CLEF 2006: Ad Hoc Track Overview

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

[1] Department of Information Engineering, University of Padua, Italy
{dinunzio, ferro}@dei.unipd.it
[2] Information Science, University of Hildesheim – Germany
mandl@uni-hildesheim.de
[3] ISTI-CNR, Area di Ricerca – 56124 Pisa – Italy
carol.peters@isti.cnr.it

Abstract. We describe the objectives and organization of the CLEF 2006 ad hoc track and discuss the main characteristics of the tasks offered to test monolingual, bilingual, and multilingual textual document retrieval systems. The track was divided into two streams. The main stream offered mono- and bilingual tasks using the same collections as CLEF 2005: Bulgarian, English, French, Hungarian and Portuguese. The second stream, designed for more experienced participants, offered the so-called "robust task" which used test collections from previous years in six languages (Dutch, English, French, German, Italian and Spanish) with the objective of privileging experiments which achieve good stable performance over all queries rather than high average performance. The performance achieved for each task is presented and the results are commented. The document collections used were taken from the CLEF multilingual comparable corpus of news documents.

1 Introduction

The ad hoc retrieval track is generally considered to be the core track in the *Cross-Language Evaluation Forum (CLEF)*. The aim of this track is to promote the development of monolingual and cross-language textual document retrieval systems. The CLEF 2006 ad hoc track was structured in two streams. The main stream offered monolingual tasks (querying and finding documents in one language) and bilingual tasks (querying in one language and finding documents in another language) using the same collections as CLEF 2005. The second stream, designed for more experienced participants, was the "robust task", aimed at finding relevant documents for difficult queries. It used test collections developed in previous years.

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

C. Peters et al. (Eds.): CLEF 2006, LNCS 4730, pp. 21–34, 2007.

cases was to retrieve relevant documents from the chosen target collection and submit the results in a ranked list.

The **Robust** task offered monolingual, bilingual and multilingual tasks using the test collections built over three years: CLEF 2001 - 2003, for six languages: Dutch, English, French, German, Italian and Spanish. Using topics from three years meant that more extensive experiments and a better analysis of the results were possible. The aim of this task was to study and achieve good performance on queries that had proved difficult in the past rather than obtain a high average performance when calculated over all queries.

In this paper we describe the track setup, the evaluation methodology and the participation in the different tasks (Section 2), 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 the Ad Hoc section of these Proceedings.

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 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 these 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 Collections

Different test collections were used in the ad hoc task in 2006. The main (i.e. non-robust) monolingual and bilingual tasks used the same document collections as in Ad Hoc 2005 but new topics were created and new relevance assessments made. As has already been stated, the test collection used for the robust task was derived from the test collections previously developed at CLEF. No new relevance assessments were performed for this task.

Documents. The document collections used for the CLEF 2006 ad hoc tasks are part of the CLEF multilingual corpus of newspaper and news agency documents

Table 1. Document collections for the main stream Ad Hoc tasks

Language	Collections
Bulgarian	Sega 2002, Standart 2002
English	LA Times 94, Glasgow Herald 95
French	ATS (SDA) 94/95, Le Monde 94/95
Hungarian	Magyar Hirlap 2002
Portuguese	Público 94/95; Folha 94/95

Table 2. Document collections for the Robust task

Language	Collections
English	LA Times 94, Glasgow Herald 95
French	ATS (SDA) 94/95, Le Monde 94
Italian	La Stampa 94, AGZ (SDA) 94/95
Dutch	NRC Handelsblad 94/95, Algemeen Dagblad 94/95
German	Frankfurter Rundschau 94/95, Spiegel 94/95, SDA 94
Spanish	EFE 94/95

described in the Introduction to these Proceedings. The Bulgarian and Hungarian collections used in these tasks were new in CLEF 2005 and consist of national newspapers for the year 2002[1]. This has meant using collections of different time periods for the ad-hoc mono- and bilingual tasks. This had important consequences on topic creation. Table 1 shows the collections used for each language.

The robust task used test collections containing data in six languages (Dutch, English, German, French, Italian and Spanish) used at CLEF 2001, CLEF 2002 and CLEF 2003. There are approximately 1.35 million documents and 3.6 gigabytes of text in the CLEF 2006 "robust" collection. Table 2 shows the collections used for each language.

Topics. Sets of 50 topics were created for the CLEF 2006 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 were 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, in CLEF 2005 the assessors found that the fact that the collections used in the CLEF 2006 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

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

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 all tend to be of broad coverage. In fact, it was difficult to construct topics that would find a limited number of relevant documents in each collection, and consequently a - probably excessive - number of topics used for the 2005 mono- and bilingual tasks have a very large number of relevant documents.

For this reason, we decided to create separate topic sets for the two different time-periods for the CLEF 2006 ad hoc mono- and bilingual tasks. We thus created two overlapping topic sets, with a common set of time independent topics and sets of time-specific topics. 25 topics were common to both sets while 25 topics were collection-specific, as follows:

- Topics C301 - C325 were used for all target collections
- Topics C326 - C350 were created specifically for the English, French and Portuguese collections (1994/1995)
- Topics C351 - C375 were created specifically for the Bulgarian and Hungarian collections (2002).

This meant that a total of 75 topics were prepared in many different languages (European and non-European): Bulgarian, English, French, German, Hungarian, Italian, Portuguese, and Spanish plus Amharic, Chinese, Hindi, Indonesian, Oromo and Telugu. Participants had to select the necessary topic set according to the target collection to be used.

Below we give an example of the English version of a typical CLEF topic:

```
<top> <num> C302 </num>
<EN-title> Consumer Boycotts </EN-title> <
EN-desc> Find documents that describe or discuss the impact of consumer
boycotts. </EN-desc>
<EN-narr> Relevant documents will report discussions or points of view on
the efficacy of consumer boycotts. The moral issues involved in such
boycotts are also of relevance. Only consumer boycotts are relevant,
political boycotts must be ignored. </EN-narr> </top>
```

For the robust task, the topic sets used in CLEF 2001, CLEF 2002 and CLEF 2003 were used for evaluation. A total of 160 topics were collected and split into two sets: 60 topics used to train the system, and 100 topics used for the evaluation. Topics were available in the languages of the target collections: English, German, French, Spanish, Italian, Dutch.

2.2 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. 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 the *Text REtrieval Conference (TREC)* (trec_eval 8.0) and available from the TREC website.

As a consequence of limited evaluation resources, we set a maximum number of runs for each task with restrictions on the number of runs that could be accepted for a single language or language combination - we try to encourage diversity.

2.3 Relevance Assessment

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 groups participating in the ad hoc tasks are used to form a pool of documents for each topic and language by collecting the highly ranked documents from all submissions. This pool is then used for subsequent relevance judgments. The stability of pools constructed in this way and their reliability for post-campaign experiments is discussed in [1] with respect to the CLEF 2003 pools. After calculating the effectiveness measures, the results are analyzed and run statistics produced and distributed. New pools were formed in CLEF 2006 for the runs submitted for the main stream mono- and bilingual tasks and the relevance assessments were performed by native speakers. Instead, the robust tasks used the original pools and relevance assessments from CLEF 2001-2003.

The individual results for all official ad hoc experiments in CLEF 2006 are given in the Appendix at the end of the on-line Working Notes prepared for the Workshop [2] and available online at www.clef-campaign.org.

2.4 Result Calculation

Evaluation campaigns such as TREC and CLEF are based on the belief that the effectiveness of *Information Retrieval Systems (IRSs)* 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 participants 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 [3]. For the robust task, we used different measures, see below Section 5.

2.5 Participants and Experiments

A total of 25 groups from 15 different countries submitted results for one or more of the ad hoc tasks - a slight increase on the 23 participants of last year.

A total of 296 experiments were submitted with an increase of 16% on the 254 experiments of 2005. On the other hand, the average number of submitted runs per participant is nearly the same: from 11 runs/participant of 2005 to 11.7 runs/participant of this year.

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 (172 out of 296, 58.11%) used this combination of topic fields, 78 (26.35%) used all fields, 41 (13.85%) used the title field, and only 5 (1.69%)

Table 3. Breakdown of experiments into tracks and topic languages

(a) Number of experiments per track, participant.

Track	# Part.	# Runs
Monolingual-BG	4	11
Monolingual-FR	8	27
Monolingual-HU	6	17
Monolingual-PT	12	37
Bilingual-X2BG	1	2
Bilingual-X2EN	5	33
Bilingual-X2FR	4	12
Bilingual-X2HU	1	2
Bilingual-X2PT	6	22
Robust-Mono-DE	3	7
Robust-Mono-EN	6	13
Robust-Mono-ES	5	11
Robust-Mono-FR	7	18
Robust-Mono-IT	5	11
Robust-Mono-NL	3	7
Robust-Bili-X2DE	2	5
Robust-Bili-X2ES	3	8
Robust-Bili-X2NL	1	4
Robust-Multi	4	10
Robust-Training-Mono-DE	2	3
Robust-Training-Mono-EN	4	7
Robust-Training-Mono-ES	3	5
Robust-Training-Mono-FR	5	10
Robust-Training-Mono-IT	3	5
Robust-Training-Mono-NL	2	3
Robust-Training-Bili-X2DE	1	1
Robust-Training-Bili-X2ES	1	2
Robust-Training-Multi	2	3
Total		**296**

(b) List of experiments by topic language.

Topic Lang.	# Runs
English	65
French	60
Italian	38
Portuguese	37
Spanish	25
Hungarian	17
German	12
Bulgarian	11
Indonesian	10
Dutch	10
Amharic	4
Oromo	3
Hindi	2
Telugu	2
Total	**296**

used the description field. The majority of experiments were conducted using automatic query construction (287 out of 296, 96.96%) and only in a small fraction of the experiments (9 out 296, 3.04%) have queries which were manually constructed from topics. A breakdown into the separate tasks is shown in Table 3(a).

Fourteen different topic languages were used in the ad hoc experiments. As always, the most popular language for queries was English, with French second. The number of runs per topic language is shown in Table 3(b).

3 Main Stream Monolingual Experiments

Monolingual retrieval was offered for Bulgarian, French, Hungarian, and Portuguese. As can be seen from Table 3(a), the number of participants and runs for each language was quite similar, with the exception of Bulgarian, which had a slightly smaller participation. This year just 6 groups out of 16 (37.5%) submitted monolingual runs only (down from ten groups last year), and 5 of these groups were first time participants in CLEF. Most of the groups submitting monolingual runs were doing this as part of their bilingual or multilingual system testing activity. Details on the different approaches used can be found in the papers in this section of the Proceedings. There was a lot of detailed work with Portuguese language processing; not surprising as we had four new groups from Brazil in Ad Hoc this year. As usual, there was a lot of work on the development of stemmers and morphological analysers ([4], for instance, applies a very deep morphological analysis for Hungarian) and comparisons of the pros and cons of so-called "light" and "heavy" stemming approaches (e.g. [5]). In contrast to previous years, we note that a number of groups experimented with NLP techniques (see, for example, papers by [6], and [7]).

Table 4. Best entries for the monolingual track

Track	Participant Rank					
	1st	**2nd**	**3rd**	**4th**	**5th**	**Diff.**
Bulgarian	unine	rsi-jhu	hummingbird	daedalus		1st vs 4th
MAP	33.14%	31.98%	30.47%	27.87%		20.90%
Run	UniNEbg2	02aplmobgtd4	humBG06tde	bgFSbg2S		
French	unine	rsi-jhu	hummingbird	alicante	daedalus	1st vs 5th
MAP	44.68%	40.96%	40.77%	38.28%	37.94%	17.76%
Run	UniNEfr3	95aplmofrtd5s1	humFR06tde	8dfrexp	frFSfr2S	
Hungarian	unine	rsi-jhu	alicante	mokk	hummingbird	1st vs 5th
MAP	41.35%	39.11%	35.32%	34.95%	32.24%	28.26%
Run	UniNEhu2	02aplmohutd4	30dfrexp	plain2	humHU06tde	
Portuguese	unine	hummingbird	alicante	rsi-jhu	u.buffalo	1st vs 5th
MAP	45.52%	45.07%	43.08%	42.42%	40.53%	12.31%
Run	UniNEpt1	humPT06tde	30okapiexp	95aplmopttd5	UBptTDrf1	

3.1 Results

Table 4 shows the results for the top five groups for each target collection, ordered by mean average precision. The table reports: the short name of the participating group; the mean average precision achieved by the run; the run identifier; and the performance difference between the first and the last participant. Table 4 regards runs using title + description fields only (the mandatory run).

4 Main Stream 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 restricted to newcomers in a CLEF cross-language task. This year, in this subtask, we focussed in particular on non-European topic languages and in particular languages for which there are still few processing tools or resources in existence. We thus offered two Ethiopian languages: Amharic and Oromo; two Indian languages: Hindi and Telugu; and Indonesian. Although, as was to be expected, the results are not particularly good, we feel that experiments of this type with lesser-studied languages are very important (see papers by [8], [9])

4.1 Results

Table 5 shows the best results for this task for runs using the title+description topic fields. The performance difference between the best and the last (up to 5) placed group is given (in terms of average precision. Again both pooled and not pooled runs are included in the best entries for each track, with the exception of Bilingual X → EN.

For bilingual retrieval evaluation, a common method to evaluate performance is to compare results against monolingual baselines. For the best bilingual systems, we have the following results for CLEF 2006:

– X → BG: 52.49% of best monolingual Bulgarian IR system;

Table 5. Best entries for the bilingual task

Track	Participant Rank					
	1st	2nd	3rd	4th	5th	Diff.
Bulgarian	daedalus					
MAP	17.39%					
Run	bgFSbgWen2S					
French	unine	queenmary	rsi-jhu	daedalus		1st vs 4th
MAP	41.92%	33.96%	33.60%	33.20%		26.27%
Run	UniNEBifr1	QMUL06e2f10b	aplbienfrd	frFSfrSen2S		
Hungarian	daedalus					
MAP	21.97%					
Run	huFShuMen2S					
Portuguese	unine	rsi-jhu	queenmary	u.buffalo	daedalus	1st vs 5th
MAP	41.38%	35.49%	35.26%	29.08%	26.50%	55.85%
Run	UniNEBipt2	aplbiesptd	QMUL06e2p10b	UBen2ptTDrf2	ptFSptSen2S	
English	rsi-jhu	depok	ltrc	celi	dsv	1st vs 5th
MAP	32.57%	26.71%	25.04%	23.97%	22.78%	42.98%
Run	aplbiinen5	UI_td_mt	OMTD	CELItitleNOEXPANSION	DsvAmhEngFullNofuzz	

- X → FR: 93.82% of best monolingual French IR system;
- X → HU: 53.13% of best monolingual Hungarian IR system.
- X → PT: 90.91% of best monolingual Portuguese IR system;

We can compare these to those for CLEF 2005:

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

While these results are good for the well-established-in-CLEF languages, and can be read as state-of-the-art for this kind of retrieval system, at a first glance they appear disappointing for Bulgarian and Hungarian. However, we must point out that, unfortunately, this year only one group submitted cross-language runs for Bulgarian and Hungarian and thus it does not make much sense to draw any conclusions from these, apparently poor, results for these languages. It is interesting to note that when *Cross Language Information Retrieval (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.

5 Robust Experiments

The robust task was organized for the first time at CLEF 2006. The evaluation of robustness emphasizes stable performance over all topics instead of high average performance [10]. The perspective of each individual user of an information retrieval system is different from the perspective of an evaluation initiative. Users are disappointed by systems which deliver poor results for some topics whereas an evaluation initiative rewards systems which deliver good average results. A system delivering poor results for hard topics is likely to be considered of low quality by a user although it may still reach high average results. The robust task has been inspired by the robust track at TREC where it was organized at TREC 2003, 2004 and 2005. A robust evaluation stresses performance for weak topics. This can be achieved by employing the *Geometric Average Precision (GMAP)* as a main indicator for performance instead of the *Mean Average Precision (MAP)* of all topics. Geometric average has proven to be a stable measure for robustness at TREC [10]. The robust task at CLEF 2006 is concerned with multilingual robustness. It is essentially an ad-hoc task which offers mono-lingual and cross-lingual sub tasks.

As stated, the roubust task used test collections developed in CLEF 2001, CLEF 2002 and CLEF 2003. No additional relevance judgements were made this year for this task. However, the data collection was not completely constant over all three CLEF campaigns which led to an inconsistency between

relevance judgements and documents. The SDA 95 collection has no relevance judgements for most topics (#41 - #140). This inconsistency was tolerated in order to increase the size of the collection. One participant reported that exploiting the knowledge would have resulted in an increase of approximately 10% in MAP [11]. However, participants were not allowed to use this information. The results of the original submissions for the data sets were analyzed in order to identify the most difficult topics. This turned out to be a very hard task. The difficulty of a topic varies greatly among languages, target collections and tasks. This confirms the finding of the TREC 2005 robust task where the topic difficulty differed greatly even for two different English collections. Topics are not inherently difficult but only in combination with a specific collection [12]. Topic difficulty is usually defined by low MAP values for a topic. We also considered a low number of relevant documents and high variation between systems as indicators for difficulty. Because no consistent definition of topic difficulty could be found, the topic set for the robust task at CLEF 2006 was arbitrarily split into two sets. Participants were allowed to use the available relevance assessments for the set of 60 training topics. The remaining 100 topics formed the test set for which results are reported. The participants were encouraged to submit results for training topics as well. These runs will be used to further analyze topic difficulty.

The robust task received a total of 133 runs from eight groups. Most popular among the participants were the mono-lingual French and English tasks. For the multi-lingual task, four groups submitted ten runs. The bi-lingual tasks received fewer runs. A run using title and description was mandatory for each group. Participants were encouraged to run their systems with the same setup for all robust tasks in which they participated (except for language specific resources). This way, the robustness of a system across languages could be explored.

Effectiveness scores for the submissions were calculated with the GMAP which is calculated as the n-th root of a product of n values. GMAP was computed using the version 8.0 of `trec_eval`[2] program. In order to avoid undefined result figures, all precision scores lower than 0.00001 are set to 0.00001.

5.1 Robust Monolingual Results

Table 6 shows the best results for this task for runs using the title+description topic fields. The performance difference between the best and the last (up to 5) placed group is given (in terms of average precision).

Hummingbird submitted the best results for five out of six sub tasks. However, the differences between the best runs are small and not always statistically significant, see [21,2].

The MAP figures were above 45% for five out of six sub tasks. These numbers can be considered as state of the art.

It is striking that the rankings based on the MAP is identical to the ranking based on the GMAP measure in most cases.

[2] `http://trec.nist.gov/trec_eval/trec_eval.8.0.tar.gz`

Table 6. Best entries for the robust monolingual task

Track	Participant Rank					
	1st	**2nd**	**3rd**	**4th**	**5th**	**Diff.**
Dutch	hummingbird	daedalus	colesir			1st vs 3rd
MAP	51.06%	42.39%	41.60%			22.74%
GMAP	25.76%	17.57%	16.40%			57.13%
Run	humNL06Rtde	nlFSnlR2S	CoLesIRnlTst			
English	hummingbird	reina	dcu	daedalus	colesir	1st vs 5th
MAP	47.63%	43.66%	43.48%	39.69%	37.64%	26.54%
GMAP	11.69%	10.53%	10.11%	8.93%	8.41%	39.00%
Run	humEN06Rtde	reinaENtdtest	dcudesceng12075	enFSenR2S	CoLesIRenTst	
French	unine	hummingbird	reina	dcu	colesir	1st vs 5th
MAP	47.57%	45.43%	44.58%	41.08%	39.51%	20.40%
GMAP	15.02%	14.90%	14.32%	12.00%	11.91%	26.11%
Run	UniNEfrr1	humFR06Rtde	reinaFRtdtest	dcudescfr12075	CoLesIRfrTst	
German	hummingbird	colesir	daedalus			1st vs 3rd
MAP	48.30%	37.21%	34.06%			41.81%
GMAP	22.53%	14.80%	10.61%			112.35%
Run	humDE06Rtde	CoLesIRdeTst	deFSdeR2S			
Italian	hummingbird	reina	dcu	daedalus	colesir	1st vs 5th
MAP	41.94%	38.45%	37.73%	35.11%	32.23%	30.13%
GMAP	11.47%	10.55%	9.19%	10.50%	8.23%	39.37%
Run	humIT06Rtde	reinaITtdtest	dcudescit1005	itFSitR2S	CoLesIRitTst	
Spanish	hummingbird	reina	dcu	daedalus	colesir	1st vs 5th
MAP	45.66%	44.01%	42.14%	40.40%	40.17%	13.67%
GMAP	23.61%	22.65%	21.32%	19.64%	18.84%	25.32%
Run	humES06Rtde	reinaEStdtest	dcudescsp12075	esFSesR2S	CoLesIResTst	

5.2 Robust Bilingual Results

Table 7 shows the best results for this task for runs using the title+description topic fields. The performance difference between the best and the last (up to 5) placed group is given (in terms of average precision).

As stated in 4.1, for bilingual retrieval evaluation, a common method is to compare results against monolingual baselines. We have the following results for CLEF 2006:

- X → DE: 60.37% of best monolingual German IR system;
- X → ES: 80.88% of best monolingual Spanish IR system;
- X → NL: 69.27% of best monolingual Dutch IR system.

5.3 Robust Multilingual Results

Table 8 shows the best results for this task for runs using the title+description topic fields. The performance difference between the best and the last (up to 5) placed group is given (in terms of average precision). The figures are lower than for multilingual experiments at previous CLEF campaigns. This shows that the multilingual retrieval problem is far from being solved and that results depend much on the topic set.

Table 7. Best entries for the robust bilingual task.

Track	Participant Rank					
	1st	**2nd**	**3rd**	**4th**	**5th**	**Diff.**
Dutch	daedalus					
MAP	35.37%					
GMAP	9.75%					
Run	nlFSnlRLfr2S					
German	daedalus	colesir				1st vs 2nd
MAP	29.16%	25.24%				15.53%
GMAP	5.18%	4.31%				20.19%
Run	deFSdeRSen2S	CoLesIRendeTst				
Spanish	reina	dcu	daedalus			1st vs 3rd
MAP	36.93%	33.22%	26.89%			37.34%
GMAP	13.42%	10.44%	6.19%			116.80%
Run	reinaIT2EStdtest	dcuitqydescsp12075	esFSesRLit2S			

Table 8. Best entries for the robust multilingual task

Track	Participant Rank					
	1st	**2nd**	**3rd**	**4th**	**5th**	**Diff.**
Multilingual	jaen	daedalus	colesir	reina		1st vs 4th
MAP	27.85%	22.67%	22.63%	19.96%		39.53%
GMAP	15.69%	11.04%	11.24%	13.25%		18.42%
Run	ujamlrsv2	mlRSFSen2S	CoLesIRmultTst	reinaES2mtdtest		

5.4 Comments on Robust Cross Language Experiments

The robust track is especially concerned with the performance for hard topics which achieve low MAP figures. One important reason for weak topics is the lack of good keywords in the query and difficulties to expand the query properly within the collection. A strategy often applied is the query expansion with external collections like the web. This or other strategies are sometimes applied depending on the topic. Only when a topic is classified as a difficult topic, additional techniques are applied. Several participants relied on the high correlation between the measure and optimized their systems as in previous campaigns. Nevertheless, some groups worked specifically for robustness. The SINAI system took an approach which has proved successful at the TREC robust task, expansion with terms gathered from a web search engine [13]. The REINA system from the University of Salamanca used a heuristic to determine hard topics during training. Subsequently, different expansion techniques were applied [14]. The MIRACLE system tried to find a fusion scheme which had a positive effect on the robust measure [16]. The results are mixed. Savoy & Abdou reported that expansion with an external search engine did not improve the results [11]. It seems that optimal heuristics for the selection of good expansion terms still

need to be developed. Hummingbird thoroughly discussed alternative evaluation measures for capturing the robustness of runs. [15].

6 Conclusions

We have reported the results of the ad hoc cross-language textual document retrieval track at CLEF 2006. This track is considered to be central to CLEF as for many groups it is the first track in which they participate and provides them with an opportunity to test their systems and compare performance between monolingual and cross-language runs, before perhaps moving on to more complex system development and subsequent evaluation. However, the track is certainly not just aimed at beginners. It also gives groups the possibility to measure advances in system performance over time. In addition, each year, we also include a task aimed at examining particular aspects of cross-language text retrieval. This year, the focus was examining the impact of "hard" topics on performance in the "robust" task.

Thus, although the ad hoc track in CLEF 2006 offered the same target languages for the main mono- and bilingual tasks as in 2005, it also had two new focuses. Groups were encouraged to use non-European languages as topic languages in the bilingual task. We were particularly interested in languages for which few processing tools were readily available, such as Amharic, Oromo and Telugu. In addition, we set up the "robust task" with the objective of providing the more expert groups with the chance to do in-depth failure analysis.

For reasons of space, in this paper we have only been able to summarise the main results; more details, including sets of statistical analyses can be found in [21,2].

Finally, it should be remembered that, although over the years we vary the topic and target languages offered in the track, all participating groups also have the possibility of accessing and using the test collections that have been created in previous years for all of the twelve languages included in the CLEF multilingual test collection. The test collections for CLEF 2000 - CLEF 2003 are about to be made publicly available on the *Evaluations and Language resources Distribution Agency (ELDA)* catalog[3].

References

1. Braschler, M.: CLEF 2003 - Overview of results. In: Peters, C., Gonzalo, J., Braschler, M., Kluck, M. (eds.) CLEF 2003. LNCS, vol. 3237, pp. 44–63. Springer, Heidelberg (2004)
2. Di Nunzio, G.M., Ferro, N.: Appendix A. Results of the Core Tracks. In: Nardi, A., Peters, C., Vicedo, J.L. (eds.) Working Notes for the CLEF 2006 Workshop (2006), Published Online at `www.clef-campaign.org`
3. Braschler, M., Peters, C.: CLEF 2003 Methodology and Metrics. In: Peters, C., Gonzalo, J., Braschler, M., Kluck, M. (eds.) CLEF 2003. LNCS, vol. 3237, pp. 7–20. Springer, Heidelberg (2004)

[3] `http://www.elda.org/`

4. Halácsy, P., Trón, V.: Benefits of Deep NLP-based Lemmatization for Information Retrieval. [In this volume]
5. Moreira Orengo, V., Buriol, L.S., Ramos Coelho, A.: A Study on the use of Stemming for Monolingual Ad-Hoc Portuguese. Information Retrieval (2006)
6. Azevedo Arcoverde, J.M., das Gracas Volpe Nunes, M.: NLP-Driven Constructive Learning for Filtering an IR Document Stream. [In this volume]
7. Gonzalez, M., de Lima, V.L.S.: The PUCRS-PLN Group participation at CLEF 2006. [In this volume]
8. Pingali, P., Tune, K.K., Varma, V.: Hindi, Telugu, Oromo, English CLIR Evaluation. [In this volume]
9. Hayurani, H., Sari, S., Adriani, M.: Query and Document Translation for English-Indonesian Cross Language IR. [In this volume]
10. Voorhees, E.M.: The TREC Robust Retrieval Track. SIGIR Forum 39, 11–20 (2005)
11. Savoy, J., Abdou, S.: Experiments with Monolingual, Bilingual, and Robust Retrieval. [In this volume]
12. Voorhees, E.M.: Overview of the TREC 2005 Robust Retrieval Track. In: Voorhees, E.M., Buckland, L.P. (eds.): The Fourteenth Text REtrieval Conference Proceedings (TREC 2005) [last visited 2006, August 4] (2005), http://trec.nist.gov/pubs/trec14/t14_proceedings.html
13. Martinez-Santiago, F., Montejo-Ráez, A., Garcia-Cumbreras, M., Ureña-Lopez, A.: SINAI at CLEF 2006 Ad-hoc Robust Multilingual Track: Query Expansion using the Google Search Engine. [In this volume] (2006)
14. Zazo, A., Berrocal, J., Figuerola, C.: Local Query Expansion Using Term Windows for Robust Retrieval. [In this volume]
15. Tomlinson, S.: Comparing the Robustness of Expansion Techniques and Retrieval Measures. [In this volume]
16. Goni-Menoyo, J., Gonzalez-Cristobal, J., Vilena-Román, J.: Report of the MIRACLE teach for the Ad-hoc track in CLEF 2006. In: Nardi, A., Peters, C., Vicedo, J.L. (eds.) Working Notes for the CLEF 2006 Workshop, Published Online (2006)
17. 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), pp. 329–338. ACM Press, New York (1993)
18. Conover, W.J.: Practical Nonparametric Statistics, 1st edn. John Wiley and Sons, New York (1971)
19. Judge, G.G., Hill, R.C., Griffiths, W.E., Lütkepohl, H., Lee, T.C.: Introduction to the Theory and Practice of Econometrics, 2nd edn. John Wiley and Sons, New York (1988)
20. Tague-Sutcliffe, J.: The Pragmatics of Information Retrieval Experimentation, Revisited. In: Sparck Jones, K., Willett, P. (eds.) Readings in Information Retrieval, pp. 205–216. Morgan Kaufmann Publisher, Inc, San Francisco, California (1997)
21. Di Nunzio, G.M., Ferro, N., Mandl, T., Peters, C.: CLEF 2006: Ad Hoc Track Overview. In: Nardi, A., Peters, C., Vicedo, J.L., eds.: Working Notes for the CLEF 2006 Workshop (2006) (last visited, March 23, 2007), http://www.clef-campaign.org/2006/working_notes/workingnotes2006/dinunzio0CLEF2006.pdf