Performance Assessment of AMRR and ADCG Metrics in MLIR and IR Systems

Raju Korra[*1], Pothula Sujatha[*2], Prof.P.Dhavachelvan*[3], Madarapu Naresh Kumar[*4], Sidige Chetana[*5]

School of Engineering and Technology, Department of Computer Science, Pondicherry University, Pondicherry-India

ABSTRACT

Multiple language documents retrieval is being done by using Multilingual Information Retrieval (MLIR) system. MLIR system deals with the use of queries in one language and retrieves the documents in various languages. The Query translation plays a central role in MLIR system research. We have used English as the source language and Hindi, French and German as the target languages. The experimental results are evaluated to analyze and compare the performance of proposed MLIR system metrics, Average Mean Reciprocal Rank (AMRR) and Average Discounted Cumulative Gain (ADCG), in Information Retrieval (IR) and MLIR system. Experimental results show that the performance of AMRR, ADCG in MLIR system has been improved 81.67%, 43.93% over IR system respectively.

Keywords

MLIR, Mean Reciprocal Rank, Cumulative Gain, Average Mean Reciprocal Rank, Average Discounted Cumulative Gain.

1. INTRODUCTION

We To measure MLIR effectiveness in the standard way, we need a test collection consisting of three things: (i) A document collection, (ii) A test suite of information needs, expressible as queries, (iii) A set of relevance judgments, standardly a binary assessment of either *relevant* or *non-relevant* for each query-document pair. The standard approach to multilingual information retrieval system evaluation revolves around the notion of *relevant* and *non-relevant* documents. With respect to a user information need for a query, a document in the test collection is given a binary classification [13] as either relevant or non-relevant. This decision is referred to as the gold standard or ground truth judgment of relevance. The test document collection and suite of information needs have to be of a reasonable size, we need to average performance over fairly large test sets, as results are highly variable over different documents and information needs for different queries. To properly evaluate a system, our test information needs must be germane to the documents in the test document collection, and appropriate for predicted usage of the system. These information needs are best designed by domain experts. Using random combinations of query terms as an information need is generally not a good idea because typically they will not resemble the actual distribution of information needs.

The goal of MLIR system is to make information accessible despite the consequences of language differences between the searcher and the information. This multilingual retrieval method involves monolingual and cross lingual language searches as well as merging their results. MLIR systems are completely deterministic. But the performance of an MLIR system for different queries can be quite different. To get a robust idea about the average performance of a system, the performance is measured over a set of queries in order to compute an average performance. Usually, the variation in retrieval performance across different queries is much larger than the variation of the averaged performance measure across systems (different hypotheses) [17] because some queries are much harder than others for all systems. This calls for hypothesis testing techniques, which are able to detect consistent and significant performance differences [18] between systems regardless of the noise introduced by query distinction.

Test collections are the principal tool used for comparison and evaluation of retrieval systems. These collections typically comprised of documents, queries (or topics), and relevance judgments have been a key part of multilingual information retrieval research for decades; the use of such collections is based on research and practice in collection formation (Sparck Jones & Van Rijsbergen, 1975; Voorhees & Harman, 1999) and measurement of retrieval effectiveness (Van Rijsbergen 1979, Ch. 7; Dunlop, 1997; Jarvelin, 2000; Buckley, 2004). Effectiveness is computed 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. In this paper we are evaluating MRR and DCG metrics i.e. Mean Reciprocal Rank is a statistic for evaluating any process that produces a list of possible responses to a query, ordered by probability of correctness and the Discounted Cumulative Gain score is a popular evaluator for multi-level relevance judgments [15]. In its indispensable form it has a logarithmic position discount, the benefit of considering a relevant document is position.

At a basic level, there are two approaches that can be taken when it comes to the design of a MLIR system. They are: translation of query language [9] or translation of the document language. Since in MLIR system the query language and document language be at variance, a query representation must be compared with each document representation in order to

determine the degree of similarity: In MLIR, either the query must be translated into the document language, or the document in the query language. Former way is a better way because translating a query one is much more efficient than translating each and every document in the collection into the query language. The advantage of query translation [10] rather than document translation is that a query translation module may be added to an existing IR system is easy, when compared with the cost of modifying the intact document base and redesigning the system for multilingual retrieval [20].

The remainder of this paper is organized as follows. Section 2 discusses related work to this study. Section 3 gives Proposed MLIR metrics along with comparative performance assessment of IR and MLIR systems and finally, Section 4 concludes the paper.

2. RELATED WORK

2.1 Mean Reciprocal Rank (MRR)

Mean Reciprocal Rank is a statistic for evaluating any process that produces a list of possible responses to a query, ordered by probability of correctness. The reciprocal rank of a query response is the multiplicative inverse of the rank of the first correct answer. The mean reciprocal rank is the average of the reciprocal ranks of results for a sample of queries Q [6].

$$MRR = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{rank_i} \quad (1)$$

2.1 Discounted Cumulative Gain (DCG)

Discounted Cumulative Gain is a measure of effectiveness of a Web search engine algorithm or related applications, often used in information retrieval. Using a graded relevance scale of documents in a search engine result set, DCG measures [14] the usefulness, or *gain*, of a document based on its position in the result list. The gain is accumulated from the top of the result list to the bottom with the gain of each result discounted at lower ranks [1].

2.1.1 Statistical Details

Two assumptions are prepared in using DCG and its correlated measures [17].

- Highly relevant documents are more useful when appearing earlier in a search engine result list (have higher ranks).
- Highly relevant documents are more useful than marginally relevant documents, which are in turn more useful than irrelevant documents.

DCG originates from a prior, more prehistoric, measure called Cumulative Gain.

2.1.2 Cumulative Gain (CG)

Cumulative Gain is the predecessor of DCG and does not include the position of a result in the consideration of the usefulness of a result set. In this way, it is the sum of the graded relevance values of all results in a search result list. The CG at a particular rank position p is defined as:

$$CG_p = \sum_{i=1}^{p} rel_i \quad (2)$$

Where rel_i is the graded relevance of the result at position *i*.

The value computed with the CG function is unaffected by changes in the ordering of search results. That is, moving a highly relevant document d_i above a higher ranked, less relevant, document d_j does not change the computed value for CG. Based on the two assumptions made above about the usefulness of search results, DCG is used in place of CG for a more accurate measure.

2.1.3 Discounted Cumulative Gain

The premise of DCG is that highly relevant documents appearing lower in a search result list should be penalized as the graded relevance [12] value is reduced logarithmically proportional to the position of the result. The discounted CG accumulated at a particular rank position p is defined as:

$$DCG_p = rel_1 + \sum_{i=2}^{p}\frac{rel_i}{log_2 i} \quad (3)$$

There has not been revealed any tentatively sound justification for using a logarithmic reduction factor [2] other than the fact that it produces a smooth reduction. An alternative formulation of DCG places stronger emphasis on retrieving relevant documents:

$$DCG_p = \sum_{i=1}^{p}\frac{2^{rel_i}-1}{log_2(1+i)} \quad (4)$$

The function is equivalent to the equation (3) DCG function when the relevance values of documents are binary,

$$rel_i \in \{0,1\}$$

2.1.4 Normalized DCG (nDCG)

Search result lists vary in length depending on the query. Comparing a search engine's performance from one query to the next cannot be consistently achieved using DCG alone, so the cumulative gain at each position for a chosen value of p should be normalized across queries. This is done by sorting documents of a result list by relevance, producing an ideal DCG (IDCG) at position p [14]. For a query, the normalized discounted cumulative gain, or nDCG, is computed *as*,

$$nDCG_P = \frac{DCG_P}{IDCG_P} \quad (5)$$

The nDCG values for all queries can be averaged to obtain a measure of the average performance of a search engine's ranking algorithm. Note that in a perfect ranking algorithm, the DCG_p will be the same as the $IDCG_p$ producing an nDCG of 1.0. All nDCG calculations are then relative values on the interval 0.0 to 1.0 and so is cross-query comparable. The main difficulty encountered in using nDCG is the unavailability of an ideal ordering of results when only partial relevance feedback is available.

The DCG measures advantages [5] are:

- It realistically weights down the gain received through documents found later in the ranked results.
- It allows modelling user persistence in examining long ranked results lists by adjusting the discounting factor.

The measures considered above, both the old and the new ones have weaknesses in two areas. Firstly, none of them take into account order effects on relevance judgments, or redundancy. In the Text Retrieval Conference (TREC) interactive track (Over 1999), instance recall is employed to handle this. The user–system pairs are rewarded for retrieving distinct instances of answers rather than multiple overlapping documents. In principle, the (D) CG measures may be used for such evaluation.

Second, the measures considered above all deal with relevance as a single dimension while it really is multidimensional (Schamber 1994) (Vakkari and Hakala 2000).In principle, such multidimensionality may be accounted for in the construction of recall bases for search topics but leads to complexity in the recall bases and in the evaluation measures. Nevertheless, such added complexity may be worth pursuing because so much effort is invested in IR evaluation.

Let l be a document cut-off value. The version of nDCG [4], they defined as:

$$nDCG = \frac{\sum_{r=1}^{l} g(r)/\,log(r+1)}{\sum_{r=1}^{l} g^{*}(r)/\,log(r+1)} \quad (6)$$

The original nDCG as defined in [1] is known to be "buggy" [3]. The above version of nDCG, first used in [2] and sometimes referred to as the Microsoft version, is free from this bug. Moreover, unlike the original nDCG, the choice of the logarithm base does not affect the Microsoft version. For evaluating system performance, they adopt the three official metrics used at NTCIR ACLIA IR4QA (NII (National Institute of Informatics) Test Collections for IR): Average Precision (AP), Q-measure (Q) and a version of nDCG [4].Q and nDCG, which can handle graded relevance, used the gain values of 3/2/1 for L3/L2/L1-relevant documents, respectively.

In these methods, the training data is composed of a set of queries, a set of documents for each query and a label or grade for each document indicating the degree of relevance [16] of this document to its corresponding query. For example, each grade can be one element in the ordinal set, and is assigned by human editors. The label can also simply be binary: relevant or irrelevant. Each query and each of its documents are paired together, and each query-document pair is represented by a feature vector,

{Perfect, excellent, good, fair, bad} (7)

In order to measure the quality of a search engine, they need some evaluation metrics. The Discounted Cumulative Gain has been widely used to assess relevance in the context of search engines (Jarvelin and Kekalainen, 2002) because it can handle multiple relevance grades such as (7).

Finally, the ranking problem is, in their opinion, more challenging when there are more than two relevance levels [15]. For this reason, they mostly focused on nDCG as an evaluation metric (because AP can only handle binary relevance) [19] and compare algorithms on the *ohsumed* dataset from the Lector benchmark (because it is the only one with more than 2 levels of relevance).

$$(N)DCG = \sum_{i=1}^{m} G(l_i)D\big(r(i)\big) \quad (8)$$

While the discount function is redefined as D(r) ← D(r)/Z with Z being the DCG obtained with the best ranking. Finally, one can also define the (N) DCG at a given truncation level k, denoted by (N)DCG@k by ignoring the documents after rank k, that is setting D(r) = 0 for r > k.

3. PROPOSED MLIR METRICS ALONG WITH COMPARATIVE PERFORMANCE ASSESSMENT

In the proposed MLIR metrics, we used dictionary based query translation using word to word translation. On the whole we have used 300 documents which include English, Hindi, German and French languages. Here we have considered English as the source language and French, German and Hindi as the target languages. The Google language translator is used for the Query translation.

3.1 Average Means Reciprocal Rank (AMRR)

Average mean reciprocal rank is a statistic for evaluating any process that produces a list of possible responses to a query, ordered by probability of appropriateness. The reciprocal rank of a query response is the multiplicative inverse of the rank of the first accurate answer. The mean reciprocal rank is the average of the reciprocal ranks of results for the given number of queries Q_T,

$$AMRR = \frac{1}{|Q_T|}\left[\sum_{i=1}^{|Q_T|}\left\{\frac{\sum_{j=1}^{|T_n|}\frac{1}{rank_j}}{|N_{rr}|}\right\}\right] \quad (9)$$

Where Q_T is the total number of queries, T_n is top n documents in the retrieved documents, $rank_j$ is j^{th} position rank in the retrieved documents and N_{rr} is the total number of relevant retrieved documents in $rank_j$.

We evaluated statistical measurement [18] of given 10 queries in IR and MLIR systems with the help of Google language translator tool; first we are giving all queries in IR system and evaluated reciprocal rank after that evaluated mean reciprocal rank. Simultaneously we evaluated for MLIR system of RR and MRR metrics. Both IR and MLIR systems statistical values are presented in table 1 and comparison of IR and MLIR systems RR, MRR and AMRR metric performance are shown in fig1, fig2 and fig3 respectively. Finally we improved the performance of MLIR system over the IR system.

Table 1. Statistical measurement of RR and MRR for IR and MLIR systems

Query no	RR_{IR}	MRR_{IR}	RR_{MLIR}	MRR_{MLIR}
1	3.1385	0.1207	0.6571	0.1314
2	2.9196	0.1327	0.2889	0.0963
3	2.8735	0.1916	0.569	0.1138
4	2.9839	0.1658	2.1855	0.1987
5	0.6614	0.0945	2.743	0.1829
6	1.6255	0.3251	0.5747	0.0958
7	2.9004	0.2072	2.3384	0.2126
8	4.3913	0.1568	1.6365	0.2728
9	2.4491	0.2449	0.1259	0.0419
10	2.0283	0.2535	1.7945	0.1994
	AMRR	0.1893		0.1546

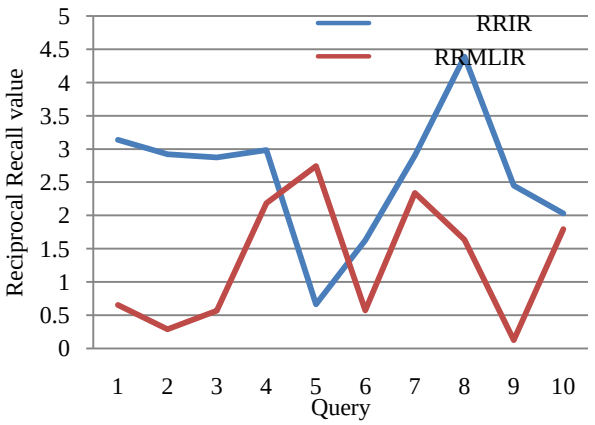

Fig 1: Comparative evaluation of RR in IR and MLIR systems

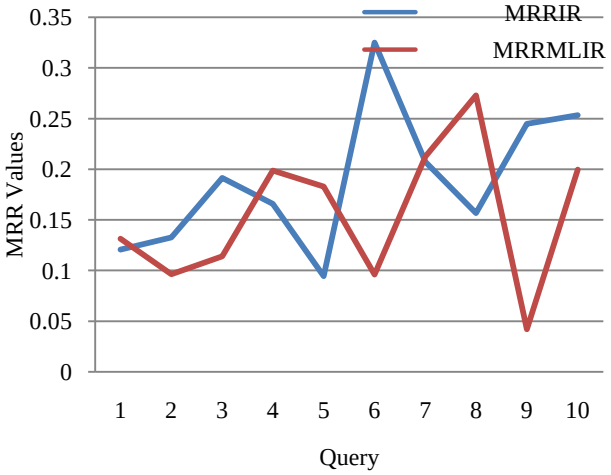

Fig 2: Comparative evaluation of MRR in IR and MLIR systems

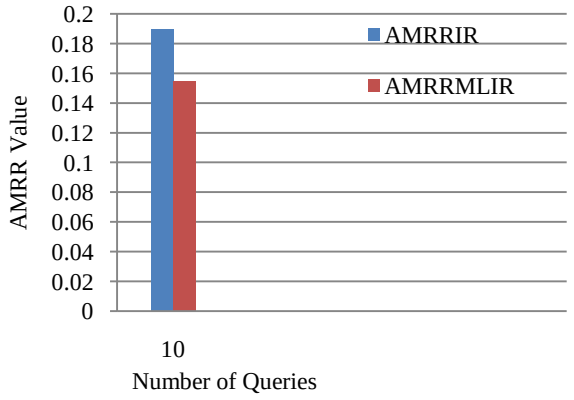

Fig 3: Comparative evaluation of AMRR in IR and MLIR systems

3.2 Average Discounted Cumulative Gain (ADCG)

The Average Discounted Cumulative Gain score (Jarvelin and Kekalainen, 2002) is a popular evaluator for multi-level relevance judgment [11]s. In its indispensable form it has a logarithmic position discount: the benefit of considering a relevant document at position j is $1/\log_2(1+j)$. Following (Burges et al, 2005), it became usual to assign exponentially high weight 2^{rel}_j to highly rated documents where rel_j is the grade of the j^{th} document going for instance from 0-irrelevant to 1- perfect relevant result. Thus the DCG for a ranking position j of a query having D_n associated documents is define as

$$ADCG_{MLIR} = \frac{1}{|Q_T|}\left[\sum_{i=1}^{|Q_T|}\frac{\left\{\sum_{j=1}^{|D_n|} 2^{rel_j-1}\right\}}{|RRD_i|}\right] \quad (10)$$

Where Q_T is the total number of queries, D_n is the top n retrieved documents, $rel_j \in \{0, 1\}$, 0-irrelevant document, 1-relevant document; RRD_i is relevant retrieved documents for each query.

Here we are using 10 queries to evaluate the ADCG metric for IR and MLIR systems with the help of Google language translator tool; firstly we are giving all queries in IR system and evaluated cumulative gain after that evaluated discounted cumulative gain. Simultaneously we evaluated for MLIR system of cumulative gain and ADCG metrics. Both IR and MLIR systems statistical evaluated values are presented in table 2 and comparison of IR and MLIR systems DCG and ADCG metrics performance are shown in fig4 and fig5 respectively. Finally we improved the performance of MLIR system over the IR system.

Table 2. Statistical measurement of DCG and ADCG for IR and MLIR systems

Query no	DCG_{IR}	DCG_{MLIR}
1	0.2909	0.2012
2	0.3043	0.1034
3	0.3549	0.3041
4	0.3354	0.1454
5	0.2754	0.0742
6	0.4692	0.296
7	0.3735	0.1753
8	0.3108	0.018
9	0.4039	0.1043
10	0.4075	0.1275
ADCG	0.3526	0.1549

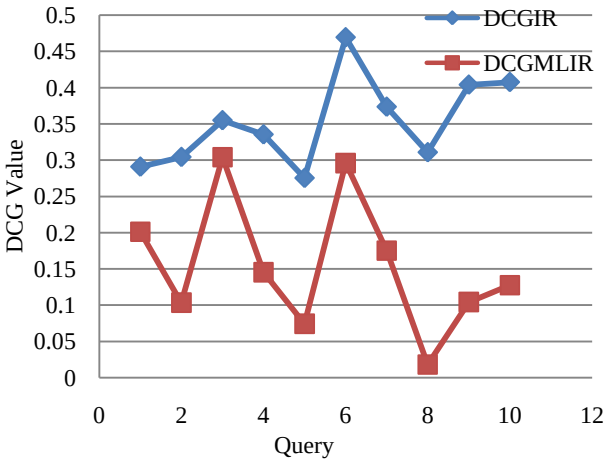

Fig 4: Comparative evaluation graph of DCG in IR and MLIR systems

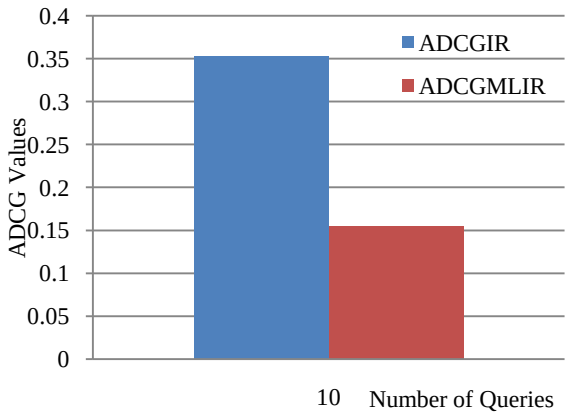

Fig 5: Comparative evaluation graph of ADCG in IR and MLIR systems

4. CONCLUSION

In this paper, we have proposed the MLIR metrics which helps in assessing the effective retrieval of documents in MLIR System. The proposed MLIR metrics AMRR, ADCG are evaluated and compared with the IR metrics MRR and DCG. Experimental results show that the performance of AMRR metric in MLIR system has been improved 81.67% over MRR metric of IR system and ADCG in MLIR system has been improved 43.93% over IR system.

5. REFERENCES

[1] Jarvelin, K. and Kekal ainen, J.: Cumulated Gain-Based Evaluation of IR Techniques, ACM TOIS, Vol. 20, No. 4, pp. 422-446, 2002.

[2] Burges, C. et al.: Learning to Rank using Gradient Descent, Proceedings of ACM ICML 2005, pp. 89-96, 2005.

[3] Sakai, T.: On Penalising Late Arrival of Relevant Documents in Information Retrieval Evaluation with Graded Relevance, Proceedings of the First Workshop on Evaluating Information Access (EVIA 2007), pp.32-43, 2007. Available at: http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings6/EVIA/1.pdf.

[4] Sakai, T., Kando, N., Lin, C.-J., Mitamura, T., Shima, H., Ji, D., Chen, K.-H., and Nyberg, E.: Overview of the NTCIR-7 ACLIA IR4QA Task, NTCIR-7 Proceedings, pp.77-114, December 2008.

[5] Kekalainen, J. & Jarvelin, K. (2002). User-oriented evaluation methods for information retrieval: A case study based on conceptual models for query expansion. In: Lakemeyer, G. & Nebel, B. (Eds.) Exploring Artificial Intelligence in the New Millennium. San Francisco: Morgan Kaufmann Publishers, pp. 355 - 379. ISBN 1-55860-811-7.

[6] E.M. Voorhees (1999). "Proceedings of the 8th Text Retrieval Conference". *TREC-8 Question Answering Track Report*. pp. 77–82.

[7] Contributors: Christian Fluhr Robert E. Frederking Doug Oard Akitoshi Okumura, Kai Ishikawa, and Kenji Chapter 2 Multilingual (or Cross-lingual) Information Retrieval,Editors: Judith Klavans and Eduard Hovy Satoh http://www.cs.cmu.edu/~ref/mlim/chapter2.html

[8] Chen-Hsin Cheng, Reuy-Jye Shue, Hung-Lin Lee, Shu-Yu Hsieh, Guann-Cyun Yeh, & Guo-Wei Bian: AINLP at NTCIR-6: *evaluations for multilingual and cross-lingual information retrieval* Proceedings of NTCIR-6 Workshop Meeting, May 15-18, 2007, Tokyo, Japan.

[9] Dan Wu, Daqing He, Huilin Wang. "Cross-Language Query Expansion Using Pseudo Relevance Feedback." Journal of the Chinese Society for Scientific and Technical Information. 29.2 (2010): 232-239.

[10] Qiang, Pu, Daqing He, Qi Li. "Query Expansion for Effective Geographic Information Retrieval." Evaluating Systems for Multilingual and Multimodal Information Access, 9th Workshop of the Cross-Language Evaluation Forum, CLEF 2008, Aarhus, Denmark, Revised Selected Papers. Springer. 2009.

[11] Sakai, T.: New Performance Metrics based on Multigrade Relevance: Their Application to Question Answering, NTCIR-4 Proceedings, 2004.

[12] Sakai, T.: On the Reliability of Information Retrieval Metrics based on Graded Relevance, Information Processing and Management, Vol. 43, Issue. 2, pp. 531-548, 2007.

[13] Kekalainen, J.: Binary and Graded Relevance in IR evaluations - Comparison of the Effects on Ranking of IR Systems, Information Processing and Management, Vol. 41, pp. 1019-1033, 2005.

[14] Sakai, T. and Robertson, S.: Modelling a User Population for Designing Information Retrieval Metrics, Proceedings of the Second International Workshop on Evaluating Information Access (EVIA 2008), pp.30-41, December 2008.

[15] Tetsuya Sakai, Noriko Kando: On information retrieval metrics designed for evaluation with incomplete relevance assessments. Information Retrieval 11 (5):447-470(2008).

[16] Olivier Chapelle, Mingrui Wu: Gradient descent optimization of smoothed information retrieval metrics. Inf. Retr. 13(3): 216-235 (2010).

[17] Jianfeng Gao, Endong Xun, Ming Zhou, Changning Huang, Jian-Yun Nie, Jian Zhang: Improving Query Translation for Cross-Language Information Retrieval Using Statistical Models. SIGIR 2001: 96-104.

[18] Marcello Federico, Nicola Bertoldi: Statistical cross-language information retrieval using n-best query translations. SIGIR 2002: 167-174.

[19] Sakai, T.: Average Gain Ratio: A Simple Retrieval Performance Measure for Evaluation with Multiple Relevance Levels, ACM SIGIR 2003 Proceedings, pp.417-418, July 2003.

[20] Manoj Kumar Chinnakotla, Karthik Raman, and Pushpak Bhattacharyya: Multilingual PRF: English lends a helping hand. SIGIR 2010: 659-666.