How Robust are Multilingual Information Retrieval Systems?

Thomas Mandl
Christa Womser-Hacker
Information Science
University of Hildesheim
Marienburger Platz 22
Germany
mandl@uni-hildesheim.de

Giorgio Di Nunzio
Nicola Ferro
Department of
Information Engineering
University of Padua
Italy
ferro@dei.unipd.it

ABSTRACT
The results of information retrieval evaluations are often difficult to apply to practical challenges. Recent research interest in the robustness of information systems tries to facilitate the application of research results for practical environments. This paper analyzes a large amount of evaluation experiments from the Cross Language Evaluation Forum (CLEF). Robustness can be interpreted as stressing the importance of difficult topics and is usually measured with the geometric mean of the topic results. Our analysis shows that a small decrease of performance of bi- and multi-lingual retrieval goes along with a tremendous difference between the geometric mean and the average of topics. Consequently, robustness is an important issue especially for cross-language retrieval system evaluation.

Categories and Subject Descriptors
H.3.4 [**Information Storage and Retrieval**]: Systems and Software – *performance evaluation.*

General Terms
Algorithms, Measurement

Keywords
Cross Language Evaluation Forum (CLEF), evaluation issues, Geometric mean

1. INTRODUCTION
Research has established a well accepted methodology to evaluate information retrieval systems [4]. Nevertheless, the discussion on appropriate metrics, test design and the human effort involved is ongoing. Evaluation initiatives compare the quality of systems by determining the mean average precision for standardized collections and topics as descriptions of information needs. The relevant documents for the topics are assessed by humans who work through all documents in a pool. The pool is constructed from the results of several systems and ultimately limits the number of relevant documents which can be encountered. The number of topics has been an issue of discussion. Obviously, if more topics are developed, the reliability of the results is higher.

Detailed analysis with small subsets of available retrieval results has led to the conclusion that 50 topics can produce a reliable result [2] and even 25 topics are sufficient [13]. On the other hand, there is research which calls for topics and smaller pools which contain much less documents [5]. Such a shallow pool requires less human effort for relevance assessment, nevertheless, increasing the number of topics is supposed to boost reliability. The human effort could even be further decreased if it is directed toward topics and documents which allow a better ranking of systems during the assessment [10]. Further research is necessary to investigate if that is the case for many evaluation settings.

Many different performance measures have been suggested for information retrieval. In recent years, binary preference (BPref) between relevant and non relevant documents has been widely adopted as a metric for test designs where only a small portion of the documents can be assessed [3]. The geometric mean has been suggested as a user oriented measures. Sometimes, these measures correlate highly. Then one might argue that new measures are not necessary. However, when the results in the system rankings differ strongly, then these metrics measure different aspects of retrieval systems [12]. It is not yet well understood what these aspects are and in which cases the use of a variety of measures makes sense.

In this paper, we analyze how the results of mono- and multi-lingual retrieval evaluation is affected by using mean average precision and a robust measure, in our case the geometric mean. The study intends to show whether a robust measure leads to different results for a multi-lingual retrieval test. Because robustness is a relevant issue for commercial systems, such an analysis is beneficial for further system development and future test design.

2. ROBUSTNESS IN INFORMATION RETRIEVAL EVALUATION
The RIA workshop [7] on reliable information access investigated the performance of systems for difficult topics in detail. Several reasons for poor performance and potential approaches for improvement were identified. An evaluation track for robust

SAC'08, March 16-20, 2008, Fortaleza, Ceará, Brazil.

retrieval has been established at the Text Retrieval Conference (TREC). This track does not only measure the average precision over all queries but it also emphasizes the performance of the systems for difficult queries. To perform well in this track it is more important for the systems to retrieve at least a few documents for difficult queries than to improve the performance in average [15]. The robust task is very user oriented because users often remember bad results better than positive search experiences. In order to allow a system evaluation based on robustness, more queries than for a normal ad-hoc track are necessary. The concept of robustness was extended in TREC 2005. Systems need to perform well over different tracks and tasks [15].

For multilingual retrieval, robustness is also an interesting evaluation concept because the performance between queries differs greatly similarly to other evaluation initiatives [[9]. Robustness in multilingual retrieval could be interpreted in several ways:

- Stable performance over all topics instead of high average performance (as at TREC)
- Stable performance over different tasks (as at TREC)
- Stable performance over different languages

A robust task has run twice within the Cross Language Evaluation Forum (CLEF). Our study analyzes the results of the first year (2006) and shows that measuring robustness is very useful for multi-lingual retrieval because the results obtained with robust measures differ more from the traditional retrieval measurements than for mono-lingual retrieval.

The robust task has been organized for the first time at CLEF 2006. The evaluation of robustness emphasizes stable performance over all topics instead of high average performance [15]. The perspective of each individual user of an information retrieval system is different from the perspective taken by an evaluation initiative. The users will be disappointed by systems which deliver poor results for some topics whereas an evaluation initiative rewards systems which deliver good average results. A system delivering poor results for hard topics is likely to be considered of low quality by a user although it may reach high average results.

A robust evaluation stresses performance for weak topics. This can be done by using the geometric average precision (GMAP) as a main indicator for performance instead of the mean average precision (MAP) of all topics [12]. Geometric average has proven to be a stable measure for robustness at TREC [15]. The robust task at CLEF 2006 is concerned with the multilingual aspects of robustness. It is essentially an ad-hoc task which offers mono-lingual and cross-lingual sub-tasks.

The robust task uses test collections previously developed at CLEF. These collections contain documents in six languages (Dutch, English, German, French, Italian and Spanish) and were used almost constantly during CLEF 2001, CLEF 2002 and CLEF 2003. There are approximately 1.35 million documents and 3.6 gigabytes of text in the collection [11].

3. ROBUSTNESS FOR TOP RUNS

The robust task at CLEF 2006 received 133 runs for ten sub-tasks and 100 topics [11]. The results prove again that the variance between the MAP between the topics is much higher than the variance between the MAP values of the systems (see figures 1a and 1b).

While the systems lie within a small performance corridor, the topics add the variance to the evaluation. The maximum is 1 and the minimum 0 for all sub-tasks. The second and the third quartile also cover much more performance difference.

Using different measures is beneficial to information retrieval when these measures are more reliable or measure a different aspect of retrieval performance. If a new measure is highly correlated with others it may not contribute much to the knowledge gained from a retrieval evaluation. As a consequence, we intend to reveal the effect of using GMAP within the robust task.

The overview of the first robust task at CLEF revealed little change in the results when MAP and GMAP rankings of systems were compared [11]. However, this was due to the fact that a maximum of five runs from the same number of groups are presented. Similar runs of the same group do not appear in the results overview. We conducted a detailed study to identify the relation between GMAP and MAP for the whole set of top runs which often perform in a quite similar way. We intended to analyze the effect of the topic set size. Maybe enlarging the topic set to 100 leads to a higher correlation between MAP and GMAP based system rankings.

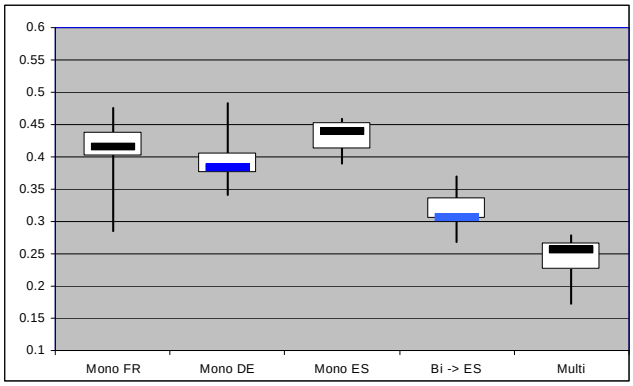

Figure 1a: Variance of systems' MAP values in the CLEF robust task 2006 (for mono-, bi- and multi-lingual runs)

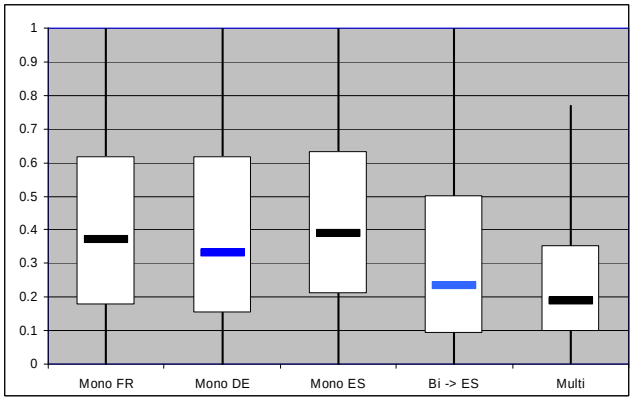

Figure 1b: Variance of topics' MAP values in the CLEF robust task 2006 (for mono-, bi- and multi-lingual runs)

In order to investigate this effect, the following methodology was used. The seven top runs for each sub-task within the robust task

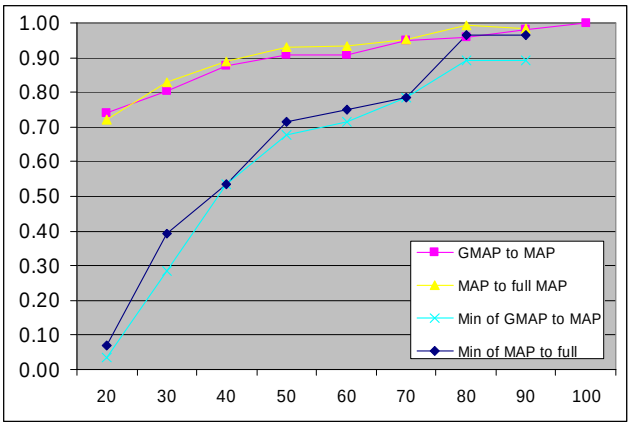

Figure 2a: Performance comparison for monolingual Dutch

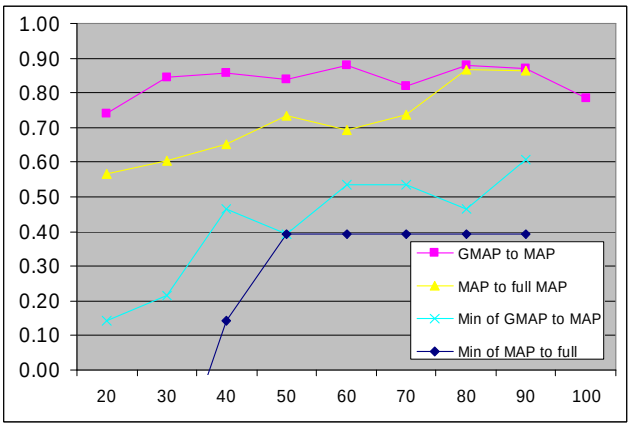

Figure 2b: Performance comparison for monolingual English

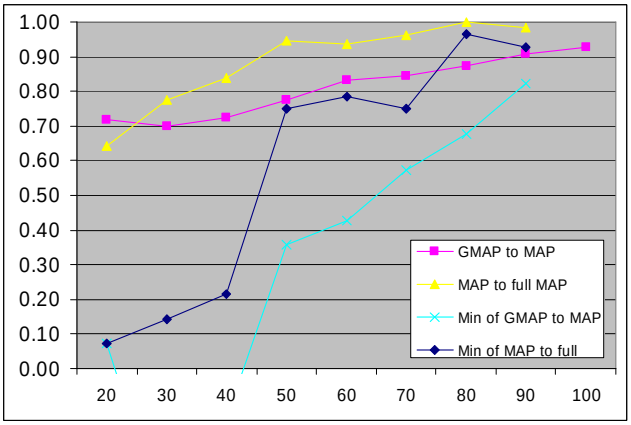

Figure 2c: Performance comparison for monolingual French

were determined based on the MAP values. This resulted in a set of different systems for each task. By doing that, weaker runs which often are quite different from the top runs are omitted. The remaining runs mostly perform quite similar. The number seven was used because for some sub-tasks, not many more runs were available. For this set, we determined the ranking of all runs (systems) based on the full topic set of 100 topics. The ranking was calculated based on the MAP as well as on the GMAP measure. The rankings of the systems were compared with the Pearson rank correlation coefficient.

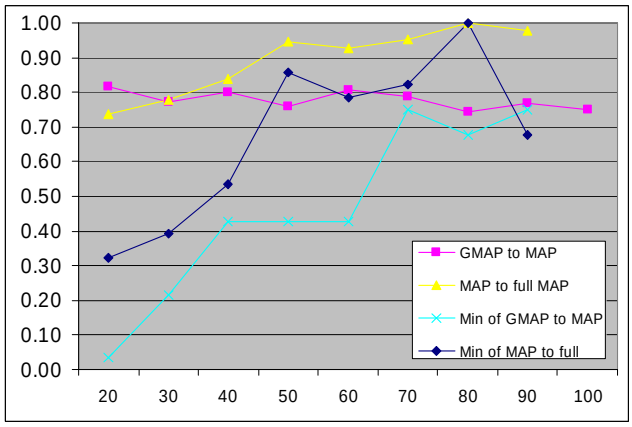

Figure 2d: Performance comparison for monolingual German

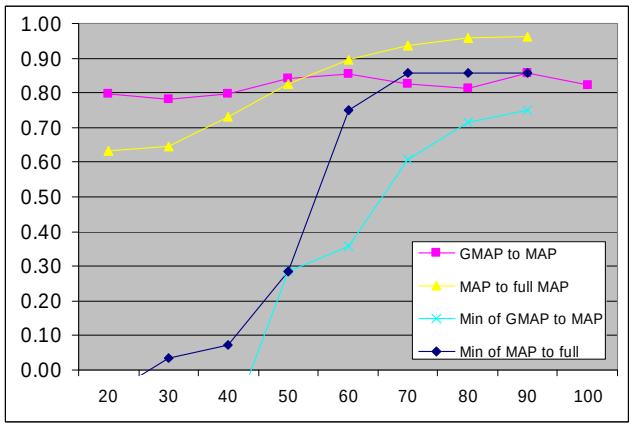

Figure 2e: Performance comparison for bilingual to Spanish

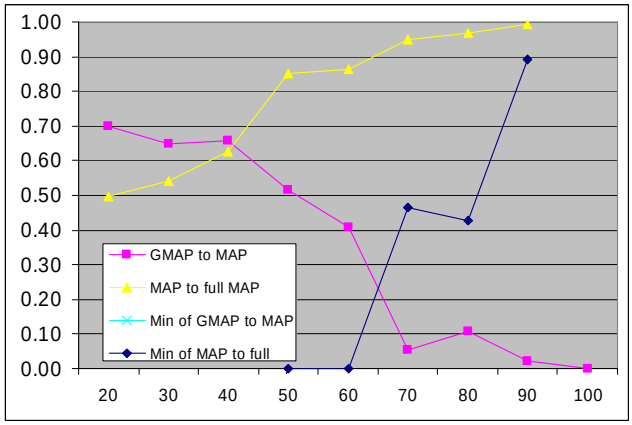

Figure 2f: Performance comparison for multilingual task

Subsequently, smaller sets of n topics were created by extracting n topics from the full set of 100 topics (with n = 20, 30, 40, 50, 60, 70, 80, 90). Due to the high number of all potential combinations of n out of 100, not all sets could be created. Instead, 100 combinations were created for each size n. Creating 200 and 400 variations for two values of n for one sub-task did not modify the results essentially. Therefore, 100 combinations were considered sufficient.

For each smaller set, we created a ranking of the systems based on MAP as well as GMAP and again the correlation between both

rankings. In addition, we calculated the correlation to the original ranking based on the full topic set for MAP. To have an idea of the lower border of the correlation, we also give the minimal value of all 100 variations for the last two measures. These four values were derived for the eight n and for eight out of the ten sub-tasks of the robust task. The numbers are presented for four mono-lingual, one bi-lingual and the multi-lingual sub-task in figures 2a through 2f.

It can be seen that the relation of partial MAP to full MAP increases for larger values of topic set size n. This can be expected. A similar analysis has previously been used to determine the minimal size of a topic set [13]. The curve reaches a high level for topic set size 50 and higher and also does so for cross-lingual retrieval. The curve for the minimal correlation is obviously below that curve. The correlation values are smaller for the multi-lingual task but also for the mono-lingual English task. This means that there are smaller topic sets which can be constructed out of the larger set which lead to very different results than the full set.

Our main interest lies in the MAP to GMAP correlation. Here mono- and cross-lingual retrieval behave quite differently. All values for mono- and bi-lingual runs lie above 0.7 which means that there is a strong correlation between the rankings. In most mono-lingual cases, the correlation increases with the topic set size. For the bi-lingual case, it remains almost the same for all n and for the multi-lingual case, it decreases from 0.7 to 0 for growing n. Rankings based on MAP and GMAP differ especially in the multi-lingual sub-task. This fact is further illustrated by showing the two rankings for the full set in figure 3. It can be observed, that some dramatic changes in ranking position occur. The Spearman rank coefficient for this particular task is 0.6 and Kendall's tau is 0.38. These values confirm that there is at most a weak correlation.

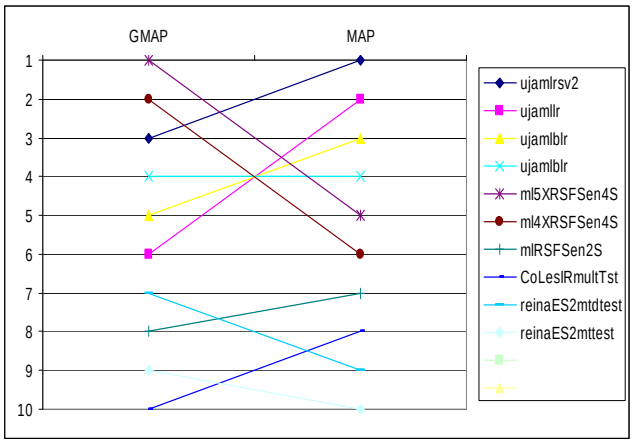

Figure 3: Comparison of ranking for GMAP and MAP

The change is quite dramatic. The performance for the multi-lingual runs decreases on average to 0.25 from around 0.4 for the mono-lingual runs. This means, that more topics become difficult in the multi-lingual set. Larger sets of multi-lingual topics contain more and more of these difficult topics and consequently, the correlation between MAP and GMAP decreases. Robustness is definitely an important issue for multi-lingual retrieval and it is not measured effectively by MAP.

The correlation values observed are also smaller than those observed for previous CLEF campaigns where 50 topics were used. These values are shown in table 1.

4. IDENTIFYING DIFFICULT TOPICS

Robustness emphasizes the performance of systems for difficult topics. Reasons for very poor performing topics have been suggested by several research groups [1,14]. Users often prefer systems which perform well for many queries. Therefore, robustness may reflect system performance better for many usage scenarios. Consequently, we did a review on topic difficulty measures.

Table 1: Rank Correlation according to the Spearman coefficient

Task	Topic Language	CLEF year	Correlation
Mono-lingual	German	2001	0.91
Multi-lingual	X	2001	0.96
Mono-lingual	Spanish	2001	0.93
Bi-lingual	English	2002	0.98
Multi-lingual	*X*	*2006*	*0.60*

4.1 Topic Difficulty

A typical approach to measure the difficulty of topics is the comparison of the estimation of experts against the actual outcome of the systems measured as the average precision which systems achieved for that topic. This approach was taken in a study of the topics of the Asian languages retrieval evaluation NTCIR [6]. No correlation was found between the two measures.

Furthermore, Eguchi et al. tried to find whether the system rankings change within the NTCIR evaluation campaign when different difficulty levels of topics were considered. They conclude, that changes in the system ranking occur, however, the Kendall correlation coefficient between the overall rankings does not drop below 0.69. For that analysis, the actual difficulty measured by the precision of the runs was used. The overall rankings remain stable; however, top ranks could be affected. It has to be noted that the number of topics was rather small after the creation of subsets.

4.2 Comparison of Definitions of Topic Difficulty

A query is usually considered as difficult when systems perform poorly for it. However, this can be measured in different ways. Mostly, average precision is used to find difficult topics [6, 8]. Furthermore, the most difficult topics could be determined by calculating the geometric average over all systems. That would increase the influence of low performing systems. On the contrary, the influence of the best systems could be increased. That would result in the topics for which even the best systems perform poorly. In order to find these topics, we considered the average precision of the best system for each topic.

Note that none of these measures is affected by the number of relevant documents present in the collection. For these measures, a difficult topic is not a topic for which there are few relevant documents in the collection. That might be a natural measure for

humans. As Voorhees has pointed out, no topic is inherently difficult. Topic difficulty is rather a complex function of topic and collection [15].

Especially for systems, it is a challenge to identify difficult topics and maybe apply specific processing methods. A typical approach is the expansion from an external collection like the web [8].

The following table 2 shows how much systems could benefit from focusing on hard topics and that there is ample room for improvement. For some examples of topics with a low average performance for all systems we examined how well the best system for that topic does. In addition, we show the performance of the best system for that topic. It can be seen that there is at least an improvement of 100% between the average and the best system.

Table 2: Examples for hard topics in the Robust task 2006

Task	Topic	Average	Best System for Topic	Best System Overall
Multi	118	0.0324	0.0682	0.0227
Multi	139	0.0412	0.0998	0.0997
Bi->ES	68	0.0090	0.0658	0.0058
Bi->ES	84	0.0538	0.2917	0.1327
Mono ES	111	0.0079	0.0221	0.0045
Mono ES	68	0.1055	0.2473	0.0904
Mono DE	111	0.0390	0.1671	0.1671
Mono DE	137	0.0393	0.1429	0.0556

5. CONCLUSION AND FUTURE WORK

The analysis presented here hints that an evaluation of multilingual retrieval focusing on robustness leads to substantially different results than standard evaluation measures. The analysis showed that cross-lingual retrieval with its inherent difficulty compared to mono-lingual retrieval greatly increases the divergence between rankings based on MAP and GMAP. Because robustness is a very relevant measure for the practical use of information retrieval systems, particular attention should be paid to robustness measures in cross-lingual retrieval.

For system developers, the challenge lies in the automatic identification and proper treatment of these cross-lingual hard topics and of hard topics in general [8].

6. REFERENCES

[1] Buckley C. Why current IR engines fail. *Proc Annual International ACM Conference on Research and Development in Information Retrieval* (SIGIR 2004) ACM Press, 584-585.

[2] Buckley, C.; Voorhees, E. The Effect of Topic Set Size on Retrieval Experiment Error. *Proc Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR '02)* (Tampere, Finland, Aug. 11-15, 2002). ACM Press, 2002, 316-323.

[3] Buckley, C.; Voorhees, E. M. Retrieval Evaluation with Incomplete Information, *Proc Annual International ACM Conference on Research and Development in Information Retrieval* (SIGIR 2004) 25-32.

[4] Buckley, C.; Voorhees, E. Retrieval System Evaluation. In: *TREC: Experiment and Evaluation in Information Retrieval*. Cambridge & London: MIT Press. (2005) 53-75.

[5] Carterette, B; Allan, J.; Sitaraman, R. Minimal test collections for retrieval evaluation. *Proc 29th Annual International ACM Conference on Research and Development in Information Retrieval* (SIGIR) Seattle. ACM Press, New York. (2006) 33-40.

[6] Eguchi, K.; Kando, N.; Kuriyama, K. Sensitivity of IR Systems Evaluation to Topic Difficulty. *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002)* (Las Palmas de Gran Canaria, Spain, May 29-31, 2002). ELRA, Paris, 585-589.

[7] Harman, D.; Buckley, C. The NRRC reliable information access (RIA) workshop. *Proc 27th annual international conference on Research and development in information retrieval* (SIGIR 2004). 528-529.

[8] Kwok, K. An Attempt to Identify Weakest and Strongest Queries. In: *Proc. Workshop Predicting Query Difficulty.*. at *28th Annual International ACM Conference on Research and Development in Information Retrieval* 2005. http://www.haifa.il.ibm.com/sigir05-qp

[9] Mandl, T.; Womser-Hacker, C. The Effect of Named Entities on Effectiveness in Cross-Language Information Retrieval Evaluation. *Proc ACM Symposium on Applied Computing (SAC).* Santa Fe, New Mexico, USA. March 13.-17. 2005. 1059-1064.

[10] Moffat, A.; Webber, W.; Zobel, J. Strategic System Comparisons via Targeted Relevance Judgments. *Proc 30th Annual International ACM Conference on Research and Development in Information Retrieval* (SIGIR) Amsterdam. 2007. 375-382.

[11] Di Nunzio, G.; Ferro, N.; Mandl, T.; Peters, C. CLEF 2006: Ad Hoc Track Overview. In: Peters, Carol et al. (Eds.). *7th Workshop of the Cross-Language Evaluation Forum, CLEF 2006, Alicante, Spain, Revised Selected Papers*. Berlin et al.: Springer [Lecture Notes in Computer Science 4730] 2007. 21-34. preprint at: http://www.clef-campaign.org/

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

[13] Sanderson, M.; Zobel, J. Information Retrieval System Evaluation: Effort, Sensitivity, and Reliability. *Proc 28th Annual International ACM Conference on Research and Development in Information Retrieval* (SIGIR) Salvador, Brazil. 2005. 162-169.

[14] Savoy, J. Why do successful search systems fail for some topics? In: *Proc ACM Symposium on Applied Computing* (SAC) Seoul, Korea. 2007. 872-877

[15] Voorhees, E. The TREC robust retrieval track. *ACM SIGIR Forum* 39 (1) 2005. 11-20.