CLEF 2004: Ad Hoc Track Overview and Results Analysis

Martin Braschler[1], Giorgio M. Di Nunzio[2], Nicola Ferro[2], and Carol Peters[3]

[1] Eurospider Information Technology AG, 8006 Zurich, Switzerland
martin_braschler@yahoo.com
[2] Department of Information Engineering, University of Padua, Italy
{dinunzio, ferro}@dei.unipd.it
[3] ISTI-CNR, Area di Ricerca, 56124 Pisa, Italy
carol.peters@isti.cnr.it

Abstract. We describe the objectives and organization of the CLEF 2004 ad hoc track and discuss the main characteristics of the experiments. The results are analyzed and commented and their statistical significance is investigated. The paper concludes with some observations on the impact of the CLEF campaign on the state-of-the-art in cross-language information retrieval.

1 Introduction

The first four CLEF campaigns, held from 2000 through 2003, focused heavily on the ad hoc text retrieval track. One of the main goals of CLEF has been to help participating groups to scale their systems successively to be able to tackle the ambitious problem presented in this track: that of simultaneous retrieval from documents written in many different languages. For this reason, the ad hoc track is structured in three tasks, testing systems for monolingual (querying and retrieving documents in one language), bilingual (querying in one language and retrieving documents in another language) and multilingual (querying in one language and retrieving documents in multiple languages) retrieval, thus helping groups to make the progression from simple to more complex tasks. However, as mentioned in the first paper in this volume [1], the emergence of new tracks in recent CLEF campaigns has changed that emphasis somewhat. CLEF today houses more diverse activities than ever, dealing with issues such as retrieval on semi-structured data, interactive retrieval, speech retrieval, image retrieval and question answering. As a consequence, the ad hoc track has been restructured, both in order to make room for these new activities, but more importantly also to present new challenging research questions, especially for those participants that submitted CLEF experiments in previous years.

On the one hand, the CLEF 2004 multilingual track was "trimmed" to four languages: English, Finnish, French and Russian (in 2003, participants had the choice of working with either four or eight languages). On the other hand, these languages were chosen not according to their political/economic influence or their global distribution (as was done in earlier campaigns), but with respect to their distinct

C. Peters et al. (Eds.): CLEF 2004, LNCS 3491, pp. 10 – 26, 2005.

linguistic characteristics[1]. The assumption was that simultaneous retrieval from such a diverse group of languages would pose (unexpected) new challenges, not least when weighting the languages against each other during retrieval. We felt that this shift, and the resulting omission of some "popular languages", was possible due to the good and stable test collections that had already been built in previous campaigns for the languages omitted this year. The bilingual and monolingual tasks reflected the choice of languages for multilingual with the addition of Portuguese, a new acquisition to the main CLEF multilingual comparable corpus[2].

In this paper we will describe the track setup, the evaluation methodology and the participation in the different tasks (Section 2), present the main characteristics of the experiments (Section 3), provide an analysis of the results (Section 4), and investigate their statistical significance (Section 5). The paper closes with some observations on the impact of the CLEF campaigns on the state-of-the-art in the cross-language information retrieval (CLIR) field.

2 Track Setup

The ad hoc track in CLEF adopts a corpus-based, automatic scoring method for the assessment of system performance, based on ideas first introduced in the Cranfield experiments [2] in the late 1960s. This methodology is widely employed and accepted by the information retrieval community. The test collection used consists of a set of "topics" describing information needs and a collection of documents to be searched to find those documents that satisfy the information needs. Evaluation of system performance is then done by judging the documents retrieved in response to a topic with respect to their relevance, and computing the measures recall and precision. The implications of adopting the Cranfield paradigm are discussed in detail in [3].

The distinguishing feature of CLEF is that it applies this evaluation paradigm in a multilingual setting. This means that the criteria normally adopted to create a test collection, consisting of suitable documents, sample queries and relevance assessments, have been adapted to satisfy the particular requirements of the multilingual context. All language dependent tasks such as topic creation and relevance judgment are performed in a distributed setting by native speakers. Rules are established and a tight central coordination is maintained in order to ensure consistency and coherency of topic and relevance judgment sets over the different collections, languages and tracks.

2.1 Tasks

The document collection used in the CLEF 2004 ad hoc track contains English, Finnish, French, Russian and Portuguese texts. As stated above, the multilingual task

[1] English: Germanic language, global distribution, well studied; French: Romance language, very good linguistic resources, rich morphology; Finnish: Finno-Ugric language group, little shared vocabulary with the other languages, complex morphology, few resources for CLIR; Russian: Cyrillic character set, few resources for CLIR.

[2] In CLEF 2004, the multilingual comparable corpus consisted of collections of news documents for the same time period for ten languages. See [1] for details.

solicited experiments retrieving documents from a collection containing documents in four of these languages (Portuguese excluded). Using a selected topic language, the goal for systems was to retrieve relevant documents for all languages in the collection, listing the results in a single, ranked list.

Similarly to CLEF 2003, the bilingual track imposed particular conditions on some of the source → target language pairs accepted. The aim was to encourage – where possible – experiments with language pairs for which existing bilingual resources are difficult to find. The following combinations were allowed:

- Italian/French/Spanish/Russian queries → Finnish target collection
- German/Dutch/Finnish/Swedish queries → French target collection
- Any query language → Russian target collection
- Any query language → Portuguese target collection

As always, newcomers to a CLEF cross-language task or groups using a new topic language were allowed to submit runs to the English target collection.

The monolingual track offered testing for four languages: Finnish, French, Russian and Portuguese.

2.2 Topics

For each of the above tasks, the participating systems constructed their queries (automatically or manually) from a common set of topics, created to simulate user information needs. Each topic consisted of three parts: a brief "title" statement; a one-sentence "description"; a more complex "narrative" specifying the relevance assessment criteria. For CLEF 2004, 50 such topics were produced on the basis of the contents of the five target collections and were then translated additionally into Amharic, Bulgarian, Chinese, Dutch, German, Italian, Japanese, Spanish and Swedish. As in previous years, for each task attempted, a mandatory run using the title and description fields had to be submitted. The objective is to facilitate comparison between the results of different systems. Here below we give the English version of a typical topic from CLEF 2004:

```
<top>
<num> C217 </num>
<EN-title> AIDS in Africa </EN-title>
<EN-desc> Find documents discussing the increase of AIDS in Africa.</EN-
desc>
<EN-narr> There has been an explosive increase of AIDS in Africa.
Relevant documents will discuss this problem. Of particular interest are
documents mentioning humanitarian organisations fighting AIDS in Africa.
</EN-narr>
</top>
```

The motivation behind using structured topics is to simulate query input for a range of different IR applications, ranging from very short ("title" field) to elaborate query formulations ("description" and "narrative" fields), and representing keyword-style input as well as natural language formulations. The latter potentially allows sophisticated systems to make use of morphological analysis, parsing, query expansion and similar features. In the cross-language context, the transfer component

must also be considered, whether dictionary or corpus-based, a fully-fledged MT system or other. Different query structures may be more appropriate for testing one or another approach.

2.3 Relevance Assessment

Relevance assessment was performed by native speakers. The practice of assessing the results on the basis of the longest, most elaborate formulation of the topic (the narrative) means that only using shorter formulations (title and/or description) implicitly assumes a particular interpretation of the user's information need that is not (explicitly) contained in the actual query that is run in the experiment. The fact that such additional interpretations are possible has influence only on the absolute values of the evaluation measures, which in general are inherently difficult to interpret. However, comparative results across systems are usually stable regardless of different interpretations. These considerations are important when using the topics to construct very short queries to evaluate a system in a web-style scenario.

The number of documents in large test collections such as CLEF makes it impractical to judge every document for relevance. Instead approximate recall values are calculated using pooling techniques. The results submitted by the participating groups were used to form a pool of documents for each topic and language by collecting the highly ranked documents from all submissions. This pool was used for subsequent relevance judgment. After calculating the effectiveness measures, the results were analyzed and run statistics produced and distributed. A discussion of the results is given in Section 4. The individual results for all official ad hoc experiments in CLEF 2004 can be found on the CLEF website in the CLEF 2004 Working Notes [4]. The stability of pools constructed in this way and their reliability for post-campaign experiments is discussed in [5] with respect to the CLEF 2003 pools.

2.4 Participation Guidelines

To carry out the retrieval tasks of the CLEF campaign, systems have to build supporting data structures. Allowable data structures include any new structures built automatically (such as inverted files, thesauri, conceptual networks, etc.) or manually (such as thesauri, synonym lists, knowledge bases, rules, etc.) from the documents. They may not, however, be modified in response to the topics, e.g. by adding topic words that are not already in the dictionaries used by their systems in order to extend coverage.

Some CLEF data collections contain manually assigned, controlled or uncontrolled index terms. The use of such terms has been limited to specific experiments that have to be declared as "manual" runs.

Topics can be converted into queries that a system can execute in many different ways. Participants submitting more than one set of results have used both different query construction methods and variants within the same method. CLEF strongly encourages groups to determine what constitutes a base run for their experiments and to include these runs (officially or unofficially) to allow useful interpretations of the results. Unofficial runs are those not submitted to CLEF but evaluated using the trec_eval package available from Cornell University[3].

[3] See ftp://ftp.cs.cornell.edu/pub/smart/

As a consequence of limited evaluation resources, a maximum of 5 runs for each multilingual task and a maximum of 10 runs overall for the bilingual tasks, including all language combinations, was accepted. The number of runs for the monolingual task was limited to 12 runs. No more than 4 runs were allowed for any individual language combination. Overall, participants were allowed to submit at most 25 runs in total for the multilingual, bilingual and monolingual tasks (higher if other tasks were attempted).

2.5 Result Calculation

The effectiveness of IR systems can be objectively evaluated by an analysis of a representative set of sample search results. For this, effectiveness measures are calculated based on the results submitted by the participant and the relevance assessments. Popular measures usually adopted for exercises of this type are Recall and Precision. Details on how they are calculated for CLEF are given in [6].

2.6 Participants and Experiments

As shown in Table 1, a total of 26 groups from 14 different countries submitted results for one or more of the ad hoc tasks. A total of 250 experiments were submitted, 40% less than in 2003 due to the reduction in size of the track plus the expansion of other tracks offered by CLEF 2004.

Table 1. CLEF 2004 ad hoc participants

CEA/LIC2M (FR) *	UC Berkeley (US) ****
CLIPS-IMAG/IPAL-CNRS (FR/SG) *	U Chicago (US) *
Daedalus/Madrid Universities (ES) *	U Evora (PT)
Dublin City U. (IE) *** (before as U.Exeter)	U Glasgow (UK) *
Hummingbird (CA) ***	U. Hagen (DE) *
IRIT-Toulouse (FR) ***	U Hildesheim (DE) **
Johns Hopkins U./APL (US) ****	U Jaen (ES) ***
Nat. Research Council - ILTG (CA)	U. Lisbon (PT)
Ricoh (JP) *	U Neuchâtel (CH) ***
SUNY Buffalo (US) *	U Oviedo (ES) *
Thomson Legal (US) ***	U Padua (IT) **
U Alicante (ES) ***	U.Stockholm/SICS (SE) ***
U Amsterdam (NL) ***	U.Surugadai/NII/NTU (JP/TW) *

* = number of previous participations in CLEF

13 different topic languages were used for experiments. As always, the most popular language for queries was English, but this year French came a fairly close second. A breakdown into the separate tasks is shown in Table 2 and of the runs per topic language in Table 3.

Table 2. CLEF 2004 ad hoc experiments

Track	# Participants	# Runs/Experiments
Multilingual	9	35
Bilingual X → FI	2	4
Bilingual X → FR	7	30
Bilingual X → PT	4	15
Bilingual X → RU	8	28
Bilingual X→ EN (restricted)	4	11
Monolingual FI	11	30
Monolingual FR	13	38
Monolingual PT	8	23
Monolingual RU	14	36

Table 3. List of experiments by topic language

Language[4]	# Runs
AM Amharic	1
BG Bulgarian	5
ZH Chinese	2
NL Dutch	7
EN English	65
FI Finnish	30
FR French	48
DE German	22
JP Japanese	2
PT Portuguese	23
ES Spanish	8
SV Swedish	1
RU Russian	36

As stated, participants were required to submit at least one title+description ("TD") run per task in order to increase comparability between experiments. In fact, the large majority of runs (205 out of 250) used this combination of topic fields, 31 used all fields and only 14 used the title field. The majority of experiments were conducted using automatic query construction. Manual runs tend to be a resource-intensive undertaking and it is likely that most participants interested in this type of work concentrated their efforts on the interactive track.

[4] Throughout the paper, language names are sometimes shortened by using their ISO-639 2-letter equivalent.

3 Characteristics of the Experiments

As expected, the choice of our target languages this year for the multilingual task seemed to pose challenges for the participants. As we had hoped to see, various approaches to tackling these challenges were proposed [7, 8, 9]. As already mentioned, the monolingual track saw the introduction of Portuguese this year, and consequently adaptations of existing approaches to this language, as well as to the previously little used Finnish and Russian, were also proposed [10, 11].

An additional consequence of the extra spotlight that the languages used in this year's multilingual track have received is the substantial work on splitting of Finnish compound words (decompounding), e.g. by [10, 12, 13].

The value of stemming and decompounding are issues that were hotly debated in previous campaigns but have now lost some attention. With the exception of the work on Finnish decompounding, the pros and cons of stemming and decompounding were not widely discussed in the participants' descriptions of their work. It could be concluded that this silent acceptance of stemmers and decompounding components as an integral part of most systems demonstrates that, in general, the value of such components for richly inflected languages is recognized by CLEF participants.

To complement these mainly linguistically motivated developments, we can discern a growing interest in new(er) weighting schemes, differing from the classical SMART Lnu.ltn [14] and OKAPI BM25 [15] weighting formulas. Some of the approaches explored by participants include deviation from randomness [12, 16, 17] and language models [13, 18].

Merging, i.e. the weighting of the different subcollections (both inter- and intra-language) from which the systems retrieve relevant documents, remains an unsolved problem from previous campaigns. Many approaches used by participants "reduce" multilingual retrieval to a sequence of bilingual retrieval runs, the results of which are then combined into a single, multilingual result. If retrieval scores are not comparable across these subcollections, the merging (combination) step proves to be difficult. It has been shown that much potential in terms of improving retrieval effectiveness lies in a better solution to the merging problem [19, 20]. Some of the merging experiments conducted this year are included in [18, 21, 22].

Information Retrieval (IR) technology has come a long way in recent years in terms of being incorporated into commercial products. Web search services based on IR approaches gain much attention, but a number of commercial enterprise IR software packages have also successfully entered the market. A different trend emerging in the last few years in the field of computer software is the successful development of "open source software", i.e. software that has liberal usage policies (often free of charge), comes with full source code, and is frequently developed by a volunteer community. The two trends start to produce collaborative results with the arrival of open source IR software. This year, in CLEF, the use of commercial and open source IR software as the basis of experiments has become more prominent, as opposed to using purely experimental tools developed during research work. Groups such as [23, 24] discuss their choice of commercial and open source systems.

In the bilingual task, participants were presented with the same target languages as in the multilingual task, plus Portuguese. It is interesting to note how similar were the performances of different groups for experiments in retrieving documents from the French document collection. French was introduced as a document language in the CLEF campaigns from the very beginning in 2000, meaning that returning participating groups, in particular, have had ample time to gain experience with this language for CLIR. It has been noticed before that the open spirit of the CLEF workshops, where participants freely share experiences and ideas, leads to a substantial pick-up of successful ideas by different groups [25]. This may explain the similarities in performance. It also underscores the value of new participants coming into the campaigns with "exotic" ideas, which minimize the danger of developing monocultures of CLIR approaches (see also [25]). A similar effect was discernible in the French monolingual track this year.

Generally speaking, both for the multilingual and bilingual tracks, query translation, as opposed to document translation, remains the method of choice for most participants. Document translation has clear advantages in terms of avoiding the merging problem, but seems to be judged as too "expensive" in terms of translation effort. Experiments in document translation have been conducted by [13, 26].

4 Results

The individual results of the participants are reported in detail in this volume and in the CLEF 2004 Working Notes [4] which were distributed to participants in the Workshop and are available on the CLEF website. In the following, we briefly summarize the main results for the multilingual, bilingual and monolingual tasks.

4.1 Multilingual Retrieval

This year, nine groups submitted 35 experiments for the multilingual task. This can be compared to the total of fourteen groups in the previous year when the ad hoc track was still the major focus of CLEF and two multilingual tasks were offered. Figure 1 shows the best entries of the top five performing groups in terms of average precision figures. Only entries using the title+description topic field were used for this comparison. Not surprisingly, the groups with the best results for this task were all veteran CLEF participants (with the exception of SUNY Buffalo) that had participated regularly in CLEF since 2001.

The top groups tended to focus a lot of attention on the merging problem [8, 18, 21, 27]. The group with the best result [13] also experimented with combination methods using runs made on various types of indexes, applying both language-dependent and language-independent tokenization techniques. Several of the groups participating in this task mentioned problems in processing and finding appropriate translation resources for the newer and less familiar CLEF languages – Finnish and Russian [17, 24].

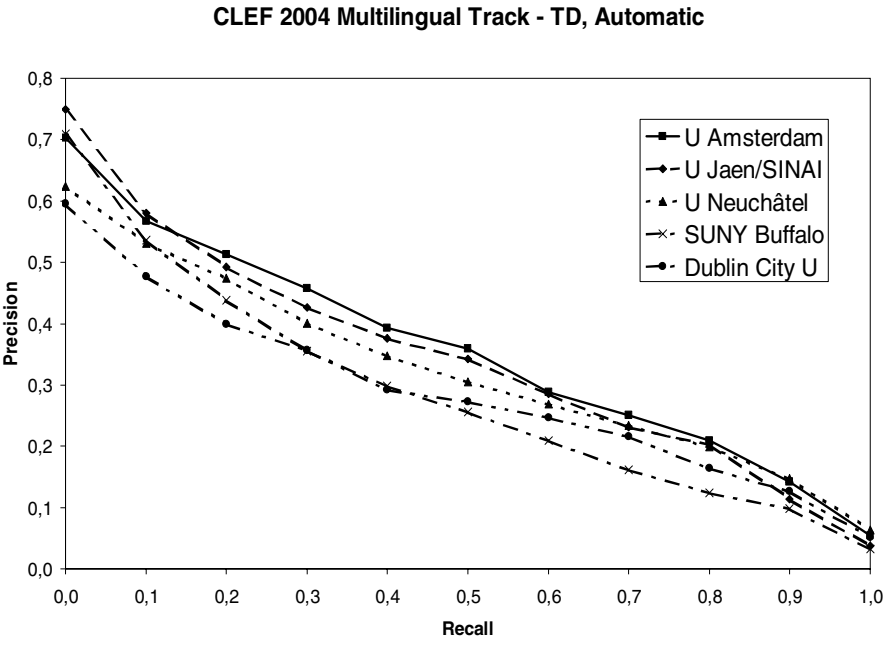

Fig. 1. Best performing entries of the top five participants in the multilingual task. The precision/recall curve, giving precision values at varying levels of recall, is shown. Only experiments using title+description topic fields are included

4.2 Bilingual Retrieval

The bilingual task was structured in four subtasks (X → FI, FR, RU or PT target collection) plus, as usual, an additional subtask with English as a target language – this last task was restricted to newcomers in a CLEF cross-language task or to groups using unusual or new topic languages (in CLEF 2004 Amharic and Bulgarian). Table 4 shows the best results for this task.

As shown in Section 2 above, some restrictions were placed on the topic languages that could be used to query the French and Finnish collections. The aim was to stimulate experiments for language pairs for which bilingual resources are scarce or non-existent. Unfortunately, this may have led to low participation in these two tracks. Only two groups tried the official bilingual to Finnish task, using French and Spanish topics. The effectiveness of these experiments was limited, with the best run scoring at only 47% of the average precision of best monolingual Finnish run. The bilingual to French task, with a choice of topic language between Dutch, Finnish, German and Swedish, was more popular with seven groups submitting a total of 30 runs. By far the most favoured topic language was German (6 groups and 22 runs), next came Dutch (3 groups and 7 runs) and finally a Swedish group contributed just one Swedish to French run. Performance was higher for this task: the best two runs (one using German and the other using Dutch topics) had a performance that was approximately 76% in terms of average precision of the monolingual results for French.

There were no restrictions on the bilingual to Russian and Portuguese target collections. This was because these languages are new additions to CLEF (Russian in 2003 and Portuguese in 2004). All groups that tried these two tasks used English as a topic language; in addition two groups also tried Spanish topics for the Portuguese target, and three different groups also used Chinese, French or Spanish topics to query the Russian target. For both languages, the group with the best monolingual results also provided the best bilingual performance. In each case, these results were obtained using English as the topic language. The difference in performance compared with monolingual was 70% in terms of average precision for Russian and a high 91% for Portuguese. From a first glance at these results, it would seem that certain target languages yield lower cross-language retrieval results. Specifically, cross-language retrieval of Finnish text, with its extremely complex morphology, and Russian text, which uses a different alphabet and encoding system from the other languages in the CLEF collection, appears to pose as yet unsolved difficulties compared to CLIR on French and Portuguese text, respectively.

Table 4. Best entries for the bilingual task (title+description topic fields only). Where applicable, the performance difference between the best and the fifth placed group is given (in terms of average precision)

Trg.	1st	2nd	3rd	4th	5th	Δ1st/5th
FI	JHU/APL	CLIPS				
FR	JHU/APL	Thomson	Daedalus	NII group	DublinCity	+12.4%
PT	U.Neuchâtel	JHU/APL	U.Amsterd.	U.Alicante		
RU	U.Alicante	U.Berkeley	DublinCity	U.Neuchâtel	JHU/APL	+138.6%
EN	U.Amsterd.	U.Oviedo				

4.3 Monolingual Retrieval

Monolingual retrieval was offered for all target collections (Finnish, French, Russian, Portuguese) with the exception of English. As can be seen from Table 2, the number of participants and runs for each language was quite similar, with the exception of Portuguese, which was added when the campaign was already well under way, leading to a somewhat smaller participation. This year just three groups submitted

Table 5. Best entries for the monolingual track (title+description topic fields only). Additionally, the performance difference between the best and the fifth placed group is given (in terms of average precision)

Trg.	1st	2nd	3rd	4th	5th	Δ1st/5th
FI	Hummingb.	Thomson LR	U.Neuchâtel	JHU/APL	U.Amsterd.	+22.4%
FR	Hummingb.	U.Neuchâtel	Daedalus	SUNY	JHU/APL	+7.5%
PT	U.Neuchâtel	Hummingb.	JHU/APL	Thomson	U.Amsterd.	+19.9%
RU	U.Alicante	Hummingb.	U.Amsterd.	U Berkeley	Dublin CU	+26.9%

monolingual runs only (down from ten groups last year), two newcomers and one veteran group [10]. Most of the groups submitting monolingual runs were doing this as part of their bilingual or multilingual system testing activity. All the groups in the top five were veteran CLEF participants (see Table 5).

One of the findings of CLEF over the years has been that successful cross-language retrieval systems are based on effective and robust monolingual processing procedures [25]. Again this year, in confirmation of a trend already observed in the past, we noted that there was very little statistical difference between the results of most of the monolingual submissions (see Table 7 below).

5 Statistical Testing

For reasons of practicality, the CLEF ad hoc track uses a limited number of queries (50 in 2004), which are intended to represent a more or less appropriate sample of all possible queries that users would want to ask from the collection. When the goal is to validate how well results can be expected to hold beyond this particular set of queries, statistical testing can help to determine what differences between runs appear to be real as opposed to differences that are due to sampling issues. We aim to identify runs with results that are significantly different from the results of other runs. "Significantly different" in this context means that the difference between the performance scores for the runs in question appears greater than what might be expected by pure chance. As with all statistical testing, conclusions will be qualified by an error probability, which was chosen to be 0.05 in the following. We have designed our analysis to follow closely the methodology used by similar analyses carried out for TREC [28].

A statistical analysis tool named IR-STAT-PAK [29] was used for the statistical analyses on the ad hoc track for the 2001 – 2003 campaigns. However, as this tool seems to be no longer supported or available on the Web, we have used the MATLAB Statistics Toolbox 5.0.1 this year, which provides the necessary functionality plus some additional functions and utilities. We continue to use the ANOVA test (Analysis of Variance). ANOVA makes some assumptions concerning the data be checked. Hull [28] provides details of these; in particular, the scores in question should be approximately normally distributed and their variance has to be approximately the same for all runs. IR-STAT-PAK uses the Hartley test to verify the equality of variances. This year two tests for goodness of fit to a normal distribution were chosen using the MATLAB statistical toolbox: the Lilliefors test [30] and the Jarque-Bera test [31]. In the case of the CLEF multilingual collection, both tests indicate that the assumption of normality is violated for most of the data samples (in this case the runs for each participant); in particular, the Lilliefors test shows that for 34 out of 35 runs the hypothesis of normality should be rejected, and the Jarque-Bera shows that the same hypothesis should be rejected for 18 runs. In such cases, a transformation of data should be performed. The transformation for measures that range from 0 to 1 is the arcsin-root transformation:

$$f(x) = \arcsin\left(\sqrt{x}\right)$$

which Tague-Sutcliffe [32] recommends for use with precision/recall measures. After the transformation the analysis of the normality of samples distribution improves significantly: the Lilliefors test claims that 15 runs are still non-normally distributed while the Jaque-Bera test indicates that only two samples are non-normally distributed. The difficulty to transform the data into normally distributed samples derives from the original distribution of run performances, which tend towards zero within the interval [0,1].

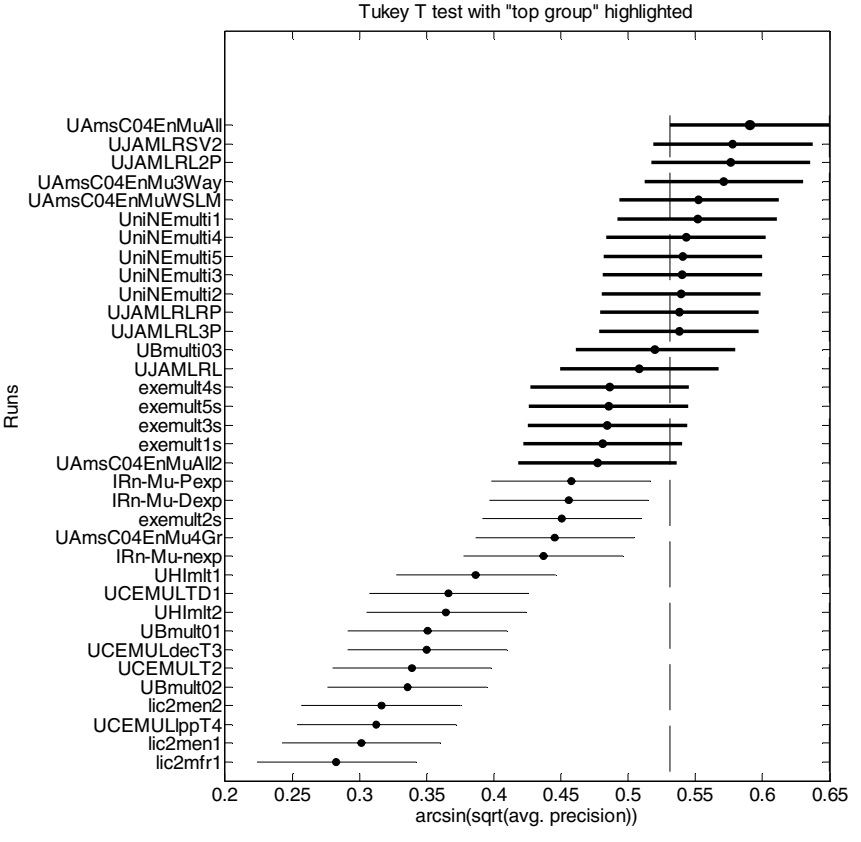

Fig. 2. Tukey T test for the multilingual track

In any case, the situation after the arcsin-root transformation allows us to perform a two-way ANOVA test that determines if there is at least one pair of runs that exhibit a statistical difference. Following a significant two-way ANOVA, various comparison procedures can be employed to investigate significant differences. The Tukey T test was used to find the statistically significant differences between participants' performances and to group runs. In particular, we used the MATLAB *multcompare* function with an *honestly significant difference* (hsd) setup for the Tukey T test.

Table 6. Results of statistical analysis (two-way ANOVA) on the experiments submitted for the multilingual task. All experiments, regardless of topic language or topic fields, are included. Results are therefore only valid for comparison of individual pairs of runs, and not in terms of absolute performance

Arcsin-transformed average precision values	Run Ids	Groups
0.5905	UAmsC04EnMuAll	X
0.5778	UJAMLRSV2	X
0.5764	UJAMLRL2P	X
0.5713	UAmsC04EnMu3Way	X X
0.5527	UAmsC04EnMuWSLM	X X X
0.5515	UniNEmulti1	X X X
0.5431	UniNEmulti4	X X X
0.5404	UniNEmulti3	X X X
0.5407	UniNEmulti5	X X X
0.5394	UniNEmulti2	X X X
0.5381	UJAMLRLRP	X X X
0.5379	UJAMLRL3P	X X X
0.5203	UBmulti03	X X X
0.5083	UJAMLRL	X X X
0.4863	exemult4s	X X X X
0.4854	exemult5s	X X X X
0.4845	exemult3s	X X X X X
0.4812	exemult1s	X X X X X X
0.4771	UAmsC04EnMuAll2	X X X X X X
0.4575	IRn-Mu-Pexp	X X X X X X
0.4558	IRn-Mu-Dexp	X X X X X X
0.4505	exemult2s	X X X X X X
0.4455	UAmsC04EnMu4Gr	X X X X X X
0.4367	IRn-Mu-nexp	X X X X X X
0.3866	UHImlt1	X X X X X X
0.3666	UCEMULTD1	X X X X X X
0.3646	UHImlt2	X X X X X
0.3505	UBmult01	X X X X
0.3504	UCEMULdecT3	X X X
0.3391	UCEMULT2	X X X
0.3355	UBmult02	X X X
0.3162	lic2men2	X X
0.3127	UCEMULlppT4	X
0.3014	lic2men1	X
0.2829	lic2mfr1	X

Two different graphs are presented to summarize the results of this test: Figure 2 shows participants' runs (y axis) and performance obtained (x axis). The circle indicates the average performance (in terms of Precision) while the segment shows the interval in which the difference in performance is not statistically significant.

Alternatively, the overall results are presented in Table 6, where all the runs that are included in the same group do not have a significantly different performance. All runs scoring below a certain group perform significantly worse than at least the top entry of that group. Likewise, all the runs scoring above a certain group perform significantly better than at least the bottom entry in that group. To determine all runs that perform significantly worse than a certain run, determine the rightmost group that includes the run. All runs scoring below the bottom entry of that group are significantly worse. Conversely, to determine all runs that perform significantly better than a given run, determine the leftmost group that includes the run. All runs that score better than the top entry of that group perform significantly better.

It is well known that it is fairly difficult to detect statistically significant differences between retrieval results based on 50 queries [32, 33]. While 50 queries remains a good choice based on practicality for doing relevance assessments, statistical testing would be the one of the areas to benefit most from having additional topics.

This fact is addressed by the measures taken to ensure stability of at least part of the document collection across different campaigns, which allows participants to run their system on aggregate sets of queries for post-hoc experiments.

For the 2004 campaign, we conducted a statistical analysis of the "pools of experiments" for all target languages. It seems that each year it is increasingly difficult to identify clearly significant differences in participants' performances. For example, in the multilingual task, the first group identified by the Tukey T test, contains a total of 19 runs submitted by 5 different participants: University of Amsterdam, University of Jaen, Université de Neuchâtel, State University New York at Buffalo, Dublin City University. From these results, it is only possible to state that this first group of participants performed significantly better than the other groups, but it is not possible to identify the top performer with any statistical validity.

Table 7. Results of statistical analysis (ANOVA) on the monolingual experiments. The table shows the number of participants submitting at least one experiment with a performance that is not statistically different to the top performance against the total number of participants submitting experiment for that target collection

Target collection	# of participants in the top group / total # of participants
Finnish	9/12
French	14/15
Portuguese	6/8
Russian	12/15

In addition to the multilingual task, we also examined non-English mono- and bilingual target collections. The analyses included both monolingual runs, and also the bilingual runs to the same target language (i.e. the French analysis contains both French monolingual and German → French bilingual experiments). Like in CLEF 2003, the monolingual tasks were very competitive. Many groups submitted experiments with very similar performances, and almost all groups that submitted at

least one run are present in the top performing group (see Table 7). It should be noted, however, that experiments of very different character are mixed in this analysis.

A complete listing and the individual results (statistics and graphs) of all the official experiments for the ad hoc track can be found in the Appendix to the CLEF Working Notes [4].

6 Impact of CLEF

This paper summarizes and analyses the results of the ad hoc track in the CLEF 2004 campaign. The size and scope of the ad hoc track in CLEF 2004 was limited somewhat in order to leave more space for new tracks addressing other issues in CLIR. However, even if the number of experiments submitted is significantly below that of 2003, the track has promoted interesting and novel work (e.g. on the problem of merging results from different collections, and experiments with different weighting formulas).

An important question is what impact the CLEF campaigns have on the current state-of-the-art in CLIR research. As test collections and tasks vary over years, it is not easy to document improvements in system performance. One common method for bilingual retrieval evaluation is to compare results against monolingual baselines. We can observe the following indications with respect to progress in bilingual retrieval over the years:

In 1997, at TREC-6, the best CLIR systems had the following results:

- EN → FR: 49% of best monolingual French IR system
- EN → DE: 64% of best monolingual German IR system

In 2002, at CLEF, with no restriction on topic and target language, the best systems obtained:

- EN →FR: 83% of best monolingual French IR system
- EN → DE: 86% of best monolingual German IR system

However, CLEF 2003 enforced the use of previously "unusual" language pairs, with the following impressive results:

- IT → ES: 83% of best monolingual Spanish IR system
- DE → IT: 87% of best monolingual Italian IR system
- FR → NL: 82% of best monolingual Dutch IR system

CLEF 2004 presented participants with a mixed set of limitations according to the respective target languages. Results include:

- ES → FI: 47% of best monolingual Finnish IR system
- DE/NL → FR: 76% of best monolingual French IR system
- EN → RU: 70% of best monolingual Russian IR system
- EN → PT: 91% of best monolingual Portuguese IR system

Again, comparisons are difficult due to increasingly complex tasks. However, it appears that a steady trend of overall improvement in CLIR performance can be

recognized as gradually systems begin to be capable of handling different and previously unusual languages pairs, finding and exploiting translation mechanisms between pairs of languages that do not include English.

It is even harder to measure progress with respect to the multilingual retrieval task. Partly for this reason, in CLEF 2005, we are proposing the CLEF 2003 multilingual-8 task again ("Multi-8 Two-years-on") The aim is to see whether there is an improvement in performance over time. In any case, CLIR systems that tackle this many languages simultaneously are clearly a great testament to the development of the field over the past years.

References

1. Peters, C.: What happened in CLEF 2004? In this volume.
2. Cleverdon, C.: The Cranfield Tests on Index Language Devices. In: Sparck-Jones, K., Willett, P. (eds.): Readings in Information Retrieval, Morgan Kaufmann (1997) 47-59.
3. Harman, D.: The TREC Conferences. In Kuhlen, R., Rittberger, M. (eds.): Hypertext - Information Retrieval - Multimedia: Synergieeffekte Elektronischer Informationssysteme, Proceedings of HIM '95, Universitätsverlag Konstanz, 9-28.
4. CLEF 2004 Working Notes. http://clef.isti.cnr.it/2004/working_notes/CLEF2004WN.
5. Braschler, M.: CLEF 2003 – Overview of results. In: Fourth Workshop of the Cross-Language Evaluation Forum, CLEF 2003, Trondheim, Norway, 2003. Revised papers. Lecture Notes in Computer Science 3237, Springer 2004, 44-63.
6. Braschler, M., Peters, C.: CLEF 2003 Methodology and Metrics. In: Fourth Workshop of the Cross-Language Evaluation Forum, CLEF 2003, Trondheim, Norway, 2003. Revised papers. Lecture Notes in Computer Science 3237, Springer 2004, 7-20.
7. Besançon, R., Ferre, O., Fluhr, C.: LIC2M Experiments at CLEF 2004. In this volume.
8. Martínez-Santiago, F., García-Cumbreras, M.A., Díaz-Galiano, M.C., Ureña, L.A. : SINAI at CLEF 2004: Using Machine Translation Resources with a Mixed 2-Step RSV Merging Algorithm. In this volume.
9. Levow, G.-A., Matveeva, I.: University of Chicago at CLEF2004: Cross-language Text and Spoken Document Retrieval. In this volume.
10. Tomlinson, S.: Portuguese and Russian Retrieval with Hummingbird SearchServer™ at CLEF 2004. In this volume.
11. Di Nunzio, G.M., Ferro, N., Orio, N.: Experiments on Statistical Approaches to Compensate for Limited Linguistic Resources. In this volume.
12. Savoy, J.: Data Fusion for Effective European Monolingual Information Retrieval. In this volume.
13. Kamps, J., Adafre, S.F., de Rijke, M.: Effective Translation, Tokenization and Combination for Cross-Lingual Retrieval. In this volume.
14. Singhal A., Buckley C., Mitra M.: Pivoted Document Length Normalization. In Frei, H.P., Harman, D., Schäuble P., Wilkinson, R., eds.: Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1996), Zurich, Switzerland, ACM, 1996, 21-29.
15. Robertson, S.E., Walker, S., Jones, S., Hancock-Beaulieu, M. M., Gatford, M.: Okapi at TREC-3. In: Overview of the Third Text Retrieval Conference (TREC-3), NIST Special Publication 500-225, 1995, 109-126.
16. Lioma, C., He, B., Plachouras, V., Ounis, I.: The University of Glasgow at CLEF 2004: French Monolingual Information Retrieval with Terrier. In this volume.

17. Serasset, G., Chevallet, J.-P.: Using Surface-Syntactic Parser and Derivation from Randomness: X-IOTA IR System used for CLIPS Mono & Bilingual Experiments for CLEF 2004. In this volume.
18. Ruiz, ME., Srikanth, M.: UB at CLEF2004: Cross-language Information Retrieval using Statistical Language Models. In this volume.
19. Chen A.: Cross-Language Retrieval Experiments at CLEF 2002. In Working Notes for the CLEF 2002 Workshop, 2002, 5-20.
20. Braschler, M.: Combination Approaches for Multilingual Text Retrieval, In Information Retrieval, Kluwer Academic Publishers, Vol. 7(1-2), 183-204.
21. Savoy, J., Berger, P.-Y.: Selection and Merging Strategies for Multilingual Information Retrieval. In this volume.
22. Goñi-Menoyo, J.M., González, J.C., Martínez-Fernández, J.L., Villena-Román, J.: MIRACLE's Hybrid Approach to Bilingual and Monolingual Information Retrieval. In this volume.
23. Nadeau, D., Jarmasz, M., Barrière, C., Foster, G., St-Jacques, C.: Using COTS Search Engines and Custom Query Strategies at CLEF. In this volume.
24. Hackl, R., Mandl, T., Womser-Hacker, C.: Mono- and Crosslingual Retrieval Experiments at the University of Hildesheim. In this volume.
25. Braschler, M., Peters, C.: Cross-Language Evaluation Forum: Objectives, Results, Achievements, Information Retrieval, Vol.7 (1-2) 5-29.
26. Gey, F.C.: Searching a Russian Document Collection using English, Chinese and Japanese Queries. In this volume.
27. Jones,G.J.F., Burke, M., Judge,J., Khasin, A., Lam-Adesina, A., Wagner, J.: Dublin City University at CLEF 2004: Experiments in Monolingual, Bilingual and Multilingual Retrieval. In this volume.
28. Hull, D.: Using Statistical Testing in the Evaluation of Retrieval Experiments. In Korfhage, R., Rasmussen, E., Willett, P., eds.: Proc. 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1993), ACM Press, New York, USA (1993) 329–338.
29. Blustein, J.: IR STAT PAK. URL: http://www.csd.uwo.ca/~jamie/IRSP-overview.html.
30. Conover, W.J.: Practical Nonparametric Statistics, (1st Ed.), John Wiley and Sons, New York, 1971.
31. Judge, G. G., R. C. Hill, W. E. Griffiths, H. Lütkepohl, and T.C. Lee: Introduction to the Theory and Practice of Econometrics, (2nd ed.), John Wiley and Sons, New York, 1988.
32. Tague-Sutcliffe, J.: The Pragmatics of Information Retrieval Experimentation, Revisited. In: Reading in Information Retrieval, Morgan Kaufmann Publishers, San Francisco, CA, USA (1997), 205-216.
33. Voorhees, E., Buckley, C.: The Effect of Topic Set Size on Retrieval Experiment Error. In: Proc. of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (1998) 307-314.