A Statistical and Graphical Methodology for Comparing 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/b – 35131 Padua – Italy
{crive, dinunzio, ferro}@dei.unipd.it

Abstract. A new methodology for the evaluation of *MultiLingual Information Access (MLIA)* systems is proposed. This two-fold methodology exploits both statistical analyses and graphical tools in order to provide MLIA researchers guidelines, hints, and directions to drive the design and development of the next generation systems, and to provide a means to interpret and compare experimental results and to present these results to other research communities. An example of the application of this methodology is applied in the real-case study of the monolingual and bilingual tasks of the CLEF 2005 and 2006.

Key words: multilingual information access, experimental evaluation

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. As an example, in the 7$^{\text{th}}$ European Community Framework Programme, the i2010 Digital Library Initiative clearly states that the improvement of multilingual and multicultural information access and search is one of the key objectives necessary to *provide access to quality digital content for all* [7, 8]. In addition, the workshop on "New Directions in Multilingual Information Access"[1] has pointed out the need for a stronger knowledge and technology transfer between the MLIA research community and other interested research communities, such as the the digital library community [12].

In this context, the experimental evaluation carried out on MLIA systems takes on a twofold meaning: on the one hand, it should provide guidelines,

[1] http://ucdata.berkeley.edu/sigir2006-mlia.htm

hints, and directions to drive the design and development of the next generation MLIA systems; on the other hand, the experimental results should be easily communicated to other research communities and effective tools for interpreting and comparing the experimental result should be made available to those research communities.

In recent years, the evaluation of MLIA systems has been carried out in important international evaluation forums which bring research groups together, provide them with the means for measuring the performances of their systems, and discuss and compare their work. In particular, the *Cross-Language Evaluation Forum (CLEF)*[2] aims at evaluating MLIA systems which operate on European languages in both monolingual and cross-lingual contexts.

We focus our attention on the study of cross-lingual *Information Retrieval System (IRS)* 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. Indeed, a common method used to evaluate performances for bilingual retrieval evaluation is to compare results against monolingual baselines. Different performance figures can be adopted to this aim: the *Mean Average Precision (MAP)* is often used as a summary indicator; for example, the recent literature reports figures where the MAP of a bilingual IRS is around 80% of the MAP of a monolingual IRS for the main European languages [5, 6, 9, 14].

The work presented in this paper aims at improving on this way of comparing bilingual and monolingual retrieval and strives to provide better methods and tools for assessing the performances. Another aspect of this work is that it can help the organizers of an evaluation forum during the topic generation process; in particular, the study of the hardness of a topic can be carried out with the goal of refining those topics which have been misinterpreted by systems. Currently, the research challenges described above pose two problems:

1. more sophisticated analysis techniques are needed to assess the performances of MLIA systems in order to effectively support the research for the next generation MLIA systems;
2. since other research communities are involved, we need effective methods and tools for communicating with other communities and to give them the means for easily assessing MLIA systems.

In this context, we propose a twofold methodology which exploits both thorough statistical analyses and graphical tools: the former will provide MLIA researchers with quantitative and more sophisticated analysis techniques, the latter will 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 CLEF 2005 and 2006 campaigns. Note that these application

[2] `http://www.clef-campaign.org/`

examples also serve the purpose of validating the proposed methodology in a real setting.

The paper is organized as follows: Sect. 2 introduces the proposed methodology; Sect. 3 describes the experimental setting used for applying the proposed methodology; Sect. 4 provides the application examples and reports the experimental results; finally, Sect. 5 draws some conclusions and provides an outlook for future work.

2 Cross-Lingual Comparison Methodology

The criteria normally adopted to create an experimental collection, consisting of suitable documents, sample topics and relevance judgements, have been adapted to satisfy the particular requirements of the multilingual context, where all language dependent tasks such as topic creation and relevance judgment are performed in a distributed setting by native speakers. In particular, the same set of topics is usually used to query all collections, whatever the task. When a monolingual task is performed, a given set of topics in the same language of the document collection is used (i.e., if a monolingual Portuguese task is performed, a collection of Portuguese documents is used as well as a set of topics in Portuguese); when a bilingual task is performed, the same document collection is used and the topics are the translation of the monolingual ones (i.e., if a bilingual Portuguese task is performed, a collection of Portuguese documents is used and a set of topics in a different language is used) [5, 6].

We exploit this way of constructing the experimental collections to go beyond the simple comparison of the MAP of a bilingual IRS with respect to a monolingual baseline given the same target language (for example, monolingual Portuguese vs. bilingual Portuguese). Indeed, we can perform an analysis on the results obtained on the single topics by monolingual and bilingual systems, because the different topics represent (in various languages) the same information needs, each bilingual topic is the direct translation of the corresponding monolingual one, and the same target test collections are used.

In particular, 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 Sect. 2.1. This kind of analysis allows us to address the problem of point 1 noted in the previous section;
- a graphical comparison of both the monolingual and the bilingual tasks, described in Sect. 2.2. This kind of comparison allows us to address the problem of point 2 noted in the previous section.

In order to present this methodology, we need to clearly define a measure that has not been used in literature. Given a task, for example monolingual Portuguese, we build a matrix $n \times m$ of n experiments and m topics, and define the following:

	t_1	t_2	...	t_m	MAP_e
e_1	AP_{e_1,t_1}	AP_{e_1,t_2}	...	AP_{e_1,t_m}	MAP_{e_1}
e_2	AP_{e_2,t_1}	AP_{e_2,t_2}	...	AP_{e_2,t_m}	MAP_{e_2}
...	...	...	...	...	...
e_n	AP_{e_n,t_1}	AP_{e_n,t_2}	...	AP_{e_n,t_m}	MAP_{e_n}
MAP_t	MAP_{t_1}	MAP_{t_2}	...	MAP_{t_m}	

where AP_{e_n,t_m} is the *Average Precision* for the m-th topic of the n-th experiment of the task; MAP_e is the mean of the average precision as known in the literature, that is to say the mean of the average precisions of an experiment across all the topics; MAP_t is the new measure that we introduce which is the mean of the average precisions for a topic across all the experiments of a task. This measure is important since for each topic it gives an indication of the average difficulty of a topic. With this measure, it is possible to compare the average performance between a monolingual task and the corresponding bilingual task (i.e. monolingual Portuguese and bilingual Portuguese) on a particular topic and study whether the translation process brought improvements in the retrieval. However, it is important to stress that the mean is a measure highly influenced by out of range values (or outliers) which, in this case, correspond to average precisions with very high or very low values; for this reason, the value of other statistics, for example the median, can be used together with the mean. The analysis of the difference between the values of the mean and the median can help to spot such cases which are worth a deeper study.

2.1 Statistical Analysis Methodology

As pointed out by [10], 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, 5, 6] 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 Portuguese experiments alone or the bilingual Portuguese 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 normal. This is the first condition to perform the following analyses;
2. the variances of the two distribution are similar. This suggests that even though the passing from one language to another causes a decrease in the performances, nevertheless 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. In fact, the general claim that "the best bilingual MAP_e is around 80% of the best monolingual MAP_e" suggests the idea that monolingual and bilingual systems behave roughly the same but there is some loss in the performances due to the translations. However, we believe that researchers need better tools to face the new challenges in the MLIA field and thus the general claim requires a deeper investigation. Indeed, the above described methodology allows us to have a more precise answer to the questions posed above; in fact, we can assess whether two distributions have comparable shapes, whether they have comparable dispersions, and finally whether they have comparable average measures.

Operatively, we use values of MAP_t calculated from the matrix $n \times m$ of average precisions for a monolingual task and the corresponding bilingual task given the same target language, and we study the conditions stated above on the two distributions, that we name $\text{MAP}_{t,mono}$ and $\text{MAP}_{t,bili}$.

In order to verify the first condition, we can adopt a normal probability plot, which allows us to compare the distribution of the monolingual experiments and the distribution of the bilingual experiments with respect to a normal distribution. In a normal probability plot, the quantiles of the distribution are increasingly ordered and compared to the quantiles of a normal distribution; if the samples do come from the same distribution, the plot will be linear. The last two conditions, same variance and same mean, 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 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 [2], or the Jarque-Bera test [11]). 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 [13] 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

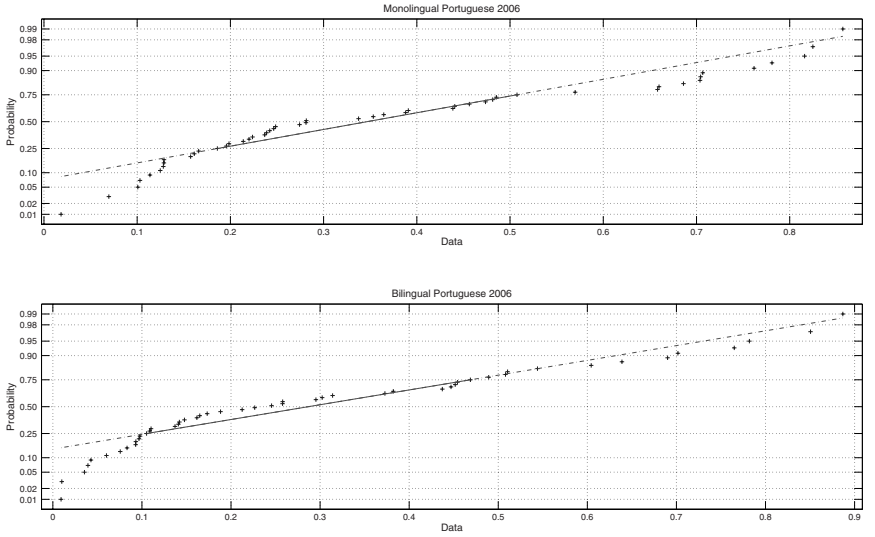


Fig. 1: Normal probability plot for monolingual and bilingual Portuguese 2006

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

In addition to the statistical analyses described in the previous section, we introduce simple and effective graphical tools which allow 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.

Before introducing specific plots for a topic by topic analysis, we can use standard statistical plots to help understand the type of distributions under consideration. In particular, an example of this a normal probability plot is shown in Fig. 1 for Portuguese monolingual and bilingual tasks. It can be seen that there is almost a perfect match with the straight line that represents the ideal line which indicates the same distribution. In order to understand if the deviation from the straight line is significant or not, statistical tests can be used to verify normality, as suggested in Sect. 2.1.

When a comparison between monolingual and bilingual results is required, a retrieval effectiveness measure to be used as a performance indicator has to first be selected; in our case, we used the average precision. Then, we compute descriptive statistics for the selected measure for each topic; in our studies we used the mean, which is useful when the distributions do not have many

outliers, and the median, which is more robust in the case of outliers. We already defined the MAP_t as the mean of average precisions per topic, and the median of average precision per topic will also be useful, here named $MEAP_t$.

We want to study the performances of specific topics and present data and results with different plots with different goals in mind: on the one hand, we want to give specific hints to the participants about how difficult a topic is, and receive feedback from them in particular for those topics where there is a *visible* difference between the MAP_t and the $MEAP_t$; on the other hand, we want to study the general results given a monolingual task and bilingual on a particular language and compare the differences, with the aim of analyzing what the most difficult topics in general are (hard topics in monolingual) and the most difficult ones to translate are (hard topics in bilingual, with a significant difference with the respective monolingual result). This graphical analysis should not only help participants in building their Information Retrieval System, but also the organizers of evaluation forums in analyzing deeply what went wrong during the topic creation phase so they can improve the quality of the queries the following year, that is to say by analyzing whether the difficulty of a topic was due to too narrow topics, misspelling errors, or misinterpreted topics.

In a topic by topic analysis, the natural ordering of topics (i.e. the identifier of the topic) can be used to represent the graphs. However, for visualization aspects, it is more convenient to order topics by one of the computed descriptive statistic. Since we are performing a topic-by-topic comparison and we want to compare a monolingual topic with the corresponding bilingual one, topics are ordered according to monolingual results first, and the bilingual topics according to the same order of topics of the monolingual task. 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.

Topic by Topic Analysis, Monolingual and Bilingual Separately

Figure 2 summarizes the results for monolingual and bilingual Portuguese 2006; topics are ordered according to the MAP_t, the upper part shows the performances for the monolingual while the bottom part the bilingual ones. For each topic, the average precision of an experiment is plotted with a star; a square represent the MAP_t and a triangle the $MEAP_t$. At a glance, we can see for each task the average performance of each topic, and if the distance between the square and the triangle is evident we can immediately draw the following conclusions: if the square is above the triangle, it means that the MAP_t was highly influenced by some experiments that performed much better than at least half of the remaining experiments; in the other case, the MAP_t was highly influenced by experiments that performed much worse than at least half of the remaining experiments. Both situations are worth a deeper

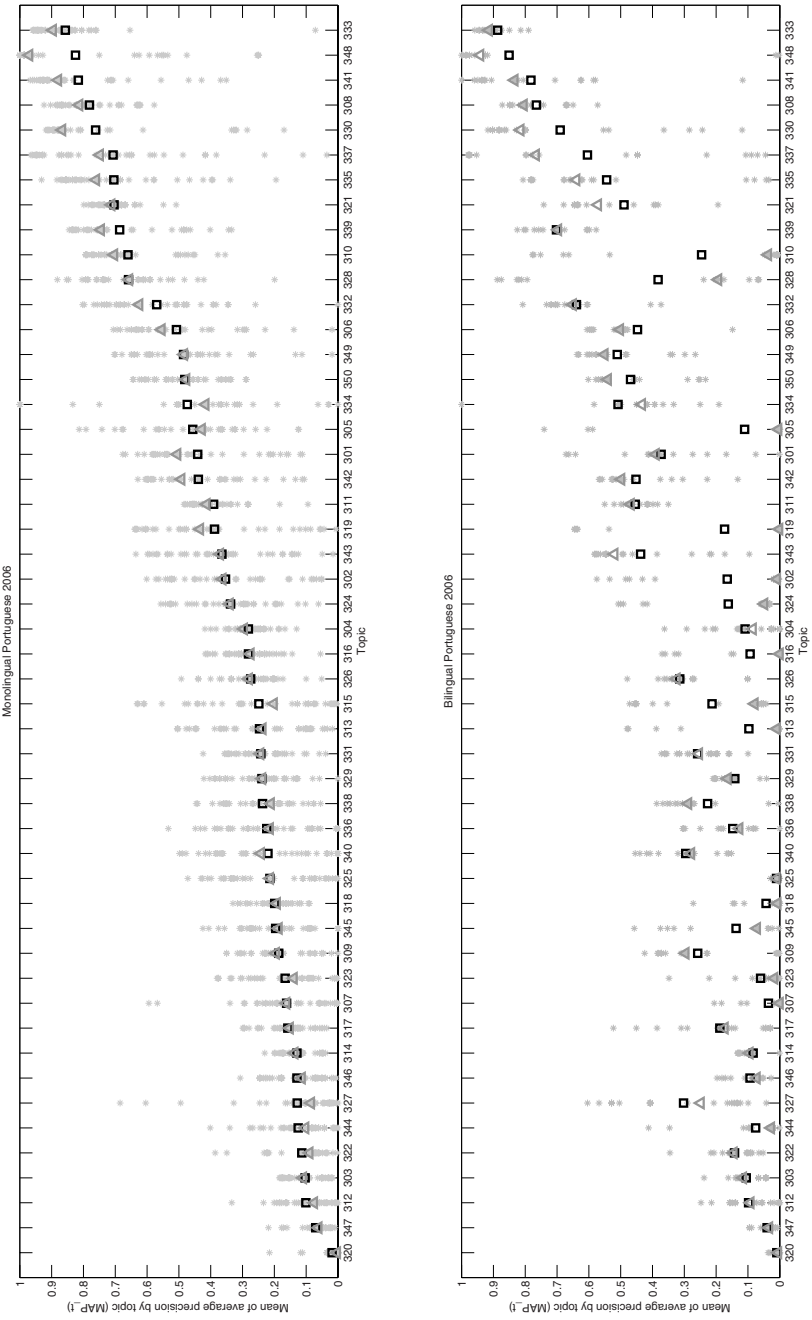


Fig. 2: Example of monolingual vs bilingual Portuguese 2006 comparison plot

study: in the former case, we would be interested in asking the participant who produced the experiment with that performance to study what are the causes of this success (i.e. feedback, query expansion, dictionary, etc.); in the latter case, we would be interested to ask that participant to perform a deeper failure analysis.

With Fig. 2 it is also possible to compare the results of monolingual with bilingual. If we pick a topic in the monolingual part and follow a vertical line, we can read the corresponding performance for the bilingual. For example, topic 327 (left part of the plot) has a MAP_t around 0.1 for monolingual and about 0.3 for the bilingual; this is a case where the process of translation produced a better result. For topic 316 (in the middle), the MAP_t of the monolingual is around 0.3 while the bilingual is about 0.1; this case is interesting for another reason, the $MEAP_t$ of the bilingual is equal to 0.0, which means that at least half of the experiments did not retrieve any relevant document for that topic. However, some experiments did very well (around 0.4), and these are the kinds of experiments that are worth a deeper analysis.

Topic by Topic Analysis, Monolingual and Bilingual Together

Figure 3 shows, for each topic and ordered by monolingual MAP_t, the monolingual MAP_t performances on the x-axis (red circle) plotted against the corresponding bilingual MAP_t performances on the y-axis (blue diamond); a line that highlights the differences between the two values is also shown. If monolingual and bilingual behaved in a similar way, the points would intersect each other and no line would be visible. 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 deeper investigation on performances can be done after a study of Fig. 3: one is about the study of difficult topics; since the left part of this figure shows the topics where the monolingual experiments perform worse, it would be important to study these topics, for example the first quartile of this distribution (topics with lowest MAP_t), and review them together with participants to see if any technological barrier avoided a better performance and with organizers to check and review the topics again to spot any flaws. Both are important because participants would be advised where to improve their systems, and organizers would be advised how to improve the evaluation forum the year after with higher quality topics. Another possible study is to order topics according to the difference in performance between monolingual and bilingual, take the first quartile of this distribution (topics where the difference is negatively higher, or where the line on the plot is longer) and make a deep failure

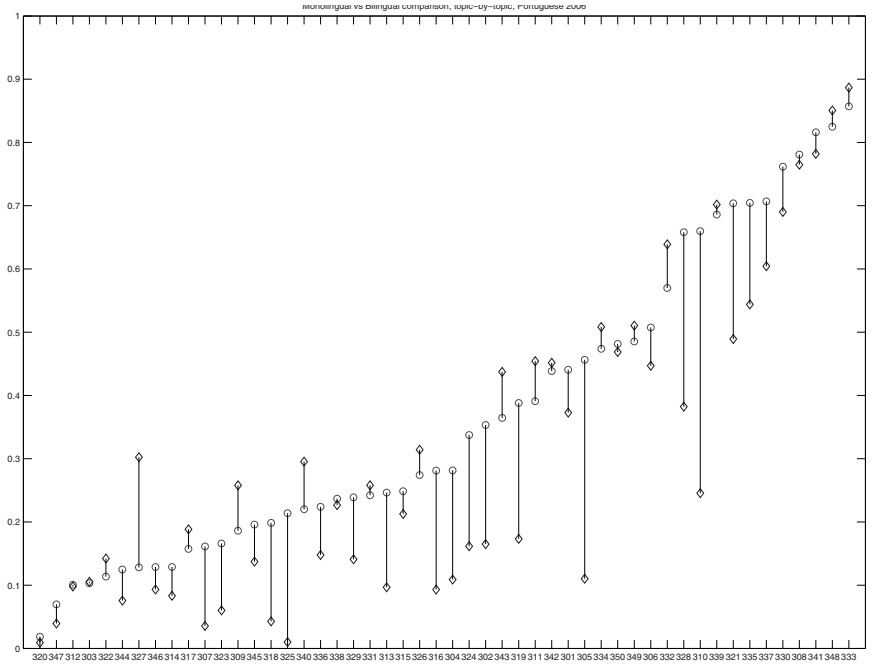


Fig. 3: Example of monolingual vs bilingual Portuguese 2006 comparison plot

analysis in order to understand why the process of translation for this set of topics introduced such a deterioration on the topic.

Monolingual and Bilingual Fitting

We can also use the plot in Fig. 3 as a starting point for a further analysis by interpolating the two series of points in order to compare their trend. In this way, not only do we strengthen the visualization of the behavior of the two distributions and improve their qualitative analysis, but we also bridge the gap between qualitative and quantitative analysis because interpolation techniques provide us with many quantitative indicators about how well an interpolation fits the data.

Figure 4 shows an example of linear fit where, for each topic and ordered by monolingual MAP_t, the monolingual MAP_t performances on the x-axis (red circle) is plotted against the corresponding bilingual MAP_t performances on the y-axis (blue diamond); the least squares fitting lines are drawn for the two tasks, solid for monolingual and dashed for bilingual. 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. Two situations are worth a deeper study: the

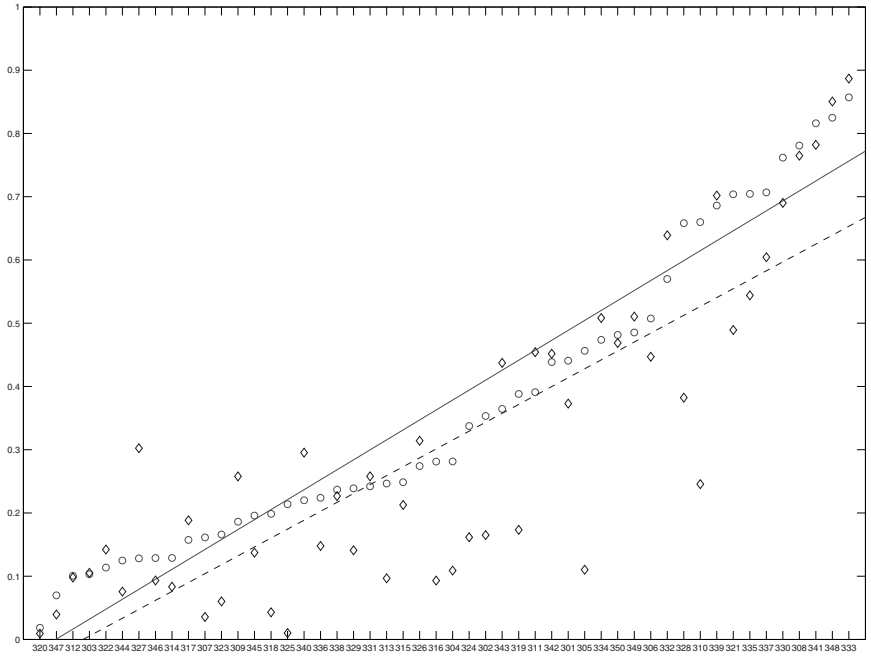


Fig. 4: Example of linear fit of the monolingual and bilingual distribution Portuguese 2006

distance between the two lines and whether the two lines intersect. In the first case, the greater the distance, the greater the difference in performance on average between the monolingual and the bilingual tasks. Figure 4 clearly shows that the monolingual performance is higher with respect to the bilingual one. An example of the second case is shown in Fig. 5 for French 2006; this is a situation where for the most difficult topics (considering monolingual performances) the bilingual task performs better on average, and viceversa for the other topics.

3 Experimental Setting

The experimental collections and experiments used are fully described in [5,6] while in [3, 4] the detailed experimental results are reported.

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. Since for the bilingual task an experiment may use as the source language one of a set of possible choices (for example, English to French, or German to Portuguese) the performance of a

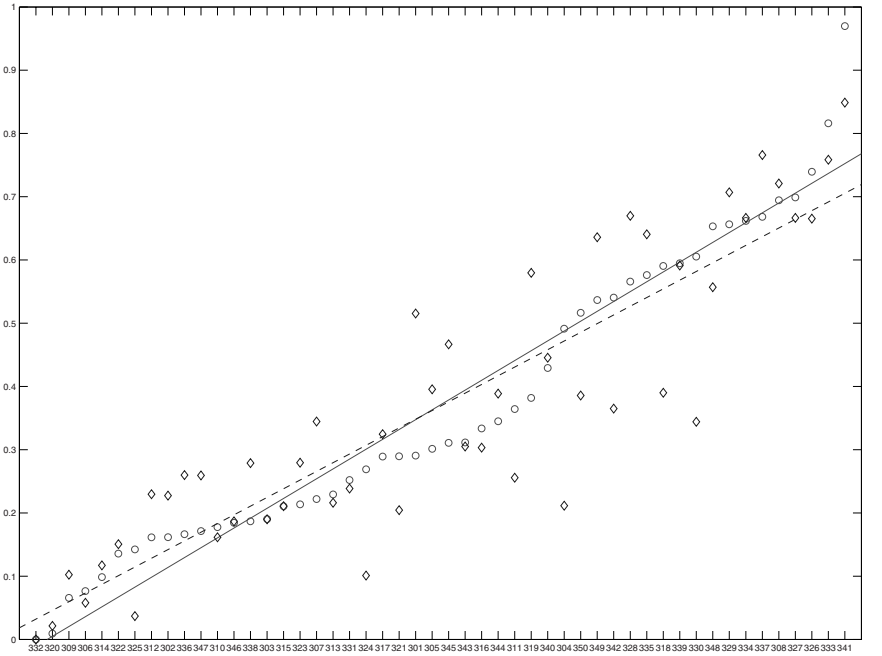


Fig. 5: Example of linear fit of the monolingual and bilingual distribution French 2006

bilingual task can be biased by languages that are particularly difficult, although interesting to study for *Information Retrieval (IR)* research goals (for example, languages such Hindi, Indonesian, or Amharic). In order to make the comparison between monolingual and bilingual performances as flawless as possible we decided to take only those experiments of a bilingual task that have English as the source language. 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, as reported in Table 1. Remember that each one of these tasks has 50 topics.

Table 1: Number of experiments per tasks

Track	# Runs	
	CLEF 2005	CLEF 2006
Monolingual French	38	27
Monolingual Portuguese	32	37
Bilingual English to French	12	8
Bilingual English to Portuguese	19	18

For each task, we built a matrix $n \times m$ of n experiments and m topics:

	t_1	t_2	...	t_m
e_1	AP_{e_1,t_1}	AP_{e_1,t_2}	...	AP_{e_1,t_m}
e_2	AP_{e_n,t_1}	AP_{e_n,t_2}	...	AP_{e_n,t_m}
...	...	...	...	...
e_n	AP_{e_n,t_1}	AP_{e_n,t_2}	...	AP_{e_n,t_m}

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:

$$v_{task}^T = [MAP_{t_1}\ MAP_{t_2}\ \ldots\ MAP_{t_m}]\ ,$$

where MAP_{t_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.

4 Application Example and Experimental Results

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).

4.1 Statistical Analysis Methodology

Since the data proved 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 2, the results for the monolingual French versus 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}.

Table 2: Variance tests (F-tests) on CLEF 2005 data

		$H_0 : \sigma^2_{mono} = \sigma^2_{bili}$	$H_0 : \sigma^2_{mono} <= \sigma^2_{bili}$	$H_0 : \sigma^2_{mono} => \sigma^2_{bili}$
French	p–value	0.8281	0.5859	0.4141
	outcome	not reject	not reject	not reject
Portuguese	p–value	0.9661	0.4831	0.5169
	outcome	not reject	not reject	not reject

The second analysis considers the means of the monolingual, μ_{mono}, and bilingual, μ_{bili}, performances. Even though the hypothesis stated in Sect. 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 3. 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 4 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.

The two-samples paired t-test on the mean of the performances, shown in Table 5, confirms 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.

Table 3: Two-samples Paired t-test on CLEF 2005 data

		$H_0 : \mu_{mono} = \mu_{bili}$	$H_0 : \mu_{mono} <= \mu_{bili}$	$H_0 : \mu_{mono} => \mu_{bili}$
French	p–value	0.8532	0.4266	0.5734
	outcome	not reject	not reject	not reject
Portuguese	p–value	0.0000	0.0000	1.0000
	outcome	**reject**	**reject**	not reject

Table 4: Variance tests (F-tests) on CLEF 2006 data

		$H_0 : \sigma^2_{mono} = \sigma^2_{bili}$	$H_0 : \sigma^2_{mono} <= \sigma^2_{bili}$	$H_0 : \sigma^2_{mono} => \sigma^2_{bili}$
French	p–value	0.8019	0.4009	0.5991
	outcome	not reject	not reject	not reject
Portuguese	p–value	0.4270	0.7865	0.2135
	outcome	not reject	not reject	not reject

4.2 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 Sect. 2.2 and we cannot report the complete set of plots here for space reasons. On the other hand, we want to comment on those plots in the light of the statistical analyses carried out in the previous section.

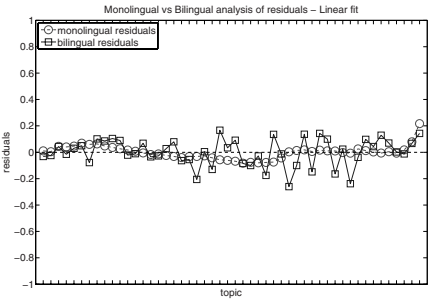
First, testing whether two distributions have similar shape and testing the normality of data can be done by means of standard tools such as the quantile-quantile plot and the normal probability plot. Quantile-quantile plots show that any monolingual-bilingual pair, both for French and Portuguese, has 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.

Figure 2 focuses on the difference in performance of the single topic for Portuguese 2006 tasks; a line that connects two points highlights the difference between the monolingual and bilingual performances. This plot shows at a glance an immediate snapshot of what the hardest topics are.

In addition, Fig. 4 and Fig. 5 show a further analysis of the same data by interpolating the two series of points in order to extrapolate and compare their trend. In Fig. 5, 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

Table 5: Two-samples Paired t-test on CLEF 2006 data

		$H_0 : \mu_{mono} = \mu_{bili}$	$H_0 : \mu_{mono} <= \mu_{bili}$	$H_0 : \mu_{mono} => \mu_{bili}$
French	p–value	0.6860	0.3430	0.6570
	outcome	not reject	not reject	not reject
Portuguese	p–value	0.0001	0.0001	0.9999
	outcome	**reject**	**reject**	not reject



(a) French 2006 analysis of residuals.

		2005		2006	
		SSE	R^2	SSE	R^2
FR	mono	0.0292	0.9865	0.1312	0.9506
	bili	0.5601	0.7645	0.5268	0.7887
PT	mono	0.1335	0.9167	0.1585	0.9421
	bili	0.6306	0.5098	0.8501	0.7084

(b) Goodness of the linear fit.

Fig. 6: Analysis of residuals of linear fitting (monolingual vs bilingual French 2006) and goodness-of-fit measures for all tasks of French (FR) and Portuguese (PT)

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. In Fig. 4 the interpolation is done on the Portuguese 2006 data. In this case, the interpolating line of the bilingual is clearly below the monolingual one, confirming the test done on the analysis of the means, the output of which was that the mean of the monolingual task was greater than the bilingual one.

Usually, when a linear interpolation is performed, it is important to assess how well the line fits the actual data. This analysis can be done by means of a graphical inspection of the plot of residuals or by means of some measures. In Fig. 6a the analysis of residuals plot is shown for French 2006 data. It is interesting to note how the ordered monolingual performance fits almost perfectly while the bilingual one is evenly distributed around the line. In general, we also noted that the tails of the residuals are usually far from the best fitting line. In Fig. 6b the sum of squares error (SSE) and the squared correlation coefficient (R^2) are shown. The SSE is close to zero both for the monolingual interpolation and the bilingual interpolation, which means a good interpolation. The R^2 is above 0.90 in many cases, which confirms that when the performances are ordered from the worst to the best, the shape of the scatterplot produced is very close to a linear one.

5 Conclusions

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.

A definition of a more general framework for the statistical analyses of results is one of the points of future work. In particular, we would not like to limit analysis to the situation where there are only two levels of independent variables, but to generalize it with techniques of the analysis of variance. However, this generalization requires a careful study of the not–so–easy situation of having together sets of repeated (an experiment tested on different topics) and paired (monolingual topic vs bilingual topic) measures. When the general framework is clearly defined, we will be able not only to answer questions such as whether monolingual is better than bilingual, but also to study the variability of the performances due to the differences among topics, the variability of performances due to the differences among experiments, as well as the variability of the interaction between these two factors.

References

1. Braschler, M., Di Nunzio, G.M., Ferro, N., Peters, C.: CLEF 2004: Ad Hoc Track Overview and Results Analysis. In: C. Peters, P. Clough, J. Gonzalo, G.J.F. Jones, M. Kluck, B. Magnini (eds.) Multilingual Information Access for Text, Speech and Images: Fifth Workshop of the Cross–Language Evaluation Forum (CLEF 2004) Revised Selected Papers, pp. 10–26. Lecture Notes in Computer Science (LNCS) 3491, Springer, Heidelberg, Germany (2005)
2. Conover, W.J.: Practical Nonparametric Statistics, 1st edn. John Wiley and Sons, New York, USA (1971)
3. Di Nunzio, G.M., Ferro, N.: Appendix A. Results of the Core Tracks and Domain-Specific Tracks. In: C. Peters, V. Quochi (eds.) Working Notes for the CLEF 2005 Workshop. http://www.clef-campaign.org/2005/working_notes/workingnotes2005/appendix_a.pdf [last visited 2007, March 23] (2005)
4. Di Nunzio, G.M., Ferro, N.: Appendix A: Results of the Ad-hoc Bilingual and Monolingual Tasks. In: A. Nardi, C. Peters, J.L. Vicedo (eds.) Working Notes for the CLEF 2006 Workshop. http://www.clef-campaign.org/2006/working_notes/workingnotes2006/Appendix_Ad-Hoc.pdf [last visited 2007, March 23] (2006)
5. Di Nunzio, G.M., Ferro, N., Jones, G.J.F., Peters, C.: CLEF 2005: Ad Hoc Track Overview. In: C. Peters, F.C. Gey, J. Gonzalo, G.J.F. Jones, M. Kluck, B. Magnini, H. Müller, M. de Rijke (eds.) Accessing Multilingual Information Repositories: Sixth Workshop of the Cross–Language Evaluation Forum (CLEF 2005). Revised Selected Papers, pp. 11–36. Lecture Notes in Computer Science (LNCS) 4022, Springer, Heidelberg, Germany (2006)
6. Di Nunzio, G.M., Ferro, N., Mandl, T., Peters, C.: CLEF 2006: Ad Hoc Track Overview. In: A. Nardi, C. Peters, J.L. Vicedo (eds.) Working Notes for the CLEF 2006 Workshop. http://www.clef-campaign.org/2006/working_notes/workingnotes2006/dinunzio0CLEF2006.pdf [last visited 2007, March 23] (2006)

7. European Commission: Commission Recommendation of 24 August 2006 on the digitisation and online accessibility of cultural material and digital preservation. Official Journal of the European Union, OJ L 236, 31.8.2006 **49**, 28–30 (2006)
8. European Commission Information Society and Media: i2010: Digital Libraries. `http://europa.eu.int/information_society/activities/digital_libraries/doc/brochures/dl_brochure_2006.pdf` [last visited 2007, March 23] (2006)
9. Gonzalo, J., Peters, C.: The Impact of Evaluation on Multilingual Text Retrieval. In: R. Baeza-Yates, N. Ziviani, G. Marchionini, A. Moffat, J. Tait (eds.) Proc. 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2005), pp. 603–604. ACM Press, New York, USA (2005)
10. Hull, D.: Using Statistical Testing in the Evaluation of Retrieval Experiments. In: R. Korfhage, E. Rasmussen, P. Willett (eds.) Proc. 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1993), pp. 329–338. ACM Press, New York, USA (1993)
11. Judge, G.G., Hill, R.C., Griffiths, W.E., Lütkepohl, H., Lee, T.C.: Introduction to the Theory and Practice of Econometrics, 2nd edn. John Wiley and Sons, New York, USA (1988)
12. Peters, C.: Multilingual Information Access for Digital Libraries: The Impact of Evaluation on System Development. In: C. Thanos (ed.) DELOS Research Activities 2006, pp. 105–107. ISTI-CNR at Gruppo ALI, Pisa, Italy (2006)
13. Tague-Sutcliffe, J.: The Pragmatics of Information Retrieval Experimentation, Revisited. In: K. Spack Jones, P. Willett (eds.) Readings in Information Retrieval, pp. 205–216. Morgan Kaufmann Publisher, Inc., San Francisco, California, USA (1997)
14. Wang, J., Oard, D.W.: Combining Bidirectional Translation and Synonymy for Cross-Language Information Retrieval. In: E.N. Efthimiadis, S. Dumais, D. Hawking, K. Järvelin (eds.) Proc. 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2006), pp. 202–209. ACM Press, New York, USA (2006)