CLEF 2009 Ad Hoc Track Overview: TEL & Persian Tasks

Nicola Ferro[1] and Carol Peters[2]

[1] Department of Information Engineering, University of Padua, Italy
ferro@dei.unipd.it
[2] ISTI-CNR, Area di Ricerca, Pisa, Italy
carol.peters@isti.cnr.it

Abstract. The 2009 Ad Hoc track was to a large extent a repetition of last year's track, with the same three tasks: Tel@CLEF, Persian@CLEF, and Robust-WSD. In this first of the two track overviews, we describe the objectives and results of the TEL and Persian tasks and provide some statistical analyses.

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, Word Sense Disambiguation

1 Introduction

From 2000 - 2007, the ad hoc track at CLEF used exclusively collections of European newspaper and news agency documents[1]. In 2008 it was decided to change the focus and to introduce document collections in a different genre (bibliographic records from The European Library - TEL[2]), a non-European language (Persian), and an IR task that would appeal to the NLP community (robust retrieval on word-sense disambiguated data). The 2009 Ad Hoc track has been to a large extent a repetition of last year's track, with the same three tasks: Tel@CLEF, Persian@CLEF, and Robust-WSD. An important objective has been to ensure that for each task a good reusable test collections is created.

[1] Over the years, this track has built up test collections for monolingual and cross language system evaluation in 14 European languages.

[2] See http://www.theeuropeanlibrary.org/

In this first of the two track overviews we describe the activities of the TEL and Persian tasks[3].

TEL@CLEF: This task offered monolingual and cross-language search on library catalog. It was organized in collaboration with The European Library and used three collections derived from the catalogs of the British Library, the Bibliothéque Nationale de France and the Austrian National Library. The underlying aim was to identify the most effective retrieval technologies for searching this type of very sparse multilingual data. In fact, the collections contained records in many languages in addition to English, French or German. The task presumed a user with a working knowledge of these three languages who wants to find documents that can be useful for them in one of the three target catalogs.

Persian@CLEF: This activity was coordinated again this year in collaboration with the Database Research Group (DBRG) of Tehran University. We chose Persian as the first non-European language target collection for several reasons: its challenging script (a modified version of the Arabic alphabet with elision of short vowels) written from right to left; its complex morphology (extensive use of suffixes and compounding); its political and cultural importance. The task used the Hamshahri corpus of 1996-2002 newspapers as the target collection and was organised as a traditional ad hoc document retrieval task. Monolingual and cross-language (English to Persian) tasks were offered.

In the rest of this paper we present the task setup, the evaluation methodology and the participation in the two tasks (Section 2). We then describe the main features of each task and show the results (Sections 3 and 4). The final section provides a brief summing up. For information on the various approaches and resources used by the groups participating in the two tasks and the issues they focused on, we refer the reader to the papers in the relevant Ad Hoc sections of these Working Notes.

2 Track Setup

As is customary in the CLEF ad hoc track, again this year we adopted a corpus-based, automatic scoring method for the assessment of the performance of the participating systems, based on ideas first introduced in the Cranfield experiments in the late 1960s [5]. The tasks offered are studied in order to effectively measure textual document retrieval under specific conditions. The test collections are made up of documents, topics and relevance assessments. The topics consist of a set of statements simulating information needs from which the systems derive the queries to search the document collections. 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 pooling methodology is used in order to limit the number of manual relevance assessments that have to be made. As always, the distinguishing

[3] As the task design was the same as last year, much of the task set-up section is a repetition of a similar section in our CLEF 2008 working notes paper.

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 The Documents

As mentioned in the Introduction, the two tasks used different sets of documents.

The TEL task used three collections:

- British Library (BL); 1,000,100 documents, 1.2 GB;
- Bibliothéque Nationale de France (BNF); 1,000,100 documents, 1.3 GB;
- Austrian National Library (ONB); 869,353 documents, 1.3 GB.

We refer to the three collections (BL, BNF, ONB) as English, French and German because in each case this is the main and expected language of the collection. However, each of these collections is to some extent multilingual and contains documents (catalog records) in many additional languages.

The TEL data is very different from the newspaper articles and news agency dispatches previously used in the CLEF ad hoc track. The data tends to be very sparse. Many records contain only title, author and subject heading information; other records provide more detail. The title and (if existing) an abstract or description may be in a different language to that understood as the language of the collection. The subject heading information is normally in the main language of the collection. About 66% of the documents in the English and German collection have textual subject headings, in the French collection only 37%. Dewey Classification (DDC) is not available in the French collection; negligible (¡0.3%) in the German collection; but occurs in about half of the English documents (456,408 docs to be exact).

Whereas in the traditional ad hoc task, the user searches directly for a document containing information of interest, here the user tries to identify which publications are of potential interest according to the information provided by the catalog card. When we designed the task, the question the user was presumed to be asking was “Is the publication described by the bibliographic record relevant to my information need?”

The Persian task used the Hamshahri corpus of 1996-2002 newspapers as the target collection. This corpus was made available to CLEF by the Data Base Research Group (DBRG) of the University of Tehran. Hamshahri is one of the most popular daily newspapers in Iran. The Hamshahri corpus consists of 345 MB of news texts for the years 1996 to 2002 (corpus size with tags is 564 MB). This corpus contains more than 160,000 news articles about a variety of subjects and includes nearly 417000 different words. Hamshahri articles vary between 1KB and 140KB in size[4].

2.2 Topics

Topics in the CLEF ad hoc track are structured statements representing information needs. Each topic typically consists of three parts: a brief "title" statement; a one-sentence "description"; a more complex "narrative" specifying the relevance assessment criteria. Topics are prepared in xml format and uniquely identified by means of a *Digital Object Identifier (DOI)*[5].

For the TEL task, a common set of 50 topics was prepared in each of the 3 main collection languages (English, French and German) plus this year also in Chinese, Italian and Greek in response to specific requests. Only the Title and Description fields were released to the participants. The narrative was employed to provide information for the assessors on how the topics should be judged. The topic sets were prepared on the basis of the contents of the collections.

In ad hoc, when a task uses data collections in more than one language, we consider it important to be able to use versions of the same core topic set to query all collections. This makes it easier to compare results over different collections and also facilitates the preparation of extra topic sets in additional languages. However, it is never easy to find topics that are effective for several different collections and the topic preparation stage requires considerable discussion between the coordinators for each collection in order to identify suitable common candidates. The sparseness of the data makes this particularly difficult for the TEL task and leads to the formulation of topics that were quite broad in scope so that at least some relevant documents could be found in each collection. A result of this strategy is that there tends to be a considerable lack of evenness of distribution in relevant documents. For each topic, the results expected from the separate collections can vary considerably. An example of a TEL topic is given in Figure 1.

For the Persian task, 50 topics were created in Persian by the Data Base Research group of the University of Tehran, and then translated into English. The rule in CLEF when creating topics in additional languages is not to produce literal translations but to attempt to render them as naturally as possible. This was a particularly difficult task when going from Persian to English as cultural differences had to be catered for. An example of a CLEF 2009 Persian topic is given in Figure 2.

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

[4] For more information, see `http://ece.ut.ac.ir/dbrg/hamshahri/`

[5] `http://www.doi.org/`

```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<topic>
    <identifier>10.2452/711-AH</identifier>

    <title lang="zh">深海生物</title>
    <title lang="en">Deep Sea Creatures</title>
    <title lang="fr">Créatures des fonds océaniques</title>
    <title lang="de">Kreaturen der Tiefsee</title>
    <title lang="el">Πλάσματα στα βάθη των ωκεανών</title>
    <title lang="it">Creature delle profondità oceaniche</title>

    <description lang="zh">
        查找有关世界上任何深海生物的出版物。
    </description>
    <description lang="en">
        Find publications about any kind of life in the depths
        of any of the world's oceans.
    </description>
    <description lang="fr">
        Trouver des ouvrages sur toute forme de vie dans les
        profondeurs des mers et des océans.
    </description>
    <description lang="de">
        Finden Sie Veröffentlichungen über Leben und
        Lebensformen in den Tiefen der Ozeane der Welt.
    </description>
    <description lang="el">
        Αναζήτηση δημοσιεύσεων για κάθε είδος ζωής στα
        βάθη των ωκεανών
    </description>
    <description lang="it">
        Trova pubblicazioni su qualsiasi forma di vita nelle
        profondità degli oceani del mondo.
    </description>
</topic>
```

Fig. 1. Example of TEL topic `http://direct.dei.unipd.it/10.2452/711-AH`.

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 selected runs according to a set of predefined criteria. One important limitation when forming the pools is the number of documents to be assessed. Traditionally, the top 100 ranked documents from each of the runs selected are included in the pool; in such a case we say that the pool is of depth 100. This pool is then used for subsequent relevance judgments. After calculating the effectiveness measures, the results are analyzed and run statistics produced and distributed. 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.

The main criteria used when constructing the pools in CLEF are:

```xml
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<topic>
    <identifier>10.2452/641-AH</identifier>

    <title lang="en">Pollution in the Persian Gulf</title>
    <title lang="fa">وضعيت آلودگي درياي خليج فارس</title>

    <description lang="en">
        Find information about pollution in the Persian Gulf and the causes.
    </description>
    <description lang="fa">
        بررسي وضعيت درياي خليج فارس از نظر آلودگي و عوامل آن
    </description>

    <narrative lang="en">
        Find information about conditions of the Persian Gulf with respect to
        pollution; also of interest is information on the causes of pollution
        and comparisons of the level of pollution in this sea against that of
        other seas.
    </narrative>
    <narrative lang="fa">
        يافتن اطلاعاتي در مورد وضعيت آلودگي درياي خليج فارس و بررسي عوامل ايجاد آل
        ودگي در اين دريا و اطلاعاتي نظير مقايسه آن با ساير درياها
    </narrative>
</topic>
```

Fig. 2. Example of Persian topic `http://direct.dei.unipd.it/10.2452/641-AH`.

- favour diversity among approaches adopted by participants, according to the descriptions of the experiments provided by the participants;
- choose at least one experiment for each participant in each task, chosen among the experiments with highest priority as indicated by the participant;
- add mandatory title+description experiments, even though they do not have high priority;
- add manual experiments, when provided;
- for bilingual tasks, ensure that each source topic language is represented.

From our experience in CLEF, using the tools provided by the DIRECT system [1], we find that for newspaper documents, assessors can normally judge from 60 to 100 documents per hour, providing binary judgments: relevant / not relevant. Our estimate for the TEL catalog records is higher as these records are much shorter than the average newspaper article (100 to 120 documents per hour). In both cases, it can be seen what a time-consuming and resource expensive task human relevance assessment is. This limitation impacts strongly on the application of the criteria above - and implies that we are obliged to be flexible in the number of documents judged per selected run for individual pools.

This year, in order to create pools of more-or-less equivalent size, the depth of the TEL English, French, and German pools was 60[6]. For each collection, we included in the pool two monolingual and one bilingual experiments from each participant plus any documents assessed as relevant during topic creation.

[6] Tests made on NTCIR pools in previous years have suggested that a depth of 60 in normally adequate to create stable pools, presuming that a sufficient number of runs from different systems have been included.

As we only had a relatively small number of runs submitted for Persian, we were able to include documents from all experiments, and the pool was created with a depth of 80.

These pool depths were the same as those used last year. Given the resources available, it was not possible to manually assess more documents. For the CLEF 2008 ad hoc test collections, Stephen Tomlinson reported some sampling experiments aimed at estimating the judging coverage [10]. He found that this tended to be lower than the estimates he produced for the CLEF 2007 ad hoc collections. With respect to the TEL collections, he estimated that at best 50% to 70% of the relevant documents were included in the pools - and that most of the unjudged relevant documents were for the 10 or more queries that had the most known answers. Tomlinson has repeated these experiments for the 2009 TEL and Persian data [9]. Although for two of the four languages concerned (German and Persian), his findings were similar to last year's estimates, for the other two languages (English and French) this year's estimates are substantially lower. These findings need further investigation. They suggest that if we are to continue to use the pooling technique, we would perhaps be wise to do some more exhaustive manual searches in order to boost the pools with respect to relevant documents. We also need to consider more carefully other techniques for relevance assessment in the future such as, for example, the method suggested by Sanderson and Joho [8] or Mechanical Turk [2].

Table 1 reports summary information on the 2009 ad hoc pools used to calculate the results for the main monolingual and bilingual experiments. In particular, for each pool, we show the number of topics, the number of runs submitted, the number of runs included in the pool, the number of documents in the pool (relevant and non-relevant), and the number of assessors.

The box plot of Figure 3 compares the distributions of the relevant documents across the topics of each pool for the different ad hoc pools; the boxes are ordered by decreasing mean number of relevant documents per topic.

As can be noted, TEL French and German distributions appear similar and are slightly asymmetric towards topics with a greater number of relevant documents while the TEL English distribution is slightly asymmetric towards topics with a lower number of relevant documents. All the distributions show some upper outliers, i.e. topics with a greater number of relevant document with respect to the behaviour of the other topics in the distribution. These outliers are probably due to the fact that CLEF topics have to be able to retrieve relevant documents in all the collections; therefore, they may be considerably broader in one collection compared with others depending on the contents of the separate datasets.

For the TEL documents, we judged for relevance only those documents that are written totally or partially in English, French and German, e.g. a catalog record written entirely in Hungarian was counted as not relevant as it was of no use to our hypothetical user; however, a catalog record with perhaps the title and a brief description in Hungarian, but with subject descriptors in French, German or English was judged for relevance as it could be potentially useful. Our assessors

Table 1. Summary information about CLEF 2009 pools.

TEL English Pool (**DOI** 10.2454/AH-TEL-ENGLISH-CLEF2009)	
Pool size	26,190 pooled documents – 23,663 not relevant documents – 2,527 relevant documents 50 topics
Pooled Experiments	31 out of 89 submitted experiments – monolingual: 22 out of 43 submitted experiments – bilingual: 9 out of 46 submitted experiments
Assessors	4 assessors
TEL French Pool (**DOI** 10.2454/AH-TEL-FRENCH-CLEF2009)	
Pool size	21,971 pooled documents – 20,118 not relevant documents – 1,853 relevant documents 50 topics
Pooled Experiments	21 out of 61 submitted experiments – monolingual: 16 out of 35 submitted experiments – bilingual: 5 out of 26 submitted experiments
Assessors	1 assessor
TEL German Pool (**DOI** 10.2454/AH-TEL-GERMAN-CLEF2009)	
Pool size	25,541 pooled documents – 23,882 not relevant documents – 1,559 relevant documents 50 topics
Pooled Experiments	21 out of 61 submitted experiments – monolingual: 16 out of 35 submitted experiments – bilingual: 5 out of 26 submitted experiments
Assessors	2 assessors
Persian Pool (**DOI** 10.2454/AH-PERSIAN-CLEF2009)	
Pool size	23,536 pooled documents – 19,072 not relevant documents – 4,464 relevant documents 50 topics
Pooled Experiments	20 out of 20 submitted experiments – monolingual: 17 out of 17 submitted experiments – bilingual: 3 out of 3 submitted experiments
Assessors	23 assessors

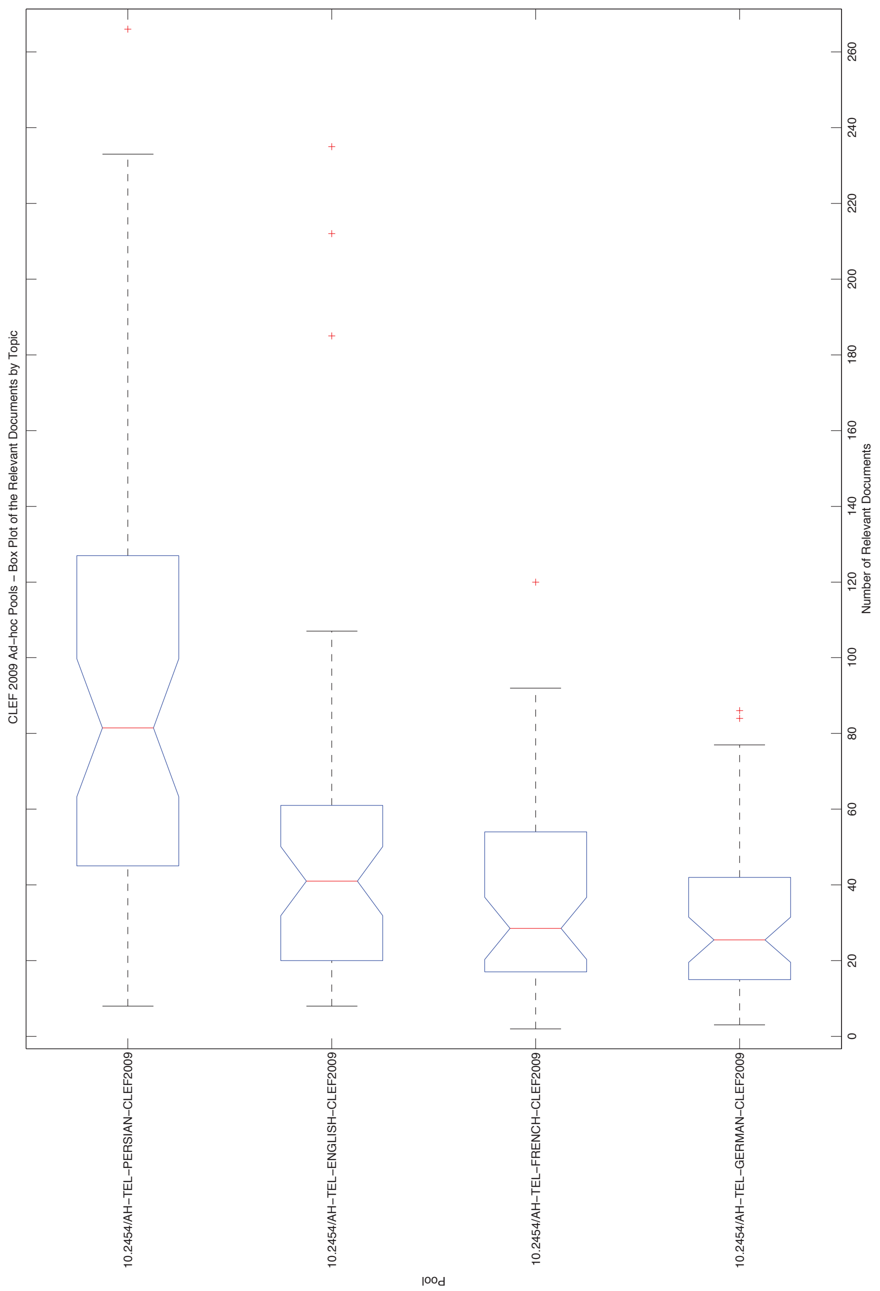

Fig. 3. Distribution of the relevant documents across the ad-hoc pools.

had no additional knowledge of the documents referred to by the catalog records (or surrogates) contained in the collection. They judged for relevance on the information contained in the records made available to the systems. This was a non trivial task due to the lack of information present in the documents. During the relevance assessment activity there was much consultation between the assessors for the three TEL collections in order to ensure that the same assessment criteria were adopted by everyone.

As shown in the box plot of Figure 3, the Persian distribution presents a greater number of relevant documents per topic with respect to the other distributions and is slightly asymmetric towards topics with a number of relevant documents. In addition, as can be seen from Table 1, it has been possible to sample all the experiments submitted for the Persian tasks. This means that there were fewer unique documents per run and this fact, together with the greater number of relevant documents per topic suggests either that all the systems were using similar approaches and retrieval algorithms or that the systems found the Persian topics quite easy.

The relevance assessment for the Persian results was done by the DBRG group in Tehran. Again, assessment was performed on a binary basis and the standard CLEF assessment rules were applied.

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 [4].

The individual results for all official Ad-hoc TEL and Persian experiments in CLEF 2009 are given in the Appendices of the CLEF 2009 Working Notes [6,7]. You can also access them online at:

- Ad-hoc TEL:
  - monolingual English: `http://direct.dei.unipd.it/DOIResolver.do?type=task&id=AH-TEL-MONO-EN-CLEF2009`
  - bilingual English: `http://direct.dei.unipd.it/DOIResolver.do?type=task&id=AH-TEL-BILI-X2EN-CLEF2009`
  - monolingual French: `http://direct.dei.unipd.it/DOIResolver.do?type=task&id=AH-TEL-MONO-FR-CLEF2009`
  - bilingual French: `http://direct.dei.unipd.it/DOIResolver.do?type=task&id=AH-TEL-BILI-X2FR-CLEF2009`
  - monolingual German: `http://direct.dei.unipd.it/DOIResolver.do?type=task&id=AH-TEL-MONO-DE-CLEF2009`
  - bilingual German: `http://direct.dei.unipd.it/DOIResolver.do?type=task&id=AH-TEL-BILI-X2DE-CLEF2009`

Table 2. CLEF 2009 Ad hoc participants.

Ad hoc TEL participants		
Participant	**Institution**	**Country**
aeb	Athens Univ. Economics & Business	Greece
celi	CELI Research srl	Italy
chemnitz	Chemnitz University of Technology	Germany
cheshire	U.C.Berkeley	United States
cuza	Alexandru Ioan Cuza University	Romania
hit	HIT2Lab, Heilongjiang Inst. Tech.	China
inesc	Tech. Univ. Lisbon	Portugal
karlsruhe	Univ. Karlsruhe	Germany
opentext	OpenText Corp.	Canada
qazviniau	Islamic Azaz Univ. Qazvin	Iran
trinity	Trinity Coll. Dublin	Ireland
trinity-dcu	Trinity Coll. & DCU	Ireland
weimar	Bauhaus Univ. Weimar	Germany

Ad hoc Persian participants		
Participant	**Institution**	**Country**
jhu-apl	Johns Hopkins Univ.	USA
opentext	OpenText Corp.	Canada
qazviniau	Islamic Azaz Univ. Qazvin	Iran
unine	U.Neuchatel-Informatics	Switzerland

- Ad-hoc Persian:
  - monolingual Farsi: `http://direct.dei.unipd.it/DOIResolver.do?type=task&id=AH-PERSIAN-MONO-FA-CLEF2009`
  - bilingual German: `http://direct.dei.unipd.it/DOIResolver.do?type=task&id=AH-PERSIAN-BILI-X2FA-CLEF2009`

2.5 Participants and Experiments

As shown in Table 2, a total of 13 groups from 10 countries submitted official results for the TEL task, while just four groups participated in the Persian task.

A total of 231 runs were submitted with an average number of submitted runs per participant of 13.5 runs/participant.

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 (216 out of 231, 93.50%) used this combination of topic fields, 2 (0.80%) used all fields[7], 13 (5.6%) used the title field. All the experiments were conducted using automatic query construction. A breakdown into the separate tasks and topic languages is shown in Table 3.

Seven different topic languages were used in the ad hoc experiments. As always, the most popular language for queries was English, with German second. However, it must be noted that English topics were provided for both the TEL

[7] The narrative field was only offered for the Persian task.

Table 3. Number of experiments by task and topic language and number of participants per task.

Task	Chinese	English	Farsi	French	German	Greek	Italian	Total	Participants
TEL Mono English	–	46	–	–	–	–	–	**46**	12
TEL Mono French	–	–	–	35	–	–	–	**35**	9
TEL Mono German	–	–	–	–	35	–	–	**35**	9
TEL Bili English	3	0	0	15	19	5	1	**43**	10
TEL Bili French	0	12	0	0	12	0	2	**26**	6
TEL Bili German	1	12	0	12	0	0	1	**26**	6
Mono Persian	–	–	17	–	–	–	–	**17**	4
Bili Persian	–	3	–	–	–	–	–	**3**	1
Total	**4**	**73**	**17**	**62**	**66**	**5**	**4**	**231**	–

and the Persian tasks. It is thus hardly surprising that English is the most used language in which to formulate queries. On the other hand, if we look only at the bilingual tasks, the most used source languages were German and French.

3 TEL@CLEF

The objective of this activity was to search and retrieve relevant items from collections of library catalog cards. The underlying aim was to identify the most effective retrieval technologies for searching this type of very sparse data.

3.1 Tasks

Two subtasks were offered: Monolingual and Bilingual. In both tasks, the aim was to retrieve documents relevant to the query. By monolingual we mean that the query is in the same language as the expected language of the collection. By bilingual we mean that the query is in a different language to the expected language of the collection. For example, in an EN → FR run, relevant documents (bibliographic records) could be any document in the BNF collection (referred to as the French collection) in whatever language they are written. The same is true for a monolingual FR → FR run - relevant documents from the BNF collection could actually also be in English or German, not just French.

Ten of the thirteen participating groups attempted a cross-language task; the most popular being with the British Library as the target collection. Six groups submitted experiments for all six possible official cross-language combinations. In addition, we had runs submitted to the English target with queries in Greek, Chinese and Italian.

3.2 Results.

Monolingual Results

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;

Table 4. Best entries for the monolingual TEL tasks.

Track	Rank	Participant	Experiment DOI	MAP
English	**1st**	inesc	10.2415/AH-TEL-MONO-EN-CLEF2009.INESC.RUN11	40.84%
	2nd	chemnitz	10.2415/AH-TEL-MONO-EN-CLEF2009.CHEMNITZ.CUT_11_MONO_MERGED_EN_9_10	40.71%
	3rd	trinity	10.2415/AH-TEL-MONO-EN-CLEF2009.TRINITY.TCDENRUN2	40.35%
	4th	hit	10.2415/AH-TEL-MONO-EN-CLEF2009.HIT.MTDD10T40	39.36%
	5th	trinity-dcu	10.2415/AH-TEL-MONO-EN-CLEF2009.TRINITY-DCU.TCDDCUEN3	36.96%
	Difference			10.50%
French	**1st**	karlsruhe	10.2415/AH-TEL-MONO-FR-CLEF2009.KARLSRUHE.INDEXBL	27.20%
	2nd	chemnitz	10.2415/AH-TEL-MONO-FR-CLEF2009.CHEMNITZ.CUT_19_MONO_MERGED_FR_17_18	25.83%
	3rd	inesc	10.2415/AH-TEL-MONO-FR-CLEF2009.INESC.RUN12	25.11%
	4th	opentext	10.2415/AH-TEL-MONO-FR-CLEF2009.OPENTEXT.OTFR09TDE	24.12%
	5th	celi	10.2415/AH-TEL-MONO-FR-CLEF2009.CELI.CACAO_FRBNF_ML	23.61%
	Difference			15.20%
German	**1st**	opentext	10.2415/AH-TEL-MONO-DE-CLEF2009.OPENTEXT.OTDE09TDE	28.68%
	2nd	chemnitz	10.2415/AH-TEL-MONO-DE-CLEF2009.CHEMNITZ.CUT_3_MONO_MERGED_DE_1_2	27.89%
	3rd	inesc	10.2415/AH-TEL-MONO-DE-CLEF2009.INESC.RUN12	27.85%
	4th	trinity-dcu	10.2415/AH-TEL-MONO-DE-CLEF2009.TRINITY-DCU.TCDDCUDE3	26.86%
	5th	trinity	10.2415/AH-TEL-MONO-DE-CLEF2009.TRINITY.TCDDERUN1	25.77%
	Difference			11.30%

the mean average precision achieved by the experiment; the DOI of the experiment; and the performance difference between the first and the last participant. Figures 4, 6, and 8 compare the performances of the top participants of the TEL Monolingual tasks.

Bilingual Results

Table 5 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 mean average precision achieved by the experiment; the DOI of the experiment; and the performance difference between the first and the last participant. Figures 5, 7, and 9 compare the performances of the top participants of the TEL Bilingual tasks.

For bilingual retrieval evaluation, a common method is to compare results against monolingual baselines. We have the following results for CLEF 2009:

- X → EN: 99.07% of best monolingual English IR system;
- X → FR: 94.00% of best monolingual French IR system;
- X → DE: 90.06% of best monolingual German IR system.

These figures are very encouraging, especially when compared with the results for last year for the same TEL tasks:

- X → EN: 90.99% of best monolingual English IR system;
- X → FR: 56.63% of best monolingual French IR system;

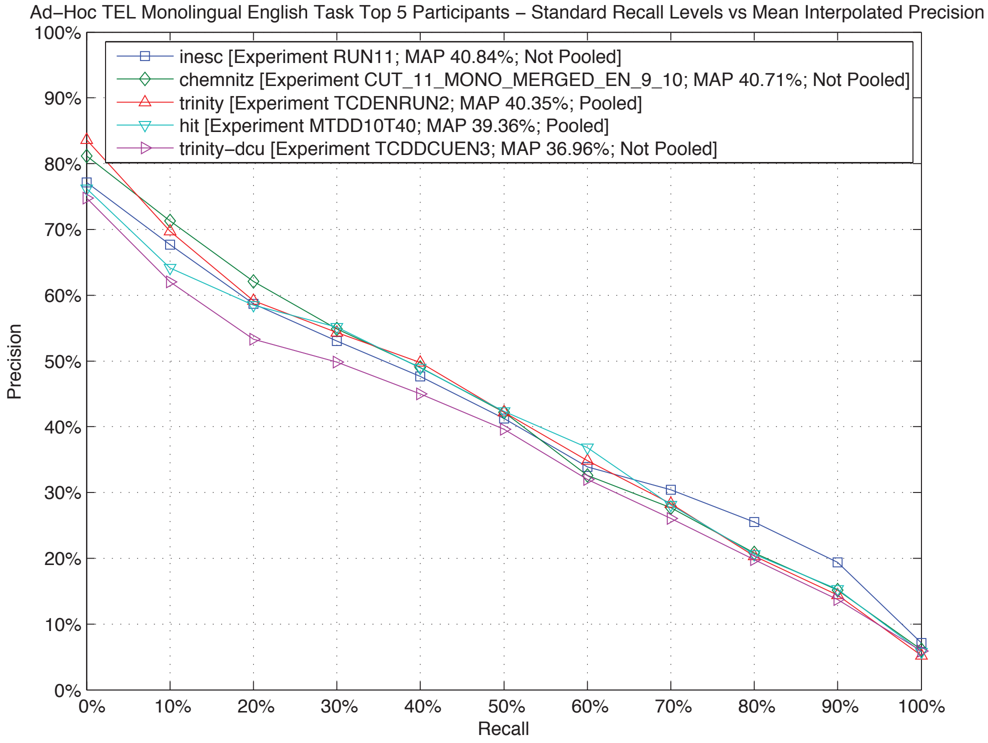

Fig. 4. Monolingual English

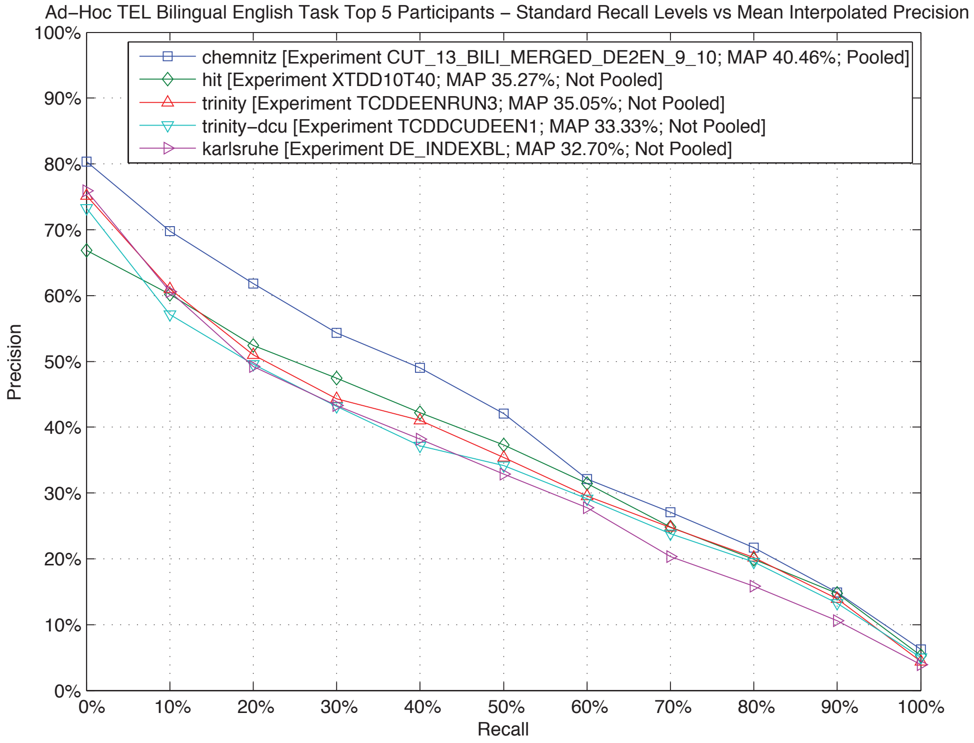

Fig. 5. Bilingual English

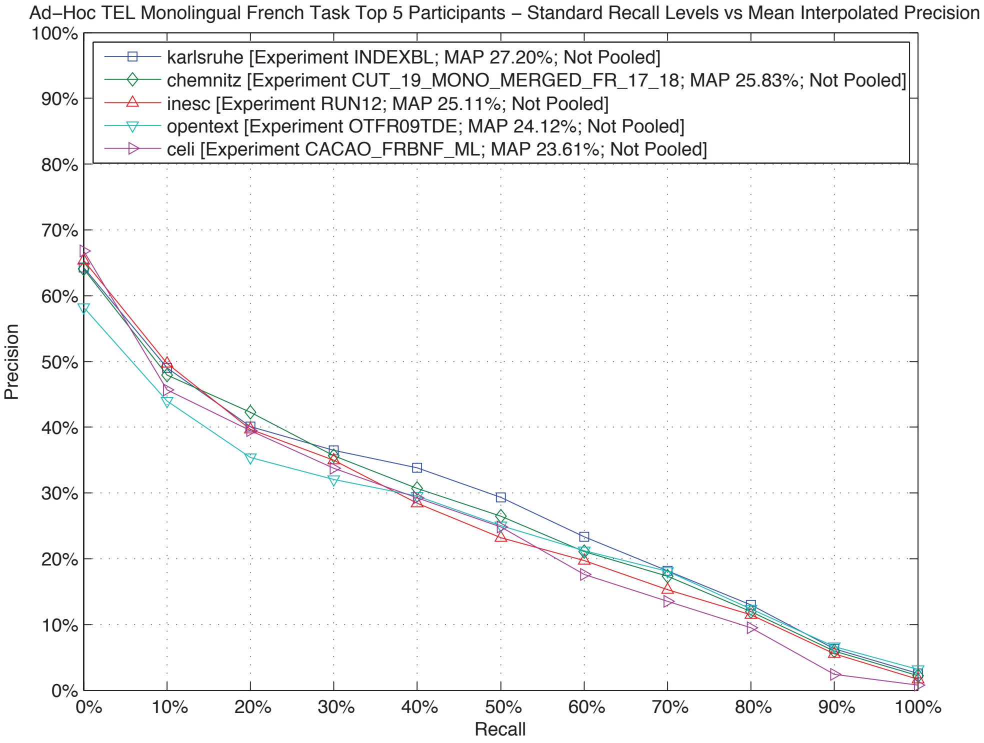

Fig. 6. Monolingual French

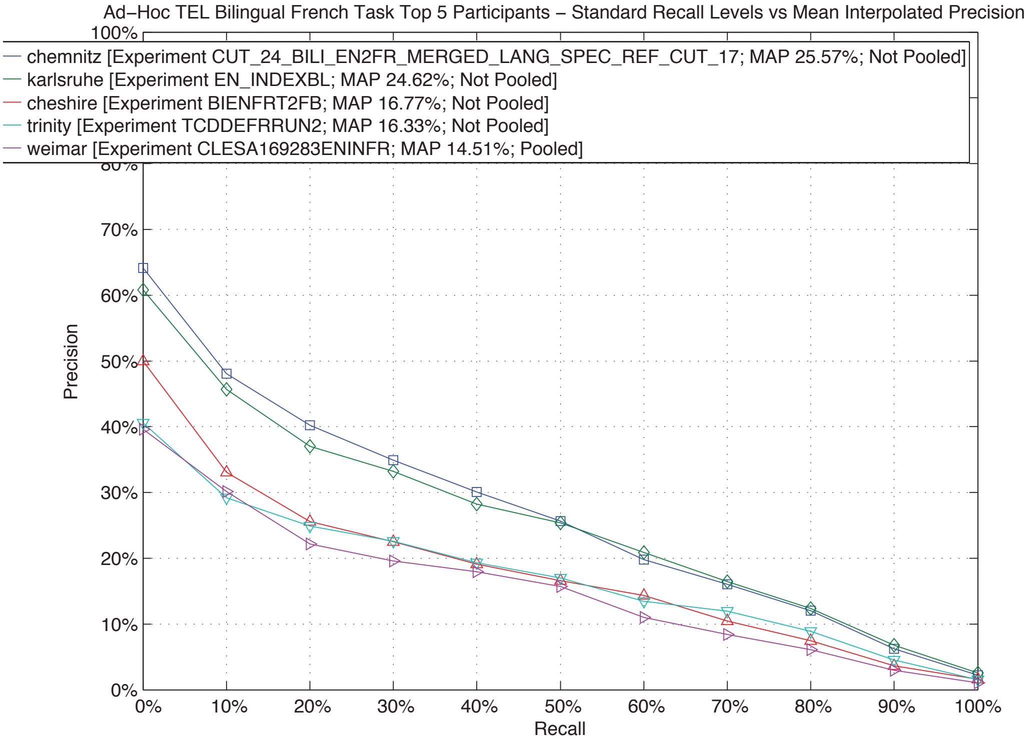

Fig. 7. Bilingual French

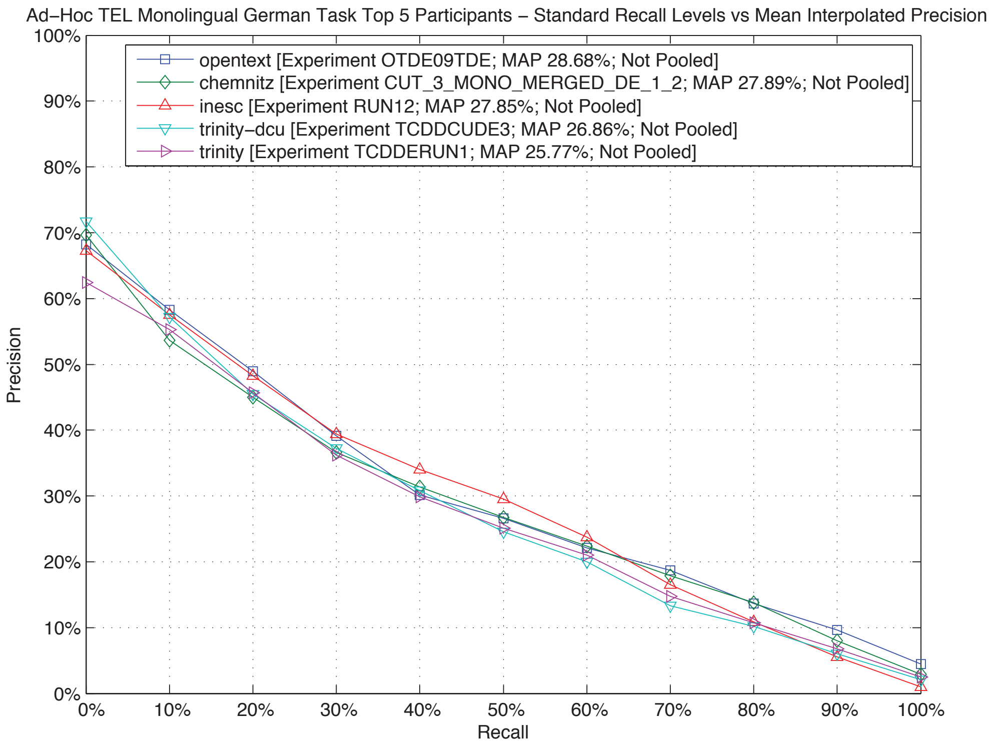

Fig. 8. Monolingual German

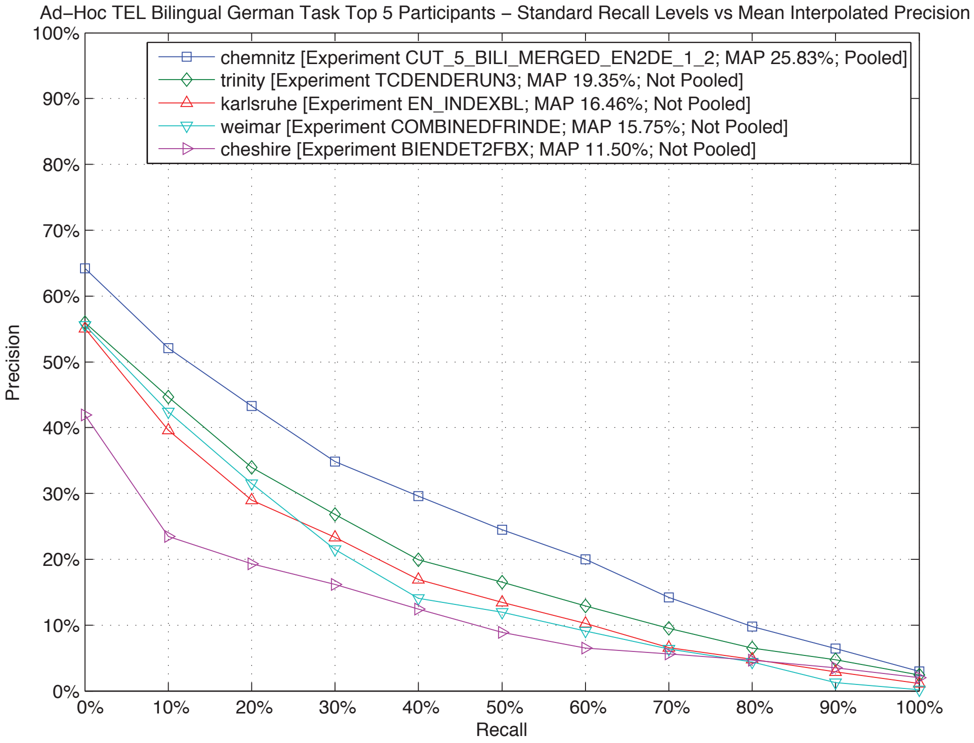

Fig. 9. Bilingual German

Table 5. Best entries for the bilingual TEL tasks.

Track	Rank	Participant	Experiment DOI	MAP
English	**1st**	chemnitz	10.2415/AH-TEL-BILI-X2EN-CLEF2009.CHEMNITZ.CUT_13_BILI_MERGED_DE2EN_9_10	40.46%
	2nd	hit	10.2415/AH-TEL-BILI-X2EN-CLEF2009.HIT.XTDD10T40	35.27%
	3rd	trinity	10.2415/AH-TEL-BILI-X2EN-CLEF2009.TRINITY.TCDDEENRUN3	35.05%
	4th	trinity-dcu	10.2415/AH-TEL-BILI-X2EN-CLEF2009.TRINITY-DCU.TCDDCUDEEN1	33.33%
	5th	karlsrhue	10.2415/AH-TEL-BILI-X2EN-CLEF2009.KARLSRUHE.DE_INDEXBL	32.70%
	Difference			23.73%
French	**1st**	chemnitz	10.2415/AH-TEL-BILI-X2FR-CLEF2009.CHEMNITZ.CUT_24_BILI_EN2FR_MERGED_LANG_SPEC_REF_CUT_17	25.57%
	2nd	karlsrhue	10.2415/AH-TEL-BILI-X2FR-CLEF2009.KARLSRUHE.EN_INDEXBL	24.62%
	3rd	chesire	10.2415/AH-TEL-BILI-X2FR-CLEF2009.CHESHIRE.BIENFRT2FB	16.77%
	4th	trinity	10.2415/AH-TEL-BILI-X2FR-CLEF2009.TRINITY.TCDDEFRRUN2	16.33%
	5th	weimar	10.2415/AH-TEL-BILI-X2FR-CLEF2009.WEIMAR.CLESA169283ENINFR	14.51%
	Difference			69.67%
German	**1st**	chemnitz	10.2415/AH-TEL-BILI-X2DE-CLEF2009.CHEMNITZ.CUT_5_BILI_MERGED_EN2DE_1_2	25.83%
	2nd	trinity	10.2415/AH-TEL-BILI-X2DE-CLEF2009.TRINITY.TCDENDERUN3	19.35%
	3rd	karlsrhue	10.2415/AH-TEL-BILI-X2DE-CLEF2009.KARLSRUHE.EN_INDEXBL	16.46%
	4th	weimar	10.2415/AH-TEL-BILI-X2DE-CLEF2009.WEIMAR.COMBINEDFRINDE	15.75%
	5th	chesire	10.2415/AH-TEL-BILI-X2DE-CLEF2009.CHESHIRE.BIENDET2FBX	11.50%
	Difference			124.60%

– X → DE: 53.15% of best monolingual German IR system.

In particular, it can be seen that there is a considerable improvement in performance for French and German This will be commented in the following section.

The monolingual performance figures for all three tasks are quite similar to those of last year but as these are not absolute values, no real conclusion can be drawn from this.

3.3 Approaches

As stated in the introduction, the TEL task this year is a repetition of the task set last year. A main reason for this was to create a good reusable test collection with a sufficient number of topics; another reason was to see whether the experience gained and reported in the literature last year, and the opportunity to use last year's test collection as training data, would lead to differences in approaches and/or improvements in performance this year. Although we have exactly the same number of participants this year as last year, only five of the thirteen 2009 participants also participated in 2008. These are the groups tagged as Chemnitz, Cheshire, Karlsruhe, INESC-ID and Opentext. The last two of these groups only tackled monolingual tasks. These groups all tend to appear in the top five for the various tasks. In the following we attempt to examine briefly the approaches adopted this year, focusing mainly on the cross-language experiments.

In the TEL task in CLEF 2008, we noted that all the traditional approaches to monolingual and cross language retrieval were attempted by the different groups. Retrieval methods included language models, vector-space and probabilistic approaches, and translation resources ranged from bilingual dictionaries, parallel and comparable corpora to on-line MT systems and Wikipedia. Groups often used a combination of more than one resource. What is immediately noticeable in 2009 is that, although similarly to last year a number of different retrieval models were tested, there is a far more uniform approach to the translation problem.

Five of the ten groups that attempted cross-language tasks used the Google Translate functionality, while a sixth used the LEC Power Translator [13]. Another group also used an MT system combining it with concept-based techniques but did not disclose the name of the MT system used [16]. The remaining three groups used a bilingual term list [17], a combination of resources including on-line and in house developed dictionaries [19], and Wikipedia translation links [18]. It is important to note that four out of the five groups in the bilingual to English and bilingual to French tasks and three out of five for the bilingual to German task used Google Translate, either on its own or in combination with another technique. One group noted that topic translation using a statistical MT system resulted in about 70% of the mean average precision (MAP) achieved when using Google Translate [20]. Another group [11] found that the results obtained by simply translating the query into all the target languages via Google gave results that were comparable to a far more complex strategy known as Cross-Language Explicit Semantic Analysis, CL-ESA, where the library catalog records and the queries are represented in a multilingual concept space that is spanned by aligned Wikipedia articles. As this year's results were significantly better than last year's, can we take this as meaning that Google is going to solve the cross-language translation resource quandary?

Taking a closer look at three groups that did consistently well in the cross-language tasks we find the following. The group that had the top result for each of the three tasks was Chemnitz [15]. They also had consistently good monolingual results. Not surprisingly, they appear to have a very strong IR engine, which uses various retrieval models and combines the results. They used Snowball stemmers for English and French and an n-gram stemmer for German. They were one of the few groups that tried to address the multilinguality of the target collections. They used the Google service to translate the topic from the source language to the four most common languages in the target collections, queried the four indexes and combined the results in a multilingual result set. They found that their approach combining multiple indexed collections worked quite well for French and German but was disappointing for English.

Another group with good performance, Karlsruhe [16], also attempted to tackle the multilinguality of the collections. Their approach was again based on multiple indexes for different languages with rank aggregation to combine the different partial results. They ran language detectors on the collections to identify the different languages contained and translated the topics to the languages recognized. They used Snowball stemmers to stem terms in ten main languages, fields in other languages were not preprocessed. Disappointingly, a baseline consisting of a single index without language classification and a topic translated only to the index language achieved similar or even better results. For the translation step, they combined MT with a concept-based retrieval strategy based on Explicit Semantic Analysis and using the Wikipedia database in English, French and German as concept space.

A third group that had quite good cross-language results for all three collections was Trinity [12]. However, their monolingual results were not so strong. They used a language modelling retrieval paradigm together with a document re-ranking method which they tried experimentally in the cross-language context. Significantly, they also used Google Translate. Judging from the fact that they did not do so well in the monolingual tasks, this seems to be the probable secret of their success for cross-language.

4 Persian@CLEF

This activity was again coordinated in collaboration with the Data Base Research Group (DBRG) of Tehran University.

4.1 Tasks

The activity was organised as a typical ad hoc text retrieval task on newspaper collections. Two tasks were offered: monolingual retrieval; cross-language retrieval (English queries to Persian target) and 50 topics were prepared (see section 2.2). For each topic, participants had to find relevant documents in the collection and submit the results in a ranked list.

Table 3 provides a breakdown of the number of participants and submitted runs by task and topic language.

4.2 Results

Table 6 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 mean average precision achieved by the experiment; the DOI of the experiment; and the performance difference between the first and the last participant.

Figures 10 and 11 compare the performances of the top participants of the Persian tasks.

For bilingual retrieval evaluation, a common method is to compare results against monolingual baselines. We have the following results for CLEF 2009:

- X → FA: 5.50% of best monolingual Farsi IR system.

This appears to be a very clear indication that something went wrong with the bilingual system that has been developed. These results should probably be discounted.

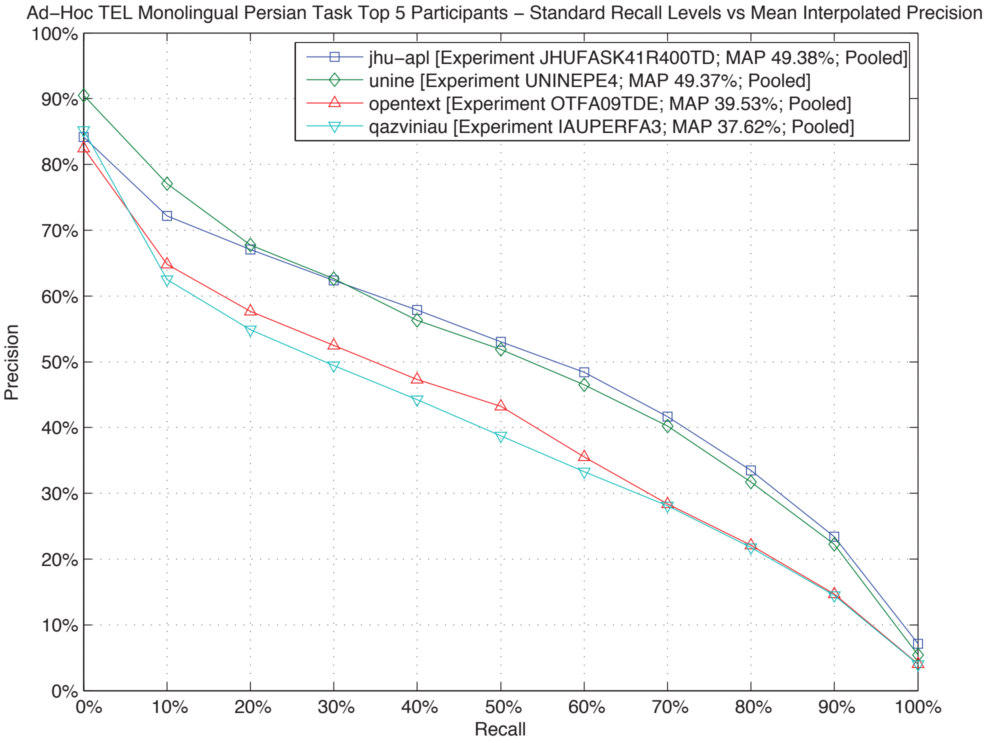

Fig. 10. Monolingual Persian

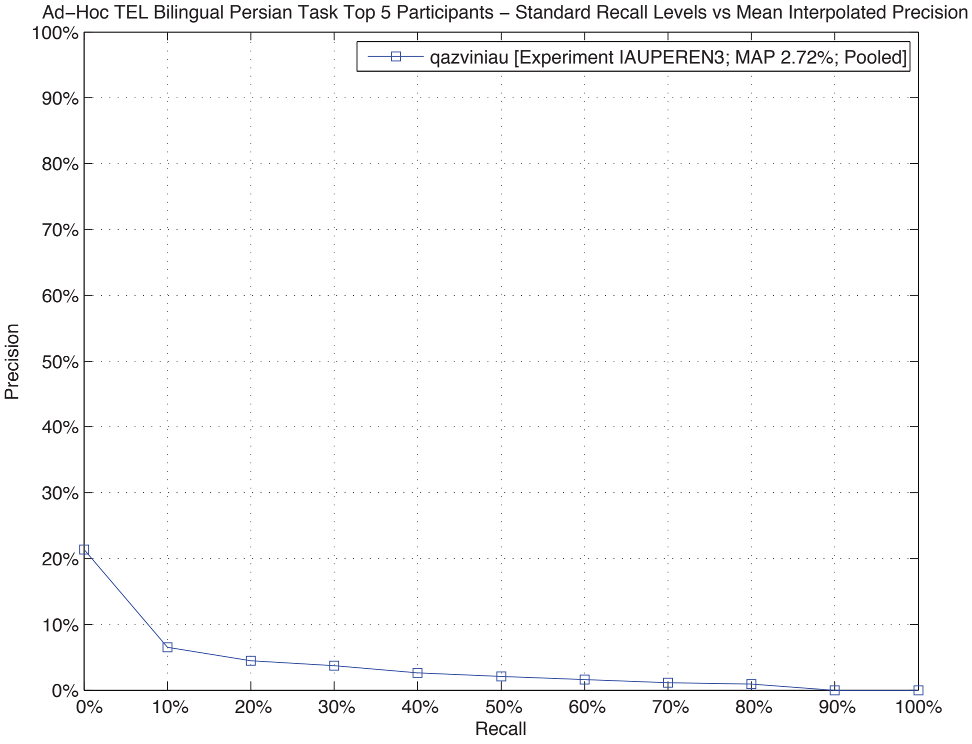

Fig. 11. Bilingual Persian

Table 6. Best entries for the Persian tasks.

Track	Rank	Participant	Experiment DOI	MAP
Monolingual	**1st**	jhu-apl	10.2415/AH-PERSIAN-MONO-FA-CLEF2009.JHU-APL.JHUFASK41R400TD	49.38%
	2nd	unine	10.2415/AH-PERSIAN-MONO-FA-CLEF2009.UNINE.UNINEPE4	49.37%
	3rd	opentext	10.2415/AH-PERSIAN-MONO-FA-CLEF2009.OPENTEXT.OTFA09TDE	39.53%
	4th	qazviniau	10.2415/AH-PERSIAN-MONO-FA-CLEF2009.QAZVINIAU.IAUPERFA3	37.62%
	5th	–	–	–%
	Difference			31.25%
Bilingual	**1st**	qazviniau	10.2415/AH-PERSIAN-BILI-X2FA-CLEF2009.QAZVINIAU.IAUPEREN3	2.72%
	2nd	–	–	–
	3rd	–	–	–
	4th	–	–	–
	5th	–	–	–
	Difference			–

4.3 Approaches

We were very disappointed this year that despite the fact that 14 groups registered for the Persian task, only four actually submitted results. And only one of these groups was from Iran. We suspect that one of the reasons for this was that the date for submission of results was not very convenient for the Iranian groups. Furthermore, only one group [18] attempted the bilingual task with the very poor results cited above. The technique they used was the same as that adopted for their bilingual to English experiments, exploiting Wikipedia translation links, and the reason they give for the very poor performance here is that the coverage of Farsi in Wikipedia is still very scarce compared to that of many other languages.

In the monolingual Persian task, the top two groups had very similar performance figures. [21] found they had best results using a light suffix-stripping algorithm and by combining different indexing and searching strategies. Interestingly, their results this year do not confirm their findings for the same task last year when the use of stemming did not prove very effective. The other group [14] tested variants of character n-gram tokenization; 4-grams, 5-grams, and skipgrams all provided about a 10% relative gain over plain words.

5 Conclusions

In CLEF 2009 we deliberately repeated the TEL and Persian tasks offered in 2008 in order to build up our test collections. Although we have not yet had sufficient time to assess them in depth, we are reasonably happy with the results for the TEL task: several groups worked on tackling the particular features of the TEL collections with varying success; evidence has been acquired on the effectiveness of a number of different IR strategies; there is a very strong indication of the validity of the Google Translate functionality.

On the other hand, the results for the Persian task were quite disappointing: very few groups participated; the results obtained are either in contradiction to those obtained previously and thus need further investigation [21] or tend to be a very straightforward repetition and confirmation of last year's results [14].

6 Acknowledgements

The TEL task was studied in order to provide useful input to The European Library (TEL); we express our gratitude in particular to Jill Cousins, Programme Director, and Sjoerd Siebinga, Technical Developer of TEL. Vivien Petras, Humboldt University, Germany, and Nicolas Moreau, Evaluation and Language Resources Distribution Agency, France, were responsible for the creation of the topics and the supervision of the relevance assessment work for the ONB and BNF data respectively. We thank them for their valuable assistance.

We should also like to acknowledge the enormous contribution to the coordination of the Persian task made by the Data Base Research group of the University of Tehran and in particular to Abolfazl AleAhmad and Hadi Amiri. They were responsible for the preparation of the set of topics for the Hamshahri collection in Farsi and English and for the subsequent relevance assessments.

Least but not last, we would warmly thank Giorgio Maria Di Nunzio for all the contributions he gave in carrying out the TEL and Persian tasks.

References

1. M. Agosti, G. M. Di Nunzio, and N. Ferro. The Importance of Scientific Data Curation for Evaluation Campaigns. In C. Thanos and F. Borri, editors, *DELOS Conference 2007 Working Notes*, pages 185–193. ISTI-CNR, Gruppo ALI, Pisa, Italy, February 2007.
2. O. Alonso and S. Mizzaro. Can we get rid of TREC assessors? Using Mechanical Turk for relevance assessment. In S. Geva, J. Kamps, C. Peters, T. Sakai, A. Trotman, and E. Voorhees, editors, *Proc. SIGIR 2009 Workshop on The Future of IR Evaluation.* `http://staff.science.uva.nl/~kamps/ireval/papers/paper_22.pdf`, 2009.
3. M. Braschler. CLEF 2002 – Overview of Results. In C. Peters, M. Braschler, J. Gonzalo, and M. Kluck, editors, *Advances in Cross-Language Information Retrieval: Third Workshop of the Cross–Language Evaluation Forum (CLEF 2002) Revised Papers*, pages 9–27. Lecture Notes in Computer Science (LNCS) 2785, Springer, Heidelberg, Germany, 2003.
4. M. Braschler and C. Peters. CLEF 2003 Methodology and Metrics. In C. Peters, M. Braschler, J. Gonzalo, and M. Kluck, editors, *Comparative Evaluation of Multilingual Information Access Systems: Fourth Workshop of the Cross–Language Evaluation Forum (CLEF 2003) Revised Selected Papers*, pages 7–20. Lecture Notes in Computer Science (LNCS) 3237, Springer, Heidelberg, Germany, 2004.
5. C. W. Cleverdon. The Cranfield Tests on Index Languages Devices. In K. Spärck Jones and P. Willett, editors, *Readings in Information Retrieval*, pages 47–60. Morgan Kaufmann Publisher, Inc., San Francisco, CA, USA, 1997.

6. G. M. Di Nunzio and N. Ferro. Appendix A: Results of the TEL@CLEF Task. In this volume.
7. G. M. Di Nunzio and N. Ferro. Appendix B: Results of the Persian@CLEF Task. In this volume.
8. M. Sanderson and H. Joho. Forming Test Collections with No System Pooling. In M. Sanderson, K. Järvelin, J. Allan, and P. Bruza, editors, *Proc. 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2004)*, pages 33–40. ACM Press, New York, USA, 2004.
9. S. Tomlinson. German, French, English and Persian Retrieval Experiments at CLEF 2009. In this volume.
10. Tomlinson, S.: Sampling Precision to Depth 10000 at CLEF 2008. Systems for Multilingual and Multimodal Information Access: 9th Workshop of the Cross-Language Evaluation Forum (CLEF 2008). Revised Selected Papers, Lecture Notes in Computer Science (LNCS) 5706, Springer, Heidelberg, Germany (2009)
11. Anderka, M., Lipka, N., Stein, B.: Evaluating Cross-Language Explicit Semantic Analysis and Cross Querying at TEL@CLEF 2009. In this volume.
12. Zhou, D., Wade, V: Language Modeling and Document Re-Ranking: Trinity Experiments at TEL@CLEF-2009. In this volume.
13. Larson, R.R.: Multilingual Query Expansion for CLEF Adhoc-TEL. In this volume.
14. McNamee, P.: JHU Experiments in Monolingual Farsi Document Retrieval at CLEF 2009. In this volume.
15. Kuersten, J.: Chemnitz at CLEF 2009 Ad-Hoc TEL Task: Combining Different Retrieval Models and Addressing the Multilinguality. In this volume.
16. Sorg, P., Braun, M., Nicolay, D., Cimiano, P.: Cross-lingual Information Retrieval based on Multiple Indexes. In this volume.
17. Katsiouli, P., Kalamboukis, T.: An Evaluation of Greek-English Cross Language Retrieval within the CLEF Ad-Hoc Bilingual Task. In this volume.
18. Jadidinejad, A.H., Mahmoudi, F.: Query Wikification: Mining Structured Queries From Unstructured Information Needs using Wikipedia-based Semantic Analysis. In this volume.
19. Bosca, A., Dini, L.: CACAO Project at the TEL@CLEF 2009 Task. In this volume.
20. Leveling, J., Zhou, D., Jones, G.F., Wade, V.: TCD-DCU at TEL@CLEF 2009: Document Expansion, Query Translation and Language Modeling. In this volume.
21. Dolamic, L., Fautsch, C., Savoy, J.: UniNE at CLEF 2009: Persian Ad Hoc Retrieval and IP. In this volume.