CLEF 2008: Ad Hoc Track Overview

Eneko Agirre[1], Giorgio Maria Di Nunzio[2], Nicola Ferro[2], Thomas Mandl[3],
and Carol Peters[4]

[1] Computer Science Department, University of the Basque Country, Spain
e.agirre@ehu.es
[2] Department of Information Engineering, University of Padua, Italy
{dinunzio,ferro}@dei.unipd.it
[3] Information Science, University of Hildesheim, Germany
mandl@uni-hildesheim.de
[4] ISTI-CNR, Area di Ricerca, Pisa, Italy
carol.peters@isti.cnr.it

Abstract. We describe the objectives and organization of the CLEF 2008 Ad Hoc track and discuss the main characteristics of the tasks offered to test monolingual and cross-language textual document retrieval systems. The track was changed considerably this year with the introduction of tasks with new document collections consisting of (i) library catalog records derived from The European Library, and (ii) and non-European language data, plus a task offering the chance to test retrieval with word sense disambiguated data. The track was thus structured in three distinct streams denominated: TEL@CLEF, Persian@CLEF and Robust WSD. The results obtained for each task are presented and statistical analyses are given.

1 Introduction

The Ad Hoc retrieval track is generally considered to be the core track in the *Cross-Language Evaluation Forum (CLEF)*. It is the one track that has been offered each year, from 2000 through 2008, and will be offered again in 2009. The aim of this track is to promote the development of monolingual and cross-language textual document retrieval systems. From 2000 - 2007, the track used exclusively collections of European newspaper and news agency documents[1] and worked hard at offering increasingly complex and diverse tasks, adding new languages each year. The results have been considerable; it is probably true to say that this track has done much to foster the creation of a strong European research community in the cross-language text retrieval area. It has provided the resources, the test collections and also the forum for discussion and comparison of ideas and approaches. Groups submitting experiments over several years have shown flexibility in advancing to more complex tasks, from monolingual to bilingual and multilingual experiments. Much work has been done

[1] Over the years, this track has built up test collections for monolingual and cross-language system evaluation in 14 European languages (see the Introduction to this volume for more details).

C. Peters et al. (Eds.): CLEF 2008, LNCS 5706, pp. 15–37, 2009.

on fine-tuning for individual languages while other efforts have concentrated on developing language-independent strategies. In fact, one of the papers in this section reports some interesting post-workshop experiments on previous CLEF Ad Hoc test collections in 13 languages, comparing the performance of different indexing approaches: word, stems, morphemes, n-gram stems and character n-grams [27].

This year the focus of the track was considerably widened: we introduced very different document collections, a non-European target language, and an information retrieval (IR) task designed to attract participation from groups interested in natural language processing (NLP). The track was thus structured in three distinct streams:

- TEL@CLEF
- Persian@CLEF
- Robust WSD

The first task was an application-oriented task, offering monolingual and cross-language search on library catalog records and was organized in collaboration with The European Library (TEL)[2]. The second task resembled the Ad Hoc retrieval tasks of previous years but this time the target collection was a Persian newspaper corpus. The third task was the robust activity which this year used word sense disambiguated (WSD) data, and involved English documents and monolingual and cross-language search in Spanish.

In this paper we first present the track setup, the evaluation methodology and the participation in the different tasks (Section 2). We then describe the main features of each task 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 rest of the 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 [10]. 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 distinguishing feature of the CLEF Ad Hoc track is that it applies this evaluation paradigm in a multilingual setting.

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

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

The three streams of the Ad Hoc track created very distinct test collections this year. The details are given in this section.

The Documents. Each of the three Ad Hoc tasks used a different set of documents.

The TEL task used three collections derived from:

- the British Library (BL); 1,000,100 documents, 1.2 GB;
- the Bibliothéque Nationale de France (BNF); 1,000,100 documents, 1.3 GB;
- the 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 language of the collection. However, each collection 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 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 information is normally in the main language of the collection. About 66% of the documents in the English and German collection have subject headings, only 37% in the French collection. Dewey Classification (DDC) is not available in the French collection; negligible (approx. 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.

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 is a Persian test collection that 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[3].

The robust task used existing CLEF news collections but with word sense disambiguation (WSD) information added. The word sense disambiguation data was automatically added by systems from two leading research laboratories, UBC [2] and NUS [9]. Both systems returned word senses from the English WordNet, version 1.6.

The document collections were offered both with and without WSD, and included the following[4]:

- LA Times 94 (with word sense disambiguated data); ca 113,000 documents, 425 MB without WSD, 1,448 MB (UBC) or 2,151 MB (NUS) with WSD;
- Glasgow Herald 95 (with word sense disambiguated data); ca 56,500 documents, 154 MB without WSD, 626 MB (UBC) or 904 MB (NUS) with WSD.

The 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 identified by means of a Digital Object Identifier (DOI)[5] of the experiment [30] which allows us to reference and cite them.

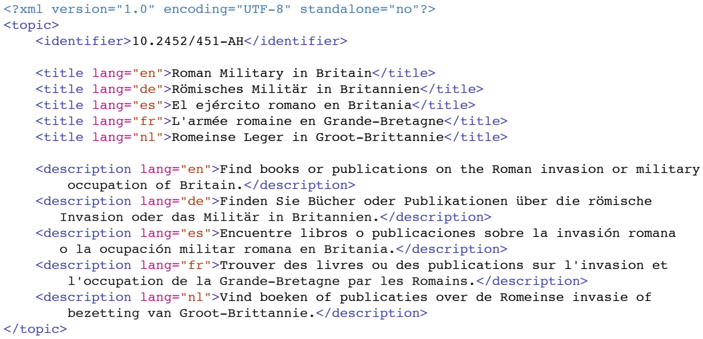
For the TEL task, a common set of 50 topics was prepared in each of the 3 main collection languages (English, French and German) plus Dutch and Spanish in response to demand. Only the Title and Description fields were released to the participants. The narrative was employed to provide information for 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 language in order to identify suitable common candidates. The sparseness of the data made this particularly difficult for the TEL task and tended to lead 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 of relevant documents over the collections. For each topic, the results expected from the separate collections can vary considerably, e.g. a topic of particular interest to Britain, such as the example given in Figure 1, can be

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

[4] A sample document and dtd are available at `http://ixa2.si.ehu.es/clirwsd/`

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

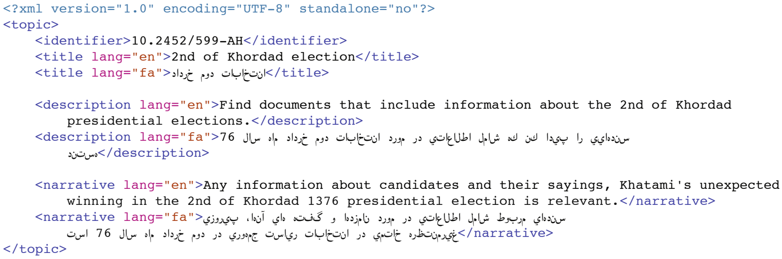


```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<topic>
    <identifier>10.2452/451-AH</identifier>

    <title lang="en">Roman Military in Britain</title>
    <title lang="de">Römisches Militär in Britannien</title>
    <title lang="es">El ejército romano en Britania</title>
    <title lang="fr">L'armée romaine en Grande-Bretagne</title>
    <title lang="nl">Romeinse Leger in Groot-Brittannie</title>

    <description lang="en">Find books or publications on the Roman invasion or military
        occupation of Britain.</description>
    <description lang="de">Finden Sie Bücher oder Publikationen über die römische
       Invasion oder das Militär in Britannien.</description>
    <description lang="es">Encuentre libros o publicaciones sobre la invasión romana
       o la ocupación militar romana en Britania.</description>
    <description lang="fr">Trouver des livres ou des publications sur l'invasion et
        l'occupation de la Grande-Bretagne par les Romains.</description>
    <description lang="nl">Vind boeken of publicaties over de Romeinse invasie of
        bezetting van Groot-Brittannie.</description>
</topic>
```

Fig. 1. Example of TEL topic in all five languages: topic `10.2452/451-AH`



```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<topic>
    <identifier>10.2452/599-AH</identifier>
    <title lang="en">2nd of Khordad election</title>
    <title lang="fa">انتخابات دوم خرداد</title>

    <description lang="en">Find documents that include information about the 2nd of Khordad
        presidential elections.</description>
    <description lang="fa">سندهايي را پيدا كن كه شامل اطلاعاتي در مورد انتخابات دوم خرداد ماه سال 76
        هستند</description>

    <narrative lang="en">Any information about candidates and their sayings, Khatami's unexpected
        winning in the 2nd of Khordad 1376 presidential election is relevant.</narrative>
    <narrative lang="fa">سندهاي مربوط شامل اطلاعاتي در مورد نامزدها و گفته هاي آنها، پيروزي
        غيرمنتظره خاتمي در انتخابات رياست جمهوري در دوم خرداد ماه سال 76 است</narrative>
</topic>
```

Fig. 2. Example of Persian topic: topic `10.2452/599-AH`

expected to find far more relevant documents in the BL collection than in BNF or ONB.

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.

For example, Iran commonly uses a different calendar from Europe and reference was often made in the Persian topics to events that are well known to Iranian society but not often discussed in English. This is shown in the example of Figure 2, where the rather awkward English rendering evidences the uncertainty of the translator.

The WSD robust task used existing CLEF topics in English and Spanish as follows:

- CLEF 2001; Topics 41-90; LA Times 94
- CLEF 2002; Topics 91-140; LA Times 94
- CLEF 2003; Topics 141-200; LA Times 94, Glasgow Herald 95

```
<top>
    <num>10.2452/141-WSD-AH</num>

    <EN-title>
        <TERM ID="10.2452/141-WSD-AH-1" LEMA="letter" POS="NNP">
            <WF>Letter</WF>
            <SYNSET SCORE="0" CODE="05115901-n"/>
            <SYNSET SCORE="0" CODE="05362432-n"/>
            <SYNSET SCORE="0" CODE="05029514-n"/>
            <SYNSET SCORE="1" CODE="04968965-n"/>
        </TERM>

        <TERM ID="10.2452/141-WSD-AH-2" LEMA="bomb" POS="NNP">
            <WF>Bomb</WF>
            <SYNSET SCORE="0.888888888888889" CODE="02310834-n"/>
            <SYNSET SCORE="0" CODE="05484679-n"/>
            <SYNSET SCORE="0.111111111111111" CODE="02311368-n"/>
        </TERM>

        <TERM ID="10.2452/141-WSD-AH-3" LEMA="for" POS="IN">
            <WF>for</WF>
        </TERM>

        ...

    </EN-title>

    <EN-desc>
        <TERM ID="10.2452/141-WSD-AH-5" LEMA="find" POS="VBP">
            <WF>Find</WF>
            <SYNSET SCORE="0" CODE="00658116-v"/>

            ...

        </TERM>

        ...

    </EN-desc>

    <EN-narr>
        ...
    </EN-narr>
</top>
```

Fig. 3. Example of Robust WSD topic: topic `10.2452/141-WSD-AH`

- CLEF 2004; Topics 201-250; Glasgow Herald 95
- CLEF 2005; Topics 251-300; LA Times 94, Glasgow Herald 95
- CLEF 2006; Topics 301-350; LA Times 94, Glasgow Herald 95

Topics from years 2001, 2002 and 2004 were used as training topics (relevance assessments were offered to participants), and topics from years 2003, 2005 and 2006 were used for the test.

All topics were offered both with and without WSD. Topics in English were disambiguated by both UBC [2] and NUS [9] systems, yielding word senses from WordNet version 1.6. A large-scale disambiguation system for Spanish was not available, so we used the first-sense heuristic, yielding senses from the Spanish wordnet, which is tightly aligned to the English WordNet version 1.6 (i.e., they share synset numbers or sense codes). An excerpt from a topic is shown in Figure 3, where each term in the topic is followed by its senses with their respective scores as assigned by the automatic WSD system[6].

Relevance Assessment. The number of documents in large test collections such as CLEF makes it impractical to judge every document for relevance. Instead

[6] Full sample and dtd are available at `http://ixa2.si.ehu.es/clirwsd/`

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 selected runs according to a set of predefined criteria. 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 [7] with respect to the CLEF 2003 pools. New pools were formed in CLEF 2008 for the runs submitted for the TEL and the Persian mono- and bilingual tasks. Instead, the robust tasks used the original pools and relevance assessments from previous CLEF campaigns.

The main criteria used when constructing the pools were:

- 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, from 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.

One important limitation when forming the pools is the number of documents to be assessed. Last year, for collections of newspaper documents, we estimated that assessors could judge from 60 to 100 documents per hour, providing binary judgments: relevant / not relevant. Our estimate this year for the TEL catalog records was 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.

Thus, in CLEF 2008, we used a depth of the top 60 ranked documents from selected runs in order to build pools of more-or-less equivalent size (approx. 25,000 documents) for the TEL English, French, and German and the Persian task[7]. Our CLEF2008 Working Notes paper reports summary information on the 2008 Ad Hoc pools used to calculate the results for the main monolingual and bilingual experiments. 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.

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

In addition the distribution of relevant documents across the topics is compared for the different Ad Hoc pools [4].

For the TEL documents, we judged for relevance only those documents that are written totally or partially in English, French and German (and Spanish for searches on the English collection), 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 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.

The relevance judgments for the Persian results were done by the DBRG group in Tehran. Again, assessment was performed on a binary basis and the standard CLEF assessment rules were applied, e.g. if in doubt with respect to the relevance of a given document, assessors are requested to ask themselves whether the document in question would be useful in any way if they had to write a report on the given topic.

As has already been stated, the robust WSD task used existing relevance assessments from previous years. The relevance assessments regarding the training topics were provided to participants before competition time.

This year, we tried a slight improvement with respect to the traditional pooling strategy adopted so far in CLEF. During the topic creation phase, the assessors express their opinion about the relevance of the documents they inspect with respect to the topic. Although this opinion may change during the various discussions between assessors in this phase, we consider these indications as potentially useful in helping to strengthen the pools of documents that will be judged for relevance. These documents are thus added to the pools. However, the assessors are not informed of which documents they had previously judged in order not to bias them in any way.

Similarly to last year, in his paper, Stephen Tomlinson, has reported some sampling experiments aimed at estimating the judging coverage for the CLEF 2008 TEL and Persian test collections. He finds that this tends to be lower than the estimates he produced for the CLEF 2007 collections. With respect to the TEL collections, the implication is that at best 55% of the relevant documents are included in the pools - however, most of the unjudged relevant documents are for the 10 or more queries that have the most known answers [33]. According to studies on earlier TREC collections which gave similar results, in any case this "level of completeness" should be acceptable. For Persian the coverage is much lower - around 25%; this could be a result of the fact that all the Persian topics tend to be relatively broad. This year's Persian collection is thus considered to be less stable than usual.

2.2 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 [8]. For the robust task, we used additional measures, see Section 5.

The individual results for all official Ad Hoc experiments in CLEF 2008 are given in the Appendices of the CLEF 2008 Working Notes [14],[15], [16].

2.3 Participants and Experiments

As shown in Table 1, a total of 24 groups from 14 different countries submitted official results for one or more of the Ad Hoc tasks - a slight increase on the

Table 1. CLEF 2008 Ad Hoc participants

Participant	Institution	Country
chemnitz	Chemnitz University of Technology	Germany
cheshire	U.C.Berkeley	United States
geneva	University of Geneva	Switzerland
imag	Inst. for Infocomm Research	France
inaoe	INAOE	Mexico
inesc	INESC ID	Portugal
isi	Indian Statistical Institute	India
ixa	Univ. Basque Country	Spain
jhu-apl	Johns Hopkins University Applied Physics Lab	United States
karlsruhe	University of Karlsruhe	Germany
know-center	Knowledge Relationship Discovery	Austria
opentext	Open Text Corporation	Canada
tehran-IRDB	IR-DB Research Group	Iran
tehran-NLP	NLP-Software Engineering Grad. Lab	Iran
tehran-NLPDB	NLP-DB Research Group	Iran
tehran-NLPDB2	NLP-DB Group	Iran
tehran-SEC	School of Electrical Computing-1	Iran
twente	Univ. of Twente	Netherlands
ucm	Universidad Complutense de Madrid	Spain
ufrgs	Univ. Fed. do Rio Grande do Sul	Brazil
uniba	Universita' di Bari	Italy
unine	U.Neuchatel-Informatics	Switzerland
xerox	Xerox Reseearch - Data Mining	France
xerox-sas	Xerox SAS	Italy

Table 2. Breakdown of experiments into tracks and topic languages

(a) Number of experiments per track, participant.

Track	# Part.	# Runs
TEL Mono English	13	37
TEL Mono French	9	29
TEL Mono German	10	30
TEL Biling. English	8	24
TEL Biling. French	5	16
TEL Biling. German	6	17
Mono Persian	8	53
Biling. Persian	3	13
Robust Mono English Test	8	20
Robust Mono English Training	1	2
Robust Biling. English Test	4	8
Robust Mono English Test WSD	7	25
Robust Mono English Training WSD	1	5
Robust Biling. English Test WSD	4	10
Total		**289**

(b) List of experiments by topic language.

Topic Lang.	# Runs
English	120
Farsi	51
German	44
French	44
Spanish	26
Dutch	3
Portuguese	1
Total	**289**

22 participants of last year[8]. A total of 289 runs were submitted with an increase of about 22% on the 235 runs of 2007. The average number of submitted runs per participant also increased: from 10.6 runs/participant of 2007 to 12.0 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 (215 out of 289, 74.40%) used this combination of topic fields, 27 (9.34%) used all fields[9], 47 (16.26%) used the title field only. The majority of experiments were conducted using automatic query construction (273 out of 289, 94.47%) and only in a small fraction of the experiments (16 out 289, 5.53%) were queries been manually constructed from topics. A breakdown into the separate tasks is shown in Table 2(a).

Seven different topic languages were used in the Ad Hoc experiments. As always, the most popular language for queries was English, with Farsi second. The number of runs per topic language is shown in Table 2(b).

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

[8] Two additional Spanish groups presented results after the deadline for the robust tasks; their results were thus not reported in the official list but their papers are included in this volume [26], [28].

[9] The narrative field was only offered for the Persian and Robust tasks.

effective retrieval technologies for searching this type of very sparse data. 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?"

3.1 Tasks

Two subtasks were offered: Monolingual and Bilingual. 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 main 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.

In CLEF 2008, the activity we simulated was that of users who have a working knowledge of English, French and German (plus wrt the English collection also Spanish) and who want to discover the existence of relevant documents that can be useful for them in one of our three target collections. One of our suppositions was that, knowing that these collections are to some extent multilingual, some systems may attempt to use specific tools to discover this. For example, a system trying the cross-language English to French task on the BNF target collection but knowing that documents retrieved in English and German will also be judged for relevance might choose to employ an English-German as well as the probable English-French dictionary. Groups attempting anything of this type were asked to declare such runs with a ++ indication.

3.2 Participants

13 groups submitted 153 runs for the TEL task: all groups submitted monolingual runs (96 runs out of 153); 8 groups also submitted bilingual runs (57 runs out of 153). Table 2(a) provides a breakdown of the number of participants and submitted runs by task.

3.3 Results

Monolingual Results. Table 3 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.

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

Table 3. Best entries for the monolingual TEL tasks

Track	Rank	Participant	Experiment DOI	MAP
English	**1st**	unine	10.2415/AH-TEL-MONO-EN-CLEF2008.UNINE.UNINEEN3	37.53%
	2nd	inesc	10.2415/AH-TEL-MONO-EN-CLEF2008.INESC.RUN3	36.23%
	3rd	chemnitz	10.2415/AH-TEL-MONO-EN-CLEF2008.CHEMNITZ.CUT_SIMPLE	35.61%
	4th	jhu-apl	10.2415/AH-TEL-MONO-EN-CLEF2008.JHU-APL.JHUMOEN4RF	35.31%
	5th	cheshire	10.2415/AH-TEL-MONO-EN-CLEF2008.CHESHIRE.BKAHTELMENTDT2F	34.66%
	Difference			8.28%
French	**1st**	unine	10.2415/AH-TEL-MONO-FR-CLEF2008.UNINE.UNINEFR3	33.27%
	2nd	xerox	10.2415/AH-TEL-MONO-FR-CLEF2008.XEROX.J1	30.88%
	3rd	jhu-apl	10.2415/AH-TEL-MONO-FR-CLEF2008.JHU-APL.JHUMOFR4	29.50%
	4th	opentext	10.2415/AH-TEL-MONO-FR-CLEF2008.OPENTEXT.OTFR08TD	25.23%
	5th	chesire	10.2415/AH-TEL-MONO-FR-CLEF2008.CHESHIRE.BKAHTELMFRTDT2FB	24.37%
	Difference			36.52%
German	**1st**	opentext	10.2415/AH-TEL-MONO-DE-CLEF2008.OPENTEXT.OTDE08TDE	35.71%
	2nd	jhu-apl	10.2415/AH-TEL-MONO-DE-CLEF2008.JHU-APL.JHUMODE4	33.77%
	3rd	unine	10.2415/AH-TEL-MONO-DE-CLEF2008.UNINE.UNINEDE1	30.12%
	4th	xerox	10.2415/AH-TEL-MONO-DE-CLEF2008.XEROX.T1	27.36%
	5th	inesc	10.2415/AH-TEL-MONO-DE-CLEF2008.INESC.RUN3	22.97%
	Difference			55.46%

Table 4. Best entries for the bilingual TEL tasks

Track	Rank	Participant	Experiment DOI	MAP
English	**1st**	chemnitz	10.2415/AH-TEL-BILI-X2EN-CLEF2008.CHEMNITZ.CUT_SIMPLE_DE2EN	34.15%
	2nd	chesire	10.2415/AH-TEL-BILI-X2EN-CLEF2008.CHESHIRE.BKAHTELBFRENTDT2FB	28.24%
	3rd	ufrgs	10.2415/AH-TEL-BILI-X2EN-CLEF2008.UFRGS.UFRGS_BI_SP_EN2	23.15%
	4th	twente	10.2415/AH-TEL-BILI-X2EN-CLEF2008.TWENTE.FCW	22.78%
	5th	jhu-apl	10.2415/AH-TEL-BILI-X2EN-CLEF2008.JHU-APL.JHUBIDEEN5	21.11%
	Difference			61.77%
French	**1st**	chesire	10.2415/AH-TEL-BILI-X2FR-CLEF2008.CHESHIRE.BKAHTELBDEFRTDT2FB	18.84%
	2nd	chemnitz	10.2415/AH-TEL-BILI-X2FR-CLEF2008.CHEMNITZ.CUT_SIMPLE_EN2FR	17.54%
	3rd	jhu-apl	10.2415/AH-TEL-BILI-X2FR-CLEF2008.JHU-APL.JHUBINLFR5	17.46%
	4th	xerox	10.2415/AH-TEL-BILI-X2FR-CLEF2008.XEROX.GER_FRE_J	11.62%
	5th	xerox-sas	10.2415/AH-TEL-BILI-X2FR-CLEF2008.XEROX-SAS.CACAOENGFREPLAIN	6.78%
	Difference			177.87%
German	**1st**	jhu-apl	10.2415/AH-TEL-BILI-X2DE-CLEF2008.JHU-APL.JHUBIENDE5	18.98%
	2nd	chemnitz	10.2415/AH-TEL-BILI-X2DE-CLEF2008.CHEMNITZ.CUT_MERGED_SIMPLE_EN2DE	18.51%
	3rd	chesire	10.2415/AH-TEL-BILI-X2DE-CLEF2008.CHESHIRE.BKAHTELBENDETDT2FB	15.56%
	4th	xerox	10.2415/AH-TEL-BILI-X2DE-CLEF2008.XEROX.FRE_GER_J	12.05%
	5th	karlsruhe	10.2415/AH-TEL-BILI-X2DE-CLEF2008.KARLSRUHE.AIFB_ONB_EN	6.67%
	Difference			184.55%

For bilingual retrieval evaluation, a common method is to compare results against monolingual baselines. We have the following results for CLEF 2008:

- X → EN: 90.99% of best monolingual English IR system;
- X → FR: 56.63% of best monolingual French IR system;
- X → DE: 53.15% of best monolingual German IR system.

While the best result for English, obtained with German topics, is very good and can be considered as state-of-the-art for a cross-language system running on well-tested languages with reliable processing tools and resources such as English and German, the results for the other two target collections are fairly disappointing.

3.4 Approaches and Discussion

In the TEL experiments, all the traditional approaches to monolingual and cross-language retrieval were attempted by the different groups. Retrieval algorithms included language models, vector-space and probabilistic approaches, and translation resources ranged from bilingual dictionaries, parallel and comparable corpora, to on-line MT systems. Groups often used a combination of more than one resource.

One of the most interesting and new features of the TEL task was the multilinguality of the collections. Only about half of each collection was in the national language (English, French or German), with virtually all other languages represented by one or more entries in one or another of the collections. However, only a few groups took this into specific consideration trying to devise ways to address this aspect and, somewhat disappointingly, their efforts do not appear to have been particularly rewarded by improved performance.

This is shown by the group from the Technical University of Chemnitz, who had overall the best results in the bilingual tasks (1st for XtoEN; 2nd for XtoFR and DE) although they did not do so well in the monolingual tasks. In their official submissions for the campaign, this group attempted to tackle the multilinguality of the collections in several ways. First, they tried to identify the language of each record in the collections using a language detector. Unfortunately, due to an error, they were unable to use the indices created in this way[10]. Second, in both their monolingual and cross-language experiments they implemented a retrieval algorithm which translated the query into the top 10 (in terms of occurrence) languages and merged these multilingual terms into a single query. They ran experiments weighting the query in different ways on the basis of estimated distribution of language content in the collections. In the monolingual experiments, rather disappointingly, the results showed that their purely monolingual baseline always out performed experiments with query translations and language weights. This finding was confirmed with the bilingual experiments where again the better results were achieved with the baseline configurations. They attributed their good overall results for bilingual to the superiority of the Google online translation service. These experiments are described in their Working Notes submission [23]. In their paper in this volume, they describe a series of post workshop experiments for both mono- and cross-language tasks. Disappointingly, they found that their experiments on generating multilingual queries actually resulted in poorer retrieval effectiveness in all cases [22].

[10] This meant that they had to recreate their indices and perform all official experiments at the very last moment; this may have impacted on their results.

Another group that attempted to tackle the multilinguality of the target collections was Xerox. In their official runs, this group built a single index containing all languages (according to the expected languages which they identified as just English, French and German although, as stated, the collections actually contain documents in other languages as well). This, of course, meant that the queries also had to be issued in all three languages. They built a multilingual probabilistic dictionary and for each target collection gave more weight to the official language of the collection [11]. Although their results for both monolingual and bilingual experiments for the French and German collections were always within the top five; they were not quite so successful with the English collection. In their post-campaign experiments described in this volume, they propose an approach to handling target collections in multiple languages. However, and similarly to the work by the group from Chemnitz, their experiments showed that exploiting information in languages different from the official language of the collection gave no advantage[12].

Most groups actually ignored the multilinguality of the single collections in their experiments. Good examples of this are three veteran CLEF groups, UniNE which had, overall the best monolingual results, JHU which appeared in the top five for all bilingual tasks, and Berkeley which figured in the top five for all experiments except for monolingual German. UniNe appeared to focus on testing different IR models and combination approaches whereas the major interest of JHU was on the most efficient methods for indexing. Berkeley tested a version of the Logistic Regression (LR) algorithm that has been used very successfully in cross-language IR by Berkeley researchers for a number of years together with blind relevance feedback [18],[27], [24].

As has been mentioned, the TEL data is structured data; participants were told that they could use all fields. Some groups attempted to exploit this by weighting the contents of different fields differently. See, for example [25]. The combination used in the experiments of this group is based on repeating the title field three times, the subject field twice and keeping the other document fields unchanged.

To sum up, it appears that the majority of groups took this task as a traditional Ad Hoc retrieval task and applied traditional methods. However, it is far too early to confirm whether this is really the best approach to retrieval on library catalog cards. This task is being repeated in CLEF 2009 and we hope that the results will provide more evidence as to which are the most effective approaches when handling catalog data of this type.

4 Persian@CLEF

This activity was coordinated in collaboration with the Data Base Research Group (DBRG) of Tehran University. It was the first time that CLEF offered a non-European language target collection. Persian is an Indo-European language spoken in Iran, Afghanistan and Tajikistan. It is also known as Farsi.

We chose Persian as our first non-European target language for a number of reasons: its challenging script (a modified version of the Arabic alphabet with elision of short vowels) which is written from right to left; its morphology (extensive use of suffixes and compounding); its political and cultural importance. However, the main influencing factor was the generous offer from DBRG to provide an important newspaper corpus (Hamshahri) as the target collection and to be responsible for the coordination of the activity. This collaboration has proved very fruitful and intellectually stimulating and is being continued in 2009.

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. For each topic, participants had to find relevant documents in the collection and submit the results in a ranked list.

4.2 Participants

Eight groups submitted 66 runs for the Persian task: all eight submitted monolingual runs (53 runs out of 66); 3 groups also submitted bilingual runs (13 runs out of 66). Five of the groups were formed of Persian native speakers, mostly from the University of Tehran; they were all first time CLEF participants. The other three groups were CLEF veterans with much experience in the CLEF Ad Hoc track. Table 2(a) provides a breakdown of the number of participants and submitted runs by task.

4.3 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.

Table 5. Best entries for the Persian tasks

Track	Rank	Participant	Experiment DOI	MAP
Monolingual	**1st**	unine	10.2415/AH-PERSIAN-MONO-FA-CLEF2008.UNINE.UNINEPE2	48.98%
	2nd	jhu-apl	10.2415/AH-PERSIAN-MONO-FA-CLEF2008.JHU-APL.JHUFASK41R400	45.19%
	3rd	opentext	10.2415/AH-PERSIAN-MONO-FA-CLEF2008.OPENTEXT.OTFA08T	42.08%
	4th	tehran-nlpdb2	10.2415/AH-PERSIAN-MONO-FA-CLEF2008.TEHRAN-NLPDB2.UTNLPDB3INEXPC2	28.83%
	5th	tehran-nlpdb	10.2415/AH-PERSIAN-MONO-FA-CLEF2008.TEHRAN-NLPDB.UTNLPDB1MT	28.14%
	Difference			74.05%
Bilingual	**1st**	jhu-apl	10.2415/AH-PERSIAN-BILI-X2FA-CLEF2008.JHU-APL.JHUENFASK41R400	45.19%
	2nd	tehran-nlpdb	10.2415/AH-PERSIAN-BILI-X2FA-CLEF2008.TEHRAN-NLPDB.UTNLPDB1BT4G	14.45%
	3rd	tehran-sec	10.2415/AH-PERSIAN-BILI-X2FA-CLEF2008.TEHRAN-SEC.CLDTDR	12.88%
	4th	–	–	–
	5th	–	–	–
	Difference			250.85%

As stated above, a common method for bilingual retrieval evaluation is to compare results against monolingual baselines. We have the following results for CLEF 2008:

– X → FA: 92.26% of best monolingual Farsi IR system.

This appears to be in line with state-of-the-art performance for cross-language systems.

4.4 Approaches

As was to be expected a common theme in a number of the papers was the most effective way to handle the Persian morphology. The group with the best results in the monolingual task tested three approaches; no stemming, a light stemmer developed in-house, and a 4-gram indexing approach. Their best performance was achieved using their light stemmer which has been made freely available on their website. However, they commented that the loss in performance with the no stemming approach was not very great. This group also tested three probabilistic models: Okapi, DFR and statistical language model (LM). The best results were obtained with the latter two [18]. The participant with the second best results compared several different forms of textual normalization: character n-grams, n-gram stems, ordinary words, words automatically segmented into morphemes, and a novel form of n-gram indexing based on n-grams with character skips. He found that that character 4-grams performed the best [27]. This participant also performed some interesting post-workshop experiments on previous CLEF Ad Hoc test collections in 13 languages comparing the results. The findings of [18] were confirmed by [34] in his Working Notes paper. This participant also tested runs with no stemming, with the UniNE stemmer and with n-grams. Similarly, he reported that stemming had relatively little impact.

Somewhat surprisingly, most of the papers from Iran-based groups do not provide much information on morphological analysis or stemming in their papers. One mentions the application of a light Porter-like stemmer but reported that the algorithm adopted was too simple and results did not improve [5]. Only one of these groups provides some detailed discussion of the impact of stemming. This group used a simple stemmer (PERSTEM[11]) and reported that in most cases stemming did improve performance but noted that this was in contrast with experiments conducted by other groups at the University of Tehran on the same collection. They suggest that further experiments with different types of stemmers and stemming techniques are required in order to clarify the role of stemming in Persian text processing [21]. Two of the Persian groups also decided to annotate the corpus with part-of-speech tags in order to evaluate the impact of such information on the performance of the retrieval algorithms [20],[21]. The results reported do not appear to show any great boost in performance.

Other experiments by the groups from Iran included an investigation into the effect of fusion of different retrieval technique. Two approaches were tested:

[11] http://sourceforge.net/projects/perstem

combining the results of nine distinct retrieval methods; combining the results of the same method but with different types of tokens. The second strategy applied a vector space model and ran it with three different types of tokens namely 4-grams, stemmed single terms and unstemmed single terms. This approach gave better results [1].

For the cross-language task, the English topics were translated into Persian. As remarked above, the task of the translators was not easy as it was both a cross-language and also a cross-cultural task. The best result - again by a CLEF veteran participant - obtained 92% of the top monolingual performance. This is well in line with state-of-the-art performance for good cross-language retrieval systems. This group used an online machine translation system applied to the queries[12] [27].

The other two submissions for the cross-language task were from Iran-based groups. We have received a report from just one of them [5]. This group applied both query and document translation. For query translation they used a method based on the estimation of translation probabilities. In the document translation part they used the Shiraz machine translation system to translate the documents into English. They then created a Hybrid CLIR system by score-based merging of the two retrieval system results. The best performance was obtained with the hybrid system, confirming the reports of other researchers in previous CLEF campaigns, and elsewhere.

5 Robust – WSD Experiments

The robust task ran for the third time at CLEF 2008. It is an Ad Hoc retrieval task based on data of previous CLEF campaigns. The robust task emphasizes the difficult topics by a non-linear integration of the results of individual topics into one result for a system, using the geometric mean of the average precision for all topics (GMAP) as an additional evaluation measure [32,35]. Given the difficulty of the task, training data including topics and relevance assessments was provided for the participants to tune their systems to the collection.

This year the robust task also incorporated word sense disambiguation information provided by the organizers to the participants. The task follows the 2007 joint SemEval-CLEF task [3], and has the aim of exploring the contribution of word sense disambiguation to monolingual and cross-language information retrieval. Note that a similar exercise was also run in the question answering track at CLEF 2008. The goal of the task is to test whether WSD can be used beneficially for retrieval systems, and thus participants were required to submit at least one baseline run without WSD and one run using the WSD annotations. Participants could also submit four further baseline runs without WSD and four runs using WSD.

The experiment involved both monolingual (topics and documents in English) and bilingual experiments (topics in Spanish and documents in English). In

[12] http://www.parstranslator.net/eng/translate.htm

addition to the original documents and topics, the organizers of the task provided both documents and topics which had been automatically tagged with word senses from WordNet version 1.6 using two state-of-the-art word sense disambiguation systems, UBC [2] and NUS [9]. These systems provided weighted word sense tags for each of the nouns, verbs, adjectives and adverbs that they could disambiguate.

In addition, the participants could use publicly available data from the English and Spanish wordnets in order to test different expansion strategies. Note that given the tight alignment of the Spanish and English wordnets, the wordnets could also be used to translate directly from one sense to another, and perform expansion to terms in another language.

5.1 Participants

Eight groups submitted 63 runs for the Robust tasks: all groups submitted monolingual runs (45 runs out of 63); 4 groups also submitted bilingual runs (18 runs out of 63). Moreover, 7 groups participated in the WSD tasks, submitting 40 out of 63 runs, 30 monolingual and 10 bilingual. Table 2(a) provides a breakdown of the number of participants and submitted runs by task. Two further groups were late, so they are not included in the official results but they do have papers in this volume [26], [28].

5.2 Results

Monolingual Results. Table 6 shows the best results for this task. The performance difference between the best and the last (up to 5) placed group is given (in terms of average precision).

Bilingual Results. Table 7 shows the best results for this task. The performance difference between the best and the last (up to 5) placed group is given (in terms of average precision). All the experiments were from English to French.

Table 6. Best entries for the robust monolingual task

Track	Rank	Participant	Experiment DOI	MAP	GMAP
English	**1st**	unine	10.2415/AH-ROBUST-MONO-EN-TEST-CLEF2008.UNINE.UNINEROBUST4	45.14%	21.17%
	2nd	geneva	10.2415/AH-ROBUST-MONO-EN-TEST-CLEF2008.GENEVA.ISILEMTDN	39.17%	16.53%
	3rd	ucm	10.2415/AH-ROBUST-MONO-EN-TEST-CLEF2008.UCM.BM25_BO1	38.34%	15.28%
	4th	ixa	10.2415/AH-ROBUST-MONO-EN-TEST-CLEF2008.IXA.EN2ENNOWSDPSREL	38.10%	15.72%
	5th	ufrgs	10.2415/AH-ROBUST-MONO-EN-TEST-CLEF2008.UFRGS.UFRGS_R_MONO2_TEST	33.94%	13.96%
	Difference			33.03%	51.64%
English WSD	**1st**	unine	10.2415/AH-ROBUST-WSD-MONO-EN-TEST-CLEF2008.UNINE.UNINEROBUST6	44.98%	21.54%
	2nd	ucm	10.2415/AH-ROBUST-WSD-MONO-EN-TEST-CLEF2008.UCM.BM25_BO1_CLAUSES_09	39.57%	16.17%
	3rd	ixa	10.2415/AH-ROBUST-WSD-MONO-EN-TEST-CLEF2008.IXA.EN2ENUBCDOCSPSREL	38.99%	15.52%
	4th	geneva	10.2415/AH-ROBUST-WSD-MONO-EN-TEST-CLEF2008.GENEVA.ISINUSLWTDN	38.13%	16.25%
	5th	ufrgs	10.2415/AH-ROBUST-WSD-MONO-EN-TEST-CLEF2008.UFRGS.UFRGS_R_MONO_WSD5_TEST	34.64%	14.17%
	Difference			29.84%	52.01%

Table 7. Best entries for the robust bilingual task

Track	Rank	Participant	Experiment DOI	MAP	GMAP
English	**1st**	ufrgs	10.2415/AH-ROBUST-BILI-X2EN-TEST-CLEF2008.UFRGS.UFRGS_R_BI3_TEST	36.38%	13.00%
	2nd	geneva	10.2415/AH-ROBUST-BILI-X2EN-TEST-CLEF2008.GENEVA.ISIESENTD	30.36%	10.96%
	3rd	ixa	10.2415/AH-ROBUST-BILI-X2EN-TEST-CLEF2008.IXA.ES2ENNOWSDPSREL	19.57%	1.62%
	4th	uniba	10.2415/AH-ROBUST-BILI-X2EN-TEST-CLEF2008.UNIBA.CROSS1TDNUS2F	2.56%	0.04%
	5th	–	–	–	–
	Difference			1,321.09%	32,400.00%
English WSD	**1st**	ixa	10.2415/AH-ROBUST-WSD-BILI-X2EN-TEST-CLEF2008.IXA.ES2EN1STTOPSUBCDOCSPSREL	23.56%	1.71%
	2nd	ufrgs	10.2415/AH-ROBUST-WSD-BILI-X2EN-TEST-CLEF2008.UFRGS.UFRGS_R_BI_WSD1_TEST	21.77%	5.14%
	3rd	geneva	10.2415/AH-ROBUST-WSD-BILI-X2EN-TEST-CLEF2008.GENEVA.ISIESPWSDTDN	9.70%	0.37%
	4th	geneva	10.2415/AH-ROBUST-WSD-BILI-X2EN-TEST-CLEF2008.UNIBA.CROSSWSD12NUS2F	7.23%	0.16%
	5th	–	–	–	–
	Difference			225.86%	3,112.50%

Evaluating the bilingual retrieval evaluation, we have the following results for CLEF 2008:

- X → EN: 80.59% of best monolingual English IR system (MAP);
- X → EN WSD: 52.38% of best monolingual English IR system (MAP).

5.3 Analysis

In this section we focus on the comparison between WSD and non-WSD runs. Overall, the best GMAP result in the monolingual system was for a run using WSD, but the best MAP was obtained for a non-WSD run. Several other participants were able to obtain their best MAP and GMAP scores using WSD information. In the bilingual experiments, the best results in MAP and GMAP were for non-WSD runs, but several participants were able to profit from the WSD annotations.

In the monolingual experiments, cf. Table 6, the best results overall in both MAP and GMAP were for unine. Their WSD runs scored very similar to the non-WSD runs, with a slight decrease of MAP (0.16 percentage points) and a slight increase of GMAP (0.27 percentage points) [17]. The second best MAP scoring team attained MAP and GMAP improvements using WSD (from 38.34 MAP – 15.28 GMAP in their best non-WSD run to 39.57 MAP – 16.18 GMAP in their best WSD run) [31]. The third best scoring team in MAP achieved lower scores on both MAP and GMAP using WSD information [19]. The fourth best team obtained better MAP results using WSD information (from 38.10 to 38.99 MAP), but lower GMAP (from 15.72 to 15.52) [29]. Regarding the rest of participants, while ufrgs and uniba obtained improvements, know-center did not, and inaoe only submitted non-WSD runs. Two additional groups (IRn and sinai) sent their results late. Both groups had their best scores for non-WSD systems. You will find more details in the relevant papers in this volume.

In the bilingual experiments, cf. Table 7, the best results overall in both MAP and GMAP were for a system which did not use WSD annotations (36.39, compared to 21.77 MAP for their best run using WSD) [13]. The second scoring team also failed to profit from WSD annotations (30.36 compared to 9.70 MAP) [19].

The other two participating groups did obtain improvements, with ixa attaining 23.56 MAP with WSD (compared to 19.57 without) [29] and uniba attaining (7.23 MAP) [6].

All in all, the exercise showed that some teams did improve results using WSD annotations (up to approx. 1 MAP point in monolingual and approx. 4 MAP points in bilingual), providing the best GMAP results for the monolingual exercise, but the best results for the bilingual were for systems which did not use WSD (with a gap of approx. 13 MAP points). In any case, further case-by-case analysis of the actual systems and runs will be needed in order to get more insight about the contribution of WSD.

6 Conclusions

The Ad Hoc task in CLEF 2008 was almost completely renovated with new collections and new tasks. It focused on three different issues:

- real scenario: document retrieval from multilingual and sparse catalogue records to meet actual user needs (TEL@CLEF)
- linguistic resources: “exotic languages” to favour the creation of new experimental collections and the growth of regional IR communities (Persian@CLEF)
- advanced language processing: assessing whether word sense disambiguation can improve system performances (Robust WSD)

For all three tasks, we were very happy with the number of participants. However, overall, the results have been fairly inconclusive.

From the results of the TEL task, it would appear that there is no need for systems to apply any dedicated processing to handle the specificity of these collections (very sparse, essentially multilingual data) and that traditional IR and CLIR approaches can perform well with no extra boosting. However, we feel that it is too early to make such assumptions; many more experiments are needed.

The Persian task continued in the tradition of the CLEF Ad Hoc retrieval tasks on newspaper collections. The first results seem to confirm that the traditional IR/CLIR approaches port well to ”new” languages - where by ”new” we intend languages which have not been subjected to a lot of testing and experimental IR studies previously.

The robust exercise had, for the first time, the additional goal of measuring to what extent IR systems could profit from automatic word sense disambiguation information. The conclusions are mixed: while some top scoring groups did manage to improve the results using WSD information by approx. 1 MAP percentage point (approx. 4 MAP percentage points in the cross-language exercise) and the best monolingual GMAP score was for a WSD run (0.27 percentage points), the best scores for the rest came from systems which did not use WSD information. Given the relatively short time that the participants had to try effective ways of using the word sense information we think that these results are fairly positive.

However, in our opinion, a further evaluation exercise is needed for participants to further develop their systems.

All three tasks are being run again in CLEF 2009 both in order to provide participants with another chance to test their systems after refinement and tuning on the basis of the CLEF 2008 experiments and also to be able to create useful and consolidated test collections.

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, GESIS Leibniz-Institut for the Social Sciences, Bonn, 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.

The robust task was partially funded by the Ministry of Education (project KNOW TIN2006-15049) and the European Commission (project KYOTO ICT-2007-211423). We would like to acknowledge the invaluable suggestions and practical assistance of Arantxa Otegi, German Rigau and Piek Vossen. In particular, Arantxa was responsible for the XML coding of the word sense disambiguation results. We also want to thank Oier Lopez de Lacalle, who runs the UBC WSD system, and Yee Seng Chan, Hwee Tou Ng and Zhi Zhong, who run the NUS WSD system. Their generous contribution was invaluable to run this exercise.

References

1. Aghazade, Z., Dehghani, N., Farzinvash, L., Rahimi, R., AleAhmad, A., Amiri, H., Oroumchian, F.: Fusion of Retrieval Models at CLEF 2008 Ad-Hoc Persian Track. In: Peters, C., et al. (eds.) CLEF 2008. LNCS, vol. 5706, pp. 97–104. Springer, Heidelberg (2009)
2. Agirre, E., Lopez de Lacalle, O.: UBC-ALM: Combining k-NN with SVD for WSD. In: Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval 2007), Prague, Czech Republic, pp. 341–345 (2007)
3. Agirre, E., Magnini, B., Lopez de Lacalle, O., Otegi, A., Rigau, G., Vossen, P.: SemEval-2007 Task 01: Evaluating WSD on Cross-Language Information Retrieval. In: Peters, C., Jijkoun, V., Mandl, T., Müller, H., Oard, D.W., Peñas, A., Petras, V., Santos, D. (eds.) CLEF 2007. LNCS, vol. 5152, pp. 908–917. Springer, Heidelberg (2008)
4. Agirre, E., Di Nunzio, G.M., Ferro, N., Peters, C., Mandl, T.: CLEF 2008: Ad Hoc Track Overview. In: Borri, F., Nardi, A., Peters, C. (eds.) Working Notes for the CLEF 2009 Workshop (2009), http://www.clef-campaign.org/

5. AleAhmad, A., Kamalloo, E., Zareh, A., Rahgozar, M., Oroumchian, F.: Cross Language Experiments at Persian@CLEF 2008. In: Peters, C., et al. (eds.) CLEF 2008. LNCS, vol. 5706, pp. 105–112. Springer, Heidelberg (2009)
6. Caputo, A., Basile, P., Semeraro, G.: SENSE: SEmantic N-levels Search Engine at CLEF 2008 Ad Hoc Robust-WSD Track. In: Peters, C., et al. (eds.) CLEF 2008. LNCS, vol. 5706, pp. 126–133. Springer, Heidelberg (2009)
7. 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)
8. 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)
9. Chan, Y.S., Ng, H.T., Zhong, Z.: NUS-PT: Exploiting Parallel Texts for Word Sense Disambiguation in the English All-Words Tasks. In: Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval 2007), Prague, Czech Republic, pp. 253–256 (2007)
10. Cleverdon, C.: The Cranfield Tests on Index Language Devices. In: Sparck Jones, K., Willett, P. (eds.) Readings in Information Retrieval, pp. 47–59. Morgan Kaufmann Publisher, Inc., San Francisco (1997)
11. Clinchant, S., Renders, J.-M.: XRCE's Participation in CLEF 2008 Ad-Hoc Track. In: Borri, F., Nardi, A., Peters, C. (eds.) Working Notes for the CLEF 2008 Workshop, http://www.clef-campaign.org/
12. Clinchant, S., Renders, J.-M.: Multi-language Models and Meta-dictionary Adaptation for Accessing Multilingual Digital Libraries. In: Peters, C., et al. (eds.) CLEF 2008. LNCS, vol. 5706, pp. 83–88. Springer, Heidelberg (2009)
13. Costa Acosta, O., Geraldo, A.P., Orengo, V.M., Villavicencio, A.: UFRGS@CLEF 2008: Indexing Multiword Expressions for Information Retrieval. In: Borri, F., Nardi, A., Peters, C. (eds.) Working Notes for the CLEF 2008 Workshop (2008), http://www.clef-campaign.org/
14. Di Nunzio, G.M., Ferro, N.: Appendix A: Results of the TEL@CLEF Task. In: Borri, F., Nardi, A., Peters, C. (eds.) Working Notes for the CLEF 2008 Workshop (2008), http://www.clef-campaign.org/
15. Di Nunzio, G.M., Ferro, N.: Appendix B: Results of the Persian Task. In: Borri, F., Nardi, A., Peters, C. (eds.) Working Notes for the CLEF 2008 Workshop (2008), http://www.clef-campaign.org/
16. Di Nunzio, G.M., Ferro, N.: Appendix C: Results of the Robust Task. In: Borri, F., Nardi, A., Peters, C. (eds.) Working Notes for the CLEF 2008 Workshop (2008), http://www.clef-campaign.org/
17. Dolamic, L., Fautsch, C., Savoy, J.: UniNE at CLEF 2008: TEL, Persian and Robust IR. In: Borri, F., Nardi, A., Peters, C. (eds.) Working Notes for the CLEF 2008 Workshop (2008), http://www.clef-campaign.org/
18. Dolamic, L., Fautsch, C., Savoy, J.: UniNE at CLEF 2008: TEL and Persian IR. In: Peters, C., et al. (eds.) CLEF 2008. LNCS, vol. 5706, pp. 178–185. Springer, Heidelberg (2009)
19. Guyot, J., Falquet, G., Radhouani, S., Benzineb, K.: Analysis of Word Sense Disambiguation-Based Information Retrieval. In: Peters, C., et al. (eds.) CLEF 2008. LNCS, vol. 5706, pp. 146–154. Springer, Heidelberg (2009)
20. Jadidinejad, A.H., Mohtarami, M., Amiri, H.: Investigation on Application of Local Cluster Analysis and Part of Speech Tagging on Persian Text. In: Borri, F., Nardi, A., Peters, C. (eds.) Working Notes for the CLEF 2008 Workshop (2008), http://www.clef-campaign.org/

21. Karimpour, R., Ghorbani, A., Pishdad, A., Mohtarami, M., AleAhmad, A., Amiri, H., Oroumchian, F.: Improving Persian Information Retrieval Systems Using Stemming and Part of Speech Tagging. In: Peters, C., et al. (eds.) CLEF 2008. LNCS, vol. 5706, pp. 89–96. Springer, Heidelberg (2009)
22. Kuersten, J., Wilhelm, T., Eibl, M.: CLEF 2008 Ad-Hoc Track: Comparing and Combining Different IR Approaches. In: Peters, C., et al. (eds.) CLEF 2008. LNCS, vol. 5706, pp. 75–82. Springer, Heidelberg (2009)
23. Kuersten, J., Wilhelm, T., Eibl, M.: CLEF 2008 Ad-Hoc Track: On-line Processing Experiments with Xtrieval. In: Borri, F., Nardi, A., Peters, C. (eds.) Working Notes for the CLEF 2008 Workshop (2008), http://www.clef-campaign.org/
24. Larson, R.: Logistic Regression for Metadata: Cheshire takes on Adhoc-TEL. In: Peters, C., et al. (eds.) CLEF 2008. LNCS, vol. 5706, pp. 38–41. Springer, Heidelberg (2009)
25. Machado, J., Martins, B., Borbinha, J.: Experiments on a Multinomial Language Model versus Lucene's off-the-shelf Ranking Scheme and Rochio Query Expansion (TEL@CLEF Monolingual Task). In: Peters, C., et al. (eds.) CLEF 2008. LNCS, vol. 5706, pp. 50–57. Springer, Heidelberg (2009)
26. Martínez-Santiago, F., Perea-Ortega, J.M., García-Cumbreras, M.A.: Evaluating Word Sense Disambiguation Tools for an Information Retrieval task. In: Peters, C., et al. (eds.) CLEF 2008. LNCS, vol. 5706, pp. 113–117. Springer, Heidelberg (2009)
27. McNamee, P.: JHU Ad Hoc Experiments at CLEF 2008. In: Peters, C., et al. (eds.) CLEF 2008. LNCS, vol. 5706, pp. 170–177. Springer, Heidelberg (2009)
28. Navarro, S., Llopis, F., Muñoz, R.: IRn in the CLEF Robust WSD Task 2008. In: Peters, C., et al. (eds.) CLEF 2008. LNCS, vol. 5706, pp. 134–137. Springer, Heidelberg (2009)
29. Otegi, A., Agirre, E., Rigau, G.: IXA at CLEF 2008 Robust-WSD Task: using Word Sense Disambiguation for (Cross Lingual) Information Retrieval. In: Peters, C., et al. (eds.) CLEF 2008. LNCS, vol. 5706, pp. 118–125. Springer, Heidelberg (2009)
30. Paskin, N. (ed.): The DOI Handbook – Edition 4.4.1. International DOI Foundation (IDF) (2006), http://dx.doi.org/10.1000/186
31. Pérez-Agüera, J.R., Zaragoza, H.: Query Clauses and Term Independence. In: Peters, C., et al. (eds.) CLEF 2008. LNCS, vol. 5706, pp. 138–145. Springer, Heidelberg (2009)
32. Robertson, S.: On GMAP: and Other Transformations. In: Yu, P.S., Tsotras, V., Fox, E.A., Liu, C.B. (eds.) Proc. 15th International Conference on Information and Knowledge Management (CIKM 2006), pp. 78–83. ACM Press, New York (2006)
33. Tomlinson, S.: Sampling Precision to Depth 10000 at CLEF 2008. In: Peters, C., et al. (eds.) CLEF 2008. LNCS, vol. 5706, pp. 163–169. Springer, Heidelberg (2009)
34. Tomlinson, S.: German, French, English and Persian Retrieval Experiments at CLEF 2008. In: Borri, F., Nardi, A., Peters, C. (eds.) Working Notes for the CLEF 2008 Workshop (2008), http://www.clef-campaign.org/
35. Voorhees, E.M.: The TREC Robust Retrieval Track. SIGIR Forum 39, 11–20 (2005)