

# Ranking Multilingual Documents using Minimal Language Dependent Resources

Santosh GSK, Kiran Kumar N, Vasudeva Varma

International Institute of Information Technology, Hyderabad, India  
{santosh.gsk,kirankumar.n}@research.iiit.ac.in, vv@iiit.ac.in

**Abstract.** This paper proposes an approach of extracting simple and effective features that enhances multilingual document ranking (MLDR). There is limited prior research on capturing the concept of multilingual document similarity in determining the ranking of documents. However, the literature available has worked heavily with language specific tools, making them hard to reimplement for other languages. Our approach extracts various multilingual and monolingual similarity features using a basic language resource (bilingual dictionary). No language-specific tools are used, hence making this approach extensible for other languages. We used the datasets provided by Forum for Information Retrieval Evaluation (FIRE) <sup>1</sup> for their 2010 Adhoc Cross-Lingual document retrieval task on Indian languages. Experiments have been performed with different ranking algorithms and their results are compared. The results obtained showcase the effectiveness of the features considered in enhancing multilingual document ranking.

## Keywords

Multilingual Document Ranking, Feature Engineering, Wikipedia, Levenshtein Edit Distance.

## 1 Introduction

Multilingual Information Retrieval (MLIR) is desirable with the increase of information in different languages. With the rapid development of globalization and digital online information in Internet, a growing demand for MLIR has emerged. MLIR involves the subtask of Cross Lingual Information Retrieval (CLIR) separately for each desired language. The clear separation of the retrieved result lists between different languages makes it necessary to have a merging step in order to produce a single result list. However, merging is intertwined with ranking step that ranks the documents of multilingual result lists as per the relevancy to the information need.

The problem of CLIR has been well studied in the past decade especially with the help of CLEF, NTCIR, TREC and FIRE forums. In the realm of CLIR the

---

<sup>1</sup> <http://www.isical.ac.in/cia/>

problem of ranking multilingual result lists is a very challenging task. The task of identifying whether two different language documents talks about the same topic is itself very challenging. There are few early attempts on ranking multilingual documents (Round robin merging [1], raw-score merging [1]). These merging processes have to make some simplifying assumptions. For example, one may assume that the similarities calculated for different language result lists are comparable; so the result lists can be merged according to their raw similarity values [1]. One can also normalize the similarities first; but this approach implicitly assumes that the highly ranked documents in different languages are similar to the query at a comparable level. These assumptions are not true. Until recent past [2], [3], [4], [5], [6], there was little focus on merging multilingual result lists. The recent work concentrated more on extracting semantic information such as multilingual topics from documents. These methods are highly dependent upon language specific tools like named-entity recognizer, part-of-speech tagger etc., hence they cannot be extended for languages with fewer resources, i.e., they do not achieve high-multilinguality.

If there is a requirement for a ranking approach to be applied across various languages, language specific development pose major challenges. Also while merging multilingual result lists, techniques that suit one language pair might not be effective for another. For example, techniques for closely-related languages (Ex: Hindi, Marathi) might not be useful for a pair of languages in widely different families (Ex: Spanish, Chinese). While some applications will only be concerned with a small number of languages, others (e.g., foreign policy or international patent law) will require systems that scale to tens of disparate languages. Progress in developing language-independent approaches will greatly benefit multilingual retrieval, and should therefore be encouraged within the MLIR community.

In this paper, we extract simple and efficient features from multilingual documents and topics that enhance the performance of Multilingual Document Ranking (MLDR). We propose to exploit the similarities among candidate documents to enhance MLDR. Because similar documents usually share similar ranks, cross-lingual relevant documents can be exploited to enhance the relevance estimation for documents of different languages. Given result lists of two different languages along with their queries, various similarity measures are calculated among documents of same language and of different languages. Same set of similarity metrics are measured among monolingual documents and multilingual documents, so that a document can be compared with any other document independent of their languages.

It is known that while a given translation tool may produce acceptable translations for a given set of queries; it may perform poorly for other queries [7]. Also the availability of language specific tools is very limited across languages. In this vein, we have eliminated the usage of any language-specific tools while measur-

ing document similarity metrics and other features. However, for calculating the multilingual document similarity, only bilingual dictionaries are used. External knowledge resource like wikipedia is also exploited in adding up to the efficiency of similarity measurement. Provided the availability of the basic language resource (bilingual dictionary), this approach can be extended to other language pairs. We carried out our experiments on FIRE 2010 corpus. Experiments are conducted by modelling several ranking algorithms on the extracted features. Their results are compared using the NDCG as the evaluation metric. The results obtained are verified against BM25 baseline ranking system and significant improvements are noticed in the ranking performance.

The rest of this paper is organized as follows. Section 2 reviews prior research on MLDR. Section 3 describes the features that generate similarity and relevance scores for all the query-document pairs. Section 4 describes the experimental results obtained using various ranking algorithms. Section 5 summarizes our findings and concludes with potential future work.

## 2 Related Work

The early attempts for merging multilingual documents were heuristic based approaches. Raw-score merging [1] tries to combine the result lists using the document scores that were previously assigned by the retrieval algorithms. As the scores were incomparable, efforts were made to normalize the scores [3] [4]. The documents with the normalized scores are then ranked. Round robin merging [1] tries to combine the result lists based on the ranks of the documents.

A 2-step merging strategy was proposed by Martinez-Santiago et al. [2] to rank multilingual result lists. In this approach, instead of directly merging the result lists, the multilingual result lists are first indexed and this indexed dataset is used to retrieve final multilingual result list. Recent work [6] [5] focused on implementing the learning approaches to the merge problem. They have extracted various features and trained the features using learning algorithms like FRank, Boltzmann. However, in the work presented by Tsai et al. [5], the constructed features identify person names, organization names and vocabulary terms. They are very much dependent upon the availability of language specific tools.

Although Gao et al. [6] laid emphasis on the importance of measuring multilingual document similarity, they have incorporated these similarities in implementing a clustering based approach to the problem of MLDR. However our approach focuses on measuring the direct influence of these similarities by incorporating them as features for a document.

The basic idea behind it is that a document can be similar to other documents in the result list of same language or in the result list of different language. Capturing the similarity between the documents gives us very useful information

regarding the importance of the documents. If a document is found to be similar to many other documents and is also relevant enough to the query, then that document needs to be placed at a higher rank in the final result list. Hence, these similarity features are studied here in the context of enhancing MLDR performance.

### 3 System Overview

In a usual scenario of MLIR, given a query in a language, CLIR is performed on separate monolingual collections. Once monolingual result lists are obtained from each collection, all the lists are merged into a multilingual result list. Our work scenario considers a query in English and Hindi, along with their corresponding monolingual result lists as the starting point. In this context, the merging process does not have any prior information on how the original result lists are produced. However there are binary relevance judgements for all query-document pairs indicating whether a document is relevant or not. Features that effectively capture the document relevancy to the query and the similarity among the documents are extracted. Same set of features are considered for all the documents. Every document is represented in terms of a vector of these features. After the vectors are constructed for all the documents, these features are modelled using various ranking algorithms. The estimated relevance probabilities assigned to the documents by the ranking algorithms are used in ordering the documents.

#### 3.1 Feature Engineering

In information retrieval and natural language processing (NLP), question answering (QA) is the task of automatically answering a question posed in natural language. To find the answer to a question, a QA computer program may use either a pre-structured database or a collection of natural language documents. Given a set of documents and a question, a QA system needs to address two interesting challenges [8].

1. *Estimating answer relevance.* Is a particular answer relevant to the question? How to identify relevant answer(s) among irrelevant ones? If irrelevant answers can be eliminated by using some knowledge base, then the remaining answers can be ranked better.
2. *Exploiting answer redundancy.* How do we exploit answer redundancy among answer candidates? If a particular relevant answer is found to repeat in a given list of answers, then that answer needs to be ranked higher.

Analyzing from this QA perspective, the problem of MLDR is indeed related to it. There is a need to identify relevant documents and also they need to be ranked higher when they are found similar to many other documents. Hence, we have emphasized on features that address these two challenges. Features are required to capture the relevancy of documents for a given query and the similarity among documents.

Such features are extracted from three levels 1) Query-Document similarity, 2) Monolingual Document similarity and 3) Multilingual document similarity. According to the extraction level, we describe these features in detail as follows.

**3.1.1) Query-Document Similarity:** There are many ways to calculate the relevancy among a document and a query. We used the tf and idf measures for measuring the relevancy of documents for a query. For every query term 'k' in a query Q, its tf-idf value is calculated with respect to a document 'j' ( $tf-idf_{kj}$ ) in the document collection  $D$  of the same language. For all the query terms, these scores are calculated and added up to get the tf-idf feature value for document 'j'. We normalized this tf-idf value within the document collection according to the following formula

$$TF - IDF = \frac{\sum_{\forall k \in Q} tf - idf_{kj}}{\sum_{\forall j \in D} \sum_{\forall k \in Q} tf - idf_{kj}} . \quad (1)$$

Each query topic has title, description and narration fields. The tf-idf measures are calculated for each of these fields and added as different features.

**3.1.2) MonoLingual Document Similarity:** Given two documents of the same language, the similarity can be measured in various ways as mentioned in the work of A. Huang [9]. Every document is represented in a vector notation. The terms of a document are assigned their tf-idf weights calculated within that document collection  $D$ . The wikipedia redirection terms corresponding to every term are also included in the vector. The concept of wikipedia redirections is explained below. The similarity measure for a document  $d_i$  is calculated using the formula

$$sim_k(d_i) = \frac{\sum_{j=1(i \neq j)}^D sim'_k(d_i, d_j)}{\sum_{\forall i} \sum_{j=1(i \neq j)}^D sim'_k(d_i, d_j)} . \quad (2)$$

Each  $sim'_k(d_i, d_j)$  is a similarity feature used to calculate document similarity between  $d_i$  and  $d_j$ . Euclidian, Cosine, Jaccard, Pearson Correlation coefficient and Averaged Kullback-Leibler Divergence [9] are the different similarity features used in calculating the document similarity.

**Concept of Wikipedia Redirection:** Wikipedia is a free, web-based, collaborative, multilingual encyclopaedia. There are 262 language editions available as of now. So, the extensibility of our techniques across certain available languages is ensured even after using wikipedia knowledge base. Since Wikipedia is web-based and therefore worldwide, contributors of a same language edition may use different dialects or may come from different countries. These differences may lead to some conflicts over spelling differences, (e.g. color vs. colour) or points of view (e.g. sachin tendulkar, master blaster, little master, SRT).

All these representations conceptually refer to the same wikipedia page. When a user enters any one of these representations, the action gets redirected to that single wikipedia page which all these terms conceptually refer to. We collect all such redirections available in English and in Hindi document collections of wikipedia. Sometimes the redirections may contain phrases, in such cases the phrases are tokenized and are added to the list of redirections fetched for a given term. It was coined in [10], that these wikipedia redirections can be referred to as synonyms. These redirections are used in both monolingual and multilingual document similarity calculations.

**3.1.3) MultiLingual Document Similarity:** In order to compare two multilingual documents, we need to map them onto a common ground representation. Given a document in English and a document in Hindi, the English document terms are mapped into Hindi representation by using bilingual dictionary and wikipedia redirections. If a term is found in dictionary, it is replaced with its synonyms in Hindi. Each word may have more than one possible synonyms. For every term in the synonyms, its wikipedia redirections are also taken into account. Finally, the terms of a English document are represented in a vector of Hindi terms. Same set of similarity features used in measuring monolingual document similarities are used here. Similarities are calculated as per the Eq. (2).

Transliteration might be highly helpful in identifying the proper nouns, but it requires parallel transliterated (English-Hindi) word lists to build even a language-independent statistical transliteration technique [11]. Acquiring such word lists is a hard task when one of the language is a minority language. Including transliteration comes at the cost of reducing the extensibility of our approach. Priority is given to the latter.

**Modified Levenshtein Edit Distance Measure:** In all of the similarities calculated above, the terms are compared using the Modified Levenshtein edit distance as a string distance measure. In many languages, words appear in several inflected forms. For example, in English, the verb 'to walk' may appear as 'walk', 'walked', 'walks', 'walking'. The base form, 'walk', that one might look up in a dictionary, is called the lemma for the word. The terms are usually lemmatized to match the base form of that term. Lemmatizers are available for English and many other European languages. But the lemmatizers support is very limited in the context of Indian Languages. So, we have modified the levenshtein edit distance metric to replace the purpose of lemmatizers by adding certain language-independent rules. Henceforth, it can be applied for any language. This modified levenshtein edit distance would help us in matching a word in its inflected form with its base form or other inflected forms. The rules are very intuitive and are based on three aspects:

1. Minimum length of the two words
2. Actual Levenshtein distance between the words
3. Length of subset string match, starting from first letter.

## 4 Experiments and Evaluation

We have conducted experiments using the FIRE 2010 dataset available for the ad-hoc cross lingual document retrieval task. The data consists of news articles collected from 2004 to 2007 for each of the English, Hindi, Bengali and Marathi languages from regional news sources. There are 50 query topics represented in each of these languages. We have considered the English and Hindi articles for our experiments. There are only binary relevance judgements given for every topic-document pair. For every topic we have worked with a set of documents which contains all its relevant documents with a certain noise i.e., irrelevant documents. The noise considered is twice the number of relevant documents.

Different ranking algorithms like SVC (SVM Classification), RSVM (Ranking SVM), SVM Regression and Logistic Regression are used to learn ranking functions by modelling the features extracted. The source codes of LibSVM<sup>2</sup>, SVM-Light<sup>3</sup>, SVM-Rank<sup>4</sup> and Logistic Regression<sup>5</sup> are used to run SVC, SVM regression, RSVM, and Logistic regression respectively. The probabilities predicted by these learning approaches are used in ranking the documents. In order to evaluate the ranking order of the documents, a small set of documents are annotated with ratings from 0(irrelevant) to 5(perfect) by human labellers. The results of NDCG@5,10,15,20 (Normalized Discounted Cumulative Gain) are used to compare the systems. MLIR ranking performance results of these learning approaches are compared in Table 1 with a BM25 baseline system.

From the table 1 it is clear that all the learning algorithms has outperformed the BM25 baseline system. The numbers indicate that the ranking functions that modelled our features stand at par with the baseline system in terms of the performance. There is an overall enhancement in the performance of MLDR. Ranking SVM and SVM regression have achieved the best results. These accuracies provide a good indication that the features considered has proved to be effective in enhancing the MLDR performance.

**Table 1.** Comparison of MLDR Performances

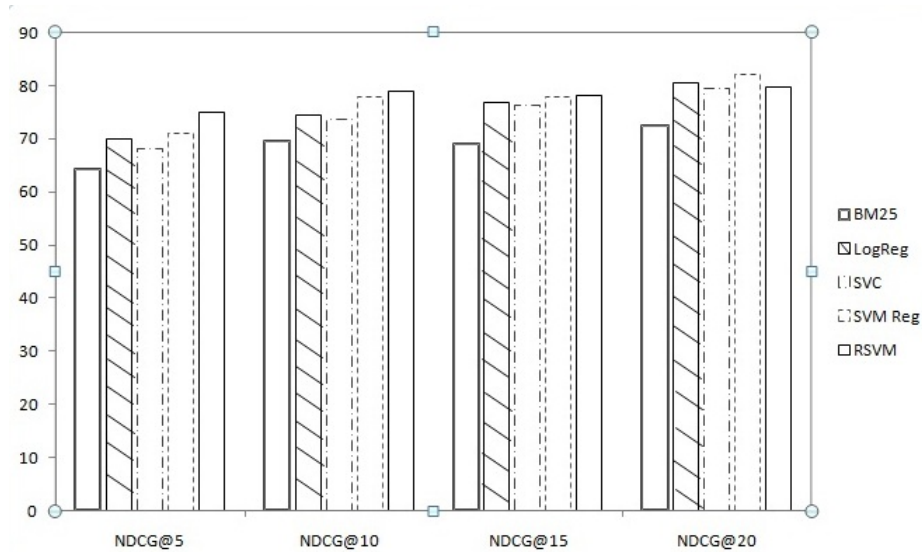
Method	NDCG@5	NDCG@10	NDCG@15	NDCG@20
<b>SVC</b>	67.99	73.53	76.28	79.50
<b>SVM-Reg</b>	71.00	77.87	77.94	<b>82.11</b>
<b>RSVM</b>	<b>75.04</b>	<b>78.84</b>	<b>78.03</b>	79.83
<b>LogReg</b>	69.89	74.53	76.69	80.37
<b>BM25</b>	64.19	69.38	68.82	72.21

<sup>2</sup> <http://www.csie.ntu.edu.tw/~cjlin/libsvm/>

<sup>3</sup> <http://svmlight.joachims.org/>

<sup>4</sup> [http://www.cs.cornell.edu/People/tj/svm.light/svm\\_rank.html](http://www.cs.cornell.edu/People/tj/svm.light/svm_rank.html)

<sup>5</sup> <http://komarix.org/ac/lr/lrtrirls>



**Fig. 1.** A Graphical Comparison of the Performances of Ranking Algorithms

## 5 Conclusion and Future Work

In this paper we have presented an approach to extract simple, effective features that enhances the MLDR performance. We have approached the problem of MLDR from a QA perspective. The features we considered extract the document relevancy to a query and various document similarity measures. These features are modelled using different ranking algorithms. From the results showcased in Table 1, it can be inferred that these features has considerably increased the MLDR performance. They dominated the BM25 baseline with a good margin.

As development of language-specific tools pose major challenges across languages, we have come up with this approach without using any language tools. Compared to the existing approaches, this approach achieves high multilinguality; it is extensible to other languages and is easily reproducible provided the availability of minimum language resource (bilingual dictionary). The cost of reproducibility of this approach is the same for any other language pairs. This approach is not specific to any dataset, it can be applied to various other datasets. This approach takes into consideration the future growth of multilingual information need.

As there are dictionaries available for few other Indian Languages <sup>6</sup>, we are currently working on extending our approach for other Indian languages using

<sup>6</sup> [http://ltrc.iiit.ac.in/onlineServices/Dictionaries/Dict\\_Frame.html](http://ltrc.iiit.ac.in/onlineServices/Dictionaries/Dict_Frame.html)

the FIRE 2010 datasets. As there are many other learning approaches available, we would like to explore their role in enhancing the performance of MLDR. We are planning to extend this approach to most-researched european languages and compare the MLDR accuracy of our approach with the existing state-of-art systems. We would like to work more on capturing query document relevancy by extracting key terms from the documents to get the topics of the documents and thereby assign more relevance to the documents that talks about the (almost)same topic as that of the query.

## References

1. Savoy, J., Calve, A.L., Vrajitoru, D.: Report on the TREC-5 experiment: Data fusion and Collection fusion. In: The Fifth Text Retrieval Conference (TREC-5). (1997) pages 489–502.
2. Martinez-Santiago, F., Urena-Lopez, L., Martin-Valdiva, M.: A merging strategy proposal: The 2-step retrieval status value method. *Information Retrieval* (2006) pages 71–93.
3. Powell, A., French, J., Callan, J., Connell, M., Viles, C.: 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, ACM (2000) pages 232–239.
4. Lin, W., Chen, H.: Merging Mechanisms in Multilingual Information Retrieval. In Peters, C., Braschler, M., Gonzalo, J., Kluck, M., eds.: Third Workshop of the CLEF. Volume 2785 of LNCS., Springer (2002) pages 175–186.
5. Tsai, M., Wang, Y., Chen, H.: A Study of Learning a Merge Model for Multilingual Information Retrieval. In: Proceedings of SIGIR 2008. SIGIR '08, ACM (2008) pages 195–202.
6. Gao, W., Niu, C., Zhou, M., Wong, K.: Joint Ranking for Multilingual Web Search. In Boughanem, M., Berrut, C., Mothe, J., Soulé-Dupuy, C., eds.: Proceedings of ECIR 2009. Volume 5478 of LNCS., Springer (2009) pages 114–125.
7. Savoy, J., Berger, P.: Selection and Merging Strategies for Multilingual Information Retrieval. In Peters, C., Clough, P., Gonzalo, J., Jones, G.J.F., Kluck, M., Magnini, B., eds.: Fifth Workshop of CLEF (Cross Language Evaluation Forum). Volume 3491 of LNCS., Springer (2004) pages 27–37.
8. Wo, J., Si, L., Nyberg, E., Mitamura, T.: Probabilistic Models for Answer-Ranking in Multilingual Question-Answering. *ACM Transactions on Information Systems* (2010)
9. Huang, A.: Similarity measures for Text Document Clustering. In: Proceedings of New Zealand Computer Science Research Student Conference. (2008) pages 49–56.
10. Wu, F., Weld, D.: Autonomously semantifying Wikipedia. In: Proceedings of sixteenth CIKM. CIKM 07, ACM (2007)
11. Ganesh, S., Harsha, S., Pingali, P., Varma, V.: Statistical Transliteration for Cross Language Information Retrieval using HMM alignment model and CRF. In: 2nd International Workshop on CLIA, 3rd International Joint Conference on Natural Language Processing (IJCNLP 2008). (2008)