CLEF 15th Birthday: What Can We Learn From Ad Hoc Retrieval?

Nicola Ferro and Gianmaria Silvello

University of Padua, Italy
{ferro,silvello}@dei.unipd.it

Abstract. This paper reports the outcomes of a longitudinal study on the CLEF Ad Hoc track in order to assess its impact on the effectiveness of monolingual, bilingual and multilingual information access and retrieval systems. Monolingual retrieval shows a positive trend, even if the performance increase is not always steady from year to year; bilingual retrieval has demonstrated higher improvements in recent years, probably due to the better linguistic resources now available; and, multilingual retrieval exhibits constant improvement and performances comparable to bilingual (and, sometimes, even monolingual) ones.

1 Motivations and Approach

Experimental evaluation has been a key driver for research and innovation in the information retrieval field since its inception. Large-scale evaluation campaigns such as Text REtrieval Conference (TREC)[1], Conference and Labs of Evaluation Forum (CLEF)[2], NII Testbeds and Community for Information access Research (NTCIR)[3], and Forum for Information Retrieval Evaluation (FIRE)[4] are known to act as catalysts for research by offering carefully designed evaluation tasks for different domains and use cases and, over the years, to have provided both qualitative and quantitative evidence about which algorithms, techniques and approaches are most effective. In addition, the evaluation campaigns have played a key role in the development of researcher and developer communities with multidisciplinary competences as well as in the development of linguistic resources and information retrieval systems.

As a consequence, some attempts have been made to determine their impact. For example, in 2010 an assessment of the economic impact of TREC pointed out that "for every \$1 that NIST and its partners invested in TREC, at least \$3.35 to \$5.07 in benefits accrued to IR researchers. The internal rate of return (IRR) was estimated to be over 250% for extrapolated benefits and over 130% for unextrapolated benefits" [11, p. ES-9]. The bibliometric impact and its effect on scientific production and literature has been studied both for TRECVid [17] and CLEF [18,19], showing how influential evaluation campaigns are.

[1] http://trec.nist.gov/

[2] http://www.clef-initiative.eu/

[3] http://research.nii.ac.jp/ntcir/

[4] http://www.isical.ac.in/~fire/

E. Kanoulas et al. (Eds.): CLEF 2014, LNCS 8685, pp. 31–43, 2014.

However, in the literature there have been few systematic longitudinal studies about the impact of evaluation campaigns on the overall effectiveness of Information Retrieval (IR) systems. One of the most relevant works compared the performances of eight versions of the SMART system on eight different TREC ad-hoc tasks (i.e. TREC-1 to TREC-8) and showed that the performances of the SMART system has doubled in eight years [5]. On the other hand, these results "are only conclusive for the SMART system itself" [20] and this experiment is not easy to reproduce in the CLEF context because we would need to use different versions of one or more systems – e.g. a monolingual, a bilingual and a multilingual system – and to test them on many collections for a great number of tasks. Furthermore, today's systems increasingly rely on-line linguistic resources (e.g. MT systems, Wikipedia, on-line dictionaries) which continuously change over time, thus preventing comparable longitudinal studies even when using the same systems.

Therefore, the goal of this paper is to carry out a longitudinal study on the Ad-Hoc track of CLEF in order to understand its impact on monolingual, bilingual, and multilingual retrieval.

To this end, we adopt the score standardization methodology proposed in [20] which allows us to carry out inter-collection comparison between systems by limiting the effect of collections (i.e. corpora of documents, topics and relevance judgments) and by making system scores interpretable in themselves. Standardization directly adjusts topic scores by the observed mean score and standard deviation for that topic in a sample of the systems. Let us say that topic t has mean $\mu_t = \bar{M}_{*t}$ and standard deviation $\sigma_t = sd(\bar{M}_{*t})$ for a given measure over a sample of systems and that system s receives a score m_{st} for that topic. Then, the standardized score m'_{st} (i.e. the *z-score* of m_{st}) is:

$$m'_{st} = \frac{m_{st} - \mu_t}{\sigma_t} \tag{1.1}$$

The z-score is directly informative in a way the unstandardized score is not: "one can tell directly from a runs score whether the system has performed well for the topic" [20]. Given that standardized scores are centered around zero and unbounded, whereas the majority of IR measures are in the interval $[0, 1]$, we map z-scores in this range by adopting the cumulative density function of the standard normal distribution; this also has the effect of reducing the influence of outlier data points:

$$F_X(m') = \int_{-\infty}^{m'} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx \tag{1.2}$$

For this study we apply standardization to Average Precision (AP) calculated for all the runs submitted to the ad-hoc tracks of CLEF (i.e. monolingual, bilingual and multilingual tasks from 2000 to 2007) and to The European Library (TEL) tracks (i.e. monolingual and bilingual tasks from 2008 to 2009). In order to use reliable standardization factors we do not consider the tasks for which less than 9 valid runs have been submitted; we consider a run as valid if it retrieves documents for each topic of the collection. In the following we indicate with sMAP the mean of the standardized AP.

All the CLEF results that we analysed in this paper are available through the Distributed Information Retrieval Evaluation Campaign Tool (DIRECT) system[5] [2]; the software library (i.e. MATTERS) used for calculating measure standardization as well as for analysing the performances of the systems is publicly available at the URL: `http://matters.dei.unipd.it/`.

The paper is organized as follows: Section 2 introduces the research questions we are investigating and provides a very short summary of the main findings for each of them; Section 3 reports the outcomes of our analyses and detailed answers to the research questions; finally, Section 4 outlines possible future directions for continuing these kinds of studies.

2 Research Questions

In this section we summarize the four research questions we tackle in this paper by reporting a brief insight of our findings.

RQ1. Do performances of monolingual systems increase over the years? Are more recent systems better than older ones?

From the analysis of sMAP across monolingual tasks we can see an improvement of performances, even if it is not always steady from year to year. The best systems are rarely the most recent ones; this may be due to a tendency towards tuning well performing systems relying on established techniques in the early years of a task while focusing on understanding and experimenting new techniques and methodologies in later years. In general, the assumption for which the life of a task is summarized by increase in system performances, plateau and termination oversimplifies reality: researchers and developers an not just incrementally adding new pieces on existing algorithms, rather they often explore completely new ways or add new components to the systems, causing a temporary drop in performances. Thus, we do not have a steady increase but rather a general positive trend.

RQ2. Do performances of bilingual systems increase over the years and what is the impact of source languages?

System performances in bilingual tasks show a growing trend across the years although it is not always steady and it depends on the number of submitted runs as well as on the number of newcomers. The best systems for bilingual tasks are often the more recent ones showing the importance of advanced linguistic resources that become available and improved over the years. Source languages have a high impact on the performances of a given target language, showing that some combinations are better performing than others – e.g. Spanish to Portuguese has a higher median sMAP than German to Portuguese.

[5] `http://direct.dei.unipd.it/`

RQ3. Do performances of multilingual systems increase over the years?
Multilingual systems show a steady growing trend of performances over the years despite the variations in target and source languages from task to task.

RQ4. Do monolingual systems have better performances than bilingual and multilingual systems?
Systems which operate on monolingual tasks prove to be more performing than bilingual ones in most cases, even if the difference between top monolingual and top bilingual systems reduces year after year and sometimes the ratio is even inverted. In some cases, multilingual systems turn out to have higher performances than bilingual ones and the top multilingual system has the highest sMAP of all the systems which participated in CLEF tasks from 2000 to 2009: the work done for dealing with the complexity of multilingual tasks pays off in terms of overall performances of the multilingual systems.

3 Experimental Analysis

RQ1. Do performances of monolingual systems increase over the years? Are more recent systems better than older ones?

With regard to monolingual tasks, there is no clear trend showing a steady improvement of sMAP over the years – see Figure 1 and Table 1. Figure 1 reports the median sMAP for all the monolingual tasks of CLEF for which more than nine valid runs were submitted; we can see that a more evident improvement over the years is shown by the languages introduced in 2004 [6] and 2005 [7]: Bulgarian, Hungarian and Portuguese – see Figure 2 where the median for Portuguese and Hungarian of the last year is higher than in the first year of the tasks. We can see that both for Portuguese and Hungarian the distribution of scores spreads out overs the years as far as the number of submitted runs and newcomers increase; on the other hand, the best system for Hungarian participated in the last year this task was performed (2007), whereas the best system for Portuguese participated in the first year of the task (2004) and it was outperformed afterwards.

The same trend is clear for French and German in the TEL tasks showing that monolingual retrieval in these languages over bibliographical records improved from 2008 [1] to 2009 [8] – see also Table 1. Note that for the TEL monolingual tasks the median increased over the years, whereas the best system participated, for both the languages, in the first year of the task. Furthermore, both for French and German, the best system for the ad-hoc tasks outperforms the best system for the TEL ones (i.e. 0.8309 versus 0.7388 for German and 0.8257 versus 0.7242 for French).

By contrast, examining the median sMAP of the monolingual tasks from 2000 to 2009 shows several examples of languages for which performances decrease – e.g. Dutch, Spanish and Italian. A closer analysis shows that for these languages the number of research groups along with the number of newcomers participating

Table 1. Statistics of the CLEF bilingual tasks started in 2000 or 2001

Task	Year	Groups(new)	Runs	Best sMAP	Median sMAP
AH Bili DE	2002	6(-)	13	.6674 (-)	**.5340** (-)
TEL Bili DE	2008	6(4)	17	.6268 (-6,08%)	.4599 (-13.88%)
	2009	6(3)	26	**.7179** (14.53%)	.4731 (+2.87%)
AH Bili EN	2000	10(-)	26	.7463 (-)	.5196 (-)
	2001	19(15)	55	.7725 (+3.51%)	.5618 (+8.12%)
	2002	5(3)	16	.6983 (-9.60%)	.4524 (-19.47%)
	2003	3(3)	15	.6980 (-0.04%)	.4074 (-9.95%)
	2004	4(4)	11	.5895 (-15.54%)	.5251 (+28.89%)
	2005	8(8)	31	**.7845** (+33.08%)	**.5667** (+7.92%)
	2006	5(4)	32	.7559 (-3.64%)	.4808 (-15.16%)
	2007	10(9)	67	.7746 (+2.47%)	.4835 (0.56%)
TEL Bili EN	2008	8(7)	24	.7611 (-1,74%)	.5382 (+11.31%)
	2009	10(7)	43	.7808 (2.59%)	.4719 (-12.32%)
AH Bili ES	2002	7(-)	16	**.6805** (-)	.4969 (-)
	2003	9(7)	15	.6737 (-1.01%)	**.5394** (+8.55%)
AH Bili FR	2002	7(-)	14	.6708 (-)	.5647 (-)
	2004	7(5)	24	.6015 (-10.33%)	.5211 (-7.72%)
	2005	9(8)	31	**.7250** (+20.53%)	**.5703** (+9.44%)
	2006	4(3)	12	.6273 (-13.47%)	.4886 (-14.33%)
TEL Bili FR	2008	5(5)	15	.6358 (+1,35%)	.4422 (-9.50%)
	2009	6(4)	23	.7151 (+12.47%)	.4355 (-1.52%)
AH Bili IT	2002	6(-)	13	.5916 (-)	.5306 (-)
	2003	8(5)	21	**.7119** (+20.34%)	**.5309** (+0.05%)
AH Bili PT	2004	4(-)	15	.6721 (-)	.4278 (-)
	2005	8(5)	24	**.7239** (+7.71%)	**.5020** (+17.34%)
	2006	6(4)	22	.6539 (-9.67%)	.4804 (-4.30%)
AH Bili RU	2003	2(-)	9	**.6894** (-)	.4810 (-)
	2004	8(7)	26	.6336 (-8.09%)	**.5203** (+8.17%)

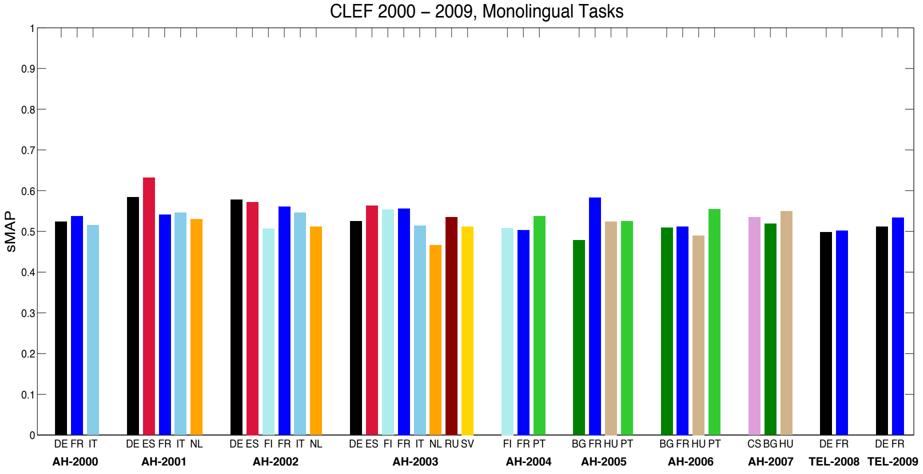


Fig. 1. Median sMAP of the CLEF monolingual tasks 2000-2009

in the tasks as well as the number of submitted runs increased over the years by introducing a high degree of variability in the performances.

The analysis of best sMAP tells us something different from the analysis of median sMAP. As an example, for the Dutch language, while the median decreases every year, the best sMAP increases showing an advancement of retrieval methods applied to this language. Also for the Italian task we can observe an improvement of best sMAP over the years given that the top systems show a

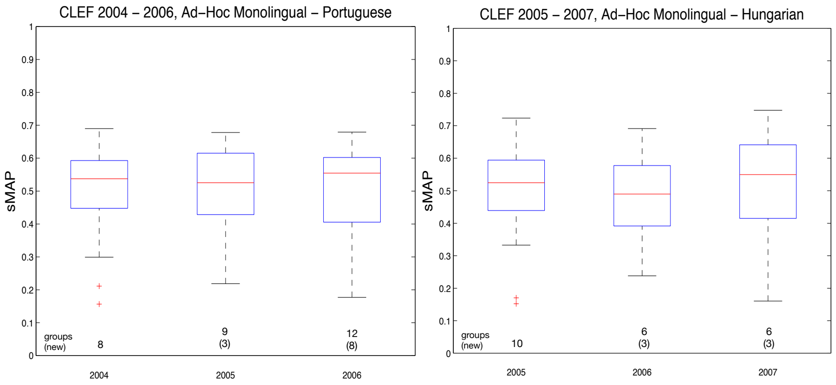


Fig. 2. Monolingual Portuguese and Hungarian Tasks Performance Breakdown

big improvement from 2000 to 2001 and then a plateau until 2003. Indeed, the best system (i.e. University of Neuchâtel [12]) in 2001 has a sMAP only 1.51% higher than the sMAP of the best system in 2002 (i.e. the PROSIT system [3] of Fondazione Ugo Bordoni) showing that the big improvement from 2000 (i.e. +22.13%) is due to a consistent advancement of retrieval techniques applied to the Italian language. In 2003 there was a 7.79% drop in sMAP for the best system with respect to the previous year; in 2003 the best system is still the one of Fondazione Ugo Bordoni, but with some differences from the system used in 2002 [4]: in 2002 they used the full enhanced PROSIT system with $B_E L2$ weighting schema, bigrams and coordination level matching, furthermore they focused only on the title of the queries and used a simple form of stemmer; in 2003 they used the same weighting schema, but focused on title plus description fields of the topics and used the Porter stemmer. From this analysis we can see that a more advanced stemmer did not improve the performances that also seem to be influenced by the topic fields considered; on the other hand, it is relevant to highlight that in 2003 the goal of this research group was to test different weighting schema in order to establish the best performing one [4], whereas in 2002 their aim was to test a fully enhanced retrieval system. This could also explain the drop in the median sMAP in 2003 with respect to 2002; in 2003 research groups that participated in previous years (i.e. ∼70%) might have been more interested in testing new techniques and retrieval settings rather than tuning already well performing systems for achieving slightly better performances. In general, this could explain why best performances are rarely achieved in the last year of a task, but one or two years before its termination; similar examples are the French and Spanish monolingual tasks.

This hypothesis is also corroborated by the best performances analysis, where we can see how in the first years of a task research groups dedicated much effort to tuning and enhancing good systems already tested in previous campaigns. The top system of all CLEF monolingual tasks is the Berkeley one [9] (i.e. 0.8309 sMAP) which participated in the German task in 2000, closely followed by the

Table 2. Statistics of the CLEF monolingual tasks started in 2000 or 2001

Task	Year	Groups(new)	Runs	Best sMAP	Median sMAP
AH Mono ES	2001	10(-)	22	.7402 (-)	**.6321** (-)
	2002	13(5)	28	**.8065** (+8.22%)	.5723 (-9.46%)
	2003	16(8)	38	.7016 (-14.95%)	.5630 (-1.62)
AH Mono DE	2000	11(-)	13	**.8309** (-)	.5235 (-)
	2001	12(9)	24	.6857 (-17.47%)	**.5839** (+11.53%)
	2002	12(5)	20	.6888 (+0.45%)	.5780 (-1.01%)
	2003	13(7)	29	.7330 (+6.42%)	.5254 (-9.10%)
TEL Mono DE	2008	10(7)	27	.7388 (+0.79%)	.4985 (-5.11%)
	2009	9(4)	34	.6493 (-12.11%)	.5123 (+2.76%)
AH Mono FR	2000	9(-)	10	.6952 (-)	.5370 (-)
	2001	9(6)	15	.6908 (-0.63%)	.5412 (+0.78%)
	2002	12(7)	16	**.8257** (+19.53%)	.5609 (+3.64%)
	2003	16(9)	35	.6758 (-18.15%)	.5565 (-0.78%)
	2004	13(4)	38	.6777 (+0.28%)	.5034 (-9.54%)
	2005	12(7)	38	.7176 (+5.89%)	**.5833** (+15.87%)
	2006	8(5)	27	.6992 (-2.56%)	.5120 (-12.22%)
TEL Mono FR	2008	9(8)	15	.7242 (+3.58%)	.5018 (-1.99%)
	2009	9(5)	23	.6838 (-5.58%)	.5334 (+6.30%)
AH Mono IT	2000	9(-)	10	.6114 (-)	.5150 (-)
	2001	8(5)	14	**.7467** (+22.13%)	**.5461** (+6.04%)
	2002	14(7)	25	.7354 (-1.51%)	**.5461** (-)
	2003	13(4)	27	.6796 (-7.59%)	.5142 (-5.84%)
AH Mono NL	2001	9(-)	18	.6844 (-)	**.5296** (-)
	2002	11(4)	19	.7128 (+4.15%)	.5118 (-3.36%)
	2003	11(4)	32	**.7231** (+1.45%)	.4657 (-10.53)

University of Neuchâtel system [13] (i.e. 0.8257 sMAP), which participated in the French task in 2002. The Berkeley system participated in several cross-lingual retrieval tasks in previous TREC campaigns; queries were manually formulated and expanded and the searcher spent about 10 to 25 minutes per topic [9]. We can see that this research group spent much time tuning an already good system by employing tested retrieval techniques enhanced with substantial manual intervention. Similarly, the Neuchâtel system is a careful improvement of techniques and methodologies introduced and tested in previous CLEF campaigns [13].

RQ2. Do performances of bilingual systems increase over the years and what is the impact of source languages?

For bilingual tasks we have to consider both the target language (i.e. the language of the corpus) and the source languages (i.e. the languages of the topics). In Figure 3 we show the median sMAP of the CLEF bilingual tasks divided by target language and on each bar we report the sources. As we can see, it is not always possible to identify a steady improvement of performances for a given target language over the years.

In Table 1 we report more detailed statistics about the bilingual tasks where we can see, unlike for the monolingual tasks, that the higher median sMAP as well as the best sMAP are achieved in the last years of each task. This is an indicator of the improvement of language resources – e.g. dictionaries, external resources like Wikipedia and the use of semantic rather than syntactic resources –

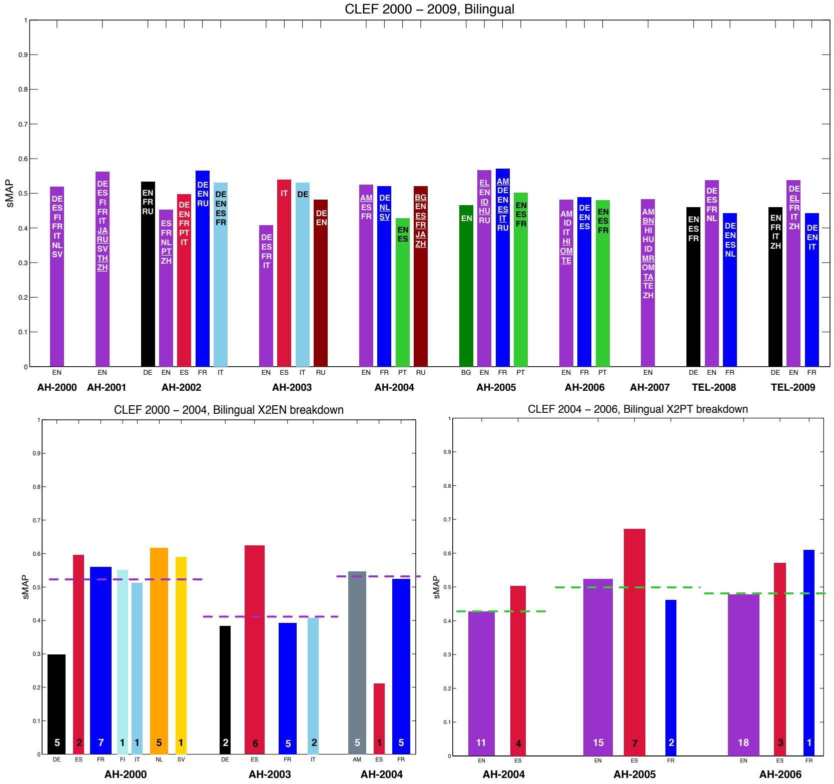
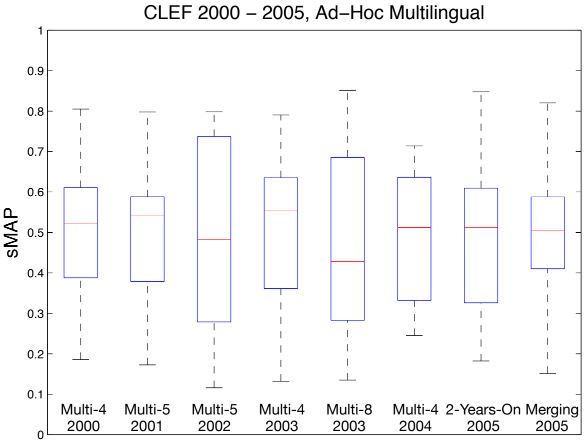


Fig. 3. Median sMAP of the CLEF bilingual tasks 2000-2009

that could be exploited by the bilingual systems. For instance, the best bilingual system for the "X2FR" task (i.e. University of Neuchâtel system [15], 0.7250 sMAP) exploited "seven different machine translation systems, three bilingual dictionaries" [15] and ten freely available translation tools; the best bilingual system in the TEL "X2DE" task (i.e. Chemnitz University of Technology [10], 0.7179 sMAP) exploited three out-the-box retrieval systems (i.e. Lucene, Lemur and Terrier) and the high quality of the Google translation service contributed substantially to achieving the final result [10].

The fluctuation of performances within the same task is due to the significant turnover of research groups and, more importantly, to the different source languages employed each year. In the lower part of Figure 3, we can see a performance breakdown for the "X2EN" and the "X2PT" tasks where we report the median sMAP achieved by the systems working on English and Portuguese target languages divided by the source language employed; inside each single bar we report the number of runs submitted for that source language and the thickness of each bar is weighted by this number. For "X2EN" we report data for the tasks carried out in 2000, 2003, and 2004; we can see that in 2003 the



	Source Languages	Target Languages	# Groups (new)	# Runs
Multi-4 2000	de, en, nl	de, en, fr, it	11(-)	26
Multi-5 2001	de, en, nl, ru, zh	de, en, es, fr, it	8(4)	26
Multi-5 2002	de, en	de, en, es, fr, it	11(6)	36
Multi-4 2003	de, en, es, fr	de, en, es, fr	14(10)	52
Multi-8 2003	en, es	de, en, es, fi, fr, it, nl, sv	7(3)	33
Multi-4 2004	en, fr	en, fi, fr, ru	9(6)	35
2-Years-On 2005	en, es, fr, nl	de, en, es, fi, fr, it, nl, sv	4(3)	21
Merging 2005	en	de, en, es, fi, fr, it, nl, sv	3(1)	20

Fig. 4. sMAP scores distribution for all the CLEF multilingual tasks

median sMAP dropped with respect to 2000 and then it recovered in 2004. In 2003, only 3 groups (all newcomers) participated by submitting fewer runs than in 2000; in 2004 the median sMAP recovered, even if there were still fewer groups (only 4 and all newcomers) than in 2000 and even fewer runs than in 2003. The main influence on performances came from the source languages used. In 2000, more than 50% of the runs used French, Spanish, Italian and Dutch languages and their performances were fairly good; the most difficult source language was German. In 2003 performances of runs using Spanish as source language further improved, but they dropped for French and Italian and showed little improvement for German. In 2004 the higher global sMAP is due to the improvement of French runs, the removal of German as source language and the introduction of Amharic for which very good runs were submitted even if this language was initiated that very year. For the "X2PT" task, we can see that global sMAP depends on the English source language for which there are more runs every year and that always performs worse than Spanish. This analysis shows that Spanish to Portuguese was always performed better than English to Portuguese; this could be due to the morphology of languages, given that Spanish and Portuguese are closer to each other than English and Portuguese; we cannot say much about French to Portuguese because there are a small number of available runs.

RQ3. Do performances of multilingual systems increase over the years?

In Figure 4 we show the boxplot of sMAP for each CLEF multilingual task from 2000 to 2005. We can identify a growing trend of performances especially for top systems. For instance, for multilingual task with four languages we can see a major improvement of median sMAP from 2002 to 2003 even if the top system of 2003 has lower sMAP than the one of 2002; at the same time, the multilingual task with 8 languages reports the lowest median sMAP and, at the same time, the best performing system of all multilingual tasks.

Standardization allows us to reconsider an important result reported in [7] while discussing the 2-Years-On task in which new systems (i.e. 2005 systems) operated on the 2003 multi-8 collection; the purpose was to compare the performances of 2003 systems with the 2005 ones on the same collection[6]. Di Nunzio et alii in [7] reported a 15.89% increase in performances for the top system of 2005 with respect to the top system of 2003; this finding showed an improvement of multilingual IR systems from 2003 to 2005. Nevertheless, analysing sMAP we draw a similar conclusion, but from a different perspective; indeed, the top system in 2003 achieved 0.8513 sMAP (i.e. University of Neuchâtel [14]), whereas the top system in 2005 achieved 0.8476 sMAP (i.e. Carnegie Mellon University [16]), reporting a 0.44% decrease in performances. On the other hand, the median sMAP in 2003 was 0.4277 and in 2005 it was 0.5117 thus reporting an overall increase of 16.41%; this result is even stronger than the findings reported in [7], since it shows that half of the participating systems in 2005 improved with respect to 2003 ones.

RQ4. Do monolingual systems have better performances than bilingual and multilingual systems?

In Figure 5 we report the median sMAP and the best sMAP of the monolingual tasks compared to the bilingual tasks for the same target language. We can see that in most cases the median sMAP of the monolingual tasks overcome the median sMAP of the corresponding bilingual task with the exception of French in 2002 and 2004 and Italian in 2003. On the other hand, the best sMAP ratio between monolingual and bilingual tasks reports another viewpoint where the gap between top monolingual and top bilingual systems is progressively reduced across the years and in several cases the trend is inverted with bilingual systems performing better than monolingual ones.

In Table 3 we report aggregate statistics where we calculated the median, best and mean sMAP for all the systems which participated in the monolingual, bilingual and multilingual tasks.

We can see that bilingual and multilingual systems have a similar median and mean sMAP even though they are slightly higher for the multilingual and

[6] Note that the multi-8 collection had 60 topics, whereas in 2005 a subset of 40 topics was actually used by the systems; the 20 remaining were employed for training purposes [7].

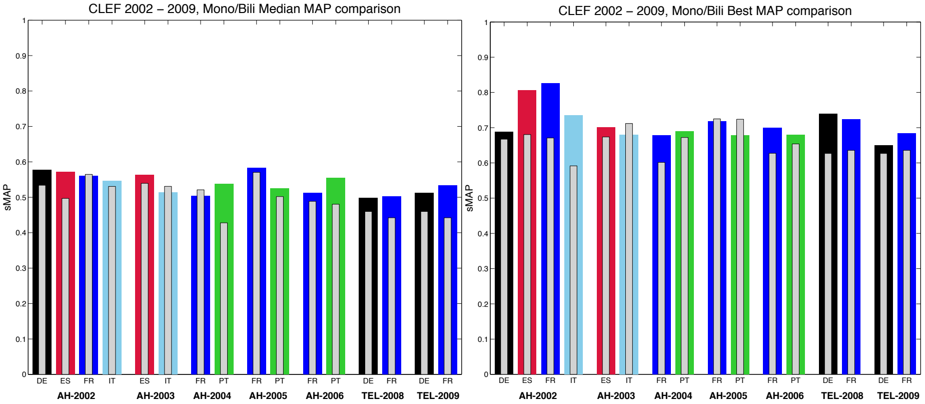


Fig. 5. Mono/Bili Median and Best sMAP comparison. The thick bars indicate monolingual tasks and thin bars bilingual tasks.

Table 3. Aggregate sMAP of mono, bili and multilingual CLEF ad-hoc and TEL tasks from 2000 to 2009

sMAP	**Monolingual**	**Bilingual**	**Multilingual**
Best	.8309	.7845	**.8513**
Median	**.5344**	.5165	.5173
Mean	**.5054**	.4898	.4914

both are exceeded by the monolingual systems. It is interesting to note that the best system is the multilingual one that has a sMAP 8.52% higher than the top bilingual and 2.46% higher than the top monolingual system.

4 Future Works

This study opens up diverse analysis possibilities and as future works we plan to investigate several further aspects regarding the cross-lingual evaluation activities carried out by CLEF; we will: (i) apply standardization to other largely-adopted IR measures – e.g. Precision at 10, RPrec, Rank-Biased Precision, bpref – with the aim of analysing system performances from different perspectives; (ii) aggregate and analyse the systems on the basis of adopted retrieval techniques to better understand their impact on overall performances across the years; and (iii) extend the analysis of bilingual and multilingual systems grouping them on a source and target language basis thus getting more insights into the role of language morphology and linguistic resources in cross-lingual IR.

References

1. Agirre, E., Di Nunzio, G.M., Ferro, N., Mandl, T., Peters, C.: CLEF 2008: Ad Hoc Track Overview. In: Peters, C., et al. (eds.) CLEF 2008. LNCS, vol. 5706, pp. 15–37. Springer, Heidelberg (2009)
2. Agosti, M., Di Buccio, E., Ferro, N., Masiero, I., Peruzzo, S., Silvello, G.: DIRECTions: Design and Specification of an IR Evaluation Infrastructure. In: Catarci, T., Forner, P., Hiemstra, D., Peñas, A., Santucci, G. (eds.) CLEF 2012. LNCS, vol. 7488, pp. 88–99. Springer, Heidelberg (2012)
3. Amati, G., Carpineto, C., Romano, G.: Italian Monolingual Information Retrieval with PROSIT. In: Peters, C., Braschler, M., Gonzalo, J. (eds.) CLEF 2002. LNCS, vol. 2785, pp. 257–264. Springer, Heidelberg (2003)
4. Amati, G., Carpineto, C., Romano, G.: Comparing Weighting Models for Monolingual Information Retrieval. In: Peters, C., Gonzalo, J., Braschler, M., Kluck, M. (eds.) CLEF 2003. LNCS, vol. 3237, pp. 310–318. Springer, Heidelberg (2004)
5. Buckley, C.: The SMART project at TREC. In: TREC — Experiment and Evaluation in Information Retrieval, pp. 301—320. MIT Press (2005)
6. Braschler, M., Di Nunzio, G.M., Ferro, N., Peters, C.: CLEF 2004: Ad Hoc Track Overview and Results Analysis. In: Peters, C., Clough, P., Gonzalo, J., Jones, G.J.F., Kluck, M., Magnini, B. (eds.) CLEF 2004. LNCS, vol. 3491, pp. 10–26. Springer, Heidelberg (2005)
7. Di Nunzio, G.M., Ferro, N., Jones, G.J.F., Peters, C.: CLEF 2005: Ad Hoc Track Overview. In: Peters, C., et al. (eds.) CLEF 2005. LNCS, vol. 4022, pp. 11–36. Springer, Heidelberg (2006)
8. Ferro, N., Peters, C.: CLEF 2009 Ad Hoc Track Overview: TEL and Persian Tasks. In: Peters, C., Di Nunzio, G.M., Kurimo, M., Mandl, T., Mostefa, D., Peñas, A., Roda, G. (eds.) CLEF 2009. LNCS, vol. 6241, pp. 13–35. Springer, Heidelberg (2010)
9. Gey, F.C., Jiang, H., Petras, V., Chen, A.: Cross-Language Retrieval for the CLEF Collections - Comparing Multiple Methods of Retrieval. In: Peters, C. (ed.) CLEF 2000. LNCS, vol. 2069, pp. 116–128. Springer, Heidelberg (2001)
10. Kürsten, J.: Chemnitz at CLEF 2009 Ad-Hoc TEL Task: Combining Different Retrieval Models and Addressing the Multilinguality. In: CLEF 2009 Working Notes, on-line
11. Rowe, B.R., Wood, D.W., Link, A.L., Simoni, D.A.: Economic Impact Assessment of NIST's Text REtrieval Conference (TREC) Program. RTI International, USA (2010)
12. Savoy, J.: Report on CLEF-2001 Experiments: Effective Combined Query-Translation Approach. In: Peters, C., Braschler, M., Gonzalo, J., Kluck, M. (eds.) CLEF 2001. LNCS, vol. 2406, pp. 27–43. Springer, Heidelberg (2002)
13. Savoy, J.: Report on CLEF 2002 Experiments: Combining Multiple Sources of Evidence. In: Peters, C., Braschler, M., Gonzalo, J. (eds.) CLEF 2002. LNCS, vol. 2785, pp. 66–90. Springer, Heidelberg (2003)
14. Savoy, J.: Report on CLEF-2003 Multilingual Tracks. In: Peters, C., Gonzalo, J., Braschler, M., Kluck, M. (eds.) CLEF 2003. LNCS, vol. 3237, pp. 64–73. Springer, Heidelberg (2004)
15. Savoy, J., Berger, P.-Y.: Monolingual, Bilingual, and GIRT Information Retrieval at CLEF-2005. In: Peters, C., et al. (eds.) CLEF 2005. LNCS, vol. 4022, pp. 131–140. Springer, Heidelberg (2006)

16. Si, L., Callan, J.: CLEF 2005: Multilingual Retrieval by Combining Multiple Multilingual Ranked Lists. In: Peters, C., et al. (eds.) CLEF 2005. LNCS, vol. 4022, pp. 121–130. Springer, Heidelberg (2006)
17. Thornley, C.V., Johnson, A.C., Smeaton, A.F., Lee, H.: The Scholarly Impact of TRECVid (2003–2009). JASIST 62(4), 613–627 (2011)
18. Tsikrika, T., de Herrera, A.G.S., Müller, H.: Assessing the Scholarly Impact of ImageCLEF. In: Forner, P., Gonzalo, J., Kekäläinen, J., Lalmas, M., de Rijke, M. (eds.) CLEF 2011. LNCS, vol. 6941, pp. 95–106. Springer, Heidelberg (2011)
19. Tsikrika, T., Larsen, B., Müller, H., Endrullis, S., Rahm, E.: The Scholarly Impact of CLEF (2000–2009). In: Forner, P., Müller, H., Paredes, R., Rosso, P., Stein, B. (eds.) CLEF 2013. LNCS, vol. 8138, pp. 1–12. Springer, Heidelberg (2013)
20. Webber, W., Moffat, A., Zobel, J.: Score standardization for inter-collection comparison of retrieval systems. In: SIGIR 2008, pp. 51–58. ACM Press (2008)