Volume 2, Issue 8, August 2012 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

A Review on Performance Evaluation Measures of Multi Lingual Information Retrieval Systems

Pothula Sujatha

Assistant Professor
Department of Computer science
School of Engineering & Technology
Pondicherry University
Pondicherry, India.

Abstract— **Multi-Lingual Information Retrieval (MLIR) can be defined as the functionality which allows one to find information that is expressed in any language. Selecting useful documents from collections that may contain several languages is a challenging issue. Multilingual retrieval is the general expression used to refer to the development, research and applications concerning the above-mentioned functionality. These systems requiring careful and thorough evaluation to demonstrate the superior performance of novel techniques on representative document collections. This paper mainly concentrates on the quantitative analysis of these systems because the evaluation of information retrieval systems has gained considerable momentum in the last few years. A number of evaluation initiatives led to a considerable increase in system performance and also identifies the major strengths and shortcomings of evaluation measures. This paper begin with a discussion of measuring the effectiveness of MLIR systems, this includes explaining the kinds of evaluation measures that are regularly used for document retrieval.**

Keywords—**Information Retrieval, Retrieval efficiency, translation ambiguity, aboutness**

I. Introduction

Information Retrieval (IR) is the area of computer science dedicated to the theory and practice of searching information. Since text is the most common medium utilized to represent and distribute information efficiently, most IR research has been focused on searches in collections of textual documents. Information searches on the Web have become a commodity thanks to the availability of efficient search technology. An ideal IR system should only return relevant documents, but “relevance” is very hard to formalize [Saracevic, 1975]. Usually relevance is defined as a function of aboutness: a document is relevant with respect to a certain information need if it is about the topics addressed in the query. A controlled indexing scheme does not allow for flexibility at retrieval time [Sparck Jones, 1999]. A document is exhaustively indexed if all concepts which are discussed are represented in the index. If a concept which is discussed in a document is not indexed, the document will not be found with a corresponding query. However, high exhaustiveness is not always desirable, since if side-issues in documents are indexed in addition to main concepts, this will deteriorate the quality of the retrieved document set. Information available in the web is not deals with many languages. User can form the query in his own Language and retrieval of the documents in various languages. This is possible with MLIR systems. This paper analyses whether the available IR metrics are enough to evaluate the MLIR systems or not. Many more are needed like how to devise new measures in order to co-up with the new MLIR systems.

The main contributions of this work are finding the relevant factors should be possessed by each MLIR system. Nearly we found forty relevant characteristics/properties. Each of these properties is described in detail in the following sections. Then we extend our work to discuss all the MLIR metrics. The organization of the paper is as follows: In section 2 we give brief descriptions of performance evaluation of MLIR systems. Section 3 the characteristics or properties of MLIR systems have been studied with definitions. Section 4 concludes the paper.

II. Performance Evaluation of MILR Systems

It is not obvious how one could measure the vague notion of quality. There are again two main approaches: system oriented and user oriented evaluation. The system-oriented evaluation has the advantage that experimental conditions can be highly controlled, using batch-mode experiments. There are however limitations to such an evaluation. A real-life information retrieval task comprises the full process of query formulation, query re-formulation and document selection. Current MLIR systems are equipped with graphical user-interfaces and offer many options for refining the query or restricting the result list. In order to measure the effectiveness of these interfaces, user oriented evaluations are required.

The effectiveness of information retrieval systems is measured by comparing performance on a common set of queries and documents. Significance tests are often used to evaluate the reliability of such comparisons. Effectiveness is compute by measuring the ability of systems to find relevant documents. The measured score is most often used as an indicator of the performance of one system relative to another; with an assumption that similar relative performance will be observed on other test collections and in operational settings [Sanderson and Zobel 2005]. Evaluation has always been a critical component of MLIR. MLIR systems have been measuring effectiveness: determining the relevance of items, retrieved by a search engine, relative to a user's query. The final component of evaluation using a test collection is the evaluation measure. The measure provides a simple simulation of a user's behaviour. A great many such measures have been discussed and described in this paper.

III. TRADITIONAL MEASURES FOR IR SYSTEM

For several other tasks, the traditional measures have been considered to be inadequate. For the Web tasks, for example, Web-user-oriented measures were sought. For the navigational tasks, the mean reciprocal rank of the target item was established. For informational tasks, early precision measures were used. Often, the precision at ten documents is used. The recall power of a system can be neglected when taking into consideration an underlying user model with the average Web user who is seeking only a few hits. Many novel retrieval measures have been developed in the past years. Nevertheless, the classic measures are still being widely used. Overall, there is a consensus that these new measures might reveal something important that is not covered by recall and precision. However, it is not yet well understood what this "something" is (Robertson 2006). Retrieval system evaluation plays an important role in judging the efficiency and effectiveness of the retrieval process.

Following selection of relevance as criterion, precision and recall became the preferred pair of measures of IR evaluation studies on the processing level. Given a relevance assessed output, precision can be directly and easily derived. Recall cannot. This is because recall depends not only on what was retrieved, but also on what was not retrieved, i.e., on what was missed. In this sense recall is a metaphysical measure: how does one know what is missed when one does not know that it is missed? Thus, establishing a set to calculate recall presents methodological challenges. Many methodological "tricks" were devised to be able to estimate recall. In toy databases all documents were assessed as to relevance for given requests beforehand. In large databases, as TREC, this is not possible. The method of polling of all the outputs for the same request is used to establish relative recall. This may be fine for comparisons.

Recall. Recall is the fraction of the documents that are relevant to the query that are successfully retrieved [Christopher 2009]. For example, a 60 percent recall rate means that 60 percent of all relevant documents in a collection have been found, and 40 percent have been missed. The recall requires knowledge of all the relevant documents in a collection that could never be put together in any real world collection. The number of known relevant documents is usually used to calculate the value. It indicates the ability of system to find relevant documents.

Precision measured at no ranking: Precision is the fraction of the documents retrieved that are relevant to the user's information need [Christopher 2009]. For example, a 65 percent precision rate means that 65 percent of the documents retrieved are relevant, while 35 percent of those documents have been misidentified as relevant. Both measures are set oriented, However, most current systems present ranked results. Precision shows how good a system is in finding only relevant documents

Precision measured at a fixed ranking. A common option for measuring is to decide that a user will choose only to examine a certain number of ranked results and calculate precision at that fixed rank position.

$$p(n) = \frac{r(n)}{n} \qquad (1)$$

where, r(n) is the number of relevant items retrieved in the top n documents. The choice of n is often influenced by the manner in which results are displayed. Precision measured at rank 10 is the commonest approach reflecting the current convention of most web search engines displaying 10 search results per web page.

This form is appealing as the number it calculates reflects the experience a user will typically encounter when using a search engine across a set of topics. Note P(n) ignores the rank position of relevant documents retrieved above the cut off and ignores all other relevant documents, also if a topic has fewer than n relevant documents in the collection being searched, P(n) for that topic will always be <1. However, there is little evidence these features of the measures are problematic.

A recall and precision value pair can be obtained for each position on the ranked list taking into account all documents from the top of the list down to that position. Plotting these values leads to the recall-precision graph [Thomas Mandl 2008]. To evaluate the effectiveness of a ranking model, many MLIR measures have been proposed. Here we give a brief introduction to several popular ones which are widely used in learning to rank. Precision@k [Voorhees et al. 2005] is a measure for evaluating top k positions of a ranked list using two levels (relevant and irrelevant) of relevance judgment: where k denotes the truncation position.

R-precision: It is the precision at R where R is the number of relevant documents in the collection for the query. It is the precision after R retrieved documents, where R is the number of relevant documents that exists for that query. An R-precision of 1.0 is equivalent to perfect relevance ranking and perfect recall. However, a typical value of R-precision which is far below 1.0 does not indicate the actual value of recall (since some of the relevant documents may be present in the hit-list beyond point R). This metric is highly correlated with MAP [Allan 2005].

Mean Reciprocal Rank (MRR): [Voorhees 1999] Reciprocal of the rank of the first relevant result averaged over a population of queries. It is possible to define for entities other than explicit absolute relevance judgments.

Binary preference (bref): Buckley and Voorhees proposed using a new metric, bpref that reduces the effect of relevant documents that are not judged through its focus on relative pair wise preference (document A vs. document B) judgments [Jensen 2006]. This takes any document into account which was judged by a human juror. This checks how many times a system retrieves a relevant document before a document is judged.

Fall-Out: It is the proportion of non-relevant documents that are retrieved, out of all non-relevant documents that are available. Fall-Out is a measure of how quickly precision drops as recall is increased. Fallout is defined as the probability to find an irrelevant among the retrieved documents.

F-measure: It is the weighted harmonic mean of precision and recall, the traditional F-measure or balanced F-score. The F-measure was derived by [van Rijsbergen 1979] so that Fβ "measures the effectiveness of retrieval with respect to a user who attaches β times as much importance to recall as precision". It is based on van Rijsbergen's effectiveness measure E = 1 − (1 / (α / P + (1 − α) / R)). Their relationship is Fβ = 1 − E where α = 1 / (β2 + 1). It is a combination of recall and precision.

Mean Average Precision (MAP): The average of precision values at certain levels of recall is calculated as the MAP, which expresses the quality of a system in one number. Commonly used measure is mean average precision. The first reference to this measure was in [Harman 1993], where the measure was called non-interpolated average precision. It is defined as follows

$$AP = \sum_{rm=1}^{N} \frac{(P(r_n) * rel(r_n))}{R} \qquad (2)$$

Here, N is the number of documents retrieved, rn is the rank number; rel(rn) is the relevance of the document retrieved at rn; P(rn) is the precision measured at rank rn and R is the number of relevant documents for this particular topic. Simply, the measure calculates precision at the rank position of each relevant document and takes the average. Note, by summing over the N retrieved documents and dividing by the R relevant documents, in effect, precision is measured as being zero for any unretrieved relevant document.

If one calculates AP for each of a set of topics and takes the mean of those average precision values, the resulting calculation is known as Mean Average Precision (MAP). Voorhees appears to be the first to describe the measure as mean average precision (1993). MAP has become one of the primary measures used in many evaluation exercises as well as a large quantity of published IR research.

Most standard among the TREC community is Mean Average Precision (MAP), which provides a single-figure measure of quality across recall levels. Among valuation measures, MAP has been shown to have especially good discrimination and stability. For a single information need, Average Precision is the average of the precision value obtained for the set of top k documents existing after each relevant document is retrieved, and this value is then averaged over information needs. MAP is to determine whether or not a given search strategy would be better than another. The average of precision values at certain levels of recall is calculated which expresses the quality of the system in one number.

Graded Relevance measures: Measures such as Precision and MAP can really only be used with binary judgments. Assuming that one can transform grades of relevance assessment to numerical values, [Järvelin and Kekäläinen 2000] proposed a suite of measures that evaluated the effectiveness of a retrieval system regardless of the number of levels of relevance. Their simplest measure, cumulative gain (CG), is the sum of relevance values (rel) measured in the top n retrieved documents.

$$CG(n) = rel(1) + \sum_{i=n}^{n} rel(i) \qquad (3)$$

Discounted Cumulative Gain (DCG): Where the relevance values are discounted progressively as one moves down the document ranking used a log-based discount function to simulate users valuing highly ranked relevant documents over the lower ranked.

$$DCG(n) = rel(1) + \sum_{i=2}^{n} \frac{rel(i)}{\log b(i)} \qquad (4)$$

Normalized Discounted Cumulative Gain (NDCG): It is designed for situations of non-binary notions of relevance like precision at k; it is evaluated over some number k of top search results.

In a follow up paper [Järvelin and Kekäläinen 2002] added a third measure, normalized DCG (nDCG)

$$nDCG(n) = \frac{DCG(n)}{IDCG(n)} \qquad (5)$$

nDCG was initially tested and proven to work well in test collection evaluations with the existence of a wide range of relevance judgments [Robertson and Stephan 2006] and [Järvelin and Kekäläinen 2002]. nDCG does not correlate strongly with other measures (such as precision, users' satisfaction with coverage, accuracy and ranking of the results) when working with a limited number of relevance judgments (i.e., the case when only judging the first 10 pages).

Generally, P-R curves demonstrate that the good performance of the expanded structured query types is due to, in particular, their ability to rank the highly relevant documents toward the top of retrieval results. The CG illustrates the value the user actually gets, but DCG can be used to forecast the system performance with regard to a user's patience in examining the result list. The advantage of the P-R measure is that it treats requests with different number of relevant documents equally, and from the system's point of view the precision at each recall level is comparable. In contrast, CG and DCG metrics gives the number of documents needed to achieve a certain gain based on user' view point [Järvelin and Kekäläinen2000].

In particular NDCG is designed for situations of non-binary notions of relevance like precision at k; it is evaluated over some number k of top search results.

NDCG is designed for multiple levels of relevance judgments. NDCG@k evaluates top k positions of a ranked list using multiple levels (labels) of relevance judgment

Inverse Document Frequency(IDF): Raw term frequency as above suffers from a critical problem: all terms are considered equally important when it comes to assessing relevancy on a query. In fact certain terms have little or no discriminating power in determining relevance [christopher 2009]. For instance, a collection of documents on the auto industry is likely to have the term auto in almost every document. To this end, we introduce a mechanism for attenuating the effect of terms that occur too often in the collection to be meaningful for relevance determination. An immediate idea is to scale down the term weights of terms with high collection frequency, defined to be the total number of occurrences of a term in the collection. The idea would be to reduce the tf weight of a term by a factor that grows with its collection frequency.

tf-idf weighting: In the classic vector space model proposed by Salton, Wong and Yang the term specific weights in the document vectors are products of local and global parameters [christopher 2009].

Overlap score: The score of a document d is the sum, over all query terms, of the number of times each of the query terms occurs in d. We can refine this idea so that we add up not the number of occurrences of each query term t in d, but instead the tf-idf weight of each term in d [christopher 2009].

Accuracy: An obvious alternative that may occur to the reader is to judge an information retrieval system by its accuracy, that is, the fraction of its classifications that are correct [christopher 2009].This seems plausible; since there are two actual classes, relevant and non-relevant, and an information retrieval system can be thought of as a two-class classifier which attempts to label them as such (it retrieves the subset of documents which it believes to be relevant). This is precisely the effectiveness measure often used for evaluating machine learning classification problems.

Interpolated precision: [christopher 2009]. The interpolated precision pinterp at a certain recall level r is defined as the highest precision found for any recall level $r' \geqslant r$.

Sensitivity: Sensitivity is just another term for recall. The false positive rate is given by fp/(fp+ tn) [christopher 2009].

Specificity: For unranked result sets, specificity, given by tn/(fp+tn), was not seen as a very useful notion. Because the set of true negatives is always so large, its value would be almost 1 for all information needs (and, correspondingly, the value of the false positive rate would be almost 0) [christopher 2009].

Kappa statistic: It is interesting to consider and measure how much agreement between judges there is on relevance judgments. In the social sciences, a common measure for agreement between judges is the kappa statistic. It is designed for categorical judgments and corrects a simple agreement rate for the rate of chance agreement [christopher 2009].

Probabilistic relevance feedback: Rather than reweighting the query in a vector space, if a user has told us some relevant and non-relevant documents, then we can proceed to build a classifier. One way of doing this is with a Naive Bayes probabilistic model. If R is a Boolean indicator variable expressing the relevance of a document, then we can estimate P (xt = 1|R), the probability of a term t appearing in a document, depending on whether it is relevant or not. [christopher 2009].

Context resemblance: We ensure that retrieval results respect this preference by computing a weight for each match. A simple measure of the similarity of a path cq in a query and a path cd in a document is the following context resemblance function CR [christopher 2009].

Simnomerge: The final score for a document is computed as a variant of the cosine measure [Tetsuya and Noriko 2008].

Q-measure: Reciprocal rank evaluates multiple answers but cannot handle answer correctness levels [Tetsuya and Noriko 2008]. Q-measure can handle single answer and multiple answer questions and can handle answer correctness levels. Q-measure can be applied to QA evaluation.

Normalized Cumulative Utility (NCP): "Normalized" means to use the ideal ranking to compute a theoretic upper bound of the measure and then normalize the actual gain value with this upper bound. "Cumulative" means to measure the overall utility of n documents by the sum of the gain of each relevant document [Tetsuya and Robertson 2008].

Robertson has recently pointed out that AP can be interpreted as a special case of NCP, computed as an expectation of precision over a population of users who eventually stop at different ranks in a ranked list of retrieved documents. He regards AP as a crude version of NCP, in that the probability distribution of the user’s stopping behavior is uniform across all relevant documents. We generalize NCP to introduce a family of metrics called Normalized Cumulative Utility (NCU), some of which are arguably more “realistic” than AP.

Kendall: Kendall’s rank correlation is a monotonic function of the probability that a randomly chosen pair of ranked systems is ordered identically in the two rankings [Tetsuya and Robertson 2008]. Hence a swap near the top of a ranked list and that near the bottom of the same list has equal impact. However, for the purpose of ranking retrieval systems, for example, in a competition-style workshop such as NTCIR and TREC, the ranks near the top of the list are arguably more important than those near the bottom.

YAR correlation: In light of this, the recently-proposed Yilmaz/Aslam/Robertson rank correlation is a monotonic function of the probability that a randomly chosen system and one ranked above it are ordered identically in the two rankings. Like Kendall’s rank correlation, YAR rank correlation lies between −1 and 1, but unlike Kendall’s, it is not symmetrical. Yilmaz, Aslam and Robertson also provides a symmetric version, but we use the raw asymmetric YAR rank correlation by taking AP as the gold standard. When the errors (i.e., pairwise swaps with respect to the gold standard) are uniformly distributed over the ranked list being examined, YAR rank correlation is equivalent to Kendall’s rank correlation [Tetsuya and Robertson 2008].

Blended Ratio BR(n): [Tetsuya and Robertson 2008] Blended Ratio for handling graded relevance. BR is suitable as a utility/cost function because, given the stopping point n, it does not matter where the relevant documents are within top n. A large β represents a very persistent user; β=0 reduces BR to P.

Collection weight: [Lin and Chen 2004] how many relevant documents there are in a document collection for a given query is also an important factor for predicting retrieval performance? This measure is to estimate the proportion of possible relevant documents of a query in a document collection. If the proportion of relevant documents is high, they have more chance be retrieved. Therefore, a document collection with higher portion of relevant documents is assigned higher weight. Given a query qi, the collection weight of a document collection Dm is estimated by the following equation.

$$CW_{im} = \frac{avg\_DF_{im}}{N_m} \qquad (7)$$

Where CWim is the collection weight of document collection Dm corresponding to query qi, avg_DFim is the average document frequency of query terms in query qi, and Nm Is the number of documents in collection Dm.

RSV (Retrieval Status Value): It is a matching score to a document reflects its degree of relevance with respect to a certain topic.

Geometric Mean Average Precision (GMAP): It is a new measure might reveal something important that Is not covered by recall and precision. When evaluating multilingual or cross-lingual retrieval runs, essentially the same measures can be used as in the monolingual case. However, work by [Mandl 2009] indicates that some measures may exhibit different behaviour for multilingual settings when compared to using them for monolingual runs. Specifically, Mandl showed that in the multilingual case, MAP calculated over a set of topics tend to be dominated by the performance that systems obtain on the "easiest" queries. This same behaviour was not observed in related monolingual experiments. Mandl showed that using the Geometric Mean (GMAP) mitigates this effect–topic difficulty has much less influence on overall system rankings in this case. These observations can be taken as strong indications on the benefits of using at least both measures when "robustness", i.e. solid retrieval effectiveness even for hard topics, is a concern. In any case, it may be beneficial to consider both measures (MAP and GMAP) when many languages are used for multilingual retrieval: in such cases, the correlation between the two measures is very low [Mandl et al. 2008], in contrast to usually high correlations for monolingual and bilingual experiments. Linguistic difference may have caused the performance differences. There is a consensus that these new measures might reveal something important that is not covered by recall and precision. However, it is not yet well understood what this “something” is [Robertson 2006].

P@10 was not as reliable as MAP. In agreement with previous work, therefore, we conclude that MAP is a more reliable measure than P@10. We believe the simple reason for this is that MAP takes into account the location in a ranking of all known relevant documents, whereas P@10 is influenced by at most 10.

IV. CONCLUSIONS

This paper discusses the IR and MLIR systems. It presents the evaluation methodology for both the systems. It also described the importance of developing new metrics for evaluating the MLIR systems. The popular performance retrieval measures with their strengths and shortfalls are also demonstrated. There is definitely a need to devise new metrics for evaluating MLIR systems since there are playing vital role with many linguistics. The available metrics and not more efficient to evaluate the performance of these systems, in future, we plan to analyze and define novel metrics for the upcoming MLIR systems.

REFERENCES

[1] Allan, James. 2005. HARD track overview in TREC 2005: High accuracy retrieval from documents. In *Proc. TREC*. 174, 519.

[2] Buckley, Chris; Voorhees, Ellen (2005): Retrieval System Evaluation. In: *TREC: Experiment and Evaluation in Information Retrieval*. Cambridge & London: MIT Press. pp. 53-75. C. J. van Rijsbergen: Information Retrieval. Butterworth 1979

[3] Buckley, C. and E. M. Voorhees. “Evaluating evaluation measure stability.” In ACM Conference on Research and Development in Information Retrieval, pages 33–40, Athens, Greece, 2000.

[4] Christopher D. Manning, Prabhakar Raghavan & Hinrich Schütze, “*An Introduction to Information Retrieval* 2009 Cambridge University Press.

[5] Eric C. Jensen, Steven M. Beitzel, Abdur Chowdhury, Ophir Frieder: Repeatable evaluation of search services in dynamic environments. ACM Trans. Inf. Syst. 26(1): (2006)

[6] Harman, D.K., 1993. Overview of the Second Text Retrieval Conference (TREC-2). In NIST Special Publication. Gaithersburg, MD, USA: Department of Commerce, National Institute of Standards and Technology.

[7] Järvelin, K. & Kekäläinen, J. (2000) IR evaluation methods for retrieving highly relevant documents, in *Proc. ACM SIGIR*, Athens, Greece, 41-48.

[8] [4] Robertson, Stephan 2006 Järvelin, K. & Kekäläinen, J. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems* 20, 422-446. 2002

[9] Mandl, T., 2009. Easy Tasks Dominate Information Retrieval Evaluation Results. In pp. 107-116.

[10] Mandl, T. et al., 2008. How robust are multilingual information retrieval systems? In Proceedings of the 2008 ACM symposium on Applied computing. ACM New York, NY, USA, pp. 1132-1136.

[11] Marchionini, G. (1995) Information Seeking in Electronic Environments Cambridge University Press.

[12] [33] Robertson, Stephan (2006). On GMAP: and other transformations. In: Proceedings of the 15th ACM International Conference on Information and Knowledge Management (CIKM) Arlington, Virginia, USA. pp. 872-877

[13] M. Sanderson, J. Zobel, “Information Retrieval System Evaluation: Effort, Sensitivity, and Reliability” , in the proceedings of the 28th ACM SIGIR conference, 2005

[14] Harman, D.K., 1993. Overview of the Second Text Retrieval Conference (TREC-2). In NIST Special Publication. Gaithersburg, MD, USA: Department of Commerce, National Institute of Standards and Technology.

[15] Saracevic, T. (1975). Relevance: Areview of and a framework for the thinking on the notion of information science. Journal of American Society for Information Science, 26(6), 321–343

[16] Sparck Jones, K. (1999). What is the role of NLP in text retrieval? In: T. Strzalkowski (ed.), Natural Language Information Retrieval, pp. 1–22.

[17] Tetsuya Sakai and Noriko Kando, On information retrieval metrics designed for evaluation with incomplete relevance assessments, Source: Information Retrieval archive, Volume 11 , Issue 5 (October 2008) table of contents, Pages: 447 – 470, Year of Publication: 2008, ISSN:1386-4564.

[18] Tetsuya Sakai Stephen Robertson “Modelling a User Population for Designing Information Retrieval Metrics,” The Second International Workshop on Evaluating Information Access (EVIA), December 16, 2008,

[19] Wen-Cheng Lin and Hisn-His Chen 2004. Merging multilingual information retrieval Results based on prediction of retrieval effectiveness, proceedings of NTCIR-4, Tokyo, April 2004.

[20] C.J. Van Rijsbergen," Information retrieval”, Butterworths, 1979.