Overview of the FIRE 2024 SqCLIR Track: Spoken Query Cross-Lingual Information Retrieval for the Indic Languages

Bhargav Dave[1,*], Prasenjit Majumder[1], Debasis Ganguly[2,*] and Evangelos Kanoulas[3]

[1]*Dhirubhai Ambani Institute of Information and Communication Technology, India*
[2]*University of Glasgow, Scotland, UK*
[3]*University of Amsterdam, Amsterdam, The Netherlands*

Abstract

This paper provides an overview of the first edition of the shared task on Spoken Query Cross-Lingual Information Retrieval for Indic Languages (SqCLIR), organized at the 16th Forum for Information Retrieval Evaluation (FIRE 2024). This year, we provided datasets for four languages from the FIRE collection: Hindi, Bengali, Gujarati, and Indian English, along with speech queries generated from the FIRE collection. This edition included two subtasks: 1) Spoken Query Ad-Hoc Retrieval: a Monolingual Retrieval 2) Spoken Query Cross-Lingual Retrieval. The SqCLIR task received an enthusiastic response, with over 26 teams registering. A total of 4 teams submitted runs across both subtasks, and 1 team ended up submitting the working notes. Standard metrics such as MRR, Recall@100, and Recall@1000 were used to evaluate both subtasks.

Keywords

Spoken Query, Information Retrieval, Indic Language, Cross-Lingual

1. Introduction

Recent advancements in Natural Language Processing (NLP) have driven substantial progress in widely spoken and resourced languages such as English, Chinese, and French. However, Indian languages, including Gujarati, Hindi, Bengali, Telugu, and Kannada, have not achieved a comparable state of resource development. The primary obstacle lies in the lack of adequate linguistic resources and tools despite India's immense linguistic diversity. Bridging this gap is essential to ensure equitable technological advancements across languages.

To address these challenges, the Forum for Information Retrieval Evaluation (FIRE) [1] has taken a leading role in promoting research on Indian languages. FIRE has made substantial contributions to various language-specific tasks, including hate speech detection [2, 3, 4, 5, 6, 7, 8], sentiment analysis [9, 10, 11], mixed-script information retrieval [12, 13], summarization [14, 15], sarcasm detection [16], fake news detection [17, 18], machine translation [19], and event detection [20, 21]. Additionally, FIRE has addressed language-independent challenges such as legal document retrieval and summarization [22, 23, 24], microblog retrieval [25], and information retrieval for software engineering [26, 27]. A notable contribution of FIRE is the release of the FIRE corpus during 2008–2012 [28, 29], which has provided researchers with critical resources for building and evaluating information retrieval systems.

India's linguistic landscape, encompassing 22 officially recognized languages under the Eighth Schedule of the Constitution—including Assamese, Bengali, Gujarati, Hindi, Kannada, and others—presents unique challenges for NLP. The diversity in scripts, grammar, and phonology, coupled with limited resources, complicates the development of robust and scalable retrieval systems. Spoken-query retrieval, a critical area for enhancing accessibility, remains particularly underexplored for Indian languages.

FIRE'24: Forum for Information Retrieval Evaluation 2024, December 12–15, 2024, India

[*] Corresponding author.

bhargavdave1@gmail.com (B. Dave); prasenjit.majumder@gmail.com (P. Majumder); debforit@gmail.com (D. Ganguly); ekanoulas@gmail.com (E. Kanoulas)
0000-0003-2742-480X (B. Dave); 0000-0003-0840-9313 (P. Majumder); 0000-0003-0050-7138 (D. Ganguly); 0000-0002-8312-0694 (E. Kanoulas)

© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR-WS.org/Vol-4054/T8-1.pdf

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

In cross-lingual contexts, the challenge grows further due to the structural and resource disparities between languages.

To address this challenge and explore new territory, we introduced a novel shared task called Spoken-Query Cross-Lingual Information Retrieval for Indic Languages (SqCLIR) as part of FIRE 2024. This task aims to support the development and evaluation of retrieval systems that process spoken queries as input to retrieve relevant documents from a corpus. The inaugural edition of SqCLIR includes two tasks:

1. Spoken Query Ad-Hoc Retrieval - Monolingual Retrieval
2. Spoken Query Cross-Lingual Retrieval

The Monolingual Retrieval Task covered languages Gujarati, Hindi, Bengali, and Indian English, while the Cross-Lingual Retrieval Task included English, Hindi, and Bengali. For this year, we utilized the FIRE dataset from 2008- 2012 as the document target retrieval collection. The Spoken Query dataset was created using these corpora, providing 50 spoken queries as training data and 150 spoken queries as test data.

The remainder of this paper is organized as follows. Section 2 provides detailed information about the shared task, followed by Section 3, which describes the dataset used. Section 4 outlines the evaluation methods employed for assessing retrieval systems. Section 5 presents the results of the participating teams, and Section 6 concludes the paper with key findings and future directions.

2. Task Definition

The first shared task on Spoken-Query Cross-Lingual Information Retrieval for the Indic languages marks a significant advancement in creating benchmark datasets for spoken-query retrieval in Indic languages. This inaugural edition covers Gujarati, Hindi, Bengali, and Indian English. This year, two subtasks are introduced: Monolingual Retrieval and Cross-Lingual Retrieval. The following subsections provide a detailed discussion of each task and the corresponding datasets.

2.1. Task 1 : Spoken Query Ad-Hoc Retrieval Data - Monolingual Task

The objective of this task is develop a Spoken Query Retrieval System to handle monolingual spoken queries within a standard text-based retrieval and ranking framework. Both the spoken queries and the documents in this task are in the same language, simplifying the retrieval process and allowing for a more direct language-specific search. The primary focus is on accurately interpreting spoken queries and retrieving relevant documents from a monolingual corpus, thus ensuring efficiency and consistency throughout the retrieval process. For this inaugural edition of SqCLIR, we have included the languages English, Gujarati, Hindi, and Bengali.

2.2. Task 2 : Spoken Query Cross-Lingual Retrieval

The objective of this task is to develop a Spoken Query Retrieval System capable of handling cross-lingual queries. Unlike monolingual retrieval, this task involves spoken queries and a corpus in different languages, introducing additional complexity to the retrieval process. The system should accurately interpret spoken queries in one language and retrieve the most relevant documents from a corpus in another language. For this inaugural edition of SqCLIR, the languages included are English, Hindi, and Bengali. The task uses various combinations of these languages as query-corpus pairs, enabling participants to tackle a range of cross-lingual retrieval challenges.

Monolingual Task Result					
Team Name	MAP	MRR	R@10	R@100	R@1000
English					
IITM_BS	0.0414	0.2414	0.0321	0.1279	0.2503
Hindi					
Awsathama	0.0032	0.0561	0.0027	0.0115	0.0388

Table 1
Results of monolingual retrieval task

3. Dataset

In this inaugural year of the SQCLIR track, we leverage the FIRE Collection[1] [28, 29], a rich and diverse dataset that includes several Indian languages, such as English, Hindi, Bengali, Gujarati, and Marathi. This corpus is sourced from reputable publications, including Anandabazar Patrika for Bengali, Gujarat Samachar for Gujarati, Indiatimes and Dainik Jagran for Hindi, and The Telegraph for English. The dataset provides a robust foundation for developing and evaluating retrieval systems, featuring a wide range of documents.

Additionally, we also provided the Spoken Query dataset queries to the participants, which is created using the FIRE Collection, with queries spoken by native speakers proficient in English, Gujarati, Hindi, and Bengali. This ensures that the spoken queries reflect natural language use and dialectal nuances in these languages, providing a robust basis for developing and evaluating retrieval systems.

For both tasks, we provided 50 spoken queries as training data and 150 spoken queries for testing data, along with the qrel for train querys. This combined dataset supports the development and evaluation of both monolingual and cross-lingual retrieval systems.

4. Evaluation

Submissions were evaluated using measures from the ir_measures tool [30], the official implementation of trec_eval for standard evaluation measures. To assess both tasks, we used the qrel files specific to the language of the documents in the corpus. For evaluation, Mean Reciprocal Rank (MRR) served as the primary metric for ranked retrieved documents. Additionally, Recall@100, and Recall@1000 metrics were employed to provide a more comprehensive assessment of the results.

5. Results and Discussion

For both tasks, we received a total of 26 team registrations, with 23 teams registered for Task 1 (Monolingual Retrieval) and 20 for Task 2 (Cross-Lingual Retrieval). In Task 1, registrations were distributed across the following languages: 9 in Gujarati, 17 in Hindi, 16 in Bengali, and 22 in English. For Task 2, registrations covered the following language pairs: 17 for English-Hindi, 11 for English-Bengali, 16 for Hindi-English, 8 for Hindi-Bengali, 8 for Bengali-Hindi, and 10 for Bengali-English. Out of the 4 teams that submitted runs, only 2 teams provided valid submissions, with 1 run each—one in English and one in Hindi, both for Task 1 only.The results are provided in Table 5.

For English Monolingual Retrieval, the IIT_BS [31] team submitted a single run. They used a pre-trained Whisper model [32] to transcribe spoken queries into text. The all-MiniLM-L6-v2 model [33] was then employed to generate embeddings for both the transcribed query and the documents. Cosine similarity was calculated between the query and document embeddings to retrieve and rank documents based on their relevance to the query. On the test queries, they achieved an MRR of 0.2414 with Recall@100 of 0.1279, and Recall@1000 of 0.2503.

[1] https://fire.irsi.org.in/fire/static/data

For Hindi Monolingual Retrieval, the Awsathama team submitted a single run. They used a pre-trained Whisper model to transcribe spoken queries into text, followed by the paraphrase-multilingual-MiniLM-L12-v2 sentence-transformers model to generate embeddings for both the transcribed queries and documents. A FAISS IndexFlatIP [34, 35] was then used to retrieve and rank documents based on their relevance to the query. On the test queries, they achieved an MRR of 0.0561 with Recall@100 of 0.0115, and Recall@1000 of 0.0388. These results were suboptimal, potentially due to implementation errors, limitations of the embedding model for Hindi, or issues with the Whisper model. However, as the team did not submit working notes, we are unable to confirm the exact cause.

6. Concluding Discussions

The SqCLIR track was introduced for the first time at FIRE 2024 to promote research in Speech Query Cross-Lingual Information Retrieval for Indian languages (Hindi, Gujarati, English, and Bengali). Although dense, single-stage retrieval systems performed well, the lack of varied methods and limited team involvement posed challenges for broader evaluation and innovation. Despite these issues, we, as the organizers, remain optimistic that future iterations will encourage greater participation and foster more diverse and robust solutions for Speech Query Cross-Lingual Information Retrieval.

Acknowledgments

We sincerely thank the organizers of FIRE 2024 for providing us with the opportunity to run the SqCLIR track as part of the conference. We also extend our heartfelt gratitude to the native speakers who played a pivotal role in the creation of the Spoken Query dataset. In particular, we acknowledge the valuable contributions of Bhavesh Baraiya for Gujarati, Avneet Sharma for Hindi and English, and Parthiv Chatterjee for Bengali. Their support and dedication were instrumental in the successful development of this resource. Additionally, we would like to thank generative AI models for assisting us in refining and improving the quality of our written materials.

Declaration on Generative AI

Either:
The author(s) have not employed any Generative AI tools.

Or (by using the activity taxonomy in ceur-ws.org/genai-tax.html):
During the preparation of this work, the author(s) used X-GPT-4 and Gramby in order to: Grammar and spelling check. Further, the author(s) used X-AI-IMG for figures 3 and 4 in order to: Generate images. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] P. Mehta, T. Mandl, P. Majumder, S. Gangopadhyay, Report on the fire 2020 evaluation initiative, in: ACM SIGIR Forum, volume 55, ACM New York, NY, USA, 2021, pp. 1–11.

[2] S. Satapara, S. Masud, H. Madhu, M. A. Khan, M. S. Akhtar, T. Chakraborty, S. Modha, T. Mandl, Overview of the hasoc subtracks at fire 2023: Detection of hate spans and conversational hate-speech, in: Proceedings of the 15th Annual Meeting of the Forum for Information Retrieval Evaluation, 2023, pp. 10–12.

[3] S. Modha, T. Mandl, P. Majumder, S. Satapara, T. Patel, H. Madhu, Overview of the hasoc subtrack at fire 2022: Identification of conversational hate-speech in hindi-english code-mixed and german language (2022).

[4] T. Mandl, S. Modha, G. K. Shahi, H. Madhu, S. Satapara, P. Majumder, J. Schäfer, T. Ranasinghe, M. Zampieri, D. Nandini, et al., Overview of the hasoc subtrack at fire 2021: Hate speech and offensive content identification in english and indo-aryan languages (2021).

[5] T. Mandl, S. Modha, G. Shahi, A. Jaiswal, D. Nandini, D. Patel, P. Majumder, J. Schäfer, Overview of the hasoc track at fire 2020: Hate speech and offensive content identification in indo-european languages, in: CEUR Workshop Proceedings, volume 2826, CEUR Workshop Proceedings, 2020, pp. 87–111.

[6] T. Mandl, S. Modha, P. Majumder, D. Patel, M. Dave, C. Mandlia, A. Patel, Overview of the hasoc track at fire 2019: Hate speech and offensive content identification in indo-european languages, in: Proceedings of the 11th Annual Meeting of the Forum for Information Retrieval Evaluation, FIRE '19, Association for Computing Machinery, New York, NY, USA, 2019, p. 14–17. URL: https://doi.org/10.1145/3368567.3368584. doi:10.1145/3368567.3368584.

[7] S. Modha, P. Majumder, T. Mandl, C. Mandalia, Detecting and visualizing hate speech in social media: A cyber watchdog for surveillance, Expert Systems with Applications 161 (2020) 113725.

[8] H. Madhu, S. Satapara, S. Modha, T. Mandl, P. Majumder, Detecting offensive speech in conversational code-mixed dialogue on social media: A contextual dataset and benchmark experiments, Expert Systems with Applications 215 (2023) 119342.

[9] B. R. Chakravarthia, R. Priyadharshinib, V. Muralidaranc, S. Suryawanshia, N. Josed, E. Sherlyd, J. P. McCraea, Overview of the track on sentiment analysis for dravidian languages in code-mixed text (2020).

[10] B. R. Chakravarthia, R. Priyadharshinib, S. Thavareesanc, D. Chinnappad, D. Thenmozhie, E. Sherlyf, J. P. McCraea, A. Handeh, R. Ponnusamyf, S. Banerjeej, et al., Findings of the sentiment analysis of dravidian languages in code-mixed text (2021).

[11] K. Shanmugavadivel, M. Subramanian, P. K. Kumaresan, B. R. Chakravarthi, B. Bharathi, S. C. Navaneethakrishnan, L. S. Kumar, T. Mandl, R. Ponnusamy, V. Palanikumar, et al., Overview of the shared task on sentiment analysis and homophobia detection of youtube comments in code-mixed dravidian languages., in: FIRE (Working Notes), 2022, pp. 80–91.

[12] S. Banerjee, K. Chakma, S. K. Naskar, A. Das, P. Rosso, S. Bandyopadhyay, M. Choudhury, Overview of the mixed script information retrieval (msir) at fire-2016, in: Text Processing: FIRE 2016 International Workshop, Kolkata, India, December 7–10, 2016, Revised Selected Papers, Springer, 2018, pp. 39–49.

[13] P. Gupta, K. Bali, R. E. Banchs, M. Choudhury, P. Rosso, Query expansion for mixed-script information retrieval, in: Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval, 2014, pp. 677–686.

[14] S. Satapara, P. Mehta, S. Modha, D. Ganguly, Key takeaways from the second shared task on indian language summarization (ilsum 2023)., in: FIRE (Working Notes), 2023, pp. 724–733.

[15] S. Satapara, B. Modha, S. Modha, P. Mehta, Findings of the first shared task on indian language summarization (ilsum): Approaches challenges and the path ahead., in: FIRE (Working Notes), 2022, pp. 369–382.

[16] B. R. Chakravarthi, N. Sripriya, B. Bharathi, K. Nandhini, S. C. Navaneethakrishnan, T. Durairaj, R. Ponnusamy, P. K. Kumaresan, K. K. Ponnusamy, C. Rajkumar, Overview of sarcasm identification of dravidian languages in dravidiancodemix@ fire-2023, in: FIRE (Working Notes), 2023.

[17] M. Amjada, S. Butta, H. I. Amjadc, A. Zhilab, G. Sidorova, A. Gelbukha, Overview of the shared task on fake news detection in urdu at fire 2021 (2021).

[18] M. Amjada, G. Sidorova, A. Zhilab, A. Gelbukha, P. Rossoc, Overview of the shared task on fake news detection in urdu at fire 2020 (2020).

[19] S. Gangopadhyaya, G. Epilia, P. Majumdera, B. Gainb, R. Appicharlab, A. Ekbalb, A. Ahsanc, D. Sharmac, Overview of mtil track at fire 2023: Machine translation for indian languages (2023).

[20] B. Dave, S. Gangopadhyay, P. Majumder, P. Bhattacharya, S. Sarkar, S. L. Devi, Fire 2020 ednil track: Event detection from news in indian languages, in: Proceedings of the 12th Annual Meeting of the Forum for Information Retrieval Evaluation, FIRE '20, Association for Computing Machinery, New York, NY, USA, 2021, p. 25–28. URL: https://doi.org/10.1145/3441501.3441516.

doi:10.1145/3441501.3441516.

[21] B. Dave, S. Gangopadhyay, P. Majumder, P. Bhattacharya, S. Sarkar, S. L. Devi, Fire 2020 ednil track: Event detection from news in indian languages, in: Proceedings of the 12th Annual Meeting of the Forum for Information Retrieval Evaluation, 2020, pp. 25–28.

[22] P. Bhattacharya, K. Ghosh, S. Ghosh, A. Pal, P. Mehta, A. Bhattacharya, P. Majumder, Overview of the fire 2019 aila track: Artificial intelligence for legal assistance (2019).

[23] P. Bhattacharyaa, P. Mehtab, K. Ghoshc, S. Ghosha, A. Pald, A. Bhattacharyae, P. Majumderf, Overview of the fire 2020 aila track: Artificial intelligence for legal assistance (2020).

[24] V. Parikh, U. Bhattacharya, P. Mehta, A. Bandyopadhyay, P. Bhattacharya, K. Ghosh, S. Ghosh, A. Pal, A. Bhattacharya, P. Majumder, Overview of the third shared task on artificial intelligence for legal assistance at fire 2021., in: Fire (working notes), 2021, pp. 517–526.

[25] M. Basu, S. Ghosh, K. Ghosh, Overview of the fire 2018 track: Information retrieval from microblogs during disasters (irmidis), in: Proceedings of the 10th annual meeting of the Forum for Information Retrieval Evaluation, 2018, pp. 1–5.

[26] S. Majumdar, S. Paul, B. Dave, D. Paul, A. Bandyopadhyay, S. Chattopadhyay, P. P. Das, P. D. Clough, P. Majumder, Generative ai for software metadata: Overview of the information retrieval in software engineering track at fire 2023 (2023).

[27] S. Majumdar, A. Bandyopadhyay, S. Chattopadhyay, P. P. Das, P. D. Clough, P. Majumder, Overview of the irse track at fire 2022: Information retrieval in software engineering., in: FIRE (Working Notes), 2022, pp. 1–9.

[28] P. Majumder, M. Mitra, D. Pal, A. Bandyopadhyay, S. Maiti, S. Pal, D. Modak, S. Sanyal, The fire 2008 evaluation exercise, ACM Transactions on Asian Language Information Processing 9 (2010). URL: https://doi.org/10.1145/1838745.1838747. doi:10.1145/1838745.1838747.

[29] S. Palchowdhury, P. Majumder, D. Pal, A. Bandyopadhyay, M. Mitra, Overview of fire 2011, in: P. Majumder, M. Mitra, P. Bhattacharyya, L. V. Subramaniam, D. Contractor, P. Rosso (Eds.), Multilingual Information Access in South Asian Languages, Springer Berlin Heidelberg, Berlin, Heidelberg, 2013, pp. 1–12.

[30] S. MacAvaney, C. Macdonald, I. Ounis, Streamlining evaluation with ir-measures, in: Advances in Information Retrieval - 44th European Conference on IR Research, ECIR 2022, Stavanger, Norway, April 10-14, 2022, Proceedings, Part II, volume 13186 of *Lecture Notes in Computer Science*, Springer, 2022, pp. 305–310. URL: https://doi.org/10.1007/978-3-030-99739-7_38. doi:10.1007/978-3-030-99739-7_38.

[31] P. R. Nagarajan, L. N. Dhasan, M. D. Thiagarajan, Spoken query retrieval in monolingual contexts with whisper and sbert models (2024).

[32] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, I. Sutskever, Robust speech recognition via large-scale weak supervision, 2022. URL: https://arxiv.org/abs/2212.04356. arXiv:2212.04356.

[33] W. Wang, F. Wei, L. Dong, H. Bao, N. Yang, M. Zhou, Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers, in: H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems, volume 33, Curran Associates, Inc., 2020, pp. 5776–5788. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[34] M. Douze, A. Guzhva, C. Deng, J. Johnson, G. Szilvasy, P.-E. Mazaré, M. Lomeli, L. Hosseini, H. Jégou, The faiss library (2024). arXiv:2401.08281.

[35] J. Johnson, M. Douze, H. Jégou, Billion-scale similarity search with GPUs, IEEE Transactions on Big Data 7 (2019) 535–547.