CLEF 2000 – 2014: Lessons Learnt from Ad Hoc Retrieval (Extended Abstract)*

Nicola Ferro and Gianmaria Silvello

Department of Information Engineering – University of Padua
{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 in the last fifteen years on the effectiveness of monolingual, bilingual and multilingual information access and retrieval systems.

1 Motivations and Approach

Experimental evaluation has been a key driver for research and innovation in the *Information Retrieval (IR)* field since its inception. Large-scale evaluation campaigns such as *Conference and Labs of Evaluation Forum (CLEF)*[1], 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.

As a consequence, some attempts have been made to determine their impact [4], however, in the literature there have been few systematic longitudinal studies about the impact of evaluation campaigns on the overall effectiveness of IR systems. One of the most relevant works compared the performances of eight versions of the SMART system on eight different *Text REtrieval Conference (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 [2]. On the other hand, these results "are only conclusive for the SMART system itself" [5] 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 online 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, we carry out a longitudinal study on the Ad-Hoc track of CLEF in order to understand its impact on monolingual, bilingual, and multilingual

* The extended version of this abstract has been published in [3].

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

retrieval by adopting the *score standardization methodology* proposed in [5]. This methodology 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.

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).

All the CLEF results that we analysed in this paper are available through the *Distributed Information Retrieval Evaluation Campaign Tool (DIRECT)* system[2] [1]; 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/`.

In the following we report the main research questions we tackled and we provide a short summary of the main findings for each of them. More detailed experimental results concerning those research questions could be found in [3]. Finally, we conclude by outlining the work we envision for the future.

2 Research Questions

The longitudinal study we carried out was aimed at tackling four research questions for which we report 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 mean standardized AP (sMAP) across monolingual tasks we can see an improvement of performances, even though it is not always steady from year to year, see Table 1. 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 do 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

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

Table 1. 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)

Table 2. Aggregate sMAP of monolingual, bilingual 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

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.

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. We can identify a growing trend of performances especially for top systems. For instance the multilingual task with four languages reports a major improvement of median sMAP from 2002 to 2003 even though the top system of 2003 has lower sMAP than the one of 2002; the multilingual task with 8 languages reports the lowest median sMAP and, at the same time, the best performing system of all multilingual tasks.

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 though 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, as shown in Table 2, 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 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. M. Agosti, E. Di Buccio, N. Ferro, I. Masiero, S. Peruzzo, and G. Silvello. DIRECTions: Design and Specification of an IR Evaluation Infrastructure. In T. Catarci, P. Forner, D. Hiemstra, A. Peñas, and G. Santucci, editors, *Proc. of the 3rd Int. Conf. of the CLEF Initiative (CLEF 2012)*, pages 88–99. Lecture Notes in Computer Science (LNCS) 7488, Springer, Heidelberg, Germany, 2012.
2. C. Buckley. The SMART project at TREC. In *TREC — Experiment and Evaluation in Information Retrieval*, pages 301—320. MIT Press, 2005.
3. N. Ferro and G. Silvello. CLEF 15th Birthday: What Can We Learn From Ad Hoc Retrieval? In E. Kanoulas, M. Lupu, P. Clough, M. Sanderson, M. Hall, A. Hanbury, and E. Toms, eds, *Proc. of the 5th Int. Conf. of the CLEF Initiative (CLEF 2014)*, pages 31–43. Lecture Notes in Computer Science (LNCS) 8685, Springer, Heidelberg, Germany, 2014.
4. B. R. Rowe, D. W. Wood, A. L. Link, and D. A. Simoni. *Economic Impact Assessment of NIST's Text REtrieval Conference (TREC) Program.* RTI International, USA, 2010.
5. W. Webber, A. Moffat, and J. Zobel. Score standardization for inter-collection comparison of retrieval systems. In SIGIR 2008, pages 51–58. ACM Press, 2008.