Roman Urdu as a Low-Resource Language: Building the First IR Dataset and Baseline

Umer Butt[1,2,3] **Stalin Varanasi**[1,2] **Günter Neumann**[1,2]
[1]Saarland Informatics Campus, D3.2, Saarland University, Germany
[2] German Research Center for Artificial Intelligence (DFKI), Saarbrücken, Germany
[3] Sequire technology GmbH, Saarbrücken, Germany
umer.butt@sequire.de, stalin.varanasi@dfki.de, guenter.neumann@dfki.de

Abstract

The field of Information Retrieval (IR) increasingly recognizes the importance of inclusivity, yet addressing the needs of low-resource languages, especially those with informal variants, remains a significant challenge. This paper addresses a critical gap in effective IR systems for Roman Urdu, a romanized version of Urdu i.e a language with millions of speakers, widely used in digital communication yet severely underrepresented in research and tooling. Roman Urdu presents unique complexities due to its informality, lack of standardized spelling conventions, and frequent code-switching with English. Crucially, prior to this work, there was a complete absence of any Roman Urdu IR dataset or dedicated retrieval work. To address this critical gap, we present the first-ever large-scale IR MS-marco translated dataset specifically for Roman Urdu, created through a multi-hop pipeline involving English-to-Urdu translation followed by Urdu-to-Roman Urdu transliteration. Using this novel dataset, we train and evaluate a multilingual retrieval model, achieving substantial improvements over traditional lexical retrieval baselines (**MRR@10**: **0.19** vs. 0.08; **Recall@10**: **0.332** vs. 0.169). This work lays foundational benchmarks and methodologies for Roman Urdu IR especially using the transformer based models, significantly contributing to inclusive information access and setting the stage for future research in informal, Romanized, and low-resource languages.

1 Introduction

Advancements in Information Retrieval (IR) have predominantly served high-resource languages, largely due to the availability of extensive training data and well-optimized models. As a result, informal and low-resource languages remain largely excluded from the benefits of modern IR systems.

Urdu is spoken by over 70 million people in South Asia and remains an important medium for written and verbal communication, especially in Pakistan and parts of India. Despite its widespread use, Urdu is underrepresented in digital language technologies due to challenges such as its Perso-Arabic script, right-to-left writing direction, and complex morphology, issues that are less severe in high-resource languages like Arabic or Chinese due to better tooling and research support.

Alongside standard Urdu, Roman Urdu (Urdu written in the Latin script) has become the dominant form of informal communication on platforms like Instagram, WhatsApp, and social media. Its popularity stems from practical constraints, such as the lack of easy-to-use Urdu keyboards and familiarity with Latin characters. (Safdar et al., 2020) However, Roman Urdu poses its own set of challenges, including inconsistent spelling, informal grammar, and frequent code-switching, making it especially difficult for information retrieval (IR) systems.

A key barrier in developing effective IR systems for both Urdu and Roman Urdu is the lack of large-scale, labeled datasets. Manual creation is often impractical, and while machine translation offers a scalable alternative, it can introduce semantic drift or misalignment. Recent work has begun addressing these issues for Urdu, but Roman Urdu remains largely overlooked. This work addresses the gap by constructing the first large-scale Roman Urdu IR benchmark. Our approach builds upon prior efforts in multilingual IR and transliteration. Following the methodology introduced in multilingual mMARCO (Nguyen et al., 2016a), we begin by translating the English MS MARCO dataset into Urdu as described by (Butt et al., 2025a), using a state-of-the-art translation model IndicTrans2 (Ramesh et al., 2022). To convert this Urdu data into Roman Urdu, we leverage a high-accuracy transliteration model as proposed in (Butt et al., 2025b) to outperform traditional approaches us-

Proceedings of the 1st Workshop on Advancing NLP for Low-Resource Languages associated with RANLP 2025, pages 82–87, Varna, Bulgaria, Sep 13, 2025.
https://doi.org/10.26615/978-954-452-100-4-009

ing transformer-based architectures and masked language modeling. Note that we go through this hopping process because there does not exist any open access model for direct translation of English to Roman-Urdu.

Our main contributions are:

- **Construction of the First Roman Urdu IR Dataset:** We generate a large-scale Roman Urdu version of MS MARCO via a multi-step translation and transliteration pipeline, maintaining semantic alignment with the original data.
- **Development of a Roman Urdu IR Model:** We fine-tune a multilingual IR model on the new dataset and demonstrate that it significantly outperforms the baseline, which struggles with informal and inconsistent spellings in Roman Urdu.
- **Scalable and Reusable Methodology:** Our approach provides a practical framework that can be adapted for other low-resource or Romanized scripts facing similar linguistic challenges.
- **Public Release of Resources:** To support future research, we make our Roman Urdu dataset, fine-tuned model, and code publicly available on Hugging Face and GitHub.[1] [2] [3]

2 Background on Romanization & Roman Urdu

Many multilingual communities, particularly across South Asia and Africa, use the Latin script (i.e., English alphabet) to write their native languages, a process known as romanization. This practice emerged from early limitations in computing and mobile technologies, where keyboards and software lacked support for non-Latin scripts. As a result, speakers of languages such as Urdu, Hindi, Bengali, and Arabic began using Roman characters to represent their native words. The trend was further reinforced by the global rise of the internet, where English remains dominant, making Romanized writing a convenient and accessible alternative for digital communication.

[1] https://huggingface.co/Mavkif/roman-urdu-mt5-mmarco

[2] https://huggingface.co/datasets/Mavkif/roman-urdu-msmarco-dataset

[3] https://github.com/UmerTariq1/MS-Marco-Translation-and-IR

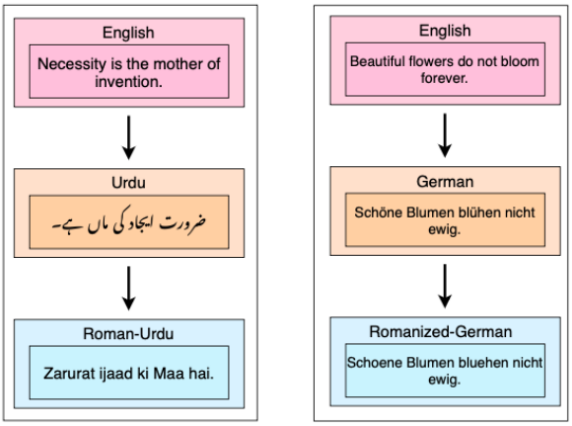

Figure 1: Examples of Romanization in English → Urdu → Roman-Urdu and English → German → Romanized-German.

Among these, Roman Urdu has seen especially widespread use across social media, e-commerce, online news, and informal messaging. Unlike standardized Latin-based scripts, however, Roman Urdu lacks any formal spelling conventions, leading to highly inconsistent, user-dependent, and phonetic spellings. The same word may appear in multiple forms based on how a speaker hears or pronounces it.

This lack of orthographic standardization poses significant challenges for NLP and IR models, which must account for noisy spelling, informal grammar, and frequent code-switching with English. As a result, building robust Roman Urdu datasets and retrieval models requires not just large-scale training data, but approaches that can handle spelling variation and contextual ambiguity in a low-resource setting. An example of a sentence in English, Urdu and Roman-Urdu and an an example of a sentence in English, Germany and Hypothethical Romanized-German language (to show example of how it would look) is given in 1

3 Related Work

Roman Urdu, despite being widely used online, has remained almost completely absent from retrieval research. Prior efforts have focused mostly on sentiment classification or dictionary creation (Smat26, 2023; Zahid et al., 2020), with no standardized datasets or IR models available for this variant. The lack of relevance-labeled resources and the informal nature of the script, non-standard spelling, code-switching with English, and inconsistent grammar, pose serious challenges for build-

ing retrieval models. This work is the first to directly address these challenges at scale. (Safdar et al., 2020)

In contrast, Urdu IR has recently seen progress through translation-based methods. (Butt et al., 2025a) translated the MS MARCO dataset (Nguyen et al., 2016b) into Urdu using IndicTrans2, producing over 8.8 million passages and 500,000+ queries. This dataset enabled fine-tuning of a multilingual mT5 reranker (Xue et al., 2021) and showed substantial performance gains over zero-shot and BM25 (Robertson and Walker, 1994) baselines. IndicTrans2 was chosen for its strong performance over Google Translate and OPUS-MT (Tiedemann, 2012), achieving a chrF++ score of 68.2.

Transliteration has also been studied in Roman Urdu, mainly using the Roman-Urdu-Parl corpus (Alam and Hussain, 2022). Early models used RNNs (Elman, 1990), but recent work by (Butt et al., 2025a) introduced a transformer-based transliteration model with MLM pretraining, significantly improving cross-domain robustness. That model forms the core of the Urdu→Roman Urdu step in our multi-hop pipeline.

Although multilingual IR research has produced general-purpose models like mBERT (Devlin et al., 2018), XLM-R (Conneau, 2019), LaBSE (Feng et al., 2020), and mT5 (Xue et al., 2021), these models struggle on underrepresented scripts like Roman Urdu due to minimal pretraining exposure. Efforts like mMARCO translated MS MARCO into 13 languages (Nguyen et al., 2016a) but excluded Urdu and Roman Urdu. Other multilingual IR benchmarks (e.g., MIRACL (Yu et al., 2021), Mr.TyDi (Clark et al., 2020)) offer limited or no coverage of these languages.

Our work bridges this gap by combining translation and transliteration to construct a Roman Urdu version of MS MARCO and training the first neural IR model for this script. It serves as a foundational benchmark for retrieval in Romanized low-resource languages.

4 Experimental Setup

4.1 Dataset Creation

We create our Roman Urdu IR dataset via a multi-hop translation process starting from the English MS MARCO dataset. Initially, we translate MS MARCO passages and queries into Urdu using the IndicTrans2 translation model, chosen for its strong performance on South Asian languages as previously demonstrated by (Butt et al., 2025a). The Urdu dataset comprises over 8.8 million passages and 500,000+ queries, serving as a reliable intermediate step.

English	Urdu (اردو) Translation	Roman-Urdu Transliteration
Query: what fruit is native to australia	**Query:** آسٹریلیا کا کون سا پھل مقامی ہے	**Query:** Australia ka kon sa phal muqami hai
Relevant Passage : Passiflora herbertiana. A rare passion fruit native to Australia. Fruits are green-skinned, white fleshed, with an unknown edible rating. Some sources list the fruit as edible, sweet and tasty.	**Relevant Passage:** پاسفلاورا ہربرٹیانا: آسٹریلیا کا ایک نایاب پیشن فروٹ۔ پھل سبز جلد والے سفیدگوشت والے ہوتے ہیں جن کی خوردنی درجہ بندی نامعلوم ہوتی ہے۔ کچھ ذرائع اس پھل کو خوردنی میٹھا اور لذیذ کے طور پر درج کرتے ہیں	**Relevant Passage :** pasflora harbrtiana. Australia ka aik nayab passion fruit. phal sabz jald walay safaid gosht walay hote hain jin ki khordani darja bandi namaloom hoti hen. kuch zaraye is phal ko khordani meetha aur lazeez ke taur par darj karte hain.
Non-Relevant Passage : The kola nut is the fruit of the kola tree, a genus (Cola) of trees that are native to the tropical rainforests of Africa.	**Non-Relevant Passage :** کولا نٹ کولا درخت کا پھل ہے جو درختوں کی ایک جینس کولا ہے جو افریقہ کے اشنکٹبندیی برساتی جنگلات سے تعلق رکھتی	**Non-Relevant Passage :** cola nutt cola darakht ka phal hai jo darakhton ki ek jeans cola hai jo Africa ke ashnktabndi barsati janglaat se taalluq rakhti ha

Figure 2: Example query and passage pair in English, Urdu (translated), and Roman Urdu (transliterated).

Next, we transliterate this whole Urdu dataset into Roman Urdu using a previously developed transliteration model (Butt et al., 2025b), which employs a transformer-based architecture (m2m100 (Fan et al., 2021)) fine-tuned on the Roman-Urdu-Parl corpus and augmented with Masked Language Modeling (MLM) pretraining. This step ensures robust handling of spelling variations common in Roman Urdu. This results in the Roman-Urdu version of the whole publically available English MS-Marco passage ranking dataset.

4.2 Potential Issues With Translated/Transliterated Dataset

Although machine translation provides a scalable solution, it can also lead to semantic drift and context loss. These issues are especially pronounced in multi-hop pipelines like ours (English → Urdu → Roman Urdu), where small inconsistencies can compound and negatively affect retrieval performance. Despite these challenges, the resulting dataset is the first large-scale Roman Urdu resource for information retrieval, making it a valuable foundation for future work.

An illustrative example of this semantic misalignment is shown in Figure 2.

4.3 Retrieval Model

We adopt a two-stage retrieval pipeline. First, a BM25 index serves as the base retriever, returning the top k=1000 candidates per query to ensure high recall. These candidates are then re-ranked using a multilingual reranker based on the mT5 architecture, which has been shown effective in multilingual IR tasks such as mMARCO.

Following the same approach as previously shown in (Butt et al., 2025a), we fine-tuned the

mMARCO model on the whole Roman Urdu dataset. While mT5 is pretrained on a diverse set of languages, it does not include Roman Urdu, making fine-tuning necessary to adapt to the script's informal and non-standard characteristics.

We frame retrieval as a binary relevance classification task in a sequence-to-sequence setup. For each query–passage pair, the model is trained to generate "yes" for relevant passages and "no" otherwise. At inference time, we compute a softmax over the generated tokens to obtain relevance scores for reranking.

The training configuration mirrors that of the mmarco model: a learning rate of 0.001, dropout of 0.1, and an effective batch size of 128 (batch size 32 with gradient accumulation over 4 steps). This consistent setup enables meaningful comparison between retrieval performance in Urdu and Roman Urdu, isolating the effects of linguistic representation.

4.4 Evaluation

We evaluate retrieval performance using standard IR metrics that reflect different aspects of effectiveness. We report **MRR@10**, **Recall@10**, **MAP@10**, **NDCG@10**, and **Precision@10** to provide a comprehensive view of overall ranking quality and the system's ability to surface relevant results.

Zero-shot multilingual models were not used as they fail to comprehend Roman Urdu's informal structure and inconsistent spelling. Since no prior baselines for Roman Urdu IR exist, we use **BM25** as the only meaningful comparison. Our significantly outperforming reranker demonstrates the effectiveness of the transliteration pipeline and model fine-tuning.

5 Results Discussion

We present the performance of our Roman Urdu IR model in Table 1, comparing our fine-tuned multilingual reranker against the BM25 baseline.

The reranker consistently outperforms BM25 across all metrics, achieving an MRR@10 of 0.1903 and Recall@10 of 0.3326, which is more than double the baseline values. Significant gains are also observed in MAP, NDCG, and Precision, indicating improvements in both early ranking and overall retrieval quality. This improvement is particularly notable given the noisy and inconsistent nature of Roman Urdu, which poses a challenge for lexical methods like BM25 that rely on exact token overlap. In contrast, the reranker benefits from contextual modeling and cross-lingual knowledge.

Compared to earlier results in Urdu IR (Butt et al., 2025a), the Roman Urdu model performs slightly lower (e.g., Urdu MRR@10: 0.248), which is expected due to the added noise introduced during transliteration. Nonetheless, the performance remains strong considering the informal nature of the script and lack of standardization.

These results validate our multi-hop pipeline and establish both the dataset and model as practical baselines for future research on retrieval in Romanized, informal, and low-resource languages.

Metric	BM25	Our Fine-tuned Reranker
MAP@10	0.0502	0.1262
MRR@10	0.0846	0.1903
NDCG@10	0.1218	0.2572
Precision@10	0.0177	0.0347
Recall@10	0.1699	0.3326

Table 1: Retrieval performance comparison on Roman Urdu MS MARCO (6980 queries).

6 Future Work and Conclusion

This paper presented the first large-scale Roman Urdu Information Retrieval dataset and benchmark, showing that fine-tuning a multilingual reranker substantially outperforms traditional methods. This approach effectively addresses the informal spelling and lack of standardization in Roman Urdu.

Future work could focus on reducing error propagation in the data pipeline, improving the transliteration model with more diverse data, and exploring subword or phonetic representations to better handle spelling variations. Our pipeline could also be extended to other Romanized scripts like Arabizi or Roman Hindi, broadening its application and fostering digital inclusion. This work provides a solid foundation and valuable resources for future research in Roman Urdu information retrieval

Limitations

While our work establishes the first large-scale Urdu and Roman Urdu resources for information retrieval, several limitations should be noted. First, the translation and transliteration pipeline introduces potential sources of semantic drift and context loss. These effects are particularly pronounced in multi-hop translation (English → Urdu → Ro-

man Urdu), where small inconsistencies can compound and reduce retrieval accuracy.

Second, our evaluation is limited to the MS MARCO-derived dataset. Although this provides a strong and widely used benchmark, it does not fully capture the diversity of information needs or linguistic phenomena in real-world Urdu and Roman Urdu usage.

Finally, due to time and resource constraints, we focused on establishing reliable baselines rather than exploring advanced modeling techniques or large-scale hyperparameter tuning. We view this work as a foundation for future improvements, such as expanding coverage to other domains, experimenting with alternative translation models, and refining retrieval strategies.

Broader Impact Statement

This work aims to improve digital information access for speakers of Roman Urdu, an informal and widely used script that has been historically ignored in language technology research. By providing the first large-scale dataset and retrieval model for Roman Urdu, our work contributes to a more inclusive digital ecosystem, enabling better access to search and knowledge for communities that rely on non-standardized, Romanized scripts.

Given the widespread use of Roman Urdu in South Asia, especially among younger and less formally educated populations, this research could help bridge digital inequality and support more equitable participation in online information spaces. Furthermore, our open-source release of models and datasets encourages transparency and reuse for similar languages and regions.

However, we also acknowledge that increased access to search and retrieval tools in informal scripts may be leveraged in unintended ways, such as misinformation or targeted advertising. Mitigating these risks requires responsible deployment and careful contextualization of the technology. We encourage future researchers and practitioners to work in collaboration with local communities to ensure that such tools are developed and used ethically.

Acknowledgement

We gratefully acknowledge the German Research Center for Artificial Intelligence (DFKI) for providing hardware and a supportive research environment. This work was also supported by the German Federal Ministry of Education and Research (BMBF) as part of the TRAILS project (01IW24005).

References

Mehreen Alam and Sibt Ul Hussain. 2022. Roman-urdu-parl: Roman-urdu and urdu parallel corpus for urdu language understanding. *Transactions on Asian and Low-Resource Language Information Processing*, 21(1):1–20.

Umer Butt, Stalin Veranasi, and Günter Neumann. 2025a. Enabling low-resource language retrieval: Establishing baselines for urdu ms marco. In *European Conference on Information Retrieval*, pages 282–289. Springer.

Umer Butt, Stalin Veranasi, and Günter Neumann. 2025b. Low-resource transliteration for roman-urdu and urdu using transformer-based models. *arXiv preprint arXiv:2503.21530*.

Jonathan H Clark, Eunsol Pfeiffer, Tom Kwiatkowski, Michael Collins, Kristina Toutanova, Patrick Lewis, Aishwarya Joshi, Pradeep Rajpurkar, and Luke Zettlemoyer. 2020. Tydi qa: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470.

A Conneau. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Jeffrey L Elman. 1990. Finding structure in time. *Cognitive science*, 14(2):179–211.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, et al. 2021. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic bert sentence embedding. *arXiv preprint arXiv:2007.01852*.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016a. Ms marco: A human generated machine reading comprehension dataset. In *Proceedings of the 2016 workshop on machine reading for question answering*, pages 180–186.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng.

2016b. Ms marco: A human-generated machine reading comprehension dataset.

Anoop K Ramesh, Deepak Raj, Dayal Tang, et al. 2022. Indictrans2: An improved neural translation model for indic languages. *arXiv preprint arXiv:2205.13431*.

Stephen E Robertson and Steve Walker. 1994. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. *Citeseer*.

Zanab Safdar, Ruqia Safdar Bajwa, Shafiq Hussain, Haslinda Binti Abdullah, Kalsoom Safdar, and Umar Draz. 2020. The role of roman urdu in multilingual information retrieval: A regional study. *The Journal of Academic Librarianship*, 46(6):102258.

Smat26. 2023. Roman urdu sentiment dataset. https://www.kaggle.com/datasets/smat26/roman-urdu-dataset.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Lrec*, volume 2012, pages 2214–2218. Citeseer.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.

Guangwei Yu, Jing Liu, Jialu Tang, Zujie Li, Shuming Bi, Yubin Pan, Peiran Huang, Bolin He, Jianhua Zhou, Xiao Zhang, et al. 2021. Miracl: Multimodal retrieval augmented with contrastive in-batch negatives. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1481–1491.

Rabail Zahid, Muhammad Owais Idrees, Hasan Mujtaba, and Mirza Omer Beg. 2020. Roman urdu reviews dataset for aspect based opinion mining. In *Proceedings of the 35th IEEE/ACM International Conference on Automated Software Engineering*, pages 138–143.