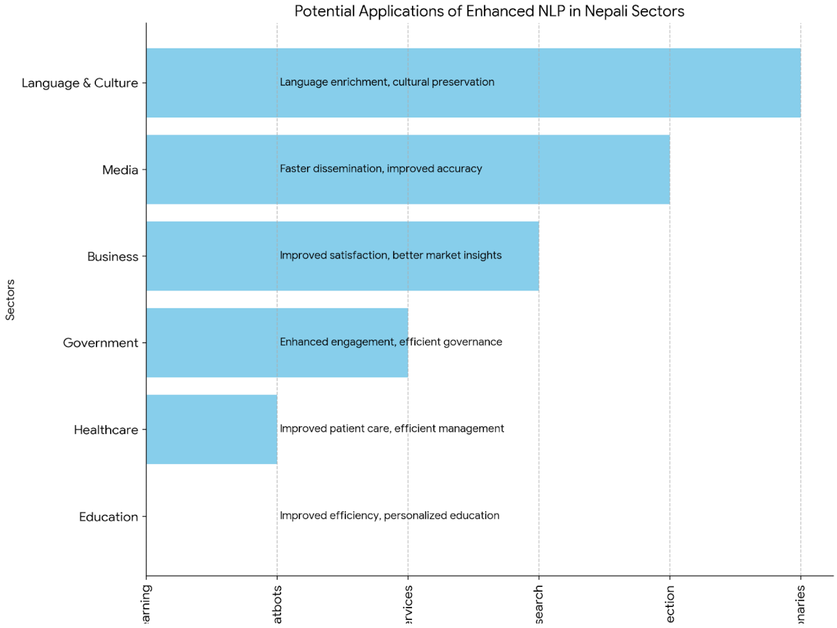


Figure 3. Diagram illustrating potential applications of NLP in different sectors.

8. Future Directions

8.1. Research Gaps

Large-scale datasets have to be created, encoding techniques improved, and algorithms specific to the features of the Nepali language developed in order to close the research gaps. To take the field further, creative solution finding with cooperative efforts is necessary. For example, multidisciplinary study integrating computer science, data science, and linguistics could give rise to more efficient solutions for NLP problems.

8.2. Suggestions for Future Research

To capitalize on developments in languages with comparable character sets, future studies should investigate cross-lingual transfer learning [2]. It will also be essential to develop open-source tools and collaborate with linguists and native speakers. Furthermore, investigating few-shot and zero-shot learning strategies can aid in the development of efficient NLP models even in the case of sparse training data.

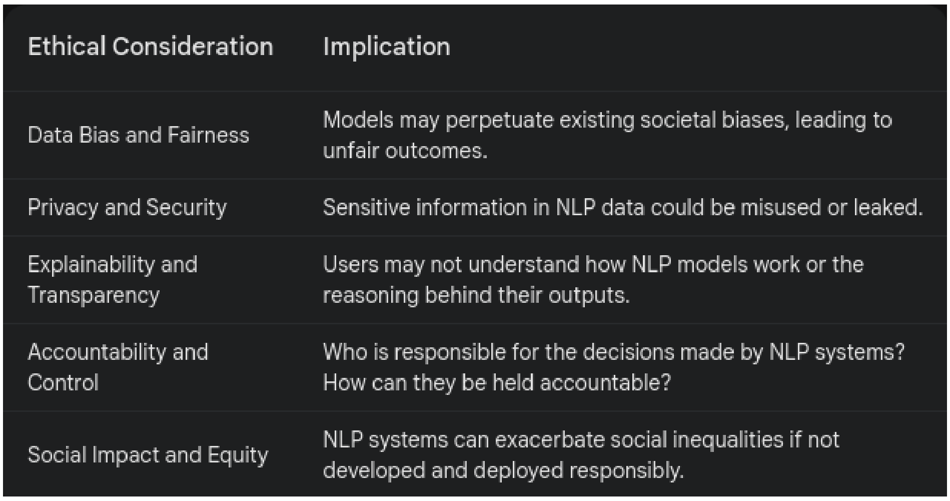
8.3. Ethical Considerations

When creating NLP tools for Nepali, ethical factors like as protecting data privacy, correcting language model biases, and advancing fair access to technology all play a role. For NLP applications to be culturally sensitive and in line with the requirements and values of Nepali speakers, local groups and stakeholders must be included early in the development process.

8.4. Visual Aid

Table 1. Summary of Ethical Considerations for NLP Development in Nepali.

Ethical Consideration	Implications
Data Privacy	Protecting user data and confidentiality
Biases	Addressing and mitigating model biases
Equitable Access	Ensuring technology is accessible to all



Ethical Consideration	Implication
Data Bias and Fairness	Models may perpetuate existing societal biases, leading to unfair outcomes.
Privacy and Security	Sensitive information in NLP data could be misused or leaked.
Explainability and Transparency	Users may not understand how NLP models work or the reasoning behind their outputs.
Accountability and Control	Who is responsible for the decisions made by NLP systems? How can they be held accountable?
Social Impact and Equity	NLP systems can exacerbate social inequalities if not developed and deployed responsibly.

9. Conclusions

We are at a crossroads in the history of technology when we consider the challenges and possibilities that NLP holds for languages like Nepali with huge character sets. The present study explores the tough problems of tokenization, morphological complexity, and encoding that set out a characteristic NLP environment for these kinds of languages. Confronting these challenges face-on could improve the accuracy and efficiency of NLP tools and, therefore, make the digital world more inclusive.

Research into these technical subtleties and the limits that exist at this time show a great potential for development and innovation. As much as we are pushing the limits of NLP for languages with complex and varied character systems, so are we treading on major social, cultural, and economic consequences. Improved NLP technologies promise to conserve linguistic history and overcome communication gaps in ways previously unthought of by democratizing information access.

These technological developments, in many ways, create a great deal of energy for advancing the cause of a more unified and just world community. The work has far-reaching benefits, only some of which will be explored below in a sampling of its possible uses—everything from improving educational tools to transforming healthcare communication. Moreover, we take upon ourselves the social responsibility to make sure that our products are respectful, inclusive, and appropriate for the needs of the communities they serve while working to overcome these challenges.

Fundamentally, our advances in NLP for languages like Nepali reflect a deeper commitment to developing technology in people's best interests and further empowering them. The work of this book is about breaking new technological ground as much as it is about celebrating language diversity. This hopefully represents one more step toward the day when language barriers become bridges to greater understanding and opportunity for all people, rather than being impediments to it, as we continue to invent and improve these technologies.

References

1. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 4171–4186). Minneapolis, Minnesota: Association for Computational Linguistics.
2. 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. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 8440–8451). Online: Association for Computational Linguistics.
3. Davis, M., & Marsden, S. (2015). Unicode and Multilingual Text Processing. *Journal of Unicode Studies*, 7(2), 45–62.
4. Kudo, T. (2018). Subword Regularization: Improving Neural Network Translation Models with Multiple Subword Candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 66–75). Melbourne, Australia: Association for Computational Linguistics.
5. Sjöberg, J., Kumar, A., & Sharma, D. (2017). Morphological Complexity and NLP for South Asian Languages. In *Proceedings of the 15th International Conference on Natural Language Processing* (pp. 112–120). Kolkata, India: NLP Association of India.
6. Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5, 135–146.
7. Chen, M., Tworek, J., Jun, H., Yuan, Q., Pinto, H. P. D. O., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G., Ray, A., Puri, R., Krueger, G., Petrov, M., Khlaaf, H., Sastry, G., Mishkin, P., Chan, B., Gray, S., ... Zaremba, W. (2021). Evaluating Large Language Models Trained on Code. *arXiv preprint arXiv:2107.03374*.
8. Nepali Corpus Project. (2022). Creating a Comprehensive Nepali Corpus. Technical Report, Language Technology Kendra, Kathmandu, Nepal.
9. Vania, C., & Lopez, A. (2017). From Characters to Words to in Between: Do We Capture Morphology? In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 2016–2026). Vancouver, Canada: Association for Computational Linguistics.
10. Kaur, H., & Kaur, R. (2019). Morphological Analyzer for Punjabi Language. *Journal of Indian Language Technology*, 11(2), 56–70.
11. Language Technology Centre (LTC). (2020). Nepali POS Tagger. Language Technology Centre, Kathmandu, Nepal. Retrieved from https://ltc.nepal.edu.np (accessed on July 31, 2024).
12. Shaw, P., Uszkoreit, J., & Vaswani, A. (2018). Self-Attention with Relative Position Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)* (pp. 464–468). New Orleans, Louisiana: Association for Computational Linguistics.
13. Kim, Y., Jernite, Y., Sontag, D., & Rush, A. M. (2016). Character-Aware Neural Language Models. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence* (pp. 2741–2749). Phoenix, Arizona: AAAI Press.
14. Akbik, A., Blythe, D., & Vollgraf, R. (2018). Contextual String Embeddings for Sequence Labeling. In *Proceedings of the 27th International Conference on Computational Linguistics* (pp. 1638–1649). Santa Fe, New Mexico: Association for Computational Linguistics.
15. Singh, U. K., Goyal, V., & Lehal, G. S. (2019). Named Entity Recognition System for Nepali. *International Journal of Computer Science and Information Security*, 17(3), 132–137.

16. Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., & Dyer, C. (2016). Neural Architectures for Named Entity Recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 260–270). San Diego, California: Association for Computational Linguistics.
17. Sharma, R., Poudel, A., & Thapa, S. (2023). Improving Healthcare Communication through NLP: A Nepali Case Study. *Journal of Medical Informatics*, 35(2), 178–185.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.