Information Processing and Management 000 (2016) 1–28

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

3.5K runs, 5K topics, 3M assessments and 70M measures: What trends in 10 years of Adhoc-*ish* CLEF?

Nicola Ferro*, Gianmaria Silvello

Department of Information Engineering, University of Padua., Via Gradenigo 6/B, 35131, Padova, Italy

ARTICLE INFO

Article history:
Available online xxx

Keywords:
Information retrieval
Multilingual information access
Longitudinal analysis
Experimental evaluation
CLEF

ABSTRACT

Multilingual information access and retrieval is a key concern in today global society and, despite the considerable achievements over the past years, it still presents many challenges. In this context, experimental evaluation represents a key driver of innovation and multilinguality is tackled in several evaluation initiatives worldwide, such as CLEF in Europe, NTCIR in Japan and Asia, and FIRE in India. All these activities have run several evaluation cycles and there is a general consensus about their strong and positive impact on the development of multilingual information access systems. However, a systematic and quantitative assessment of the impact of evaluation initiatives on multilingual information access and retrieval over the long period is still missing.

Therefore, in this paper we conduct the first systematic and large-scale longitudinal study on several CLEF Adhoc-*ish* tasks – namely the Adhoc, Robust, TEL, and GeoCLEF labs – in order to gain insights on the performance trends of monolingual, bilingual and multilingual information access systems, spanning several European and non-European languages, over a range of 10 years.

We learned that monolingual retrieval exhibits a stable positive trend for many of the languages analyzed, even though the performance increase is not always steady from year to year due to the varying interests of the participants, who may not always be focused on just increasing performances. Bilingual retrieval demonstrates higher improvements in recent years – probably due to the better language resources now available – and it also outperforms monolingual retrieval in several cases. Multilingual retrieval shows improvements over the years and performances are comparable to those of bilingual and monolingual retrieval, and sometimes even better. Moreover, we have found evidence that the rule-of-thumb of a 3-year duration for an evaluation task is typically enough since top performances are usually reached by the third year and sometimes even by the second year, which then leaves room for research groups to investigate relevant research issues other than top performances.

Overall, this study provides quantitative evidence that CLEF has achieved the objective which led to its establishment, i.e. making multilingual information access a reality for European languages. However, the outcomes of this paper not only indicate that CLEF has steered the community in the right direction, but they also highlight the many open challenges for multilinguality. For instance, multilingual technologies greatly depend on language resources and targeted evaluation cycles help not only in developing and improving them, but also in devising methodologies which are more and more language-independent.

* Corresponding author.
E-mail address: ferro@dei.unipd.it (N. Ferro).

http://dx.doi.org/10.1016/j.ipm.2016.08.001

Another key aspect concerns multimodality, intended not only as the capability of providing access to information in multiple media, but also as the ability of integrating access and retrieval over different media and languages in a way that best fits with user needs and tasks.

1. Introduction

MultiLingual Information Access (MLIA) is a prominent area of research concerned with the design and development of *Information Retrieval (IR)* systems able to seamlessly search and retrieve information in multiple languages (Nie, 2010; Peters, Braschler, & Clough, 2011). The growing number of users and content in multiple languages on the Web[1] (Montalvo, Martinez, Fresno, & Capilla, 2015), the raise of social media and online communities, where code-mixed languages[2] are more and more widespread (Raghavi, Chinnakotla, & Shrivastava, 2015) as well as the need of crossing the barriers between media and languages make MLIA a primary research challenge in the IR field.

IR is a research field strongly rooted in experimental and evaluation has been always representing a key driver of innovation in the field. Indeed, large-scale evaluation campaigns at the international level, such as the *Text REtrieval Conference (TREC)*[3] (Harman & Voorhees, 2005) in the United States since 1992, the *Conference and Labs of the Evaluation Forum (CLEF)*[4] (Braschler & Peters, 2004; Ferro, 2014) in Europe since 2000, the *NII Testbeds and Community for Information access Research (NTCIR)*[5] in Japan and Asia since 1999, and the *Forum for Information Retrieval Evaluation (FIRE)*[6] in India since 2008, have been fostering research and innovation in the IR field for decades. They have done this by providing both evidence about which models, algorithms, techniques and solutions have been performing best and by producing huge amounts of experimental data (Ferro, Hanbury, Müller, & Santucci, 2011), which represent an extremely valuable asset for past, current, and future research.

An open question for MLIA is to understand and assess the impact that large-scale evaluation campaigns have had on its development and to have quantitative evidence about it. This is a crucial question both to understand whether the field has evolved in a positive way and, learning from this, to envision strategic directions which will shape the future evolution of the area.

To this end, one approach is to look at the scholarly and scientific impact as it has been done for TRECVid (Thornley, Johnson, Smeaton, & Lee, 2011) and CLEF (Angelini et al., 2014; Tsikrika, Garcia Seco de Herrera, & Müller, 2011; Tsikrika, Larsen, Müller, Endrullis, & Rahm, 2013). Another approach is to estimate the economic impact as done in the case of TREC where "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 was estimated to be over 250% for extrapolated benefits and over 130% for unextrapolated benefits" (Rowe, Wood, Link, & Simoni, 2010).

Finally, another relevant means for understanding the impact of an evaluation campaign is to conduct a longitudinal study over different editions of the campaign in order to mine, analyze, and interpret system performances over time. This kind of study is often hard to carry out because of, on the one hand, the huge (and sometimes sparse) amount of experimental data to process and, on the other hand, the inherent difficulty in making comparisons over years due to intrinsic differences in the collections. As a consequence, there is a lack of systematic longitudinal investigations and, to the best of our knowledge, there have only been two limited attempts so far, both concerning TREC (Armstrong, Moffat, Webber, & Zobel, 2009a; Buckley, 2005) while nothing has been done with a specific focus on MLIA.

This paper carries out the first extensive longitudinal study on the "CLEF Classic" period (Ferro, 2014), i.e. the first ten years of CLEF since its establishment in 2000, where main stream research has been carried out on multilingual information access by considering different kinds of Adhoc-*ish* search tasks, namely the Adhoc, Robust, TEL, and GeoCLEF labs. In this period, different angles of monolingual, bilingual and multilingual information access have been explored for many European and non-European languages. The objective of the paper is to observe performances over different years and in several languages in order to gain an understanding of the impact of CLEF in the development of MLIA systems.

The paper aims at "letting the data speak" by reporting different trends/phenomena which we have observed over the different editions of a task – for example, improvement in the best or median performances, changes in the performances due to language resources, impact of experienced and less experienced participating groups, and so on – and which help us in getting an appreciation of CLEF overall influence.

[1] http://www.internetworldstats.com/stats7.htm

[2] Code-mixed languages are the embedding of linguistic units such as phrases, words, and morphemes of one language into an utterance of another language.

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

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

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

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

Please cite this article as: N. Ferro, G. Silvello, 3.5K runs, 5K topics, 3M assessments and 70M measures: What trends in 10 years of Adhoc–*ish* CLEF? Information Processing and Management (2016), http://dx.doi.org/10.1016/j.ipm.2016.08.001

In particular, the main areas where CLEF may have had an impact are monolingual, bilingual and multilingual information access. Therefore, we investigate the following specific research questions:

RQ1 What performance trends can we observe for monolingual systems over the years? What is the influence of language resources?
RQ2 What performance trends can we observe for bilingual systems over the years? What is the influence of source languages?
RQ3 What performance trends can we observe for multilingual systems over the years?
RQ4 What is the relationship between the performances of monolingual systems and those of bilingual systems?
RQ5 Is the typical 3-year duration of an evaluation task enough to improve the participating systems?

We rely on score standardization (Webber, Moffat, & Zobel, 2008) to normalize scores within each edition of a task and to set different editions side-by-side. We then study both best and median performances over the years and describe what happened in the different editions of a task in order to gather evidence for answering the above research questions and to get an idea of the heritage of CLEF and an outlook of future directions in MLIA.

The paper is organized as follows: Section 2 introduces the methodology adopted for carrying out the longitudinal study, discusses its limitations and compares it to other possible approaches; Section 3 describes the experimental setup; Sections 4–7 report the outcomes of our analyses and detailed answers to the research questions; finally, Section 8 draws conclusions and wraps up the discussion on the main findings of the study.

2. Methodology

2.1. Longitudinal studies

Longitudinal studies are typically intended in two main ways: either as distilling lessons learned over the years or as a comparison and analysis of performances over the years.

Best practices and lessons learned about which approaches are better for a given task or which language resources are most appropriate are already available both for TREC (Harman & Voorhees, 2005) and CLEF (Braschler et al., 2012; Peters et al., 2011), not to mention notable examples beyond them, such as Robertson and Spärck Jones (1994) and Spärck Jones (1981). This approach is outside of the scope of the present paper, since we already have such kinds of studies for CLEF.

The other approach is to study performances over the years but this is difficult because results on one collection are not directly comparable to results on another collection due to differences in topics, documents, and their interaction with the tested systems.

Buckley (2005) performed a study on the SMART systems for eight different TREC Adhoc tasks (TREC-1 to TREC-8) by freezing eight different versions of SMART and running each of them on each edition of the Adhoc task. In this way, the same system has been tested over all the collections, thus ensuring comparability of the results. However, these results "are only conclusive for the SMART system itself" (Webber et al., 2008) and they are not representative of a whole evaluation campaign and the different participating groups.

Moreover, the approach by Buckley (2005) is difficult, if not impossible, to replicate and apply to whole evaluation campaigns, such as CLEF. Indeed, we would need to have available all the different versions of the systems that participated in the evaluation campaigns over the years, which is almost impossible for practical reasons, and to test them on many collections for a great number of tasks, which is extremely demanding resource-wise. Furthermore, today MLIA systems are increasingly reliant upon online linguistic resources (e.g. online machine-translation services, Wikipedia, online dictionaries) which continuously change over time, thus preventing comparable longitudinal studies.

Therefore, we adopted the score standardization methodology (Webber et al., 2008) which is, to the best of our knowledge, the only other technique available for this purpose. Indeed, score standardization allows us to carry out inter-collection comparison between systems by limiting the effect of collections and by making system scores interpretable on their own.

This methodology was applied by Armstrong et al. (2009a) to perform a brief longitudinal study on TREC Adhoc data on six TREC Adhoc tasks (TREC-3 to TREC-8) plus three TREC Robust tasks (TREC-2003 to TREC-2005), with the conclusion that no clear improvement trend was found. In this study, Armstrong et al. (2009a) also used off-the-shelf open source IR systems (Lucene, Terrier, Indri, Zettair, and MG) to contrast them with TREC research prototypes; the authors found that they "have not captured the effectiveness achievements observed in the better historical TREC runs". This finding is consistent with a recent one on reproducible baselines (Lin et al., 2016), where the authors have run several off-the-shelf open source IR systems on different TREC collections and observed that historical TREC systems still outperform them.

However, the study by Armstrong et al. (2009a) was focused on finding a steady performance improvement from year to year. This research question may fit the context of TREC well, where English IR relies on consolidated linguistic resources – tokenizers, stop lists, stemmers and so on – which are widely available to the community. As a consequence, in a well-understood task like the Adhoc one, it is somehow natural for participating groups to mostly look for improving performances, e.g. with new retrieval models, finer tuning of the parameters and so on.

In the context of CLEF the situation can be more complex because it acted as a catalyst for creating MLIA systems in many European (and in some cases non European) languages, which might have been severely under-resourced (Ferro, 2014). The goal of the CLEF Adhoc-*ish* task is to improve MLIA system for European languages which not only transforms into increasing

Please cite this article as: N. Ferro, G. Silvello, 3.5K runs, 5K topics, 3M assessments and 70M measures: What trends in 10 years of Adhoc–*ish* CLEF? Information Processing and Management (2016), http://dx.doi.org/10.1016/j.ipm.2016.08.001

performances from year to year. Indeed, each task involves a community interested not only in achieving top performances but also in building language resources for a specific language or in experimenting with alternative ways for dealing with the complexity and corner-cases of a given language. This was also apparent form an initial study we carried out on some CLEF Adhoc data (Ferro & Silvello, 2014), where we learned that it is difficult to see steady performance improvements. Instead it is better to look for performance trends that inform us how MLIA systems have been affected by CLEF over the time.

Moreover, CLEF has also acted as a driving force behind new and local research communities which have learned how to deal with the specificity of their own language and with multilingualism more in general. Therefore, there is a not-negligible turn-over with new (and often inexperienced) groups joining every year and experienced ones leaving for other tasks, which obviously impacts the observed performances.

All these considerations motivate why the research questions laid out in the previous section are focused on observing performances and their phenomena rather than just seeking for steady performance improvements.

Finally, Armstrong, Moffat, Webber, and Zobel (2009b) used the score standardization methodology, although with a different goal from the one of the present paper. Instead of analyzing the trends within the different editions of an evaluation task, they used it to analyze and compare the results, which were built on the TREC Adhoc experimental collections and were reported in the literature at major venues such as SIGIR and CIKM, in order to highlight the problem of weak baselines. This has been further investigated very recently by Kharazmi, Scholer, Vallet, and Sanderson (2016), who show that a significant improvement over a weak baseline is not a predictor of a possible improvement against a strong baseline.

Note that longitudinal studies are seldom conducted because of both their difficulty in terms of technical challenges to carry them out and the effort needed to conduct them. As a consequence, it is more common to see reports that consolidate the main findings in one or two cycles of an evaluation activity, as for example Pal, Mitra, and Kamps (2011), Tang, Geva, Trotman, Xu, and Itakura (2014) or systematic studies which apply a given set of techniques over several collections, as for example Ferro and Silvello (2016a), Ferro and Silvello (2016b), Lin et al. (2016) and Tax, Bockting, and Hiemstra (2015). Since all these kinds of systematic studies are not longitudinal studies, they are out of the scope of the present paper.

As previously discussed, our goal is to observe how performances evolve and change over the years in order to assess the role of CLEF in the evolution of MLIA. Nevertheless, comparing editions of a task to understand the overall impact of an evaluation campaign is rather different from many of the just discussed previous studies, which basically compare systems, versions of a system over the years, systems against baselines, and so on. Indeed, we are interested in describing as one group of systems behaved with respect to another one rather than comparing two systems either in the same group or in different groups to understand whether there are significant differences between them.

2.2. Why score standardization is needed

It is known that it is not possible to safely compare raw scores of performance measures across tasks/collections because of intrinsic differences among tasks/collections. In the context of the CLEF Adhoc-*ish* tasks, like in the corresponding TREC ones, these intrinsic differences are mainly due to the topics, which vary from year to year, since the goal of the tasks is more or less stable over time and the used document collections are more or less the same or vary in an incremental way, as further discussed in Section 3.

System performances are typically broken down (Robertson & Kanoulas, 2012; Tague-Sutcliffe & Blustein, 1994) to a reasonable approximation as

$$\text{system performances} = \text{topic effect} + \text{system effect} + \text{topic/system interaction}$$

to which other effects might be added, such as a document collection effect, but they typically have a small impact on system performances. Moreover, the effect of the topics is typically much bigger than the one of the systems (Banks, Over, & Zhang, 1999; Tague-Sutcliffe & Blustein, 1994).

This break-down explains why it is not possible to compare raw scores directly from year to year. Indeed, what we would like to compare is the system effect but this is confounded by the topic and interaction effects.

As an example, let us consider a sample task run over two editions of a campaign: by inspecting raw scores we may see that the group of systems in edition 2 has better performances than the group in edition 1, but from this observation we cannot conclude that systems improved from year to year. Indeed, in the second year the effect of the topics might be much bigger than the one in the first year and/or they may have a more positive interaction with the systems, boosting their performances.

Similar considerations also hold if you run exactly the same system over different years. By inspecting table 13.1 of Buckley (2005), you can see that the performances of each version of the SMART system increase/decrease a lot across the different editions of the TREC Adhoc tasks and, since for each version the system effect is the same, all this variability is due to the topic and interaction effects.

The topic and interaction effects also affect the comparison between two different systems and it may produce different results on different years. Again from table 13.1 of Buckley (2005), you can observe that SMART-TREC-4 (version of SMART developed in TREC 4) performs better than SMART-TREC-8 in the case of the TREC-1, TREC-2, and TREC-3 editions of the Adhoc task while SMART-TREC-8 performs better than SMART-TREC-4 for TREC-4, TREC-5 and TREC-8 and they are very close in TREC-6 and TREC-7.

Please cite this article as: N. Ferro, G. Silvello, 3.5K runs, 5K topics, 3M assessments and 70M measures: What trends in 10 years of Adhoc–*ish* CLEF? Information Processing and Management (2016), http://dx.doi.org/10.1016/j.ipm.2016.08.001

Score standardization basically removes the topic effect and smoothes as much as possible the topic/system interaction effect, leaving us mainly with the system effect which is what we aim at comparing across years. Indeed, standardization directly adjusts topic scores by the observed mean score and standard deviation for that topic across the systems.

2.3. Score standardization

Consider an edition of a task and a matrix $M = \{m_{s_j t_k}\}$ whose elements represent the performance for a given measure m of system s_j on topic t_k. Let us say that topic t_k has mean $\mu_{t_k} = \overline{M}_{\cdot t_k}$ and standard deviation $\sigma_{t_k} = sd\left(\overline{M}_{\cdot t_k}\right)$ across the systems $j = 1, 2, \ldots, J$ that participated in the considered edition of a task. The *z-score* is given by:

$$m'_{s_j t_k} = \frac{m_{s_j t_k} - \mu_{t_k}}{\sigma_{t_k}} \tag{1}$$

The z-score is directly informative in a way that the raw score is not: "one can tell directly from a run score whether the system has performed well for the topic" (Webber et al., 2008).

Given that z-scores are centered around zero and unbounded, whereas the majority of IR measures are in the interval [0, 1], Webber et al. (2008) map them in this range by adopting the cumulative density function of the standard normal distribution to obtain the final *standardized scores*:

$$sm_{s_j t_k}\left(m'_{s_j t_k}\right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{m'_{s_j t_k}} e^{-\frac{x^2}{2}}\, dx \tag{2}$$

which also has the effect of reducing the influence of outliers.

The standardization process described above works topic-by-topic. When, in the following, we reason in terms of mean performances of a system s_j, it means that we compute the average of the standardized version of a measure over all the topics $k = 1, 2, \ldots, K$ for that system:

$$\overline{sm}_{s_j} = \frac{1}{K} \sum_{k=1}^{K} sm_{s_j t_k}\left(m'_{s_j t_k}\right) \tag{3}$$

So, for example, if we consider *Average Precision (AP)* as evaluation measure, Eq. (1) computes $ap'_{s_j t_k}$, which is the z-score of the AP values for topic t_k across all the systems. This z-score is then normalized by using Eq. (2) which produces $sap_{s_j t_k}$, i.e. the standardized score of AP for each topic and system. Finally, Eq. (3) computes, for each system, the average of $sap_{s_j t_k}$ across the topics and it corresponds to what *Mean Average Precision (MAP)* is when you use raw scores.

2.4. Score standardization for comparison between years

To be applied properly, score standardization requires only that performance scores come from some common distribution.

However, to ease the following discussion, let us make the usual assumption of normal distribution of the data[7] that is, for edition Y_i of a task, the raw performance scores $m_{Y_i} \sim \mathcal{N}\left(\mu_{Y_i}, \sigma_{Y_i}\right)$ belong to a normal distribution[8] with mean μ_{Y_i} and standard deviation σ_{Y_i}.

Given two editions of a task Y_1 and Y_2, Eq. (1) maps both the raw performance scores $m_{Y_1} \sim \mathcal{N}\left(\mu_{Y_1}, \sigma_{Y_1}\right)$ and $m_{Y_2} \sim \mathcal{N}\left(\mu_{Y_2}, \sigma_{Y_2}\right)$ to the standard normal distribution $m'_{Y_i} \sim \mathcal{N}(0, 1)$. This is what allows us to set them side by side and compare across the years, since they are all now expressed in terms of the same distribution and thus directly comparable.

Afterwards, Eq. (2) maps the z-scores in the interval [0, 1] by transforming by the cumulative density function of the standard normal distribution. Under the normality assumption of the data, this corresponds to computing the probability $sm_{Y_i} = \mathbb{P}[x \leq m'_{Y_i}]$ of observing a score less than or equal to the considered one, i.e. of finding a system performing worse than or equal to the considered one. Conversely, we have a probability $1 - sm_{Y_i}$ of observing a score greater than the considered one. Therefore, an increase of $sm_{Y_1} < sm_{Y_2}$ from year 1 to year 2 represents an improvement in the performances because in year 1 is less probable to find lower performances than in year 2 or, conversely, it is more probable to find higher performances than in year 2.

When we reason in terms of mean performances of a system s_j, we are basically reasoning in terms of a kind of "expected probability" of observing less performing systems over the averaged topics.

As pointed out above, the normality assumption is not indispensable to ensure that standardization works properly. When the normality assumption is not met, the main difference is in the interpretation of the scores produced by Eq. (2). The closer to the normality assumption, the more sm_{Y_i} matches the probability $\mathbb{P}[x \leq m'_{Y_i}]$ of observing a score less than or equal to the considered one; the farther from the normality assumption, the less sm_{Y_i} is a probability and the more it is just a transformed score in the range [0, 1].

[7] IR measures data typically meet the normality assumption only to a certain extent.

[8] Note that here we use a lighter notation from the one of Eqs. (1)–(3), not explicitly reporting also s_j and t_k in all the formulas, since it is clear from the context the dependency on them.

Please cite this article as: N. Ferro, G. Silvello, 3.5K runs, 5K topics, 3M assessments and 70M measures: What trends in 10 years of Adhoc–*ish* CLEF? Information Processing and Management (2016), http://dx.doi.org/10.1016/j.ipm.2016.08.001

2.5. Limitations of score standardization

The score standardization methodology was designed to compensate for collection effects, in particular the high variability due to topics, and to allow the actual contributions of the systems to better emerge by making performance figures interpretable on their own. Clearly, score standardization does not fully compensate for all the possible effects, e.g. the corpora effect or second-order interactions between systems and topics. Nevertheless, Webber et al. (2008) have shown that it is a robust enough tool for conducting analyses and highlighting trends.

In particular, topic variability is a major effect for CLEF collections while corpora are much more comparable and shared across years and tasks, as discussed in Section 3.1; this makes score standardization a suitable methodology to be applied in our case. Moreover, for many different languages, we plan to study performance trends across years for each language but not across languages. This keeps the variation among collections under control, making it somewhat similar to analyzing many TRECs in parallel, one for each language, while still being within the boundaries of the base assumptions of score standardization.

Score standardization needs a minimum number of systems to be sampled to provide reliable scores: Webber et al. (2008) report that as few as 5 systems are needed to achieve consistent results while between 10 and 15 systems are enough for better results. Therefore, as Section 3 will show, we will not analyze all the possible tasks that fall in the period under examination but only those for which there are enough systems to be sampled.

The major drawback of score standardization is that it tends to flatten out data, in particular outliers, because of the smoothing due to Eq. (2). This impacts top performing systems, the effects of which prove result to be less marked. However, since score standardization is a monotone transformation, the best systems continue to be the top performing ones. Therefore, this methodology is sub-optimal for studies only focused on steady performance improvement or quantifying the achieved performance gap over the years, although it is still suitable for our main objective, which is detecting and analyzing performance trends in order to appreciate the impact and influence of CLEF.

3. Experimental setup

We considered the CLEF Adhoc-*ish* labs with informational intents from 2000 to 2009, which corresponds to the "CLEF Classic" period. These labs are: Adhoc (AH) (Agirre, Di Nunzio, Ferro, Mandl, & Peters, 2008; Agirre, Di Nunzio, Ferro, Mandl, & Peters, 2009; Braschler, 2001, 2002, 2003, 2004; Braschler, Di Nunzio, Ferro, & Peters, 2005; Di Nunzio, Ferro, Jones, & Peters, 2005; 2006a; Di Nunzio, Ferro, Mandl, & Peters, 2006; Di Nunzio, Ferro, Mandl, & Peters, 2007a; 2007b; 2008; Ferro & Peters, 2009; 2010), GeoCLEF (GC) (Gey, Larson, Sanderson, Bischoff, et al., 2006; Gey et al., 2007; Gey, Larson, Sanderson, Joho, et al., 2006; Mandl, Carvalho, et al., 2008; Mandl et al., 2007; Mandl, Gey, et al., 2008), Robust (ROB) (Agirre et al., 2008; Agirre Di Nunzio, Ferro, Mandl, & Peters, 2009; Agirre, Di Nunzio, Mandl, & Otegi, 2009; Agirre, Di Nunzio, Mandl, & Otegi, 2010; Di Nunzio, Ferro, Mandl, & Peters, 2006; Di Nunzio et al., 2007a; 2007b; 2008), and "The European Library" (TEL) (Agirre et al., 2008; Agirre Di Nunzio, Ferro, Mandl, & Peters, 2009; Agirre, Di Nunzio, Mandl, & Otegi, 2009; Agirre et al., 2010).

Fig. 1 shows the timeline 2000–2009 of all the CLEF Adhoc-*ish* tasks broken down by their language, identified by their ISO 639:1 two letters code, and their kind, i.e. monolingual, bilingual and multilingual tasks.

Monolingual tasks use the same source and target languages, i.e. the same language for topics (source) and documents (target); for example, "AH MONO DE" refers to a monolingual task of the Adhoc lab for the German language where the systems used German topics against German documents.

Bilingual tasks use a target language for the documents which is different from the source language for the topics; for instance, "AH BILI X2EN" refers to a bilingual task where the systems retrieve documents in English but use topics in different source languages (e.g. French or Chinese) chosen from the available ones.

Multilingual tasks use more than one target language at the same time for the documents starting from a source language for the topics; for instance, "AH MULTI-4" refers to a multilingual task using a corpora made up of 4 different target languages, e.g. English, French, German, and Italian, where the systems retrieve documents in all these 4 languages at the same time using topics in a source language (e.g. Dutch) chosen from the available ones.

The Adhoc (AH) tasks ran from 2000 to 2009 and the documents composing their corpora are newspaper articles and news agency dispatches; all these tasks have an informational search intent. The Robust (ROB) tasks have the same search intent and are based on the same corpora and topics used in the Adhoc ones, with a large number of hard topics selected from the previous years.

GeoCLEF (GC) is an Adhoc task which ran from 2005 to 2008 with an emphasis on geographic search. It was aimed at evaluating *Geographical Information Retrieval (GIR)* systems where topics contained spatial hints that may or may not be exploited for additional reasoning. GC tasks employed the same corpora as the Adhoc ones.

The TEL tasks ran in 2008 and 2009; they have an informational search intent as well, but they are based on large numbers of bibliographical records. TEL adds a further level of indirection: whereas in the traditional Adhoc task the user searches directly for a document containing information of interest, within the TEL tasks the user tries to identify which publications are of potential interest according to the information provided by the bibliographic record.

Appendix A reports detailed information about the analyzed tasks. In particular, Tables A.1–A.3 summarize all the details about the analyzed tasks: the corpora and number of documents used; the number of topics; the size of the pool; the

Please cite this article as: N. Ferro, G. Silvello, 3.5K runs, 5K topics, 3M assessments and 70M measures: What trends in 10 years of Adhoc–*ish* CLEF? Information Processing and Management (2016), http://dx.doi.org/10.1016/j.ipm.2016.08.001

		2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Monolingual											
	Bulgarian (bg)						AH	AH	AH		
	Czech (cs)								AH		
	Dutch (nl)		AH	AH	AH			ROB			
	Farsi (fa)									AH	AH
	Finnish (fi)			AH	AH	AH					
	French (fr)	AH	AH	AH	AH	AH	AH	AH ROB	ROB	TEL	TEL
	German (de)	AH	AH	AH	AH		GC	GC	GC	GC TEL	GC TEL
	Hungarian (hu)						AH	AH	AH		
	Italian (it)	AH	AH	AH	AH			ROB			
	Portuguese (pt)					AH	AH	AH GC	GC ROB	GC	
	Russian (ru)				AH	AH					
	Spanish (es)		AH	AH	AH			GC ROB			
	Swedish (sv)			AH	AH						
Bilingual											
	Bulgarian (bg)						AH	AH	AH		
	Czech (cs)								AH		
	Dutch (nl)		AH	AH	AH			ROB			
	English (en)	AH	AH	AH	AH	AH	AH GC	AH GC	AH GC	ROB TEL	ROB TEL
	Farsi (fa)									AH	AH
	Finnish (fi)			AH		AH					
	French (fr)			AH		AH	AH	AH			
	German (de)			AH	AH		GC	GC ROB	GC	GC TEL	TEL
	Hungarian (hu)						AH	AH	AH		
	Italian (it)			AH	AH						
	Portuguese (pt)					AH	AH	AH GC	AH GC	AH GC	
	Russian (ru)				AH	AH					
	Spanish (es)			AH	AH			GC ROB			
	Swedish (sv)			AH							
Multilingual											
MULTI-4	de, en, fr, it	AH									
MULTI-4	de, en, es, fr				AH	AH					
MULTI-5	de, en, es, fr, it		AH	AH							
MULTI-8	de, en, es, fi, fr, it, nl, sv				AH						
2-YEARS-ON	de, en, es, fi, fr, it, nl, sv						AH				
MERGING	de, en, es, fi, fr, it, nl, sv						AH				

Fig. 1. Timeline of all the CLEF Adhoc-*ish* monolingual, bilingual and multilingual tasks from 2000 to 2009 broken down by language, identified by their ISO 639:1 two letters code. AH stands for Adhoc; GC for GeoCLEF, ROB for Robust; TEL for "The European Library".

Please cite this article as: N. Ferro, G. Silvello, 3.5K runs, 5K topics, 3M assessments and 70M measures: What trends in 10 years of Adhoc–*ish* CLEF? Information Processing and Management (2016), http://dx.doi.org/10.1016/j.ipm.2016.08.001

number of participating groups with new groups within brackets; the number of submitted runs; in the case of bilingual and multilingual tasks we also report the source and target languages. All the systems that participated in the tasks reported in Tables A.1–A.3 have been used to conduct this study.

As noted in the discussion in Section 2.5, score standardization requires at least 5 runs to provide consistent results and 10 or more for better ones. Thus, there is a trade-off between the minimum number of runs needed to apply score standardization and the sparsity of the data which are spread across many years and tasks. In order to avoid leaving out too many of the tasks shown in Fig. 1, hampering the possibility of detecting trends year after year, we analyze tasks for which at least 9 valid runs have been submitted. The only exceptions are for some Robust tasks and one TEL bilingual task, where there are less than 9 runs but at least 5. We consider a run as valid if it retrieves documents for each topic of the collection.

All the CLEF results that we analyzed in this paper are available through the *Distributed Information Retrieval Evaluation Campaign Tool (DIRECT)* system[9] (Agosti et al., 2012; Agosti & Ferro, 2009; Silvello, Bordea, Ferro, Buitelaar, & Bogers, 2016); the software library (i.e. MATTERS) used for processing raw data, calculating measures, analyzing them and plotting the graphs, the source code we produced and all the data (original measures, standardized measures and z-scores) discussed in the next sections are available at the following address: http://matters.dei.unipd.it/.

3.1. Corpora

The CLEF corpora are formed by document sets in different European languages but with common features: the same genre and time period, comparable content. Indeed, the large majority of the corpora are composed by newspaper articles from 1994–1995 with the exception of the Bulgarian, Czech and Hungarian corpora composed of newspaper articles from 2002 and the Persian corpus (i.e. Farsi) which spans the period from 1996 to 2002.

The French, German and Italian news agency dispatches – i.e. ATS, SDA and AGZ – are all gathered from the Swiss news agency and are the same corpus translated in different languages. The Spanish corpus is composed of news agencies (i.e. EFE) from the same time period as the Swiss news agency corpus and thus it is very similar in terms of structure and content.

All these corpora are very similar to one another and are stable across the various tasks using them, i.e. Adhoc, Robust and GeoCLEF. Moreover, within the same task the variations in the adopted corpus are minimal or not present. This is a positive fact for standardization because it limits the effect of the corpus in inter-collection comparisons, as discussed in Section 2.5

The only exception to this is the TEL corpus, which consists of bibliographic records from the English, French and German national libraries expressed using an expanded version of Dublin Core. Bibliographic records tend to be shorter than news articles and more sparse, i.e. there is some variations in the Dublin Core fields used from record to record. Therefore, TEL data is a bit different from the data used in the other Adhoc-*ish* labs and this introduces a limited collection effect which is less compensated by the score standardization. Therefore, TEL labs are comparable over the years with each other but a little more care is needed when comparing them to the other CLEF Adhoc-*ish* labs.

Tables A.1–A.3 in Appendix A report the corpora used in each of the analyzed tasks together with their total number of documents.

3.2. Topics

CLEF topics follow the typical TREC structure composed of three fields: title, description and narrative. The topic creation process in CLEF has had to deal with specific issues related to the multilingualism as described in Kluck and Womser-Hacker (2002) and it has been based on three main steps: (i) for each language considered, a team of native speakers generates a certain number of topics accordingly to a predefined set of rules, e.g. the topics must be real-life and must meet the content of all the considered corpora in multiple languages; (ii) in a plenary meeting, topic creators revise the different subsets of topics created by each language team and select a common set of topics pooled from the subsets in the different languages in such a way that maximizes the topic appropriateness across all the used corpora; (iii) the experts translate the selected topics in a pivot language, typically English, and then back to all the other languages which are offered. Even with this careful topic design, for few topics there may not be enough relevant documents in the corpora for some specific languages; in such a case, those specific topics are discarded for that language and for this reason the number of topics reported in Tables A.1–A.3 in Appendix A may vary from language to language within the same edition of CLEF.

In general, the CLEF Adhoc-*ish* tasks have used 50 or more topics. Only GeoCLEF used 25 topics, based on the findings (Sanderson & Zobel, 2005).

Fig. 2 shows an example of topics in four languages – English, French, Chinese and Bulgarian – used in the Adhoc 2007 bilingual to English task (Braschler et al., 2005).

Fig. 3 shows an example of GeoCLEF topic in two languages – English and Portuguese – taken from the 2008 bilingual to German GC task. It can be noted how it shares the same informational intent as the typical Adhoc topic previously shown but with some additional geographical indications

Please cite this article as: N. Ferro, G. Silvello, 3.5K runs, 5K topics, 3M assessments and 70M measures: What trends in 10 years of Adhoc–*ish* CLEF? Information Processing and Management (2016), http://dx.doi.org/10.1016/j.ipm.2016.08.001

```
<topic lang="en">
    <identifier>401-AH</identifier>
    <title>Euro Inflation</title>
    <description>
        Find documents about rises in prices after the introduction of the Euro.
    </description>
    <narrative>
        Any document is relevant that provides information on the rise of prices in any
        country that introduced the common European currency.
    </narrative>
</topic>
<topic lang="fr">
    <identifier>401-AH</identifier>
    <title>Inflation de l'Euro</title>
    <description>
        Trouver des documents qui parlent de la hausse des prix après l'introduction de l'Euro.
    </description>
    <narrative>
        Les documents pertinents fourniront des informations sur la hausse des prix dans
        n'importe quel pays ayant introduit la monnaie unique européenne.
    </narrative>
</topic>
<topic lang="zh">
    <identifier>401-AH</identifier>
    <title>歐元通貨膨脹</title>
    <description>尋找有關使用歐元後物價上漲的文件</description>
    <narrative>相關的文件會提到使用歐元的國家物價上漲的情形</narrative>
</topic>
<topic lang="bg">
    <identifier>401-AH</identifier>
    <title>Инфлацията на еврото</title>
    <description>
        Намерете документи за повишаването на цените след въвеждане на еврото.
    </description>
    <narrative>
        Подходящ е всеки документ, който дава информация за повишаването на цените в
        която и да е държава, въвела единната европейска валута.
    </narrative>
</topic>
```

Fig. 2. Topic 401-AH of the AH-BILI-X2EN-2007 task.

Fig. 4 shows an example of a TEL topic in four languages – English, Italian, German and Chinese – taken from the 2009 bilingual to English TEL task. It can be noted how it shares the same informational intent as the typical Adhoc topic previously shown, but it is instead trying to understand the relevance of a book from the description contained in its bibliographic record.

As the above examples show, all the topics used in the different Adhoc-*ish* tasks are quite close in nature and genre, with slight modifications due to the specificities of each task. These circumstances together with the comparability of the corpora make the CLEF Adhoc-*ish* tasks suitable for the application of the score standardization method.

3.3. Ground truth

As far as relevance assessments are concerned, CLEF adopted they standard approach based on the pooling method and the assessment based on the longest, most elaborate formulation of the topic, i.e. the narrative (Sanderson, 2010). Typical pool depths are between 60 and 100 documents. The stability of pools constructed in this way and their reliability for post-campaign experiments is discussed in Braschler (2004) and Tomlinson (2007); 2009).

Information about the pool sizes for all the CLEF experimental collections employed in this paper are reported in Tables A.1–A.3 in Appendix A.

3.4. Participation

Tables A.1 –A.3 in Appendix A report the number of groups participating in each analyzed task as well as the number of newcomers with respect to the previous year. This information is particularly useful for understanding the CLEF performance trends, because newcomers have a significant impact on performances from year to year. For instance, in the “AH-Bili-X2EN” task in the 2002–2005 timespan all the participants were newcomers.

[9] http://direct.dei.unipd.it/

Please cite this article as: N. Ferro, G. Silvello, 3.5K runs, 5K topics, 3M assessments and 70M measures: What trends in 10 years of Adhoc–*ish* CLEF? Information Processing and Management (2016), http://dx.doi.org/10.1016/j.ipm.2016.08.001

```
<topic lang="en">
    <identifier>94-GC</identifier>
    <title>Demonstrations in German cities</title>
    <description>Documents mentioning demonstrations in German cities</description>
    <narrative>
        Relevant documents contain information about participants, and number of
        participants, reasons, type (peaceful or riots) and consequences of
        demonstrations in German cities.
    </narrative>
</topic>
<topic lang="pt">
    <identifier>94-GC</identifier>
    <title>Manifestações em cidades alemãs</title>
    <description>
        Documentos que mencionem a realização de manifestações em cidades da Alemanha
    </description>
    <narrative>
        Documentos relevantes devem conter informação sobre o número de manifestantes,
        as razões que levaram à manifestação, o tipo de manifestação (pacífica, motim)
        e as consequências da mesma.
    </narrative>
</topic>
```

Fig. 3. Topic 94-GC of the GC-BILI-X2DE-2008 task.

```
<topic lang="en">
    <identifier>750-AH</identifier>
    <title>Contemporary French Philosophers</title>
    <description>
        Find books by 20th or 21st century French philosophers or on their philosophy.
    </description>
</topic>
<topic lang="it">
    <identifier>750-AH</identifier>
    <title>Filosofi francesi contemporanei</title>
    <description>
        Trova libri su filosofi francesi del XX o XXI secolo, oppure sulla loro filosofia.
    </description>
</topic>
<topic lang="de">
    <identifier>750-AH</identifier>
    <title>Zeitgenössische französische Philosophen</title>
    <description>
        Finden Sie Bücher von französischen Philosophen des 20. oder 21. Jahrhunderts
        oder über ihre Philosophie.
    </description>
</topic>
<topic lang="zh">
    <identifier>750-AH</identifier>
    <title>当代法国哲学家</title>
    <description>查找20世纪或者21世纪法国哲学家的著作或者有关其哲学的书籍。</description>
</topic>
```

Fig. 4. Topic 750-AH of the AH-TEL-BILI-X2EN-2009 task.

In general, the average ratio of newcomers in CLEF Adhoc-*ish* tasks is 58%, indicating a high renewal rate from year to year. Bilingual tasks show the higher renewal rate at 73%, whereas the renewal rate for monolingual tasks is lower, i.e. 49%. As shown in Table A.2, bilingual tasks changed the source languages from year to year for the same target language, thus attracting different groups according to the specific language pair being investigated in the task; vice versa, monolingual tasks do not present this variability given that both the source and target languages are fixed.

3.5. Measures

Average Precision (AP) (Buckley & Voorhees, 2005) represents the "gold standard" measure in IR, known to be stable and informative, with a natural top-heavy bias and an underlying theoretical basis as approximation of the area under the precision/recall curve. MAP is the average of AP over all the topics for a given system.

AP is the reference measure in this study for all CLEF tasks and it is the measure originally adopted by CLEF for evaluating the systems participating in the campaigns. In the following, with sAP we refer to the standardized version of AP and with sMAP to the standardized version of MAP, as also discussed in Section 2.3.

Please cite this article as: N. Ferro, G. Silvello, 3.5K runs, 5K topics, 3M assessments and 70M measures: What trends in 10 years of Adhoc – ***ish*** CLEF? Information Processing and Management (2016), http://dx.doi.org/10.1016/j.ipm.2016.08.001

Table 1
Best and median sMAP of the CLEF monolingual tasks. In bold the best performances for each language across the different tasks and in brackets the variation in the performances with respect to the previous year.

Task	Year	Best sMAP	Median sMAP
AH mono BG	2005	69.79% (–)	47.84% (–)
	2006	58.03% (−20.25%)	50.93% (+6.06%)
	2007	**75.91%** (+23.55%)	**51.85%** (+1.79%)
AH mono ES	2001	74.02% (–)	**63.21%** (–)
	2002	**80.65%** (+8.22%)	57.23% (−9.46%)
	2003	70.16% (−14.95%)	56.30% (−1.62%)
Robust mono ES	2006	59.61% (–)	52.55% (–)
AH mono DE	2000	**83.09%** (–)	52.35% (–)
	2001	68.57% (−17.47%)	**58.39%** (+11.53%)
	2002	68.88% (+0.45%)	57.80% (−1.01%)
	2003	73.30% (+6.42%)	52.54% (−9.10%)
GC mono DE	2005	64.38% (–)	46.71% (–)
	2006	58.17% (−10.66%)	51.45% (+9.22%)
	2007	64.02% (+9.13%)	50.07% (−2.76%)
	2008	61.13% (−4.73%)	54.18% (+7.60%)
Robust mono DE	2006	74.03% (–)	45.73% (–)
TEL mono DE	2008	73.88% (–)	49.85% (–)
	2009	64.93% (−12.11%)	51.23% (+2.76%)
AH mono FA	2008	**84.14%** (–)	45.63% (–)
	2009	67.93% (−23.86%)	**55.95%** (+14.44%)
AH mono FI	2002	68.13% (–)	50.68% (–)
	2003	65.10% (−4.64%)	**55.29%** (+8.33%)
	2004	**74.53%** (+12.64)	50.86% (−8.71%)
AH mono FR	2000	69.52% (–)	53.70% (–)
	2001	69.08% (−0.63%)	54.12% (+0.78%)
	2002	**82.57%** (+19.53%)	56.09% (+3.64%)
	2003	67.58% (−18.15%)	55.65% (−0.78%)
	2004	67.77% (+0.28%)	50.34% (−9.54%)
	2005	71.76% (+5.89%)	58.33% (+15.87%)
	2006	69.92% (−2.56%)	51.20% (−12.22%)
Robust mono FR	2006	65.76% (–)	52.47% (–)
	2007	**71.57%** (+8.12%)	51.64% (−1.60%)
TEL mono FR	2008	72.42% (–)	50.18% (–)
	2009	68.38% (−5.58%)	53.34% (+6.30%)
AH mono HU	2005	72.35% (–)	52.45% (–)
	2006	69.13% (−4.65%)	48.98% (−7.07%)
	2007	**74.77%** (+7.54%)	**54.96%** (+10.88%)
AH mono IT	2000	61.14% (–)	51.50% (–)
	2001	**74.67%** (+22.13%)	**54.61%** (+6.04%)
	2002	73.54% (−1.51%)	**54.61%** (–)
	2003	67.96% (−7.59%)	51.42% (−5.84%)
Robust mono IT	2006	62.58% (–)	48.17% (–)
AH mono NL	2001	68.44% (–)	**52.96%** (–)
	2002	71.28% (+4.15%)	51.18% (−3.36%)
	2003	**72.31%** (+1.45%)	46.57% (−10.53%)
Robust mono NL	2006	69.95% (–)	51.91% (–)
AH mono PT	2004	68.97% (–)	53.74% (–)
	2005	67.79% (−1.75%)	52.52% (−2.32%)
	2006	67.91% (+0.18%)	**55.47%** (+5.31%)
GC mono PT	2006	**74.77%** (–)	42.18% (–)
	2007	69.14% (−8.15%)	49.75% (+15.21%)
	2008	63.32% (−9.19%)	54.18% (+8.43%)
Robust mono PT	2007	72.47% (–)	60.42% (–)

4. Monolingual tasks

Table 1 summarizes the performances achieved in the monolingual tasks in terms of best and median sMAP. This information is shown graphically also in Fig. 5 which reports the distribution of sMAP scores by means of box plots.

4.1. Highlights

We can observe two general trends corresponding to a focus shift of CLEF over the years.

Please cite this article as: N. Ferro, G. Silvello, 3.5K runs, 5K topics, 3M assessments and 70M measures: What trends in 10 years of Adhoc–*ish* CLEF? Information Processing and Management (2016), http://dx.doi.org/10.1016/j.ipm.2016.08.001

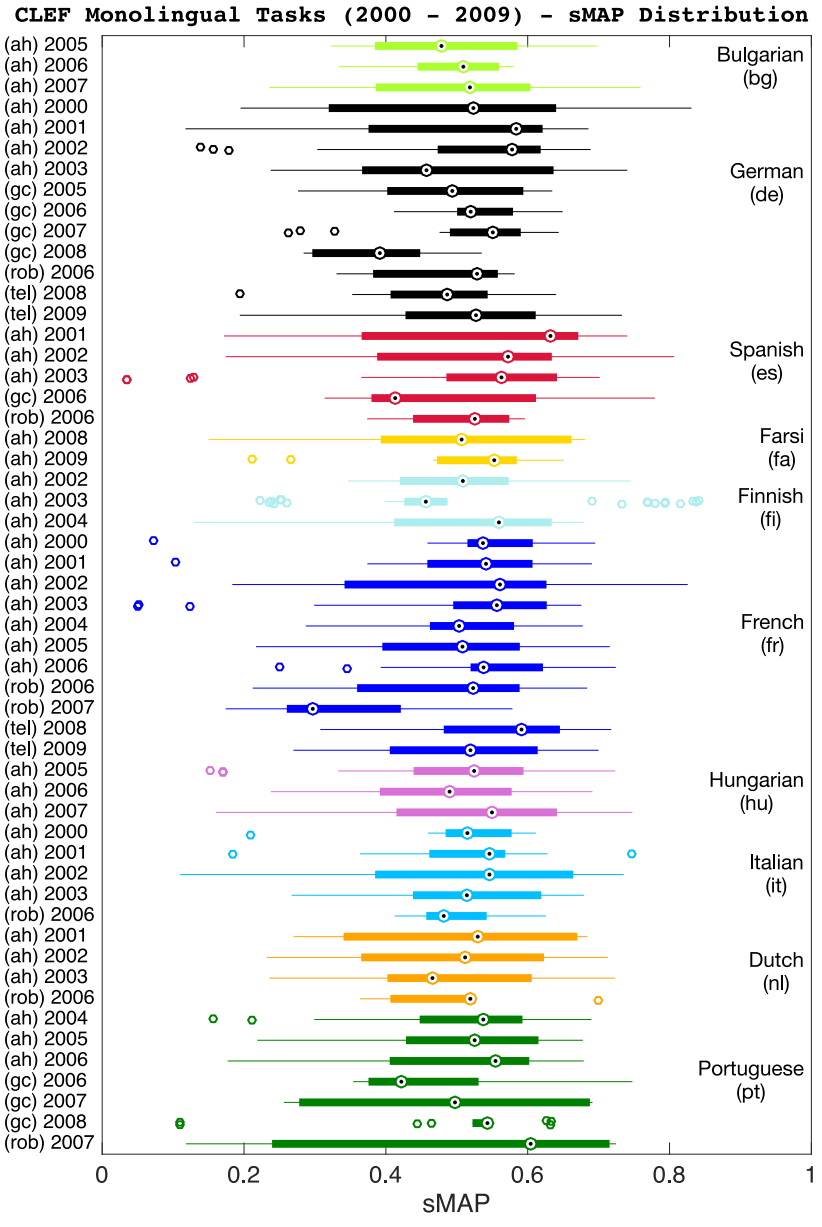

Fig. 5. sMAP distribution for the selected CLEF monolingual tasks 2000–2009 grouped by language.

In the period from 2000 to 2004, CLEF was starting up and it was more focused on European languages like French, German, Italian and others, for which there were more consolidated language resources and approaches as well as previous experimental results deriving from the TREC *Cross Language Information Retrieval (CLIR)* tracks (Schaüble & Sheridan, 1997).

In the period from 2004 to 2009, the focus of CLEF moved to other European languages like Bulgarian, Hungarian and Portuguese for which there were less consolidated language resources and approaches and for which there was a need to support the growth of a multidisciplinary research community as well as an alternative viewpoint on Adhoc retrieval, such as geographical intents or bibliographic emphasis.

In the first period (2000–2004), a recurring pattern can be observed where median performances increase in the early years of the task, reach a plateau and then decrease or assume an oscillatory behavior. An exception to this pattern is the Spanish monolingual task that reports a steady decrease of median performances over the years; in this case, an important factor to be considered is the large proportion of newcomers participating in the task from year to year. Moreover, the best performances were achieved in the early years of the task where the more experienced groups worked on new language

Please cite this article as: N. Ferro, G. Silvello, 3.5K runs, 5K topics, 3M assessments and 70M measures: What trends in 10 years of Adhoc–*ish* CLEF? Information Processing and Management (2016), http://dx.doi.org/10.1016/j.ipm.2016.08.001

resources and on the tuning of already tested and working systems; whereas, in the later years of the tasks their focus shifted to testing new techniques and retrieval settings that may or may not had positive effects on system effectiveness.

In the second period (2004–2009), a steady growing trend in median performances across the years can be seen. Moreover, the best performances were achieved in the last year of the tasks and in many cases by the most experienced groups participating in CLEF. The research on new language resources conducted in the early years of the tasks had a decisive role in the top performances increase across the years.

In this second period we can also observe that median performances in the Adhoc tasks are typically higher than those in the Robust, GeoCLEF, and TEL tasks. This provides experimental evidence that these tasks have actually introduced some challenges to the basic Adhoc task: Robust by selecting hard topics, GeoCLEF by introducing geographical constraints and TEL by the further level of indirection represented by bibliographic records, which are also very sparse.

Furthermore, the role of newcomers is central for interpreting the performance trends in CLEF monolingual tasks. For tasks with a high percentage of newcomers, clear trends are more difficult to identify, e.g. AH-MONO-ES, whereas when the percentage of newcomers is lower (below 50%) and there is a consolidated class of returning groups, the results are more stable, e.g. AH-MONO-PT and GC-MONO-DE.

Overall, the development of new and better language resources had a sizable impact on the improvement of MLIA systems. Moreover, the work of experienced groups, which participated in many editions of CLEF, contributed decisively to the advancement of information retrieval techniques in languages other than English while the high number of newcomers contributed to the creation of a large and multidisciplinary research community with shared skills and competencies.

Finally, the typical 3-year duration for an evaluation cycle seems to be appropriate for the monolingual case, even though in cases of languages with less consolidated resources a fourth round might have made it possible to observe behavior more similar to that of more resourced languages.

4.2. Keystone cases

In this section, we explore more in detail three relevant cases which represent specific examples of some of the observed trends for monolingual tasks and more clearly show the behavior and interests of participating groups as well as the impact of the CLEF campaigns.

4.2.1. The first period (2000–2004): detailed trends for best systems

Let us analyze the French case in details by focusing on the best performing systems in the AH-MONO-FR tasks. We can observe that the best sMAP in 2001 is similar to that achieved in 2000, whereas in 2002 it increases notably (+19.53%) to reach the best performances ever achieved in the French monolingual task, only to decrease in 2003 (−18.15%). The best runs from 2001 to 2003 were all submitted by the University of Neuchâtel. In 2001, the University of Neuchâtel took part in CLEF for the first time and its goal was to define a general stopword list for the European languages as well as to provide simple and efficient stemming procedures (Savoy, 2002). In 2002 (Savoy, 2003), the goal was overtly to improve the results obtained in the previous year by employing a better and more general stopword list and an improved, simpler and efficient stemmer; furthermore, they leveraged on the SMART system (Buckley, 2005) as a testbed for implementing the Okapi probabilistic model (Robertson, Walker, & Beaulieu, 2000) as well as other vector space models. The growth in performances from 2000 to 2002 was mainly due to a constant improvement of the language resources available for the French language as reported by the University of Neuchâtel which focused on improving the stopword list and the stemmer from year to year.

In 2003 (Savoy, 2004b), the focus was on providing good stopword lists and stemmers for Finnish, Swedish and Russian languages whereas the focus on other European languages including French was less marked. The goal was not to further improve already tested approaches for the French language, but to investigate novel techniques; indeed, they investigated retrieval by adopting fusion techniques such as CombMNZ (Fox & Shaw, 1993) and by proposing improvements to these techniques tested here for the first time. This approach continued also in 2004 where the goal was to propose and investigate novel data fusion techniques such as NormRSV or Z-score (Savoy, 2004a); indeed, the performances did not increase notably with respect to the previous year given that the goal was overtly to try different approaches without necessarily improving techniques which already worked well.

The behavior of the best systems for French is consistent with the behavior for Spanish. Indeed, the best system in 2001 and 2002 for this language was also the one adopted by the University of Neuchâtel, which applied the techniques described for French also for Spanish and achieved good results. In 2003, the last year of the Spanish monolingual task, the University of Neuchâtel focused on newly introduced languages and tested novel fusion techniques as described above; in 2003 the best system was the one adopted by Fondazione Ugo Bordoni which participated in the Spanish task for the first time, but with its consolidated experience from the Italian monolingual task. The performances obtained by Fondazione Ugo Bordoni in 2003 (Amati, Carpineto, & Romano, 2003) for the Spanish monolingual task are consistent with those obtained for the Italian monolingual task. This fact also casts light on the affinity between Spanish and Italian from the retrieval point-of-view as we discuss more in detail in the analysis of bilingual tasks. As explained in detail by Ferro and Silvello (2014), the Italian case also follows the pattern observed in the French task.

Please cite this article as: N. Ferro, G. Silvello, 3.5K runs, 5K topics, 3M assessments and 70M measures: What trends in 10 years of Adhoc–*ish* CLEF? Information Processing and Management (2016), http://dx.doi.org/10.1016/j.ipm.2016.08.001

4.2.2. The first period (2000–2004): detailed trends for median systems

Comparable behavior to that of best systems can be seen by examining the performances of systems which participated each year in a task with good (even though not necessarily the best) results.

To this end, let us consider the Hummingbird Fulcrum SearchServer system (Hummingbird in the following) for the monolingual French task from 2001 to 2003; Hummingbird always performed above the median, indeed its sMAP value is 18% higher than the median in 2001, 3% higher than the median in 2002 and 20% higher than the median in 2003. Hummingbird achieved an sMAP 8% lower than the best in 2001, 30% lower than the best in 2002 and only 1% lower than the best in 2003.

We can see that the only time when this system was not amongst the top performing ones was 2002. In 2001 Hummingbird focused on linguistic expansion which had a positive impact especially on French (Tomlinson, 2001). In 2002 Hummingbird experimented new techniques to deal with French by introducing the indexing of accented words and treating apostrophes as word separators and it discovered that the first technique was harmful for French whereas the second was beneficial (Tomlinson, 2002). In 2003 Hummingbird learned from the previous years and the system employed a lexical stemmer which tolerated missing accents for French (Tomlinson, 2003); as a consequence Hummingbird performances were very close to those of the best system. In the following years, Hummingbird participated in the monolingual French tasks with good results, but there was a clear research focus shift to other newly introduced languages such as Portuguese in 2004 (Tomlinson, 2004) and Bulgarian in 2005 (Tomlinson, 2005).

4.2.3. The second period (2004–2009): detailed trends for best systems

It is interesting to highlight that the best performances for the Portuguese task were achieved by a different group each year (i.e. University of Neuchâtel, Hummingbird and Johns Hopkins University). These three groups have broad experience, have participated in many tasks over the years in CLEF and in particular they participated each year in the Portuguese task contributing with a sizable number of runs – 40% of the runs submitted in 2004 were from these three groups and 30% in 2005 and 2006. This could also explain the small variation in median sMAP from year to year for the Portuguese monolingual task.

The research for improving retrieval performances for Bulgarian and Hungarian started later than for other European languages and the focus of research groups participating in CLEF remained oriented towards the improvement of performances. For instance, for Hungarian the performances of the best system steadily improved from 2005 to 2007. The best run of 2005 was submitted by a research group from the Johns Hopkins University (McNamee, 2005) which exploited its lengthy experience with language neutral methods (e.g. n-grams) for cross-language IR, but had never worked before with Hungarian. In the 2006 the University of Neuchâtel system (Savoy & Abdou, 2006) obtained the best sMAP, 6% higher than the best performance in 2005; its goal was to "propose and evaluate various indexing and search strategies for the Hungarian language in order to produce better retrieval effectiveness than language-independent approach (n-gram)", thus to overcome the issues recognized in the previous year evaluation cycle. In 2007 (Dolamic & Savoy, 2007), the goal of the University of Neuchâtel which achieved the best sMAP for Hungarian (+1.79% w.r.t. the previous year), was to get a better picture of the relative merit of various search engines in exploiting Hungarian documents; they applied several state-of-the-art retrieval models such as Okapi, Divergence from Randomness (Amati & van Rijsbergen, 2002) and some Language Models (Ponte & Croft, 1998).

5. Bilingual tasks

Table 2 summarizes the performances achieved in the bilingual tasks in terms of best and median standardized MAP. This information is shown graphically also in Fig. 6 which reports the distribution of sMAP scores by means of box plots.

5.1. Highlights

As a general trend, we can observe that top performances in terms of both median and best sMAP are typically achieved in the second or third year of bilingual tasks. This is probably an indicator of the progressive availability of better language resources, such as dictionaries, parallel corpora and on-line resources, which is crucial for a bilingual task.

However, it is somewhat more difficult to identify recurring patterns in system performances for bilingual tasks than for monolingual ones due to the sizable turnover in group participation (73% on average of newcomers from year to year) and the variation of source languages from year to year.

Moreover, variations in system performances are often related to specific source and target language pairs: pairs of languages with a greater affinity, e.g. German to English or Spanish to Portuguese, often obtain higher performances than those with a lesser affinity, such as French to English or English to Portuguese.

English represents somewhat of a special case: it served as the entry point for introducing new source languages, which afterwards might have also been tried towards other European target languages, and it acted as the easiest task for those groups approaching cross-lingual IR for the first time. This is also witnessed by the large number of newcomers which each year participated in the task.

Moreover, it can be divided into two sub-cases. The tasks from 2000 to 2004 were mainly focused on well-resourced European languages to English, which also benefited from previous research at the TREC CLIR track, and achieved top

Please cite this article as: N. Ferro, G. Silvello, 3.5K runs, 5K topics, 3M assessments and 70M measures: What trends in 10 years of Adhoc–*ish* CLEF? Information Processing and Management (2016), http://dx.doi.org/10.1016/j.ipm.2016.08.001

Table 2
Best and median sMAP of the CLEF bilingual tasks. In bold the best performances for each language across the different tasks and enclosed in brackets the performances variation with respect to the previous year.

Task	Year	Best sMAP	Median sMAP
AH bili DE	2002	66.74% (–)	53.40% (–)
GC bili DE	2005	66.69% (–)	43.32% (–)
	2006	55.27% (−20.66%)	48.42% (+10.53%)
	2008	63.94% (+13.56%)	45.62% (−6.14%)
ROB bili DE	2006	60.13% (–)	**57.20%** (–)
TEL bili DE	2008	62.68% (−6,08%)	45.99% (−13.88%)
	2009	**71.79%** (14.53%)	47.31% (+2.87%)
AH bili EN	2000	74.63% (–)	51.96% (–)
	2001	77.25% (+3.51%)	56.18% (+8.12%)
	2002	69.83% (−9.60%)	45.24% (−19.47%)
	2003	69.80% (−0.04%)	40.74% (−9.95%)
	2004	58.95% (−15.54%)	52.51% (+28.89%)
	2005	**78.45%** (+33.08%)	**56.67%** (+7.92%)
	2006	75.59% (−3.64%)	48.08% (−15.16%)
	2007	77.46% (+2.47%)	48.35% (0.56%)
Robust bili EN	2008	53.79% (–)	65.76% (–)
	2009	62.95% (−13.24%)	55.77% (+3.56%)
GC bili EN	2005	64.54% (–)	46.63% (–)
	2006	64.96% (0.64%)	55.09% (15.36%)
	2007	59.17% (−9.78%)	49.18% (−12.02%)
	2008	63.11% (+6.23%)	52.02% (+5.46%)
TEL bili EN	2008	76.11% (−1,74%)	53.82% (+11.31%)
	2009	78.08% (2.59%)	47.19% (−12.32%)
AH bili ES	2002	**68.05%** (–)	49.69% (–)
	2003	67.37% (−1.01%)	**53.94%** (+8.55%)
ROB bili ES	2006	61.42% (–)	46.07% (–)
AH bili FR	2002	67.08% (–)	56.47% (–)
	2004	60.15% (−10.33%)	52.11% (−7.72%)
	2005	**72.50%** (+20.53%)	**57.03%** (+9.44%)
	2006	62.73% (−13.47%)	48.86% (−14.33%)
ROB bili FR	2007	69.73% (–)	51.62% (–)
TEL bili FR	2008	63.58% (–)	44.22% (–)
	2009	71.51% (+12.47%)	43.55% (−1.52%)
AH bili IT	2002	59.16% (–)	53.06% (–)
	2003	**71.19%** (+20.34%)	**53.09%** (+0.05%)
AH bili PT	2004	67.21% (–)	42.78% (–)
	2005	**72.39%** (+7.71%)	**50.20%** (+17.34%)
	2006	65.39% (−9.67%)	48.04% (−4.30%)
AH bili RU	2003	**68.94%** (–)	48.10% (–)
	2004	63.36% (−8.09%)	**52.03%** (+8.17%)

performances early on. The tasks from 2004 to 2009 focused on less-resourced European languages and on non-European languages. In this case, we can observe comparable enough performances over the years, even though the source languages changed markedly from year to year, and this can be considered as an indicator of the maturity of the technology for cross-lingual IR to English.

As in the case of monolingual tasks, we can observe that Robust, GeoCLEF and TEL tasks typically have slightly lower performances than the standard Adhoc task.

Finally, we can also report the positive interaction between CLEF and the other evaluation campaigns for the bilingual tasks.

The Chinese to English task, which ran in 2001, 2002 and 2007, shown an improvement of median performances from 2002 to 2007, even though there was no Chinese task in CLEF in-between. Many factors may have contributed to this improvements: for example, the work done in NTCIR for the Chinese language between 2002 and 2006 and the increased interest in Chinese from the machine translation community in this period may have generated a positive feedback on CLEF as well.

The opposite also holds true. The positive impact of CLEF is evident for the non-European languages of the bilingual English task where there is a constant growth in median performances for Indonesian to English as well as Indian languages (Hindi, Oromo, Telugu) to English, which seeded the FIRE evaluation campaign series.

Finally, the fact that top performances are achieved in the second or third year is an indicator of the appropriateness of the 3-year duration for a typical evaluation cycle.

Please cite this article as: N. Ferro, G. Silvello, 3.5K runs, 5K topics, 3M assessments and 70M measures: What trends in 10 years of Adhoc–*ish* CLEF? Information Processing and Management (2016), http://dx.doi.org/10.1016/j.ipm.2016.08.001

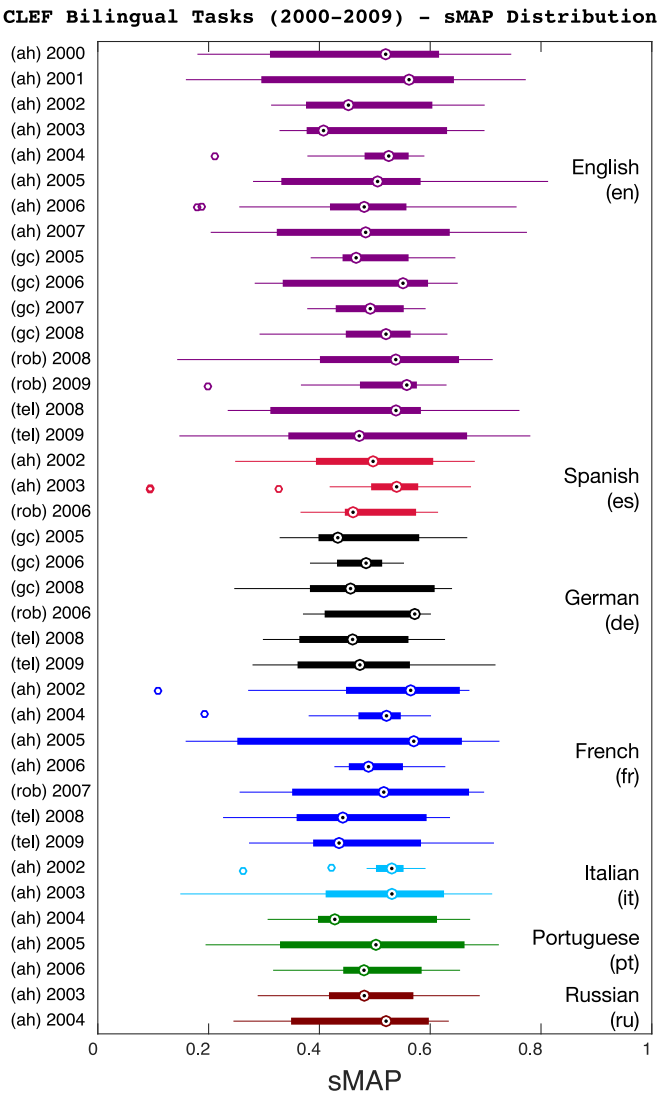

Fig. 6. sMAP distribution of the CLEF bilingual tasks 2000–2009 grouped by language/task.

5.2. Keystone cases

In this section, we dig into some relevant cases which provide a more thorough understanding of some general trends.

5.2.1. The role of source languages and language resources: the Portuguese case

The Adhoc bilingual task to Portuguese (AH X2PT) shows a similar trend to the monolingual task by improving both the median and the best sMAP from the first to the second year and then decreasing the third year. The first factor influencing the performances of bilingual systems is source language; we can see that the best sMAP in 2004 was achieved by a system using Spanish as source language as well as in 2005 – the best sMAP for Portuguese for the whole evaluation cycle – whereas in 2006 the best system used English as source language. The second factor, which is also valid for the monolingual tasks, is the research focus – i.e. investigating new techniques, consolidating state-of-the-art, fine tuning of already tested techniques – of the research groups participating in the task. In the Portuguese case in 2005 the best run was submitted by the Johns Hopkins University which used the well-established HAIRCUT retrieval system (McNamee & Mayfield, 2004), which was not specifically tuned for Portuguese and adopted language-neutral techniques relying on language similarity when linguistic resources were scarce; they highlighted the difficulties working with Portuguese queries since many query terms were discarded due to the absence of a good stemmer for this language. A solution to this problem was working with Spanish queries (for which there was much more previous research and better language resources) and perform translation which worked very well due to the similarity between the two languages. In 2005 the best sMAP was achieved by the

Please cite this article as: N. Ferro, G. Silvello, 3.5K runs, 5K topics, 3M assessments and 70M measures: What trends in 10 years of Adhoc–*ish* CLEF? Information Processing and Management (2016), http://dx.doi.org/10.1016/j.ipm.2016.08.001

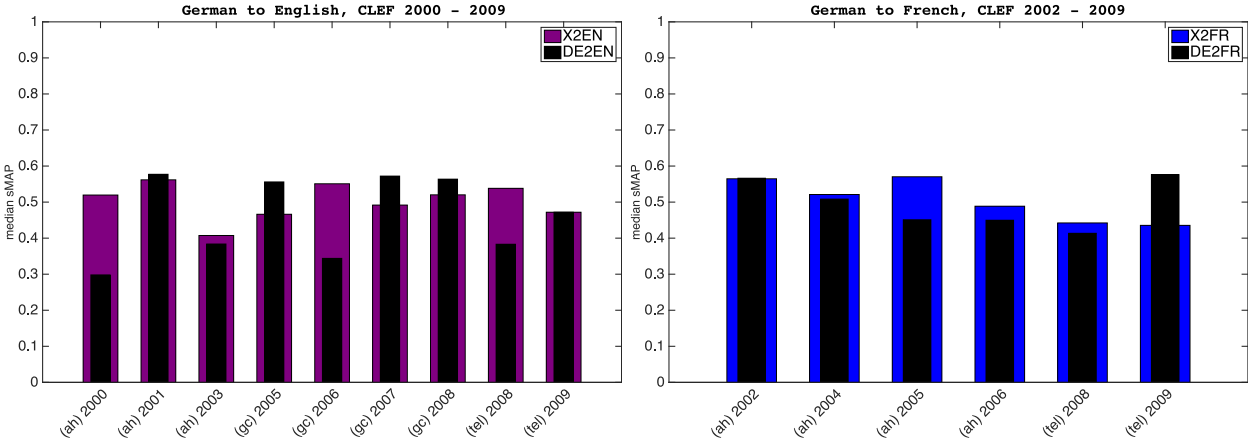

Fig. 7. Bilingual tasks performance breakdown, German (source language) to English and French (target languages).

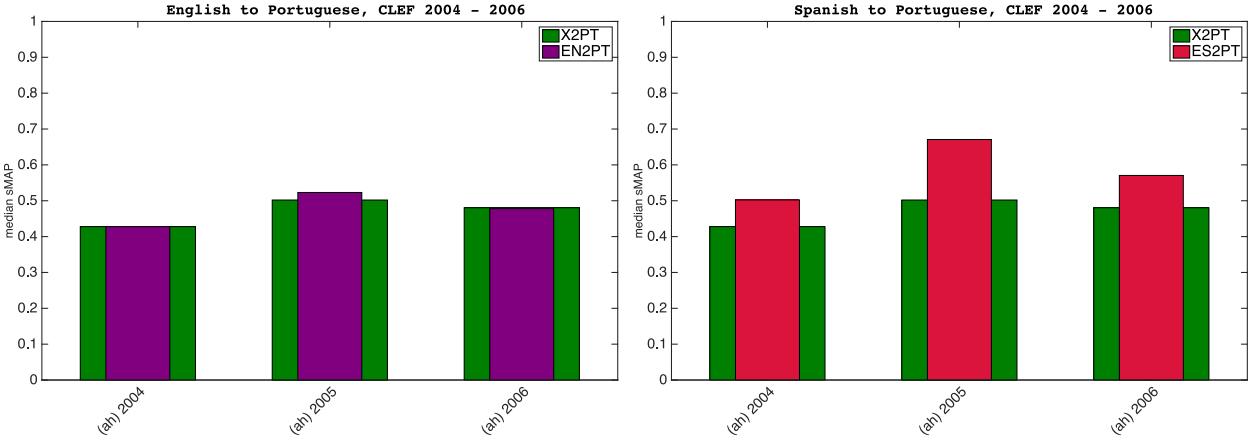

Fig. 8. Bilingual tasks performance breakdown, Spanish and English (source languages) to Portuguese (target language).

Miracle team of the University of Madrid by using Spanish as source language (González, Goñi-Menoyo, & Villena-Román, 2005); the goal of this group was to continue and consolidate already initiated research based on combining results coming from the monolingual tasks and by using reliable available translations. In 2006 the best sMAP was achieved by the University of Neuchâtel using English as source language (Savoy & Abdou, 2006); the main goal of this research group was to test the effectiveness of different machine translation techniques from English to French and Portuguese. They did not test their system by using Spanish as source language and this could explain the decrease in performances with respect to the previous year; furthermore, we can see from Table A.2 in Appendix A that the Portuguese task has seen a noteworthy turnover of participating groups.

5.2.2. Relationships between source and target European languages

In Fig. 7 we can see a bar chart contrasting the median sMAP for a bilingual task considering all the runs participating in it with respect to the median sMAP considering only the run employing a specific source language. We show how German to English (DE2EN) behaves with respect to German to French (DE2FR). We can see that DE2EN runs behave better than X2EN runs 44% of the time whereas FR2EN behaves better than X2EN only 16% of the time. This shows a correlation between the source and the target languages which favors languages with a greater affinity for each other such as German and English over language pairs such as French and English.

Similar conclusions are attained by analyzing Fig. 8 which shows the comparison between EN2PT and ES2PT; as discussed above, the affinity between Spanish and Portuguese is a factor directly influencing the performance of a bilingual retrieval system and the best performances using Portuguese as a target language have been achieved by systems using Spanish as source language. In Fig. 8 we can see that ES2PT runs outperform X2EN runs for all the tasks, whereas EN2PT runs outperforms X2EN runs only in one out of three cases.

Please cite this article as: N. Ferro, G. Silvello, 3.5K runs, 5K topics, 3M assessments and 70M measures: What trends in 10 years of Adhoc–*ish* CLEF? Information Processing and Management (2016), http://dx.doi.org/10.1016/j.ipm.2016.08.001

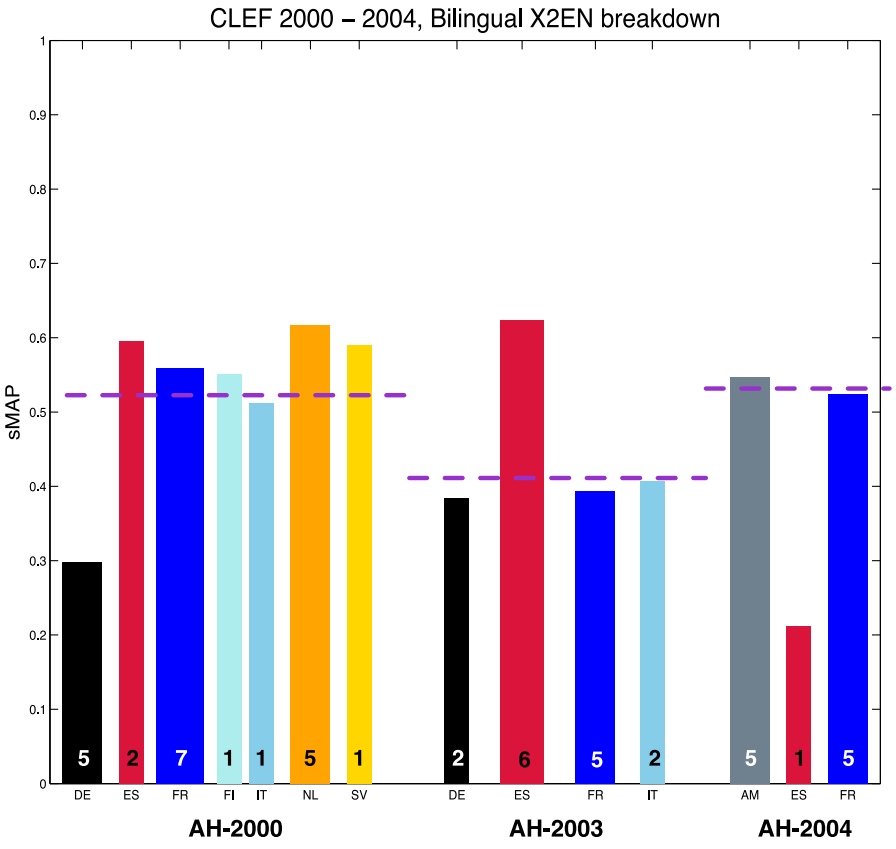

Fig. 9. AH BILI X2EN task performance breakdown for a three-years evaluation cycle. We report the median sMAP achieved by the systems working on English target language divided by the source language employed; within each single bar we report the number of runs submitted for that source language whereas thickness of each bar is weighted by this number.

5.2.3. European languages versus non-European languages to English

In Fig. 9 we can see a performance breakdown for the "X2EN" tasks where we report the median sMAP achieved by the systems working on English target language divided by the source language employed; within each single bar we report the number of runs submitted for that source language whereas thickness of each bar is weighted by this number. We report data for the tasks carried out in 2000, 2003, and 2004; we can see that in 2003 the median sMAP dropped with respect to 2000 and then it recovered in 2004. In 2003 (Braschler, 2003), only 3 groups (all newcomers) participated by submitting fewer runs than in 2000; in 2004 the median sMAP recovered, even though 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 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 in French runs, the removal of German as source language and the introduction of Amharic for which very good runs were submitted even though this language was initiated that very year.

In the upper-left part of Fig. 10 we can see the ratio between Chinese and English, which shows a constant increasing trend from 2000 to 2007. In particular, we can see the effect of NTCIR evaluation campaigns in promoting the development of novel retrieval techniques and richer linguistic resources for the Chinese which turns out in a median sMAP of ZH2EN runs to overcome X2EN runs in 2007. A similar positive trend can be seen for the Indonesian to English runs which sMAP constantly increased from 2005 to 2007 and also for the GeoCLEF task which overtook the X2EN sMAP. A clear improvement of performances is also evident in the lower-right plot of the figure where we show the contrast between X2EN median sMAP and Indian languages to English (HI2EN, OM2EN, TE2EN) that are the seed from which the FIRE campaigns began their evaluation activities. Oromo (OM) and Telugu (TE) show important growth from 2006 to 2007, whereas Hindi remains almost constant, overtaking X2EN performances in both the considered years. Lastly, in the upper-right plot we show how French to English runs behave over the years, reporting a behavior close to that mentioned for the German to French runs where DE2EN overtook X2EN only for TEL 2009 as well as here.

Please cite this article as: N. Ferro, G. Silvello, 3.5K runs, 5K topics, 3M assessments and 70M measures: What trends in 10 years of Adhoc–*ish* CLEF? Information Processing and Management (2016), http://dx.doi.org/10.1016/j.ipm.2016.08.001

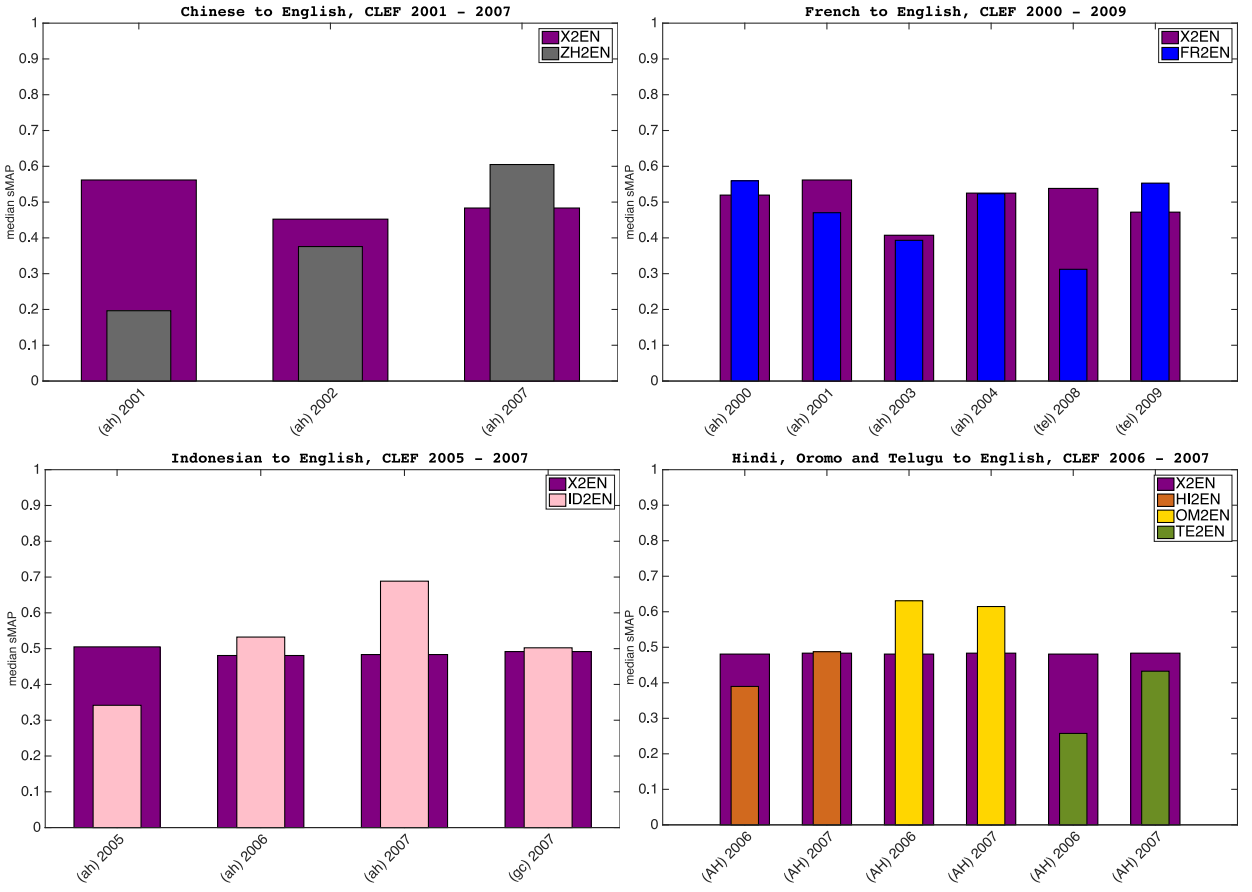

Fig. 10. Bilingual tasks performance breakdown, different source languages (Chinese, French, Indonesian, Hindi, Oromo and Telugu) to English (target language).

Table 3
Best and median sMAP of the CLEF multilingual tasks. In bold the best performances for each task over the years and in brackets the performances variation with respect to the previous year.

Task	Year	Best sMAP	Median sMAP
MULTI-4	2000	**80.51%** (–)	52.10% (–)
	2003	79.05% (−1.81%)	**55.30%** (+5.78%)
	2004	71.39% (−10.73%)	51.23% (−7.36%)
MULTI-5	2001	79.79% (–)	**54.29%** (–)
	2002	**79.82%** (0.03%)	48.32% (−12.36%)
MULTI-8	2003	**85.13%** (–)	42.77% (–)
MULTI-8 2-Y-ON	2005	84.76% (−0.44%)	**51.17%** (+16.41%)
MULTI-8 MERGING	2005	82.04% (−3.63%)	50.37% (−1.56%)

6. Multilingual tasks

Table 3 summarizes the performances achieved in the monolingual tasks in terms of best and median sMAP. This information is shown graphically in Fig. 11.

6.1. Highlights

Multilingual tasks are the most complex tasks offered at CLEF and they require the development of sophisticated and carefully tuned systems. We can see the effects of this complexity by looking at both the median and best sMAP performances for the Multi-4 and Multi-5 tasks.

Please cite this article as: N. Ferro, G. Silvello, 3.5K runs, 5K topics, 3M assessments and 70M measures: What trends in 10 years of Adhoc–*ish* CLEF? Information Processing and Management (2016), http://dx.doi.org/10.1016/j.ipm.2016.08.001

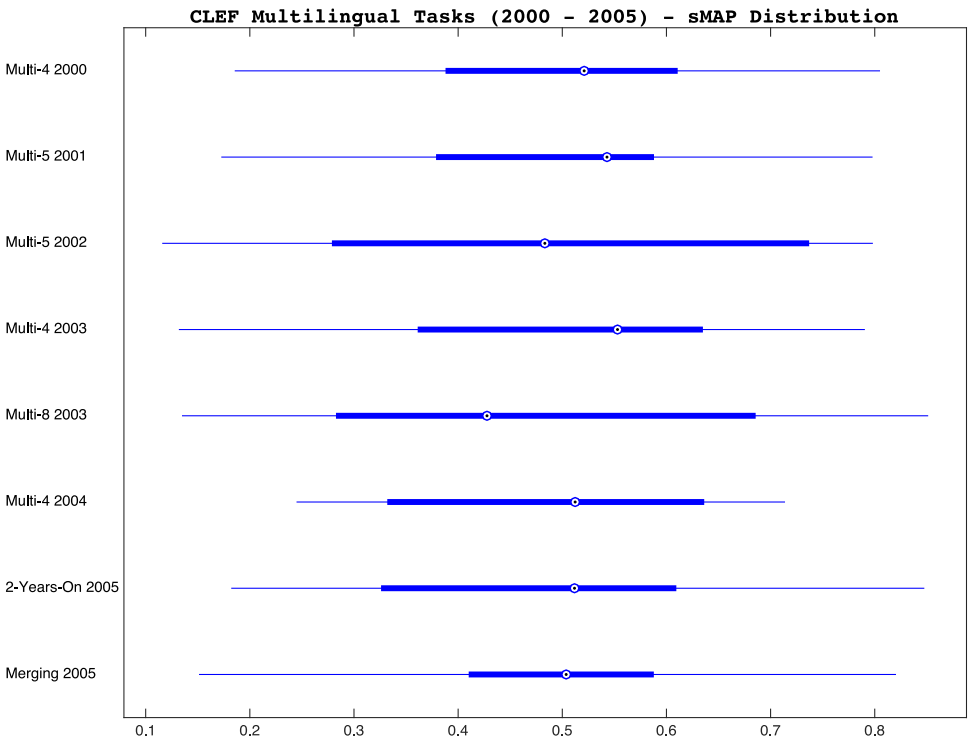

Fig. 11. Distribution of sMAP for all the CLEF multilingual tasks.

The median performances are quite stable over the years, even though the combination of source and target languages has changed from year to year. For the first year of the task, the core technology needed to address multilingualism could rely on the prior research conducted in the CLIR task carried out in TREC; in the subsequent years the median performances remained quite stable indicating that the research carried out in the CLEF monolingual and bilingual tasks produced advanced methods and language resources applicable also in the multilingual case. Over years the best performances increased reaching the peak for the Multi-8 in 2003 where several machine translation techniques and advanced language resources were successfully exploited together to address this challenging task.

In general, Multi-8 in 2003 was a much more challenging task. Indeed, the median performances of Multi-8 dropped after three years of more or less stable median performances in Multi-4 and Multi-5. This gap was bridged two years later with the Multi-8 Two Years On task, where new systems were experimented using the same collection, and the Merging task was added whereby data fusion techniques were applied for mixing the original 2003 runs; both these tasks reached performances comparable with the Multi-4 and Multi-5.

The best performances of Multi-8 in 2003 are not only the highest of all the multilingual tasks but they are also the top performances ever achieved in all the analyzed CLEF tasks. This supports the above idea that, even if it is challenging, bridging a high gap pays off and it is a clear indicator of the high quality of the developed systems.

We can find confirmation of the 3-years length for an evaluation cycles also in the case of multilingual tasks: the peak for Multi-5 was reached in 2001 and afterwards in 2003 for Multi-4 followed by a slight decline.

As recalled also in Ferro (2014), the underlying motivation for starting CLEF was the “Grand Challenge”, formulated at the *Association for the Advancement of Artificial Intelligence (AAAI)* 1997 Spring Symposium on Cross-Language and Speech Retrieval (Hull & Oard, 1997), which was calling for the development of fully multilingual and multimodal information access systems. The analyses conducted provide us with evidence that CLEF has achieved its main objective and has made multilingual IR for European languages a reality, with performances as satisfactory as or even better than monolingual ones.

6.2. Keystone cases: multi-8 improvements over the years

Standardization allows us to reconsider an important result reported in Di Nunzio, Ferro, Jones, and Peters (2006b) 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. Di Nunzio et al. (2006b) reported a 15.89% increment of performances for the top system of 2005 with respect to the top system of 2003; this finding showed an improvement in multilingual IR systems from 2003 to 2005. Nevertheless, analyzing sMAP we draw a similar conclusion, but from a different perspective; indeed, the top system in 2003 achieved 0.8513 sMAP (i.e. University

Please cite this article as: N. Ferro, G. Silvello, 3.5K runs, 5K topics, 3M assessments and 70M measures: What trends in 10 years of Adhoc–*ish* CLEF? Information Processing and Management (2016), http://dx.doi.org/10.1016/j.ipm.2016.08.001

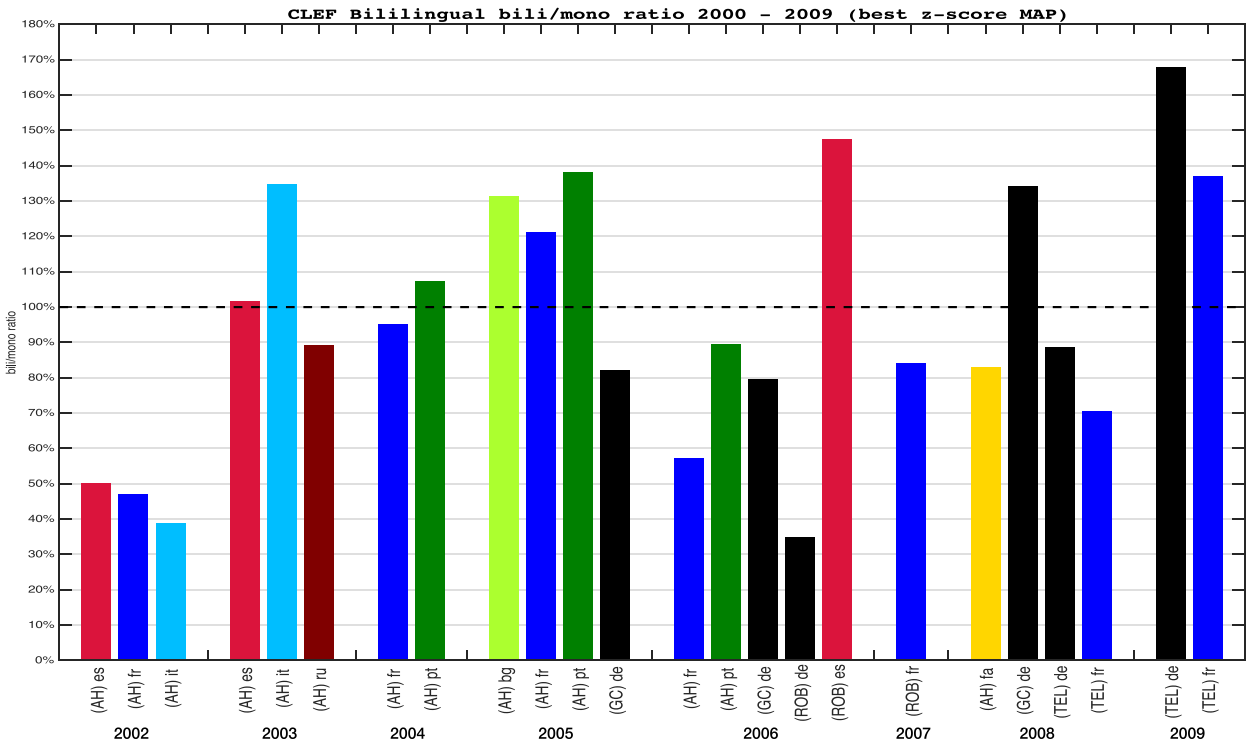

Fig. 12. Mono/bili best z-score MAP comparison.

of Neuchâtel Savoy, 2004b), whereas the top system in 2005 achieved 0.8476 sMAP (i.e. Carnegie Mellon University Si & Callan, 2006), 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 increment of 16.41%; this result is even stronger than the findings reported in Di Nunzio et al. (2006b), since it shows that half of the participating systems in 2005 improved with respect to those in 2003.

7. Bilingual to monolingual performances

A commonly used approach to compare bilingual to monolingual performances is to compute the ratio between the performances of the best bilingual system towards a given target language with respect to the best monolingual system in the same target language; this is usually done comparing the best bilingual MAP to the best monolingual one.

As discussed in Section 2, the transformation of the z-scores by the cumulative density function of the standard normal distribution in Eq. (2) reduces the effect of outlier data points making them closer to each other. In order to highlight as much as possible the differences between top performing systems in monolingual and bilingual tasks, in this case we compute the bilingual to monolingual ratios between top systems using the z-scores given by Eq. (1).

Fig. 12 shows the ratio of the z-score MAP of the best bilingual system with respect to the best monolingual one for the same target language across different tasks and over several years.

A first insight from Fig. 12 is the presence of several languages for which bilingual systems perform as good as or even quite better than the monolingual ones, as for example in the case of Bulgarian (about 130%), French (about 135%), German (between 135% and 170%), Italian (about 135%), Portuguese (about 140%), and Spanish (about 150%).

Moreover, we can see clear improvements over the years, for example Spanish spans from a ratio of about 50% in the Adhoc 2002 to a ratio of about 150% in the Robust 2006, which means almost 3 times the initial performances. Other languages exhibit similar behavior: French started at a nearly 50% ratio in the Adhoc 2002 and climbed to slightly more than 120% in the Adhoc 2005 and about 135% in the TEL 2009, which means around 2.5 times the initial performances; Italian moved from about 40% in the Adhoc 2002 to around 135% the next year, with a more than 3-fold improvement; German progressed from 80% in GeoCLEF 2005 to 135% in GeoCLEF 2008 and almost 170% in TEL 2009, which is a 1.5–2-fold increase in the initial performances.

We can find evidence also in this case of the appropriateness of 3-year evaluation cycles: for example, in the case of Portuguese, the bilingual to monolingual ratio improved for two subsequent years (2004 and 2005) and the dropped the next year, probably indicating a shift of interest of participants from top performances to other issues for Portuguese. A similar trend can also be seen for French from 2002 to 2006.

Please cite this article as: N. Ferro, G. Silvello, 3.5K runs, 5K topics, 3M assessments and 70M measures: What trends in 10 years of Adhoc–*ish* CLEF? Information Processing and Management (2016), http://dx.doi.org/10.1016/j.ipm.2016.08.001

Finally, as in the multilingual case, these findings support the case that CLEF has achieved its main objective also regarding CLIR with European languages since it has contributed to the development of bilingual systems that are as good as or even better than monolingual ones.

8. Conclusions

In this paper we conducted the first systematic and large-scale longitudinal analysis on 10 years of Adhoc-*ish* tasks at CLEF. The main goal of the paper was to identify and analyze performance trends over the years for different tasks and in different languages. We conducted the study on the Adhoc, GeoCLEF, Robust and TEL labs which are in turn divided into three main tasks, namely monolingual, bilingual and multilingual.

Since it is not possible to directly make comparisons over years and collections, we used score standardization (Webber et al., 2008), which makes performance figures interpretable on their own thus allowing for inter-collection system comparison. Score standardization has its own limitations: (i) it mainly compensates for topic effects rather than collection effects; (ii) it requires a minimum number of runs to provide reliable results; and (iii) it tends to flatten out data, performing a monotone transformation which preserves the ordering among systems. Overall, these limitations did not hamper the findings of the present study. Regarding (i), the analyzed tasks rely on topics with the same informational intents and use very comparable corpora, which are also shared across different tasks; this greatly reduces the collections effects and make score standardization appropriate. Regarding (ii), we analyzed only those tasks for which there were enough runs; even if this caused a slight sparsity of the data, it was still possible to clearly detect relevant trends. Finally, regarding (iii), the main goal of this work has been detecting performance trends rather than pointing out the largest performance gap possible.

Coming back to our initial research questions, reported here for convenience, we can sum up some lessons learned.

RQ1 What performance trends can we observe for monolingual systems over the years? What is the influence of language resources?
RQ2 What performance trends can we observe for bilingual systems over the years? What is the influence of source languages?
RQ3 What performance trends can we observe for multilingual systems over the years?
RQ4 What is the relationship between the performances of monolingual systems and those of bilingual systems?
RQ5 Is the typical 3-year duration of an evaluation task enough to improve the participating systems?

Regarding RQ1–RQ3, we observed similar improvement trends in monolingual, bilingual and multilingual tasks, basically due to two main benefits of repeated evaluation cycles: firstly, support for the development of better IR systems through carefully designed evaluation tasks; secondly, motivation for the creation of better language resources, which are crucial in the CLEF context.

We have seen that for monolingual tasks the performance trends can be divided into two main recurring patterns, one applicable for already well-resourced languages and the other for less-resourced languages. In the first case, median performances increase in the early years of the task, reach a plateau and then decrease or tend to oscillate. In the second case, it is possible to identify a steady growing trend in median performances over the years.

For bilingual tasks the highest performances are achieved in the second or third year of the tasks, with the improvement of language resources being decisive for increasing system effectiveness. In this context, we have seen that variations in system performances are correlated with specific source and target language pairs; pairs of languages with greater affinity obtain higher performances than languages with less affinity.

An interesting case is the Multi-8 task, which achieved both the lowest median sMAP and the highest sMAP among all the multilingual tasks. This shows that multilingual tasks are more difficult as the number of target languages increases, although broad availability of good and heterogeneous linguistic resources can contribute to developing very effective retrieval systems for multilingual retrieval. Moreover, the performances achieved in the Multi-8 task are the highest among those of all tasks analyzed in this study, which indicates the high quality of the developed systems.

One important aspect to be considered is the turnover between expert groups and newcomers. This is an especially important and valuable feature of large-scale evaluation campaigns, which act as catalysts for the creation of large and multidisciplinary communities where competencies are transferred from one group to the other and cross-fertilization among expert and new groups produces extremely positive effects.

The positive effect of evaluation campaigns is visible also outside CLEF boundaries. This is evident in the Chinese to English task run in 2001, 2002 and 2007; there was a noteworthy improvement in median performances from 2002 to 2007 even though there was no Chinese task in CLEF. This shows that the work done in NTCIR for the Chinese language between 2002 and 2006 produced a sizable improvement in retrieval techniques for this language.

Regarding RQ4, we learned that there are several bilingual systems that outperform monolingual ones and that over the years a 1.5–3-fold improvement of the bilingual system performances with respect to the monolingual ones is possible. These improvements are basically due to two factors both prompted by repeated evaluation cycles: one is the advancement of techniques and methods for bilingual retrieval and the other is the creation of better language resources.

Finally, regarding RQ5, we found evidence that a 3-year evaluation cycle for a task is typically enough for developing and improving new systems and solutions while a longer duration leaves room for participants to shift their interest from the initial goals of the task to other relevant research issues.

Please cite this article as: N. Ferro, G. Silvello, 3.5K runs, 5K topics, 3M assessments and 70M measures: What trends in 10 years of Adhoc–*ish* CLEF? Information Processing and Management (2016), http://dx.doi.org/10.1016/j.ipm.2016.08.001

Overall, this study helps also to deal with the problem of the definition of well-known baselines to compare against and the reproducibility of experimental results in IR (Armstrong et al., 2009b; Carterette, 2015; Ferro & Silvello, 2015). In particular, the "Open-Source IR Reproducibility Challenge" at the SIGIR 2015 RIGOR workshop (Arguello, Crane, Diaz, Lin, & Trotman, 2015), which provided reproducible baselines of several open-source IR systems in a common environment, experimented on many CLEF Adhoc-*ish* monolingual tasks and this paper contributes to putting these experimental results in the broader perspective of 10 years of CLEF evaluation campaigns.

As explained in Ferro (2014), two main periods can be identified in CLEF since its inception in 2000: the "classic" period, from 2000 to 2009, and the "CLEF Initiative" period, which started in 2010 and is currently ongoing.

The "classic" period was driven by the "Grand Challenge" formulated at the AAAI 1997 Spring Symposium on Cross-Language and Speech Retrieval (Hull & Oard, 1997), the ambitious goal of which was to develop fully multilingual and multimodal information access systems (Gey, Kando, & Peters, 2005). Overall, this study has provided quantitative evidence for supporting the claim that CLEF "classic" has fully achieved this bold objective by delivering truly bilingual and multilingual systems for European languages with performances that are as satisfactory as, or even better than, the monolingual ones.

The "CLEF Initiative" period maintains the core interest on multilingualism but with an increased focus on the *multimodal* aspect, intended not only as the ability to deal with information coming in multiple media but also in different modalities, e.g. the Web, social media, news streams, specific domains and so on. These different modalities should ideally be addressed in an integrated way; rather than building vertical search systems for each domain/modality, the interaction between the different modalities, languages and user tasks needs to be exploited to provide comprehensive and aggregated search systems. We hope that after the end of this new period for CLEF a new longitudinal study will be able to show that it has achieved its objectives as successfully as CLEF "classic" did.

Acknowledgments

The authors would like to express their gratitude to Carol Peters – who has initiated CLEF and coordinated it during the "CLEF Classic" period reported in this paper – for having shared her know-how and experience in running large-scale evaluation activities. Martin Braschler and Giorgio Maria Di Nunzio deserve special thanks for the time spent together in co-organizing several of the CLEF Adhoc-*ish* tasks. Maristella Agosti merits all our appreciation for having always pursued CLEF and evaluation as a core activity of our research group.

Appendix A. Detailed Information about the Analyzed CLEF Adhoc-*ish* Tasks

In Table A.1 we report the details about the monolingual tasks we consider in this study, in Table A.2 information about the bilingual tasks and in Table A.3 information about the multilingual tasks.

As explained in Section 3, we analyze tasks for which at least 9 valid runs have been submitted with the only exception of some Robust tasks and one TEL bilingual task, where there are less than 9 runs but at least 5.

In each table we report: used corpora and number of documents, described in Section 3.1; number of topics, described in Section 3.2; size of the pool, described in Section 3.3; number of participating groups with new groups within brackets; number of submitted runs. Languages are expressed as ISO 639:1 two letters code. In the case of bilingual and multilingual tasks we report also the source and target languages. In the following tables, AH stands for Adhoc; GC for GeoCLEF, ROB for Robust; TEL for "The European Library"

Table A.1
Analyzed CLEF monolingual tasks: used corpora; number of documents; number of topics; size of the pool; number of participating groups with new groups within brackets; number of submitted runs. Languages are expressed as ISO 639:1 two letters code. AH stands for Adhoc; GC for GeoCLEF, ROB for Robust; TEL for "The European Library".

Task	Year	Corpora	Docs	Topics	Pool	Groups	Runs
AH mono BG	2005	SEGA 2002 STANDART 2002		49	20,130	7 (–)	20
	2006		69,195	50	17,308	4 (2)	11
	2007			50	19,441	8 (5)	16
AH mono DE	2000	FRANKFURTER 1994	139,715	49	11,335	11 (–)	22
	2001	FRANKFURTER 1994		49	16,726	12 (9)	22
	2002	SDA 1994	225,371	50	19,394	12 (5)	28
	2003	SPIEGEL 1994 & 1995		57	21,534	16 (8)	38
GC mono DE	2005	FRANKFURTER 1994		21	15,664	6 (–)	20
	2006	SDA 1994	294,809	25	14,094	4 (2)	11
	2007	SDA 1995		24	14,863	4 (2)	16
	2008	SPIEGEL 1994 & 1995		25	15,079	3 (–)	9

(*continued on next page*)

Table A.1 (*continued*)

Task	Year	Corpora	Docs	Topics	Pool	Groups	Runs
ROB mono DE	2006	FRANKFURTER 1994					
		SDA 1995	223,132	95	35,859	3 (–)	7
		SPIEGEL 1994 & 1995					
TEL mono DE	2008	TELONB	869,353	50	28,734	10 (–)	27
	2009			50	25,441	9 (4)	34
AH mono ES	2001	EFE 1994	215,738	49	14,268	10 (–)	22
	2002			50	19,668	13 (5)	28
	2003	EFE 1994 & 1995	454,045	57	23,822	16 (8)	38
ROB mono ES	2006			97	36,970	5 (–)	11
AH mono FA	2008	HAMSHAHRI 2002	166,744	50	26,814	8 (–)	53
	2009			50	23,536	4 (1)	15
AH mono FI	2002	AMULEHTI 1994 & 1995	55,344	30	9825	7 (–)	11
	2003			45	10,803	7 (3)	13
	2004			45	20,124	11 (6)	30
AH mono FR	2000	LEMONDE 1994	44,013	34	7003	9 (–)	10
	2001	LEMONDE 1994 ATS 1994	87,191	49	12,263	9 (6)	15
	2002			50	17,465	12 (7)	16
	2003	LEMONDE 1994 ATS 1994 & 1995	129,806	52	16,785	16 (9)	35
	2004	LEMONDE 1995 ATS 1995	90,261	49	23,541	13 (4)	38
	2005	LEMONDE 1994 & 1995 ATS 1994 & 1995	177,452	50	23,999	12 (7)	38
	2006			49	17,882	8 (5)	27
ROB mono FR	2006	LEMONDE 1994 ATS 1994 & 1995	129,806	97	28,227	7 (–)	18
	2007			93	20,445	5 (1)	11
TEL mono FR	2008	TELBNF	1,000,100	50	24,530	9 (–)	17
	2009			50	21,971	9 (5)	23
AH mono HU	2005	MAGYAR 2002		50	20,561	10 (–)	30
	2006		49,530	48	20,435	6 (3)	17
	2007			50	18,704	6 (3)	19
AH mono IT	2000	AGZ 1994 LASTAMPA 1994	108,578	34	6760	9 (–)	10
	2001			47	10,697	8 (5)	14
	2002			49	17,822	14 (7)	25
	2003	AGZ 1994 & 1995 LASTAMPA 1994	157,558	51	20,902	13 (4)	27
ROB mono IT	2006			90	34,812	5 (–)	8
AH mono NL	2001	ALGEMEEN 1994 & 1995 NRC 1994 & 1995	190,604	50	16,774	9 (–)	18
	2002			50	20,957	11 (4)	19
	2003			50	20,332	11 (4)	32
ROB mono NL	2006			96	36,746	3 (–)	7
AH mono PT	2004	PUBLICO 1994 & 1995	106,821	46	20,103	8 (–)	22
	2005	FOLHA 1994 & 1995 PUBLICO 1994 & 1995	210,734	50	20,539	9 (3)	32
	2006			50	20,154	12 (8)	34
GC mono PT	2006			25	15,145	3 (–)	11
	2007			25	15,572	3 (2)	10
	2008			25	14,780	3 (–)	17
ROB mono PT	2006			96	31,593	4 (–)	17

Table A.2
Analyzed CLEF bilingual tasks: used corpora; number of documents; source languages of the topics; number of topics; size of the pool; number of participating groups with new groups within brackets; number of submitted runs. Languages are expressed as ISO 639:1 two letters code. AH stands for Adhoc; GC for GeoCLEF, ROB for Robust; TEL for "The European Library".

Task	Year	Corpora	Docs	Sources	Topics	Pool	Groups	Runs
AH bili X2DE	2002	FRANKFURTER 1994		en, fr, ru	50	19,394	6 (–)	13
GC bili X2DE	2005	SDA 1994	225,371	en, ru	21	15,664	3 (–)	17
	2006	SPIEGEL 1994		en, es, pt	25	14,094	3 (2)	10
	2008	SPIEGEL 1995		en, pt	25	15,079	3 (1)	17
ROB bili X2DE	2006	FRANKFURTER 1994						
		SDA 1995	223,132	en	95	35,859	3 (–)	5
		SPIEGEL 1994 & 1995						
TEL bili X2DE	2008	TELONB	869,353	en, es, fr	50	28,734	6 (–)	17
	2009			en, fr, it, zh	50	25,441	6 (3)	26
AH bili X2EN	2000	LATIMES 1994	113,005	de, es, fi, fr, it, nl, sv	33	11,999	10 (–)	26
	2001			de, es, fi, fr, it, ja, ru, sv, th, zh	47	23,290	19 (15)	55
	2002			es, fr, nl, pt, zh	42	17,888	5 (3)	16
	2003	LATIMES 1994 GLASGOW 1995	169,477	de, es, fr, it	54	21,317	3 (3)	15
	2004			am, es, fr	42	16,651	4 (4)	11
	2005			el, en, id, hu, ru	50	19,790	8 (8)	31
	2006			am, id, it, hi, om, te	49	22,582	5 (4)	32
	2007	LATIMES 2002	135,153	am, bn, hi, hu, id, mr, om, ta, te, zh	50	24,855	10 (9)	67

(*continued on next page*)

Please cite this article as: N. Ferro, G. Silvello, 3.5K runs, 5K topics, 3M assessments and 70M measures: What trends in 10 years of Adhoc–*ish* CLEF? Information Processing and Management (2016), http://dx.doi.org/10.1016/j.ipm.2016.08.001

Table A.2 (*continued*)

Task	Year	Corpora	Docs	Sources	Topics	Pool	Groups	Runs
ROB bili X2EN	2008	LATIMES 1994 GLASGOW 1995	169,477	es	153	63,689	4 (–)	5
	2009			es	153	63,689	5 (1)	9
TEL bili X2EN	2008	TELBL	1,000,100	de, es, fr, nl	153	28,104	8 (–)	5
	2009			de, el, fr, it, zh	153	26,190	10 (7)	9
AH bili X2ES	2002	EFE 1994	215,738	de, en, fr, pt, it	50	19,668	7 (–)	16
	2003	EFE 1994 & 1995	454,045	it	57	23,822	9 (7)	15
ROB bili ES	2006			it	97	36,970	3 (–)	8
AH bili X2FR	2002	LEMONDE 1994 ATS 1994	87,191	de, en, ru	49	17,475	7 (–)	14
	2004	LEMONDE 1995 ATS 1995	90,261	de, nl, sv	49	23,541	7 (5)	24
	2005	LEMONDE 1994 & 1995 ATS 1994 & 1995	177,452	am, de, en, es, it, ru	50	23,999	9 (8)	31
	2006			de, en, es	49	17,882	4 (3)	12
ROB bili X2FR	2007	LEMONDE 1994 ATS 1994 & 1995	129,806	en, es, pt	97	20,445	3 (–)	18
TEL bili X2FR	2008	TELBNF	1,000,100	de, en, es, nl	50	24,530	5 (–)	15
	2009			de, en, it	50	21,971	6 (4)	23
AH bili X2IT	2002	AGZ 1994 LASTAMPA 1994	108,578	de, en, es, fr	49	17,822	6 (–)	13
	2003	AGZ 1994 & 1995 LASTAMPA 1994	157,558	de	50	20,902	8 (5)	21
AH bili X2PT	2004	PUBLICO 1994 & 1995	106,821	en, es	46	20,103	4 (–)	15
	2005	FOLHA 1994 & 1995 PUBLICO 1994 & 1995	210,734	en, es, fr	50	20,539	8 (5)	24
	2006			en, es, fr	50	20,154	6 (4)	22
AH bili X2RU	2003	IZVESTIA 1995	16,716	de, en	28	11,042	2 (–)	9
	2004			bg, en, es, fr, ja, zh	34	16,816	8 (7)	26

Table A.3
Analyzed CLEF multilingual tasks: used corpora; number of documents; source languages of the topics employed by the submitted runs; target languages of the documents; number of topics; size of the pool; number of participating groups with new groups within brackets; number of submitted runs. Languages are expressed as ISO 639:1 two letters code.

Task	Year	Corpora	Docs	Sources	Targets	Topics	Pool	Groups	Runs
Multi-4	2000	FRANKFURTER 1994 (de) LATIMES 1994 (en) LEMONDE 1994 (fr) LASTAMPA 1994 (it)	354,784	de, en, nl	de, en, fr, it	40	43,566	11 (–)	26
Multi-5	2001	FRANKFURTER 1994 (de) SDA 1994 (de) SPIEGEL 1994 & 1995 (de) LATIMES 1994 (en) EFE 1994 (es) ATS 1994 (fr) LEMONDE 1994 (fr) AGZ 1994 (it) LASTAMPA 1994 (it)	749,883	de, en, nl, ru, zh	de, en, es, fr, it	50	80,624	8 (4)	26
Multi-5	2002	FRANKFURTER 1994 (de) SDA 1994 (de) SPIEGEL1994 & 1995 (de) LATIMES 1994 (en) EFE 1994 (es) ATS 1994 (fr) LEMONDE 1994 (fr) AGZ 1994 (it) LASTAMPA1994 (it)	749,883	de, en	de, en, es, fr, it	50	96,420	11 (6)	36
Multi-4	2003	FRANKFURTER 1994 (de) SDA 1994 & 1995 (de) SPIEGEL 1994 & 1995 (de) GLASGOW 1995 (en) LATIMES 1994 (en) EFE 1994 & 1995 (es) ATS 1994 & 1995 (fr) LEMONDE 1994 (fr)	1,048,137	de, en, es, fr	de, en, es, fr	60	92,808	14 (10)	52
Multi-8	2003	FRANKFURTER 1994 (de) SDA 1994 & 1995 (de) SPIEGEL 1994 & 1995 (de) GLASGOW 1995 (en) LATIMES 1994 (en) EFE 1994 & 1995 (es) AMULEHTI 1994 & 1995 (fi) ATS 1994 & 1995 (fr) LEMONDE 1994 (fr) AGZ 1994 & 1995 (it) LA STAMPA 1994 (it) ALGEMEEN 1994 & 1995 (nl) NRC 1994 & 1995 (nl) TT 1994 & 1995 (sv)	1,451,643	en, es	de, en, es, fi, fr, it, nl, sv	60	173,406	7 (3)	33
Multi-4	2004	FRANKFURTER 1994 (de) SDA 1994 & 1995 (de) SPIEGEL 1994 & 1995 (de) GLASGOW 1995 (en) LATIMES 1994 (en) EFE 1994 & 1995 (es) ATS 1994 & 1995 (fr) LEMONDE 1994 (fr)	1,048,137	en, fr	de, en, es, fr	60	92,035	9 (3)	35

(*continued on next page*)

Please cite this article as: N. Ferro, G. Silvello, 3.5K runs, 5K topics, 3M assessments and 70M measures: What trends in 10 years of Adhoc–*ish* CLEF? Information Processing and Management (2016), http://dx.doi.org/10.1016/j.ipm.2016.08.001

Table A.3 (*continued*)

Task	Year	Corpora	Docs	Sources	Targets	Topics	Pool	Groups	Runs
Multi-8 Two Years On	2005	FRANKFURTER 1994 (de) SDA 1994 & 1995 (de) SPIEGEL 1994 & 1995 (de) GLASGOW 1995 (en) LATIMES 1994 (en) EFE 1994 & 1995 (es) AMULEHTI 1994 & 1995 (fi) ATS 1994 & 1995 (fr) LEMONDE 1994 (fr) AGZ 1994 & 1995 (it) LA STAMPA 1994 (it) ALGEMEEN 1994 & 1995 (nl) NRC 1994 & 1995 (nl) TT 1994 & 1995 (sv)	1,451,643	en, es	de, en, es, fi, fr, it, nl, sv	60	173,406	4	21
Multi-8 Merging	2005	FRANKFURTER 1994 (de) SDA 1994 & 1995 (de) SPIEGEL 1994 & 1995 (de) GLASGOW 1995 (en) LATIMES 1994 (en) EFE 1994 & 1995 (es) AMULEHTI 1994 & 1995 (fi) ATS 1994 & 1995 (fr) LEMONDE 1994 (fr) AGZ 1994 & 1995 (it) LA STAMPA 1994 (it) ALGEMEEN 1994 & 1995 (nl) NRC 1994 & 1995 (nl) TT 1994 & 1995 (sv)	1,451,643	en, es	de, en, es, fi, fr, it, nl, sv	60	173,406	3	20

References

Agirre, E., Di Nunzio, G. M., Ferro, N., Mandl, T., & Peters, C. (2008). CLEF 2008: Ad hoc track overview. In Borri, Nardi, Peters, and Ferro (2008).
Agirre, E., Di Nunzio, G. M., Ferro, N., Mandl, T., & Peters, C. (2009). CLEF 2008: Ad hoc track overview. In Peters et al. (2009).
Agirre, E., Di Nunzio, G. M., Mandl, T., & Otegi, A. (2009). CLEF 2009 ad hoc track overview: Robust – WSD task. In Borri, Nardi, Peters, and Ferro (2009).
Agirre, E., Di Nunzio, G. M., Mandl, T., & Otegi, A. (2010). CLEF 2009 ad hoc track overview: Robust-WSD task. In Peters et al. (2010).
Agosti, M., Di Buccio, E., Ferro, N., Masiero, I., Peruzzo, S., & Silvello, G. (2012). DIRECTions: Design and specification of an IR evaluation infrastructure. In T. Catarci, P. Forner, D. Hiemstra, A. Peñas, & G. Santucci (Eds.), *Information access evaluation. Multilinguality, multimodality, and visual analytics. Proceedings of the third international conference of the CLEF initiative (CLEF 2012)*. In *Lecture notes in computer science (LNCS): 7488* (pp. 88–99). Heidelberg, Germany: Springer.
Agosti, M., & Ferro, N. (2009). Towards an evaluation infrastructure for DL performance evaluation. In G. Tsakonas, & C. Papatheodorou (Eds.), *Evaluation of digital libraries: An insight into useful applications and methods* (pp. 93–120). Oxford, UK: Chandos Publishing.
Amati, G., Carpineto, C., & Romano, G. (2003). Italian monolingual information retrieval with PROSIT. In Peters, Braschler, Gonzalo, and Kluck (2003).
Amati, G., & van Rijsbergen, C. J. (2002). Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Transactions on Information Systems (TOIS), 20*(4), 357–389.
Angelini, M., Ferro, N., Larsen, B., Müller, H., Santucci, G., & Silvello, G. (2014). Measuring and analyzing the scholarly impact of experimental evaluation initiatives. In M. Agosti, T. Catarci, & F. Esposito (Eds.), *Proc. 10th italian research conference on digital libraries (IRCDL 2014): vol. 38* (pp. 133–137). Procedia Computer Science
Arguello, J., Crane, M., Diaz, F., Lin, J., & Trotman, A. (2015). Report on the SIGIR 2015 workshop on reproducibility, inexplicability, and generalizability of results (RIGOR). *SIGIR Forum, 49*(2), 107–116.
Armstrong, T. G., Moffat, A., Webber, W., & Zobel, J. (2009a). Has ad hoc retrieval improved Since 1994?. In J. Allan, J. A. Aslam, M. Sanderson, C. Zhai, & J. Zobel (Eds.), *Proc. 32nd annual international ACM SIGIR conference on research and development in information retrieval (SIGIR 2009)* (pp. 692–693). New York, USA: ACM Press.
Armstrong, T. G., Moffat, A., Webber, W., & Zobel, J. (2009b). Improvements that don't add up: Ad-hoc retrieval results since 1998. In D. W.-L. Cheung, I.-Y. Song, W. W. Chu, X. Hu, & J. J. Lin (Eds.), *Proc. 18th international conference on information and knowledge management (CIKM 2009)* (pp. 601–610). New York, USA: ACM Press.
Banks, D., Over, P., & Zhang, N.-F. (1999). Blind men and elephants: Six approaches to TREC data. *Information Retrieval, 1*, 7–34.
Borri, F., Nardi, A., Peters, C., & Ferro, N. (Eds.). (2008). *CLEF 2008 working notes*. CEUR workshop proceedings (CEUR-WS.org). ISSN 1613-0073, http://ceur-ws.org/Vol-1174/
Borri, F., Nardi, A., Peters, C., & Ferro, N. (Eds.). (2009). *CLEF 2009 working notes*. CEUR workshop proceedings (CEUR-WS.org). ISSN 1613-0073, http://ceur-ws.org/Vol-1175/
Borri, F., Peters, C., & Ferro, N. (Eds.). (2004). *CLEF 2004 working notes*. CEUR workshop proceedings (CEUR-WS.org). ISSN 1613-0073, http://ceur-ws.org/Vol-1170/
Braschler, M. (2001). CLEF 2000 – overview of results. In C. Peters (Ed.), *Cross-language information retrieval and evaluation: Workshop of cross-language evaluation forum (CLEF 2000)*. In *Lecture notes in computer science (LNCS): 2069* (pp. 89–101). Heidelberg, Germany: Springer.
Braschler, M. (2002). CLEF 2001 – overview of results. In Peters, Braschler, Gonzalo, and Kluck (2002).
Braschler, M. (2003). CLEF 2002 – overview of results. In Peters et al. (2003).
Braschler, M. (2004). CLEF 2003 – overview of results. In Peters, Braschler, Gonzalo, and Kluck (2004).
Braschler, M., Di Nunzio, G. M., Ferro, N., & Peters, C. (2005). CLEF 2004: Ad hoc track overview and results analysis. In Peters, Clough, Gonzalo, Jones, Kluck, Magnini (2005).
Braschler, M., & Peters, C. (2004). Cross-language evaluation forum: Objectives, results, achievements. *Information Retrieval, 7*(1–2), 7–31.
Braschler, M., Reitberger, S., Imhof, M., Järvelin, A., Hansen, P., Lupu, M., et al.(2012). Deliverable D2.3 – best practices report. PROMISE Network of Excellence, EU 7FP, Contract N. 258191. http://www.promise-noe.eu/documents/10156/086010bb-0d3f-46ef-946f-f0bbeef305e8.
Buckley, C. (2005). The SMART project at TREC. In Harman and Voorhees (2005).
Buckley, C., & Voorhees, E. M. (2005). Retrieval system evaluation. In Harman and Voorhees (2005).
Carterette, B. A. (2015). Bayesian inference for information retrieval evaluation. In J. Allan, W. B. Croft, A. P. de Vries, C. Zhai, N. Fuhr, & Y. Zhang (Eds.), *Proc. 1st ACM SIGIR international conference on the theory of information retrieval (ICTIR 2015)* (pp. 31–40). New York, USA: ACM Press.
Di Nunzio, G. M., Ferro, N., Jones, G. J. F., & Peters, C. (2005). CLEF 2005: Ad hoc track overview. In Peters, Quochi, and Ferro (2005).
Di Nunzio, G. M., Ferro, N., Jones, G. J. F., & Peters, C. (2006a). CLEF 2005: Ad hoc track overview. In Peters et al. (2006).
Di Nunzio, G. M., Ferro, N., Jones, G. J. F., & Peters, C. (2006b). CLEF 2005: Ad hoc track overview. In Peters et al. (2006).
Di Nunzio, G. M., Ferro, N., Mandl, T., & Peters, C. (2006). CLEF 2006: Ad hoc track overview. In Nardi, Peters, Vicedo, and Ferro (2006).

Please cite this article as: N. Ferro, G. Silvello, 3.5K runs, 5K topics, 3M assessments and 70M measures: What trends in 10 years of Adhoc–*ish* CLEF? Information Processing and Management (2016), http://dx.doi.org/10.1016/j.ipm.2016.08.001

Di Nunzio, G. M., Ferro, N., Mandl, T., & Peters, C. (2007a). CLEF 2006: Ad hoc track overview. In Peters et al. (2007).
Di Nunzio, G. M., Ferro, N., Mandl, T., & Peters, C. (2007b). CLEF 2007: Ad hoc track overview. In Nardi, Peters, and Ferro (2007).
Di Nunzio, G. M., Ferro, N., Mandl, T., & Peters, C. (2008). CLEF 2007: Ad hoc track overview. In Peters et al. (2008).
Dolamic, L., & Savoy, J. (2007). Stemming approaches for east European languages. In Nardi et al. (2007).
Ferro, N. (2014). CLEF 15th birthday: Past, present, and future. *SIGIR Forum, 48*(2), 31–55.
Ferro, N., Hanbury, A., Müller, H., & Santucci, G. (2011). Harnessing the scientific data produced by the experimental evaluation of search engines and information access systems. In M. Sato, S. Matsuoka, P. M. Sloot, G. D. van Albada, & J. Dongarra (Eds.), *Proc. 11th international conference on computational science (ICCS 2011)* (pp. 740–749). Procedia Computer Science, vol. 4
Ferro, N., & Peters, C. (2009). CLEF 2009 ad hoc track overview: TEL & Persian tasks. In Borri et al. (2009).
Ferro, N., & Peters, C. (2010). CLEF 2009 ad hoc track overview: TEL & Persian tasks. In Peters et al. (2010).
Ferro, N., & Silvello, G. (2014). CLEF 15th birthday: What can we learn from ad hoc retrieval?. In E. Kanoulas, M. Lupu, P. Clough, M. Sanderson, M. Hall, & A. Hanbury (Eds.), *Information access evaluation – multilinguality, multimodality, and interaction. Proceedings of the fifth international conference of the CLEF initiative (CLEF 2014)* In *Lecture notes in computer science (LNCS): 8685* (pp. 31–43). Heidelberg, Germany: Springer.
Ferro, N., & Silvello, G. (2015). Rank-biased precision reloaded: Reproducibility and generalization. In N. Fuhr, A. Rauber, G. Kazai, & A. Hanbury (Eds.), *Advances in information retrieval. Proc. 37th european conference on IR research (ECIR 2015)*. In *Lecture Notes in Computer Science (LNCS): 9022* (pp. 768–780). Heidelberg, Germany: Springer.
Ferro, N., & Silvello, G. (2016a). A general linear mixed models approach to study system component effects. In R. Perego, F. Sebastiani, J. Aslam, I. Ruthven, & J. Zobel (Eds.), *Proc. 39th annual international ACM SIGIR conference on research and development in information retrieval (SIGIR 2016)* (pp. 25–34). New York, USA: ACM Press.
Ferro, N., & Silvello, G. (2016b). The CLEF monolingual grid of points. In N. Fuhr, P. Quaresma, T. Gonçalves, B. Larsen, K. Balog, & C. Macdonald (Eds.), *Experimental IR meets multilinguality, multimodality, and interaction. Proceedings of the seventh international conference of the CLEF association (CLEF 2016)*. In *Lecture notes in computer science (LNCS): 9822* (pp. 13–24). Heidelberg, Germany: Springer.
Fox, E. A., & Shaw, J. (1993). Combination of multiple searches. In D. K. Harman (Ed.), *The second text retrieval conference (TREC-2)* (pp. 243–252). Washington, USA: National Institute of Standards and Technology (NIST). Special Publication 500-215
Gey, F., Larson, R., Sanderson, M., Bischoff, K., Mandl, T., Womser-Hacker, K., et al.(2006). GeoCLEF 2006: The CLEF 2006 cross-language geographic information retrieval track overview. In Nardi et al. (2006).
Gey, F., Larson, R., Sanderson, M., Bischoff, K., Mandl, T., Womser-Hacker, K., et al. (2007). GeoCLEF 2006: The CLEF 2006 cross-language geographic information retrieval track overview. In Peters et al. (2007).
Gey, F. C., Kando, N., & Peters, C. (2005). Cross-language information retrieval: The way ahead. *Information Processing & Management, 41*(3), 415–431.
Gey, F. C., Larson, R. R., Sanderson, M., Joho, H., Clough, P., & Petras, V. (2006). GeoCLEF: The CLEF 2005 cross-language geographic information retrieval track overview. In Peters et al. (2006).
González, J. C., Goñi-Menoyo, J. M., & Villena-Román, J. (2005). MIRACLE's 2005 approach to cross-lingual information retrieval. In Peters, Quochi, et al. (2005).
Harman, D. K., & Voorhees, E. M. (Eds.) (2005). *TREC. Experiment and evaluation in information retrieval*. Cambridge (MA), USA: MIT Press.
Hull, D. A., & Oard, D. W. (1997). Cross-language text and speech retrieval – papers from the AAAI spring symposium. *Technical report SS-97-05*. Association for the Advancement of Artificial Intelligence (AAAI). http://www.aaai.org/Press/Reports/Symposia/Spring/ss-97-05.php
Kharazmi, S., Scholer, F., Vallet, D., & Sanderson, M. (2016). Examining additivity and weak baselines. *ACM Transactions on Information Systems (TOIS), 34*(4), 23:1–23:18.
Kluck, M., & Womser-Hacker, C. (2002). Inside the evaluation process of the cross-language evaluation forum (CLEF): Issues of multilingual topic creation and multilingual relevance assessment. In European Language Resources Association (ELRA) (Ed.), *Proc. 3rd international language resources and evaluation conference (LREC 2002)*.
Lin, J., Crane, M., Trotman, A., Callan, J., Chattopadhyaya, I., & Foley, J. (2016). Toward reproducible baselines: The open-source IR reproducibility challenge. In N. Ferro, F. Crestani, M.-F. Moens, J. Mothe, F. Silvestri, & G. M. Di Nunzio (Eds.), *Advances in information retrieval. Proc. 38th european conference on IR research (ECIR 2016)*. In *Lecture notes in computer science (LNCS): 9626* (pp. 357–368). Heidelberg, Germany: Springer.
Mandl, T., Carvalho, P., Gey, F., Larson, R., Santos, D., Womser-Hacker, K., et al. (2008). GeoCLEF 2008: The CLEF 2008 cross-language geographic information retrieval track overview. In Borri et al. (2008).
Mandl, T., Gey, F., Di Nunzio, G. M., Ferro, N., Larson, R., Sanderson, M., et al. (2007). GeoCLEF 2007: The CLEF 2007 cross-language geographic information retrieval track overview. In Nardi et al. (2007).
Mandl, T., Gey, F., Di Nunzio, G. M., Ferro, N., Larson, R., Sanderson, M., et al. (2008). GeoCLEF 2007: The CLEF 2007 cross-language geographic information retrieval track overview. In Peters et al. (2008).
McNamee, P. (2005). Exploring new languages with HAIRCUT at CLEF 2005. In Peters, Quochi, et al. (2005).
McNamee, P., & Mayfield, J. (2004). Cross-language retrieval using HAIRCUT for CLEF 2004. In Borri, Peters, and Ferro (2004).
Montalvo, S., Martinez, R., Fresno, V., & Capilla, R. (2015). Multilingual information access on the web. *IEEE Computer, 48*(7), 73–75.
Nardi, A., Peters, C., & Ferro, N. (Eds.). (2007). *CLEF 2007 working notes CEUR workshop proceedings (CEUR-WS.org)*. ISSN 1613-0073, http://ceur-ws.org/Vol-1173/
Nardi, A., Peters, C., Vicedo, J. L., & Ferro, N. (Eds.). (2006). *CLEF 2006 working notes CEUR workshop proceedings (CEUR-WS.org)*. ISSN 1613-0073, http://ceur-ws.org/Vol-1172/
Nie, J.-Y. (2010). *Cross-language information retrieval*. USA: Morgan & Claypool Publishers.
Pal, S., Mitra, M., & Kamps, J. (2011). Evaluation effort, reliability and reusability in XML retrieval. *Journal of the American Society for Information Science and Technology (JASIST), 62*(2), 375–394.
Peters, C., Braschler, M., & Clough, P. (2011). *Multilingual information retrieval*. Germany: Springer-Verlag, Heidelberg.
Peters, C., Braschler, M., Gonzalo, J., & Kluck, M. (Eds.) (2002). Evaluation of cross-language information retrieval systems: Second workshop of the cross–language evaluation forum (CLEF 2001) Revised Papers. *Lecture notes in computer science (LNCS)*: 2406. Heidelberg, Germany: Springer.
Peters, C., Braschler, M., Gonzalo, J., & Kluck, M. (Eds.) (2003). Advances in cross-language information retrieval: Third workshop of the cross–language evaluation forum (CLEF 2002) revised papers. *Lecture notes in computer science (LNCS)*: 2785. Heidelberg, Germany: Springer.
Peters, C., Braschler, M., Gonzalo, J., & Kluck, M. (Eds.) (2004). Comparative evaluation of multilingual information access systems: Fourth workshop of the cross–language evaluation Forum (CLEF 2003) revised selected papers. *Lecture notes in computer science (LNCS)*: 3237. Heidelberg, Germany: Springer.
Peters, C., Clough, P., Gey, F. C., Karlgren, J., Magnini, B., & Oard, D. W. (Eds.) (2007). Evaluation of multilingual and multi-modal information retrieval : Seventh workshop of the cross–language evaluation forum (CLEF 2006). Revised selected papers. *Lecture notes in computer science (LNCS)*: 4730. Heidelberg, Germany: Springer.
Peters, C., Clough, P., Gonzalo, J., Jones, G. J. F., Kluck, M., & Magnini, B. (Eds.) (2005). Multilingual information access for text, speech and images: Fifth workshop of the cross–language evaluation forum (CLEF 2004) Revised Selected Papers. *Lecture notes in computer science (LNCS)*: 3491. Heidelberg, Germany: Springer.
Peters, C., Deselaers, T., Ferro, N., Gonzalo, J., Jones, G. J. F., & Kurimo, M. (Eds.) (2009). Evaluating systems for multilingual and multimodal information access: Ninth workshop of the cross–language evaluation forum (CLEF 2008). Revised selected papers. *Lecture notes in computer science (LNCS)*: 5706. Heidelberg, Germany: Springer.
Peters, C., Di Nunzio, G. M., Kurimo, M., Mandl, T., Mostefa, D., & Peñas, A. (Eds.) (2010). Multilingual information access evaluation vol. I text retrieval experiments – Tenth workshop of the cross–language evaluation forum (CLEF 2009). Revised selected papers. *Lecture notes in computer science (LNCS)*: 6241. Heidelberg, Germany: Springer.

Please cite this article as: N. Ferro, G. Silvello, 3.5K runs, 5K topics, 3M assessments and 70M measures: What trends in 10 years of Adhoc–*ish* CLEF? Information Processing and Management (2016), http://dx.doi.org/10.1016/j.ipm.2016.08.001

Peters, C., Gey, F. C., Gonzalo, J., Jones, G. J. F., Kluck, M., & Magnini, B. (Eds.) (2006). Accessing multilingual information repositories: Sixth workshop of the cross–language evaluation forum (CLEF 2005). Revised selected papers. *Lecture notes in computer science (LNCS)*: 4022. Heidelberg, Germany: Springer.

Peters, C., Jijkoun, V., Mandl, T., Müller, H., Oard, D. W., & Peñas, A. (Eds.) (2008). Advances in multilingual and multimodal information retrieval: Eighth workshop of the cross–language evaluation forum (CLEF 2007). Revised selected papers. *Lecture notes in computer science (LNCS)*: 5152. Heidelberg, Germany: Springer.

Peters, C., Quochi, V., & Ferro, N. (Eds.). (2005). *CLEF 2005 working notes CEUR workshop proceedings (CEUR-WS.org)*. ISSN 1613-0073, http://ceur-ws.org/Vol-1171/

Ponte, J. M., & Croft, W. B. (1998). A language modeling approach to information retrieval. In W. B. Croft, A. Moffat, C. J. van Rijsbergen, R. Wilkinson, & J. Zobel (Eds.), *Proc. 21st annual international ACM SIGIR conference on research and development in information retrieval (SIGIR 1998)* (pp. 275–281). New York, USA: ACM Press.

Raghavi, K. C., Chinnakotla, M., & Shrivastava, M. (2015). "Answer ka type kya he?" Learning to classify questions in code-mixed language. In *Proceedings of the 24th international conference on world wide web (WWW 2015). Companion volume. Republic and Canton of Geneva, Switzerland, 2015. International world wide web conferences steering committee* (pp. 865–870).

Robertson, S. E., & Kanoulas, E. (2012). On per-topic variance in IR evaluation. In W. Hersh, J. Callan, Y. Maarek, & M. Sanderson (Eds.), *Proc. 35th annual international ACM SIGIR conference on research and development in information retrieval (SIGIR 2012)* (pp. 891–900). New York, USA: ACM Press.

Robertson, S. E., & Spärck Jones, K. (1994). Simple, proven approaches to text retrieval. *Technical report*. University of Cambridge, Computer Laboratory. UCAM-CL-TR-356. https://www.cl.cam.ac.uk/techreports/UCAM-CL-TR-356.pdf

Robertson, S. E., Walker, S., & Beaulieu, M. (2000). Experimentation as a way of life: Okapi at TREC. *Information Processing & Management, 36*(1), 95–108.

Rowe, B. R., Wood, D. W., Link, A. L., & Simoni, D. A. (2010). Economic impact assessment of NIST's text retrieval conference (TREC) Program. RTI project number 0211875, RTI International, USA. http://trec.nist.gov/pubs/2010.economic.impact.pdf.

Sanderson, M. (2010). Test collection based evaluation of information retrieval systems. *Foundations and Trends in Information Retrieval (FnTIR), 4*(4), 247–375.

Sanderson, M., & Zobel, J. (2005). Information retrieval system evaluation: Effort, sensitivity, and reliability. 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. 162–169). New York, USA: ACM Press.

Savoy, J. (2002). Report on CLEF-2001 experiments: Effective combined query-translation approach. In Peters et al. (2002).

Savoy, J. (2003). Report on CLEF 2002 experiments: Combining multiple sources of evidence. In Peters et al. (2003).

Savoy, J. (2004a). Data fusion for effective monolingual information retrieval. In Borri et al. (2004).

Savoy, J. (2004b). Report on CLEF-2003 multilingual tracks. In Peters et al. (2004).

Savoy, J., & Abdou (2006). UniNE at CLEF 2006: Experiments with monolingual, bilingual, domain-specific and robust retrieval. In Nardi et al. (2006).

Schaüble, P., & Sheridan, P. (1997). Cross-language information retrieval (CLIR) track overview. In E. M. Voorhees, & D. K. Harman (Eds.), *The sixth Text retrieval conference (TREC-6)* (pp. 31–44). Washington, USA: National Institute of Standards and Technology (NIST). Special publication 500-240

Si, L., & Callan, J. (2006). CLEF 2005: Multilingual retrieval by combining multiple multilingual ranked lists. In Peters et al. (2006).

Silvello, G., Bordea, G., Ferro, N., Buitelaar, P., & Bogers, T. (2016). Semantic representation and enrichment of information retrieval experimental data. *International Journal on Digital Libraries (IJDL)*. doi:10.1007/s00799-016-0172-8.

Spärck Jones, K. (Eds.) (1981). *Information retrieval experiment*. London, United Kingdom: Butterworths.

Tague-Sutcliffe, J. M., & Blustein, J. (1994). A statistical analysis of the TREC-3 data. In D. K. Harman (Ed.), *The third text retrieval conference (TREC-3)* (pp. 385–398). Washington, USA: National Institute of Standards and Technology (NIST). Special publication 500-225

Tang, L.-X., Geva, S., Trotman, A., Xu, Y., & Itakura, K. Y. (2014). An evaluation framework for cross-lingual link discovery. *Information Processing & Management, 50*(1), 1–23.

Tax, N., Bockting, S., & Hiemstra, D. (2015). A cross-benchmark comparison of 87 learning to rank methods. *Information Processing & Management, 51*(6), 757–772.

Thornley, C. V., Johnson, A. C., Smeaton, A. F., & Lee, H. (2011). The scholarly impact of TRECVid (2003–2009). *Journal of the American Society for Information Science and Technology (JASIST), 62*(4), 613–627.

Tomlinson, S. (2001). Stemming evaluated in 6 languages by hummingbird searchserver™ at CLEF 2001. In Peters et al. (2002).

Tomlinson, S. (2002). Experiments in 8 European languages with hummingbird searchserver™ at CLEF 2002. In Peters et al. (2003).

Tomlinson, S. (2003). Lexical and algorithmic stemming compared for 9 European languages with hummingbird searchserver™ at CLEF 2003. In Peters et al. (2004).

Tomlinson, S. (2004). Finnish, Portuguese and Russian retrieval with hummingbird searchserver™ at CLEF 2004. In Peters et al. (2005).

Tomlinson, S. (2005). Bulgarian and hungarian experiments with hummingbird searchserver™ at CLEF 2005. In Peters et al. (2006).

Tomlinson, S. (2007). Sampling precision to depth 10000: Evaluation experiments at CLEF 2007. In Nardi et al. (2007).

Tomlinson, S. (2009). Sampling precision to depth 10000 at CLEF 2008. In Peters et al. (2009).

Tsikrika, T., Garcia Seco de Herrera, A., & Müller, H. (2011). Assessing the scholarly impact of imageCLEF. In P. Forner, J. Gonzalo, J. Kekäläinen, M. Lalmas, & M. de Rijke (Eds.), *Multilingual and multimodal information access evaluation. Proceedings of the second international conference of the cross-language evaluation forum (CLEF 2011)*. In *Lecture notes in computer science (LNCS): 6941* (pp. 95–106). Heidelberg, Germany: Springer.

Tsikrika, T., Larsen, B., Müller, H., Endrullis, S., & Rahm, E. (2013). The scholarly impact of CLEF (2000–2009). In P. Forner, H. Müller, R. Paredes, P. Rosso, & B. Stein (Eds.), *Information access evaluation meets multilinguality, multimodality, and visualization. Proceedings of the fourth international conference of the CLEF initiative (CLEF 2013)*. In *Lecture notes in computer science (LNCS): 8138* (pp. 1–12). Heidelberg, Germany: Springer.

Webber, W., Moffat, A., & Zobel, J. (2008). Score standardization for inter-collection comparison of retrieval systems. In T.-S. Chua, M.-K. Leong, D. W. Oard, & F. Sebastiani (Eds.), *Proc. 31st annual international ACM SIGIR conference on research and development in information retrieval (SIGIR 2008)* (pp. 51–58). New York, USA: ACM Press.

Please cite this article as: N. Ferro, G. Silvello, 3.5K runs, 5K topics, 3M assessments and 70M measures: What trends in 10 years of Adhoc–*ish* CLEF? Information Processing and Management (2016), http://dx.doi.org/10.1016/j.ipm.2016.08.001