Spoken-Query Cross-Lingual Information Retrieval for the Indic Languages (SqCLIR) using BM25 and Indic-BERT

Pranesh TT, Thamizhmathi KK and Bharathi B

Department of Computer Science and Engineering, Sri Sivasubramaniya Nadar College of Engineering, Chennai, Tamil Nadu, India

Abstract

Voice-based search is becoming more common, but spoken queries are challenging for retrieval systems because they often contain transcription errors and, in many cases, need to work across languages. The SqCLIR task at FIRE 2025 addresses this challenge with two subtasks: a monolingual setup and a cross-lingual setup. For our participation, we focused on the Hindi monolingual track, where the goal was to retrieve relevant passages from a Hindi text collection using queries spoken in Hindi. To explore variation in speech input, we generated queries using both male and female voices. Our experiments combined traditional BM25 retrieval with neural embedding approaches using Indic-BERT and FAISS indexing. The results were encouraging, showing that embedding-based retrieval can improve recall and ranking compared to baselines. We also discuss how transcription errors and speaker characteristics influence retrieval effectiveness and how multilingual embeddings can provide benefits even within a monolingual setup.

Keywords

Spoken Query Retrieval, Cross-Language Information Retrieval, SqCLIR, Hindi Monolingual, Speech Recognition, Indic-BERT, FAISS, BM25

1. Introduction

1.1. Background and Motivation

Voice-based technologies such as virtual assistants, search engines, and conversational systems are now a part of everyday life. As more users interact with machines through speech, spoken query retrieval has become an important research challenge. Compared to typed queries, spoken queries are often noisy—they may include disfluencies, mispronunciations, or background interference that make retrieval harder [1]. In a multilingual country like India, this challenge is even greater, since users often ask queries in their native language. This makes the development of robust Spoken Query Cross-Language Information Retrieval (SqCLIR) systems particularly relevant [2].

1.2. The SqCLIR Task at FIRE 2025

The Forum for Information Retrieval Evaluation (FIRE 2025) introduced the second edition of the SqCLIR shared task to promote research on speech-driven information retrieval systems [3, 4]. The task provides both monolingual and cross-lingual tracks, where spoken queries must be matched against large collections of text documents. The official task website https://sites.google.com/view/sqclir-2025 offers detailed guidelines, dataset specifications, and evaluation protocols. Queries are released as audio recordings spoken by both male and female speakers, and are accompanied by transcriptions generated using Automatic Speech Recognition (ASR). In our work, we focus on the Hindi monolingual track, where both queries and documents are in Hindi.

Forum for Information Retrieval Evaluation, December 17-20, 2025, Varanasi, India

*Corresponding author.

†These authors contributed equally.

pranesh2370060@ssn.edu.in (P. TT); thamizhmathi2370055@ssn.edu.in (T. KK); bharathib@ssn.edu.in (B. B)



© 2025 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

1.3. Related Work

Prior research in information retrieval has explored both lexical ranking models and neural embedding-based approaches. While these methods have shown effectiveness for text and cross-lingual retrieval, their adaptation to spoken queries—especially for Indic languages—remains limited. We provide a detailed survey of these approaches in Section 2.

1.4. Our Contributions

To address these issues, we study the Hindi monolingual spoken query retrieval task at FIRE 2025. Our work makes the following contributions:

- We evaluate both traditional lexical retrieval (BM25) and embedding-based approaches using IndicBERT combined with FAISS indexing.
- We investigate the impact of speaker variation by comparing retrieval performance across queries spoken by male and female utterances.
- We highlight the trade-offs between lexical and embedding-based retrieval in handling noisy spoken queries, offering insights for developing more robust SqCLIR systems in low-resource settings.

The remainder of this paper is organized as follows. Section 2 presents a literature survey of related work in spoken query retrieval, lexical methods, and embedding-based approaches. Section 3 describes the methodology, including data preparation, retrieval workflow, and evaluation setup. Section 4 reports the experimental results of BM25 and Indic-BERT on the Hindi SqCLIR task, followed by a comparative discussion. Finally, Section 5 concludes the paper and outlines possible directions for future research.

2. Literature Survey

In recent years, research in the field of spoken query retrieval has gained momentum due to the growing use of speech-enabled technologies such as virtual assistants, search engines, and conversational systems. Spoken queries differ from typed queries because they often contain disfluencies, mispronunciations, or background noise, which makes retrieval more challenging [1]. In multilingual settings like India, cross-lingual information retrieval is further complicated by the diversity of languages and dialects. Several recent works have addressed these challenges, focusing on either lexical or embedding-based approaches for robust retrieval. The SqCLIR task introduced by FIRE further emphasizes the importance of developing effective speech-driven retrieval methods for Indian languages [2, 5, 6].

2.1. Lexical Retrieval Methods: BM25

Robertson et al. [7] introduced BM25, a probabilistic relevance framework widely used in traditional information retrieval tasks. BM25 ranks documents based on term frequency and inverse document frequency, providing strong baselines for text retrieval. In the context of spoken queries, BM25 has been applied to transcriptions obtained from automatic speech recognition (ASR) systems. Previous studies in FIRE SqCLIR tasks have shown that BM25 can achieve competitive performance on monolingual Hindi queries when ASR errors are low [2]. However, lexical methods struggle with semantic mismatches caused by paraphrasing or vocabulary variations in spoken language.

2.2. Embedding-Based Retrieval: IndicBERT and Multilingual Models

To handle semantic variations and cross-lingual queries, embedding-based methods have been widely explored. Multilingual sentence embeddings, such as LaBSE [8] and IndicBERT [9], map queries and documents into a shared semantic space, enabling similarity-based retrieval. IndicBERT has been fine-tuned on Indian language corpora and shown to improve retrieval performance over lexical methods by

capturing semantic context and mitigating ASR errors. Embedding-based approaches are particularly effective for low-resource languages and for queries that exhibit code-mixing or regional vocabulary differences.

Beyond Indic-specific models, dense retrieval methods developed in the international IR community have significantly advanced the state of the art. DPR (Dense Passage Retrieval) [10] introduced a dual-encoder framework for efficient passage retrieval, while ColBERT (Contextualized Late Interaction BERT) [11] proposed a late interaction mechanism that balances efficiency with fine-grained token-level matching. Further, ANCE (Approximate Nearest Neighbor Negative Contrastive Estimation) [12] improved training through dynamic negative sampling, and TAS(Topic Aware Sampling) [13] demonstrated the effectiveness of knowledge distillation and topic-aware sampling for lightweight dense retrievers. These models highlight the broader progress in dense retrieval for English and other high-resource languages, offering insights that can inform spoken query retrieval in Indic settings.

2.3. Spoken Query Retrieval Challenges in FIRE

Previous FIRE SqCLIR tracks have highlighted challenges specific to spoken query retrieval in Indian languages [5, 6, 2]. ASR errors, speaker variations (male vs. female utterances), and domain-specific vocabulary significantly affect retrieval performance. Some approaches combine lexical and embedding-based methods to balance precision and semantic recall. These studies provide a foundation for further improvements by analyzing both transcription quality and embedding effectiveness.

2.4. Current Work

Our study extends prior FIRE SqCLIR efforts by focusing on the Hindi monolingual track. We evaluate both lexical (BM25) and embedding-based (IndicBERT) retrieval, with a specific emphasis on speaker variability (male vs. female queries). This positions our work as a step toward more robust spoken query retrieval in low-resource settings.

3. Methodology

3.1. Objective and Research Questions

The primary goal of this research is to improve the retrieval of spoken Hindi queries from a large collection of Hindi documents. Specifically, we compare the effectiveness of a lexical method (BM25 [7]) with an embedding-based method (IndicBERT [14]) when applied to Automatic Speech Recognition (ASR) transcripts of spoken queries. Our research is guided by two main questions. First, how does BM25 compare to IndicBERT in retrieving relevant documents for noisy spoken Hindi queries? Second, how does speaker variability, particularly the difference between male and female utterances, influence retrieval performance? Answering these questions allows us to analyze the trade-offs between lexical and embedding-based retrieval approaches in the context of monolingual Hindi SqCLIR.

3.2. Data Preparation

The dataset used in this study is based on the Hindi resources released as part of the FIRE 2025 SqCLIR shared task [3, 4], combined with our own additional recorded queries. Spoken queries were provided in two evaluation sets (DL19 and DL20), with balanced representation from male and female speakers: 43 male and 43 female queries in DL19, and 54 male and 54 female queries in DL20. In addition, FIRE released official text queries for both DL19 and DL20 (43 and 54 queries respectively). To introduce additional speaker variability, we recorded 96 new spoken queries from male speakers only. In total, the dataset consists of 291 queries spanning spoken, text, and recorded sources. The document collection against which retrieval was performed consists of 8,841,823 Hindi news articles released as part of FIRE 2025. Table 1 summarizes the distribution of queries.

Table 1
Distribution of Hindi queries across sources, speaker groups, and evaluation years.

Query Type	Subset	DL19	DL20	Total
Spoken (Hindi)	Male	43	54	97
	Female	43	54	97
Text (Hindi)	—	43	54	97
Own Recorded (Hindi)	Male	97		97
Total		129	162	291

All spoken queries were transcribed into text using the Whisper ASR model [15]. Whisper was selected over alternatives such as Wav2Vec2 and Google Speech-to-Text due to its robustness to background noise and disfluencies, its multilingual coverage, and its open-source availability, making it particularly suitable for Indic language research.

The document collection and transcribed queries were normalized by tokenizing text, converting all words to lowercase, and removing Hindi stopwords. For embedding-based retrieval, we used IndicBERT, a transformer-based model pretrained on 12 major Indic languages (including Hindi) and English [14]. Its multilingual pretraining on large-scale web and Wikipedia corpora makes it particularly well-suited for Hindi retrieval tasks, offering a stronger representation baseline than generic multilingual models such as mBERT.

3.3. Retrieval Workflow

The retrieval workflow is illustrated in Figure 1. Spoken queries were first transcribed by Whisper ASR to generate textual inputs. Both the queries and the documents then underwent preprocessing. Retrieval was carried out using two complementary approaches. The first approach, BM25, is a probabilistic lexical method based on term frequency and inverse document frequency [7], which provides a strong baseline. The second approach, IndicBERT, is a transformer-based embedding model pretrained on a large corpus of 12 Indic languages [14]. IndicBERT produces dense vector representations of queries and documents, and retrieval is performed by computing cosine similarity using FAISS indexing. Finally, the ranked outputs from both methods were formatted into TREC-style run files to enable standardized evaluation and comparison.[1]

3.4. Evaluation Setup

The evaluation considered three types of queries: FIRE male queries, FIRE female queries, and our own recorded male queries. Retrieval effectiveness was assessed using Mean Reciprocal Rank (MRR), Recall@K, and nDCG [16]. This setup provided a systematic comparison between BM25 and IndicBERT under varying speaker conditions, while also allowing us to analyze the effect of ASR transcription quality on retrieval performance.

3.5. Summary

In summary, our methodology follows a structured pipeline that begins with transcription of spoken queries using Whisper ASR, followed by preprocessing of both queries and documents. Retrieval is then performed using BM25 as a lexical baseline and IndicBERT as an embedding-based method. The results are evaluated using standard IR metrics to compare the two approaches under different speaker conditions. BM25 provides an interpretable baseline, while IndicBERT leverages pretrained multilingual embeddings specifically tailored for Indic languages. Together, these methods enable a structured comparison between lexical and semantic retrieval strategies for Hindi SqCLIR.

[1]The source code and TREC run files are available at https://github.com/Pranesh4950/fire2025-sqclir.

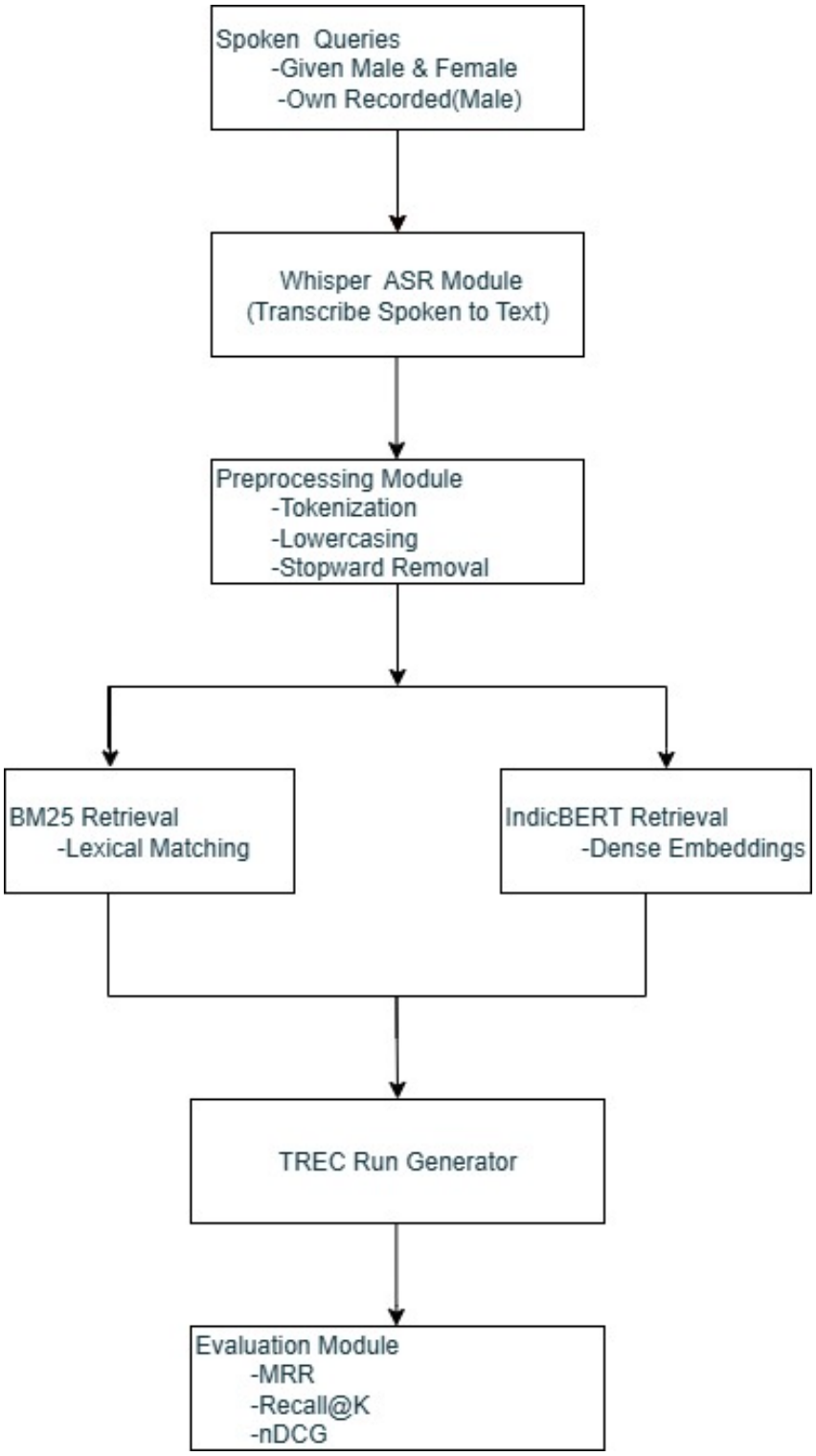


Figure 1: Workflow of our SqCLIR Hindi monolingual retrieval system.

4. Results and Comparative Analysis

The following tables present the evaluation of two retrieval models, BM25 and IndicBERT, on the SQCLIR task using spoken and text queries. The metrics reported include nDCG@10, Reciprocal Rank (RR), and Recall at different cutoffs (10, 100, and 1000). The results highlight differences in retrieval effectiveness across query types and between the two models.

Table 2
BM25 Results on Hindi SqCLIR Task

Query Type	nDCG@10	MRR	R@10	R@100	R@1000
Given spoken query (male)	0.0751	0.1872	0.0218	0.0782	0.1834
Given spoken query (female)	0.0951	0.2462	0.0330	0.1050	0.1976
Given text query	0.2024	0.4497	0.0578	0.1767	0.3358
Recorded spoken query (male)	0.0814	0.1992	0.0255	0.0776	0.1584

Table 3
IndicBERT Results on Hindi SqCLIR Task

Query Type	nDCG@10	MRR	R@10	R@100	R@1000
Given spoken query (male)	0.0586	0.1323	0.0915	0.0640	0.1348
Given spoken query (female)	0.0715	0.1964	0.0241	0.0674	0.1475
Given text query	0.1638	0.3618	0.0453	0.1241	0.2608
Recorded spoken query (male)	0.0744	0.1786	0.0231	0.0629	0.1371

4.1. Discussion of Results

4.1.1. BM25

BM25 demonstrates consistently stronger retrieval effectiveness compared to IndicBERT across all query types. As shown in Table 2, for given text queries BM25 achieves the highest nDCG@10 (0.2024) and RR (0.4497), indicating that relevant documents are ranked highly. Recall also improves significantly at higher cutoffs (R@1000 = 0.3358), suggesting that BM25 is able to capture a broader set of relevant documents.

For spoken queries, both male and female, BM25 achieves moderate results, with female queries slightly outperforming male queries (nDCG@10: 0.0951 vs. 0.0751, RR: 0.2462 vs. 0.1872). This indicates that female spoken queries are transcribed in a way that aligns more effectively with the document collection. Recorded spoken queries (male) yield results similar to given spoken queries, with small improvements at larger recall cutoffs.

4.1.2. IndicBERT

IndicBERT performs weaker than BM25 across all metrics. From Table 3, we observe that given text queries achieve the best performance among all IndicBERT results (nDCG@10 = 0.1638, RR = 0.3618). However, these values remain lower than BM25. Similarly, recall at 1000 documents (0.2608) falls short of BM25's coverage.

For spoken queries, IndicBERT shows a noticeable performance gap. Male spoken queries achieve the lowest nDCG@10 (0.0586), while female spoken queries perform slightly better (0.0715), but still trail BM25. Recorded spoken queries yield marginal improvement (nDCG@10 = 0.0744), but remain below BM25.

4.1.3. Comparative Analysis

Across all experiments, BM25 consistently outperforms IndicBERT. One clear trend is that both models perform best on given text queries, reflecting the absence of ASR transcription errors. However, the number of relevant tokens differs significantly between query types (e.g., 2296 for text vs. 1066 for spoken). This discrepancy arises because ASR-transcribed queries often lose or alter terms, reducing the overlap with relevant documents. As a result, spoken queries inherently lead to lower retrieval effectiveness.

The weaker performance of IndicBERT can be attributed to several factors. First, the model was not fine-tuned on the SqCLIR dataset, which limits its ability to capture domain-specific query–document relevance. Second, IndicBERT is sensitive to ASR noise, where transcription errors distort embeddings and reduce similarity with relevant documents. Third, while IndicBERT is pretrained on multiple Indic languages, its representations may not fully capture Hindi-specific nuances, especially when compared against BM25's direct term-matching mechanism.

4.1.4. Overall

In summary, BM25 demonstrates robustness to transcription noise and provides higher ranking effectiveness across all query types, making it the stronger retrieval model in our SqCLIR setting. IndicBERT remains a valuable semantic baseline but highlights the challenges of embedding-based methods for noisy, real-world Hindi speech queries without task-specific fine-tuning.

5. Conclusion

In this work, we evaluated BM25 and IndicBERT for the Hindi monolingual SQCLIR task at FIRE 2025. Our results show that BM25 consistently outperforms IndicBERT, with both models performing best on text queries and facing degradation on spoken queries due to ASR errors. Female spoken queries yield slightly higher effectiveness than male queries, but overall performance remains limited.

These findings underscore two key limitations: (i) ASR noise significantly reduces retrieval quality, and (ii) IndicBERT embeddings underperform without task-specific fine-tuning. Addressing these challenges motivates future directions such as hybrid retrieval that combines lexical precision with semantic embeddings, fine-tuning IndicBERT on Hindi speech-text pairs, and incorporating re-ranking strategies to mitigate noise.

Beyond benchmarking, advances in SqCLIR systems hold practical value for speech-enabled information access in Indian languages, supporting inclusive search tools in domains such as digital libraries, e-governance, and education.

6. Declaration on Generative AI

The authors declare that GPT-5 was used for grammar and spelling corrections.

Acknowledgments

The authors would like to acknowledge the support and resources provided by the Department of Computer Science and Engineering at Sri Sivasubramaniya Nadar College of Engineering, Chennai, Tamil Nadu, India.

References

[1] F. Crestani, Spoken query processing for interactive information retrieval, Data & Knowledge Engineering 41 (2002) 105–124.
[2] B. Dave, P. Majumder, Sqcliril: Spoken query cross-lingual information retrieval in indian languages, Pattern Recognition Letters (2025).
[3] B. Dave, P. Majumder, D. Ganguly, E. Kanoulas, Overview of the second shared task on spoken query cross-lingual information retrieval for indic languages sqclir at fire 2025 (2025).
[4] B. Dave, P. Majumder, D. Ganguly, E. Kanoulas, Findings from the second shared task on spoken query cross-lingual information retrieval for indic languages sqclir at fire 2025, in: Proceedings of the 17th Annual Meeting of the Forum for Information Retrieval Evaluation, 2025.

[5] B. Dave, P. Majumder, D. Ganguly, E. Kanoulas, Overview of the fire 2024 sqclir track: Spoken query cross-lingual information retrieval for the indic languages (2024).
[6] B. Dave, P. Majumder, D. Ganguly, E. Kanoulas, Findings of shared task on spoken query cross-lingual information retrieval for the indic languages at fire 2024, in: Proceedings of the 16th Annual Meeting of the Forum for Information Retrieval Evaluation, 2024.
[7] S. Robertson, H. Zaragoza, et al., The probabilistic relevance framework: Bm25 and beyond, Foundations and Trends® in Information Retrieval 3 (2009) 333–389.
[8] C. Feng, et al., Language-agnostic bert sentence embedding, in: ACL, 2020.
[9] A. Aravapalli, M. Marreddy, S. R. Oota, R. Mamidi, M. Gupta, Indicsenteval: How effectively do multilingual transformer models encode linguistic properties for indic languages?, arXiv preprint arXiv:2410.02611 (2024).
[10] V. Karpukhin, B. Oguz, S. Min, P. S. Lewis, L. Wu, S. Edunov, D. Chen, W.-t. Yih, Dense passage retrieval for open-domain question answering., in: EMNLP (1), 2020, pp. 6769–6781.
[11] O. Khattab, M. Zaharia, Colbert: Efficient and effective passage search via contextualized late interaction over bert, in: Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, 2020, pp. 39–48.
[12] L. Xiong, C. Xiong, Y. Li, K.-F. Tang, J. Liu, P. Bennett, J. Ahmed, A. Overwijk, Approximate nearest neighbor negative contrastive learning for dense text retrieval, arXiv preprint arXiv:2007.00808 (2020).
[13] S. Hofstätter, S.-C. Lin, J.-H. Yang, J. Lin, A. Hanbury, Efficiently teaching an effective dense retriever with balanced topic aware sampling, in: Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval, 2021, pp. 113–122.
[14] D. Kakwani, et al., Indicbert: A multilingual language model for indian languages, in: LREC, 2020.
[15] C. Graham, N. Roll, Evaluating openai's whisper asr: Performance analysis across diverse accents and speaker traits, JASA Express Letters 4 (2024).
[16] K. Järvelin, J. Kekäläinen, Cumulated gain-based evaluation of ir techniques, in: ACM Transactions on Information Systems (TOIS), volume 20, ACM, 2002, pp. 422–446.