How to Compare Bilingual to Monolingual Cross-Language Information Retrieval

Franco Crivellari, Giorgio Maria Di Nunzio, and Nicola Ferro

Department of Information Engineering – University of Padua
Via Gradenigo, 6/a – 35131 Padova – Italy
{crive, dinunzio, ferro}@dei.unipd.it

Abstract. The study of cross-lingual *Information Retrieval Systems (IRSs)* and a deep analysis of system performances should provide guidelines, hints, and directions to drive the design and development of the next generation *MultiLingual Information Access (MLIA)* systems. In addition, effective tools for interpreting and comparing the experimental results should be made easily available to the research community. To this end, we propose a twofold methodology for the evaluation of *Cross Language Information Retrieval (CLIR)* systems: statistical analyses to provide MLIA researchers with quantitative and more sophisticated analysis techniques; and graphical tools to allow for a more qualitative comparison and an easier presentation of the results. We provide concrete examples about how the proposed methodology can be applied by studying the monolingual and bilingual tasks of the *Cross-Language Evaluation Forum (CLEF)* 2005 and 2006 campaigns.

1 Introduction

The growing interest in *MultiLingual Information Access (MLIA)* is witnessed by the international activities which promote the access, use, and search of digital contents available in multiple languages and in a distributed setting, that is, digital contents held in different places by different organisations. In particular, the *Cross-Language Evaluation Forum (CLEF)*[1] aims at evaluating MLIA systems which operate on European languages in both monolingual and cross-lingual contexts. The experimental evaluation carried out on MLIA systems takes on a twofold meaning: on the one hand, it should provide guidelines, hints, and directions to drive the design and development of the next generation MLIA systems; on the other hand, effective tools for interpreting and comparing the experimental results should be made easily available to the community.

We focus our attention on the study of cross-lingual IRSs and on a deep analysis of performance comparison between systems which perform monolingual tasks, i.e. querying and finding documents in one language, with respect to those which perform bilingual tasks, i.e. querying in one language and finding documents in another language.

[1] http://www.clef-campaign.org/

G. Amati, C. Carpineto, and G. Romano (Eds.): ECIR 2007, LNCS 4425, pp. 533–540, 2007.

The work presented in this paper aims at improving the current way of comparing bilingual and monolingual retrieval, which is basically a comparison of two averages of performance measures, and strives to provide better methods and tools for assessing the performances. A twofold methodology for the evaluation of CLIR systems is proposed: statistical analyses to provide MLIA researchers with quantitative and more sophisticated analysis techniques; graphical tools to allow for a more qualitative comparison and an easier presentation of the results. We discuss an example of application of the proposed methodology by studying the monolingual and bilingual tasks of the CLEF 2005 and 2006 campaigns. Note that these application examples also serve the purpose of validating the proposed methodology in a real setting.

The paper is organized as follows: Section 2 introduces the proposed methodology; Section 3 describes the experimental setting used for applying the proposed methodology, provides the application examples and reports the experimental results; finally, Section 4 draws some conclusions and provides an outlook for future work.

2 Cross-Lingual Comparison Methodology

A common method used to evaluate how good bilingual retrieval systems are is to compare results against monolingual baselines. The *Mean Average Precision (MAP)* is often used as a summary indicator among different performance figures available in Information Retrieval. For example, the overviews of the last CLEF workshops report figures where the MAP of a bilingual IRS is around 80% of the MAP of a monolingual IRS for the main European languages [1,2,3]. Even the recent literature on CLIR evaluation [5] compares performances between a monolingual baseline MAP and a MAP of bilingual approach. However, some hints of a deeper analysis can be found: statistical and query-by-query performance analyses.

In order to go beyond the simple comparison of MAP values between monolingual and bilingual performances, we propose a comparison methodology consisting of two complementary techniques which are both based on a comparison of results on single topics: a deep statistical analysis of both the monolingual and the bilingual tasks, described in Section 2.1; and a graphical comparison of both the monolingual and the bilingual tasks, described in Section 2.2.

2.1 Statistical Analysis Methodology

As pointed out by [4], a statistical methodology for judging whether measured differences between retrieval methods can be considered statistically significant is needed and, in line with this, CLEF usually performs statistical tests on the collected experiments [1,2] to assess their performances. On the other hand, these statistical tests are usually aimed at investigating the differences among the experiments within the same task, e.g. the monolingual French experiments alone or the bilingual French experiments alone, but they do not perform any kind of cross-task analysis, i.e. some kind of direct comparison between monolingual and bilingual tasks.

Given the average performance for each single topic of the monolingual and bilingual task, we want to study the distribution of these performances and employ different statistical tests to verify the following conditions:

1. the distributions of the performances are similar. This suggests that bilingual systems behave in a similar way with respect to monolingual ones, which represent our empirical baseline of the best attainable performances;
2. the variances of the two distributions are similar. This suggests that even though the passing from one language to another causes a decrease in the performances, the effect of the translation does not increase the dispersion of performances, which would add more uncertainty;
3. the mean of the two distributions are different and, in particular, the mean of the monolingual distribution is greater than the mean of the bilingual one. This suggests some loss of performances due to the effect of the translations from one language to another.

Note that we do not aim to demonstrate whether all these conditions simultaneously hold or not. Rather, we want to develop an analysis methodology which allows researchers to gain better insights into these conditions. We can anticipate here from Section 3 that these conditions do not hold simultaneously for all the monolingual and bilingual tasks we have analysed even though the general claim is usually complied with.

In order to verify the first condition, since the distribution of the experiments is unknown, we can adopt a quantile-quantile plot, which allows us to compare the distribution of the monolingual experiments with respect to the distribution of the bilingual experiments. In a quantile-quantile plot the quantiles of the two distributions are increasingly ordered and compared and, if the samples do come from the same distribution, the plot will be linear.

The last two conditions are analyzed and studied by means of statistical tests for the equality of two variances and for the equality of two means; the tests that are used in the paper (the F-test and the t-test, respectively) assume that collected data are normally distributed. Therefore, before proceeding, we need to verify the normality of the involved distributions by using graphical tools for inspection (i.e. the boxplot, or the normality plot) or normality tests (i.e. the Lilliefors test, or the Jarque-Bera test). However, if the normality assumption is violated, a transformation of the data should be performed. The transformation for proportion measures that range from 0 to 1 is the arcsin-root transformation which Tague-Sutcliffe recommends for use with precision/recall measures.

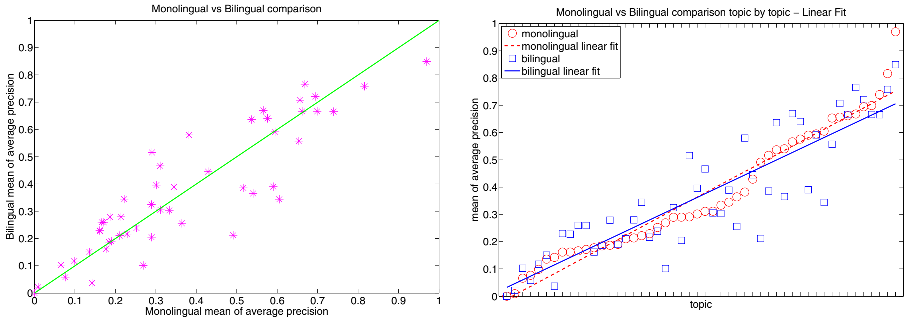
After the check on the normality of data, a test for the equality of variances, the F-test, is carried out to check whether the distributions have the same variance or not, and this step allows us to verify the second condition. Finally, in order to assess whether the mean of the monolingual performances is greater than the bilingual one, a t-test is used. In particular, since we have two paired sets (monolingual and bilingual) of m measured values, where m is the number of topics, the paired t-test is used to determine whether they differ from each other in a significant way under the assumptions that the paired differences are independent and identically normally distributed. This step allows us to verify the third condition reported above.

2.2 Graphical Comparison Methodology

The graphical tool which allows us to easily compare the performances for each topic of the monolingual and bilingual tasks, and to gain a visual explanation of the behavior of the two distributions, first needs a retrieval effectiveness measure to be used as a performance indicator. Then, we compute a descriptive statistic for the selected measure for each topic. In our case, the average precision and the mean were used, respectively. Finally, we increasingly order the monolingual topics by the computed descriptive statistic; the bilingual topics are ordered in the same order as the monolingual ones, because we are performing a topic-by-topic comparison and we want to compare a monolingual topic with the corresponding bilingual one. Note that ordering of the bilingual topics is usually different from what we would obtain if we increasingly ordered the bilingual topics by the computed descriptive statistic. From this ordered data we can produce two plots which provide us with an indicator for summarizing the trend of the monolingual and bilingual distributions. Examples of these plots are shown in Figure 1. For space reasons, the graphical examples are limited to only two plots.

Figure 1a shows the first plot where, for each topic, the monolingual performances on the x-axis, ordered increasingly, are plotted against the corresponding bilingual performances on the y-axis. If monolingual and bilingual behave in a similar way, the points are placed close and around the bisector of the first and third quadrant; on the other hand, if the monolingual performs better than the bilingual, the points are shifted towards the right whereas they are shifted symmetrically if the bilingual performs better than the monolingual. This plot provides us with a qualitative estimate about whether the three conditions introduced in the previous section hold: in that case, the plot would appear roughly linear and the points would be shifted towards the right. It is important to stress that this plot resembles a quantile-quantile plot but with an important difference: in a quantile-quantile plot the two distributions are independently ordered by their increasing quantiles, while in this plot the bilingual distribution is in the same order as the monolingual one.

Figure 1b presents a different view of the same data: for each topic on the x-axis, we plot on the y-axis both the monolingual (circles) and the bilingual performances (squares). This representation allows us to directly inspect the differences of the performances in a topic-by-topic fashion and provides us with hints about which topics require a deeper investigation because, for example, performances are too low or differences in the performances are too great. Moreover, this plot also allows us to qualitatively assess the three conditions reported in the previous section: in that case, the bilingual points would have a trend roughly similar to the monolingual ones and they would be below the monolingual ones.



(a) Example of direct comparison.

(b) Comparison on French 2006 data.

Fig. 1. Example of monolingual vs bilingual French 2006 comparison plots

Figure 1b also shows an example of linear fit: if the three conditions introduced in the previous section hold, the two straight lines would have roughly the same slope and the bilingual line would be right shifted with respect to the monolingual one.

3 Experimental Setting

The experimental collections and experiments used are fully described in [1,2].

In the CLEF 2005 and 2006 campaigns the languages of the target collection used for the monolingual and bilingual tasks were the same: Bulgarian, English, French, Hungarian, and Portuguese. However, we decided to take only those experiments of a bilingual task that have English as the source language in order to remove outliers from the data (for example extremely low performances due to very difficult languages such as Hindi or Amharic). Moreover, since we needed a sufficient number of experiments for each task to have reliable statistical analyses, we selected the tasks with the most experiments. These constraints led use to choose monolingual French (38 experiments for 2005 and 27 for 2006), monolingual Portuguese (32 experiments for 2005, 37 for 2006), bilingual French (12 experiments for 2005, 8 for 2006), and bilingual Portuguese (19 experiments for 2005, 18 for 2006). Remember that each one of these tasks has 50 topics.

For each task, we built a matrix $n \times m$ of n experiments and m topics where at position (i, j), with $1 \leq i \leq n$ and $1 \leq j \leq m$, we have the average precision (AP) of experiment e_1 on topic t_j. Then, we took the mean of the transformed performances by columns, that is, we took the average performances for each topic. As a result we had a vector for each task, like: $v_{task}^T = [mean_1 \; mean_2 \; \ldots \; mean_m]$, where $mean_1$ is the mean calculated for the first column, that is, the first topic of the task. The aim of the experimental analysis is to study the distribution of the mean of both the monolingual and bilingual tasks and compare them.

Table 1. Variance tests (F-tests) and Two-samples Paired t-test on CLEF 2005 data. Values in bold indicate hypotheses rejected.

		$H_0 : \sigma^2_{mono} = \sigma^2_{bili}$	$H_0 : \sigma^2_{mono} <= \sigma^2_{bili}$	$H_0 : \sigma^2_{mono} => \sigma^2_{bili}$
French	p-value	0.8281	0.5859	0.4141
Portuguese	p-value	0.9661	0.4831	0.5169
		$H_0 : \mu_{mono} = \mu_{bili}$	$H_0 : \mu_{mono} <= \mu_{bili}$	$H_0 : \mu_{mono} => \mu_{bili}$
French	p-value	0.8532	0.4266	0.5734
Portuguese	p-value	**0.0000**	**0.0000**	1.0000

The results presented are divided into years (2005 and 2006) and language (French and Portuguese). First the result of the normality test is presented, then the results of the analysis of variance are shown, and finally the analysis of the mean is discussed. Each calculation was carried out using MATLAB (version 7.2 R2006a) and MATLAB Statistics Toolbox (version 5.2 R2006a).

3.1 Statistical Analysis Methodology

Since the data resulted normal after a normality test, no arcsin-root transformation was adopted. In all the analyses, an alpha level of 5% was used.

CLEF 2005. The first analysis examines the variances of the data of the monolingual and the bilingual tasks. In Table 1, the results for the monolingual French vs bilingual French and the monolingual Portuguese vs bilingual Portuguese are presented. All the hypotheses are shown, starting from the most important one: the variances of the monolingual, σ^2_{mono}, and the bilingual, σ^2_{bili}, are equal. The other two hypotheses are important because the outcome shows that it is better not to reject them instead of accepting the alternative hypothesis which is, in those cases, σ^2_{mono} is either greater or less than σ^2_{bili}.

The second analysis considers the means of the monolingual, μ_{mono}, and bilingual, μ_{bili}, performances. Even though the hypothesis stated in Section 2.1, that is, the mean of the monolingual performances are better than the bilingual ones, is the main one, we believe it is important to consider all the aspects of the analysis. For this reason, we have presented the results for all the hypotheses in Table 1. It is interesting to see the differences between the French tests that result all in favor of the null hypothesis, that is to say it is preferable never to accept the alternative hypotheses that μ_{mono} is either greater or less than μ_{bili}. On the other hand, the analysis of Portuguese tasks shows that with the combination of all the hypotheses there is strong evidence that the mean of the performance of the monolingual Portuguese is greater than the bilingual one.

CLEF 2006. The analyses of the variances of the data of the monolingual and the bilingual tasks are shown in Table 2 for both the monolingual French vs bilingual French and the monolingual Portuguese vs bilingual Portuguese. All the tests confirm the hypothesis that the variances of the monolingual and bilingual tasks are equal.

Table 2. Variance tests (F-tests) and Two-samples Paired t-test on CLEF 2006 data. Values in bold indicate hypotheses rejected.

		$H_0 : \sigma^2_{mono} = \sigma^2_{bili}$	$H_0 : \sigma^2_{mono} <= \sigma^2_{bili}$	$H_0 : \sigma^2_{mono} => \sigma^2_{bili}$
French	p-value	0.8019	0.4009	0.5991
Portuguese	p-value	0.4270	0.7865	0.2135
		$H_0 : \mu_{mono} = \mu_{bili}$	$H_0 : \mu_{mono} <= \mu_{bili}$	$H_0 : \mu_{mono} => \mu_{bili}$
French	p-value	0.6860	0.3430	0.6570
Portuguese	p-value	**0.0001**	**0.0001**	0.9999

The two-samples paired t-test on the mean of the performances, shown in Table 2, confirm the outcome of the CLEF 2005: the tests on the French tasks are all in favor of the null hypothesis, that is to say the means are equal; the tests on the Portuguese tasks confirm that there is strong evidence that the mean of the performance of the monolingual Portuguese is greater than the bilingual one.

3.2 Graphical Comparison Methodology

In addition to the statistical analyses, we also present an effective graphical tool that gives a visual explanation of the behavior of the distributions of the monolingual and bilingual performances. Figures and plots were already shown in Section 2.2 and we cannot report the complete set of plots here for space reasons. On the other hand, we would like to comment on those plots in the light of the statistical analyses carried out in the previous section.

First, testing whether two distributions have similar shape and testing the normality of data can be done by means of standard tools such as the quantile-quantile plot and the normal probability plot. The quantile-quantile plots show that any monolingual-bilingual pair, both for French and Portuguese, has a regular linear trend, that is to say the shapes of the distributions are similar. The normal probability plot also shows the same regularity, which is sometimes violated along the tails of the distributions.

Second, the analysis of the performance on single topics can be appreciated in Figure 1: Figure 1a shows the performances on French 2006 data, ordered in the way explained in Section 2.2. This plot shows that there is a strong correlation between monolingual and bilingual performances; this correlation is above 0.80 for each pair of French, Portuguese tasks. Moreover, the cloud of points is located around the bisector of the first and third quadrant (above and below it) which means that, in this case, it is difficult to decide whether the monolingual is better than the bilingual or not. This confirms the results of the test on the means for French. Instead, the plot for the Portuguese (not shown here) shows a cloud of points which is more below than above the line, confirming that the monolingual Portuguese performs generally better than the bilingual.

In Figure 1b, a linear interpolation of the French 2006 tasks is performed. The two lines are very close and cross themselves; this figure clearly shows that even the linear interpolation of the monolingual and bilingual French data gives a positive response to the question that, in this case, the monolingual and bilingual performances are equal. Notice that we also have an indication of when the monolingual performance is better or worse than the bilingual; for example, for low performances bilingual performs better than monolingual while for high performances monolingual performs better.

4 Conclusions and Future Work

In this paper, we proposed a methodology which exploits both statistical analyses and graphical tools for the evaluation of MLIA systems. The statistical analysis provides MLIA researchers guidelines to drive the design and development of the next generation MLIA systems; the graphical tool provides a means to interpret experimental results and to present the results to other research communities easily. We provided concrete examples about how the proposed methodology can be applied by the analysis of the monolingual and bilingual tasks of the CLEF 2005 and 2006 campaigns.

Acknowledgements

The work reported in this paper has been partially supported by the DELOS Network of Excellence on Digital Libraries, as part of the Information Society Technologies (IST) Program of the European Commission (Contract G038-507618). The authors would like to warmly thank Carol Peters, coordinator of CLEF, for her continuous support and advice.

References

1. G. M. Di Nunzio, N. Ferro, G. J. F. Jones, and C. Peters. CLEF 2005: Ad Hoc Track Overview. In *Accessing Multilingual Information Repositories: Sixth Workshop of the Cross–Language Evaluation Forum (CLEF 2005). Revised Selected Papers*, pages 11–36. LNCS 4022, Springer, Heidelberg, Germany, 2006.
2. G. M. Di Nunzio, N. Ferro, T. Mandl, and C. Peters. CLEF 2006: Ad Hoc Track Overview. In *Working Notes for the CLEF 2006 Workshop*. `http://www.clef-campaign.org/2006/working_notes/workingnotes2006/dinunzioOCLEF2006.pdf`, 2006.
3. J. Gonzalo and C. Peters. The Impact of Evaluation on Multilingual Text Retrieval. In *Proc. 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2005)*, pages 603–604. ACM Press, New York, USA, 2005.
4. D. Hull. Using Statistical Testing in the Evaluation of Retrieval Experiments. In *Proc. 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1993)*, pages 329–338. ACM Press, New York, USA, 1993.
5. J. Wang and D.W. Oard. Combining Bidirectional Translation and Synonymy for Cross-Language Information Retrieval. In *Proc. 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2006)*, pages 202–209. ACM Press, New York, USA, 2006.