CLEF - IST-2000-31002
Cross-Language Evaluation Forum

Results Interpretation and Report (Campaign 1)

D4.2.1

Deliverable Type: REPORT
Number: D4.2.1
Nature: Public
Contractual Date of Delivery: month 15
Actual Date of Delivery: month 15
Task WP4.2

Name of responsible: Eurospider Information Technology AG (EIT)

Authors: Martin Braschler, EIT

contact info
martin.braschler@eurospider.com

Abstract

Deliverable 4.2.1 presents an overview of the results obtained by the participants at the CLEF 2002 campaign, and discusses the validity of the results based on statistical significance and the stability of the evaluation measures.

Keyword List

Information Retrieval, Evaluation Results, CLIR, Statistical Testing, Relevance Assessments, Pool Validity

Executive Summary

In this deliverable, we present an overview of the results obtained by the participants at the CLEF 2002 campaign, and discuss the validity of the results based on statistical significance and the stability of the evaluation measures.
CLEF 2002 has seen the largest number of participants and experiments submitted among the campaigns organized under the name "CLEF" so far. In particular, it should be noted that a steadily increasing number of European participants submit results to CLEF. We report both on the main characteristics of the experiments and on trends seen in the approaches and methodologies used by the participants. An exhaustive report on all individual results is outside the scope of this report for practical reasons, but the main results are summarized and analyzed.
Careful statistical analysis allows potential generalization of claims based on findings inside CLEF. We describe how to carry out such statistical analysis and give results for the main task in CLEF, the multilingual track.
CLEF relies heavily on relevance assessments for the calculation of its performance measures. To ensure validity of the results published by CLEF, we investigate the quality of the relevance assessments by computing their coverage.

Main findings of the 2002 campaign are:

1. Participants' systems have matured, and the top performing systems are either complex combination systems built by long-time participants (multilingual track) or systems that have been carefully fine-tuned for specific target languages (bilingual and monolingual tracks)
2. Statistical analysis of the results underscores the importance of ensuring reusability of the test collection prepared by CLEF, since combining data from multiple previous campaigns can help to solidify results and obtain better statistical significance
3. The investigation into the quality of the relevance assessment pools shows that they are very stable, and that the test collection should therefore be well suited for later post-hoc experiments. This also ensures that results as published by CLEF should be valid within the inherent limitations of the testing methodology.

From the standpoint of the increased participation, the large diversity of the submitted experiments and the quality of the resulting test collection, we consider the 2002 campaign a big success.

Table of Contents

1 Introduction

The campaigns organized as centerpiece of the CLEF project build on work carried out in earlier years inside the TREC campaigns [6], organized in the United States by the National Institute of Standards and Technology (1997-1999), and on work carried out as part of the DELOS Network of Excellence under the banner "CLEF" (2000-2001) [2], [3]. In this sense, much of this work is a continuation of earlier efforts, and it is possible to draw comparisons to the campaigns organized as part of these earlier activities. The deliverable does this freely, where appropriate. While building on this earlier work, the CLEF 2002 campaign far eclipses the previous campaigns in the amount of participants involved and data processed.
This deliverable reports on general characteristics of the experiments submitted for CLEF, such as the participants and the languages/topic fields used, as well as on trends observed in the work by the participating groups. Only a summary of the hundreds of different results obtained by the participants is given, since an exhaustive listing is well beyond the scope of this report. Interested readers can refer to the complete working notes of the CLEF workshop, which contains nearly 300 pages of individual result listings [8].
From necessity, CLEF uses a limited set of "constructed" information needs that serve as a representation of the type of queries real users might ask a CLIR system, known as topics. Consequently, we must investigate how the results generalize beyond this limited setting. Statistical analysis provides us tools for this task. We describe how to carry out statistical analysis on CLEF results and present results for the main, multilingual task.
CLEF relies heavily on relevance assessments to compute the published performance measures. We investigate the quality of the relevance assessments by carrying out an analysis of their coverage (completeness).
The deliverable is structured into three main sections, Sections 2-4, giving details on the experiments (Section 2), on statistical analysis (Section 3), and on the quality of the relevance assessments (Section 4). Conclusions are given in Section 5.

2 Overview of Results

2.1 Participants

In all, 37 participants from 12 different countries participated in one or more activities offered under the CLEF umbrella. This is a very substantial growth compared to all previous campaigns except the last one, which already had 34 participants. However, the main growth occurred in the amount of data that was submitted by the participants and processed by the CLEF consortium: a total of 282 experiments was submitted for the main tasks[1]. Many of the participants had already taken part in earlier CLEF campaigns or in related activities, such as TREC (North America) and/or NTCIR (East Asia). However, there was also a healthy number of newcomers. The number of European groups is still growing, and is now a clear majority (27,5 out of 37, showing the importance of CLEF in fostering interest in CLIR research in Europe) (see Table 1).

[1] Additionally, a substantial number of "interactive experiments" were submitted, which are not included in this total, because, while carried out under the CLEF umbrella, they are not part of the official project work as defined by the technical annex and not funded as part of the CLEF project.

City University (UK[2])	SINAI/U Jaen * (ES)
Clairvoyance Corp. (US)	Tagmatica (FR)
COLE Group/U La Coruna (ES)	Thomson Legal ** (US)
CWI/CNLP * (NL/US)	U Alicante * (ES)
Eurospider ** (CH)	U Amsterdam * (NL)
Fondazione Ugo Bordoni (IT)	U Dortmund * (DE)
Hummingbird * (CA)	U Exeter * (UK)
IMBIT (DE)	U Hildesheim (DE)
IMS U Padova (IT)	U Maryland ** (US)
IRIT ** (FR)	U Montreal/RALI ** (CA)
ITC-irst ** (IT)	U Neuchâtel * (CH)
JHU-APL ** (US)	U Salamanca ** (ES)
Lexware (SV)	U Sheffield ** (UK)
Medialab * (NL)	U Tampere ** (FI)
Middlesex U (UK)	U Twente/TNO ** (NL)
National Taiwan U * (TW)	UC Berkeley (2 groups) ** (US)
OCE Tech. BV *	UNED *
SICS/Conexor *	Xerox *

Table 1. Participants to CLEF 2002. One star (*) denotes a participant that has taken part in one previous campaign (2000 or 2001), two stars (**) denote participants that have taken part in both previous campaigns.

2.2 Collection and Tasks

For 2002, the CLEF consortium considerably expanded the test collection used for the experiments in every respect: more documents (~20%), more languages covered (8, with Finnish and Swedish being new) and most importantly, more relevance assessments (~40% more). Aspects of the additional documents and languages are covered in deliverable 2.3.1 "Multilingual Collection for Campaign 1", consigned month 6 [4]. Relevance assessment procedures are detailed in deliverable 3.2.1 "Test Collection Report for Campaign 1" [5].

For the 2002 campaign, CLEF tasks were structured as follows:

1. Multilingual Track: Choice of 11 topic languages, search multilingual document collection containing documents each written in one of five languages (DE, EN, ES, FR, IT; ~750,000 documents). This was the "main", hardest task.
2. Bilingual Track: Choice of 11 topic languages, choice of 7 document languages (DE, ES, FI, FR, IT, NL, SV), newcomers could also choose EN as target language. Target collection contains only documents written in the chosen language.
3. Monolingual Track. Choice of 7 topic languages (DE, ES, FI, FR, IT, NL, SV). Documents in same language as topic language.

CLEF de-emphasizes retrieval on English language documents (only included in the multilingual track), as it is already covered in the TREC evaluation campaigns.

CLEF 2002 also offered domain-specific retrieval:

[2] In this paper, we use ISO 3166 2-letter country codes to denote the countries of origin of participants, and ISO 639 2-letter language codes to abbreviate references to topic and document languages.

4. GIRT track: Choice of three topic languages (DE, EN, RU). Retrieval of German documents from the domain of social sciences.
5. Amaryllis track. Choice of two topic languages (EN, FR). Retrieval of French scientific documents.

2.3 Experiments and their Characteristics

A total of 282 experiments were officially submitted for these 5 tracks. This is an increase of more than 40% compared to CLEF 2001, making the 2002 campaign an unprecedented success. Submissions were divided among the tracks as follows (Table 2, Figure 1):

Track	# Participants	# Runs/Experiments
Multilingual	11	36
Bilingual to DE	6	13
Bilingual to EN	5	16
Bilingual to ES	7	16
Bilingual to FI	2	2
Bilingual to FR	7	14
Bilingual to IT	6	13
Bilingual to NL	7	10
Bilingual to SV	1	1
Monolingual DE	12	21
Monolingual ES	13	28
Monolingual FI	7	11
Monolingual FR	12	16
Monolingual IT	14	25
Monolingual NL	11	19
Monolingual SV	6	9
Domain-specific Amaryllis	3	15
Domain-specific GIRT	5	17

Table 2. Different tracks/tasks, and the respective number of particpants/experiments.

This is a fairly even distribution, both in terms of the tasks and the languages covered (see also Table 3). Obviously, it is very difficult to "steer" the distribution, since it reflects the participants' interest. Considering this, the fact that all nearly all tasks/combinations are well represented (apart from bilingual to the new languages FI and SV) is encouraging. The participation in the domain-specific tasks was somewhat below what we have hoped, but this is in line with earlier experiences which showed that while a lot of interest is initially expressed by many participants, groups tend to drop these tasks when they run out of resources for their experiments.

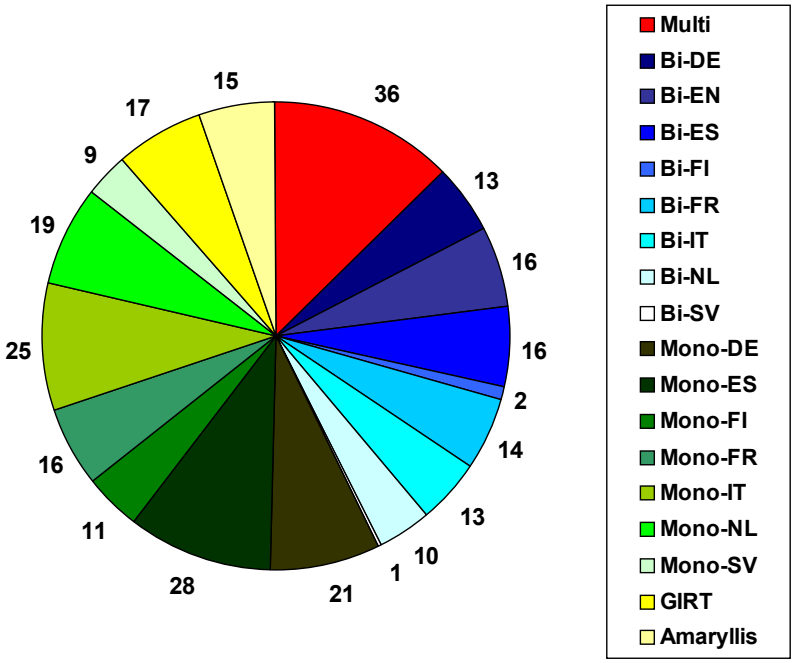

Figure 1. Distribution of the experiments across different tracks/tasks.

CLEF offers a choice of using short, medium-length and long queries for all experiments. All three choices were used by participants, with the medium-length queries dominating (participants were required to submit at least one experiment per task using medium length in order to boost comparability across sites) (Table 4). Queries could be constructed either manually or automatically out of the statements of information need (topics) distributed by the CLEF organization. The overwhelming majority of participants used automatic query construction.

Topic Language	# Experiments
DE German	38
EN English	99
ES Spanish	35
FI Finnish	11
FR French	32
IT Italian	26
NL Dutch	20
PT Portuguese	6
RU Russian	5
SV Swedish	9
ZH Chinese	4

Table 3. Distribution of tasks across topic (query) languages.

Topic fields	# Experiments
TDN – long queries	34
TD – medium-length queries	227
T – short queries	19
Other	2

Table 4. Different topic fields used for query construction. T=title, D=description, N=narrative.

2.4 Main Trends

With CLEF building on earlier campaigns organized both under the same name and under other umbrellas (TREC in North America, NTCIR in East Asia), there are participants that have worked on this type of evaluation for several years. Therefore, CLEF acts as a "trendsetter", and methods that work well one year are adapted eagerly by other participants in following campaigns. This is clearly a valuable contribution that CLEF plays in distributing successful ideas.
For the 2002 campaign, we discern the following main trends:

- Participants were using fewer "corpus-based" CLIR approaches, i.e. methods that extract translation resources from suitable training data. However, such approaches were still popular in combination systems (Latent Semantic Indexing, Similarity Thesauri, Statistical Models).
- A few MT systems proved to be very popular, mainly for query translation (Systran, LH Power Translator).
- Participants invested a lot of effort into work on (blind) query expansion, such as blind relevance feedback, the use of concepts and synonyms, association thesauri, similarity thesauri for expansion and others.
- A fairly new trend was the added emphasis on fine-tuned weighting per language (as opposed to using same parameters for all languages). It will remain a challenge to prove how the findings based on the CLEF collection generalize to a particular language, however.
- Continuing from the previous year, diverse work on stemming was submitted, using simple and elaborate stemming (morphological analyzers), a programming language expressly for stemmers and other ideas. These efforts were supplemented by interesting work on decompounding (however, different conclusions were reached for different languages on this issue).
- The merging problem, i.e. the combination of multiple translation resources, translation approaches or target languages into a single retrieval result was very much a main focus for participants this year. While simple methods were still widely used in this area, new ideas were proposed as well, such as the use of an unified index, reindexing, prediction based on translation quality, and feedback-based merging.

2.5 The Results

The individual results of the participants are reported in detail in the CLEF 2002 Working Notes [8] distributed to the participants at the CLEF workshop in Rome and are also available on the CLEF website. The focus of this report and the number of experiments submitted make it impossible to provide exhaustive lists of all individual results in this deliverable. In the following, we summarize the results for the multilingual, bilingual and monolingual track briefly.

Multilingual Track

The multilingual track is the hardest task to complete in CLEF and is therefore the main focus of the activities. Eleven groups submitted 36 runs to the track. Figure 2 shows the best entries of the five top performing groups in terms of average precision figures. Only entries using the title+description topic field combination were used for this comparison.

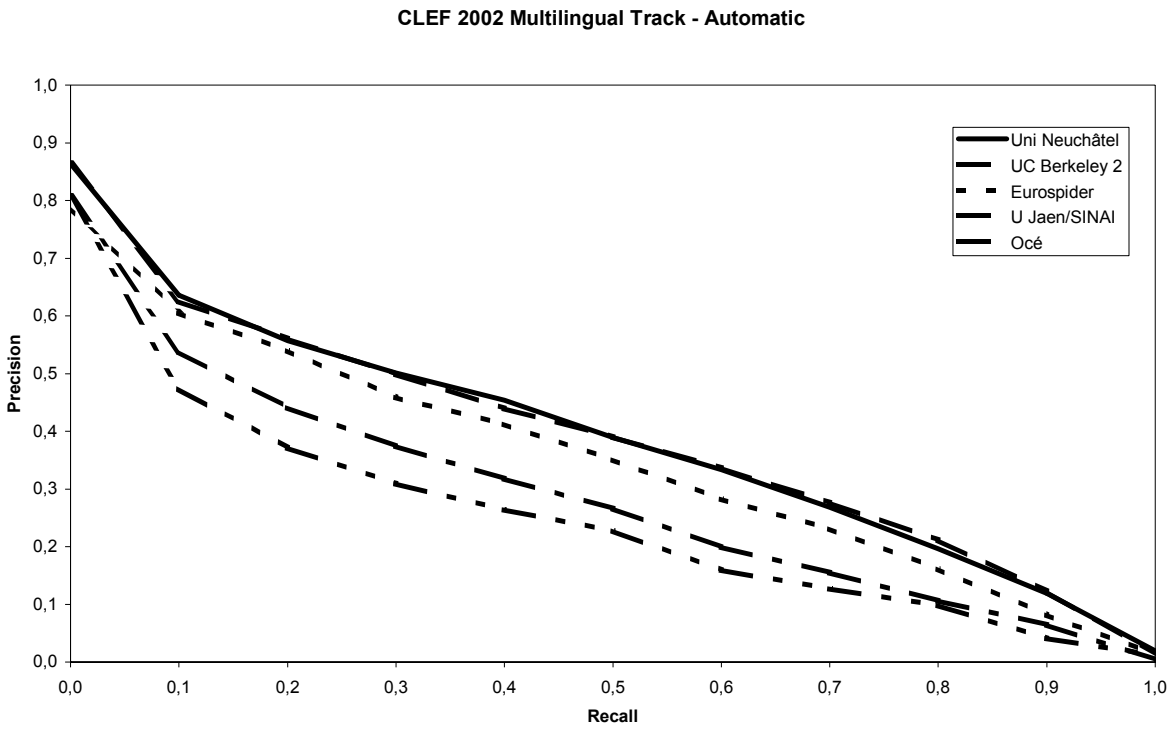

Figure 2. Best performing entries of the top five participants for the multilingual track. Shown is the precision/recall curve, giving precision values at varying levels of recall. Only experiments using the title+description topic fields are included.

As can be seen, the top two performances, by Université de Neuchâtel and by University of California at Berkeley Group 2, respectively, are very close in performance across all recall levels. A third entry by Eurospider follows closely behind (~10% below). After the top three entries, there is a considerable gap of around 20% to other participants. The top three entries all used elaborate combination approaches and come from groups that are long-time participants in these form of evaluation. Clearly, experience in building such sophisticated combination systems with considerable complexity helps in performing well in such a hard task. The two groups that round out the top five, SINAI/University of Jaen and Océ have participated in a CLEF-style evaluation before, but not in the multilingual track.

Bilingual track

The 2002 campaign offered a bilingual track that was far more extensive than the one offered in previous campaigns. In 2001, participants were free to chose between two target languages, English and Dutch. In 2002, the CLEF consortium responded to numerous requests from participants and opened the bilingual track to all eight target languages (DE, EN; ES, FI, FR, IT, NL, SV; EN for newcomers or under special conditions only). While allowing for added flexibility in testing the systems on the participant's part, this decision makes comparing different bilingual experiments somewhat harder, since experiments on different target languages use different document sets. It is therefore necessary to investigate eight different result sets, one for each target language. Table 5 shows the best entries by the top five performing participants for each target language, including only runs using the mandatory title+description topic field combination.

Target Language	1st	2nd	3rd	4th	5th
DE German	UC Berkeley 2	UC Berkeley 1	U Neuchâtel	JHU/APL	U Jaen/SINAI
EN English	JHU/APL	Clairvoyance	Océ	IRIT	Middlesex U
ES Spanish	U Neuchâtel	U Berkeley 2	U Exeter	Océ	JHU/APL
FI Finnish	U Tampere	JHU/APL	-	-	-
FR French	U Neuchâtel	UC Berkeley 2	UC Berkeley 1	JHU/APL	U Jaen/SINAI
IT Italian	UC Berkeley 2	U Exeter	U Neuchâtel	ITC-IRST	JHU/APL
NL Dutch	JHU/APL	U Twente/TNO	UC Berkeley 2	U Amsterdam	Océ
SV Swedish	JHU/APL	-	-	-	-

Table 5. Best entries for the bilingual track. Shown are the top five participants for each target language (title+description topic fields only).

While it is hard to compare results from English, Finnish and Swedish with the other languages – the former because of the restrictions in participation, and the later because of the smaller number of participants for the new language – there are one or two interesting trends that can be derived from the results of the remaining five languages. Firstly, the two groups University of California at Berkeley Group 2 and Université de Neuchâtel, both among the top groups in the multilingual track, again perform well for most languages. The differences between the top group and the runner-up are generally more pronounced than in the multilingual track (roughly 5-10%). We conclude that general knowledge of how to build CLIR systems seems to help in performing well across a variety of languages, but the languages still have individual potential for fine-tuning that results in different placement of the groups across the languages. Secondly, it is also interesting to see that for the Dutch target language, where we have an active group of participants from the Netherlands, three Dutch groups placed very well. Clearly, it is advantageous for fine-tuning to have detailed knowledge of the characteristics of a language.

Monolingual track

The CLEF 2002 campaign offered monolingual retrieval for all target languages besides English. Again, Table 6 summarizes the best entries of the top five performing groups for the title+description topic field combination.

Target Language	1st	2nd	3rd	4th	5th
DE German	UC Berkeley 2	U Amsterdam	U Neuchâtel	JHU/APL	Eurospider
ES Spanish	U Neuchâtel	UC Berkeley 2	JHU/APL	Thomson Legal	U Alicante
FI Finnish	U Neuchâtel	U Twente/TNO	Hummingbird	JHU/APL	U Amsterdam
FR French	UC Berkeley 2	U Neuchâtel	UC Berkeley 1	U Amsterdam	JHU/APL
IT Italian	F. U. Bordoni	ITC-IRST	UC Berkeley 2	U Neuchâtel	JHU/APL
NL Dutch	JHU/APL	U Neuchâtel	UC Berkeley 2	U Amsterdam	Hummingbird
SV Swedish	JHU/APL	U Amsterdam	Hummingbird	Thomson Legal	SICS/Conexor

Table 6. Best entries for the monolingual track. Shown are the top five participants for each target language (title+description topic fields only).

Again, the results for Finnish and Swedish are somewhat hard to interpret, because these two languages were offered in CLEF for the first time this year. The five "older" languages allow some interesting observations however. As already stated for the bilingual track, knowing a language well seems to be an advantage in designing good retrieval for it, as proven by the top two entries for Italian, which both come from groups from Italy. As for bilingual, there are

some groups that do well for many languages, but again the ranking changes from language to language. Competition for the top spots in the monolingual track this year was fierce, with participants that submitted the best performing entries last year dropping six or seven ranks this year for the same language. Generally, results in German, Spanish French, Italian and Dutch are very close, with the careful fine-tuning of little details providing the difference. This seems to be a clear indication that monolingual non-English text retrieval has matured, and that participants have better knowledge of how to squeeze optimum performance out of their systems. It might also indicate, however, that participants potentially start overtuning for the CLEF document collection. It will be an important challenge in future experiments to investigate how much of the differences seen across the target languages are due to language-specific characteristics, and how much can be attributed to collection-specific artifacts.

Domain-specific

As already stated, we had less entries for the domain-specific tracks than for the other tracks. We again give a summary of the best entries of the top five performing groups for the title+description topic field combination (Table 7).

Track	1st	2nd	3rd	4th	5th
GIRT Bilingual	UC Berkeley 1	U Amsterdam	-	-	-
GIRT Monolingual	UC Berkeley 1	U Amsterdam	U Dortmund	U Hildesheim	-
Amaryllis Bilingual	UC Berkeley 1	U Amsterdam	-	-	-
Amaryllis Monolingual	U Neuchâtel	UC Berkeley 1	U Amsterdam	-	-

Table 7. Best entries for the domain-specific tracks. Shown are the top five participants for each target language (title+description topic fields only).

The smaller number of participants makes it difficult to draw overall conclusions. Indeed, the performance obtained by the groups was very dissimilar, probably due to the different degree of tuning for the characteristic of the domain-specific data. A detailed description of the different experiments for the domain-specific tracks can be found in the CLEF 2002 working notes [8].

3 Statistical Testing

CLEF uses, for reasons of practicality, a limited number of queries (50 in 2002), which are intended to represent a more or less appropriate sample of the population 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 determine what differences between runs appear to be real as opposed to differences that are due to sampling variation. As with all statistical testing, conclusions will be qualified by an error probability, which was chosen to be 0.05 in the following.

Using the IR-STAT-PAK tool [1], a statistical analysis of the results for the multilingual track was carried out for the first time after the 2001 campaign. We have repeated this analysis for 2002. The tool provides an Analysis of Variance (ANOVA) which is the parametric test of choice in such situations but requires that some assumptions concerning the data are checked. Hull [7] 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. In the case of the CLEF

multilingual collection, it indicates that the assumption is violated. For such cases, the program offers an arcsine transformation,

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

which Tague-Sutcliffe [10] recommends for use with Precision/Recall measures, and which we have therefore applied.

The ANOVA test proper only determines if there is at least one pair of runs that exhibit a statistical difference. Following a significant ANOVA, various comparison procedures can be employed to investigate significant differences. IR-STAT-PAK uses the Tukey T test for grouping the runs.

Looking at the result (Table 8), all runs that are included in the same group (denoted by "X") do not have significantly different performance. All runs scoring below a certain group perform significantly worse than at least the top entry of that group. Likewise, all 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 those runs that score better than the top entry of that group perform significantly better.

As mentioned, it is well-known that it is fairly difficult to detect statistically significant differences between retrieval runs based on 50 queries. While 50 queries remains a good choice based on practicality for doing relevance assessments, statistical testing would be 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 2002 campaign, we have observed a fairly clear division of runs into performance groups for the multilingual track. The top three groups (Université de Neuchâtel, University of California at Berkeley – Group 2 and Eurospider) are within 10% of each other in terms of average precision, but then a considerable gap opens of roughly 20% to the next best group (University of Jaen/SINAI Group). A further 10% down, Océ places fifth, and the Johns Hopkins APL group and Thomson Legal come in at sixth and seventh respectively, another 10% behind Océ. All in all, there is a performance drop of nearly 50% between the top and the seventh entry.

This considerable difference facilitates the detection of significant differences, and groups with similar performance emerge. Interpreting Table 8, we see that the top three groups significantly outperformed all entries of all other groups except University Jaen/SINAI Group. The difference between this, the fourth placing group, and the top three is large, but narrowly misses significance. The top four groups significantly outperform all groups ranked eighth and lower, while the best entries of the top seven show significant difference from at least the bottom three groups.

Arcsine-transformed average precision values	Run ID									
0.64145	bky2muen1	X								
0.63544	UniNEm2	X	X							
0.63227	UniNEm4	X	X							
0.62259	UniNEm5	X	X							
0.61933	bky2muen2	X	X							
0.61808	EIT2MNF3	X	X							
0.61540	EAN2MDF4	X	X							
0.60960	EIT2MNU1	X	X							
0.60692	UniNEm3	X	X							
0.60413	UniNEm1	X	X							
0.59978	EIT2MDF3	X	X							
0.59777	EIT2MDC3	X	X							
0.51680	UJAMLTDRSV2		X	X						
0.51560	UJAMLTD2RSV2		X	X						
0.46963	oce02mulMSlo			X	X					
0.46050	oce02mulRRlo			X	X					
0.43769	oce02mulMSbf			X	X					
0.43746	aplmuenb			X	X					
0.43714	tlren2multi			X	X					
0.43348	aplmuena			X	X					
0.43130	UJAMLTDRR			X	X					
0.42789	UJAMLTDNORM			X	X	X				
0.42484	oce02mulRRbf			X	X	X				
0.38359	oce02mulRRloTO				X	X	X			
0.36336	tremu1				X	X	X			
0.30544	run2					X	X			
0.30063	run1						X	X		
0.29724	tremu2						X	X		
0.28162	UJAMLTDRSV2RR						X	X		
0.26705	run3						X	X	X	
0.23093	iritMEn2All							X	X	X
0.15715	NTUmulti04								X	X
0.15540	NTUmulti05								X	X
0.14879	NTUmulti03								X	X
0.13168	NTUmulti02									X
0.10864	NTUmulti01									X

Table 8. Results of statistical analysis (ANOVA) on the experiments submitted for the multilingual track. 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.

4 Pool Quality and Result Validity

The results reported in the CLEF campaigns rely heavily on the concept of judging the relevance of documents with respect to given topics. The relevance of a document is judged by human assessors, making this a costly undertaking. These relevance assessments are then used for the calculation of the recall/precision figures that underlie the graphs and figures presented to the participants.

Their central importance for the calculation of many popular evaluation measures means that relevance assessments are not without critics. Generally, concerns mentioned focus mostly on two aspects: the "quality" and the "coverage" ("completeness") of the assessments. The first concern stems from the subjective nature of relevance, which can lead to disagreements between different assessors or even when the same assessor judges a document twice. Such disagreements can emerge from, among other things, personal bias of the judge, or a lack of understanding of the topics and documents. There is no "solution" for obtaining universal relevance judgments. Rather, researchers that rely upon the results from an evaluation campaign such as CLEF have to be aware of this issue and its implications. Numerous studies have analyzed the impact of disagreement in judging on the validity of evaluation results. These studies generally conclude that as long as sufficient consistency is maintained during judging, the ranking and comparison of systems is stable even if the absolute performance values calculated on the basis of the assessments change. The quality and consistency of the assessments in CLEF is ensured by following a well-proven methodology based on TREC experience. More details of relevance assessment processes can be found in [9] and in deliverable 3.2.1 [5].

The problem of coverage arises from practical considerations in the production of the relevance assessments. While it is comparatively easy to judge a substantial part of the top-ranked results submitted by participants, it is much harder to judge the documents that were not part of any of the submitted result sets, since the number of such documents is usually far greater than that of the documents retrieved in result sets. This is especially the case with today's large test collections. Judging the non-retrieved documents is necessary to calculate some evaluation measures such as recall.

In order to keep costs manageable, only documents included and highly ranked in at least one result set are judged for relevance (with the union of all judged result sets forming a "document pool"). This implies that some relevant documents potentially go undetected if they are not retrieved by any of the participating systems. The assertion is that a sufficient number of diverse systems will turn up most relevant documents this way. Figures calculated based on these "limited" assessments are then a good approximation of theoretical figures based on complete assessments. A potential problem is the usability of the resulting test collection for the evaluation of a system that did not contribute to this "pool of judged documents". If such a system retrieves a substantial number of unjudged documents that are relevant, but went undetected, it is unfairly penalized when calculating the evaluation measures. An investigation into whether the assessments for the CLEF multilingual collection provide sufficient coverage follows below.

One way to analyze the coverage of the relevance judgments is by focusing on the "unique relevant documents" [11]. For this purpose, a unique relevant document is defined as a document that was judged relevant with respect to a specific topic, but that would not have been part of the pool of judged documents had a certain group not participated in the evaluation, i.e., only one group retrieved the document with a score high enough to have it included in the judgment pool. This addresses the concern that systems not directly participating in the evaluation are unfairly penalized. Subtracting relevant documents only found by a certain group, and then reevaluating the results for this group, simulates the scenario that this group was a non-participant. The smaller the change in performance that is observed, the higher is the probability that the relevance assessments are sufficiently complete.

This kind of analysis has been run by the CLEF consortium since the 2000 campaign for the multilingual track. This year, we have expanded the analysis to include an investigation of the subcollections formed by the individual target languages. A total of $n+1$ sets of relevance assessments are used: the original set, and n sets that are built by taking away the relevant documents uniquely found by one specific participant. The results for every experiment are

then recomputed using the set without the group-specific relevant documents. The key figures obtained after rerunning the evaluations can be found in Table 9.

Multilingual	Mean difference	Max difference	StdDev difference
Absolute	0.0008	0.0030	0.0018
Percentage	0.48%	1.76%	1.0133%

DE German	Mean difference	Max difference	StdDev difference
Absolute	0.0025	0.0095	0.0054
Percentage	0.71%	5.78%	1.7055%

EN English	Mean difference	Max difference	StdDev difference
Absolute	0.0023	0.0075	0.0051
Percentage	1.14%	3.60%	2.5950%

ES Spanish	Mean difference	Max difference	StdDev difference
Absolute	0.0035	0.0103	0.0075
Percentage	0.87%	2.52%	1.8622%

FI Finnish	Mean difference	Max difference	StdDev difference
Absolute	0.0021	0.0100	0.0049
Percentage	0.82%	4.99%	2.0495%

FR French	Mean difference	Max difference	StdDev difference
Absolute	0.0019	0.0050	0.0038
Percentage	0.54%	1.86%	1.0828%

IT Italian	Mean difference	Max difference	StdDev difference
Absolute	0.0008	0.0045	0.0016
Percentage	0.22%	0.93%	0.4608%

NL Dutch[3]	Mean difference	Max difference	StdDev difference
Absolute	0.0045	0.0409	0.0116
Percentage	1.26%	9.15%	3.0907%

SV Swedish	Mean difference	Max difference	StdDev difference
Absolute	0.0082	0.0306	0.0182
Percentage	3.32%	10.19%	7.5054%

Table 9. Key values of the pool quality analysis: mean and maximum change in average precision when removing the pool contribution of one participant, and associated standard deviation.

The quality of a document pool can therefore be judged by the mean performance difference in terms of average precision that is obtained if the pool had been missing the contribution of a specific group. This difference should be as small as possible, indicating that the pool is "sufficiently exhaustive" and that adding more documents to the pool, such as documents

[3] One experiment that was an extreme outlier in terms of performance was removed before calculation of the Dutch figures to avoid a non-representative skew in the numbers.

found by an additional participant, does not substantially influence results and/or rankings. As we also found in 2000 and 2001, the pool used for the multilingual track is very stable. The maximum change in performance scores is 1.76%. These small differences influence only direct comparisons between systems that have practically identical performance, and where the original performance differences are not considered significant in any case. The value of the multilingual pool for reuse in post-hoc experiments should thus be assured, and the validity of the results reported by CLEF should be given within the inherent limits of interpretation (restricted set of queries, characteristics of evaluation measure and others).
The pools for individual target languages are smaller, since they are restricted to the document set of that language. Only runs for that language, and therefore a smaller number than for the multilingual pool, contributed. It is therefore not surprising that differences found for the individual languages are therefore a little higher than for the multilingual pool. We feel, however, that they are still comfortably within acceptable limits, and they do indeed compare favorably with numbers reported for comparable collections in the past [2], [3]. Not surprisingly, the pool for Swedish is a little less stable than the others, owing to having less contributions and Swedish being a new language in CLEF. For Dutch, we had issues with a few outliers that obfuscate the measures somewhat, but we believe that the pool should be of comparable quality to the other languages.

5 Conclusions

We have reported on the results obtained for the 2002 campaign and their interpretation. CLEF 2002 experienced growth both in the number of participants, and more noticeably, in the number of experiments submitted. This deliverable summarizes the main characteristics of the 282 experiments submitted for the campaign, and discusses trends observed and the main results. The statistical significance of the results is then explored for the main, multilingual track. Lastly, we investigate on the validity of the results by analyzing the completeness of the relevance assessment pools, which is critical for calculating the performance measures used by CLEF.
In summary, we can conclude that people adopt each other's ideas and methods across campaigns, and that those returning groups that have the experience to build complex combination systems have performed well in the main, multilingual track. This demonstrates clearly the learning curve that these participants have completed. For the bilingual and monolingual track we observe that good performance in a wide variety of target languages requires careful fine tuning for all these languages. The main challenge is to do this in a manner that respects the characteristics of each language without overturning to artifacts of using the CLEF document collection.
A further encouraging development is a clear trend that returning groups move from simpler to harder tasks from one campaign to the next. CLEF clearly helps finding these groups an entry into the more challenging CLIR tasks. This is especially valuable considering the large number of new European groups attracted by CLEF in the last campaigns.
Statistical analysis allows to qualify and better interpret the results as published by CLEF. As evidenced by an analysis of the multilingual experiments that we present, fairly large performance differences are needed to reach a level of statistical significance. For this kind of testing, having a maximum number of queries available is of great benefit. The CLEF consortium strives for stability in the test collections to allow post-hoc experiments with combined resources from several campaigns for this reason.
Finally, the results published by CLEF are only as good as the data they build on. We investigate the quality of the relevance assessments by investigating their completeness through pool quality evaluation. We find that the CLEF relevance assessments seem to be very stable, making them suitable for reuse in post-hoc experiments, and further validating the results published during the campaigns.

References

[1] Blustein, J.: IR STAT PAK. URL: http://www.csd.uwo.ca/~jamie/IRSP-overview.html
[2] Braschler, M.: CLEF 2000 – Overview of Results. In Peters, C. (Ed.) Cross-Language Information Retrieval and Evaluation. Workshop of the Cross-Language Evaluation Forum, CLEF 2000, Revised Papers. Pages 89-101
[3] Braschler, M.: CLEF 2001 – Overview of Results. In Peters, C., Braschler, M., Gonzalo, J., Kluck, M. (Eds): Evaluation of Cross-Language Information Retrieval Systems, Second Workshop of the Cross-Language Evaluation Forum, CLEF 2001, pages 9-26, 2002.
[4] CLEF Consortium: Deliverable 2.3.1, Multilingual Collection for Campaign 1.
[5] CLEF Consortium: Deliverable 3.2.1, Test Collection Report for Campaign 1.
[6] Harman, D., Braschler, M., Hess, M., Kluck, M., Peters, C., Schäuble, P., Sheridan P.: CLIR Evaluation at TREC. In Peters, C. (Ed.) Cross-Language Information Retrieval and Evaluation. Workshop of the Cross-Language Evaluation Forum, CLEF 2000, Revised Papers. Pages 7-23.
[7] Hull, D. A: Using Statistical Testing in the Evaluation of Retrieval Experiments. In Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Pittsburg, USA, 1993.
[8] Peters, C. (Ed).: Results of the CLEF 2002 Cross-Language System Evaluation Campaign. Working Notes for the CLEF 2002 Workshop.
[9] Peters, C., Braschler, M.: European research letter: Cross-language system evaluation: The CLEF campaigns, In Journal of the American Society for Information Science and Technology, Volume 52, Issue 12, pages 1067-1072
[10] Tague-Sutcliffe, J., Blustein, J.: A Statistical Analysis of the TREC-3 Data. In Proceedings of the Third Text REtrieval Conference (TREC-3), NIST Special Publication 500-226. Page 385ff.
[11] Zobel, J.: How reliable are the results of large-scale information retrieval experiments? In Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (1998)