JCBI
Journal of Computing & Biomedical Informatics Volume 05 Issue 01
ISSN: 2710 - 1606 2023

Review Article
https://doi.org/10.56979/501/2023

Cross-Lingual Information Retrieval in a Hybrid Query Model for Optimality

Abdul Basit[1], Israr Hanif[1], Muhammad Sajid Maqbool[1], Wahid Qayyum[2], Muhammad Adnan Hasnain[3], Rubaina Nazeer[4]

[1]Department of Computer Science, Bahauddin Zakariya University,Multan, Pakistan.
[2]Department of Software Engineering, Faculty of Computer Science, Lahore Garrison University, Lahore, 54000, Pakistan.
[3]Department of Computer Science, National College of Business Administration & Economics, Lahore, Pakistan.
[4]Department of Information Sciences, University of Education, Lahore, Pakistan.
Corresponding Author: Muhammad Sajid Maqbool. Email: sajidmaqbool7638@gmail.com.

Received: April 02, 2023 **Accepted:** May 29, 2023 **Published:** June 05, 2023.

Abstract: Cross-Lingual Information Retrieval (CLIR) allows users to get the documents in the language other than the query language. It is accomplished in two ways: In first method the query is translated in target language while in second method the documents are translated in query's language. Usually, the query translation is done due to translation complexity. In query translation method a query in language A is translated and compared against the document index in language B. Text RErieval Conference (TREC) is a forum to evaluate performance of an information retrieval system. Different tracks are designed to address different domains. Each track normally provides a corpus which contains collection of documents, few query topics and a set of related documents against each topic to perform the evaluation task. Mono lingual information retrieval in Urdu-Urdu domain is addressed by the researchers up to some extent but cross lingual Urdu-English retrieval is not focused yet. Our research addresses this area by using UIR-21 corpus composed of Urdu news documents designed for Urdu information retrieval task. We used this corpus for modeling hybrid query impacts on retrieval. Proposed CLIR model supports query in three languages Urdu, English and Roman-Urdu and provides the documents in Mono-Lingual as well as Cross-Lingual (Urdu to English and vice versa) contexts. For evaluation purpose we computed the Precision, Recall and F-1 Score of each mode. The highest precision is achieved by the Roman Urdu Retrieval Model (RURM) and the lowest precision by the Urdu Retrieval Model (URM).

Keywords: Information Retrieval System, Urdu information retrieval, Cross-Lingual Information Retrieval, Roman-Urdu Information Retrieval

1. Introduction

Cross-Lingual Information Retrieval (CLIR) systems are designed to provide users with access to information written in languages different from the one they are using for their queries. These systems allow users to send requests in one language and receive documents in another language. Developing a reliable CLIR system is a challenging task for researchers, especially when it comes to languages like Urdu, which has significant morphological complexities and a large number of native speakers. Urdu is widely spoken in Pakistan and has approximately 163 million native speakers worldwide. However, most of the information available on the internet is in English. This language barrier restricts Urdu-speaking

individuals from accessing the vast amount of information on the web. Therefore, it is crucial to focus on developing effective CLIR systems for Urdu-English language pair.

Our research team is focusing on creating a hybrid query model that maximises cross-linguistic context in order to meet this need. There aren't many end-effective CLIR solutions available right now in the Urdu-English domain. With support for questions in Urdu, English, and Roman-Urdu, our proposed CLIR model provides documents in both monolingual and cross-lingual situations. We used the UIR-21 corpus, which is a current and pertinent collection of documents, in the review process. On April 30th, 2022, a corpus of 37,000 TREC-formatted documents was uploaded online after being collected from several publications. Particularly, it has 2,887,169 Urdu documents. We used the Google Translator API and the Ijanoon website to translate these documents into English and Roman-Urdu, respectively, to make them available for cross-lingual retrieval. Using common metrics like Precision, Recall, and F-1 Score, we assessed the effectiveness of our suggested model. In accordance with the findings, the Urdu Retrieval Model (URM) had the lowest precision while the Roman Urdu Retrieval Model (RURM) had the highest precision. These evaluations were conducted according to the guidelines provided by TREC, which emphasize the inclusion of a document set, a query set, and relevance judgments for accurate assessment. In conclusion, the development of CLIR systems is of utmost importance, particularly for languages like Urdu, where a significant portion of the population is not proficient in English and thus lacks access to a vast pool of information on the internet. Our research focuses on bridging this language gap by proposing a hybrid query model for cross-linguistic context optimization. By evaluating our model using the TREC guidelines and employing the UIR-21 corpus, we have made significant progress in addressing the challenges of CLIR in the Urdu-English domain.

2. Related Work

Research efforts in IR have focused on two key areas: Cross-Lingual Information Retrieval (CLIR) systems and monolingual IR systems. CLIR systems aim to bridge language barriers, enabling users to search for information in languages different from their own. The initial CLIR test collections were prepared in anticipation of Salton's seminal work on CLIR, involving the manual translation of English queries into German back in 1970. However, the contemporary landscape of CLIR poses significant challenges, primarily due to the exponential growth of research articles available on the internet. Searching with less commonly used languages often yields limited and less relevant results. The primary objective of any IR system is to provide users with accurate answers to their queries. However, there are instances when the required information may not be accessible in the user's native language. To tackle this issue, researchers have proposed diverse approaches. For instance, Bisma Ayaz et al. [4] developed a semantic search engine tailored specifically for Urdu e-novels. Their system incorporates three layers: ontology modeling, information retrieval, and dynamic ontology modifications. By leveraging semantic queries on the ontology, the system effectively retrieves the necessary information.

A CLIR approach for Afaan Oromo-English translation was introduced by Daniel Bekele et al. [5] in a similar spirit. Through queries in their original language, this system enables native Afaan Oromo speakers to browse and obtain information that is available in English. Measures like precision and recall were used to assess the system's performance. An essential component of CLIR systems is query translation. Deep learning approaches, notably neural machine translation (NMT), were proposed by Tianchi Bi et al. [6] for query translation. Although it is often trained on large-scale out-of-domain data rather than query translation pairs, NMT has demonstrated potential in a number of applications. Researchers have also concentrated on particular language combinations. Maryamah Maryamah et al. [7] presented a study on Arabic-Indonesian CLIR. They modified Google Translate by incorporating a dictionary and word embedding techniques. Their proposed technique achieved a higher average BLEU score compared to standard translation software.

Comparison between query translation methodologies has also been investigated. Shadi Saleh and Pavel Pecina [8] conducted a study comparing query translation (QT) and document translation (DT) methodologies. They examined various machine translation techniques and their combinations, analyzing large-scale test data collected during the CLEF eHealth tasks. Additionally, researchers have explored specific challenges in CLIR systems. Rabih Zbib et al. [9] focused on Urdu-English CLIR and found that transliteration of identified entities, particularly missing vowels, contributed to incorrect query results. They highlighted the importance of accurate transliteration. Improving word translation probabilities is another area of research. Manaal Faruqui et al. [10] developed an artificial neural model for estimating CLIR word translation probabilities. Their model improved the prediction of word translation probabilities by incorporating source word context and encoding input word strings.

Researchers have made significant contributions to the field of IR systems, focusing on specific requirements and challenges. For instance, Elizabeth Boschee et al. [11] presented SARAL, an IR method designed for low-resource languages. SARAL enables English speakers to explore foreign language documents or audio directories using English queries, providing summaries and comprehensive transcriptions when needed. In the domain of medical information retrieval across languages, Eugenio Picchi and Carol Peters [12] proposed a technique that combines machine translation and linear regression for query expansion. Their method translates queries from the source language to the document language using machine translation, while linear regression predicts the performance of expanded queries with candidate words.

Creating suitable test collections is crucial for evaluating Cross-Lingual Information Retrieval (CLIR) systems. Dawn Lawrie et al. [13] developed HC4, a corpus that includes documents in languages such as Chinese, Persian, and Russian, along with graded relevance assessments. This collection facilitates the evaluation of CLIR systems. To support the evaluation of CLIR systems, Shuo Sun et al. [14] introduced CLIReval, a toolset that assesses machine translation performance through the proxy task of CLIR. CLIReval does not rely on annotated CLIR datasets, making it a valuable resource for evaluating system performance. In terms of framework development, Elham Ghanbari and Azadeh Shakery [2] proposed an LTR-based CLIR framework that leverages knowledge from training queries in both the source and target languages to extract features and create a ranking model. Marco Jungwirth and Allan Hanbury [18] replicated an experiment by Philipp Sorg and Philipp Cimiano, focusing on CLIR with Explicit Semantic Analysis.

Furthermore, researchers have explored new challenges in CLIR. Juntao Li et al. [15] investigated the set-to-description retrieval challenges CLIR. Despite the progress made, query translation remains a critical aspect of CLIR systems, accompanied by various obstacles. Liang Yao Baosong et al. [18] proposed a methodology to overcome these challenges by addressing issues such as adequate translation, limited in-domain parallel training data, and the need for low latency in CLIR systems.

In summary, extensive research has been conducted in the field of CLIR systems, focusing on query translation, system evaluation, specific language pairs, and specialized domains. These studies have advanced the effectiveness and usability of CLIR systems, facilitating access to information across languages and bridging the language gap.

Table 1. Comparison of different research work

Ref.	Language	Type	Models
Maryamah Maryamah et al. (2021)	Arabic-Indonesian	Cross-lingual Information	Arabic-Indonesian CLIR
Peng Shi et al. (2021)	Chinese, Arabic, French, Hindi, Bengali, Spanish	Multilingual	dense retrieval
Shuo Sun et al. (2020)	Different languages	Crosslingual	CLIReval,
Elham Ghanbari and Azadeh Shakery (2021)	Multi_languages	Crosslingual	CLIR
Shadi Saleh and Pavel Pecina (2020)	English and 7 European languages	Cross-lingual Information	English -Europeans CLIR
In Juntao Li et al. (2020)	English and other languages	Crosslingual	CLMN
Ketan Rajshekhar Shahapure (2016)	English to Japanese or Korean, Urdu, Tamil, Persian	Crosslingual	CLIR system
Chayapathi A R et al. (2021)	Hindi-English Kannada-English	Multilingual	Multi-lingual information retrieval system (MLIR)
Vijay Sharma and Namita Mittal (2019)	Hindi-English	Crosslingual	Hindi-English CLIR
Daniel Bekele et al. (2015)	Oromo-English	Crosslingual	Afaan Oromo-English (CLIR)
Tianchi Bi et al. (2020)	Russian and English	Crosslingual	CLIR system
Masha Yarmohammadi et al. (2019)	Somali, Swahili, Tagalog and English	Crosslingual	CLIR
Elizabeth Boschee et al. (2019)	Swahili	Crosslingual	SARAL
In D Thenmozhi and ChandraBose Aravindan (2014)	Tamil-English	Crosslingual	Tamil-English CLIR
Bisma Ayaz (2016)	Urdu	Monolingual	Urdu E-novel IR
Manaal Faruqui et al. (2011)	Urdu-English	Crosslingual	CLIR for Urdu English
Shuo Sun (2020)	Urdu-English	Crosslingual	CLIReval
Marco Jungwirth et al. (2019)	Wikipedia	Multilingual	CL-ESA

There are many academics who work on the CLIR in the aforementioned papers, and some of them are listed in (Table 1), which includes "Urdu, English, Hindi, French, Chinese, Arabic, Bengali, Somali, Tamil, Spanish, Swahili, and Tagalog". Most of our research is CLIR in Urdu—English. A CLIR for Urdu—English must be completed after reading all of the aforementioned studies. In the following chapter, we suggested a model that can handle three various question kinds and languages and return documents in the challenging language. Urdu, English, and Roman—Urdu static document collections are included in our suggested approach. For this research, we used the UIR-21 Urdu dataset and the Relevancy Judgement

provided in the used corpus.

3. Proposed Framework

Figure 1 illustrates our proposed framework, which consists of three main modules. The framework incorporates a static corpus collection comprising three different types of languages such as ((1)Urdu,(2)English,(3)Roman-Urdu). In the first module, when the user submits a query in Urdu, the model retrieves results in Urdu because it already possesses a collection in the Urdu language. Similarly, if the user submits a query in English or Roman-Urdu, they will receive results in the respective query language.

The second module addresses the scenario where the user submits a query in Urdu but desires results in English. In this case, the framework utilizes the Google Translate API to translate the query into English. The translated query is then matched against the English collection, and relevant results are retrieved. The proposed paradigm initially translates the query into the language of the document collection (Urdu) in the third module if the user submits a query in English and requests results in Urdu. The converted query is then used to retrieve desired and relevant results by matching it with the Urdu collection. Overall, our framework facilitates cross-linguistic information retrieval by allowing users to input queries in different languages and retrieve results in the same language or in a desired target language. It leverages translation capabilities and language matching to bridge the gap between different languages and provide users with access to relevant information in their preferred language.

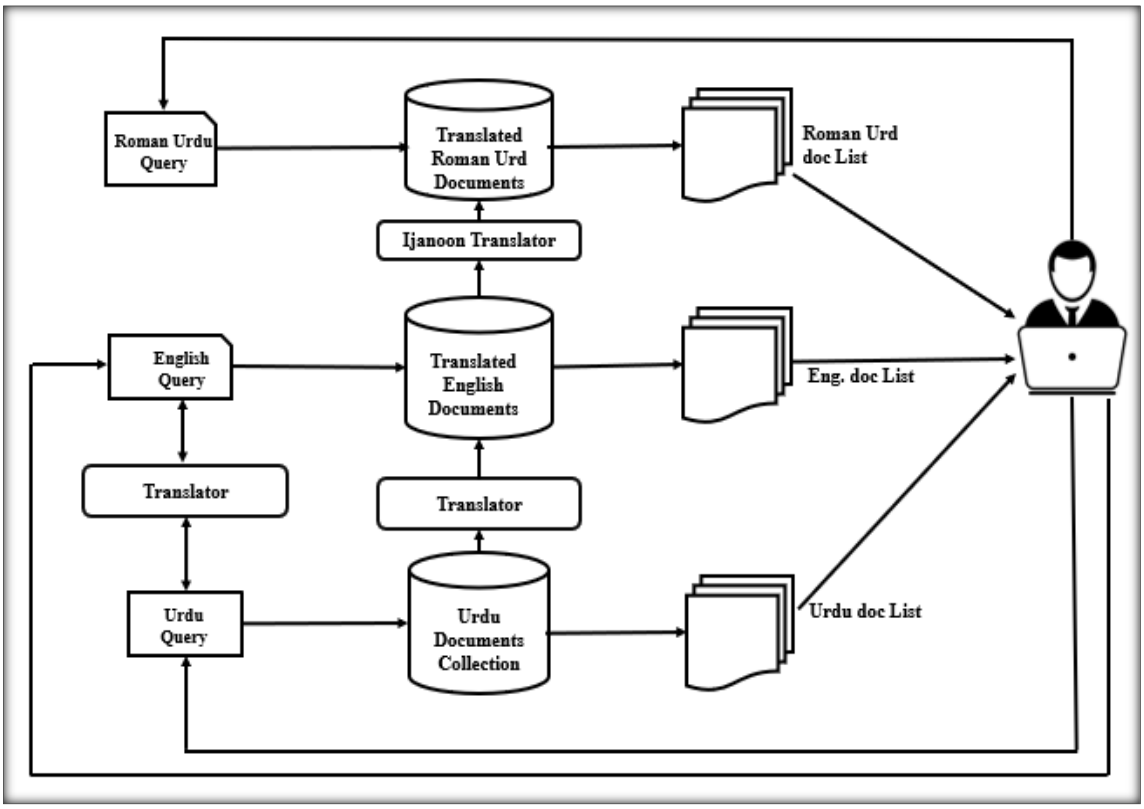

Figure 1. Framework of Model

The test collection for corpus UIR-21 consists of three components, the first component is a base collection of Urdu News Documents from which the appropriate information must be extracted, and second component is a set of Urdu Enquires used to discover relevant information and and final component is a relevance of individual pair judgements.

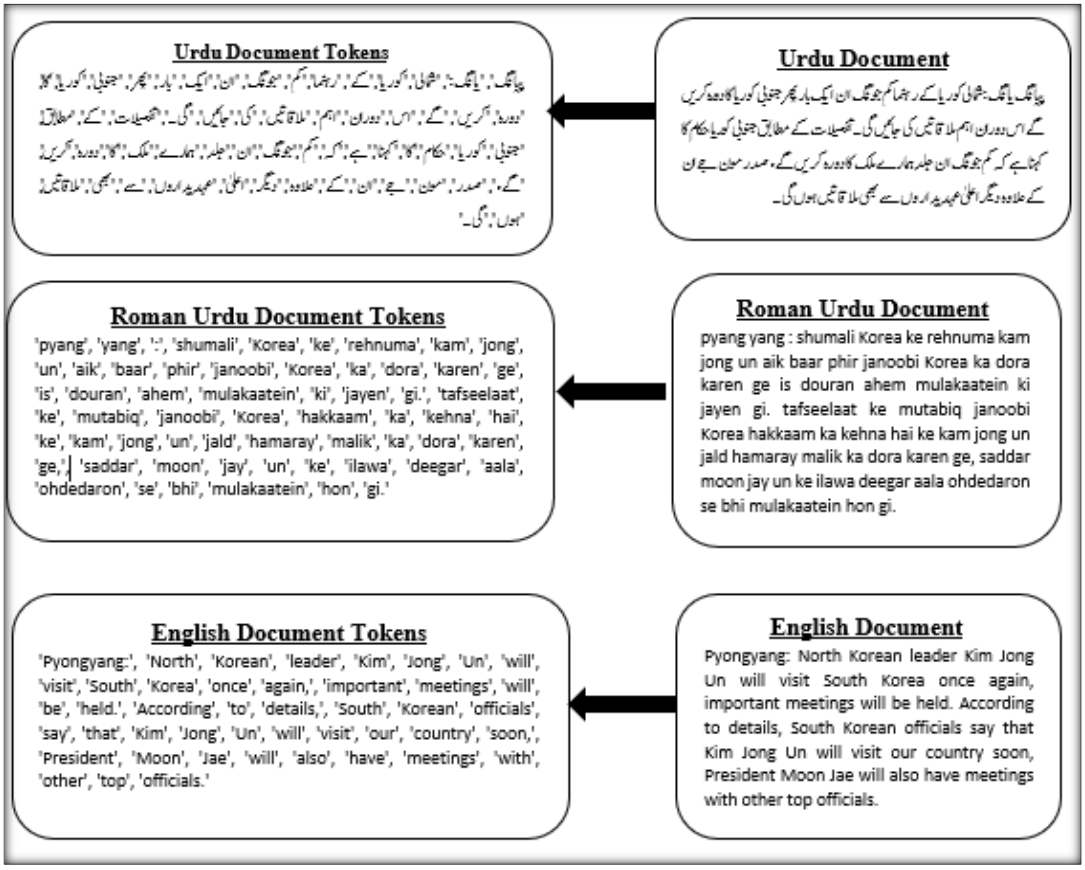

Figure 2. Steps of Preprocessing

Three texts (Urdu, translated English, and Roma-Urdu) are shown in (Figure 2) together with the tokens that are used to match them with the query tokens. First, the model reads each text file in the collection of documents one at a time, creates tokens, and stores them in the vector before matching the query tokens with the document collection tokens. Three languages were supported by our suggested model: English, Urdu, and Urdu-Roman. There are three options for submitting the query: English, Roman-Urdu, and Urdu. The proper language is chosen for the response to each enquiry.

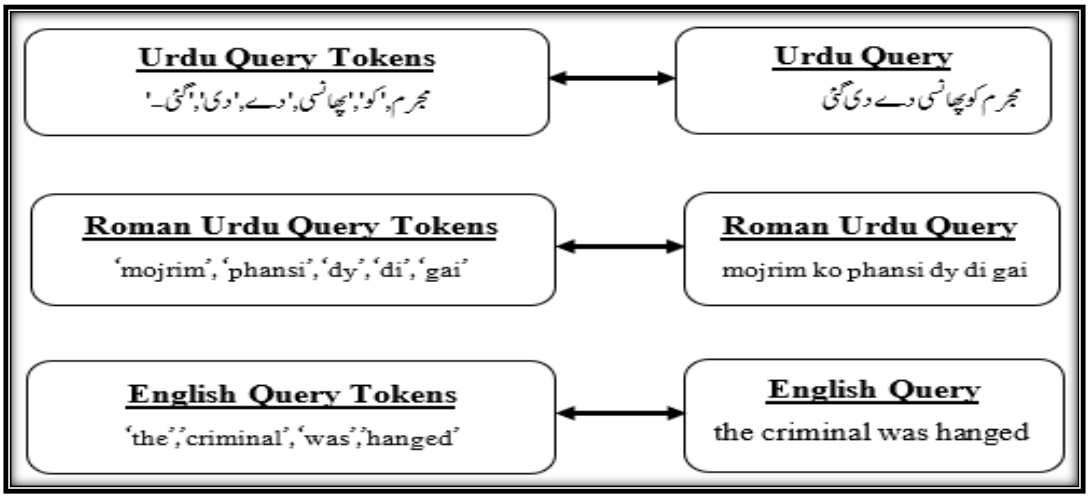

Figure 3. Query_Preprocessing

Figure 3 illustrates how the query was prepared before being matched with the documents. The prepared query is shown here along with its tokens. The first box shows the Urdu query, and the second

box shows the tokens that can be used to match the query with the tokens from the documents collection. The query is presented in Roman-Urdu in the second box, and the documents collection's corresponding tokens are displayed in the adjacent box. The English query is contained in the third box, and the English query tokens are contained in the parallel box and can be used to match the tokens from the documents collection.

4. Results and Discussion

Here are the findings from our trial. The "Urdu Retrieval Model (URM), English Retrieval Model (ERM), and Roman-Urdu Retrieval Model (RURM)" compare the documents list Relevance Judgement (RJ) to the retrieved documents.

4.1 Performance evaluation of URM vs RJ

The results in the (Table 2) shows the total number of Retrieved Documents (RD), Retrieved (RD), and the total number of RD for each query to allow users to verify the accuracy of the URM model. The total number of papers found shows that the Q6 found the most (60), while the Q3 found the least (three). The Q5 is able to find 18 total relevant papers, however the Q1, Q2, and Q3 can only find 2, respectively, the fewest RD. The total No of pertinent documents for each query is shown in the table's last column.

Table 2. Relevancy_Judgment vs Urdu_Retrieved

Sr.	No. Retrieved Doc	Total Retrived RD	Total No. RD
Q3	3	2	3
Q1	4	2	3
Q2	4	2	3
Q4	19	3	3
Q5	24	18	25
Q7	40	6	7
Q6	60	4	6

The graph of each query's precision, recall, and F-1 Score is shown in (Figure 4). The graph demonstrates that query 5 provides the most precision with a score of 75%, whereas query 6 only provides the minimal precision with a score of 6%. The Q4 provides the maximum recall with a score of 100%, and the Q3 provides the minimum recall with a score of 60%. The F-1 Score of Q5 provides the highest level of accuracy (67%), and the F-1 Score of Q6 provides the lowest level of accuracy (12%).

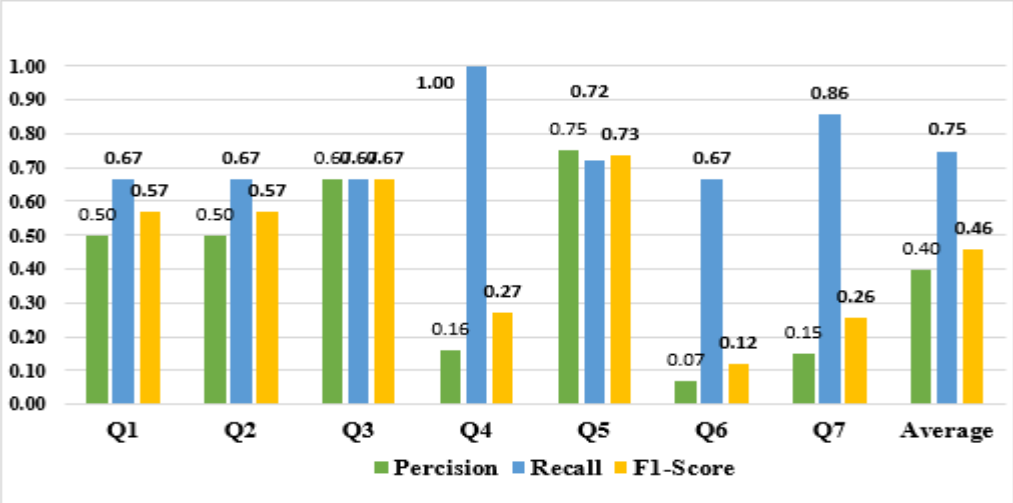

Figure 4. Relevancy_Judgment vs Urdu_Retried

4.2 Performance evaluation of ERM vs RJ

To assess the accuracy of the ERM model, (Table 3) displays the total number of RD, Relevant RD, and the total number of RD for each query. Total RD indicate that the Q6 retrieved the most documents, 64, while the Q2 retrieved the fewest documents, 2, according to the data. The Q5 is able to retrieve the most relevant papers, with 21, while the Q1, Q2, and Q3 can only retrieve the fewest relevant documents, with only 2. The last column of the table lists the total number of RD for each query. Figure 5 depicts a graph with each query's precision, recall, and F-1 score.

Table 3. Relevancy_Judgment vs English_Retrived

Sr.	No. Retrieved Doc	Total Retrived RD	Total No. RD
Q2	2	2	3
Q3	3	2	3
Q1	4	2	3
Q4	26	3	3
Q5	26	21	25
Q7	28	5	7
Q6	64	5	6

The graph in (Figure 5)demonstrates that Q2 performed well in terms of precision with a result of 100%, Q4 gives better results with recall of 100%, and Q5 provided the best performance on the F-1 Score with a result of 80%. The graph demonstrates that Q6 performs poorly in terms of precision with an accuracy of only eight percent, Q1 performs poorly in terms of recall with a score of only sixty seven percent, and Q6 provides the lowest result on the F-1 Score with a score of just fourteen percent.

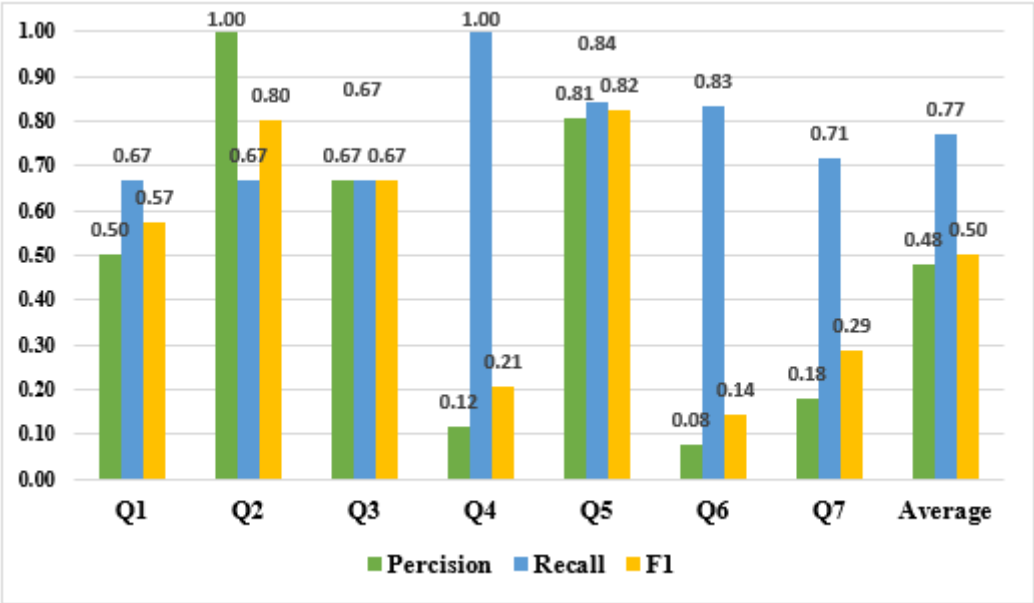

Figure 5. Relevancy_Judgment vs English_Retrieved

4.3 Performance evaluation of RURM vs RJ

To evaluate the RURM model's accuracy, (Table 4) lists the total number of RD, Relevant_RD, and the total number of RD for each query. Total RD indicate that the Q6 retrieved the most documents (50), while the Q3 retrieved the fewest documents. The Q5 retrieves the most relevant papers, a total of 16, while the

Q1, Q2, and Q3 retrieve the fewest, a total of two, relevant materials. The last column of the table lists the total number of RDfor each query.

Table 4. Relevancy_Judgment vs Roman_Retrieved

Sr.	No. Retrieved Doc	Total Retrived RD	Total No. RD
Q3	2	2	3
Q1	3	2	3
Q2	3	2	3
Q4	14	3	3
Q5	20	16	20
Q7	31	5	7
Q6	50	4	6

The graph in (Figure 6) shows that Q6 performed poorly in recall with a score of just 67 percent, Q1 and Q6 performed poorly in precision with an accuracy of just 2%, and Q6 provided the lowest result on the F-1 Score with a score of just 14 percent.

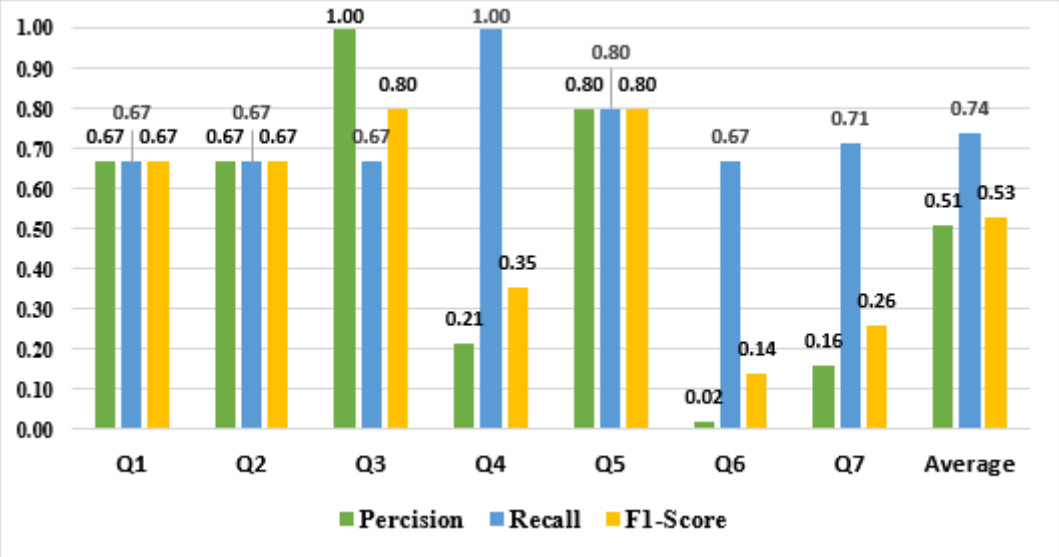

Figure 6. Relevancy_Judgment vs Roman_Retrieved

Figure 6 displays a graph with each query's (precision, recall,&F-1 score). "The graph demonstrates that Q3 performed well in terms of precision with a result of 100%, Q5 performed well in terms of recall with a result of 80%", and Q5 provided the best F-1 Score result with a result of 80%.4.4 Average

4.4 Comparison of precision, recall and F-1

This section compares the three models that have been proposed using the dataset's relevancy assessment. Table 5 provides each model's precision, recall, and F-1 Score. The model RURM provides 62 percent of the maximum precision whereas the model URM offers 40 percent of the minimum precision. The model ERM has the highest recall rate (77%) while the model RURM has the lowest; the model RURM provides the best F-1 Score (58%) while the model URM provides the lowest F-1 Score (46%).

Table 5. RJ vs other Models

Used_Models	Precision	Recall	F-1 Score

RURM vs RJ	62%	74%	58%
URM vs RJ	40%	75%	46%
ERM vs RJ	48%	77%	50%

After reviewing the IRS models and their results once accuracy is attained, we arrived to the following findings. The Precision, Recall, and F-1 Score of the IRS Model influence the model's accuracy.

- The Urdu Retrieval Model achieves the Best Precision (URM).
- Using Google Translate works better than the English Retrieval Model (ERM).
- To maintain current performance, the precision, recall, and F1-score of each model are assessed.
- Roman-Urdu retrieval modal and Urdu retrieval model were shown to be highly correlated.
- The model performs both well and poorly in the recall test.
- The F-1 score of the positive class is greater than that of the negative class.

5. Conclusions

Users who used CLIR systems could retrieve data that was written in languages other than the ones they used in their search terms. It could transmit requests in one language and receive papers in another. There are over 163 million native speakers of Urdu worldwide, which is also the national language of Pakistan. Due to the great number of morphological characteristics and speakers of Urdu, the research community required to pay special attention to (IR). Because there was no efficient CLIR in the Urdu-English domain, the task was done by creating a hybrid query model for cross-linguistic context optimisation.

The exhibited CLIR paradigm accepts inquiries in Urdu, English, and Roman-Urdu and provides documents in both a monolingual and cross-lingual context (Urdu to English and vice versa). Each IR system must be evaluated and compared using the TREC, which is made up of the corpus, the query, and its relevance assessment. To evaluate the suggested model, we computed the precision, recall, and f-1 score. The roman Urdu retrieval model (RU) had the highest preciosion in contrast to the Urdu retrieval model (URM), which had the lowest preciosion. The English retrieval model (ERM) had the highest remember (77%), whereas the Roman-Urdu retrieval model (RURM) had the lowest (74%).

Data availability: The Dataset that translated from urdu to english or urdo to ruman is available on demand.

Competing Interests: No financial or interpersonal conflicts on the research presented in this study.

Funding: It is declared that this work is not funded by any agency.

References

1. Shaukat, S., Shaukat, A., Shahzad, K., & Daud, A. (2022). Using TREC for developing semantic information retrieval benchmark for Urdu. Information Processing & Management, 59(3), 102939.
2. Ghanbari, E., & Shakery, A. (2022). A Learning to rank framework based on cross-lingual loss function for cross-lingual information retrieval. Applied Intelligence, 52(3), 3156-3174.
3. Ghanbari, E., & Shakery, A. (2019). Query-dependent learning to rank for cross-lingual information retrieval. Knowledge and Information Systems, 59, 711-743.
4. Ayaz, B., Altaf, W., Sadiq, F., Ahmed, H., & Ismail, M. A. (2016, December). Novel Mania: A semantic search engine for Urdu. In 2016 International Conference on Open Source Systems & Technologies (ICOSST) (pp. 42-47). IEEE.
5. Nekemte, E., & Addis Ababa, E. A Cross Lingual Information Retrieval (CLIR) System for Afaan Oromo-English using a Corpus Based Approach.
6. Bi, T., Yao, L., Yang, B., Zhang, H., Luo, W., & Chen, B. (2020). Constraint translation candidates: A bridge between neural query translation and cross-lingual information retrieval. arXiv preprint arXiv:2010.13658.
7. Maryamah, M., Arifin, A. Z., Sarno, R., & Hasan, A. M. (2021, January). Adapting Google Translate using Dictionary and Word Embedding for Arabic-Indonesian Cross-lingual Information Retrieval. In 2020 IEEE International Conference on Internet of Things and Intelligence System (IoTaIS) (pp. 205-209). IEEE.
8. Saleh, S., & Pecina, P. (2020, July). Document translation vs. query translation for cross-lingual information retrieval in the medical domain. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 6849-6860).
9. Zbib, R., Zhao, L., Karakos, D., Hartmann, W., DeYoung, J., Huang, Z., ... & Makhoul, J. (2019, July). Neural-network lexical translation for cross-lingual IR from text and speech. In Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 645-654).
10. Faruqui, M., Majumder, P., & Padó, S. (2011, November). Soundex-based translation correction in urdu–english cross-language information retrieval. In Proceedings of the Fifth International Workshop On Cross Lingual Information Access (pp. 25-29).
11. Boschee, E., Barry, J., Billa, J., Freedman, M., Gowda, T., Lignos, C., ... & Miller, S. (2019, July). SARAL: A low-resource cross-lingual domain-focused information retrieval system for effective rapid document triage. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations (pp. 19-24).
12. Picchi, E. & Peters, C, (2000) "Cross-language information retrieval: a system for comparable corpus querying, in Cross-Language Information Retrieval," G. Greenstreet, Editor. 2000, Kluwer Academic Publishing: Massachusetts. pp. 81-90.
13. Lawrie, D., Mayfield, J., Oared, D., & Yang, E. (2022). HC4: A New Suite of Test Collections for Ad Hoc CLIR. arXiv preprint arXiv:2201.09992.
14. S. Sia, "CLIReval: Evaluating Machine Translation as a Cross-Lingual Information Retrieval Task," pp. 134–141, 2020.
15. Li, J., Liu, C., Wang, J., Bing, L., Li, H., Liu, X., ... & Yan, R. (2020, April). Cross-lingual low-resource set-to-description retrieval for global e-commerce. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 34, No. 05, pp. 8212-8219).
16. Zhao, L., Zbib, R., Jiang, Z., Karakos, D., & Huang, Z. (2019, November). Weakly supervised attentional model for low resource ad-hoc cross-lingual information retrieval. In Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019) (pp. 259-264).
17. Zhang, L., Karakos, D., Hartmann, W., Srivastava, M., Tarlin, L., Akodes, D., ... & Makhoul, J. (2020, May). The 2019 bbn cross-lingual information retrieval system. In Proceedings of the workshop on Cross-Language Search and Summarization of Text and Speech (CLSSTS2020) (pp. 44-51).
18. Yao, L., Yang, B., Zhang, H., Luo, W., & Chen, B. (2020). Exploiting neural query translation into cross lingual information retrieval. arXiv preprint arXiv:2010.13659.
19. Rasheed, I., Banka, H., & Khan, H. M. (2021). Building a text collection for Urdu information retrieval. ETRI Journal, 43(5), 856-868.
20. Maqbool, M. S., Hanif, I., Iqbal, S., Basit, A., & Shabbir, A. (2022). Optimized Feature Extraction and Cross-Lingual Text Reuse Detection using Ensemble Machine Learning Models.
21. Maryamah, M., Arifin, A. Z., Sarno, R., & Hasan, A. M. (2021, January). Adapting Google Translate using Dictionary and Word Embedding for Arabic-Indonesian Cross-lingual Information Retrieval. In 2020 IEEE International Conference on Internet of Things and Intelligence System (IoTaIS) (pp. 205-209). IEEE.
22. Faruqui, M., Majumder, P., & Padó, S. (2011, November). Soundex-based translation correction in urdu–english cross-language information retrieval. In Proceedings of the Fifth International Workshop On Cross Lingual Information Access (pp. 25-29).
23. Yousaf, F., Iqbal, S., Fatima, N., Kousar, T., & Rahim, M. S. M. (2023). Multi-class disease detection using deep learning and human brain medical imaging. Biomedical Signal Processing and Control, 85, 104875.
24. Fazal, U., Khan, M., Maqbool, M. S., Bibi, H., & Nazeer, R. (2023). Sentiment Analysis of Omicron Tweets by using Machine Learning Models.
25. Sansone, C., & Sperlí, G. (2022). Legal Information Retrieval systems: State-of-the-art and open issues. Information Systems, 106, 101967.
26. Ateyah, S., & Al-Augby, S. (2023, March). Proposed information retrieval systems using LDA topic modeling for answer finding of COVID 19 pandemic: A brief survey of approaches and techniques. In AIP Conference Proceedings (Vol. 2591, No. 1). AIP Publishing.

27. Krieg, K., Parada-Cabaleiro, E., Medicus, G., Lesota, O., Schedl, M., & Rekabsaz, N. (2023, March). Grep-BiasIR: A Dataset for Investigating Gender Representation Bias in Information Retrieval Results. In Proceedings of the 2023 Conference on Human Information Interaction and Retrieval (pp. 444-448).
28. Zhou, T., Li, Y., Zhang, Y., & Wang, L. (2022). Pattern Matching Method for Q&A Information Retrieval System. In Advances in Intelligent Information Hiding and Multimedia Signal Processing: Proceeding of the IIH-MSP 2021 & FITAT 2021, Kaohsiung, Taiwan, Volume 2 (pp. 101-112). Singapore: Springer Nature Singapore.
29. Sharma, A., & Kumar, S. (2023). Machine learning and ontology-based novel semantic document indexing for information retrieval. Computers & Industrial Engineering, 176, 108940.
30. Ogundepo, O., Zhang, X., Sun, S., Duh, K., & Lin, J. (2022, December). AfriCLIRMatrix: Enabling cross-lingual information retrieval for african languages. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (pp. 8721-8728).