International Journal of Science and Research Archive
eISSN: 2582-8185
Cross Ref DOI: 10.30574/ijsra
Journal homepage: https://ijsra.net/

(RESEARCH ARTICLE)

# MarQO: A query optimizer in multilingual environment for information retrieval in Marathi language

Suhas D. Pachpande [1,*] and Parag U. Bhalchandra [2]

[1] *Department of Computer Science, Sant Gadge Baba Amravati University, Amravati, India.*
[2] *School of Computational Sciences, S.R.T.M. University, Nanded, India.*

International Journal of Science and Research Archive, 2023, 09(02), 986–996

Publication history: Received on 15 July 2023; revised on 24 August 2023; accepted on 27 August 2023

Article DOI: https://doi.org/10.30574/ijsra.2023.9.2.0712

## Abstract

Information retrieval is a crucial component of modern information systems. A significant portion of the vast amount of information stored worldwide is in local languages. While most information retrieval systems are designed primarily for English, there is a growing need for these systems to work with data in languages other than English. Cross Language Information Retrieval (CLIR) systems play a pivotal role in enabling information retrieval across multiple languages. However, these systems often face challenges due to ambiguities in query translation, impacting retrieval accuracy. This paper introduces "MarQO," a query optimizer designed to address these challenges in the context of *Marathi* language. MarQO employs a multi-stage approach, including lexical processing, extraction of multi-word terms, synonym addition, phrasal translations, utilization of word co-occurrence statistics, and more. By disambiguating query keyword translations, MarQO significantly improves the accuracy of translations, thereby leading to more relevant document retrieval results.

**Keywords:** Cross Language Information Retrieval; Multi-word terms; Phrasal translation

## 1. Introduction

### 1.1. Cross Language Information Retrieval (CLIR) and the Importance of Translation

The introduction lays the groundwork by highlighting the role of information retrieval in contemporary information systems. It emphasizes the need for CLIR systems to bridge the language gap and allow effective retrieval of information in languages other than English (Suhas and Parag 2020).Cross Language Information Retrieval (CLIR) is a pivotal aspect of modern information systems that seeks to overcome the linguistic barriers inherent in our diverse global information landscape. As the world becomes increasingly interconnected, the volume of information generated and stored in various languages has surged (Hadni et al., 2014; Mahalakshmi et al., 2022; Safdar et al., 2020). This multilingual reality underscores the importance of CLIR, which is essentially the practice of enabling information retrieval systems to effectively search and retrieve data across languages different from the system's native language, typically English (Jacques Savoy2004, Monika and Sudha 2015, Suhas et. al., 2022).

In the current era of information abundance, retrieving relevant and accurate information swiftly is of paramount importance (Archana et al., 2023). However, the majority of conventional information retrieval systems are developed with a primary focus on English. This poses a significant challenge when users seek information in languages other than English. Herein lies the crucial role of CLIR systems – to bridge this language gap and facilitate the seamless retrieval of information regardless of the language in which it is stored (Chauhan et al., 2023; Lomonosov MSU, Moscow, Russia and Potemkin, 2019; Spasic, 2018).

*Corresponding author: Suhas D. Pachpande

The significance of translation within the realm of CLIR cannot be overstated. Translation serves as the linchpin that connects individuals to the information they seek, irrespective of linguistic barriers (Suhas et. al., 2022). It transforms a query or search request in one language into queries that match content in another language (Spasic, 2018). This process demands more than just literal word-for-word translation; it requires an understanding of the semantics, context, and cultural nuances that underpin language usage (Dan and Daqing 2008).

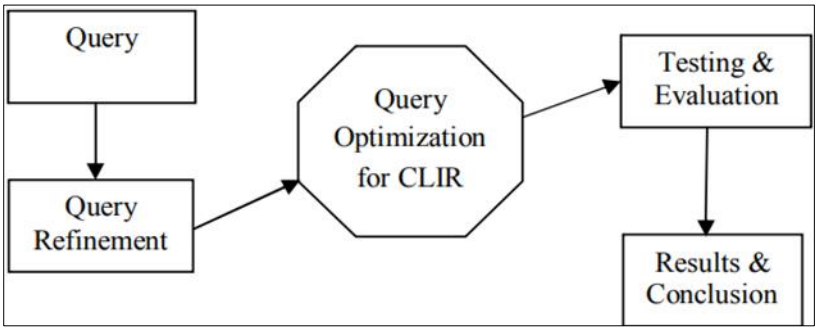



**Figure 1** Probable Model for CLIR

The role of CLIR is to empower users to access information in languages that are unfamiliar to them, thus broadening the accessibility of knowledge across the global landscape. It enables researchers, professionals, students, and individuals from all walks of life to explore resources and insights from different cultures and languages, fostering a more inclusive and diverse information ecosystem (Turid Hedlund 2003). In addition to its practical applications, CLIR also carries implications for diplomacy, international relations, business, and academia. Governments, businesses, and organizations can utilize CLIR to monitor global trends, sentiment, and opinions across languages, aiding in decision-making and strategic planning. Academics can benefit from the ability to access research literature in various languages, thereby enriching their understanding of diverse perspectives and methodologies (Pu-Jen Cheng et al., 2004).

To sum up, CLIR stands as a technological bridge that transcends linguistic barriers and brings together the rich tapestry of human knowledge stored in different languages. As the world continues to interconnect, the importance of CLIR and accurate translation becomes increasingly evident (Jian-Yun 2003). It enables information systems to truly serve as conduits of global understanding and collaboration, fostering a world where knowledge knows no boundaries.

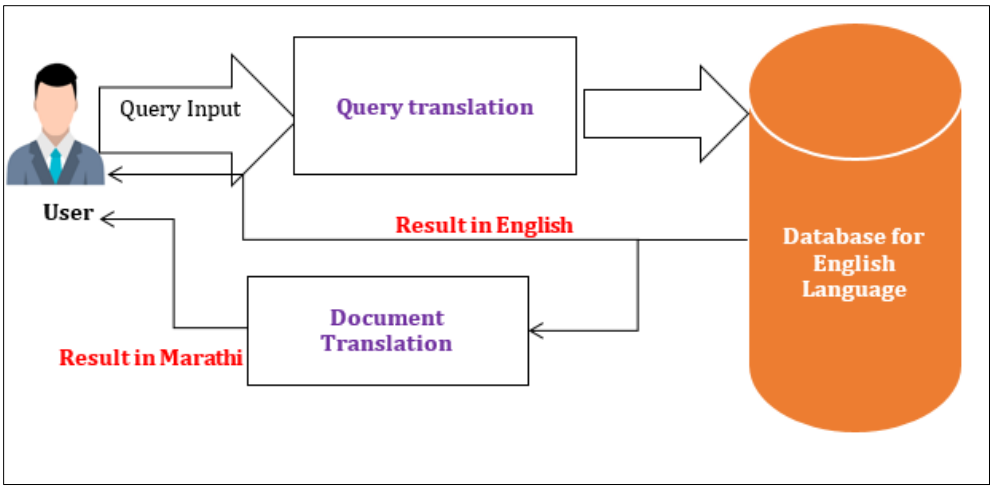



**Figure 2** Cross Lingual Text Retrieval from Marathi to English

## 2. Challenges in Translation

### 2.1. Challenges in Translation within Cross Language Information Retrieval (CLIR) Systems

Translation, particularly within the context of CLIR systems, introduces a host of intricate challenges that necessitate careful consideration and innovative solutions. This section delves into the scientific intricacies of these challenges, highlighting the complexities associated with accurate and contextually relevant translation.

*2.1.1. Out of Vocabulary (OOV) Terms*

One of the foremost challenges in translation lies in dealing with Out of Vocabulary (OOV) terms. OOV terms are words or phrases that do not have direct equivalents in the target language's vocabulary. In CLIR systems, encountering OOV terms is common due to the vast linguistic diversity present in the world's languages. These terms could be domain-specific, colloquial, or entirely novel concepts. Translating OOV terms requires a robust understanding of the term's context and intent to generate meaningful and accurate translations. Addressing OOV terms is critical to ensure comprehensive information retrieval, as their presence can lead to information loss and retrieval inaccuracies (Robert et al., 2016).

*2.1.2. Translation Ambiguity*

Translation ambiguity stems from the inherent nature of language, where words and phrases can carry multiple meanings depending on context. This challenge is amplified when translating across languages with distinct linguistic structures and nuances. The potential for misunderstanding arises due to polysemy (multiple meanings for a single word) and homonymy (different words with identical spellings/pronunciations) (Korn´el et al., 2005). Resolving translation ambiguity demands a deep understanding of the source language's contextual cues to accurately infer the intended meaning and select the appropriate target language equivalent.

*2.1.3. Multi-Word Terms (MWTs) and Phrases*

Multi-Word Terms (MWTs) and phrases add a layer of complexity to translation. These units often carry specific, domain-specific, or culturally ingrained meanings that cannot be deciphered by directly translating individual words. The significance of MWTs transcends the sum of their parts, necessitating the translation of their collective meaning rather than just individual word translations. Achieving this requires a profound understanding of the term's semantics, contextual connotations, and potential cultural variations. The challenge is to preserve the intended meaning while ensuring coherence in the target language (Wessel et al., 2009).

*2.1.4. Contextual Preservation*

Preserving contextual meanings across languages is a fundamental objective of translation, particularly within CLIR systems. Words and phrases derive much of their meaning from the surrounding text and cultural nuances (Dan and Daqing 2008). Transferring this meaning accurately is challenging, as the same word might hold different connotations in different contexts. Contextual preservation involves not only selecting appropriate synonyms or equivalents but also understanding the subtle shifts in meaning that can occur during translation. This challenge is amplified when translating idiomatic expressions, metaphors, and cultural references.

## 3. Handling multi-word terms (MWTs) and associated meanings

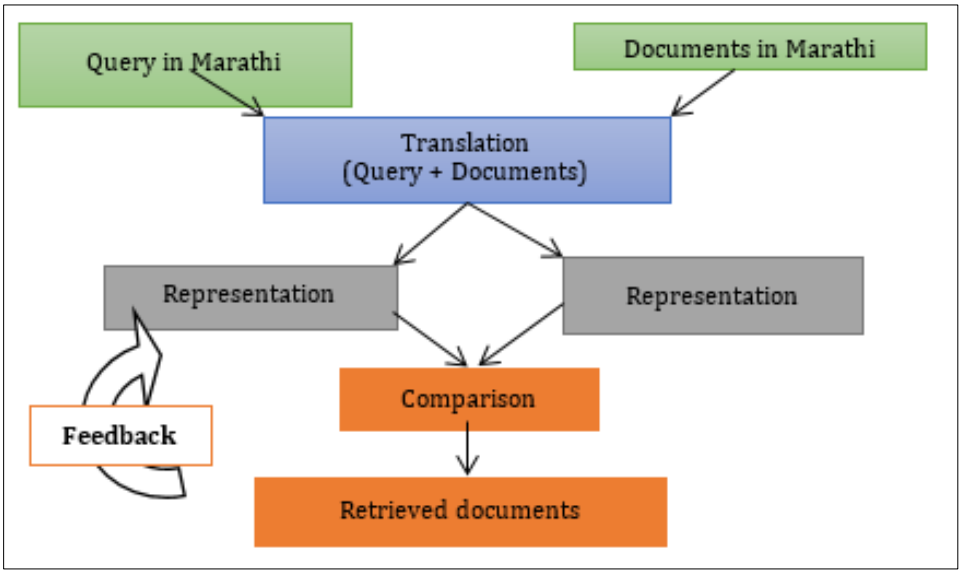



**Figure 3** Typical architecture of a CLIR system

In this section, the paper delves into the intricate task of managing Multi-Word Terms (MWTs) and the nuanced meanings they carry. MWTs are expressions formed by combining multiple words that often carry specific meanings distinct from their individual constituents. Translating MWTs requires a sophisticated understanding of the underlying concepts and their interrelationships (Ismail et al., 2004, Cristina et al., 2005). The paper explores strategies to tackle this challenge, emphasizing the need for contextual comprehension during translation. It underscores that the meaning of a phrase extends beyond mere word translation; it encompasses the collective significance that emerges from how words interact within a specific context. By capturing the holistic essence of MWTs, translation efforts ensure that the intended message is accurately conveyed, preserving the contextual depth and ensuring that information retrieval remains accurate and meaningful across languages (Krister Lind´en 2006, Mandeep *et al.,* 2014).

## 4. Query Disambiguation

This section introduces the pivotal concept of query disambiguation, a fundamental step in refining query translation accuracy within CLIR systems. The paper elucidates the various techniques used to resolve ambiguities present in translated queries. Ambiguities can arise due to the multiple meanings a word may hold, and disambiguation involves selecting the most contextually appropriate meaning. The spotlight is on MarQO's contribution to this process, showcasing its role in enhancing the precision of translated queries. By dissecting the intricacies of ambiguous terms and employing context-based analysis, MarQO addresses the challenge of query ambiguity and raises the accuracy of translated queries, thereby improving the overall retrieval process (Jacques Savoy 2004, Cristina et al., 2005).

## 5. Phrasal Translation

This section immerses into the technique of phrasal translation, an essential component of accurate cross-language information retrieval. Phrasal translation involves translating multi-word phrases as cohesive units rather than treating them as isolated words. The paper presents a spectrum of methods to achieve precise phrasal translations, emphasizing the significant role MarQO plays in this context. Accurate phrasal translation ensures that the essence, nuances, and contextual meaning of the original phrases are preserved, contributing to more meaningful information retrieval across languages. MarQO's contributions to refining phrasal translation techniques underscore its value in enhancing the quality of translation outputs.

## 6. Utilizing Parallel Corpora for Phrasal Translation and Addressing OOV Terms

This segment delves into the strategic utilization of parallel corpora, bilingual datasets containing equivalent texts in multiple languages, to enhance translation accuracy. The paper explores how parallel corpora aid in refining phrasal translation by providing aligned examples for contextually accurate translations. Moreover, this section highlights the role of parallel corpora in addressing Out of Vocabulary (OOV) terms, terms without direct equivalents in the target language. By analyzing parallel corpora, MarQO can identify contextually suitable translations for OOV terms, enriching the quality of translated output and enhancing information retrieval precision.

## 7. Experimental Work on Marathi Dataset using Vector Space Model (VSM)

This part showcases the empirical experimentation conducted on a Marathi dataset using the Vector Space Model (VSM) for information retrieval. The paper presents the results achieved by MarQO in terms of various metrics such as Term Frequency (TF), Document Frequency (DF), Inverse Document Frequency (IDF), and Cosine Similarity. These metrics collectively demonstrate MarQO's effectiveness in enhancing the accuracy and relevance of retrieved documents in the Marathi language.

Two versions of CLIR models were implemented and tested

- Basic CLIR model
- CLIR with MarQO

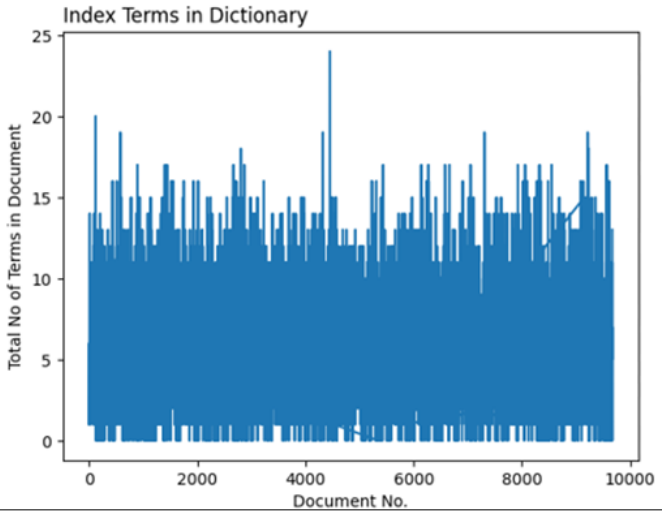
The model is expected to retrieve documents containing desired information from English as well as Marathi datasets. Although the basic CLIR model retrieves relevant documents from English dataset, it fails to retrieve all relevant documents from Marathi due to improper translations. Whereas the other model which incorporates the proposed MarQO can retrieve more optimized results because of its ability to handle OOV terms, Multi word Terms, Phrasal translation, etc.

The experimental results are demonstrated below

**Table 1** No. of index terms in documents

**Term Frequency (TF)**

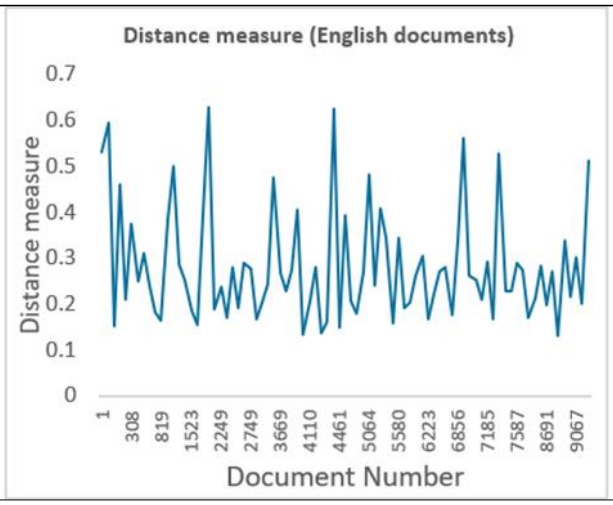
| Document No | No. of Index terms |
|---|---|
| 1 | 1 |
| 2 | 6 |
| 3 | 6 |
| 4 | 4 |
| 5 | 5 |
| … | … |
| … | … |
| … | … |
| 9672 | 7 |
| 9673 | 10 |
| 9674 | 13 |




### 7.1. Document Weights compute using distance measure vector (English)

**Table 2** Distance measure for documents

| Document No | Distance measure |
|---|---|
| 1 | 0.53191 |
| 164 | 0.59412 |
| 184 | 0.15214 |
| 242 | 0.46029 |
| 290 | 0.21006 |
| … | … |
| … | … |
| … | … |
| 9067 | 0.30099 |
| 9110 | 0.20149 |
| 9165 | 0.51240 |



### 7.2. Retrieve result (English Datasets)

**Table 3** Retrieval results (Cosine Similarity)

The final results retrieved by the model contain the list of documents ranking according to degree of relevance.

| Document Rank | Distance measure |
|---|---|
| 2060 | 0.62954 |
| 4377 | 0.62528 |
| 164 | 0.59412 |
| 6948 | 0.56330 |
| 1 | 0.53191 |
| … | … |
| … | … |
| … | … |
| 4348 | 0.13853 |
| 3963 | 0.13416 |
| 8871 | 0.13231 |

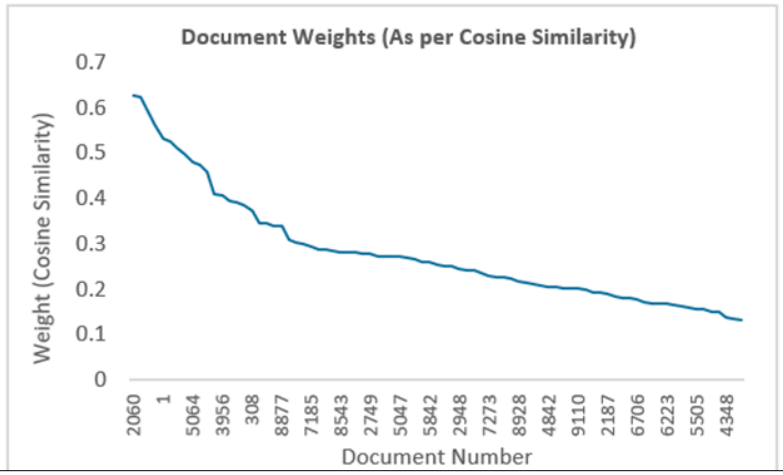



**Table 4** Distance measure (for Marathi Documents)

**Cross lingual information retrieval model (without Query Optimization)**

a) Results on Marathi Dataset: Using same query in English translated into Marathi

Document weights computed using the distance measure vector:

| Document Number | Distance measure |
|---|---|
| 35 | 0.16095 |
| 76 | 0.18050 |
| 97 | 0.38317 |
| 114 | 0.17544 |
| 150 | 0.54417 |
| … | … |
| … | … |
| … | … |
| 9389 | 0.20479 |
| 9501 | 0.19946 |
| 9618 | 0.19400 |

Table 4: Distance measure (for Marathi documents)

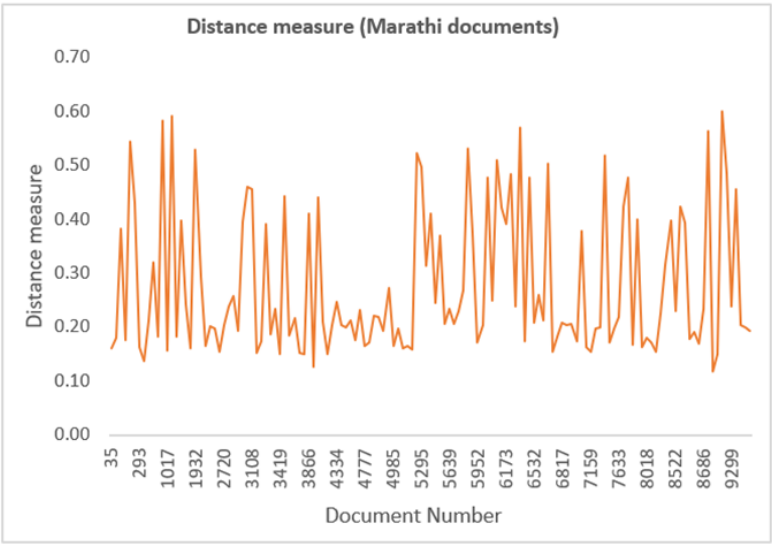


**Table 5** Retrieval results- Cosine Similarity (for Marathi documents)

**Retrieval Results:**

The final results retrieved by the model contain the list of documents ranking according to degree of relevance.

| Document Number | Weight |
|---|---|
| 9045 | 0.60118 |
| 1432 | 0.59231 |
| 924 | 0.58334 |
| 6311 | 0.57123 |
| 8694 | 0.56503 |
| … | … |
| … | … |
| … | … |
| 363 | 0.13705 |
| 3963 | 0.12608 |
| 8705 | 0.11782 |

Table 5: Retrieval Results - Cosine Similarity (for Marathi documents)

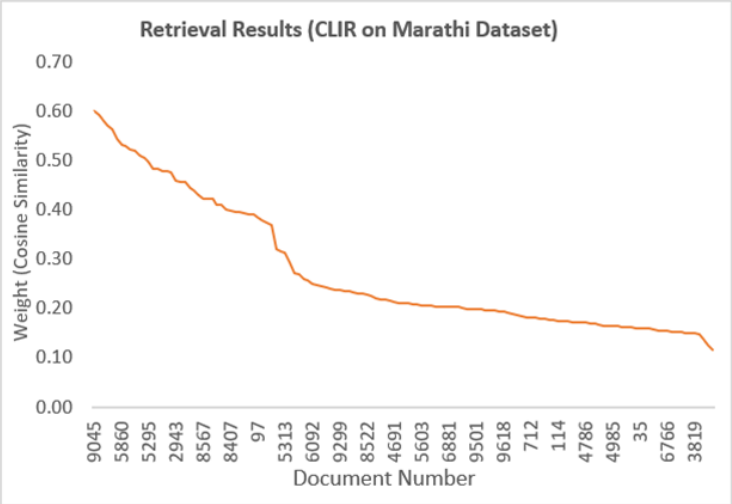



**Table 6** Distance measure (for Marathi documents)

**Cross lingual information retrieval model with Query Optimization**

Results on same Marathi Dataset:

Document weights computed using the distance measure vector:

| Document Number | Weight |
|---|---|
| 26 | 0.18001 |
| 35 | 0.19293 |
| 76 | 0.41507 |
| 97 | 0.18952 |
| 114 | 0.51431 |
| … | … |
| … | … |
| … | … |
| 9618 | 0.22795 |
| 9673 | 0.52938 |
| 3054 | 0.08026 |

Table 6: Distance measure (for Marathi documents)

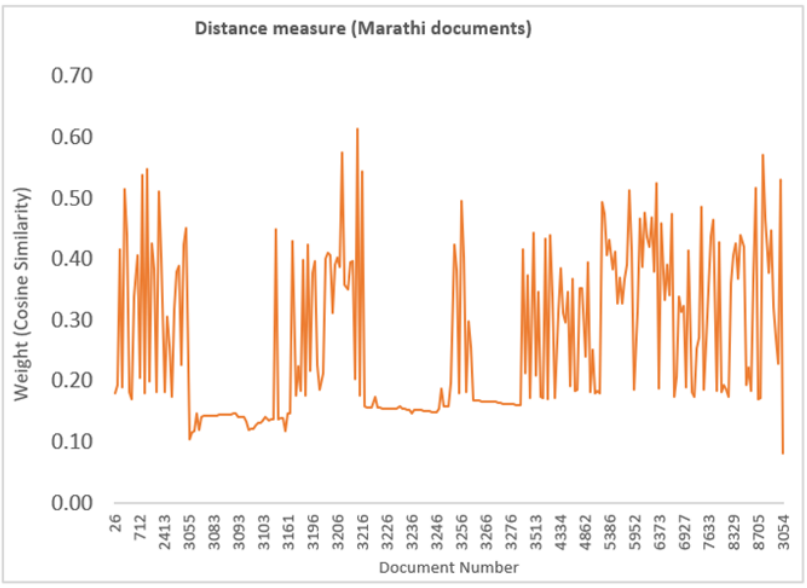


**Table 7** Retrieval results- Cosine Similarity (for Marathi documents)

**Retrieval Results:**

The final results retrieved by the model contain the list of documents ranking according to degree of relevance.

| Document Number | Weight |
|---|---|
| 3214 | 0.61332 |
| 3208 | 0.57534 |
| 9045 | 0.56990 |
| 1432 | 0.54754 |
| 3216 | 0.54394 |
| … | … |
| … | … |
| … | … |
| 3074 | 0.11562 |
| 3055 | 0.10416 |
| 3054 | 0.08026 |

Table 7: Retrieval Results - Cosine Similarity (for Marathi documents)

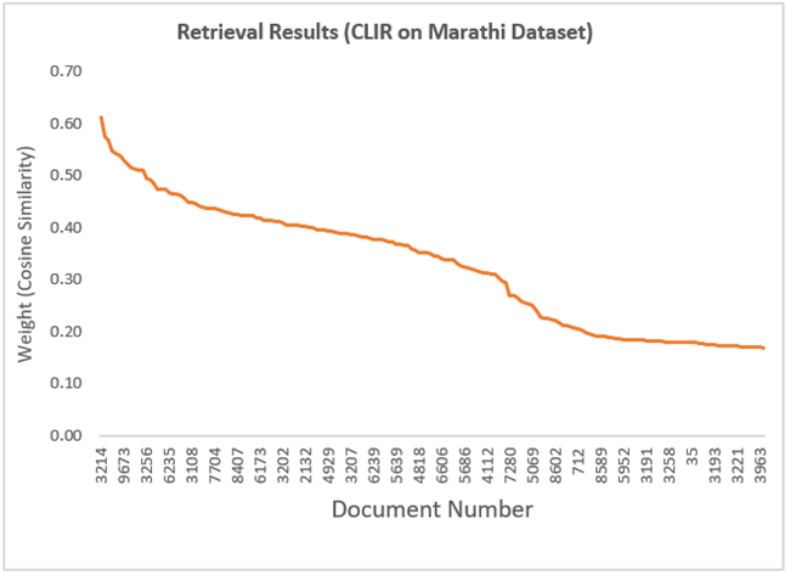
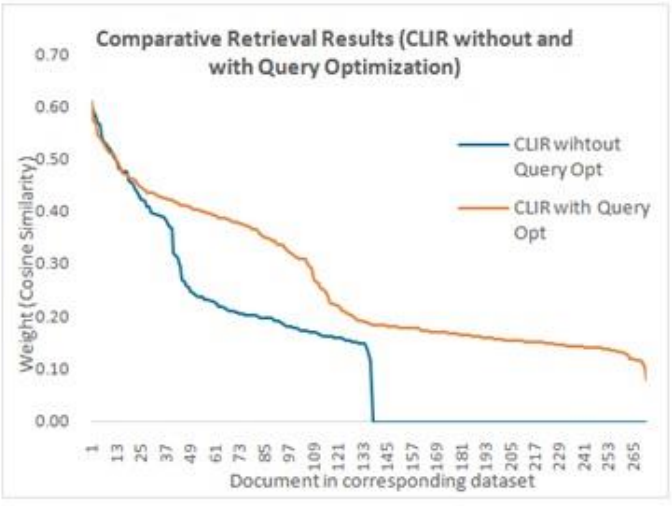



**Table 8** Retrieval results-cosine Similarity (for English and Marathi documents)

**Comparison of performance of CLIR Model without Query Optimization and CLIR model with Query Optimization**

| S. No. of result | CLIR without Query Optimization (Weight ) | CLIR with Query Optimization (Weight ) |
|---|---|---|
| 1 | 0.60118 | 0.61332 |
| 2 | 0.59231 | 0.57534 |
| 3 | 0.58334 | 0.56990 |
| 4 | 0.57123 | 0.54754 |
| 5 | 0.56503 | 0.54394 |
| … | … | … |
| … | … | … |
| … | … | … |
| 268 | 0.00000 | 0.11706 |
| 269 | 0.00000 | 0.11562 |
| 270 | 0.00000 | 0.10416 |
| 271 | 0.00000 | 0.08026 |




## 8. Evaluating Performance: Precision, Recall and Visual Representation

The final section critically evaluates MarQO's performance using established metrics like Precision and Recall. These metrics gauge the precision of retrieved information and the comprehensiveness of the retrieval process, respectively.

The paper further employs visual representations, such as charts, to provide a tangible and accessible illustration of MarQO's impact on improving retrieval results. This visual presentation enhances the paper's clarity and reinforces the quantitative enhancements MarQO brings to cross-language information retrieval.

**Table 9** Ten quires execution for Precision and Recall for datasets

**Precision and Recall for ten queries executed using the same dataset:**

| Query No | No of relevant documents in dataset | No of relevant documents retrieved | Total No of retrieved documents | Precision | Recall |
|---|---|---|---|---|---|
| 1 | 86 | 82 | 86.1 | 95.2380952 | 95.3488372 |
| 2 | 396 | 374 | 396.44 | 94.3396226 | 94.4444444 |
| 3 | 144 | 137 | 144.535 | 94.7867299 | 95.1388889 |
| 4 | 158 | 145 | 149.06 | 97.2762646 | 91.7721519 |
| 5 | 280 | 271 | 281.84 | 96.1538462 | 96.7857143 |
| 6 | 177 | 163 | 170.172 | 95.7854406 | 92.0903955 |
| 7 | 396 | 362 | 387.34 | 93.4579439 | 91.4141414 |
| 8 | 29 | 27 | 29.16 | 92.5925926 | 93.1034483 |
| 9 | 51 | 45 | 46.575 | 96.6183575 | 88.2352941 |
| 10 | 21 | 20 | 21.02 | 95.1474786 | 95.2380952 |

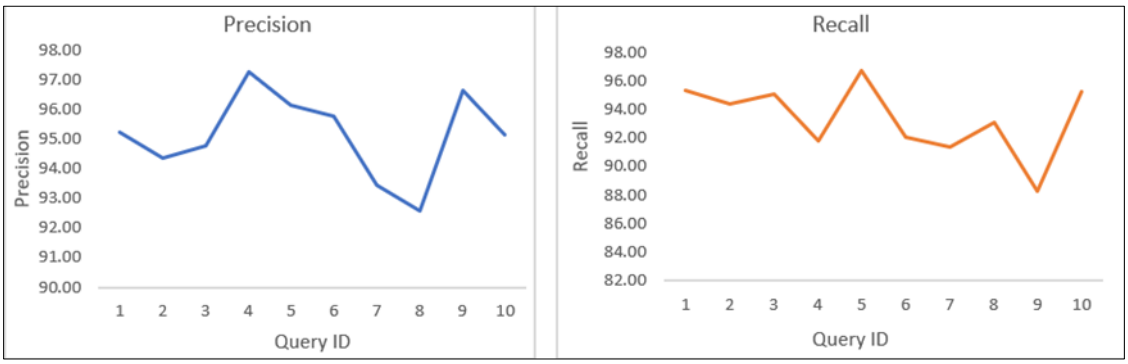



**Figure 4** Representation of Precision and Recall dataset

## 9. Conclusion

Article sheds light on the importance of accurate translation in multilingual information retrieval, particularly in the Marathi language context. The introduction of MarQO as a query optimizer demonstrates its effectiveness in overcoming challenges associated with query translation and enhancing retrieval accuracy. By following a comprehensive approach that encompasses various stages of optimization, MarQO serves as a valuable tool in improving the quality of cross-lingual information retrieval systems.

## Compliance with ethical standards

### *Disclosure of conflict of interest*

No conflict of interest to be disclosed.

## References


[1] Chauhan, D.S., Singh, G.V., Ekbal, A., Bhattacharyya, P., (2023). : A Multilingual Humor-aided Multiparty Dialogue Generation in multimodal conversational setting. Knowledge-Based Systems 278, 110840. https://doi.org/10.1016/j.knosys.2023.110840

[2] Cristina Lopez, Vicente P. GuemeroBote (2005), University of Extremadura, Badajoz, Spain, Proc. Of ACM SigmodIntl.conf.of Management of Data. 14 (1) 14-20.

[3] Dan Wu, Daqing He (2008), ICE-TEA: an Interactive Cross-language Search Engine with Translation Enhancement, SIGIR'08 (ACM), 20–24, 882-84.

[4] Hadni, M., Lachkar, A., Ouatik, S.A., (2014). Multi-Word Term Extraction Based on New Hybrid Approach for Arabic Language, in: Computer Science & Information Technology (CS & IT). Presented at the Second International Conference on Computational Science and Engineering, Academy & Industry Research Collaboration Center (AIRCC), pp. 109–120. https://doi.org/10.5121/csit.2014.4410

[5] Ismail H. Toroslu , Ahmet Cosar (2004) , Dynamic programming solution for multiple query optimization problem, Information Processing Letters, Science Direct , Elsevier 92 149-155.

[6] Jacques Savoy (2004), Combining Multiple Strategies for Effective Monolingual and Cross Language Retrieval, Information Retrieval, 8 (7) 121-148.

[7] Jian-Yun Nie (2003), Cross-Language Information Retrieval, IEEE Computational Intelligence Bulletin, 6 (4) 19-23.

[8] Korn´el Mark´o, Stefan Schulz, OlenaMedelyan, OlenaMedelyan (2005), Bootstrapping Dictionaries for Cross Language Information Retrieval, SIGIR'05 (ACM), 15 (19), 528-535.

[9] Krister Lind´en, (2006), Multilingual modeling of cross-lingual spelling variants, Journal of Information Retrieval, Springer, 4, 295-310.

[10] Lomonosov MSU, Moscow, Russia, Potemkin, S., (2019). Multiword Terms and Machine Translation, in: Proceedings of the Third International Conference, Europhras 2019, Computational and Corpus-Based Phraseology. Presented at the Third International Conference, Europhras 2019, Computational and Corpus-Based Phraseology, Editions Tradulex, Geneva, pp. 133–139. https://doi.org/10.26615/978-2-9701095-6-3_018

[11] Mahalakshmi, P., Sabiyath Fatima, N., Balaji, R., Patel, M.J., (2022). An effective multilingual retrieval with query optimization using deep learning technique. Advances in Engineering Software 173, 103244. https://doi.org/10.1016/j.advengsoft.2022.103244

[12] Mandeep Pannu, Anne James, Robert Bird (2014), A Comparison of Information Retrieval Models, WCCCE'14 (ACM), 4 (9), 22-29.

[13] Monika Sharma, SudhaMorwal (2015), A Survey on Cross Language Information Retrieval, International Journal of Advanced Research in Computer and Communication Engineering 4 (2), 384-387.

[14] Pu-Jen Cheng, Jei-Wen Teng, Ruei-Cheng Chen, Jenq-Haur Wang, Wen-Hsiang Lu, Lee-Feng Chien (2004), Translating Unknown Queries with Web Corpora for CrossLanguage Information Retrieval, SIGIR'04 ACM 25 (9) 146-153.

[15] Robert Krajewski, HenrykRybinski, Marek Kozlowski (2016), A novel method for dictionary translation, Journal of Intelligent Information Systems, Springer, 47, 491-514.

[16] Safdar, Z., Bajwa, R.S., Hussain, S., Abdullah, H.B., Safdar, K., Draz, U., (2020). The role of Roman Urdu in multilingual information retrieval: A regional study. The Journal of Academic Librarianship 46, 102258. https://doi.org/10.1016/j.acalib.2020.102258

[17] Spasic, I., (2018). Acronyms as an Integral Part of Multi-Word Term Recognition – A Token of Appreciation. IEEE Access 6, 8351–8363. https://doi.org/10.1109/ACCESS.2018.2807122

[18] Suhas D. Pachpande, Parag U. Bhalchandra (2020). Cross Language Information Retrieval (CLIR): A Survey of Approaches for Exploring Web Across Languages. International Journal of Innovative Technology and Exploring Engineering (IJITEE), Volume-10 Issue-1, November 2020, 326-332.

[19] Suhas D. Pachpande, Parag U. Bhalchandra and Ashok Gingine (2022). Framework of an Expert System for Intelligent Information Retrieval across Languages using CLIR Techniques. AJOMC – Special issue on Research in Applied Science, Management and Technology, Vol. 7 No. 1. pp. 1662-1667

[20] Turid Hedlund (2003), Dictionary-based Cross-Language Information Retrieval, Thesis submitted for Academic dissertation in the University of Tampere, 1- 84.

[21] U. Archana, A. Khan, A. Sudarshanam, C. Sathya, A. K. Koshariya and R. Krishnamoorthy (2023). "Plant Disease Detection using ResNet," International Conference on Inventive Computation Technologies (ICICT), Lalitpur, Nepal, 2023, pp. 614-618, IEEE 01 June 2023 doi: 10.1109/ICICT57646.2023.10133938. https://ieeexplore.ieee.org/document/10133938

[22] Wessel Kraaij, Jian-Yun Nie, Michel Simard (2009), Embedding Web-Based Statistical Translation Models in Cross-Language Information Retrieval, Computational Linguistics 29 (3), 381-419.