CLEF 2005: Ad Hoc Track Overview

Giorgio M. Di Nunzio[1], Nicola Ferro[1], Gareth J.F. Jones[2] and Carol Peters[3]

[1]Department of Information Engineering, University of Padua, Italy
{dinunzio|ferro}@dei.unipd.it
[2]School of Computing, Dublin City University, Ireland
gjones@computing.dcu.ie
[3]ISTI-CNR, Area di Ricerca, 56124 Pisa, Italy
carol.peters@isti.cnr.it

Abstract. We describe the objectives and organization of the CLEF 2005 ad hoc track and discuss the main characteristics of the tasks offered to test monolingual, bilingual and multilingual textual document retrieval. The performance achieved for each task is presented and a preliminary analysis of results is given. The paper focuses in particular on the multilingual tasks which reused the test collection created in CLEF 2003 in an attempt to see if an improvement in system performance over time could be measured, and also to examine the multilingual results merging problem.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval; H.3.4 [**Systems and Software**]: Performance evaluation.

General Terms

Experimentation, Performance, Measurement, Algorithms.

Additional Keywords and Phrases

Multilingual Information Access, Cross-Language Information Retrieval

1 Introduction

The Ad Hoc retrieval track is generally considered as the core track in CLEF. The aim of this track is to promote the development of monolingual and cross-language textual document retrieval systems. As in past years, the CLEF 2005 ad hoc track was structured in three tasks, testing systems for monolingual (querying and finding documents in one language), bilingual (querying in one language and finding documents in another language) and multilingual (querying in one language and finding documents in multiple languages) retrieval, thus helping groups to make the progression from simple to more complex tasks. The document collections used were taken from the CLEF multilingual comparable corpus of news documents.

The **Monolingual** and **Bilingual** tasks were principally offered for Bulgarian, French, Hungarian and Portuguese target collections. Additionally, in the bilingual task only, newcomers (i.e. groups that had not previously participated in a CLEF cross-language task) or groups using a "new-to-CLEF" query language could choose to search the English document collection. The aim in all cases was to retrieve relevant documents from the chosen target collection and submit the results in a ranked list.

The **Multilingual** task was based on the CLEF 2003 multilingual-8 test collection which contained news documents in eight languages: Dutch, English, French, German, Italian, Russian, Spanish, and Swedish. There were two subtasks. a traditional multilingual retrieval task requiring participants to carry out retrieval and merging (Multi-8 Two-Years-On), and a new task focussing only on the multilingual results merging problem using standard sets of ranked retrieval output (Multi-8 Merging Only). One of the goals for the first task was to see whether it is possible to measure progress over time in multilingual system performance at CLEF by reusing a test collection created in a previous campaign. In running the merging only task our aim was to encourage participation by researchers interested in exploring the multilingual merging problem without the need to build retrieval systems for the document languages.

In this paper we describe the track setup, the evaluation methodology and the participation in the different tasks (Section 2), and present the main characteristics of the experiments and show the results (Sections 3 - 5). The final section provides a brief summing up. For information on the various approaches and resources used by the groups participating in this track and the issues they focused on, we refer the reader to the other papers in these Working Notes.

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 [1] in the late 1960s. 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 recall and precision measures. 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 Test Collection

This year, for the first time, separate test collections were used in the ad hoc track: the monolingual and bilingual tasks were based on document collections in Bulgarian, English, French, Hungarian and Portuguese, whereas the two multilingual tasks reused a test collection – documents, topics and relevance assessments - created in CLEF 2003.

Documents: The document collections used for the CLEF 2005 ad hoc tasks are part of the CLEF multilingual corpus of news documents described in the Introductory paper to these Working Notes [2]. In the monolingual and bilingual tasks, the English, French and Portuguese collections consisted of national newspapers and news agencies for the period 1994 and 1995. Different variants were used for each language. Thus, for English we had both US and British newspapers, for French we had a national newspaper of France plus Swiss French news agencies, and for Portuguese we had national newspapers from both Portugal and Brazil. This meant that, for each language, there were significant differences in orthography and lexicon over the sub-collections. This is a real world situation and system components, i.e. stemmers, translation resources, etc., should be sufficiently robust to handle such variants. The Bulgarian and Hungarian collections used in these tasks were new in CLEF 2005 and consisted of national newspapers for the year 2002[1]. This meant that the collections we used in the ad hoc mono- and bilingual tasks this year were not all for the same time period. This had important consequences on topic creation. For the multilingual tasks we reused the CLEF 2003 multilingual document collection. This consisted of news documents for 1994-95 in the 8 languages listed above in the Introduction.

Topics: Topics in CLEF are structured statements representing information needs; the systems use the topics to derive their queries. Each topic consists of three parts: a brief "title" statement; a one-sentence "description"; a more complex "narrative" specifying the relevance assessment criteria. Sets of 50 topics were created for the CLEF 2005 ad hoc mono- and bilingual tasks. One of the decisions taken early on in the organization of the CLEF ad hoc tracks was that the same set of topics would be used to query all collections, whatever the task. There are a number of reasons for this: it makes it easier to compare results over different collections, it means that there is a single master set that is rendered in all query languages, and a single set of relevance assessments for each language is sufficient for all tasks. However, the fact that the collections used in the CLEF 2005 ad hoc mono- and bilingual tasks were from two different time periods (1994-1995 and 2002) made topic creation particularly difficult. It was not possible to create time-dependent topics that referred to particular date-specific events as all topics had to refer to events that could have been reported in any of the collections, regardless of the dates. This meant that the CLEF 2005 topic set is somewhat different from the sets of previous years as the topics tend to be of broad coverage. For this reason, it was difficult to construct topics that would find a limited number of relevant documents in each collection, and a – probably excessive – number of topics used for the

[1] It proved impossible to find national newspapers in electronic form for 1994 and/or 1995 in these languages.

2005 mono- and bilingual tasks have a very large number of relevant documents. We have yet to analyze the possible impact of this fact on results calculation, but we suspect that it has meant that this year's ad hoc test collection is less effective in "discriminating" between the performance of different systems.

The topic sets for the mono- and bilingual tasks were prepared in thirteen languages: Amharic, Bulgarian, Chinese, English, French, German, Greek, Hungarian, Indonesian, Italian, Portuguese, Russian, and Spanish. Twelve were actually used and, as usual, English was by far the most popular. To counter this, in previous years, we placed restrictions on the possible topic languages for the bilingual task. We will probably reinstate some such constraint in CLEF 2006 in order to promote the testing of systems with less common languages.

For the multilingual task, the CLEF 2003 Dutch, English and Spanish sets of 60 topics were used. They were divided into two sets: 20 topics for training and 40 for testing.

Here below we give the English version of a typical topic from CLEF 2005:

```
<top><num> C254 </num>
<EN-title> Earthquake Damage </EN-title>
<EN-desc> Find documents describing damage to property or persons caused by an
earthquake and specifying the area affected. </EN-desc>
<EN-narr> Relevant documents will provide details on damage to buildings and
material goods or injuries to people as a result of an earthquake. The geographical
location (e.g. country, region, city) affected by the earthquake must also be
mentioned. </EN-narr></top>
```

2.3 Relevance Assessment

Relevance assessment for the mono- and bilingual tasks was performed by native speakers. The multilingual tasks used the relevance assessments of 2003. 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.

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 2005 are given in Appendix at the end of these Working Notes. The stability of pools constructed in this way and their reliability for post-campaign experiments is discussed in [3] 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. This year we have used the new package written by Chris Buckley for TREC (trec_eval 7.3) and available from the TREC website

As a consequence of limited evaluation resources, a maximum of 4 runs for each multilingual task and a maximum of 12 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 32 runs in total for the multilingual, bilingual and monolingual tasks.

2.5 Result Calculation

Evaluation campaigns such as TREC and CLEF are based on the belief that 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 [4].

2.6 Participants and Experiments

As shown in Table 1, a total of 23 groups from 15 different countries submitted results for one or more of the Ad-hoc tasks - a slight decrease on the 26 participants of last year. A total of 254 experiments were submitted, nearly the same as the 250 experiments of 2003. Thus, there is a slight increase in the average number of submitted runs per participant: from 9.6 runs/participant of 2004 to 11 runs/participant of this year.

Table 1. CLEF 2005 ad hoc participants – new groups are indicated by *

Budapest U. Tech.&Econom (Hungary)*	U.Amsterdam - Informatics (Netherlands)
CLIPS-Grenoble (France)	U.Buffalo - SUNY - Informatics (USA)
CMU - Lang.Tech. (USA)	U.Geneva - Inf.Systems (Switzerland)*
Daedalus & Madrid Univs (Spain)	U.Glasgow - IR (UK)
Dublin City U. - Computing. (Ireland)	U.Hildesheim - Inf.Sci (Germany)
ENS des Mines St Etienne (France)*	U.Indonesia - Comp.Sci (Indonesia)*
Hummingbird Core Tech. (Canada)	U.Jaen - Intell.Systems (Spain)
Johns Hopkins U. (USA)	U.Lisbon (Portugal)
Moscow State U.-Computing (Russia)*	U.Neuchatel (Switzerland)
Swedish Inst.Comp.Sci (Sweden)	U.Stockholm, NLP (Sweden)
Thomson Legal Regulatory (USA)	U.Surugadai - Cultural Inf. (Japan)
U. Alicante - Comp.Sci + (Spain)	

Table 2. CLEF 2005 ad hoc experiments

Track	# Participants	# Runs (%)	
AH-2-years-on	4	21	8.27%
AH-Merging	3	20	7.87%
AH-Bilingual-X2BG	4	12	4.72%
AH-Bilingual-X2EN	4	13	5.12%
AH-Bilingual-X2FR	9	31	12.20%
AH-Bilingual-X2HU	3	7	2.76%
AH-Bilingual-X2PT	8	28	11.02%
AH-Monolingual-BG	7	20	7.87%
AH-Monolingual-FR	12	38	14.97%
AH-Monolingual-HU	10	32	12.60%
AH-Monolingual-PT	9	32	12.60%
TOTAL		**254**	**100.00%**

As stated, participants were required to submit at least one title+description ("TD") run per task in order to increase comparability between experiments. The large majority of runs (188 out of 254, 74.02%) used this combination of topic fields, 54 (21.27%) used all fields, 10 (3.94%) used the title field, and only 2 (0.79%) used the description 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. A breakdown into the separate tasks is shown in Table 2.

Thirteen different topic languages were used for ad hoc experiments– the Dutch run was in the multilingual tasks and used the CLEF 2003 topics. As always, the most popular language for queries was English, and French was second. Note that Bulgarian and Hungarian, the new collections added this year, were also quite popular as new monolingual tasks – Hungarian was also used in one case a topic language in a bilingual run. The number of runs per topic language is shown in Table 3.

Table 3. List of experiments by topic language

Language[2]		# Runs	(%)
EN	English	85	33.47%
FR	French	42	16.54%
HU	Hungarian	33	12.99%
PT	Portuguese	32	12.60%
BG	Bulgarian	20	7.87%
ES	Spanish	15	5.91%
ID	Indonesian	8	3.15%
DE	German	6	2.36%
AM	Amharic	4	1.57%
GR	Greek	3	1.18%
IT	Italian	3	1.18%
RU	Russian	2	0.79%
NL	Dutch	1	0.39%
TOTAL		**254**	**100.00%**

3 Monolingual Experiments

As stated, monolingual retrieval was offered for the following target collections: Bulgarian, French, Hungarian, and Portuguese. As can be seen from Table 2, the number of participants and runs for each language was quite similar, with the exception of Bulgarian, which has a slightly smaller participation. This year just 5 groups out of 16 (31,25%) submitted monolingual runs only (down from ten groups last year), and just one of these groups was a first time participant in CLEF. This is in contrast with previous years where many new groups only participated in monolingual experiments. This year, most of the groups submitting monolingual runs were doing this as part of their bilingual or multilingual system testing activity.

Table 4 shows the top five groups for each target collection, ordered by mean average precision. The table reports: the short name of the participating group; the run identifier, specifying whether the run has participated in the pool or not, and the page in Appendix A containing all figures and graphs for this run; the mean average precision achieved by the run; and the performance difference between the first and the last participant. The pages of appendix A containing the overview graphs are indicated under the name of the sub-task. Table 4 regards runs using title + description fields only (the mandatory run).

All the groups in the top five had participated in previous editions of CLEF. Both pooled and not pooled runs are in the best entries for each track. Finally, it can be noticed that the trend observed in the previous editions of CLEF is confirmed: differences for top performers for tracks with languages introduced in past campaigns are small: in particular only 5.35% in the case of French (French monolingual has been offered in CLEF since 2000) and 7.55% in the case of Portuguese, which was introduced last year. However, for the new languages, Bulgarian and Hungarian, the differences are much greater, in the order of 25%, showing that there should be room for improvement if these languages are offered in future campaigns.

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

Table 4. Best entries for the monolingual track (title+description topic fields only). Additionally, the performance difference between the best and the last (up to 5) placed group is given (in terms of average precision).

Track		Participant Rank					
		1st	2nd	3rd	4th	5th	Diff.
Monolingual Bulgarian (A.45 – A.46)	Part.	jhu/apl	hummingbird	Unine	miracle	u.glasgow	
	Run	aplmobgd pooled (A.232)	humBG05tde pooled (A.230)	UniNEbg3 not pooled (A.242)	ST pooled (A.235)	glabgtdqe not pooled (A.239)	
	Avg. Prec.	32.03%	29.18%	28.39%	26.76%	25.14%	27.41%
Monolingual French (A.49 – A.50)	Part.	jhu/apl	unine	u.Glasgow	hummingbird	tlr	
	Run	aplmofra pooled (A.261)	UniNEfr1 pooled (A.278)	glafrtdqe1 pooled (A.275)	humFR05tde not pooled (A.260)	tlrTDfrRFS1 pooled (A.273)	
	Avg. Prec.	42.14%	42.07%	40.17%	40.06%	40.00%	5.35%
Monolingual Hungarian (A.53 – A.54)	Part.	jhu/apl	unine	miracle	hummingbird	hildesheim	
	Run	aplmohud pooled (A.294)	UniNEhu3 not pooled (A.312)	xNP01ST1 pooled (A.297)	humHU05tde pooled (A.288)	UHIHU2 pooled (A.285)	
	Avg. Prec.	41.12%	38.89%	35.20%	33.09%	32.64%	25.98%
Monolingual Portuguese (A.57 – A.58)	Part.	Unine	hummingbird	Tlr	jhu-apl	alicante	
	Run	UniNEpt2 pooled (A.338)	humPT05tde not pooled (A.322)	tlrTDptRF2 not pooled (A.332)	aplmopte not pooled (A.326)	IRn−pt−vexp pooled (A.314)	
	Avg. Prec.	38.75%	38.64%	37.42%	36.54%	36.03%	7.55%

4 Bilingual Experiments

The bilingual task was structured in four subtasks (X → BG, FR, HU or PT target collection) plus, as usual, an additional subtask with English as target language – this last task was restricted to newcomers in a CLEF cross-language task or to groups using unusual or new topic languages (Amharic, Greek, Indonesian, and Hungarian) Table 5 shows the best results for this task. Note that both pooled and not pooled runs are in the best entries for each track, with the exception of Bilingual X → EN.

For bilingual retrieval evaluation, a common method is to compare results against monolingual baselines. We have the following results for CLEF 2005:

- X -> FR: 85% of best monolingual French IR system
- X -> PT: 88% of best monolingual Portuguese IR system
- X -> BG: 74% of best monolingual Bulgarian IR system
- X -> HU: 73% of best monolingual Hungarian IR system

Similarly to monolingual, this is an interesting result. Whereas, the figures for French and Portuguese reflect those of recent literature [5], it can be seen that for the new languages where there has been little CLIR system experience and testing so far, there is much room for improvement. It is interesting to note that when CLIR system evaluation began 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

Table 5. Best entries for the bilingual task (title+description topic fields only). The performance difference between the best and the last (up to 5) placed group is given (in terms of average precision).

Track		Participant Rank					
		1st	2nd	3rd	4th	5th	Diff.
Bilingual Bulgarian (A.25 – A.26)	Part.	miracle	unine	u.glasgow	jhu/apl		
	Run	ENXST pooled (A.135)	UniNEbibg3 not pooled (A.143)	glaenbgtd pooled (A.136)	aplbienbge pooled (A.133)		
	Avg. Prec.	23.55%	13.99%	12.04%	9.59%		145.57%
Bilingual French (A.33 – A.34)	Part.	alicante	unine	hildesheim	jhu/apl	miracle	
	Run	IRn−enfr−vexp not pooled (A.158)	UniNEbifr2 not pooled (A.186)	UHIENFR2 not pooled (A.169)	aplbienfrc pooled (A.171)	ENSST not pooled (A.172)	
	Avg. Prec.	35.90%	34.67%	34.65%	34.42%	30.76%	16.71%
Bilingual Hungarian (A.37 – A.38)	Part.	miracle	unine	jhu/apl			
	Run	ENMST not pooled (A.190)	UniNEbihu3 not pooled (A.194)	aplbienhue not pooled (A.189)			
	Avg. Prec.	30.16%	28.82%	24.58%			22.70%
Bilingual Portuguese (A.41 – A.42)	Part.	unine	jhu/apl	miracle	alicante	tlr	
	Run	UniNEbipt1 pooled (A.216)	aplbiesptb not pooled (A.204)	ESAST not pooled (A.209)	IRn−enpt−vexp not pooled (A.197)	tlrTDfr2ptRFS1 pooled (A.212)	
	Avg. Prec.	34.04%	31.85%	31.06%	29.18%	23.58%	44.36%
Bilingual English (A.29 – A.30)	Part.	jhu/apl	u.glasgow	depok			
	Run	aplbiidena pooled (A.152)	glagrentdqe pooled (A.156)	UI−TD10 pooled (A.146)			
	Avg. Prec.	33.13%	29.35%	12.85%			157,82%

5 Multilingual Experiments

Table 6 shows results for the best entries for the multilingual tasks and is organized similarly to Table 4. Additional rows for each task show the difference in the MAP for this run compared to the best performing run at this rank in the original CLEF 2003 Multi-8 track.. The final row of the table shows the results of the top 5 group submissions of the CLEF 2003 Multi-8 track for comparison with the 2-Years-On and Merging tracks of this year.

Since the CLEF 2005 tracks used only 40 topics of the original 60 topics of the 2003 as the test set (topics 161 to 200), while the first 20 topics (topics 141 to 160) were used as a training set, the average precision of the original 2003 runs had to be recomputed for the 40 test topics used this year. These revised MAP figures are reported in Table 6. These figures are slightly different from the original results which appear in the CLEF 2003 proceedings [3], although the ranking of these runs is unchanged.

It can be seen from Table 6 that the performance difference between the first and the last participant for the 2-Years-On track is much greater (nearly 3 times) than the corresponding difference in 2003, even if the task performed in these two tracks is the same. On the other hand, the performance difference for the Merging track is nearly one third of the corresponding difference in 2003: it seems that merging the results of the run reduces the gap between the best and the last performer, even though there is still a considerable difference (35.63%) if compared to the small differences of most popular monolingual languages, e.g. 5.35% of monolingual French.

We can note that the top participant of the 2-Years-On track achieves a 15.89% performance improvement with respect to the top participant of CLEF 2003 Multi-8. On the other hand, the fourth participant of the 2-Years-On track has a 59.15% decrease in performance with respect to the fourth participant of CLEF 2003 Multi-8. Similarly, we can note that the top participant of the Merging track achieves a 6.24% performance improvement with respect to the top participant of 2003.

In general, we can note that for the 2-Years-On track there is a performance improvement only for the top participant, while the performances deteriorate quickly for the other participants with respect to 2003. On the other hand, for the Merging track the performance improvement of the top participant with respect to 2003 is less than in the case of the 2-Years-On track. . There is also less variation between the submissions for the Merging task than seen in the earlier 2003 runs. This is probably due to the fact that the participants were using the same ranked lists, and that the variation in performance arises only from the merging strategies adopted.

Table 6. Best entries for the multilingual task (title+description topic fields only). The performance difference between the best and the last (up to 5) placed group is given (in terms of average precision).

Track		Participant Rank					
		1st	2nd	3rd	4th	5th	Diff.
Multilingual 2 Years On (A.17 – A.18)	Part.	Cmu	jaen	miracle	isi-unige		
	Run	adhocM5Trntes not pooled (A.93)	UJAPRFRSV2RR not pooled (A.101)	esml9XstiSTp not pooled (A.110)	AUTOEN not pooled (A.96)		
	Avg. Prec.	44.93%	29.57%	26.06%	10.33%		334.95%
	Diff. 2003	+15.89%	–17.34%	–12.02%	–59.15%		
Multilingual Merging (A.21 – A.22)	Part.	Cmu	dcu	Jaen			
	Run	UNET150w05test (A.118)	dcu.Prositqgm2 (A.121)	UJAMENEDFRR (A.129)			
	Avg. Prec.	41.19%	32.86%	30.37%			35.63%
	Diff. 2003	+6.24%	–7.93%	+2.53%			
Multilingual CLEF 2003	Part.	UC Berkely	U. Neuchatel	U. Amsterdam	jhu/apl	U. Tampere	
	Run	bkmul8en3 pooled	UniNEml1 not pooled	UAmsC03EnM8SS4G not pooled	aplmuen8b not pooled	UTAmul1 pooled	
	Avg. Prec.	38.77%	35.69%	29.62%	25.29%	18.95%	104.59%

Figures 1 and 2 compare the performances of the top participants of the 2-Years-On track with respect to the top and the fifth performer of CLEF 2003 Multilingual-8. Figure 1 shows the average precision at the different interpolated recall levels, while Figure 2 shows the precision at different document cut-off values. Figures 3 and 4 show corresponding results for the Multilingual Merging task. Trends in these figures are similar to those seen in Table 6. The top performing submission to the Multilingual 2-Years-On and Merging tasks are both clearly higher than the best submission to the CLEF 2003 task. The variation between submissions for 2-Years-On is also greater than that observed for the Merging only task.

6 Conclusions

From a first rapid glance at the reports from the groups that participated in the bilingual ad hoc tasks, it appears that this year's experiments provide a good overview of most of the traditional approaches to CLIR when matching between query and target collection, including n-gram indexing, machine translation, machine-readable bilingual dictionaries, multilingual ontologies, pivot languages, query and document translation – perhaps corpus-based approaches were less used than in previous years continuing a trend first noticed in CLEF 2004. Veteran groups were mainly concerned with fine tuning and optimizing strategies already tried in previous years. The issues examined were the usual ones: word-sense disambiguation, out-of-dictionary vocabulary, ways to apply relevance feedback, results merging, etc.

Figure 1. Comparison between Multilingual 2-Years-On and CLEF 2003 Multilingual-8. Interpolated Recall vs Average Precision.

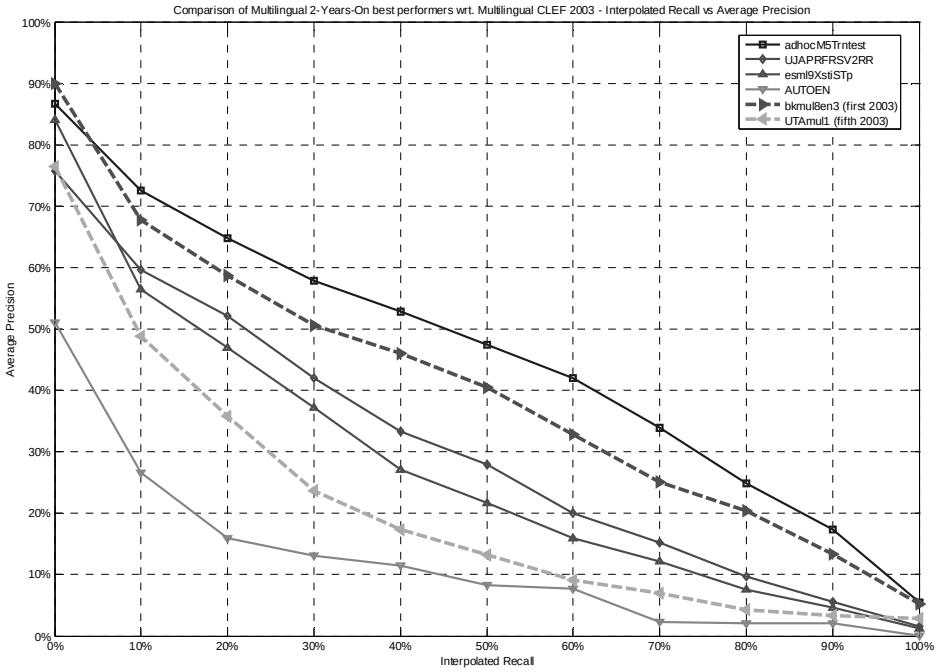


Figure 2. Comparison between Multilingual 2-Years-On and CLEF 2003 Multilingual-8. Document Cut-off Values vs Precision.

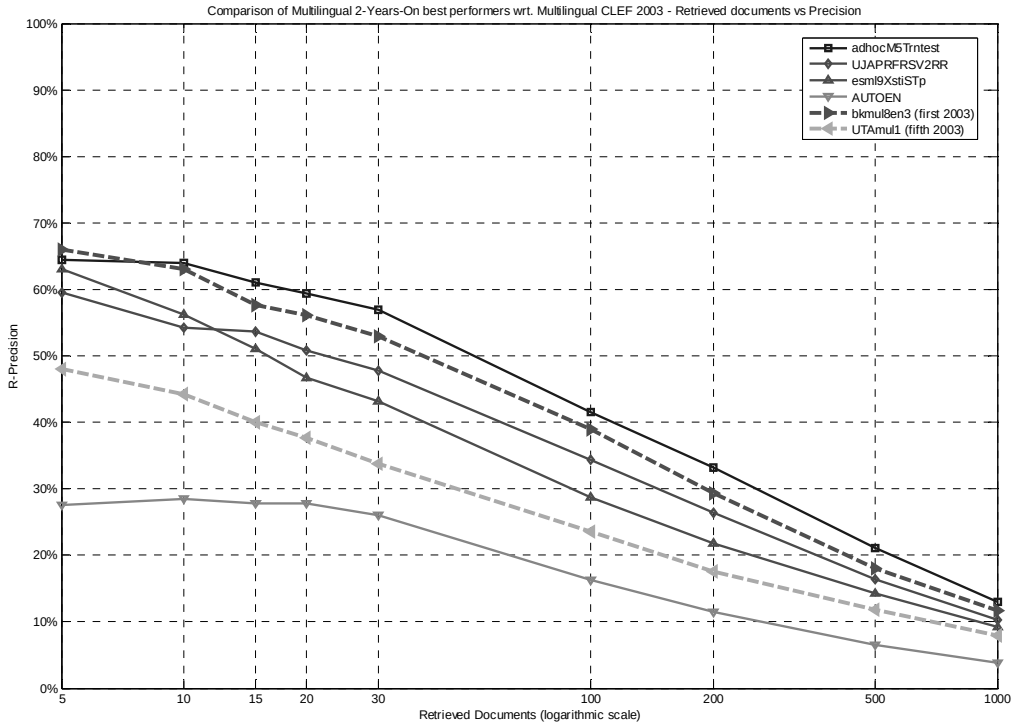


Figure 3. Comparison between Multilingual Merging and CLEF 2003 Multilingual-8. Interpolated Recall vs Average Precision.

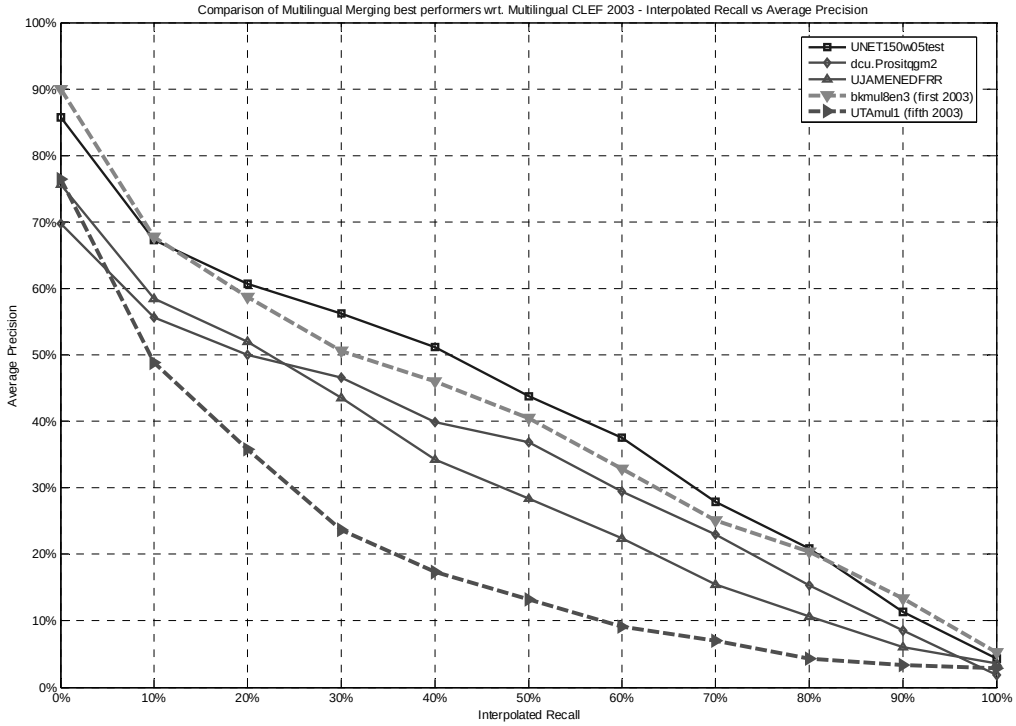
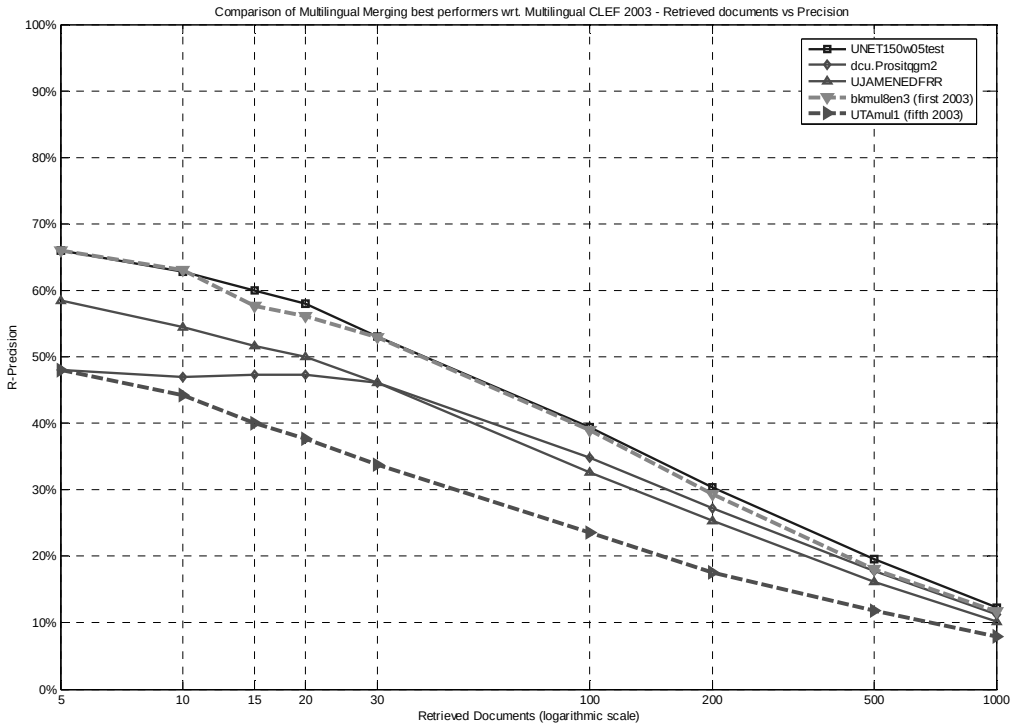


Figure 4. Comparison between Multilingual Merging and CLEF 2003 Multilingual-8. Document Cut-off Values vs Precision.



Although, as has already been mentioned, English was by far the most popular language for queries, some less common and interesting query to target language pairs were tried, e.g. Amharic, Spanish and German to French, and French to Portuguese. The track overview paper in the post-workshop proceedings will provide a more in depth analysis of the approaches adopted for these tasks in CLEF 2005. One of the objectives will be to see if the hypothesis concerning a "blueprint for a successful CLIR system" proposed in [6] can be confirmed.

A main focus in the monolingual tasks was the development of new or the adaptation of existing stemmers and/or morphological analysers for the "new" CLEF languages: Bulgarian and Hungarian. Any comments on the outcomes?

The multilingual tasks at CLEF 2005 were intended to assess whether re-use of the CLEF 2003 Multi-8 task data could be used as an indication of progress in multilingual information retrieval and to provide common sets of ranked lists to enable specific exploration of merging strategies for multilingual IR. The submissions to these tasks show that multilingual performance can indeed be improved beyond that reported at CLEF 2003 both when performing the complete retrieval process and when merging ranked result lists generated by other groups. The initial running of this task suggests that there is scope for further improvement in multilingual IR from exploiting ongoing improvements in IR methods, but also from focused exploration of merging techniques.

Encouraged by the results of the multilingual tasks, we are currently considering running a similar X-years-on task for the mono- and/or bilingual experiments in CLEF 2006, again with the aim of seeing if it is possible to measure progress over time by testing new or updated systems against existing test collections.

References

1. 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.
2. Peters, C.: What happened in CLEF 2005? In this volume.
3. 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.
4. 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.
5. Gonzalo, J., Peters, C. (2005). The Impact of Evaluation on Multilingual Text Retrieval, Proc. SIGIR 2005, 603-604
6. Braschler, M., Peters, C. (2004). Cross-Language Evaluation Forum: Objectives, Results, Achievements, Information Retrieval, Vol.7 (1-2), pp.5-29