Language-Preference-Based Re-ranking for Multilingual Swahili Information Retrieval

Joseph P. Telemala
Sokoine University of Agriculture
Morogoro, Tanzania
josephmasamaki@sua.ac.tz

Hussein Suleman
University of Cape Town
Cape town, Western Cape, South Africa
hussein@cs.uct.ac.za

ABSTRACT

Approaches for merging results in multilingual information retrieval (MLIR) systems strive for topical relevance, regardless of whether they are heuristic or machine learning (ML)-based. However, to build on topical relevance, current MLIR results merging approaches largely ignore other factors derived from user interaction behaviours, which, if used, could potentially improve the relevance of the merged results. MLIR user behaviour studies suggest that users' language preferences differ depending on the topic of search. In this paper, we propose to use language preferences driven by search topics, i.e., topic-language (T-L) preferences. Specifically, we create a T-L-based algorithm for merging results in a multilingual Swahili IR system. The approach promotes a certain number of results in the preferred language to the top of the results list, while the remaining results in the preferred language and those in the non-preferred language are interleaved in a round-robin fashion. Using a multilingual Swahili IR data set, the evaluation results show that the T-L-based approach improves the relevance of results for T-L preference-sensitive topics in general. Our findings also show that the T-L-based approach outperforms the other approaches for queries with a strong T-L association. According to these findings, incorporating user behaviour into the merging equation in MLIR systems has the potential to improve the relevance of results for some topics.

CCS CONCEPTS

• **Information systems → Information retrieval**; Specialized information retrieval; Structure and multilingual text search; **Multilingual and cross-lingual retrieval**;

KEYWORDS

Topic-Language preferences, Re-ranking, Swahili-speaking, Multilingual Information Retrieval, Swahili Information Retrieval.

ACM Reference Format:
Joseph P. Telemala and Hussein Suleman. 2022. Language-Preference-Based Re-ranking for Multilingual Swahili Information Retrieval. In *Proceedings of the 2022 ACM SIGIR International Conference on the Theory of Information Retrieval (ICTIR '22), July 11–12, 2022, Madrid, Spain.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3539813.3545131

ICTIR '22, July 11–12, 2022, Madrid, Spain.

ACM ISBN 978-1-4503-9412-3/22/07.
https://doi.org/10.1145/3539813.3545131

1 INTRODUCTION

This study focuses on the Swahili-speaking Web users in Tanzania, a multilingual country where English and Kiswahili are official languages. However, citizens' use of the two languages in daily business varies greatly. It is unusual to see people conversing in English on the streets or in office corridors, even among highly educated people, government officials, university students, and judicial officers. Specifically, Kiswahili dominates almost all domains of life as a communication language, such as politics, mass media, local business, worship, and literature, whereas English is used alongside Kiswahili for education and commerce [16]. The Swahili-speaking community in Tanzania has a clear distinction in how the languages are used, with English serving as merely a language of record (documents). This is in contrast to other multilingual communities around the world, which may exhibit different characteristics, such as equal language use in all aspects of life. The Swahili-speaking community in Kenya, for example, can use both Kiswahili and English equally in parliamentary debates, whereas this is not the case in Tanzania.

This distinction in how Swahili speakers in Tanzania use the two languages complicates the mechanism by which users interact with information on the Web. This exacerbates the consequences and implications for MLIR design and implementation. Given that polyglots are mostly fluent in multiple languages, the MLIR system design is centered on achieving topical relevance while ignoring other factors. However, given the use of English and Swahili in Tanzania, the information needs for work-related tasks and the language of information they use may differ from those for non-work-related tasks. The unique characteristics of Tanzania's polyglots necessitate a study that not only focuses on achieving topical relevance, but also incorporates human behaviours to improve the relevance of MLIR system results.

Various studies in MLIR, such as by Ling et al. [12], Steichen and Lowe [25], and our previous study (Telemala and Suleman [26]), suggest that users have language preferences for certain topics of search. For example, in our study [26], it was observed that while users significantly preferred Swahili results for the *Music* topic, they significantly preferred English results for the *Computer* hardware topic. We can refer to these preferences as topic-language (T-L) associations/preferences. The current MLIR results merging approaches do not incorporate T-L preferences in search result ranking. As a result, the system hides potential relevant results further down the list, and users either miss them or expend extra effort to find them. Users are more likely to be satisfied if they see top-ranked results in the language they want/prefer right away, rather than having to scroll [14].

In this paper, therefore, we propose and evaluate a *T-L-based* approach for merging or re-ranking results that uses the users' language preferences. The proposed T-L-based approach primarily re-ranks a small set of the top ranked results before presenting them to the users, while the remaining results are interleaved. Because users are typically interested in the first few results, it is critical to ensure that the first few results present the most relevant documents. Thus, we demonstrate the types of scenarios where the T-L preferences can and cannot improve the relevance of ranked results in MLIR. The evaluation of our proposed T-L-based approach is based on a multilingual Swahili IR system. Thus, this paper addresses the primary research question, which asks: *How can topic-language preferences improve the relevance of ranked results in a multilingual Swahili information retrieval (IR) system?* Our answer to this question forms the contribution of this paper through the following research objectives: i) to assess the overall performance of the proposed T-L-based approach in T-L association-sensitive topics; ii) to examine the factors that influence the performance of the T-L-based approach in T-L association-sensitive topics; and iii) to determine whether there is a minimum threshold of promoted results that can provide optimal performance of the T-L-based approach in T-L association-sensitive topics.

The following is the paper's organizational structure. Section 2 presents the related literature. The proposed T-L-based algorithm is presented in Section 3. This is followed by an experimental setup for evaluating the T-L-based algorithm in Section 4. The evaluation results are presented and discussed in Section 5. Section 6 provides a summary and concludes the paper.

2 LITERATURE REVIEW

2.1 Traditional Merging Approaches

One family of traditional MLIR merging strategies employ document comparable scores and/or ranks. Prominent approaches include: round-robin [24, 31], where the merged list is obtained by interleaving a single result from each of the intermediate result lists until all of the intermediate result lists are exhausted; raw-score [24], where the documents are arranged in a descending order of raw relevance scores; and normalized score [11, 20, 23], where there exists a variety of normalization styles, such as *normalized-by-top1* [20, 23], which divides each document's score value by the highest score in the list, and *normalized-by-topk* [11] where normalization is based on a cut-off point determined by a certain number of documents with the highest scores. They all assume either a similar distribution of relevant documents across individual collections or that individual collection scores are comparable [21].

Other traditional approaches utilize more latent information from the retrieved result lists or individual sub-collections. This group includes: the weighted-score [19, 22], and sub-collection based merging techniques [3, 10, 15]. The weighted-score merging approaches assign a score to each document in the collection based on its relevance and the corpus to which it belongs [19, 22]. High-scoring relevant documents from a low-scoring corpus rank lower than relevant high-scoring documents from a high-scoring corpus. Another approach called centralized architecture for MLIR results merging [17] combines all of the documents in the targeted collections into a single corpus. Each document in the mixed documents collection is then assigned a language tag, and both the original and translated queries are tagged with their respective languages. As a result of the mixed document corpus having a single central index, the retrieval process can proceed normally, as in classical IR. The primary advantage of traditional approaches is that they can be used in low-resource MLIR settings.

2.2 Machine Learning Merging Approaches in MLIR

The application of machine learning (ML) approaches to solve ranking problems is known as learning to rank (LTR) [14]. LTR is widely used in federated search, distributed IR, and meta-search for merging/fusing results from various information sources and/or search engines. Some works on LTR for results merging used hand-crafted features, which are then used to train the ranking/merging models [6, 27, 29]. One approach to developing the feature list is to create it based on the documents' similarities with the search query, where a joint relevance probability is used [6]. Another approach used by Tsai and colleagues is to extract named-entity, document length, and the number of query terms [27, 29]. To learn the weights of these features, the authors used a LTR algorithm called FRANK [28]. The learned weights for each feature were used in combination with the BM25 ranking model scores to calculate the final ranking score for each document, then the documents were sorted based on these scores to generate the final ranked list.

A semi-supervised approach based on the multi-view architecture [30] proposed to consider each language in the collection as a *view* of a document. ML approaches are typically data hungry, requiring a large amount of data to train, validate, and test. Unfortunately, not all languages, such as Kiswahili, have a large number of resources to train on, making heuristic approaches continue to a viable option for low resourced languages. However, the success of transformers-based models (e.g. BERT [5]), has opened up the possibility of transferring rankers trained in resource-rich languages such as English to other languages using multilingual encoders. For example, Litschko et al.'s [13] recent work, tested on Kiswahili and Somali, suggests improving ranking competitively to machine-translation-based models.

3 TOPIC-LANGUAGE-BASED APPROACH

The T-L-based algorithm (see Algorithm 1) for T-L preference-sensitive topics is based on two simple ideas of *promoting* and *interleaving*, where the preferred language's top results are pushed to the top of the results list. The number of promoted results (batch size) n can be varied i.e., from 1, 2, 3,..., 10. The remaining results in the preferred language and those in the non-preferred language are then interleaved in a round-robin style until the result lists are exhausted. The T-L-based approach aims to present (more) results in a preferred language for a query in a specific topic first.

3.1 Formulations and Analysis

We assume an MLIR system that supports two languages – languages A and B – where if one language is preferred for a topic, the other language is referred to as a non-preferred language for that topic. There are two possible scenarios when comparing the relevance of results between the preferred and non-preferred language

Algorithm 1: T-L-Based Approach

```
if T-L preferences then
    specify the language to start with;
    specify the batch size e.g., 3;
    push the results in a specified batch size from the
      preferred language to the top of the merged list;
    Interleave the remaining results between the ones in a
      preferred language and the non-preferred in a
      round-robin, based on a language chosen to start with;
    Iterate through the lists until all the results are
      exhausted;
    Terminate and return the merged list;
Result: An interleaved result list per T-L Preferences
if no T-L preferences then
    choose the language to start with;
    interleave the results using round-robin-based approach
      with a batch size of 1;
    Terminate and return the merged list;
Result: An interleaved results list
```

results lists. First, the preferred language's results list contains equal or more relevant top *n* results than the non-preferred language's list. Second, the results list for the preferred language contains fewer or no relevant top *n* results than the results list for the non-preferred language.

3.1.1 Equal or More Relevant Top Results in the Preferred Language. Suppose language B is preferred over language A for a specific topic G; and suppose the top *n* results in language B's list are relevant, while we do not know which results are relevant in language A's list; pushing the top *n* results from language B's list to the top of the merged list guarantees that the T-L-based approach T achieves better MLIR ranking performance than any other system S, which does not promote results in a preferred language. Using examples, we want to show that for T to promote top *n* results from the preferred language implies that the overall performance, in terms of average precision (AP), of the merged list is better than that of another system S i.e., $AP(T) > AP(S)$. Since we do not know what top *n* results on language A's list are relevant and what are not, there are mainly three possibilities. First, all the top *n* results are relevant; second, all the top *n* results are irrelevant; and third, top *n* is a mixture of relevant and irrelevant results.

In the first case, if all of the top *n* results from language A's list are relevant, and given that all of the top *n* results from language B's list are also relevant, then $AP(T) = AP(S)$. In the second case, if all of the top *n* results from language A's list are irrelevant, and given that all of the top *n* results from language B's list are relevant, then $AP(T) > AP(S)$. In the third case, if top *n* results from language A's list is a mixture of relevant and irrelevant results, and given that top *n* results from language B's list are relevant, then $AP(T) \geq AP(S)$.

This analysis and the example in Table 3 (in Appendix A.1), which assumes that each list has two top *n* results and that the results in approach S are interleaved, generally means that the performance of the T-L-based approach will always be better or equal to that of any system S, i.e., $AP(T) \geq AP(S)$. The analysis demonstrates that the T-L-based approach improves relevance of ranked results of the MLIR system if the preferred language's results list contains more relevant results at the top than the non-preferred language list.

3.1.2 Few or No Relevant Top Results in the Preferred Language. Reversing our assumption in Section 3.1.1, we assume that the top *n* results in language B's list are irrelevant, while we do not know which results are relevant in language A's list. Using an example detailed in Table 4 (in the Appendix A.2), it can generally be seen that the performance of a T-L-based approach will always be poor or equal to that of other systems that do not promote results in a preferred language, i.e., $AP(T) \leq AP(S)$. Therefore, this analysis demonstrates that the T-L-based approach does not improve relevance of results in an MLIR system if the preferred language's results list contains fewer or no relevant results at the top compared to the non-preferred language list.

4 EVALUATION

4.1 Dataset

We used click-through log data from our controlled multilingual search engine, which supported English and Kiswahili. The multilingual Swahili speakers in Tanzania interacted with this system, and their click-through data was logged. To provide control over the topics of interest in our study, we created queries that were organized into topics, thanks to Google Trends [7] and Tanzania-specific Web directories such as Alexa [2], 123Tanzania [1], Yalwa [32], and the WWW Virtual Library [9]. We had a total of 123 topics and 1184 queries all pertaining to the Tanzania geographical location. The system that stored the topics and queries generated five randomly generated topics from the 123 topics for each user, after which the user could click on their topic of interest and specify the language in which they wanted to view the queries. Each displayed query contained an embedded link to the MLIR search engine, from which they could access non-simulated results using the MS Bing Web Search API. The Bing search engine generated the individual results lists and their original rankings.

For each user, the system displayed the results in an interleaving style, randomly alternating the language of results to begin with – English or Kiswahili. The instructions directed users to click (check on the most relevant result(s) based on the snippet assessment) using a checkbox on the left-hand side of each result. We collected the logs for queries and click-through. For this evaluation, we only used the click-through log data, which consisted of a total of 3493 query records. Following pre-processing, the click-through logs contained information about the query name, the topic to which it belongs, the language of each clicked result, and a list of the clicked and non-clicked URLs – which were treated as relevance judgements. We assumed that the users' judgments were absolute, so the relevance judgements were purely binary, i.e., a click implies a relevant result, otherwise not. We removed all queries and their associated clicked results for topics that did not have language preferences because users did not have language preferences for all of them. Pre-processing and removing topics with no language preference reduced the click-through log data to 99 and 45 unique query records for English and Swahili preferred topics, respectively.

4.2 Baseline

We used the *Round-Robin (R-R)* merging approach as the only baseline among the traditional merging approaches due to its ability to work with result lists that do not have comparable relevance scores, as we did not have access to relevance scores in this study's setting. This baseline acts as a representative of all other traditional results merging strategies, which essentially assume that the relevant documents are distributed uniformly across the individual result lists [21].

4.3 Measures, Notations and Analysis

The performance measures used in the evaluation were gain-based (i.e., Normalized Discounted Cumulative Gain (NDCG)) and precision-based (i.e., Average Precision (AP) and Mean Average Precision (MAP)). The notations R-R$_{En}$ and R-R$_{Sw}$ represent the Round-Robin (R-R) approach, where English and Kiswahili were the starting languages for interleaving. T-Ln_{En} and T-Ln_{Sw}, on the other hand, represent the T-L-based algorithm, with English and Kiswahili as the starting languages for interleaving, respectively, where n represents the number of promoted results, where $n \in \mathbb{Z} : n \in [1, 10]$. For averaged scores, we use the notations R-R$_{Av}$ and T-Ln_{Av}.

The query-level analysis took into account users' query-clicking behaviours (T-L association). The T-L association in the queries means that certain queries had their results clicked only in one language, others had their results clicked more in one language than the other, and still others had their results clicked in a language other than the one revealed as the preferred language. That is: i) clicking solely on English results; ii) clicking solely on Swahili results; iii) clicking an equal number of English and Swahili results, e.g., 1 English result and 1 Swahili result; iv) clicking more English than Swahili results, e.g., 2 English results and 1 Swahili result; v) and clicking more Swahili than English results, e.g., 1 English result and 3 Swahili results.

5 RESULTS AND DISCUSSIONS

5.1 Overall Performance of the Proposed T-L-Based Approach

In the first research objective, we wanted to "*assess the overall performance of the proposed T-L-based approach in T-L association-sensitive topics*". Thus, we divide the findings into two categories: performance of the T-L-based approach per preferred language in topics; and performance of the T-L-based approach in individual queries.

5.1.1 Performance at Topic Level. The findings in Table 1 show that the T-L-based approach generally improves MLIR performance by outperforming the baseline in almost all cases for both English and Swahili preferred topics. However, there were minor variations in T-L-based approach's performance, owing primarily to the metrics used, and the number of promoted results. The MAP scores, in particular, show a significant improvement from the baseline, as opposed to the NDCG scores. For example, considering the English preferred topics, while the MAP@10 score improved the performance by 33.5%, the NDCG@10 score improved the performance by only 0.9% for T-L3$_{Sw}$. The performance of the T-L-based algorithm for NDCG@5 slightly deteriorated as the number of promoted results increased. For example, considering Swahili preferred topics, when the algorithm promoted 4 and 5 English results (T-L4$_{En}$ and T-L5$_{En}$), the performance dropped by 3.4% and 22%, respectively. The T-L-based approach's blurred improvement when using the NDCG measure could be attributed to the nature of our dataset. MAP is better suited to the type of data we were working with (binary relevance judgement) than NDCG, which was originally designed for graded relevance [8, 18].

Our findings also revealed slight differences in results based on the starting language for interleaving. For example, Table 1 shows a minimal difference between T-L2$_{En}$ and T-L2$_{Sw}$ for MAP@5 scores, which were 0.73 and 0.72, respectively. One reason for such findings could be that the starting language for interleaving may be the same as the preferred language; in this case, the starting language for interleaving acts as a continuation of the preferred language. This means that, unlike the R-R algorithm, the performance of the T-L-based algorithm remains stable regardless of the starting language for interleaving. This means that once the results have been promoted in a preferred language, the starting language for interleaving is no longer relevant. This stability is reflected in the averaged scores; and as a result, one can use T-L-based approach with random choice of the starting language for interleaving, while the performance remains relatively better.

5.1.2 Performance at Query Level. The performance improvement by the T-L-based approach in topics was minor and/or deteriorated in several cases, particularly when the NDCG measure was used, as previously discussed. This necessitated a closer analysis for individual queries rather than relying on the analysis of queries bundled in a topic. To do this, we sorted and grouped queries based on how users clicked on their results/URLs, as explained in Section 4.3.

The results in Figure 1 show that the T-L-based approach performs well for queries that have all clicked results in the estimated preferred language, as well as those that have more clicked results in the estimated preferred language. For queries in English preferred topics, the T-L-based algorithm outperformed the R-R algorithm for queries with: i) purely clicked English results (En); and ii) more clicked English results than Swahili results (<En). For queries in Swahili preferred topics, the T-L-based approach improved performance for queries with: i) only clicked Swahili results (Sw); and ii) more clicked Swahili results than English results (<Sw).

The T-L-based approach performed poorly for queries with an equal number of clicked results from the two languages, more clicked results in the other language than the estimated preferred language, and all results in the other language than the estimated preferred language. For queries in English preferred topics, the T-L-based approach failed to improve results performance for queries with: i) an equal number of clicked English and Swahili results (Equal); ii) more clicked Swahili results than English results (<Sw); and iii) queries with entirely clicked Swahili results (Sw). For queries in Swahili preferred topics, the T-L-based approach failed to improve results performance for queries with: i) an equal number of clicked English and Swahili results (Equal); ii) more clicked English results than Swahili results (<En); and iii) queries with entirely clicked English results (En).

Table 1: The MAP@5, MAP@10 and NDCG@5, NDCG@10 scores for English, and Swahili preferred topics.

	English Preferred Topics				**Swahili Preferred Topics**			
	MAP@		NDCG@		MAP@		NDCG@	
	5	**10**	**5**	**10**	**5**	**10**	**5**	**10**
R-R$_{En}$	0.69	0.65	**0.64**	**0.69**	0.53	0.53	**0.51**	**0.61**
T-L1$_{En}$	0.72 (+5.7%)	0.67 (+3.7%)	0.64 (0.00%)	**0.69** (0.90%)	0.63 (+17.6%)	0.61 (+15.1%)	**0.53** (+2.7%)	**0.61** (0.00%)
T-L2$_{En}$	**0.73** (+5.9%)	**0.68** (+4.8%)	0.64 (-0.5%)	**0.69** (+0.9%)	**0.65** (+21.8%)	0.62 (+17.1%)	**0.53** (+2.7%)	**0.61** (+0.5%)
T-L3$_{En}$	0.72 (+5.5%)	0.67 (+3.8%)	0.64 (-0.5%)	**0.69** (+0.7%)	**0.65** (+21.2%)	0.62 (+18.1%)	0.50 (-3.4%)	**0.61** (+0.5%)
T-L4$_{En}$	0.71 (+3.9%)	0.67 (+3.6%)	0.59 (-8.8%)	**0.69** (+0.7%)	**0.65** (+21.6%)	**0.63** (+20.1%)	0.50 (-3.4%)	**0.61** (+0.7%)
T-L5$_{En}$	0.71 (+3.9%)	0.67 (+3.7%)	0.59 (-8.8%)	**0.69** (+0.1%)	**0.65** (+22.0%)	**0.63** (+20.1%)	0.40 (-22.0%)	**0.61** (+0.7%)
R-R$_{Sw}$	0.53	0.51	0.60	**0.69**	0.63	0.61	**0.53**	**0.61**
T-L1$_{Sw}$	0.69 (+29.7%)	0.65 (+27.4%)	**0.64** (+7.0%)	**0.69** (0.00%)	**0.65** (+3.5%)	0.62 (+1.9%)	0.52 (-0.4%)	**0.61** (0.30%)
T-L2$_{Sw}$	0.72 (+37.1%)	0.67 (+32.1%)	**0.64** (+7.0%)	**0.69** (+0.9%)	**0.65** (+3.1%)	0.62 (+2.6%)	0.50 (-5.9%)	**0.61** (+0.5%)
T-L3$_{Sw}$	**0.73** (+37.4%)	**0.68** (+33.5%)	**0.64** (+6.4%)	**0.69** (+0.9%)	**0.65** (+3.4%)	**0.63** (+4.4%)	0.50 (-5.9%)	**0.61** (+0.7%)
T-L4$_{Sw}$	0.72 (+36.9%)	0.67 (+32.3%)	**0.64** (+6.4%)	**0.69** (+0.7%)	**0.65** (+3.8%)	**0.63** (+4.4%)	0.40 (-24.0%)	**0.61** (+0.7%)
T-L5$_{Sw}$	0.71 (+34.8%)	0.67 (+32.0%)	**0.64** (+6.4%)	**0.69** (+0.7%)	**0.65** (+3.8%)	**0.63** (+3.9%)	0.40 (-24.0%)	**0.61** (+0.7%)
R-R$_{Av}$	0.61	0.58	0.62	**0.69**	0.58	0.57	0.52	**0.61**
T-L1$_{Av}$	0.70 (+16.1%)	0.66 (+14.1%)	**0.64** (+3.4%)	**0.69** (+0.4%)	0.63 (+9.00%)	0.61 (+7.50%)	**0.53** (+1.1%)	**0.61** (+0.1%)
T-L2$_{Av}$	**0.73** (+19.5%)	**0.68** (+16.8%)	**0.64** (+3.1%)	**0.69** (+0.9%)	**0.65** (+11.7%)	0.62 (+9.40%)	0.51 (-1.7%)	**0.61** (+0.5%)
T-L3$_{Av}$	0.72 (+19.4%)	**0.68** (+16.9%)	**0.64** (+2.8%)	**0.69** (+0.8%)	**0.65** (+11.6%)	**0.63** (+10.8%)	0.50 (-4.6%)	**0.61** (+0.6%)
T-L4$_{Av}$	0.72 (+18.3%)	0.67 (+16.2%)	0.61 (-1.5%)	**0.69** (+0.7%)	**0.65** (+12.0%)	**0.63** (+11.7%)	0.45 (-13.8%)	**0.61** (+0.7%)
T-L5$_{Av}$	0.71 (+17.4%)	0.67 (+16.2%)	0.61 (-1.5%)	**0.69** (+0.4%)	**0.65** (+12.2%)	**0.63** (+11.4%)	0.40 (-23.0%)	**0.61** (+0.7%)

The R-R$_{En}$ and R-R$_{Sw}$ are the R-R approach with English and Kiswahili used as the starting languages for interleaving, respectively. The T-L1$_{En}$, ..., T-L5$_{En}$ and T-L1$_{Sw}$, ..., T-L5$_{Sw}$ are the T-L-based approach with different number of promoted results ranging from 1 to 5 and English and Kiswahili are the starting languages for interleaving, respectively. R-R$_{Av}$ and T-Ln$_{Av}$ stand for the averaged R-R and T-Ln scores, respectively.

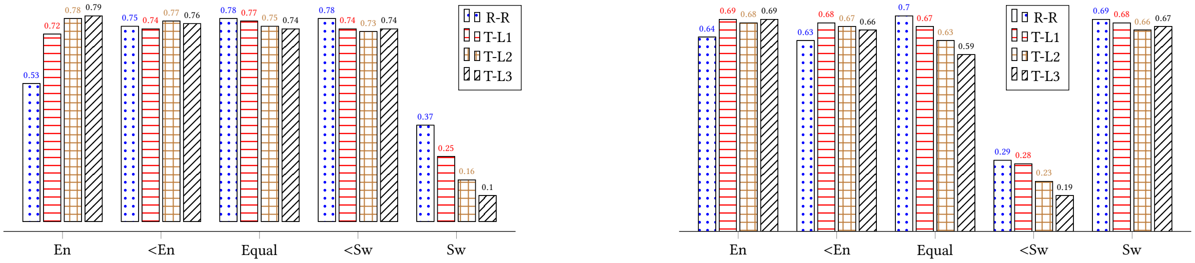
This generally shows that, for example, if the estimated preferred language for a specific topic is Kiswahili, then the T-L-based approach performs better if all or most of the results clicked for a given query are in Kiswahili; otherwise, the performance is poor. These findings are consistent with our analysis in Section 3.1 that, in order to improve T-L-based performance, the top *n* results in a preferred language must be relevant.

5.2 Factors Influencing the Performance of the T-L-based Approach

In the second research objective, we wanted to: "*examine the factors that influence the performance of the T-L-based approach in T-L association-sensitive topics.*" Performance of the T-L-based approach presented in the previous section suggests that it is only strong at the fine grained level of query as opposed to the abstract level of topic. The clear improvement made by the T-L-based algorithm in individual queries suggests that the performance of the T-L-based approach is primarily determined by the proportion of clicked results/URLs in the estimated preferred language. That is, if the proportion of clicked results in the estimated preferred language is high, the T-L-based approach performs well; otherwise, it performs poorly. This correlates with our analytical assessment presented in Section 3.1, which suggested that for the T-L-based approach to perform better, there must be enough top *n* relevant results in the preferred language. Users clicking more or all results from the non-preferred language suggests that the preferred language's result list did not contain enough top *n* results.

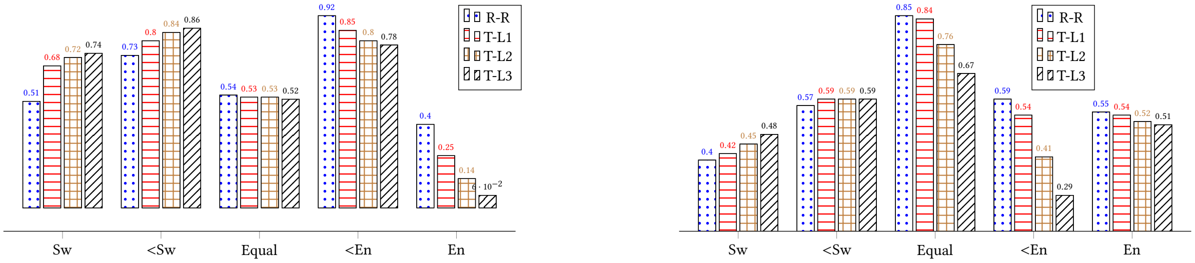
Thus, the T-L-based approach's poor performance for queries that do not conform to the estimated T-L associations is understandable, as the T-L-based approach is premised on the notion of T-L association. That is, it works best for T-L association-sensitive topics and, more specifically, queries with such associations. This observation indicates that calculating T-L-based performance without taking into account the fact that individual queries have varying click behaviour is a bad idea. That is one of the reasons why the T-L-based approach improved slightly for some cases at the topic level analysis in Section 5.1.1 above. The T-L-based approach's performance should be calculated for individual queries based on their click behaviour. This implies that the performance improvement of the T-L-based approach is primarily determined by the strength of the T-L association for a query.

The strength of the T-L association is determined by whether a query has: i) all the clicked results in the estimated preferred language – *very strong T-L association*; ii) more clicked results in the estimated preferred language than the non-preferred language



(a) *MAP@5 for English preferred topics*

(b) *NDCG@5 for English preferred topics*



(c) *MAP@5 for Swahili preferred topics*

(d) *NDCG@5 for Swahili preferred topics*

Figure 1: The MAP@5 and NDCG@5 scores for queries in English and Swahili preferred topics grouped based on how their results were clicked.

– *strong T-L association*; iii) equal number of clicked results in the estimated preferred language and the non-preferred language – *neutral T-L association*; iv) more clicked results in a non-preferred language than the estimated preferred language – *weak negative T-L association*; v) and all the clicked results in a non-preferred language – *negative T-L association.*

To demonstrate that the stronger the T-L association, the stronger the T-L-based approach to improving MLIR performance, we separated queries with actual T-L associations (very strong, and strong) from queries without T-L associations (neutral, weak, and negative). Table 2 shows that, while the T-L-based approach vastly improves relevance of results for queries with actual T-L associations, the same approach vastly degrades relevance of results for queries without T-L associations.

5.3 Minimum Threshold of Promoted Results for Optimal T-L-based Approach Performance

The third research objectives seeks to "*determine whether there is a minimum threshold of promoted results that can provide optimal performance of the T-L-based approach in T-L association-sensitive topics.*" The T-L-based approach improves relevance of results for queries that conform to the estimated T-L association. However, the exact number of results to be promoted is not yet known. To estimate this figure, we take the average precision (AP) for queries with actual T-L association (i.e., those with very strong, and strong association). There were 74 and 31 queries from English and Swahili preferred topics, respectively. The results in Figure 2 show that the threshold for achieving optimal performance of the T-L-based approach varies depending on the preferred language as well as the evaluation measure used. However, the results suggest that promoting at least three to four English results for English preferred topics, and two to five Swahili results for Swahili preferred topics could ensure the best T-L-based performance.

5.4 Limitations of the Study

Despite the fact that we used data from actual users interacting with the MLIR system, our evaluation was entirely based on IR evaluation metrics such as MAP and NDCG. Due to time and financial constraints, we were only able to conduct a system evaluation. It would have been interesting to see how actual users evaluate the effectiveness of our proposed T-L-based approach in terms of relevance improvement. Furthermore, as Carterette and Jones [4] pointed out, our simplified assumption that a click implies relevance may be skewed because the relationship between a click and relevance is always complex and multifaceted. In order for the T-L-based approach to improve results relevance, the T-L associations must be known, for example using statistical approaches (at the topic level) from historical click-through logs [26] – the *implicit approach.* It might also be interesting to ask MLIR users, explicitly, via the search engine interface, what language preferences they have for a query – *explicit approach.* A user expressing language preferences for a query may indicate that there is a strong T-L association for such a query while also reducing bias caused by click-based estimation of relevance.

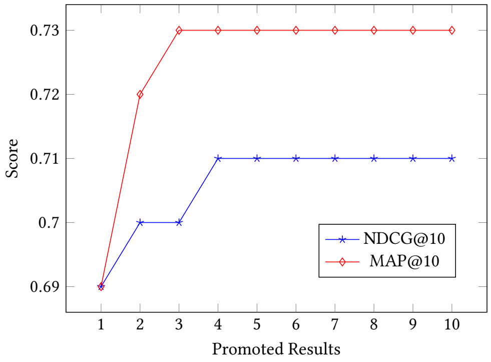
6 CONCLUSION

The paper addressed a significant research problem in multilingual information retrieval (MLIR) systems, which occurs in some contexts where users' language preferences are driven by the topic of search. Thus, we proposed and evaluated an approach for MLIR

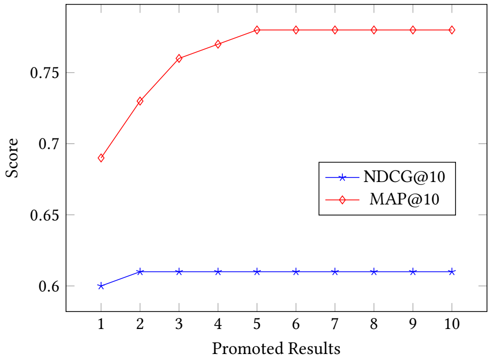
Table 2: The MAP@10 scores for queries with T-L associations vs. queries without T-L associations for English and Swahili preferred topics.

	R-R$_{Av}$	T-L1$_{Av}$	T-L2$_{Av}$	T-L3$_{Av}$	T-L4$_{Av}$	T-L5$_{Av}$
			English Preferred Topics			
with	0.60	0.69 (+14.8%)	0.72 (+19.2%)	**0.73** (+20.7%)	**0.73** (+21.0%)	**0.73** (+20.9%)
without	**0.62**	0.56 (-9.3%)	0.51 (-16.9%)	0.48 (-22.8%)	0.45 (-27.2%)	0.44 (-28.3%)
			Swahili Preferred Topics			
with	0.59	0.70 (+17.8%)	0.73 (+23.1%)	0.76 (+27.4%)	0.77 (+30.5%)	**0.78** (+31.3%)
without	**0.61**	0.53 (-11.9%)	0.49 (-19.1%)	0.46 (-24.1%)	0.44 (-28.2%)	0.42 (-31.0%)

The *with* and *without* stand for queries with actual T-L association and without T-L association, respectively.



(a) ***English Preferred Topics***



(b) ***Swahili Preferred Topics***

Figure 2: The minimum number of promoted results for optimal T-L-based algorithm performance.

results merging called T-L-based algorithm. We show that taking the correct language preferences into account during the results merging process could improve retrieval performance. Given the language preferences of the search topic, our proposed method is a simple yet effective way of merging results from the MLIR system.

The experimental results show that: the T-L-based approach improves multilingual Swahili IR performance by outperforming the baseline in almost all cases for English and Swahili preferred topics; the strength of the T-L association dramatically affects the effectiveness of the T-L-based approach; and the threshold for achieving optimal performance of the T-L-based approach varies depending on the preferred language and the evaluation measure used.

We propose to investigate several opportunities, including the use of learning-to-rank in exploring and estimating the T-L association and using them to improve ranking, as part of future work. This is crucial because our findings indicate that it is critical to correctly assess and disclose the preferred language given a query. We also propose involving users in the evaluation rather than relying solely on IR system evaluation metrics.

REFERENCES

[1] 123Tanzania. 2022. 123 Tanzania Business Directory. http://www.123tanzania.com/. [Online; accessed 25-April-2022].

[2] Amazon. 2020. Alexa - Top sites by Category. https://www.alexa.com/topsites/category/Regional/Africa/Tanzania. [Online; accessed 20-June-2020].

[3] Martin Braschler and Peter Schäuble. 2000. Using corpus-based approaches in a system for multilingual information retrieval. *Information Retrieval* 3, 3 (Oct. 2000), 273–284. https://doi.org/10.1023/A:1026525127581

[4] Ben Carterette and Rosie Jones. 2007. Evaluating Search Engines by Modeling the Relationship Between Relevance and Clicks. In *Advances in Neural Information Processing Systems 20*, J. Platt, D. Koller, Y. Singer, and S. Roweis (Eds.), Vol. 20. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2007/file/872488f88d1b2db54d55bc8bba2fad1b-Paper.pdf

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805 https://arxiv.org/abs/1810.04805v2

[6] Wei Gao, Cheng Niu, Ming Zhou, and Kam-Fai Wong. 2009. Joint ranking for multilingual web search. In *Advances in Information Retrieval (ECIR 2009. Lecture Notes in Computer Science, Vol. 5478)*, Catherine Berrut, Mohand Boughanem, Josiane Mothe, and Chantal Soule-Dupuy (Eds.). Springer, Berlin, Heidelberg, 114–125. https://doi.org/10.1007/978-3-642-00958-7_13

[7] Google. 2022. Google Trends. https://trends.google.com/trends/?geo=TZ. [Online; accessed 23-April-2022].

[8] Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated Gain-Based Evaluation of IR Techniques. *ACM Transactions on Information Systems (TOIS)* 20, 4 (Oct. 2002), 422–446. https://doi.org/10.1145/582415.582418

[9] The WWW Virtual Library. 2020. The WWW Virtual Library. http://vlib.org/. [Online; accessed 25-April-2022].

[10] Wen-Cheng Lin and Hsin-Hsi Chen. 2004. Merging Multilingual Information Retrieval Results Based on Prediction of Retrieval Effectiveness. In *Proceedings of the Fourth NTCIR Workshop on Research in Information Access Technologies Information Retrieval, Question Answering and Summarization (NTCIR-4, National Center of Sciences)*. National Institute of Informatics (NII), Tokyo, Japan, 1–7. http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings4/CLIR/NTCIR4-CLIR-LinWC.pdf

[11] Wen-Cheng Lin and Hsin-Hsi Chen. 2003. Merging Mechanisms in Multilingual Information Retrieval. In *Advances in Cross-Language Information Retrieval*, Carol Peters, Martin Braschler, Julio Gonzalo, and Michael Kluck (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 175–186.

[12] Chenjun Ling, Ben Steichen, and Silvia Figueira. 2020. Multilingual News – An Investigation of Consumption, Querying, and Search Result Selection Behaviors. *International Journal of Human–Computer Interaction* 36, 6 (2020), 516–535. https://doi.org/10.1080/10447318.2019.1662636

[13] Robert Litschko, Ivan Vulić, and Goran Glavaš. 2022. Parameter-Efficient Neural Reranking for Cross-Lingual and Multilingual Retrieval. https://doi.org/10.48550/ARXIV.2204.02292

[14] Tie-Yan Liu. 2011. *Learning to rank for information retrieval.* Springer Science & Business Media, Berlin, Heidelberg, Germany.

[15] Fernando Martínez-Santiago, Alfonso Urena-López, and Maite Martín-Valdivia. 2006. A merging strategy proposal: The 2-step retrieval status value method. *Information Retrieval* 9, 1 (Jan. 2006), 71–93. https://doi.org/10.1007/s10791-005-5722-4

[16] Deo Ngonyani. 1995. Language shift and national identity in Tanzania. *Ufahamu: A Journal of African Studies* 23, 2 (1995), 69–92. https://escholarship.org/content/qt8072719q/qt8072719q.pdf

[17] Jian-Yun Nie and Fuman Jin. 2002. A multilingual approach to multilingual information retrieval. In *A multilingual approach to multilingual information retrieval (CLEF 2002. Lecture Notes in Computer Science, Vol. 2785)*, C. Peters, M. Braschler, J. Gonzalo, and M. Kluck (Eds.). Springer, Heidelberg, Berlin, 101–110. https://doi.org/10.1007/978-3-540-45237-9_8

[18] Shuzi Niu, Jiafeng Guo, Yanyan Lan, and Xueqi Cheng. 2012. Top-k Learning to Rank: Labeling, Ranking and Evaluation. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval. Portland, Oregon, USA* (Portland, Oregon, USA) *(SIGIR '12)*. Association for Computing Machinery, New York, NY, USA, 751–760. https://doi.org/10.1145/2348283.2348384

[19] Georgios Paltoglou, Michail Salampasis, and Maria Satratzemi. 2007. Hybrid Results Merging. In *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management* (Lisbon, Portugal) *(CIKM '07)*. Association for Computing Machinery, New York, NY, USA, 321–330. https://doi.org/10.1145/1321440.1321487

[20] Allison L. Powell, James C. French, Jamie Callan, Margaret Connell, and Charles L. Viles. 2000. The Impact of Database Selection on Distributed Searching. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Athens, Greece) *(SIGIR '00)*. Association for Computing Machinery, New York, NY, USA, 232–239. https://doi.org/10.1145/345508.345584

[21] Razieh Rahimi, Azadeh Shakery, and Irwin King. 2015. Multilingual information retrieval in the language modeling framework. *Information Retrieval Journal* 18, 3 (May 2015), 246–281. https://doi.org/10.1007/s10791-015-9255-1

[22] Yves Rasolofo, Faïza Abbaci, and Jacques Savoy. 2001. Approaches to Collection Selection and Results Merging for Distributed Information Retrieval. In *Proceedings of the Tenth International Conference on Information and Knowledge Management* (Atlanta, Georgia, USA) *(CIKM '01)*. Association for Computing Machinery, New York, NY, USA, 191–198. https://doi.org/10.1145/502585.502618

[23] Jacques Savoy. 2002. Report on CLEF-2001 Experiments: Effective Combined Query-Translation Approach. In *Evaluation of Cross-Language Information Retrieval Systems (CLEF 2001. Lecture Notes in Computer Science)*, Carol Peters, Martin Braschler, Julio Gonzalo, and Michael Kluck (Eds.). Springer, Berlin, Heidelberg, 27–43. https://doi.org/10.1007/3-540-45691-0_3

[24] Jacques Savoy, Anne Le Calvé, and Dana Vrajitoru. 1997. Report on the TREC-5 Experiment: Data Fusion and Collection Fusion. *NIST Special Publication* ., . (Nov. 1997), 489–502. https://cs.iusb.edu/~danav/papers/trec5.pdf

[25] Ben Steichen and Ryan Lowe. 2021. How do multilingual users search? An investigation of query and result list language choices. *Journal of the Association for Information Science and Technology* 72, 6 (2021), 759–776. https://doi.org/10.1002/asi.24443

[26] Joseph P. Telemala and Hussein Suleman. 2021. Exploring Topic-Language Preferences in Multilingual Swahili Information Retrieval in Tanzania. *ACM Transactions on Asian and Low-Resource Language Information Processing* 20, 6, Article 96 (Aug. 2021), 30 pages. https://doi.org/10.1145/3458671

[27] Ming-Feng Tsai, Hsin-Hsi Chen, and Yu-Ting Wang. 2011. Learning a merge model for multilingual information retrieval. *Information Processing & Management* 47, 5 (Sept. 2011), 635–646. https://doi.org/10.1016/j.ipm.2009.12.002

[28] Ming-Feng Tsai, Tie-Yan Liu, Tao Qin, Hsin-Hsi Chen, and Wei-Ying Ma. 2007. FRank: A Ranking Method with Fidelity Loss. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Amsterdam, The Netherlands (SIGIR '07)*. Association for Computing Machinery, New York, NY, USA, 383–390. https://doi.org/10.1145/1277741.1277808

[29] Ming-Feng Tsai, Yu-Ting Wang, and Hsin-Hsi Chen. 2008. A Study of Learning a Merge Model for Multilingual Information Retrieval. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Singapore, Singapore) *(SIGIR '08)*. Association for Computing Machinery, New York, NY, USA, 195–202. https://doi.org/10.1145/1390334.1390370

[30] Nicolas Usunier, Massih-Reza Amini, and Cyril Goutte. 2011. Multiview Semi-supervised Learning for Ranking Multilingual Documents. In *Machine Learning and Knowledge Discovery in Databases (ECML PKDD 2011. Lecture Notes in Computer Science, Vol. 6913)*, Dimitrios Gunopulos, Thomas Hofmann, Donato Malerba, and Michalis Vazirgiannis (Eds.). Springer, Berlin, Heidelberg, 443–458. https://doi.org/10.1007/978-3-642-23808-6_29

[31] Ellen M. Voorhees, Narendra K. Gupta, and Ben Johnson-Laird. 1995. Learning Collection Fusion Strategies. In *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Seattle, Washington, USA) *(SIGIR '95)*. Association for Computing Machinery, New York, NY, USA, 172–179. https://doi.org/10.1145/215206.215357

[32] Yalwa. 2020. Yalwa Business Directory. https://www.yalwa.co.tz/. [Online; accessed 20-June-2020].

Appendix A FORMULATION AND ANALYSIS

A.1 Equal or More Relevant Top Results in the Preferred Language

	List A	**List B**	*S*	**Precision**	**T**	**Precision**
Best case	✓	✓	✓	1.000	✓	1.000
	✓	✓	✓	1.000	✓	1.000
			✓	1.000	✓	1.000
			✓	1.000	✓	1.000
			AP	**1.000**	**AP**	**1.000**
Worst case	X	✓	X	0.000	✓	1.000
	X	✓	✓	0.500	✓	1.000
			X	0.667	X	0.333
			✓	0.500	X	0.500
			AP	**0.500**	**AP**	**1.000**

Table 3: Performance of the T-L-based approach when the preferred language has either equal or more relevant top *n* results than the non-preferred language.

A.2 Few or No Relevant Top Results in the Preferred Language

	List A	**List B**	*S*	**Precision**	**T**	**Precision**
Best case	✓	X	✓	1.000	X	0.000
	✓	X	X	0.500	X	0.000
			✓	0.667	✓	0.333
			X	0.500	✓	0.500
			AP	**0.833**	**AP**	**0.417**
Worst case	X	X	X	0.000	X	0.000
	X	X	X	0.000	X	0.000
			X	0.000	X	0.000
			X	0.000	X	0.000
			AP	**0.000**	**AP**	**0.000**

Table 4: Performance of the T-L-based when the preferred language has a few or no relevant top *n* results than the non-preferred language.