Towards Building a Standard Benchmark for Low-Resource Nepali Information Retrieval

Praveen Acharya
praveen.acharya2@mail.dcu.ie
Dublin City University
Dublin, Ireland

Bal Krishna Bal
bal@ku.edu.np
Kathmandu University
Dhulikhel, Nepal

Abstract

Despite being spoken by over 26 million people, Nepali lacks a standardized Information Retrieval (IR) benchmark, preventing systematic evaluation of retrieval models in this low-resource setting. While progress has been made in several Nepali NLP tasks, no publicly available test collection with queries and relevance judgments exists for ad-hoc retrieval task. We propose the creation of a first standardized Nepali IR benchmark supporting monolingual (Nepali queries → Nepali documents), cross-lingual (English queries → Nepali documents), and code-mixed (Nepali-English queries → Nepali documents) retrieval. We argue that Nepali provides a compelling testbed for studying morphology-aware retrieval, code-mixed query processing, and multilingual embedding transfer.

CCS Concepts

• **Information systems** → **Multilingual and cross-lingual retrieval**; **Test collections**; *Relevance assessment*; **Specialized information retrieval**; **Structure and multilingual text search**; **Evaluation of retrieval results**.

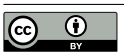
Keywords

Nepali, Low-resource, Nepali Information Retrieval, Ad-hoc Retrieval

ACM Reference Format:
Praveen Acharya and Bal Krishna Bal. 2026. Towards Building a Standard Benchmark for Low-Resource Nepali Information Retrieval. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '26), July 20–24, 2026, Melbourne, VIC, Australia.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3805712.3808533

1 Introduction

Low-resource languages remain underrepresented in Information Retrieval (IR) research, primarily due to the lack of standardized evaluation benchmarks. Although multilingual pretrained models have improved cross-lingual transfer, systematic evaluation is not possible without test collections that include queries and relevance judgments. Nepali, the official language of Nepal, is spoken by approximately 26 million people [18]. While tools such as stemmers, morphological analyzers, POS taggers, and transformer-based pretrained embedding models are available for Nepali NLP tasks, there is no standardized benchmark for ad-hoc IR evaluation. Consequently, the effectiveness of retrieval models for Nepali remains unknown. The absence of an evaluation dataset prevents systematic analysis of retrieval models, as well as multilingual retrieval under structural and data constraints. We therefore propose creating a standardized Nepali IR benchmark to evaluate the effectiveness and robustness of various retrieval models.



This work is licensed under a Creative Commons Attribution 4.0 International License.
SIGIR '26, July 20–24, 2026, Melbourne, VIC, Australia
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2599-9/2026/07
https://doi.org/10.1145/3805712.3808533

2 Challenges and Opportunities

2.1 Scarce/Limited Data

Research into language technologies for the Nepali language faces a major challenge due to the lack of language resources. This is problematic for the development of local IR technologies in Nepali. Although some domain-specific resources are available for other natural language processing (NLP) tasks (*Section 2.3*), they cannot be used for IR research due to the absence of query sets and relevance judgments.

2.2 Linguistic Diversity

Nepali language linguistic diversity, shaped by culture, geography, and more recently, social media and the web, has led to increased code-mixing in written texts [15, 20]. The lack of equivalent vocabulary in some cases further complicates the development of computational tools. As the need for systems that can process multiple languages and scripts grows, researchers must create tools that can handle diverse linguistic forms, including code-mixed content. Addressing this challenge requires building computational capabilities to support this linguistic variety.

2.3 Limited Availability of Technological Resources

The linguistic diversity and distinct characteristics of the Nepali language pose substantial challenges for processing digital content [14]. Limited domain-specific tools exist, such as morphological analyzers, stemmers [5] and Part-Of-Speech (POS) tagsets [23], with existing research in areas including sentiment analysis [2, 13], machine translation [1, 9, 12] and domain-specific machine translation [21]. Shahi and Sitaula [26] presents an overview of the NLP research for the Nepali language. Advanced technologies such as Large Language Models (LLMs) are also being studied [17, 28], with applications in POS and Named-Entity-Recognition (NER) [27]. Some of these technological resources, such as stemmers, morphological analyzers, and stopword lists, could potentially be helpful for IR research. However, since there is no evidence of their use in Nepali IR research so far, it is uncertain whether they are truly beneficial, and if they are, to what extent.

3 Nepali as a Low-Resource Testbed

Nepali exhibits linguistic and structural properties that make it a meaningful stress-test for retrieval systems in low-resource conditions.

Morphological variation: Nepali language has rich inflectional and derivational morphology, leading to significant variation in word forms derived from the same root [6, 22]. A standardized benchmark would allow controlled evaluation of morphology-aware normalization techniques, including stemming and morphological analysis, and their effect on retrieval effectiveness.

Script and representation gap: Nepali is written in the Devanagari script and remains under-represented in large multilingual pretraining corpora. Consequently, embedding alignment and zero-shot transfer reliability may differ from high-resource languages. Evaluating retrieval under these conditions offers insights into the robustness of multilingual encoders and dense retrievers trained primarily on high-resource data.

Code-mixing: Digital Nepali content frequently exhibit code-mixing with English [15, 20]. A benchmark incorporating controlled code-mixed queries would allow for systematic evaluation of retrieval models under code-mixed linguistic signals.

4 Scope of Research and Questions

We propose to investigate ad-hoc document retrieval for Nepali, a low-resource language. Our study covers three retrieval settings: *(1) monolingual* (Nepali queries → Nepali documents), *(2) cross-lingual* (English queries → Nepali documents), and *(3) code-mixed* (Nepali–English queries → Nepali documents). Across these settings, we propose to systematically evaluate and compare sparse, dense, and hybrid retrieval methods to assess their effectiveness.

RQ: How do sparse, dense, and hybrid retrieval methods perform in Nepali ad-hoc retrieval across monolingual, cross-lingual, and code-mixed queries?

5 Proposed Benchmark Construction

5.1 Positionality Statement

This work is informed by the authors' familiarity with the Nepali language and access to available data sources and annotation resources, which may influence certain design choices in dataset construction. We also recognize that Nepali exhibits influences resulting from ethnic, linguistic, and cultural diversity, making it challenging to fully capture its variation within a single benchmark.

5.2 Data Collection and Corpus Development

A key challenge in developing IR systems for Nepali is the scarcity of large-scale, high-quality datasets. Efforts are underway to consolidate benchmark datasets for language understanding tasks [19] and to develop question–answer datasets [7]. However, constructing a comprehensive IR test collection set with a large number of documents, query sets, and relevance judgments remains beyond the available resources.

One potential solution is to use automatic approaches: a short-term strategy [4] to bootstrap the creation of initial resources with the long-term goal of creating a standard evaluation benchmark dataset with human-annotated query sets and relevance judgments. For example, Frej et al. [11] proposed a framework for creating IR datasets from resources like Wikipedia, which could be adapted for Nepali by leveraging resources such as Nepali Wikipedia and other public domain materials. This approach has been successfully used in languages other than English [10]. Furthermore, recent advancements in multilingual IR [25, 29, 30] and cross-lingual IR [3, 8, 16] using pre-trained LLMs, enabling transfer learning, could also be explored.

5.2.1 Document Collection. We propose collecting a 50K–100K document corpus from online Nepali news archives, government publications, Nepali Wikipedia articles, and public domain materials. The documents collected will be deduplicated and normalized for consistency across different formats.

5.2.2 Topic Collection. We propose to manually create a set of 30–40 information needs. The queries will be designed to reflect informational and exploratory needs across diverse domains (e.g. politics, health, education, law, and culture). In the absence of large-scale user query logs, the queries will be manually curated and validated by fluent Nepali speakers. For cross-lingual evaluation, the English queries will be created by expert translators who will translate the Nepali queries.

5.2.3 Relevance Judgments. Collecting relevance judgments is resource intensive. To address this, we propose two complementary approaches: *(1) a Semi-automatic approach*, used to bootstrap the relevance annotation process and reduce initial manual effort, and *(2) a Pooling-based manual annotation approach.* Query–document pairs will be assessed and assigned graded relevance scores (e.g., on a 1–4 scale) by subject matter experts fluent in Nepali (e.g. university professors, graduate and senior undergraduate students).

Semi-automatic approach: Adopt the methodology of Frej et al. [11] to generate graded relevance labels. Alternatively, use an LLM-based approach, followed by a manual evaluation of the assigned annotation, to create graded relevance labels [24].

Pooling-based approach: The document and topic collections can be released as part of evaluation tracks at IR workshops (e.g., FIRE). The top-ranked documents from multiple retrieval systems will be pooled to form a candidate set, and this pooled set will then be manually annotated to obtain graded relevance judgments.

6 Expected Contributions

This work is expected to deliver the first standardized Nepali IR benchmark dataset. It will provide baseline comparisons of various retrieval models, along with empirical insights into morphology-aware and cross-lingual retrieval in Nepali. The research will also release open-source resources to support future IR research in Nepali. Beyond Nepali, the proposed framework will serve as a blueprint for IR evaluation in other under-represented languages of Nepal.

7 Conclusion

The absence of a standardized IR benchmark has limited retrieval research for Nepali despite its large speaker base and growing digital presence. We propose a structured, reproducible evaluation framework for monolingual and cross-lingual retrieval, enabling systematic comparison of sparse, dense, and hybrid approaches. By establishing foundational infrastructure and defining clear research

questions, this initiative aims to catalyze low-resource IR research in Nepal and contribute broader insights into multilingual retrieval robustness.

Acknowledgments

This work was conducted with the financial support of Taighde Éireann – Research Ireland under Grant No. 18/CRT/6223. The authors acknowledge the peoples of the Woi Wurrung and Boon Wurrung language groups of the eastern Kulin Nation on whose unceded lands ACM SIGIR 2026 was hosted. We pay our respects to their Elders past and present, and extend that respect to all Aboriginal and Torres Strait Islander peoples today and their continuing connection to land, sea, sky, and community.

References

[1] Praveen Acharya and Bal Krishna Bal. 2018. A Comparative Study of SMT and NMT: Case Study of English-Nepali Language Pair. In *Proc. SLTU 2018*. 90–93.

[2] Abiral Adhikari, Prashant Manandhar, Reewaj Khanal, Samir Wagle, Praveen Acharya, and Bal Krishna Bal. 2024. Profanity and Offensiveness Detection in Nepali Language Using Bi-directional LSTM Models. In *Proceedings of the 21st International Conference on Natural Language Processing (ICON)*. 515–521.

[3] Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the association for computational linguistics* 7 (2019), 597–610.

[4] Bal Krishna Bal, Balaram Prasain, Rupak Raj Ghimire, and Praveen Acharya. 2024. Strategies for Corpus Development for Low-Resource Languages: Insights from Nepal. *Automatic Speech Recognition and Translation for Low Resource Languages* (2024), 297–330.

[5] Bal Krishna Bal and Prajol Shrestha. 2004. A morphological analyzer and a stemmer for Nepali. *PAN Localization, working papers* 2007 (2004), 324–31.

[6] Santa B Basnet and Shailesh B Pandey. [n. d.]. Morphological analysis of verbs in Nepali. *Nepalese Linguistics* 24 ([n. d.]), 21–30.

[7] Funghang Limbu Begha, Praveen Acharya, and Bal Krishna Bal. 2026. Nepali Passport Question Answering: A Low-Resource Dataset for Public Service Applications. *arXiv preprint arXiv:2603.13320* (2026).

[8] Marta R Costa-Jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672* (2022).

[9] Sharad Duwal and Bal Krishna Bal. 2019. Efforts in the development of an augmented english–nepali parallel corpus. In *Proceedings of the 1st International Conference on Language Technologies for All*. European Language Resources Association Paris, France, 375–378.

[10] Jibril Frej, Didier Schwab, and Jean-Pierre Chevallet. 2020. MLWIKIR: A Python toolkit for building large-scale Wikipedia-based Information Retrieval Datasets in Chinese, English, French, Italian, Japanese, Spanish and more. In *Joint Conference of the Information Retrieval Communities in Europe*.

[11] Jibril Frej, Didier Schwab, and Jean-Pierre Chevallet. 2020. WIKIR: A Python Toolkit for Building a Large-scale Wikipedia-based English Information Retrieval Dataset. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*. 1926–1933.

[12] Rupak Raj Ghimire, Bipesh Subedi, Balaram Prasain, Prakash Poudyal, Praveen Acharya, Nischal Karki, Rupak Tiwari, Rishikesh Kumar Sharma, Jenny Poudel, and Bal Krishna Bal. 2026. NepTam: A Nepali-Tamang Parallel Corpus and Baseline Machine Translation Experiments. In *Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, Stelios Piperidis, Núria Bel, Henk van den Heuvel, Nancy Ide, Simon Krek, and Antonio Toral (Eds.). European Language Resources Association (ELRA), Palma, Mallorca, Spain, 8850–8861. https://doi.org/10.63317/37edei5qcjb3

[13] Chandan Prasad Gupta and Bal Krishna Bal. 2015. Detecting sentiment in nepali texts: A bootstrap approach for sentiment analysis of texts in the nepali language. In *2015 international conference on cognitive computing and information processing (ccip)*. IEEE, 1–4.

[14] Pat Hall, Bal Krishna Bal, Sagun Dhakhwa, and Bhim Narayan Regmi. 2014. Issues in encoding the writing of Nepal's languages. In *Computational Linguistics and Intelligent Text Processing: 15th International Conference, CICLing 2014, Kathmandu, Nepal, April 6-12, 2014, Proceedings, Part I 15*. Springer, 52–67.

[15] Jang Sen Thakuri Khadga Chandra. 2012. *English Code-Mixing in The Nepal Samacharpatra Daily*. Ph. D. Dissertation. Central department of English Education.

[16] Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, et al. 2021. Muril: Multilingual representations for indian languages. *arXiv preprint arXiv:2103.10730* (2021).

[17] Utsav Maskey, Manish Bhatta, Shiva Bhatt, Sanket Dhungel, and Bal Krishna Bal. 2022. Nepali encoder transformers: An analysis of auto encoding transformer language models for nepali text classification. In *Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages*. 106–111.

[18] National Statistics Office. 2025. Languages in Nepal. Kathmandu: National Statistics Office. (National Population and Housing Census 2021).

[19] Jinu Nyachhyon, Mridul Sharma, Prajwal Thapa, and Bal Krishna Bal. 2024. Consolidating and Developing Benchmarking Datasets for the Nepali Natural Language Understanding Tasks. *arXiv preprint arXiv:2411.19244* (2024).

[20] Sagar Poudel. 2019. Code-Mixing and literal translation in Nepal's English newspapers. *Journal of NELTA* 24, 1-2 (2019), 191–203.

[21] Shabdapurush Poudel, Bal Krishna Bal, and Praveen Acharya. 2024. Bidirectional English-Nepali machine translation (MT) system for legal domain. In *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages@ LREC-COLING 2024*. 53–58.

[22] Balaram Prasain. 2011. *A computational analysis of Nepali morphology: A model for natural language Processing*. Ph. D. Dissertation. Faculty of Linguistics.

[23] B Prasain, LP Khatiwada, BK Bal, and P Shrestha. 2008. Part-of-speech Tagset for Nepali. *Madan Puraskar Pustakalaya* (2008).

[24] Hossein A Rahmani, Emine Yilmaz, Nick Craswell, Bhaskar Mitra, Paul Thomas, Charles LA Clarke, Mohammad Aliannejadi, Clemencia Siro, and Guglielmo Faggioli. 2024. Llmjudge: Llms for relevance judgments. *arXiv preprint arXiv:2408.08896* (2024).

[25] Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084* (2019).

[26] Tej Bahadur Shahi and Chiranjibi Sitaula. 2022. Natural language processing for Nepali text: A review. *Artificial Intelligence Review* 55, 4 (2022), 3401–3429.

[27] Bipesh Subedi, Sunil Regmi, Bal Krishna Bal, and Praveen Acharya. 2024. Exploring the potential of large language models (LLMs) for low-resource languages: A study on named-entity recognition (NER) and part-of-speech (POS) tagging for Nepali language. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. 6974–6979.

[28] Prajwal Thapa, Jinu Nyachhyon, Mridul Sharma, and Bal Krishna Bal. 2025. Development of Pre-Trained Transformer-based Models for the Nepali. *COLING 2025* (2025), 9.

[29] Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual e5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672* (2024).

[30] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mT5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934* (2020).