Project no. 215231

TrebleCLEF

Treble-CLEF: Evaluation, Best Practices and Collaboration for Multilingual Information Access
IST: ICT-1-4-1, Digital libraries and technology-enhanced learning

Deliverable 2.3.1
Analysis of Evaluation Campaign Results: CLEF 2008

Start Date of Project: 01 January 2008
Duration: 24 Months

Organisation Name of Lead Contractor for this Deliverable: UNIPD

Version 1.00, December 2008 [final]

Project co-funded by the European Commission within the Seventh Framework programme

Document Information

Deliverable number:	2.3.1
Deliverable title:	Analysis of Evaluation Campaign Results – CLEF 2008
Due date of deliverable:	31/12/2008
Actual date of deliverable:	31/12/2008
Author(s):	Maristella Agosti, Martin Braschler, Paul Claugh, Franco Crivellari, Giorgio Maria Di Nunzio, Nicola Ferro, Julio Gonzalo, Anselmo Peñas, Carol Peters, Mark Sanderson
Participant(s):	ISTI-CNR, UNED, UNIPD, USFD, ZHAW
Workpackage:	2
Workpackage title:	Evaluation Infrastructure
Workpackage leader:	UNIPD
Dissemination Level:	PU
Version:	1.00
Keywords:	Performance Evaluation, Multilingual Textual Document Retrieval, Cross-language Image Retrieval, Interactive Cross-Language Retrieval, Multiple Language Question Answering

History of Versions

Version	Date	Status	Author (Partner)	Description/Approval Level
0.10	05/11/2008	Draft	UNIPD	Outline circulated to all partners
0.11	07/11/2008	Draft	UNIPD	Outline updated according to partners' comments
0.20	19/12/2008	Draft	All	First draft with contributions from all partners
0.30	30/12/2008	Draft	All	Final draft with updated contributions from all partners
1.00	31/12/2008	Final	UNIPD	Final version updated according to partners' comments

Abstract

This document reports and analyses the results of the CLEF 2008 campaign and, specifically, deals with the outcomes of four main tracks of CLEF: the Ad hoc track, which deals with multilingual textual document retrieval; the ImageCLEF track, which concerns cross-language retrieval in image collections; the iCLEF track, which addresses interactive cross-language retrieval; and the QA@CLEF track, which covers multiple language question answering. Moreover, the document discusses these tracks and their achievements in the light of the main activities conducted during the first year of the TrebleCLEF project.

Table of Contents

Executive Summary

The Cross-Language Evaluation Forum (CLEF) has been running for almost a decade now; the tenth birthday will be celebrated at the CLEF 2009 workshop. When we launched this activity as a European initiative in early 2000, our declared objectives were the following: "to develop and maintain an infrastructure for the testing and evaluation of information retrieval systems operating on European languages, in both monolingual and cross-language contexts, and to create test-suites of reusable data that can be employed by system developers for benchmarking purposes" [42]. Although it is true to say that this basic idea remains at the core of our activity, over the years our range of interest and our interpretation of these initial objectives have both widened and deepened.

The intention of this document is to provide a panorama of the CLEF results so far and to show how we have created the foundations which now allow us to shift the focus of our activity, and enable us to concentrate not only on widening our coverage of the various building blocks involved in multilingual system development (tools, components, resources, lexicons) but also on the acquisition of a deeper understanding of the underlying issues. In particular, we focus our attention on the analysis and discussion of the results of four main tracks:

- *Multilingual Textual Document Retrieval (Ad Hoc)*: The Ad Hoc track is considered as our core track. 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. 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 CLIR area.
- *Cross-Language Retrieval in Image Collections (ImageCLEF)*: This track evaluated retrieval of images from multilingual collections; both text and visual retrieval techniques were exploitable with a major focus on the combination of text and image features to improve search. ImageCLEF has become the most popular of all tracks, even though (or maybe because) it is the track that deals the least with language and linguistic issues. One of the secrets of its popularity is that image search has a number of well-known applications.
- *Interactive Cross-Language Retrieval (iCLEF)*: In iCLEF, cross-language search capabilities are studied from a user-inclusive perspective. A central research question is how best to assist users when searching information written in unknown languages, rather than how best an algorithm can find information written in languages different from the query language.
- *Multilingual Question Answering (QA@CLEF)*: This track has been offering monolingual and cross-language question answering tasks since 2003. QA@CLEF 2008 proposed both main and pilot tasks. The main scenario was event-targeted QA on a heterogeneous document collection (news articles and Wikipedia). A large number of questions were topic-related, i.e. clusters of related questions possibly containing anaphoric references.

For each of these tracks, we present the experimental results, consider the strategies and approaches adopted by the participants, discuss the problems and the issues we have encountered, and outline our plans for the forthcoming CLEF 2009 campaign.

Finally, we reason about the results of the CLEF 2008 campaign from a different angle, relating them to some of the main activities organized and carried out during the first year of the TrebleCLEF project. TrebleCLEF not only aims at distilling the knowledge gathered during the CLEF campaigns and transferring it to relevant application, developer, and user communities but also strives to obtain input from those communities in order to improve the design of the campaigns and activate positive feedback cycles.

The overall outcome of CLEF 2008 is that of a successful activity which is able to involve a large research community whose achievements are advancing the MLIA/CLIR field and which both represents a huge repository of knowledge for the application and developer communities and is ready to address and satisfy the needs and requirements emerging from those communities.

Introduction

When we launched CLEF in 2000, our focus was on text and document retrieval. However, over the years our scope has gradually expanded to include different kinds of text retrieval across languages (i.e. not just document retrieval but question answering and geographic information retrieval as well) and different kinds of media (i.e. not just plain text but collections also containing images and speech). The goal has been not only to meet but also to anticipate the emerging needs of the R&D community and to encourage the development of next generation multilingual IR systems. This has meant that the number of tracks offered by CLEF has increased over the years, from just two in 2000 to nine separate tracks in 2008. Each track is run by a coordinating group with specific expertise in the area covered by the track. Most tracks offer several different tasks and these tasks normally vary each year, according to the interests of the track coordinators and participants. Figure 1 shows when tracks have been introduced and when they have been terminated.

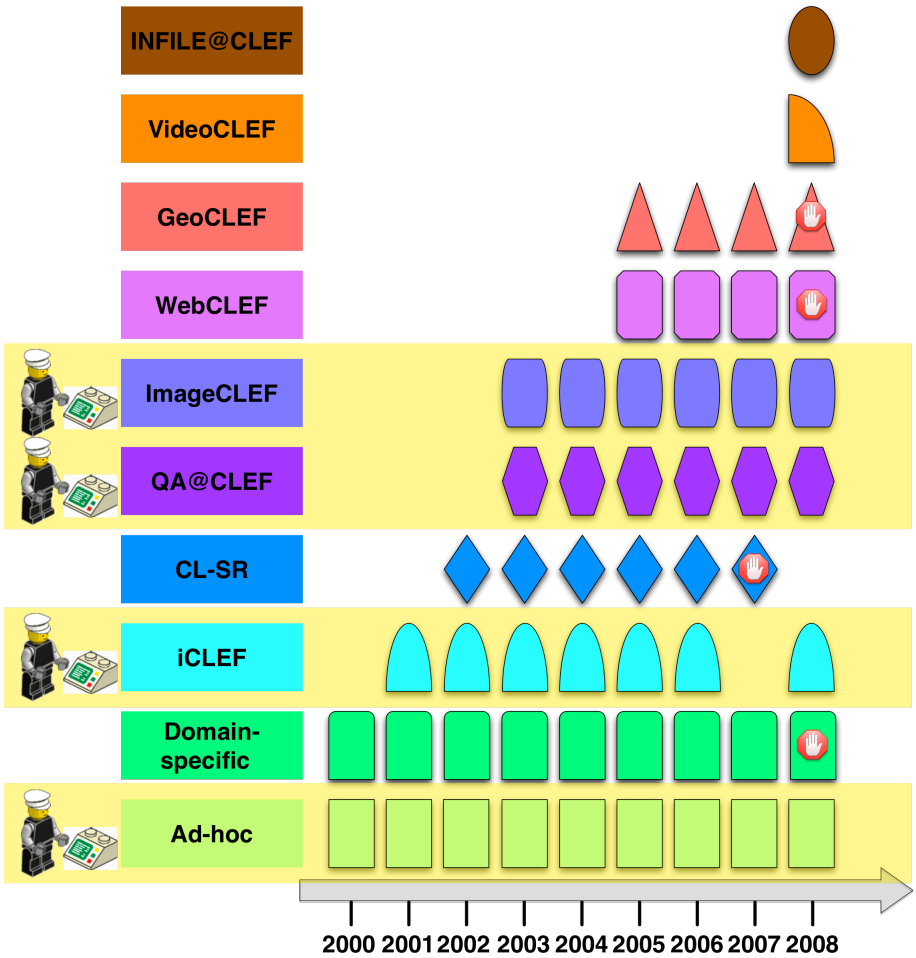

Figure 1: CLEF 2000-2008 tracks with tracks under analysis in this report highlighted.

The organisation of the CLEF 2008 campaign [18] has already been described in Deliverable 2.1.1. In this first section we briefly summarise the main details necessary to provide background information for the rest of this report. Indeed, this deliverable aims at providing an in-depth analysis of the results for a selection of these tracks: Ad Hoc; ImageCLEF; iCLEF and Question Answering, as shown in Figure 1.

These tracks have been chosen mainly for two reasons. The first is pragmatic: members of TrebleCLEF have been responsible for their coordination and for the analysis of the results.. The

second, and more relevant, is their strategic importance in the general multilingual information access paradigm and within the context of TrebleCLEF. A major objective of TrebleCLEF is to study more closely the needs of the real world in order promote the development of system functionality designed to meet these needs. In different ways, as will be explained below, three of the tracks (Ad Hoc, iCLEF and ImageCLEF) offered tasks that directly involved user or application communities. All four tracks strongly encourage interaction and collaboration between different scientific communities, e.g. information retrieval, image processing, natural language processing, human computer interaction, and user groups, such as librarians, cultural heritage professionals, radiographers.

The document is organized as follows: the following section briefly introduces the tracks under examination in this report; then Sections 2 to 5 provide a detailed analysis of each of them; Section 6 places the CLEF 2008 campaign in the context of the other TrebleCLEF activities; finally, Section 7, wraps up the discussion and provides some concluding remarks.

1.1 Multilingual Textual Document Retrieval (Ad Hoc)

The Ad Hoc track is considered as our core track. 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 to 2007, the track exclusively used collections of European newspaper and news agency documents and worked hard at offering increasingly complex and diverse tasks, adding new languages each year. Table 1 shows that in the first eight years of the Ad Hoc track, monolingual and bilingual tasks were offered for target collections in twelve different European languages, with bilingual tasks often being proposed for unusual pairs of languages, such as Finish to German, or French to Dutch. In addition multilingual tasks were offered with varying numbers of languages in the target collections.

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 CLIR area. It has provided the resources, the test collections and also the forum for discussion and comparison of ideas and results. 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 on fine-tuning for individual languages while other efforts have concentrated on developing language-independent strategies.

There is also substantial proof of significant increase in retrieval effectiveness in multilingual settings by the systems of CLEF participants. [7] provides a comparison between effectiveness scores from the 1997 TREC-6 campaign and the CLEF 2003 campaign in which retrieval tasks were offered for eight European languages. While in 1997 systems were performing at about 50%–60% of monolingual effectiveness for multilingual settings, that figure had risen to 80%–85% by 2003 for languages that had been part of multiple evaluation campaigns. In the recent campaigns, we commonly see a figure of about 85%–90% for most languages.

In 2008 there was a big change in focus in this track: 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.

The first task was an application-oriented task, offering monolingual and cross-language search on library catalog records. It was organized in collaboration with The European Library (TEL)[1] and used three collections from the catalogs of the British Library, the Bibliothèque 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 and semi-structured multilingual data.

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

	Monolingual	Bilingual	Multilingual
CLEF 2000	de;fr;it	x→en	x→de;en;fr;it
CLEF 2001	de;es;fr;it;nl	x→en x→nl	x→de;en;es;fr;it
CLEF 2002	de;es;fi;fr;it;nl;sv	x→de;es;fi;fr;it;nl;sv x→en (newcomers only)	x→de;en;es;fr;it
CLEF 2003	de;es;fi;fr;it;nl;ru;sv	it→es de→it fr→nl fi→de x→ru x→en (newcomers only)	x→de;en;es;fr x→de;en;es;fi;fr;it;nl;sv
CLEF 2004	fi;fr;ru;pt	es;fr;it;ru→fi de;fi;nl;sv→fr x→ru x→en (newcomers only)	x→fi;fr;ru;pt
CLEF 2005	bg;fr;hu;pt	x→bg;fr;hu;pt	Multi8 2yrson (as in CLEF 2003) Multi8 Merge (as in CLEF 2003)
CLEF 2006	bg;fr;hu;pt Robust de;en;es;fr;it;nl	x→bg;fr;hu;pt am;hi;id;te;or→en Robust it→es fr→nl en→de	 Robust x→de;en;es;fr;it;nl
CLEF 2007	bg, cz, hu Robust en;fr;pt	x→bg;cz;hu am;id;or;zh→en bn;hi;mr;ta;te→en Robust x→en;fr;pt	
CLEF 2008	fa TEL de;en;fr Robust WSD en	en→fa TEL x→de;en;fr Robust WSD es→en	

Table 1: Ad Hoc 2000–2008 Tasks. The following ISO 639-1 language codes have been used: am=Amharic; bg=Bulgarian; bn=Bengali; de=German; en=English; es=Spanish; fa=Farsi; fi=Finnish; fr=French; hi=Hindi; hu=Hungarian; id=Indonesian; it=Italian; mr=Marathi; nl=Dutch; or=Oromo; pt=Portuguese; ru=Russian; sv=Swedish; ta=Tamil; te=Telugu.

had been part of multiple evaluation campaigns. In the recent campaigns, we commonly see a figure of

The other two tasks were more classical CLEF research-oriented activities. The Persian@CLEF activity was coordinated in collaboration with the Database Research Group (DBRG) of Tehran University. It was the first time that CLEF offered a non-European language target collection. We chose Persian 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. Monolingual and cross-language (English to Persian) tasks were offered.

The robust task ran for the third time at CLEF 2008. This year it used English test data from previous campaigns but, in addition to the original documents and topics, the organizers provided word sense

disambiguated (WSD) documents and topics. Both monolingual and bilingual experiments (topics in Spanish) were activated.

1.2 Cross-Language Retrieval in Image Collections (ImageCLEF)

ImageCLEF has become the most popular of all tracks, even though (or maybe because) it is the track that deals the least with language and linguistic issues. One of the secrets of its popularity is that image search has a number of well known applications. In CLEF we focus on one of the most important: medical image processing and analysis.

Five challenging tasks were offered in 2008:

- A photo retrieval task: a good image search engine ensures that duplicate or near duplicate documents retrieved in response to a query are hidden from the user. Ideally the top results of a ranked list will contain diverse items representing different sub-topics within the results. This task focused on the study of successful clustering to provide diversity in the top-ranked results. The target collection contained images with captions in English and German; queries were in English.
- A medical image retrieval task: this is a domain-specific retrieval task in a domain where many ontologies exist; the target collection was a subset of the Goldminer collection containing images from English articles published in Radiology and Radiographics with captions and html links to the full text articles. Queries were provided in English, French and German.
- A visual concept deception task: the objective was to identify language-independent visual concepts that would help in solving the photo retrieval task. A training database was released with approximately 1,800 images classified according to a concept hierarchy. This data was used to train concept detection/annotation techniques. For each of the 1,000 images in the test database, participating groups were required to determine the presence/absence of the concepts.
- An automatic medical image annotation task: automatic image annotation or image classification can be an important step when searching for images from a database of radiographs. The aim of the task was to find out how well current language-independent techniques can identify image modality, body orientation, body region, and biological system on the basis of the visual information provided by the images.
- A Wikipedia image retrieval task: this was an ad hoc image search task where the information structure can be exploited for retrieval. The aim was to investigate retrieval approaches in the context of a larger scale and heterogeneous collection of images (similar to those encountered on the Web) that are searched for by users with diverse information needs.

A major focus in all tasks is on the combination of text and image features to improve search. In this respect, it is interesting to note that one of the findings of the coordinators of the medical tasks this year was that from an examination of mixed media runs that had corresponding text-only runs, it was clear that combining good textual retrieval techniques with questionable visual retrieval techniques can negatively affect system performance. This demonstrates the difficulty of usefully integrating both textual and visual information, and the fragility that such combinations can introduce into retrieval systems.

1.3 Interactive Cross-Language Retrieval (iCLEF)

In iCLEF, cross-language search capabilities are studied from a user-inclusive perspective. iCLEF 2008 dealt with Flickr[2], a large-scale, online image database based on a large social network of WWW users, with the potential for offering both challenging and realistic multilingual search tasks for interactive experiments. The search interface provided by the iCLEF organizers was a basic cross-language retrieval system for the Flickr image database presented as an online game: the user is given an image, and must find it again without any a priori knowledge of the language(s) in which the image

[2] See `http://www.flickr.com/`

is annotated. The game was publicized on the CLEF mailing list and prizes were offered for the best results in order to encourage participation.

The main novelty of the iCLEF 2008 experiments was the shared analysis of a search log from a single search interface provided by the organizers (i.e. the focus was on log analysis, rather than on system design). Search logs were harvested from the search interface described above and iCLEF participants could essentially do two things:

- *Search log analysis*: participants had access to the search logs, and could freely perform data mining studies on them, such as looking for differences in search behaviour according to language skills, or looking for correlations between search success and search strategies, etc.
- *Interactive experiments*: participants could recruit their own users and conduct their own experiments with the interface. For instance, they could recruit a set of users with passive abilities and another with active abilities in certain languages and, besides studying the search logs, could perform observational studies on how they search, conduct interviews, etc.

1.4 Multilingual Question Answering (QA@CLEF)

This track has been offering monolingual and cross-language question answering tasks since 2003. QA@CLEF 2008 proposed both main and pilot tasks. The main scenario was event-targeted QA on a heterogeneous document collection (news articles and Wikipedia). A large number of questions were topic-related, i.e. clusters of related questions possibly containing anaphoric references. Besides the usual news collections, articles from Wikipedia were also considered as sources of answers. Many monolingual and cross-language sub-tasks were offered: Basque, Bulgarian, Dutch, English, French, German, Italian, Portuguese, Romanian and Spanish were proposed as both query and target languages; not all were used in the end.

After 6 years, a lot of resources and know-how have been accumulated. However, the tasks offered have proved to be difficult for the systems which have not shown a very good overall performance, even those that have participated year by year. In addition, a result of offering so many language possibilities has meant that there have always been very few systems participating in the same task, with the same languages. This has meant that comparative analysis is extremely problematic.

QA@CLEF also promoted a number of additional exercises:

- The Answer Validation Exercise (AVE) in its third edition was aimed at evaluating answer validation systems based on recognizing textual entailment.
- QAST was focused on Question Answering over Speech Transcriptions of seminars. In this 2nd year pilot task, answers to factual and definitional questions in English were to be extracted from spontaneous speech transcriptions related to separate scenarios in English, French and Spanish. For the tasks where the word error rate was low enough (around 10%) the loss in accuracy compared to manual transcriptions was under 5%, suggesting that QA in such documents is potentially feasible. However, even where automatic speech recognition (ASR) performance is reasonably good, there remain many challenges when dealing with spoken language. The results from the QAST evaluation indicate that if a QA system performs well on manual transcriptions it will also performs reasonably well on high quality automatic transcriptions.
- QA-WSD provided questions and collections with already disambiguated Word Senses in order to study their contribution to QA performance. Unfortunately, just one groups participated in this task so the results were not significant.

2 The Ad-hoc Track

As anticipated in section 1.1, the CLEF 2008 Ad Hoc track focused on three different issues:

- *real scenario*: document retrieval from multilingual and sparse catalogue records to meet actual user needs, described in section 2.1;

- *linguistic resources*: "exotic languages" to favour the creation of new experimental collections and the growth of regional IR communities, described in section 2.2;
- *advanced language processing*: robust and WSD to strengthen system performances, briefly reported in section 2.3.

A detailed description of the Ad hoc track and all the experimental results can be found in [2], [14], [15], and [16].

2.1 TEL Task

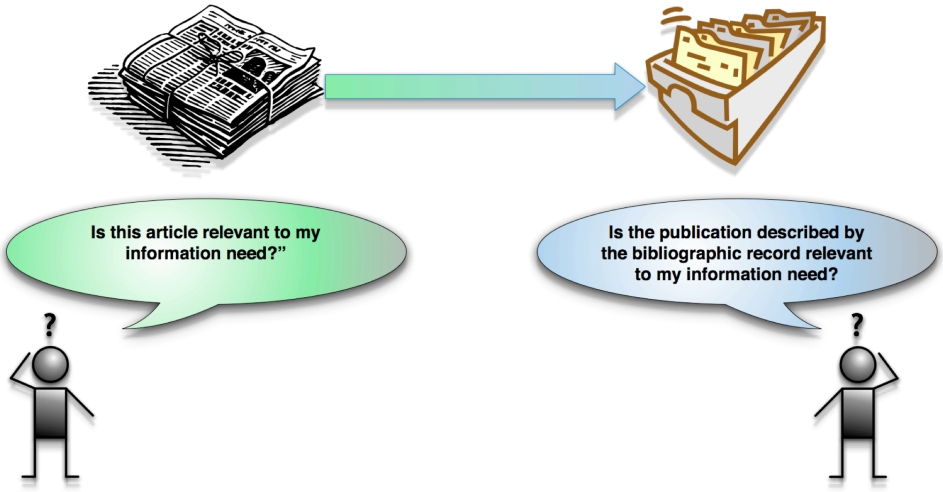

Figure 2: Shift from searching for documents of interest to searching for surrogates that represent documents of interest.

As shown in Figure 2, whereas in the traditional ad hoc task, the user searches directly for a document containing information of interest, in the TEL task 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?".

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.

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.

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).

2.1.1 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, as shown in Figure 3.

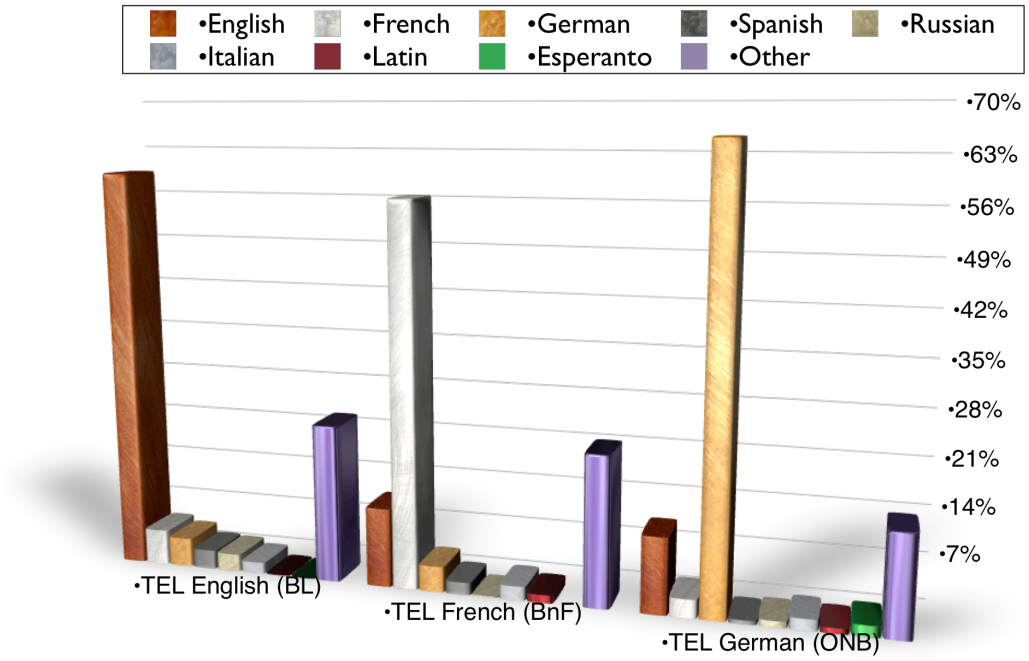

Figure 3: Distribution of the languages in the TEL collections.

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. Figure 4 shows the distribution of the content bearing elements in the TEL collections. Since the same element may be repeated more than once in a document, percentages greater than 100% indicate this fact; for example, 200% for the subject element in the English collections indicates that English records usually have two subject headings describing them.

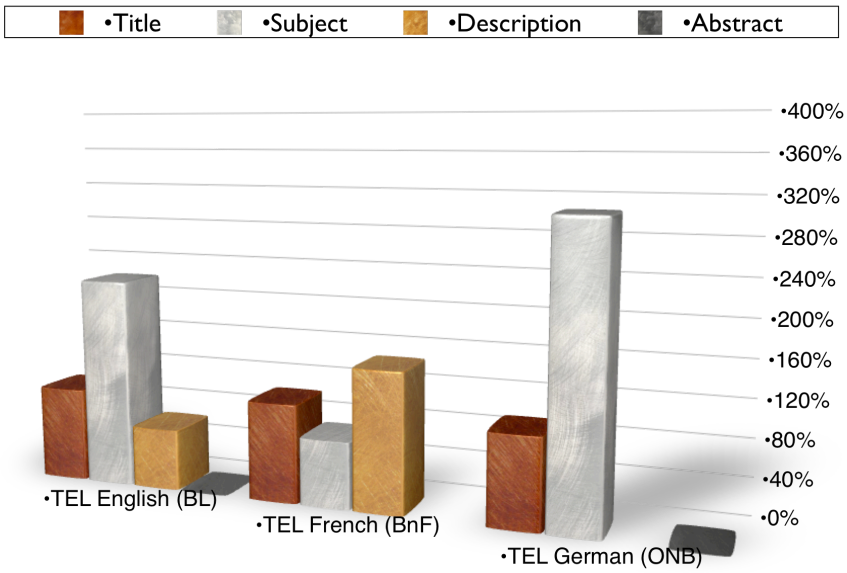

Figure 4: Distribution of the content in the TEL collections.

2.1.2 Topics

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.

```
<?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>
```

Figure 5: Example of TEL topic in all five languages: topic 10.2452/451-AH.

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 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 in relevant documents. For each topic, the results expected from the separate collections can vary considerably, e.g. in the case of the TEL task, a topic of particular interest to Britain, such as the example given in Figure 5, can be expected to find far more relevant documents in the BL collection than in BNF or ONB.

2.1.3 Relevance Assessments

Table 2 reports summary information on the TEL 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.

TEL English Pool (DOI 10.2454/AH-TEL-ENGLISH-CLEF2008)	
Pool size	28,104 pooled documents – 25,571 not relevant documents – 2,533 relevant documents 50 topics
Pooled Experiments	21 out of 61 submitted experiments – monolingual: 13 out of 37 submitted experiments – bilingual: 8 out of 24 submitted experiments
Assessors	3 assessors
TEL French Pool (DOI 10.2454/AH-TEL-FRENCH-CLEF2008)	
Pool size	24,530 pooled documents – 23,191 not relevant documents – 1,339 relevant documents 50 topics
Pooled Experiments	14 out of 45 submitted experiments – monolingual: 9 out of 29 submitted experiments – bilingual: 5 out of 16 submitted experiments
Assessors	3 assessors
TEL German Pool (DOI 10.2454/AH-TEL-GERMAN-CLEF2008)	
Pool size	28,734 pooled documents – 27,097 not relevant documents – 1,637 relevant documents 50 topics
Pooled Experiments	16 out of 47 submitted experiments – monolingual: 10 out of 30 submitted experiments – bilingual: 6 out of 17 submitted experiments
Assessors	4 assessors

Table 2: Summary information about TEL pools.

The box plot of Figure 6 compares the distributions of the relevant documents across the topics of each pool for the different TEL pools; the boxes are ordered by decreasing mean number of relevant documents per topic.

As can be noted, TEL English, French and German distributions appear similar and are asymmetric towards topics with a greater number of relevant documents. Both the English and French 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.

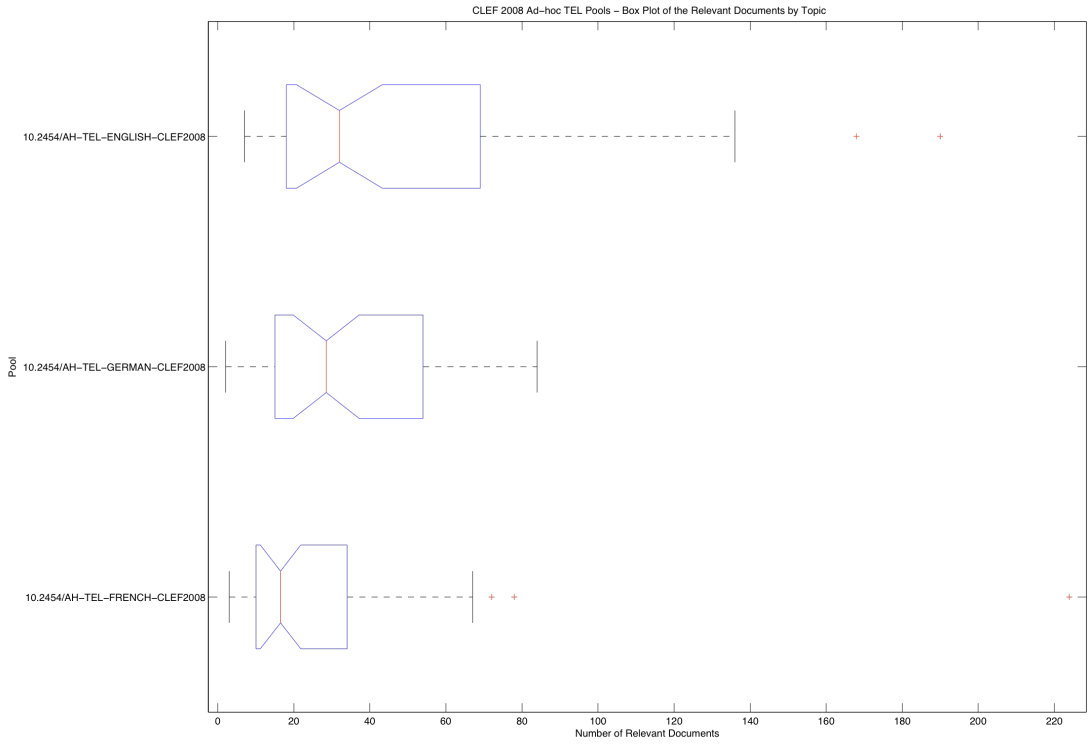

Figure 6: Distribution of the relevant documents in the TEL pools.

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 as we expected this language to be used only for ES to EN runs), 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.

2.1.4 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.

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 3: Best entries for the monolingual TEL tasks.

Figure 7 compares the performances of the top participants of the TEL Monolingual tasks.

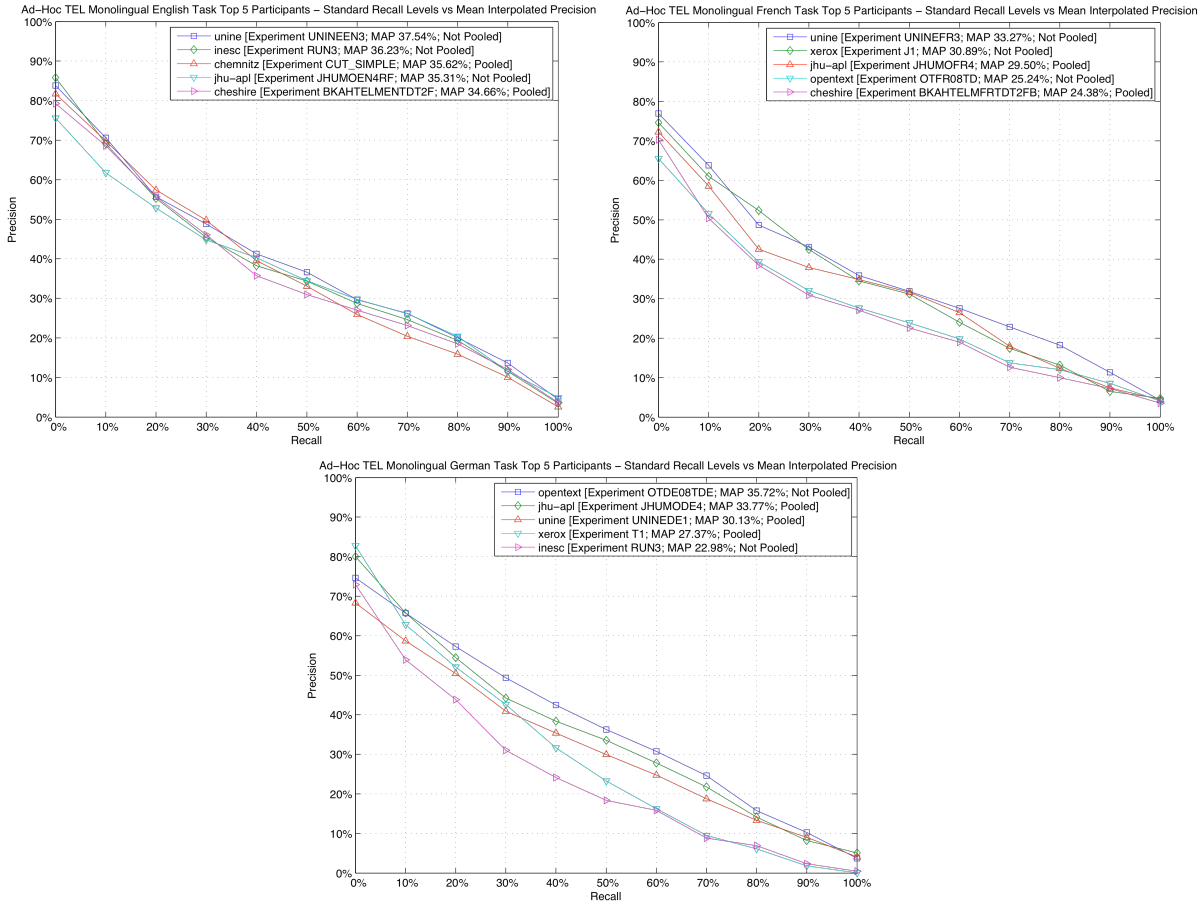

Figure 7: Comparison of the performances of the top participants to the monolingual TEL tasks.

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%

Table 4: Best entries for the bilingual TEL tasks.

Figure 8 compares the performances of the top participants of the TEL Bilingual tasks.

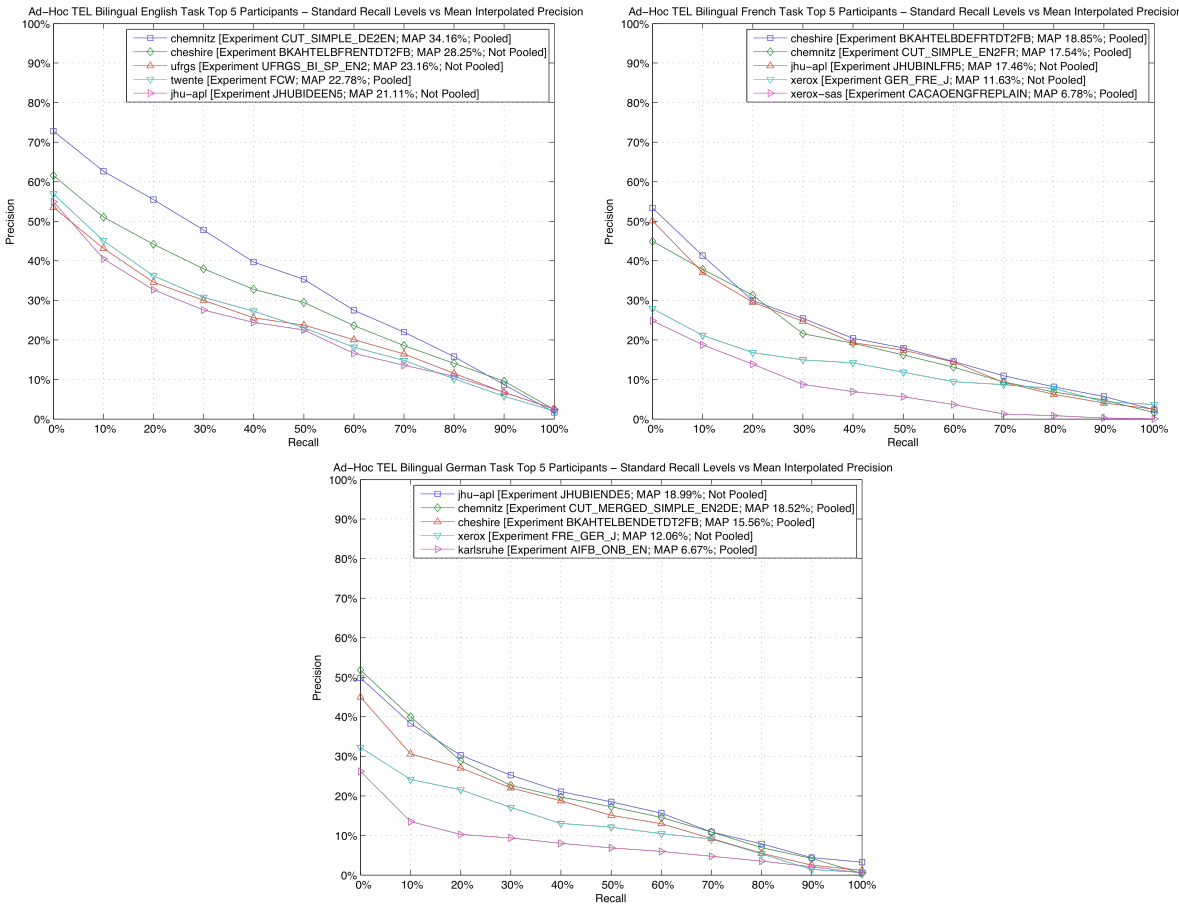

Figure 8: Comparison of the performances of the top participants to the bilingual TEL tasks.

2.1.5 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.

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 good 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. We have no explanation for this at the present.

2.1.6 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 and Wikipedia. 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.

An example of this is the group from the Technical University of Chemnitz, which 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. 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[3]. 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 [29].

Another group that attempted to tackle the multinguality of the target collections was Xerox. 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 [10]. 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.

[3] 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.

However, most groups appear to have 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 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 [17], [31], and [36].

As was mentioned in Section 2.1, 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 [35].

To sum up, from a preliminary scanning of the results of this task, it appears that the majority of groups took it as a traditional ad hoc retrieval task and applied traditional methods. However, the data acquired in this year's campaign is insufficient to be able to confirm whether this is really the best approach to retrieval on library catalog cards. We will thus be offering this task again in 2009 making it obligatory for groups to provide runs in which they have applied only traditional IR retrieval methods and also runs in which they have attempted to take the specificity of the data into account. The aim is to motivate the research community to deal as much as possible with the real settings of the (digital) library scenario and to find out innovative means to address its shortcomings. Moreover, running the task again in 2009 will give us the opportunity of developing a set of 50 additional topics, which is needed to conduct different kinds of analyses and obtain stable results.

2.2 Persian Task

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

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.

2.2.1 Documents

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[4].

2.2.2 Topics

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 9, where the rather awkward English rendering evidences the uncertainty of the translator.

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

```
<?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">                                  76  سندهایی را پیدا کن که شامل اطلاعاتی در مورد انتخابات دوم خرداد ماه سال
          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 اطلاعاتی در مورد نامزدها و گفته هاي آنها، پيروزي
          است  76  غيرمنتظره خاتمي در انتخابات رياست جمهوري در دوم خرداد ماه سال                  </narrative>
</topic>
```

Figure 9: Example of Persian topic: topic 10.2452/599-AH.

2.2.3 Relevance Assessments

Table 5 reports summary information on the Persian pool used to calculate the results for the main monolingual and bilingual experiments.

Persian Pool (DOI 10.2454/AH-PERSIAN-CLEF2008)	
Pool size	26,814 pooled documents – 21,653 not relevant documents – 5,161 relevant documents 50 topics
Pooled Experiments	66 out of 66 submitted experiments – monolingual: 53 out of 53 submitted experiments – bilingual: 13 out of 13 submitted experiments
Assessors	22 assessors

Table 5: Summary information about the Persian pool.

As shown in the box plot of Figure 10, the Persian distribution presents a greater number of relevant documents per topic with respect to the distributions of the TEL pools and is more symmetric between topics with lesser or greater number of relevant documents. This greater symmetry in distribution of relevant documents is probably due to the fact that the topic set was created just on the basis of the contents of the Persian collection, rather than needing to reflect the contents of multiple collections. In addition, as can be seen from Table 5, 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 (however this is not so - see Section 4 below) 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.

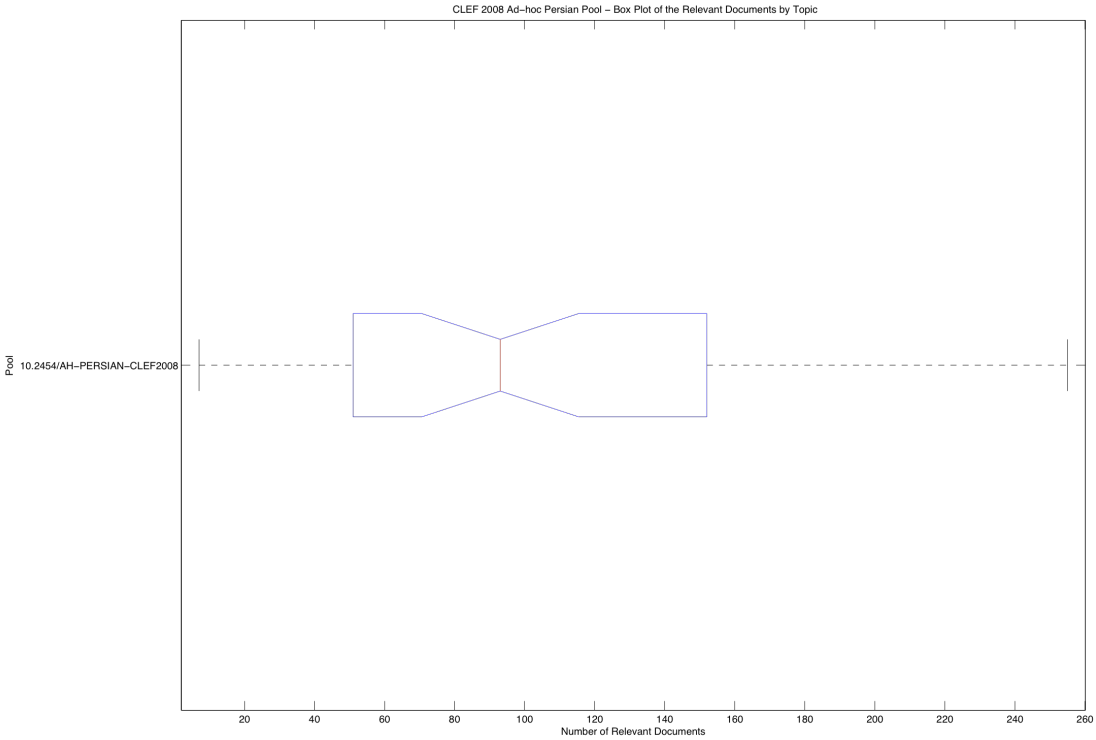

Figure 10: Distribution of the relevant documents in the Persian pool.

2.2.4 Monolingual 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.

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%

Table 6: Best entries for the monolingual Persian task.

Figure 11 compares the performances of the top participants of the Persian task.

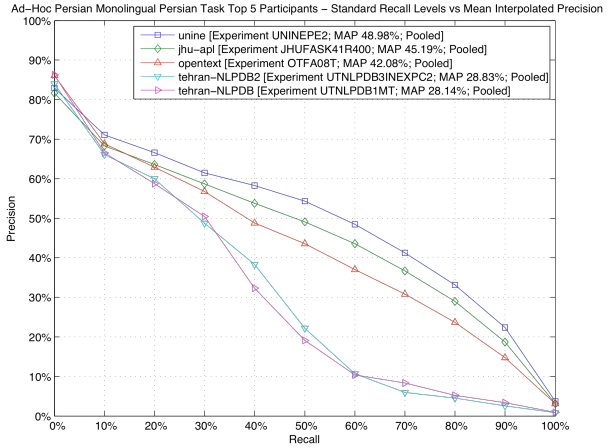

Figure 11: Comparison of the performances of the top participants to the monolingual Persian task.

2.2.5 Bilingual Results

Table 7 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.

Track	Rank	Participant	Experiment DOI	MAP
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%

Table 7: Best entries for the bilingual Persian task.

Figure 12 compares the performances of the top participants of the Persian tasks.

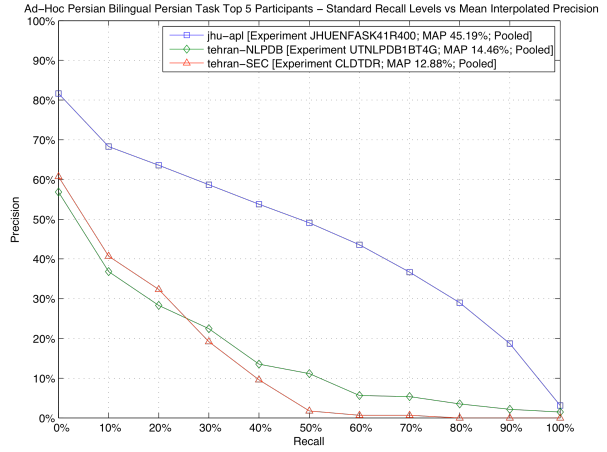

Figure 12: Comparison of the performances of the top participants to the bilingual Persian task.

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

- EN → 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.

2.2.6 Approaches and Discussion

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 results were 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 [17]. The group 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. They found that that character 5-grams and skipgrams performed the best [36]. The findings of [17] were confirmed by [44]. This group also tested runs with no stemming, with the UniNE stemmer and with n-grams. Similarly, they reported that stemming had relatively little impact.

Somewhat surprisingly, most of the papers from the Iran-based groups do not provide much information wrt 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 [4]. Only one of these groups provides some detailed discussion of the impact of stemming. This group used a simple stemmer, PERSTEM[5], 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. 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 [27] and [28]. 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: 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 best 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[6] [36].

The other two submissions for the cross-language task were from Iran-based groups. We have received a report from just one of them. 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.

In general, the Iran-based groups retrieved only 100 documents for each topic in their runs instead of the usual 1000 ones. This may have negatively impacted their performances, e.g. in terms of MAP, with respect to other groups which have retrieved 1000 documents and found relevant ones after rank 100. This issue has been clearly pointed out during the CLEF workshop and it will be addressed in the CLEF 2009 campaign.

[5] See http://sourceforge.net/projects/perstem
[6] See http://www.parstranslator.net/eng/translate.htm

Indeed, we will be offering this task again in 2009 in order to gain a better understanding of the best strategies needed to deal with the Persian morphology; to further investigate how cultural differences impact retrieval performances, e.g. by providing dates according to the Shamsi calendar in the Farsi topics and mapping them to the Gregorian calendar in their English translation; to develop an additional set of 50 topics for having enough topics to conduct various kinds of analyses; and, finally, to give Iran-based groups the opportunity to deal with the "100 documents" issue and provide runs really comparable with the ones of the other participants.

2.3 Robust WSD

The robust task ran for the third time at CLEF 2008. This year it used English test data from previouscampaigns but, in addition to the original documents and topics, the organizers provided word sense disambiguated (WSD) documents and topics. Both monolingual and bilingual experiments (topics in Spanish) were activated. Full details about this task can be found in [2].

The results for the eight participating groups were mixed: while some top scoring groups did manage to improve the results using WSD information in both monolingual and bilingual settings, and the best monolingual robustness (GMAP) score was for a WSD run, 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 can be considered positive; a subsequent evaluation exercise would be needed for participants to further develop their systems. Therefore, this task will be offered again in 2009.

2.4 Comparing Bilingual to Monolingual Results

The experimental evaluation carried out on MLIA systems has a double value: on the one hand, it should provide guidelines, hints, and directions to drive the design and development of the next generation MLIA systems; on the other hand, the experimental results should be easily communicated to other research communities and effective tools for interpreting and comparing the experimental result should be made available to those research communities.

The CLEF initiative has had, among the other things, the important function of spreading and circulating results to different research communities. However, there are still some common practices in the presentation of MLIA system performances which can be improved from both the point of view of the mathematical analysis and the graphical impact showing the advantages and the drawbacks of an approach compared to another.

To give an example, a common method used to evaluate performances for bilingual retrieval evaluation is to compare results against monolingual baselines. Mean Average Precision (MAP) is often used as a summary indicator. In the case of CLEF 2008, for instance, for the TEL English task we have that the best bilingual English system (X → EN) performed 90.99% compared to the best monolingual English IR system and similar comparisons are carried out for the other tasks. These measures are easy to understand, and for this reason important as they convey information on performances quickly. Nevertheless, at the same time, they tend to hide a lot of information which could enrich this simple number and give a more complete picture of the comparison between bilingual and monolingual systems.

In [12] and [13], a comparison methodology has been proposed which is based on a comparison of results on single topics in both the monolingual and the bilingual tasks instead of a comparison of single measures (like the one presented previously),. Here, we focus our attention on the graphical tools since they provide an intuitive means to compare the performances of MLIA systems, to gain a visual explanation of the behavior of different systems, and to communicate them to other communities, such as the digital library one, which may not have all the expertise needed to deal with complex statistical analyses or with the details entailed by the different performance measures. The goal is to give these communities the means to easily assess MLIA systems and understand how to best fit them in their applicative context.

We proceed as follows:

- given a set of experiments e_1, …, e_n and a set of topics t_1, …, t_m, the average precision of the experiment e_1 for the topic t_1 is indicated with $AP_{e1,t1}$
- for each topic t_i, we compute the mean of the average precisions $AP_{e1,ti}$, …, $AP_{en,ti}$ across the experiments e_1, …, e_n and we indicate this mean with MAP_{ti}[7] ;
- we increasingly order, on the x-axis, the topics by monolingual MAP_t and we plot, on the y-axis, the monolingual MAP_t performances with a red circle and the corresponding bilingual MAP_t performances with a blue diamond;
- finally, the least squares fitting lines are drawn for the two tasks, solid for monolingual and dashed for bilingual. In this way we are able to extrapolate and compare the trend of the monolingual and bilingual tasks.

The general claim that bilingual is X% of the monolingual suggests that , the two fitting lines would have roughly the same slope and the bilingual line would be right shifted with respect to the monolingual one due to the loss in performances for crossing the language barriers. As we will see in the following sections, this is not always the case and other types of behaviour can emerge.

2.4.1 Ad Hoc TEL: Monolingual vs Bilingual German

Figure 13 clearly shows that the bilingual line is below the monolingual line, which indicates that the bilingual performances are lower than the monolingual ones, and the slopes of the two lines are slightly different, which indicates that monolingual gains more in the "easy" topics, i.e. those topics in which, on average, the performances are higher.

It is also possible analyze the sets of bilingual experiments (English to German, French to German, Spanish to German) singularly. However, there would be no difference in the outcome: monolingual is by far better than bilingual no matter what source language is used.

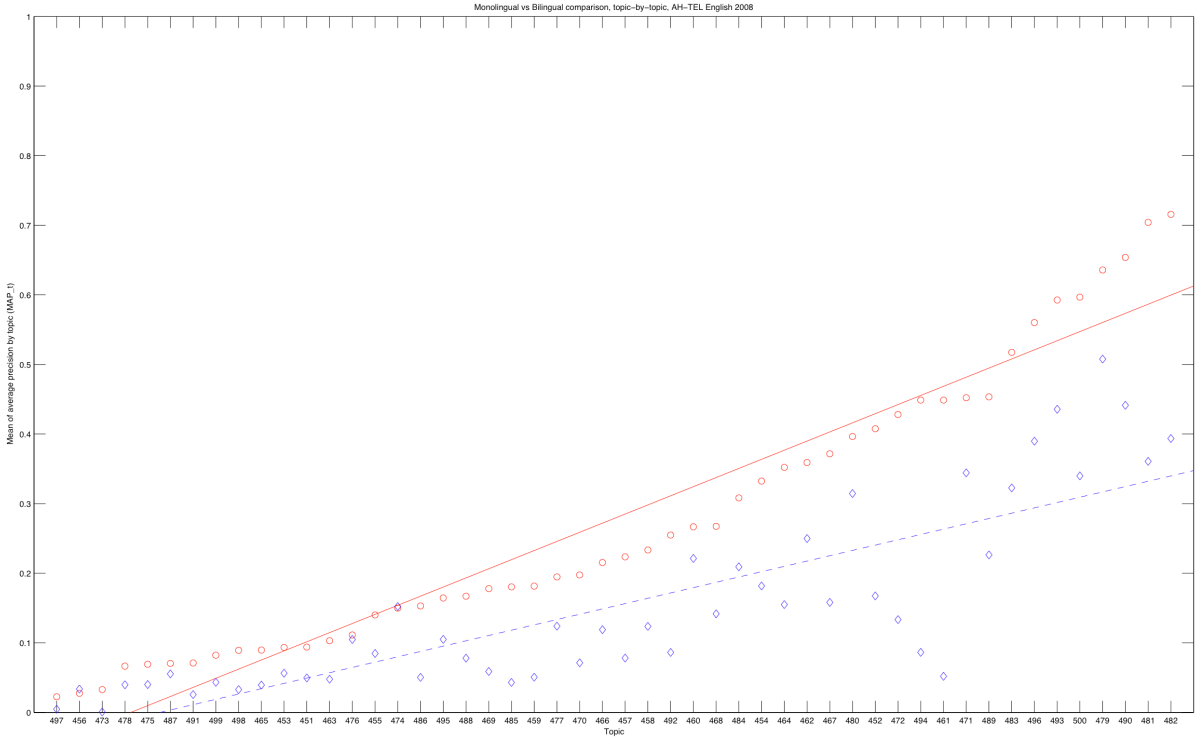

Figure 13: German TEL tasks: comparison of monolingual (red circle, solid line) with respect to bilingual (blue diamond, dashed line) performances and linear fit.

[7] Note that MAP_{ti} differs from which is usually called MAP, since MAP is, given an experiment e_j, the mean of the average precisions $AP_{ej,t1}$, …, $AP_{ej,tm}$ across the topics t_1, …, t_m.

2.4.2 Ad Hoc TEL: Monolingual vs Bilingual French

The linear fit shows a different trend of the interpolating line, as reported in Figure 14. There is a point where the lines cross, and this point is close to the left part of the picture where the topics with the lowest monolingual MAP are concentrated. This particular case highlights one important thing: performances of the bilingual experiments are, on average, higher for the topics where the monolingual experiments fail (or perform worse). There are four topics (reading from left to right 10.2452/468-AH, 10.2452/455-AH, 10.2452/482-AH, and 10.2452/484-AH) which show a significant positive difference in favor of the bilingual experiments. This means that, even if in general bilingual performances are lower, there are cases where the translation process helps the retrieval of documents, and these cases are all concentrated in those topics which obtain lower monolingual performances. Also in this case, the analysis on the single sets of experiments with the same source language shows no difference in the outcome.

Interestingly enough, even if the general claims for German (bilingual is about 53% of monolingual) and French (bilingual is about 56% of monolingual) are quite similar, from Figure 13 and Figure 14 it turns out that they actually behave in different ways since in French bilingual tends to perform better than monolingual for the difficult topics while in German monolingual constantly performs better than bilingual.

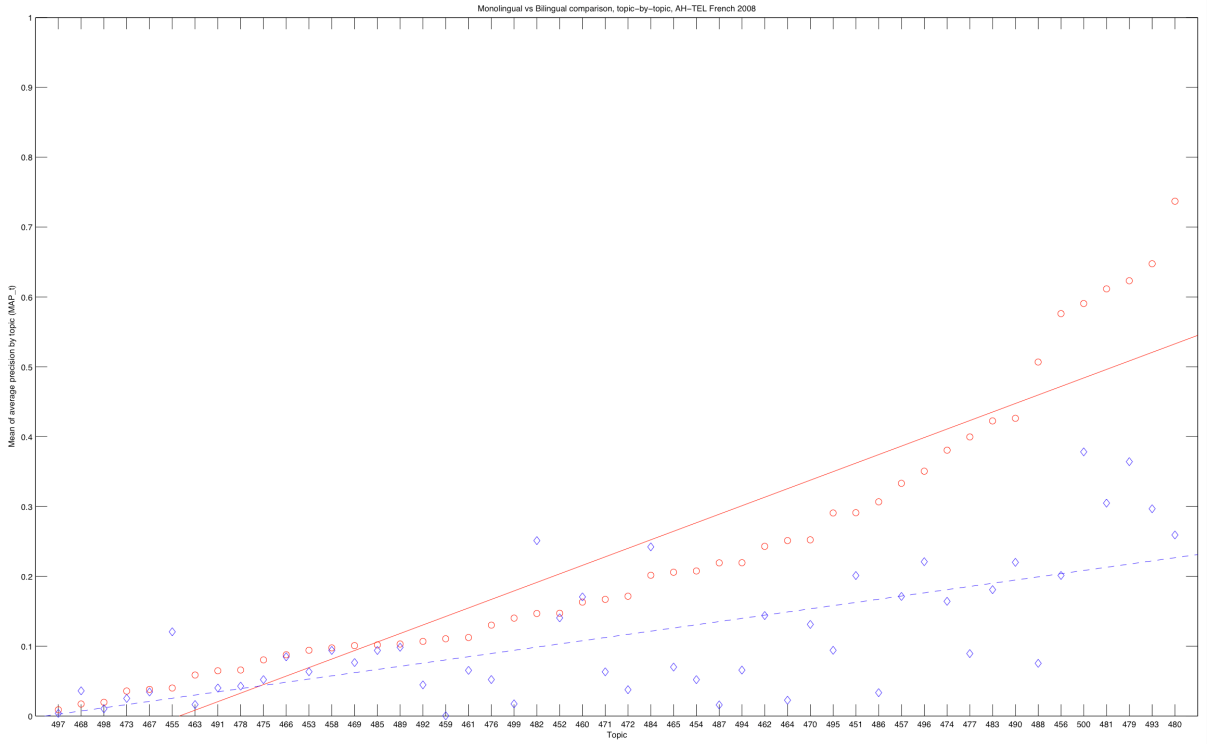

Figure 14: French TEL tasks: comparison of monolingual (red circle, solid line) with respect to bilingual (blue diamond, dashed line) performances and linear fit.

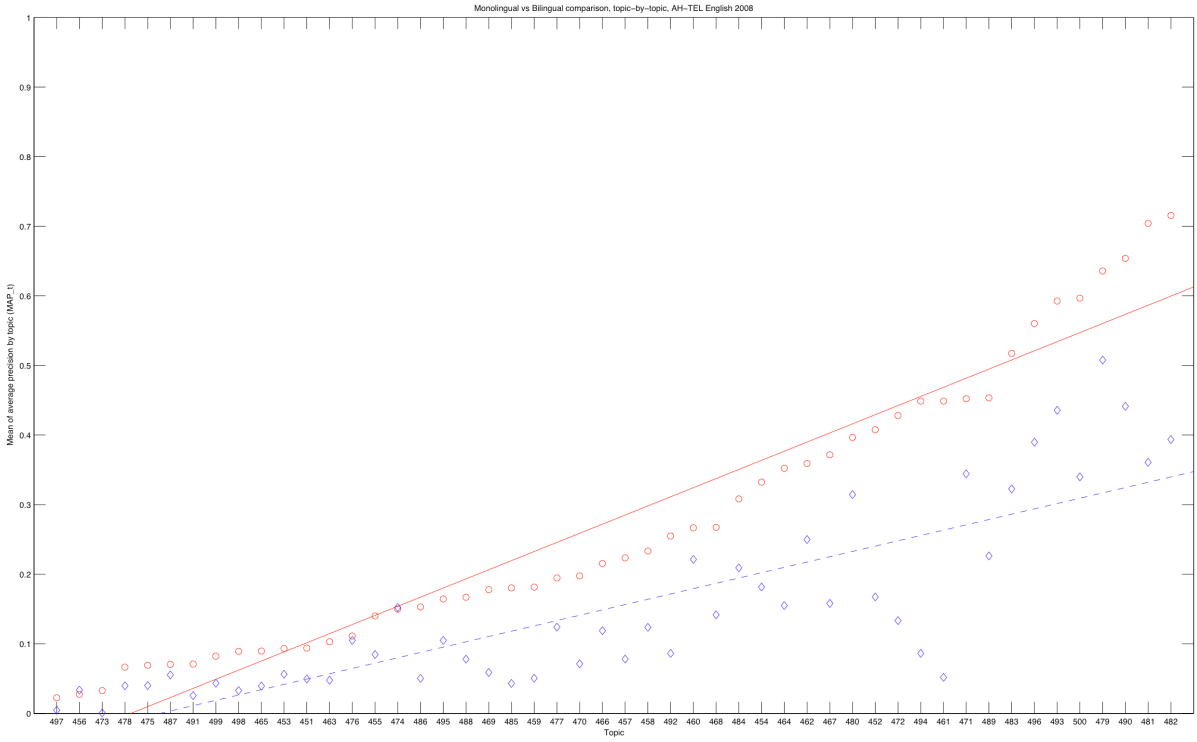

Figure 15: English TEL tasks: comparison of monolingual (red circle, solid line) with respect to bilingual (blue diamond, dashed line) performances and linear fit.

2.4.3 Ad Hoc TEL: Monolingual vs Bilingual English

Figure 15 shows the interpolating lines for the two tasks. In this picture it is clear that it is the group of topics on the right side that put the line of the bilingual task below and far the monolingual line. The analysis on the single sets of experiments with the same source language shows no difference in the outcome.

Moreover, even if the general claims for German (bilingual is about 53% of monolingual) and English (bilingual is about 91% of monolingual) are quite different, from Figure 13 and Figure 15 it turns out that they actually behave in similar ways since in both German and English monolingual constantly performs better than bilingual.

2.4.4 Ad Hoc Persian: Monolingual vs Bilingual Farsi

The linear fit in Figure 16 shows the two crossing interpolating lines. Similarly to the case of Ad-Hoc TEL French, the point where the two lines cross is close to the left part of the figure, where the topics with a low performance are. In fact, for almost one third of the topics among the first ten (from left to right) the performance of the bilingual is better than the monolingual one.

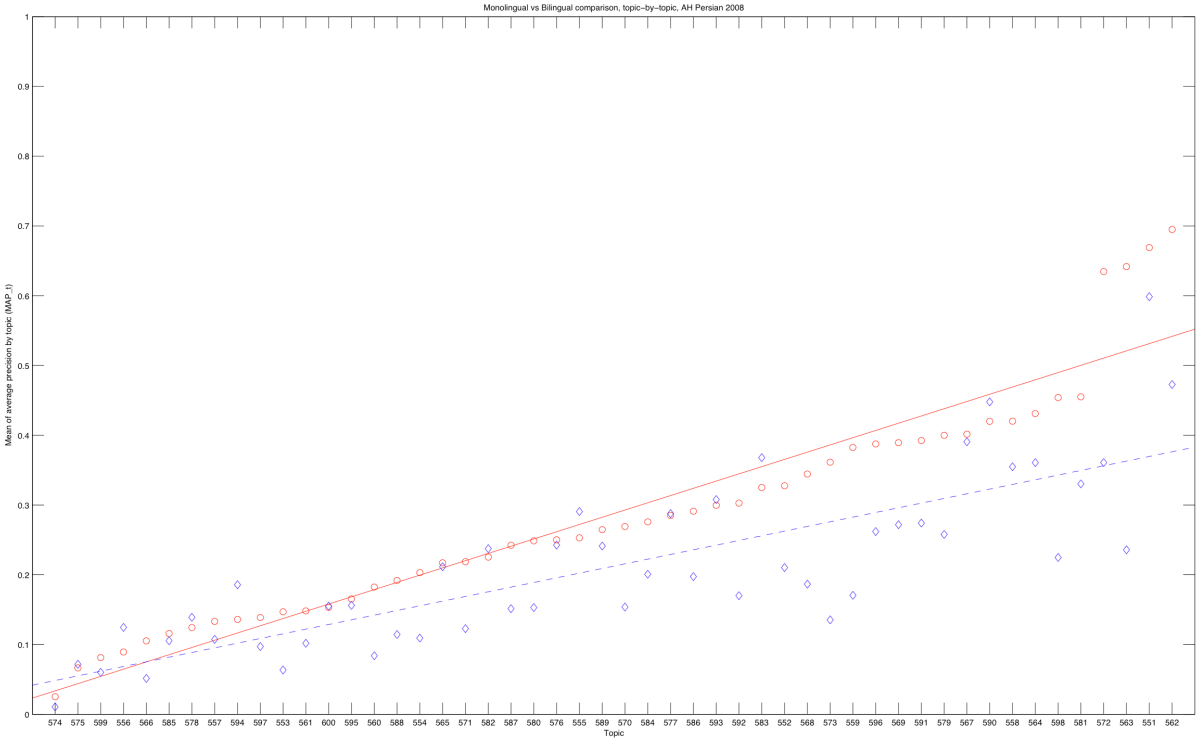

Figure 16: Persian tasks: comparison of monolingual (red circle, solid line) with respect to bilingual (blue diamond, dashed line) performances and linear fit.

3 ImageCLEF

3.1 Overview

The ImageCLEF[8] track of CLEF has run since 2003 with the aim of providing resources to evaluate (mainly within a lab-based context) the effectiveness of (cross-language) image retrieval systems (see, e.g. [11]). Although the focus in ImageCLEF has been predominantly on system-centred evaluation, the organisers have also been actively involved in contributing to user-centred evaluation, within iCLEF[9] (the interactive track of CLEF). The organisers of ImageCLEF typically provide the following resources: document (image and text) collections, example search tasks (topics), relevance judgements for each topic and utility resources (e.g. output from a default CBIR system). A total of 63 groups registered for five tasks in 2008: ad-hoc photographic retrieval (ImageCLEFphoto), medical retrieval from scientific literature sources (ImageCLEFmed), Wikipedia retrieval task (WikipediaMM), medical image annotation and a visual concept detection task.

The track has seen a consistent (and rapid) growth in participation: going from one task and 4 participants submitting results in 2003, to five tasks and 45 participants submitting results in 2008. The main highlights of 2008 campaign include: strong participation in the photographic retrieval task (evaluation of diversity-based ranking methods), the inclusion of a new cross-language image retrieval task based on images with semi-structured user-generated annotations from Wikipedia (derived from the INEX multimedia track), the development of a new database for medical retrieval based on a collection of scientific medical literature and a successful pre-CLEF workshop[10] on multimedia retrieval evaluation sponsored by the EU-funded Quaero[11] project.

[8] See `http://imageclef.org`
[9] See `http://np.uned.es/iCLEF/`
[10] See `http://imageclef.org/2008/preWorkshop`
[11] See `http://www.quaero.org/`

The aim of the photographic retrieval task for 2008[12] was to promote diversity within search results based on a general collection of 20,000 images (with queries and captions offered in German and English). Although a current "hot-topic" in information retrieval, less research has been conducted on generating resources to evaluate diversity. In total, 24 groups submitted 1042 runs to this task, typically using a two stage process: ad-hoc retrieval followed by the clustering of results to produce a final ranking. Analysis of results (and further experiments) indicated that standard retrieval does not promote diversity, that the choice of query (and collection) language made negligible differences to results, that (based on further experiments) users prefer diverse results, and that combining concept and content-based retrieval methods gives the overall best results.

The medical retrieval task[13] has seen considerable interest in the past 4 years as one of the premier forums for comparative evaluation of medical retrieval systems. The task in 2008 made use of a new database of radiology/radiography documents (scientific articles that contain diagrams, photos and illustrations). This is particularly significant because this kind of resource is exactly the kind of information that clinicians realistically search on a regular basis. The task attracted submissions from 15 groups (111 runs) for a bilingual search task. Analysis of results indicated that topics defined by the organisers may need to be re-designed for future years; that image analysis in this domain had to be performed with care and that visual retrieval could be seen to improve precision earlier in the ranking.

An exciting new development for 2008 was the WikipediaMM Retrieval task, which attracted 15 groups (submitting 77 runs). Based on an image collection created for the INEX Multimedia (MM) Track[14] (2006-2007), organisers provided a collection of 150,000 images from Wikipedia which come with associated unstructured and noisy textual annotations in English [48]. The task for 2008 was a purely English monolingual task, but highly challenging due to the nature of user-generated annotations. Analysis of results indicated that text retrieval alone performed well, but the use of concept labels (defined for visual concepts) could be used to improve performance. An interesting aspect of this task was the use of participants in the judging of relevance for the supplied topics.

The aim of the image annotation task for medical images[15] is to automatically assign class labels to radiological images using a hierarchical medical classification scheme. The labels can then be used for storage and retrieval purposes. The challenges of this task are the domain (medical - radiology), the use of a hierarchical classification scheme and the distribution of classes in the training and text data are not equal (this requires that participants use a confidence level on each hierarchy level). The organizers have developed (over the years) a database of 12,089 fully categorized radiographs (and 1,000 training images). The task witnessed 6 groups submitting a total of 24 runs using purely visual-based techniques. Analysis of results indicated that local features outperform global ones and that use of machine learning techniques are key to succeeding at this task. The results from the 2007 task were also published in a special issue of Pattern Recognition.

Finally, the 2008 visual concept detection task[16] continued the 2006 object annotation task and the 2007 object retrieval task. The aim of the task was identifying within a collection of 20,000 general photographs (the same collection as used in ImageCLEFphoto) a set of simple visual concepts, organised into a simple hierarchy (e.g. labelling images as 'indoor' vs. 'outdoor'). The goal was to associate this task with the photo retrieval task with the expectation that a labelling step would enhance ad-hoc retrieval (this was performed by only one participant). The task achieved participation from 11 groups (53 runs) and involved purely visual techniques. Analysis of results indicated that purely visual concept detection works well (i.e. using a simple and small set of classes), and that local features (such as SIFT) outperform other techniques.

[12] See http://www.imageclef.org/2008/photo
[13] See http://www.imageclef.org/2008/medical
[14] See http://inex.is.informatik.uni-duisburg.de/2007/mmtrack.html
[15] See http://www.imageclef.org/2008/medaat
[16] See http://www.imageclef.org/2008/vcdt

Overall the evaluation campaign was highly successful, with positive feedback from participants at the CLEF workshop in Arhus and a record number of participating groups (both academic and commercial). For 2009, the organizers are planning several enhancements to the current evaluation campaign. These include: for the medical retrieval task using cases instead of images as unit to retrieve (this models more realistically clinical practice), the potential use of 3D images for the medical classification task, multiples query languages for the Wikipedia task, an object recognition task using a possible robot vision database, and a larger and more diverse set of images and queries for the photographic retrieval task. Some of the ImageCLEF organisers have also been involved in preparing a Special Issue on Image and Video Retrieval Evaluation for the journal: Computer Visual and Image Understanding[17] (CVIU). This aims to promote the work of ImageCLEF beyond the traditional CLEF audience to computer vision communities. We now describe one of the ImageCLEF tasks in more detail: ImageCLEFphoto.

3.2 ImageCLEFphoto

The main objective of ImageCLEFphoto for 2008 comprised the evaluation of ad-hoc multilingual visual information retrieval systems from a general collection of annotated photographs (i.e. image with accompanying semi-structured captions such as the title, location, description, date or additional notes). However, this year focused on a particular aspect of retrieval: diversity of the results set. Research in image search has recently concentrated on ensuring that duplicate or near-duplicate documents retrieved in response to a query are hidden from the user. This should ideally lead to a ranked list where images are both relevant *and* diverse. More details of the task and analysis of results can be found in [5], [6], and [43].

3.2.1 Document collection

Similar to the 2006 and 2007 ImageCLEFphoto, we generated a subset of the IAPR TC-12 Benchmark as an evaluation resource for 2008. The IAPR TC-12 Benchmark consists of 20,000 colour photographs taken from locations around the world and comprises a varying cross-section of still natural images. Figure 17 illustrates a number of sample images from a selection of categories [22].

The majority of images have been provided by Viventura[18], an independent travel company that organises adventure and language trips to South America. Travel guides accompany the tourists and maintain a daily online diary including photographs of trips made and general pictures of each location including accommodation, facilities and ongoing social projects. In addition to these photos, a number of photos from a personal archive have also been added to form the collection used in ImageCLEF. The collection is publicly available for research purposes and, unlike many existing photographic collections, can be used to evaluate image retrieval systems. The collection is general in content with many different images of similar visual content, but varying illumination, viewing angle and background. This makes it a challenge for the successful application of techniques involving visual analysis.

Sports *Landscapes* *Animals* *People*

Figure 17: Sample images from the IAPR TC-12 collection

[17] See http://www.imageclef.org/cviusi
[18] See http://www.viventura.de/

Each image in the collection has a corresponding semi-structured caption consisting of the following six fields: (1) a unique identifier, (2) a title, (3) a free-text description of the semantic and visual contents of the image, (4) notes for additional information, (5) where and (6) when the photo was taken. Figure 18 shows a sample image with its corresponding textual annotation (in English).

<DOC>
<DOCNO>annotations/16/16392.eng</DOCNO>
<TITLE>Sunset in Salvador</TITLE>
<DESCRIPTION>a sandy beach at the sea with dark rocks behind it; the setting sun in an orange sky in the background;
</DESCRIPTION>
<NOTES></NOTES>
<LOCATION>Salvador, Brazil</LOCATION>
<DATE>10 October 2004</DATE>
<IMAGE>images/16/16392.jpg</IMAGE>
<THUMBNAIL>thumbnails/16/16392.jpg</THUMBNAIL>
</DOC>

Figure 18: Sample image caption.

By using a custom-built application for managing the images, various subsets of the collection can be generated with respect to a variety of particular parameters (e.g. using a selected subset of caption fields). For 2008, the following data was provided: two sets of annotations in (1) English and (2) Random. In the random set, the annotation language was randomly selected from for each of the images (i.e. annotations are either German or English image captions), all caption fields were provided for the 2008 task and each image caption exhibited the same level of annotation completeness - there were no images without annotations (as experimented with in 2006).

3.2.2 Topics

From an existing set of 60 topics (those used in 2006-07), 39 were selected and distributed to participants representing varying search requests (many of these are realistic and based on queries extracted during log file analysis – see [6] for more detailed information). We found that for the new retrieval challenge (promoting diversity), not all of the existing topics were suitable and therefore some were removed (see [6] for further details). Although 21 topics were removed, the remaining 39 topics are well-balanced, diverse and should present a retrieval challenge to participants wishing to use either text and/or low-level visual analysis techniques for creating clusters. Similar to TREC, topics were provided as structured statements of user needs. The full description of a topic consists of (1) a topic titles (2) a topic narrative, (3) a newly added *cluster type* and (4) three example relevant images for that topic. An additional field (cluster type) was added for easier assessment of the clusters as well as to facilitate the quantification of the result set diversity [5]. Below is an example augmented topic:

```
<top>
<num> Number: 48 </num>
<title> vehicle in South Korea </title>
<cluster> vehicle </cluster>
...
</top>
```

The cluster type in topic 48 is vehicle *(*in the <cluster> tag), which clearly defines how relevant images from this topic should be clustered. Different from previous years, topics were available in English only.

3.2.3 Relevance assessments

The relevance assessments, with the exception of removing any additional images considered as non-relevant, are exactly the same as in year 2007 [18]. To enable diversity to be quantified, it was necessary to classify images relevant to a given topic to one or more sub-topics or clusters. This was performed by two assessors. In the case of inconsistent judgements, a third assessor was used to resolve the inconsistencies. The resulting cluster assessment judgements are then used in combination with the normal relevance assessment to determine the retrieval effectiveness of each submitted system run (for further details see [5]).

3.2.4 Evaluating submissions

Once relevance judgements and cluster relevance assessments were completed, the performance of individual systems and approaches were evaluated. The results for submitted runs were computed using the latest version of treceval, as well as a custom-built tool to calculate diversity of the results set[19]. Submissions were evaluated using two metrics: (1) precision at rank 20 (P20) and (2) cluster recall at rank 20 (CR20)[20]. The classic harmonic mean combination measure in IR, f was used to combine CR20 and P20 into a single measure. Both measures are defined below:

$$CR(K) = \frac{\left| \cup_{i=i}^{K} subtopics(d_i) \right|}{n_A} \qquad f = 2\frac{P \cdot CR}{P + CR}$$

where K represents the rank (20 in our case), d_i represents the i^{th} document, $subtopics(d_i)$ is the number of sub-topics d_i belongs to, and n_A is the total number of sub-topics in a particular topic.

Rank 20 was selected as the cut-off point to measure precision and cluster recall because most online image retrieval engines (e.g. Google, Yahoo! and AltaVista) display 18 to 20 images by default. Further measures considered included uninterpolated (arithmetic) Mean Average Precision (MAP), Geometric Mean Average Precision (GMAP) to test system robustness and binary preference (bpref), which is a good indicator of how complete relevance judgements are. To enable an absolute comparison between individual runs, a single metric is required: the F1-measure was used to combine scores from P20 and CR20 (representing the harmonic mean of P20 and CR20).

3.2.5 Results

In 2008, 43 groups registered for ImageCLEFphoto (32 in 2007, 36 in 2006), with 24 groups eventually submitting a total of 1,042 runs[21] (all of which were evaluated by the organisers). This is an increase in the number of runs from previous years (20 groups submitting 616 runs in 2007, 12 groups submitting 157 runs in 2006, and 11 groups 349 runs in 2005 respectively). The 24 participating groups are affiliated to 21 different institutions in 11 countries. New participants submitting in 2008 include joint work from four French labs (AVEIR), University of Waseda (GITS), Laboratory of Informatics of Grenoble (LIG), System and Information Science Lab (LSIS), Meiji University (Meiji), University of Ottawa (Ottawa), Telecom ParisTech (PTECH), University of Alicante (TEXTMESS) and Piere & Marie Curie University (UPMC). Overall, 65% of the participants in 2007 returned and participated in 2008.

Overall, 1042 runs were submitted and categorised with respect to the following dimensions: (1) annotation language, (2) modality (text only, image only or combined) and (3) run type (automatic or manual). Table 8 provides an overview of all submitted runs according to these dimensions. Most

[19] See `http://imageclef.org/ClusterEval`

[20] Based on S-Recall - Zhai, C., Cohen, W. W. and Lafferty, J. (2003) Beyond independent relevance: Methods and evaluation metrics for subtopic retrieval. In Proceedings of ACM SIGIR 2003, pp. 10–17.

[21] For 2008 we did not put a limit on the number of submissions, hence some institutions submitted many runs, e.g. Dublin City University (DCU) submitted a total of 733 runs.

submissions (96.8%) used the provided image annotations, with 22 groups submitting a total of 404 purely concept-based (textual) runs and 19 groups a total of 605 runs using a combination of content-based (visual) and concept-based features. A total of 11 groups submitted 33 purely content-based runs. Of all retrieval approaches, 61.2% involved the use of image retrieval (53.4% in 2007 and 31% in 2006), 79% of all groups used content-based (i.e. visual) information in their runs (60% in 2007 and 58% in 2006). Almost all of the runs (99.7%) were automatic (i.e. involving no human intervention); only 3 submitted runs were manual.

Dimensions	Type	2008		2007		2006	
		Runs	Groups	Runs	Groups	Runs	Groups
Annotation language	EN	514	24	271	17	137	2
	RND	495	2	32	2		
Modality	Text Only	404	22	167	15	121	2
	Mixed (text and image)	605	19	255	13	21	1
	Image Only	33	11	52	12		
Run type	Manual	3	1	19	3		
	Automatic	1039	25	455	19	142	2

Table 8: Submission overview by dimension.

Table 9 and Table 10 show the runs which achieved highest F1-measure scores for the two annotation languages: ENG and RND. Taking into account that only two groups submitted 495 runs with a random annotation language, the result shows the same trend as in previous years: the highest monolingual run still outperforms the highest bilingual run, which consists of a random annotation language. However, as in previous years, the margin of difference is low and can be attributed to significant progress of the translation and retrieval methods using these languages. The best performing runs using random annotations performed with an F1-measure score at 97.4% of the highest monolingual run. Hence, the language barrier is no longer a critical factor in achieving good retrieval results.

Query language	Annotation language	Group	Run-ID	Run type	Modality	P20	CR20	F1-Measure
English	English	PTECH	PTECH-EN-EN-MAN-TXTIMG-MMBQI.run	MAN	TXTIMG	0.6885	0.6801	0.6843
English	English	PTECH	PTECH-EN-EN-MAN-TXTIMG-MMBMI.run	MAN	TXTIMG	0.6962	0.6719	0.6838
English	English	PTECH	PTECH-EN-EN-MAN-TXT-MTBTN.run	MAN	TXT	0.5756	0.5814	0.5785
English	English	XRCE	xrce_tilo_nbdiv_15	AUTO	TXTIMG	0.5115	0.4262	0.4650
English	English	DCU	DCU-EN-EN-AUTO-TXTIMG-qe.txt	AUTO	TXTIMG	0.4756	0.4542	0.4647
English	English	XRCE	xrce_tilo_nbdiv_10	AUTO	TXTIMG	0.5282	0.4146	0.4646
English	English	XRCE	xrce_cm_nbdiv_10	AUTO	TXTIMG	0.5269	0.4111	0.4619
English	English	DCU	DCU-EN-EN-AUTO-TXTIMG.txt	AUTO	TXTIMG	0.4628	0.4546	0.4587
English	English	XRCE	xrce_cm_mmr_07	AUTO	TXTIMG	0.5282	0.4015	0.4562
English	English	XRCE	xrce_tfidf_nbdiv_10	AUTO	TXTIMG	0.5115	0.4081	0.4540

Table 9: Systems with the highest F1-Measure for English annotations.

Query language	Annotation language	Group	Run-ID	Run type	Modality	P20	CR20	F1-Measure
English	RND	DCU	DCU-EN-RND-AUTO-TXTIMG-tr.txt	AUTO	TXTIMG	0.4397	0.4673	0.4531
English	RND	DCU	DCU-EN-RND-AUTO-TXTIMG-tr-qe.txt	AUTO	TXTIMG	0.4423	0.4529	0.4475
English	RND	DCU	DCU-EN-RND-AUTO-TXTIMG-tr-d50-k40-tf-all.txt	AUTO	TXTIMG	0.4038	0.4967	0.4455
English	RND	DCU	DCU-EN-RND-AUTO-TXTIMG-tr-d50-k40-tfidf-all.txt	AUTO	TXTIMG	0.3974	0.4948	0.4408
English	RND	DCU	DCU-EN-RND-AUTO-TXTIMG-tr-d50-kx-tfidf-all.txt	AUTO	TXTIMG	0.3897	0.5049	0.4399
English	RND	DCU	DCU-EN-RND-AUTO-TXTIMG-tr-d50-kx-tf-qe-all.txt	AUTO	TXTIMG	0.4013	0.4806	0.4374
English	RND	DCU	DCU-EN-RND-AUTO-TXTIMG-tr-d50-kx-tf-all.txt	AUTO	TXTIMG	0.3910	0.4936	0.4363
English	RND	DCU	DCU-EN-RND-AUTO-TXTIMG-tr-d50-k40-tfidf-qe-all.txt	AUTO	TXTIMG	0.4013	0.4766	0.4357
English	RND	DCU	DCU-EN-RND-AUTO-TXTIMG-tr-d50-kx-tfidf-qe-all.txt	AUTO	TXTIMG	0.3897	0.4768	0.4289
English	RND	DCU	DCU-EN-RND-AUTO-TXTIMG-tr-d50-k40-tf-qe-all.txt	AUTO	TXTIMG	0.3897	0.4678	0.4252

Table 10: Systems with the highest F1-Measure for Random annotations (German / English).

In 2006 and 2007, the results showed that by combining visual features from the image and semantic knowledge derived from the captions offered optimum performance for retrieval from a general photographic collection with fully annotated images. As indicated in Table 11, the results of ImageCLEFphoto 2008 show that this also applies for our modified task, which promotes diversity in the results set. However, contrary to 2007 (24% MAP improvement over averages for combining techniques over solely text-based approaches), the improvement is not as clearly visible when combining visual features from the image and semantic information. The difference between "Mixed" and "Text Only" runs is across the averages from all runs, and differs only marginally. However, looking at the best runs in each modality, the "Mixed" runs (F1-Measure = 0.4650) outperform the "Text Only" runs by 16% (F1-measure = 0.4008). Purely content-based approaches still lag behind, although with a smaller gap than in previous years. The best "Image Only" runs (F1-Measure = 0.3396) is higher than both averages for the "Mixed" and "Text only" runs.

Modality	Precision at 20		Cluster Recall at 20		F1-measure (P20/CR20)	
	Mean	SD	Mean	SD	Mean	SD
Mixed	0.2538	0.1023	0.3998	0.0977	0.3034	0.0932
Text Only	0.2431	0.0590	0.3915	0.0819	0.2957	0.0576
Image Only	0.1625	0.1138	0.2127	0.1244	0.1784	0.1170

Table 11: Results by retrieval modality.

Table 12 shows the average scores and the standard deviations across all systems runs with respect to the run type. Unsurprisingly, F1-Measure results of manual approaches are significantly higher than purely automatic runs. All submitted manual runs are done with English annotation, whereas the average of the automatic runs is both from English as well as Random annotation. However, as previously shown the translation does not have a big impact and can therefore be neglected. In case of the automatic runs the F1-measure is practically identical for the English (ENG) annotations and those with the language randomly selected (RND).

Technique	Precision at 20		Cluster Recall at 20		F1-measure (P20/CR20)	
	Mean	SD	Mean	SD	Mean	SD
Manual	0.6534	0.0675	0.6445	0.0548	0.6489	0.0610
Automatic	0.2456	0.0873	0.3899	0.0975	0.2955	0.0829
Automatic RND Only	0.2353	0.0651	0.4191	0.0731	0.2992	0.0679
Automatic ENG Only	0.2609	0.0990	0.3731	0.1002	0.2994	0.0879
Automatic IMG Only	0.1625	0.1138	0.2127	0.1244	0.1784	0.1170

Table 12: Results by run type.

3.2.6 Approaches used by participants

Some of the participating groups started by using a baseline run, carried out using different weighting methods (e.g. BM25, DFR, LM), with or without query expansion (e.g. using Local Content Analysis, Pseudo Relevance Feedback, thesaurus-based query expansion, Conceptual Fuzzy Sets, using a location hierarchy, and using Wordnet), and using content- and/or concept-based retrieval methods. The aim of this initial step was obtaining the best possible ranking (i.e. maximising the number of relevant documents returned in the top *n*). The most common following step was to re-rank the initial baseline run in order to promote diversity. One approach of re-ranking is to cluster the top *n* documents into sub-topics or clusters and then select the highest ranked document in each cluster and promote higher in the ranked list (i.e. to the top *n*). Clustering was mostly based on the associated textual information using various clustering algorithms (e.g. k-means, k-medoids, knn-density, and latent dirichlet allocation) and different weighting parameters. Some groups also tried to re-rank results using Maximal Marginal Relevance. Other approaches included merging different kind of runs (e.g. calculating image ranking with average/min/mean) or combining scores (novelty/ranking score) to get a diverse and relevant results list. Overall, a majority of approaches applied post-processing methods in one way or another.

3.2.7 Further analysis

We carried out further analysis of the results to investigate further aspects of precision and cluster recall based on submissions to ImageCLEFphoto 2008 [40]. This included a user experiment based on submitted runs to establish whether users preferred diverse search results for image retrieval. The challenge for participants in 2008 was to maximise both the number of relevant images, as well as the number of relevant image clusters represented within the top 20 results. The key results from our further analysis include:

1. A comparison between the 2008 and 2006/7 versions of the task showed that retrieval systems not explicitly built to maximise diversity (as typified in the 2006/7 tasks) had significantly lower CR than the systems that explicitly supported diversity. In other words “standard” retrieval systems do not by default support diversity well.
2. Although there was a concern that building a retrieval system that maximises diversity was likely to impact on precision, there was little evidence to support this. However, experiments to establish upper bound and random baselines for diversity showed that there is much potential to improve diversity in the future.
3. Finally, a user experiment showing pairs of search results at different levels of cluster recall produced significant evidence showing that users preferred the search results that were more diverse.

The results shown here strongly suggest that support for diversity is an important and currently largely overlooked aspect of information retrieval. However, these results were derived from just one test collection. This is because the collection is one of a very small number of existing public resources that can be used to test diversity. Consequently, a key goal of our future work is to create more a substantially larger collection for ImageCLEFPhoto 2009 to enable a broader range of diversity experiments to be conducted.

4 iCLEF

iCLEF stands for the interactive CLEF track, where MLIA systems are studied from a user-inclusive perspective. The central research question is how best to assist users when searching information written in foreign languages, rather than how best an algorithm can find information written in languages different from the query language.

Since 2006, iCLEF moved away from news collections in order to explore user behaviour in scenarios where the necessity for multilingual search arises more naturally for the average user. We chose Flickr as target collection: a large-scale image database based on a social network of WWW users sharing over two billion images, with the potential for offering both challenging and realistic multilingual search tasks for interactive experiments.

Over the last years, iCLEF participants have typically designed one or more cross-language search interfaces for tasks such as document retrieval (focusing on document selection [38] and query translation/refinement [32]), question answering [34] or text-based image retrieval [20]. Experiments were hypothesis-driven, and interfaces were studied and compared using controlled user populations under laboratory conditions. This experimental setting has provided valuable research insights into the problem, but has a major limitation: user populations are necessarily small in size, and the cost of training users and of scheduling and monitoring search sessions is very high. In addition, the target notion of relevance does not cover all aspects that make an interactive search session successful; other factors include user satisfaction with the results and usability of the interface, for instance.

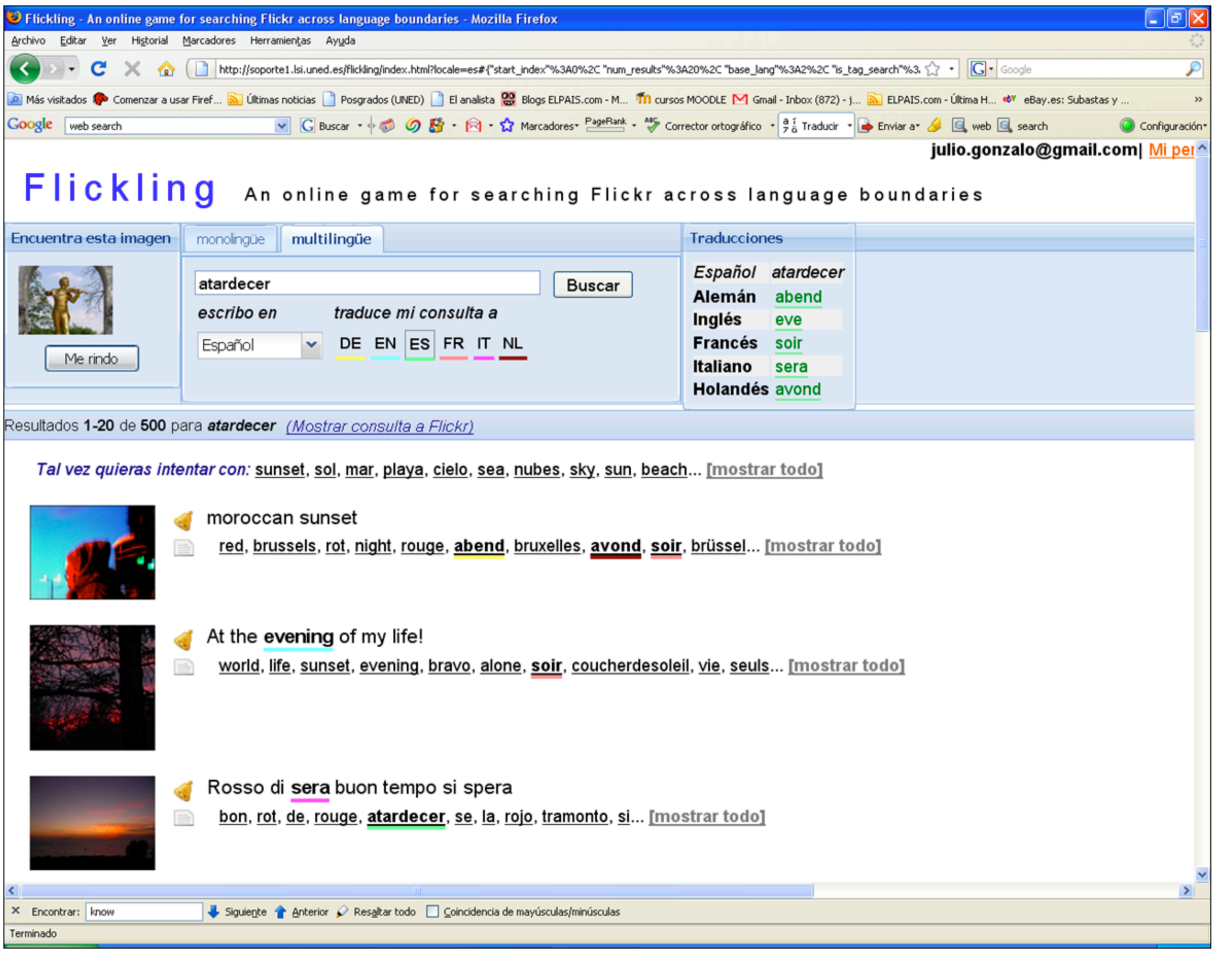

Figure 19: Flickr gaming interface available to iCLEF participants.

Thus, in iCLEF 2008 we made a major methodological move: the task focused on the shared analysis of a large search log from a single search interface provided by the iCLEF organizers, as shown in Figure 19. The focus is, therefore, on search log analysis (how users behave) rather than on system design (which design is better). Details on the task design can be found in the iCLEF overview [19]. The motivation is to study the behaviour of users in a naturalistic search scenario on a much larger data set than in previous iCLEF campaigns. The search interface provided by iCLEF organizers is a basic multilingual retrieval system to access images in Flickr, with the mechanics of an online game: the user is given an image, and it must be found in Flickr without any prior knowledge of the language(s) in which the image is annotated. Game-like features are intended to engage casual users and therefore increase the chances of achieving a large, representative search log. Details on the interface design can be found in [40].

In terms of the collected dataset, the experience was a success. The harvesting of the search log resukted in in the largest (known to us) available log of complete, and truly multilingual, search sessions. The dataset comprises

- 103 topics, which consists of a Flickr image and an ordered set of hints. The first hint is always the (primary) target language, and the rest of hints are keywords used to annotate the image in the collection. Therefore, before asking for the first hint, the search sessions are strictly multilingual (the user is simultaneously searching in six languages), and become monolingual or bilingual (depending on the match between the user language profile and the annotations in the target image) once the first hint is requested.
- A target collection which consists of all images uploaded in Flickr (available only via Flickr API).
- More than 230 active users from four continents with a large variety of language profiles and backgrounds (with a bias towards photography fans, IR researchers and university students).
- 5101 complete search sessions explicitly ending in success (the user finds the image) or failure (the user explicitly gives up), where all interactions between the user and the system are included in the log.
- 5101 post-search questionnaires (filled in after each search session) and over 70 post-experience questionnaires (filled in by users which had already searched for at least 15 topics).

In terms of participation, our feelings are mixed. Fifteen groups registered for the task, and six actually submitted results: Manchester Metropolitan University, University of Padua, University of Westminster, Indian Institute of Information Technology Hyderabad, Swedish Institute of Computer Science and UNED. Both are record figures for iCLEF, but still below our expectations. Probably most significant is the fact that many CLEF groups were really active as subjects for the log harvesting phase. For us, this implies that many MLIA researchers who had never actually seen the problem from the other side of the road (that of users) might now incorporate user-inclusive concerns in their research.

Overall, the main point is that, for the first time, iCLEF managed to produce a truly reusable data set for researchers interested in MLIA from the user's perspective. Study of these logs has already produced valuable results for the field in the context of the iCLEF 2008 task:

- A quantitative (and reliable) estimation of the task difficulty when the target language is unknown to the user, a passive language (the user can understand results but not make queries) or an active language (the user can understand results and make queries directly in the target language). UNED results, for instance, indicate that a passive knowledge of the target language may end up in a similar success rate (as compared with users having active knowledge of the target language), but at a higher cognitive effort (expressed as the number of interactions with the system before successfully completing the task). When the target language is unknown, however, the success rate decreases significantly and simultaneously the cognitive effort increases. This is an indication that more sophisticated cross-language search assistance is needed to bridge that gap.

- Observation studies made by some participating groups on controlled subgroups from the user population. For instance, Manchester Metropolitan University studied how users considered language and cross-linguistic issues during a session and how they switched between the cross-lingual and mono-lingual interfaces available, and University of Padua monitored a subset of users in a constrained version of the task, requiring users to make a rapid decision as to whether an image was relevant or not. These observational studies provided additional input on future variants of the task definition.
- An exploration (made by the University of Westminster) of users' interaction with the facility provided by the interface to add user-specific translation terms. By exploring the user's perceived language skills and usage of the personal dictionary feature, experiments demonstrated that even with modest language skills, users were interacting with and using the dictionary-edit feature. Results point towards further study of collaborative translation in the global web space.
- Evidence of different levels of user confidence and competence in the behaviour exhibited and recorded in them (Swedish Institute of Computer Science).

The results of the experiments will be used to inform more usage-oriented tasks for future cycles; the methodology has proven to be lightweight and should be helpful for future participants, and the logs will be a sustainable and reusable resource for future user-orientated studies of cross-language search behaviour.

When planning iCLEF 2009, there was a general consensus – between task participants and other CLEF observers – that the current log analysis approach was fruitful and should have a continuation next year. The most controversial issue was to choose the search task. In iCLEF 2008 we chose a known-item retrieval task (which is similar to the "stuff I've seen before" search scenario) for various reasons: (i) it was simple to understand and at the same time challenging, essential requisites when it is not possible to train and monitor users; (ii) it had a clear notion of success which does not involve costly relevance assessments; (iii) it was suitable for the game-like features – and especially the online ranking – that we wanted to have in order to attract and engage users. Initially, in iCLEF 2009 we would like to expand into other search scenarios, such as illustrating a text or ad-hoc facet retrieval (e.g. find as many images of *different* gothic cathedrals in Spain as possible). But we haven't found a clear way of incorporating this kind of task without losing some of the listed properties which are, in our opinion, essential to collect a substantial search log. Our decision has finally been to keep the task identical (with a few minor improvements), and focus on having larger participation in the track and harvesting a larger log than in iCLEF 2008.

5 QA@CLEF

QA@CLEF[22] 2008 was carried out following the methodology consolidated in previous years. Beside the traditional main task, three additional exercises were proposed:

- the *Answer Validation Exercise* (AVE)[23]: in its third round was aimed at evaluating answer validation systems based on textual entailment recognition. In this task, systems were required to emulate human assessment of QA responses and decide whether an *Answer* to a *Question* is correct or not according to a given *Text*. Results were evaluated against the QA human assessments.
- the *Question Answering on Speech Transcripts* (QAST)[24] [45]: which continued last year's successful pilot task, aimed at providing a framework in which QA systems could be evaluated when the answers to factual and definition questions must be extracted from spontaneous speech transcriptions.
- the *Word Sense Disambiguation for Question Answering* (QA-WSD)[25], a pilot task which provided the questions and collections with already disambiguated Word Senses in order to study their contribution to QA performance.

[22] See `http://clef-qa.itc.it/`
[23] See `http://nlp.uned.es/QA/ave/`
[24] See `http://www.lsi.upc.edu/~qast/`

As far as the main task is concerned, following last year experience, the exercise consisted of topic-related questions, i.e. clusters of questions which were related to the same topic and contained co-references between one question and the others. The requirement for questions related to a topic necessarily implies that the questions refer to common concepts and entities within the domain in question. This is accomplished either by co-reference or by anaphoric reference to the topic, implicit or explicitly expressed in the first question or in its answer.

In addition, besides the usual news collections provided by ELRA/ELDA, articles from Wikipedia were considered as an answer source. Some questions could have answers only in one collection, i.e. either only in the news corpus or in Wikipedia.

As a general remark, this year we had the same number of participants as in 2007 campaign, but the number of submissions went up. Due to the complexity of the innovation introduced in 2007 - the introduction of topics and anaphora, list questions, Wikipedia corpus - the questions tended to be more difficult and the performance of systems dropped dramatically, so, people were disinclined to continue the following year (i.e. 2008), inverting the positive trend in participation registered in the previous campaigns.

As reflected in the results, the task proved to be even more difficult than expected. Results improved in the monolingual subtasks but are still very low in the cross-lingual subtasks.

5.1 Task Description

As far as the main task is concerned, the consolidated procedure was followed, capitalizing on the experience of the task proposed in 2007.

The exercise consisted of topic-related questions, i.e. clusters of questions which were related to the same topic and contained co-references between one question and the others. Neither the question types nor the topics were given to the participants.

The systems were fed with a set of 200 questions – which could concern facts or events (Factoid questions), definitions of people, things or organisations (Definition questions), or lists of people, objects or data (List questions) – and were asked to return up to three exact answers per question, where *exact* meant that neither more nor less than the information required was given.

The answer needed to be supported by the identifier of the document in which the exact answer was found, and by portion(s) of text, which provided enough context to support the correctness of the exact answer. Supporting texts could be taken from different sections of the relevant documents, and could sum up to a maximum of 700 bytes. There were no particular restrictions on the length of an answer-string, but unnecessary pieces of information were penalized, since the answer was marked as *ineXact*. As in previous years, the exact answer could be exactly copied and pasted from the document, even if it was grammatically incorrect (e.g.: inflectional case did not match the one required by the question). Anyway, systems were also allowed to use natural language generation in order to correct morpho-syntactical inconsistencies (e.g., in German, changing *dem Presidenten* into *der President* if the question implies that the answer is in nominative case), and to introduce grammatical and lexical changes (e.g., QUESTION*: What nationality is X*? TEXT: *X is from the Netherlands* → EXACT ANSWER: Dutch).

The subtasks were both:

- monolingual, where the language of the question (Source language) and the language of the news collection (Target language) were the same;
- cross-lingual, where the questions were formulated in a language different from that of the news collection.

Two new languages were added, i.e. Basque and Greek both as source and target languages. In total eleven source languages were considered, namely, Basque, Bulgarian, Dutch, English, French,

[25] See `http://ixa2.si.ehu.es/qawsd/`

German, Greek, Italian, Portuguese, Romanian and Spanish. All these languages were also considered as target languages.

As shown in Table 13, 43 tasks were proposed:

- 10 Monolingual – i.e. Bulgarian (BG), German (DE), Greek (EL), Spanish (ES), Basque (EU), French (FR), Italian (IT), Dutch (NL), Portuguese (PT) and Romanian (RO);
- 33 Cross-lingual (as customary in recent campaigns, in order to prepare the cross-language subtasks, for which at least one participant had registered, some target language question sets were translated into the combined source languages).

however, not all the proposed tasks were then carried out by the participants. As long-established, the monolingual English (EN) task was not available as it has already been thoroughly investigated in the TREC campaigns. English was still both offered source and target language in the cross-language tasks.

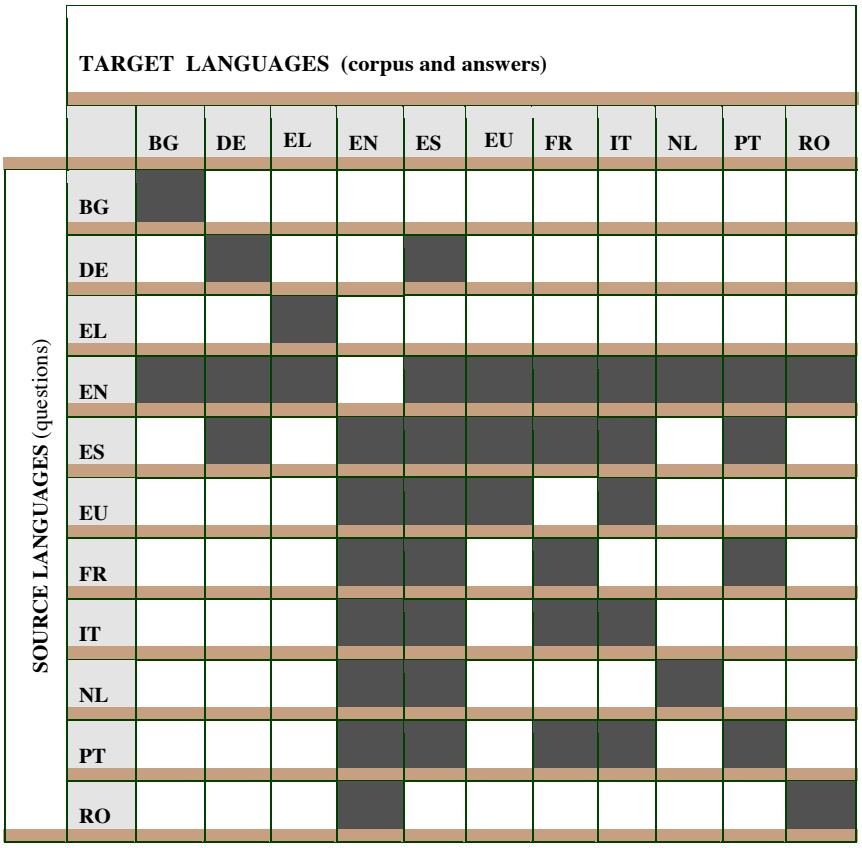

Table 13: QA tasks activated in 2008.

5.2 Document Collections

In addition to the data collections composed of news articles provided by ELRA/ELDA (see Table 14), also Wikipedia was considered.

Wikipedia pages in the target languages, as found in the version of November 2006, could be used. Romanian had Wikipedia[26] as the only document collection, because there was no Romanian news corpus available. The "snapshots" of Wikipedia were made available for download both in XML and HTML versions. The answers to the questions had to be taken from actual entries or articles of

[26] See `http://static.wikipedia.org/downloads/November_2006/ro/`

Wikipedia pages. Other types of data such as images, discussions, categories, templates, revision histories, as well as any files with user information and meta-information pages, had to be excluded.

One of the major reasons for using Wikipedia was to make a first step towards web formatted corpora in which to search for answers. In fact, as nowadays such large information sources are available on the web, this may be considered a desirable next level in the evolution of QA systems. An important advantage of Wikipedia is that it is freely available for all languages so far considered. However, the variation in size of Wikipedia, depending on the language, is still problematic.

TARGET LANG.	COLLECTION	PERIOD	SIZE
[**BG**] Bulgarian	Sega	2002	120 MB (33,356 docs)
	Standart	2002	93 MB (35,839 docs)
	Novinar	2002	48 MB (18,086 docs)
[DE] German	Frankfurter Rundschau	1994	320 MB (139,715 docs)
	Der Spiegel	1994/1995	63 MB (13,979 docs)
	German SDA	1994	144 MB (71,677 docs)
	German SDA	1995	141 MB (69,438 docs)
[EL] Greek	The Southeast European Times	2002	
[EN] English	Los Angeles Times	1994	425 MB (113,005 docs)
	Glasgow Herald	1995	154 MB (56,472 docs)
[ES] Spanish	EFE	1994	509 MB (215,738 docs)
[EU] Basque	EFE	1995	577 MB (238,307 docs)
	Egunkaria	2001/2003	
[FR] French	Le Monde	1994	157 MB (44,013 docs)
	Le Monde	1995	156 MB (47,646 docs)
	French SDA	1994	86 MB (43,178 docs)
	French SDA	1995	88 MB (42,615 docs)
[IT] Italian	La Stampa	1994	193 MB (58,051 docs)
	Italian SDA	1994	85 MB (50,527 docs)
	Italian SDA	1995	85 MB (50,527 docs)
[NL] Dutch	NRC Handelsblad	1994/1995	299 MB (84,121 docs)
	Algemeen Dagblad	1994/1995	241 MB (106,483 docs)
[PT] Portuguese	Público	1994	164 MB (51,751 docs)
	Público	1995	176 MB (55,070 docs)
	Folha de São Paulo	1994	108 MB (51,875 docs)
	Folha de São Paulo	1995	116 MB (52,038 docs)

Table 14: Document collections used in QA@CLEF 2008.

5.3 Topics and Questions

The procedure followed to prepare the test set was the same as that used in the 2007 campaign. First of all, each organizing group, responsible for a target language, freely chose a number of topics. For each

topic, one to four questions were generated. Topics could be not only named entities or events, but also other categories such as objects, natural phenomena, etc. (e.g. George W. Bush; Olympic Games; notebooks; hurricanes; etc.). The set of ordered questions were related to the topic as follows:

- the topic was named either in the first question or in the first answer
- the following questions could contain co-references to the topic expressed in the first question/answer pair.

Topics were not given in the test set, but could be inferred from the first question/answer pair. For example, if the topic was *George W. Bush*, the cluster of questions related to it could have been:

Q1: *Who is George W. Bush?*; Q2: *When was he born?*; Q3: *Who is his wife?*

As far as the question types are concerned, as in previous campaigns, the three following categories were considered:

- *factoid questions*, fact-based questions, asking for the name of a person, a location, the extent of something, the day on which something happened, etc.;
- *definition questions*, questions such as "What/Who is X?";
- *closed list questions*: i.e., questions that require one answer containing a determined number of items, e.g.: Q20: Name all the airports in London, England. A20: Gatwick, Stansted, Heathrow, Luton and City.

As only one answer was allowed, all the items had to be present in sequence in the document and copied, one next to the other, in the answer slot.

Moreover, all types of questions could contain a temporal restriction, i.e. a temporal specification that provided important information for the retrieval of the correct answer, for example:

> Q21: *Who was the Chancellor of Germany from 1974 to 1982?*
>
> A21: *Helmut Schmidt.*

Some questions could have no answer in the document collection, and in that case the exact answer was "NIL" and the answer and support docid fields were left empty. A question was assumed to have no right answer when neither human assessors nor participating systems could find one.

5.4 Evaluation

As far the evaluation process is concerned, no changes were made with respect to the previous campaigns. Human judges assessed the exact answer (i.e. the shortest string of words which is supposed to provide the exact amount of information to answer the question) as:

- R (Right) if correct;
- W (Wrong) if incorrect;
- X (ineXact) if contained less or more information than that required by the query;
- U (Unsupported) if either the docid was missing or wrong, or the supporting snippet did not contain the exact answer.

Most assessor-groups managed to guarantee a second judgement of all the runs.

As regards the evaluation measures, the main one was accuracy, defined as the average of SCORE(q) over all 200 questions q, where SCORE(q) is 1 in the first answer to q in the submission file is assessed as R, and 0 otherwise. In addition most assessor groups computed the following measures:

- Confident Weighted Score (CWS). Answers are in a decreasing order of confidence and CWS rewards systems that give correct answers at the top of the ranking [16]
- the Mean Reciprocal Rank (MRR) over N assessed answers per question (to consider the three answers). That is, the mean of the reciprocal of the rank of the first correct label over all questions. If the first correct label is ranked as the 3rd label, then the reciprocal rank (RR) is 1/3. If

none of the first N responses contains a correct label, RR is 0. RR is 1 if the highest ranked label matches the correct label.

5.5 Results

As far as accuracy is concerned, scores were generally low, as Figure 20 shows. Although comparison between different languages and years is not possible, in Figure 1 we can observe some trends which characterized this year's competition: best accuracy in the monolingual task increased with respect to last year, going up again to the values recorded in 2006. But systems - even those that participated in all previous campaigns - did not achieve a brilliant overall performance. Apparently systems could not manage the new challenges suitably, although they improved their performances when tackling issues already treated in previous campaigns.

More in detail, best accuracy in the monolingual task scored 63,5 almost ten points up with respect to last year, meanwhile the overall performance of the systems was quite low, as average accuracy was 23,63, practically the same as last year. On the contrary, the performances in the cross-language tasks recorded a drastic drop: best accuracy reached only 19% compared to 41,75% in the previous year, which means more than 20 points lower, meanwhile average accuracy was more or less the same as in 2007 - 13,24 compared to 10,9.

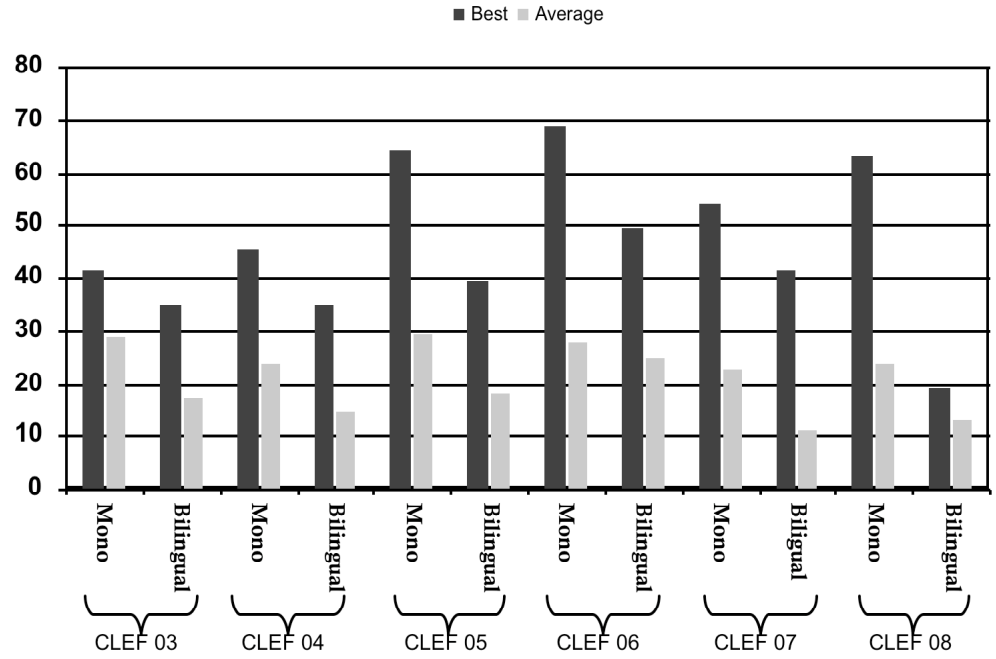

Figure 20: Best and average scores in QA@CLEF campaigns.

It seems that the high level of difficulty of the question sets particularly impacted the bilingual tasks and the task proved to be still difficult also for veterans.

5.5.1 Basque as target

In the first year working with Basque as target only one research group submitted runs for evaluation in the track having Basque as target language, the Ixa group from the University of the Basque Country. They sent four runs: one monolingual, one English-Basque and two Spanish-Basque.

The Basque question set consisted of 145 factoid questions, 39 definition questions and 16 list questions. 39 questions contained a temporal restriction, and 10 had no answer in the Gold Standard. 40 answers were retrieved from Wikipedia, the rest from the news collections. Half of the questions were linked to a topic, so the second (and sometimes the 3rd) question was more difficult to answer.

The news were from the Egunkaria newspaper 2000, 2001 and 2002 and the information from Wikipedia was the export corresponding to Wikipedia content in 2006.

Table 15 shows the evaluation results for the four submitted runs (one monolingual and three cross-lingual). The table shows the number of Right, Wrong, ineXact and Unsupported answers, as well as the percentage of correctly answered Factoids, Temporally restricted questions, Definition and List questions.

Run	R #	W #	X #	U #	%F [145]	%T [23]	%D [39]	L% [16]	NIL		CWS	Overall accuracy
									#	% [*]		
ixag081eueu	26	163	11	0	15.9	8.7	7.7	0	4	7.0	0.023	13
ixag081eneu	11	182	7	0	5.5	4.3	7.7	0	6	6.2	0.004	5.5
ixag081eseu	11	182	7	0	6.9	4.3	2.6	0	4	4.8	0.004	5.5
ixag082eseu	7	185	8	0	4.8	4.3	0	0	3	3.5	0.003	3.5

Table 15: Evaluation of the results for the Basque target.

The monolingual run (ixag081eueu.xml) achieved accuracy of 13%, lower than the most systems for other target languages during the evaluation of 2007 but better than some of them. It is necessary to underline that Basque is a highly flexional language, which makes the matching of term and entities more complex, and that this is the first participation. The system achieved better accuracy in factoid questions (15.9%). No correct answers was retrieved for list questions. 57 answers were NIL (only four of them were corrects), it is to be expected that the participants could improve this aspect.

Looking at the cross-lingual runs the loss of accuracy with respect to the monolingual system is slightly more than 50% for the two best runs. This percentage is quite similar to runs for other target languages in 2007. The overall accuracy is the same for both (English and Spanish to Basque) but they only agree for five correct answers (each system gives six other correct answers). The second system for Spanish-Basque got poorer results and only is slightly better in inexact answers. These runs get also a lot of NIL answers.

5.5.2 Bulgarian as Target

Run	R #	W #	X #	U #	% F [*]	% T [*]	% D [*]	% L [*]	NIL		CWS	MRR	Overall accuracy
									#	% [*]			
btb1	20	173	7	0	8.80	7.14	25.00	0.00	-	0.00	0.01	-	10 %

Table 16: Evaluation of the results for the Bulgarian target.

This year, contrary to our optimistic expectations, only one run by one group (BTB) was performed for Bulgarian. As Table 16 shows, the result is far from satisfying. Again, the results for definitions were better than those for other question types. Also, the difference between the detection of factoids and of temporally restricted questions is negligible. The results from the previous years decreased in both directions – for participating groups and for system performance.

5.5.3 Dutch as Target

This year, only one team took part in the question answering task with Dutch as target language: the University of Groningen. The team submitted two monolingual runs and two cross-lingual runs (English to Dutch). All runs were assessed twice by a single assessor. This resulted in a total of eight conflicts (1%). These were corrected. The results of the assessment can be found in Table 17.

Run	R #	W #	X #	U #	%F [151]	%T [13]	%D [39]	L% [10]	NIL		CWS	Overall accuracy
									#	% [*]		
gron081nlnl	50	138	11	1	24.5	15.4	33.3	0.0	19	5.3	0.342	25.0
gron082nlnl	51	136	10	3	24.5	15.4	35.9	0.0	15	6.7	0.331	25.5
gron081ennl	27	157	10	6	13.2	7.7	17.9	0.0	30	3.3	0.235	13.5
gron082ennl	27	157	10	6	13.2	7.7	17.9	0.0	30	3.3	0.235	13.5

Table 17: Evaluation of the results for the Dutch target.

The two cross-lingual runs gron081ennl and gron082ennl produced exactly the same answers.

The best monolingual run (gron082nlnl) achieved exactly the same score as the best run of 2007 (25.5%). The same is true for the best monolingual run (13.5%). The fact that the two scores are in the same range as last year is no big surprise since the task has not changed considerably this year and all scores have been achieved by the same system.

Like in 2007, the system performed better for definition questions than for other question types. The definition questions could be divided into two subtypes: those that asked for a definition (26) and those that contained a definition and asked for the name of the defined object (12). The monolingual runs performed similarly for both subtypes but the cross-lingual runs did not contain a correct answer to any question of the second subtype.

None of the runs obtained any points for the list questions. The answers contained some parts that were correct but none of them were completely correct. We were unable to award points for partially correct answers in the current assessment scheme.

All the runs were produced by the same system and the differences between the runs are small. The cross-lingual runs contained seven correct answers that were not present in any of the monolingual runs (for questions 20, 25, 120, 131, 142, 150 and 200). Eight questions were only answered correctly in a single monolingual run (1, 28, 54, 72, 83, 143, 193 and 199). Thirty-five questions were answered correctly in two runs, three in three runs and seventeen in all four runs. 137 questions failed to receive any correct answer.

5.5.4 English as Target

Five cross-lingual runs with English as target were submitted this year, as compared with eight in 2007 and thirteen in 2006. Four groups participated in three languages, Dutch, German and Romanian. Each group worked with only one source language, and only DCUN submitted two runs. The rest submitted only one run.

Run	R #	W #	X #	U #	% F [160]	% T [12]	% D [30]	% L [10]	NIL		CWS	K1	Overall accuracy
									#	%[0]			
dcun081deen	16	168	7	9	5.00	8.33	26.67	0.00	0	0.00	0.00516	0.10	8.00
dcun082deen	1	195	3	1	0.63	0.00	0.00	0.00	0	0.00	0.00013	0.03	0.50
dfki081deen	28	164	5	3	6.25	8.33	60.00	0.00	0	0.00	0.01760	N/A	14.00
ilkm081nlen	7	182	2	9	4.38	0.00	0.00	0.00	0	0.00	0.00175	N/A	3.50
wlvs081roen	38	155	2	5	11.25	0.00	66.67	0.00	0	0.00	0.05436	0.13	19.00

Table 18: Evaluation of the results for the English target.

Of the five runs with English as target, wlvs081roen was the best with an accuracy of 19.00% overall. They also did very well on the definitions, scoring 66.67%. The only source language for which there

was more than one run was German, for which there were three submissions from two groups. dfki081 scored the best with 14.00% and this was followed by dcun081deen with 8.00% and dcun082deen with 0.50%. DFKI also did very well on definitions with an accuracy of 60.00. Interestingly, none of the systems answered any of the list questions correctly. Only dcun082deen answered one list question inexactly.

If we compare the results this year with those of last year when the task was very similar, performance has improved here. The best score in 2007 was wolv071roen with 14.00% (the best score) which has now improved to 19.00%. Similarly, dfki071deen scored 7.00% in 2007 but increased this to 14.00% this year in dfki081deen. An attempt was made to set easier questions this year, which might have affected performance. In addition, many more questions came from the Wikipedia in 2008 with only a minority being drawn from the newspaper corpora.

5.5.5 French as Target

This year only one group took part in the evaluation tasks using French as a target language: the French group *Synapse Développement.* Last year's second participant, the *Language Computer Corporation* (LCC, USA) didn't send any submission this time.

Synapse submitted three runs in total:

- one monolingual run: French to French (FR-to-FR),
- two bilingual runs: English-to-French (EN-to-FR) and Portuguese-to-French (PT-to-FR).

Run	Assessed Answers (#)	R #	W #	X #	U #	%F [135]	%T [66]	%D [30]	L% [35]	NIL Answers		CWS	Overall accuracy
										#	% [12]		
syn08 frfr	200	131	77	9	1	54.8	51.5	86.7	37.1	20	50.0	0.30937	56.5
syn08 enfr	200	36	157	6	1	15.6	15.1	50.0	0.0	60	8.3	0.02646	18.0
syn08 ptfr	200	33	163	4	0	14.1	13.6	43.3	2.9	67	11.9	0.02387	16.5

Table 19: Evaluation of the results for the French target.

For the monolingual task, the Synapse system returned 113 correct answers (accuracy of 56.5%), slightly more than last year (accuracy of 54.0%). The bilingual runs performance is quite low, with an accuracy of 18.0% for EN-to-FR and 16.5% for PT-to-FR. It cannot be fairly compared to the results of CLEF2007, because Synapse didn't submit bilingual runs last year. Last year, LCC obtained an accuracy of 41.7% for EN-to-FR, but did not submit anything this year.

It appears that the level of performance strongly depends on the type of questions. The monolingual run scores very high on the definition questions (86.7%). The lowest performance is obtained with closed list questions (37.1%).

It is even more obvious when looking at the bilingual runs. If the systems performed pretty well on the definition questions (50.0% and 43.3% for EN-to-FR and PT-to-FR respectively), they could not cope with the closed list questions. The PT-to-FR system could only give one close list correct answer. The EN-to-FR system could not even answer to any of these questions. The bilingual runs did not reach high accuracy with factoid and temporally restricted questions (50.0% and 43.3% for EN-to-FR and PT-to-FR respectively). This year, the complexity of the task, in particular regarding closed list questions, seems to have been hard to cope with for the bilingual systems.

The complexity of the task is also reflected by the number of NIL answers. The monolingual system returned 20 NIL answers (to be compared with the 12 expected). The bilingual systems returned 60 (EN-to-FR) and 67 (EN-to-FR) NIL answers, i.e. at least 5 times more as expected.

It is also interesting to look at the results when categorizing questions by the size of the topic they belong to. This year, topics could contain from 1 single question to 4 questions. The CLEF 2008 set consists of: 52 single question topics; 33 topics with 2 questions (66 questions in total); 18 topics with 3 questions (54 questions in total) ;7 topics with 4 questions (28 questions in total).

The monolingual system is not sensitive to the size of the topic question set. On the contrary, the performances of the bilingual runs decrease by a half, when comparing the 1- and 2-question sets to the 3- and 4-question sets. A possible explanation is that the bilingual systems perform poorly with questions containing anaphoric references (which are more likely to occur in the 3- and 4-question sets).

In conclusion, there was unfortunately only one participant for French this year. In particular; it would have been interesting to see how the LCC group, which submitted a bilingual run last year, would have performed this year. This decrease in participation can be explained by the discouragement of some participants. Some have complained that the task is each year harder (e.g. this year, there were more closed list questions and anaphoric references than last year) which can result in a decrease in the systems performances.

This year, the number and complexity of closed list questions was clearly higher than the previous year. In the same way, there were more temporally restricted questions, more topics (comprising from 2 to 4 questions) and more anaphoric references. It seems that this higher level of difficulty particularly impacted the bilingual tasks. In spite of this, the monolingual Synapse system performed slightly better than last year.

5.5.6 German as Target

Three research groups submitted runs for evaluation in the task having German as target language: The German Research Center for Artificial Intelligence (DFKI), Fern Universität Hagen (FUHA) and Universität Koblenz-Landau (LOGA). All groups provided system runs for the monolingual scenario, DFKI and FUHA submitted runs for the cross-language English-German scenario and FUHA also had runs for the Spanish-German scenario.

Run	R # 	W # 	X # 	U # 	% F [160] 	% T [9] 	% D [30] 	% L [10] 	NIL # 	NIL % [10] 	CWS 	MRR 	Overall accuracy
dfki081dede$_M$	73	119	2	6	30.62	44.44	80	0	0	0	0.16	0	36.5
dfki082dede$_M$	74	120	2	4	31.25	33.33	80	0	0	0	0.16	0	37
fuha081dede$_M$	45	141	8	6	24.37	44.44	20	0	1	4.76	0.05	0.29	22.5
fuha082dede$_M$	46	139	11	4	25.62	33.33	16.66	0	21	4.76	0.048	0.29	23
loga081dede$_M$	29	159	11	1	13.75	0	20	10	55	5.45	0.031	0.19	14.5
loga082dede$_M$	27	163	9	1	13.12	0	16.66	10	48	4.16	0.029	0.17	13.5
dfki081ende$_C$	29	164	2	5	10	0	43.33	0	0	0	0.038	0	14.5
fuha081ende$_C$	28	163	6	3	15	11.11	13.33	0	81	7.4	0.023	0.24	14
fuha082ende$_C$	28	160	6	6	15	11.11	13.33	0	81	7.4	0.019	0.22	14
fuha081esde$_C$	19	169	9	2	9.43	0	13.33	0	9	0	0.015	0.15	9.54
fuha082esde$_C$	17	173	5	5	8.12	0	13.33	0	61	3.27	0.007	0.13	8.5

Table 20: Evaluation of the results for the German target.

Compared to the previous editions of the evaluation forum, this year an increase in the accuracy of the best performing system and of an aggregated virtual system for monolingual and a decrease in the accuracy of the best performing system and of an aggregated virtual system for cross-language tasks was registered. The details of systems' results can be seen in Table 20.

5.5.7 Portuguese as Target

The Portuguese track had six different participants: beside the veteran groups of Priberam, Linguateca, Universidade de Évora, INESC and FEUP, we had a new participants this year, Universidade Aberta. No bilingual runs were submitted this year.

In this fourth year of Portuguese participation, Priberam repeated the top place of its previous years, with University of Évora second. Again we added the classification X-, meaning incomplete, keeping the classification X+ for answers with extra text or other kinds of inexactness. In Table 21 we present the overall results .

Run Name	R (#)	W (#)	X+ (#)	X- (#)	U (#)	Overall Accuracy (%)	NIL Accuracy		
							#	Precision (%)	Recall (%)
diue081	93	94	8	1	2	46.5%	21	9.5	20
esfi081	47	134	5	7	5	23.5%	20	20.0	20
esfi082	39	137	7	9	6	19.5%	20	15.0	10
feup081	29	165	2	2	2	14.5%	142	8.5	90
feup082	25	169	3	1	2	12.5%	149	8.1	90
idsa081	65	119	8		8	32.5%	12	16.7	20
ines081	40	150	2	1	5	20.0%	123	9.7	90
ines082	40	150	2	1	5	20.0%	123	9.7	90
prib081	127	55	9	3	4	63.5%	8	12.5	10

Table 21: Evaluation of the results for the Portuguese target.

On the whole, compared to last year, Priberam and Senso (UE) improved their results, which were already the best. The INESC system and Esfinge (Linguateca) also showed some improvement, at a lower level Raposa (FEUP) showed similar results. The system of Universidade Aberta appeared with good results compared to some veteran systems. We leave it to the participants to comment on whether this might have been caused by harder questions or changes (or lack thereof) in the systems.

5.5.8 Romanian as Target

In the third year of Romanian participation in QA@CLEF, and the second one with Romanian addressed as a target language, the question generation was based on the collection of Wikipedia Romanian pages frozen in November 2006[27]- the same corpus as in the previous edition[28].

As in the 2007 edition, this year two Romanian groups took part in the monolingual task with Romanian as a target language: the Faculty of Computer Science from the Al. I. Cuza University of Iasi (UAIC), and the Research Institute for Artificial Intelligence from the Romanian Academy (ICIA), Bucharest. Each group submitted two runs, the four systems having an average of 2.4 answers per question for ICIA, and 1.92 for UAIC. The 2008 general results are presented in Tables 31 below.

The statistics includes a system, named *combined*, obtained through the combination of the 4 participating RO-RO systems. Because at the evaluation time we observed that there are correct

[27] See `http://static.wikipedia.org/downloads/November_2006/ro/`

[28] At `http://static.wikipedia.org/downloads/` the frozen versions of Wikipedia exist for April 2007 and June 2008, for all languages involved in QA@CLEF.

answers not only in the first position, but also on the second or the third, the *combined* system considers that an answer is R if there exists at least one R answer among all the answers returned by the four systems. If there is no R answer, the same strategy is applied to X, U and finally W answers. This "ideal" system permits to calculate the percentage of the questions (and their type), answered by at least one of the four systems in any of the maximum 3 answers returned for a question.
All three systems crashed on the LIST questions. The best results were obtained by ICIA for DEFINITION questions, whereas UAIC performed best with the FACTOID questions. The *combined* system suggests that a joint system, developed by both groups, would improve substantially the general results for Romanian.

Run	R	W	X	U	F	T	D	L	NIL		CWS	MRR	Overall accuracy
	#	#	#	#	[162]	[47]	[28]	[10]	#	% [8]			
icia081roro	10	179	11	0	4.938	8.511	7.143	0.0	15	6.667	0.00812	0.08217	5.0
icia082roro	21	168	11	0	6.173	8.511	39.286	0.0	15	6.667	0.02191	0.14319	10.5
uaic081roro	41	128	27	3	24.691	25.532	3.571	0.0	65	7.692	0.03679	0.34324	20.5
uaic082roro	45	125	26	4	26.543	27.660	3.571	10.0	64	9.375	0.04892	0.36799	22.5

Table 22: Evaluation of the results for the Romanian target.

5.5.9 Spanish as Target

The participation in the Spanish as Target subtask has decreased from 5 groups in 2007 to 4 groups this year. 6 runs were monolingual and 3 runs were crosslingual. **Table 23** shows the summary of systems results with the number of Right (R), Wrong (W), Inexact (X) and Unsupported (U) answers. The table shows also the accuracy (in percentage) of factoids (F), factoids with temporal restriction (T), definitions (D) and list questions (L). Best values are marked in bold face.

Run	R	W	X	U	% F	% T	% D	% L	NIL	F	CWS	MRR	Overall accuracy
	#	#	#	#	[124]	[36]	[20]	[20]	#	[10]			
prib081eses	**86**	105	5	4	**41,13**	**41,67**	75	**20**	3	0,17	0,178	**0,4483**	**42,5**
inao082eses	44	152	3	1	19,35	8,33	80	5	4	0,10	0,068	0,2342	22
inao081eses	42	156	1	1	15,32	8,33	**95**	5	3	0,13	0,053	0,2375	21
qaua082eses	39	156	4	1	22,58	13,89	30	-	6	0,15	0,041	0,2217	19,5
mira081eses	32	156	3	9	12,90	2,78	75	-	3	0,21	0,032	0,1766	16
mira082eses	29	159	3	9	11,29	2,78	70	-	3	0,23	0,026	0,1591	14,50
qaua081enes	25	173	-	2	11,29	16,67	20	5	6	0,19	0,011	0,1450	12,50
qaua082enes	18	176	3	3	9,68	8,33	15	-	8	0,15	0,006	0,1108	9
mira081fres	10	185	2	3	5,65	-	15	-	3	0,12	0,008	0,0533	5

Table 23: Evaluation of the results for the Spanish target.

5.6 Overall Comments

This year we proposed the same evaluation setting as in 2007 campaign. In fact, last year the task was changed considerably and this affected the general level of results and also the level of participation in the QA task. This year participation increased slightly but the task proved to be still very difficult. Wikipedia increased its presence as a source of questions and answers. Following last year's conclusions Wikipedia seemed to be a good source for finding answers to simple factoid questions.

Moreover, the overall decrease in accuracy was probably due to linked questions. This fact confirms that topic resolution is a weak point for QA systems.

Only 5 out of 11 target languages had more than one different participating group. Thus from the evaluation methodology perspective, a comparison between systems working under similar circumstances cannot be accomplished and this impedes one of the major goals of campaigns such as QA@CLEF, i.e. the systems comparison which could determine an improvement in approaching QA problematic issues.

After 6 years, a lot of resources and know-how have been accumulated. However, the tasks offered have proved to be difficult for the systems which have not shown a very good overall performance, even those that have participated year by year. In addition, a result of offering so many language possibilities has meant that there have always been very few systems participating in the same task, with the same languages. This has meant that comparative analysis is extremely problematic. Moreover, discussion has also concerned the collections themselves since it is easier to find questions with IR techniques in Wikipedia than in the other collections and they do not enforce any user model.

Consequently, the QA organisers have decided to redefine the task for CLEF 2009 to permit the evaluation and comparison of systems even when they are working in different languages. Therefore, in the ResPubliQA task the JRC collection of European treaties will be used and the task will be simplified and made closer to passage retrieval. The new setting will also take as reference a real user scenario, in this document collection in which multilinguality is more natural and comparative analysis is less problematic.

6 CLEF 2008 in the TrebleCLEF Context

The previous sections discussed and analysed in detail the achievements and the results of four tracks of the CLEF 2008 and presented some future directions for these tracks, as gained from the experience with this year campaign.

Here we look at the CLEF 2008 campaign from a different perspective by placing it in the context of some of the main activities organized and carried out by the TrebleCLEF project during 2008. TrebleCLEF not only aims at distilling the knowledge gathered during the CLEF campaigns and transferring it to relevant application, developer, and user communities but also strives to obtain input from those communities in order to improve the design of the CLEF campaign and activate positive feedback cycles.

6.1 The Evaluation Experts Viewpoint

On March 30th, 2008, TrebleCLEF organised the "Workshop on Novel Methodologies for Evaluation in Information Retrieval". The workshop had 11 refereed short and long papers along with 2 invited speakers. A common theme to a great many of the papers and invited talks was the importance of moving away from the traditional test collection measures and conventional test collection search tasks and documents. Many papers illustrated how distinct search tasks require distinct evaluation test beds and measures. Such work points to the challenges of evaluating information retrieval systems and the continual need for evaluation campaigns to investigate different search challenges and consider how best to evaluate systems.

Two further papers described the challenges of building collections to evaluate related tasks such as web link analysis. There was also a paper describing a tool to interactively measure and visualise search results encouraging researchers to engage in a more exploratory approach to assessing experimental results.

This workshop provided valuable input and suggested new angles to be taken into consideration when designing the tasks of the CLEF campaign. In particular, the discussions about how to create test beds and define measures more suitable to represent realistic tasks provided useful suggestions and hints for designing CLEF tasks such as the ImageCLEFphoto task which looks for diversity in the retrieval results.

6.2 The System Developers Viewpoint

Partners of TrebleCLEF organised a System Developer's Workshop (2nd - 3rd October 2008) at ZHAW (Winterthur, Switzerland) to provide a forum in which to discuss the more operational/applied aspects of CLEF; details about this workshop are given in Deliverable 3.1. A number of leading figures from both academic and commercial institutions participated in the event and provided helpful input regarding the operational impact of CLEF and test collections as a whole (e.g. realism of the evaluation resources with respect to use cases and scenarios, performance, verifiability, robustness, presentation of results, transferability of results to operational settings etc). A helpful distinction was presented by a researcher from Xerox between considering the needs of different individuals for the outputs of CLEF: researches (e.g. wanting to find the best algorithms), developers (e.g. interested in architectures, reusable components and services), and end users (e.g. wanting simple, reliable, usable and stable interfaces). This may help in the future planning of CLEF in defining the goals and outputs in terms of benefits to different communities.

Suggestions by participants of the workshop for future CLEF events included: to find realistic use cases for MLIA/CLIR (e.g. English call centre agents required to talk to Spanish customers), to focus on the science and not the systems (e.g. set tasks as sets of hypotheses/questions to address and have participants write up these rather than system descriptions – seen as less beneficial/useful to the developer), to assess the contribution of different technologies for MLIA/CLIR (e.g. NLP), to consider user interaction and user interfaces, to provide (and maintain) comprehensive lists of publicly-accessible resources (including access details, availability, coverage and performance in different contexts), to evaluate cross-language technologies with respect to databases (commonly used for IR in commercial contexts), to focus on domains other than news, to create documents detailing the design decisions to consider when developing MLIA/CLIR applications (e.g. decisions, process flows, best practices and component performance and dependencies), to provide more accessible and synthesised versions of the CLEF proceedings (digestible for the developer), and to assess the success of MT systems (e.g. on various object types and for different domains/contexts).

Some of these suggestions have been already taken into consideration, at least to some extent, in the design of the CLEF tracks. For example, the recommendation to consider user interaction and user interfaces is the core objective of the iCLEF track, presented in Section 4.

Others will be addressed in the forthcoming CLEF 2009 campaign. This is the case, for example, of the advice about assessing the contribution of different technologies to MLIA/CLIR. To this end, in 2009, a new pilot task, called Grid@CLEF[29], will be organised. The objectives will be to look at differences across a wide set of languages; to identify best practices for each language; and to help other countries to develop their expertise in the IR field and create IR groups.

6.3 The Users Viewpoint

Treble-CLEF organized a workshop in June 2008 bringing together system developers and current/potential user communities of MLIA services, with the goals of (i) providing input for the Treble-CLEF (user-oriented) best-practices report and (ii) bringing CLEF closer to current/potential user communities for MLIA services by identifying user communities and application scenarios, increasing the awareness of CLEF results in user communities, and helping to shape future iCLEF campaigns (and possibly other CLEF tasks as well); details about this workshop are given in Deliverable 3.2.

An objective of this workshop was to reach a common vision on two specific issues: (i) features that an MLIA system should have from the users' perspective and (ii) strategies to provide MLIA technology with these features and transfer these technologies and the related know-how to the society and relevant application and developer communities, considering the use of evaluation forums such as CLEF.

[29] See `http://ims.dei.unipd.it/gridclef/`

The input from user communities was very valuable: in particular, a full list of desirable features for MLIA systems emerged as a consensus of all workshop participants, and will be reflected in the best practices deliverable 3.3 which is due for June 2009.

In addition, the workshop produced a few immediate consequences for the CLEF campaign:

- An initiative to involve user communities by including representatives in the CLEF steering committee. Denis Teysou (AFP) and Bruno Pouliquen (JRC) were invited to become part of the committee, and are actually active members of it.
- Tighter bonds with the patent searching community were established, via the Information Retrieval Facility in Vienna and its representative, John Tait. This resulted in a proposal for a multilingual patent retrieval track at CLEF, which attracted attention both from CLEF participants and from patent experts, which attended CLEF in order to interact with CLEF researchers and shape a both practical and challenging task.
- At least one concrete possibility for additional iCLEF scenarios was detected; JRC offered the possibility of using an MLIA search log with around 1M interactions per day for a log analysis task at CLEF. Search log analysis in general – the new methodological approach of iCLEF – was regarded by the user communities representatives as a fruitful way to go.

6.4 The European (Digital) Library Scenario

The quality of the services and documents a digital library supplies are very important. In particular, when the users of a digital library system come from different cultural and social groups, understanding the different categories of users becomes a key point in order to offer the best service possible. The evaluation process of the quality of a digital library system together with the analysis of the preferences of users that use a digital library system can be observed implicitly by means of log data and explicitly by means of user studies. In fact, the insights gained by analyzing log data together with data from controlled studies are more informative than those obtained when the two types of studies are conducted alone.

In the European Digital scenario one reality of particular interest is The European Library (TEL) because it is a service intended to give access to the worldwide users to the Cultural Heritage artifacts that are maintained in the National Libraries of all Europe. In the context of the TELplus[30] project, funded by the European Commission under the *eContentplus* Programme, one of the main tasks concerns the analysis of user preferences and how the functionalities of The European Library portal are perceived by its users.

This kind of analysis require complementary strategies: the analysis of log data and the analysis of data collected through user studies. The study designed to carry out this cross-analysis was conducted by UNIPD in a controlled setting during the end of 2007 and the beginning of 2008. A Questionnaire consisting of 17 main points organized in a pdf form was prepared to gather user preferences explicitly, students were asked to fill-in the questionnaire while browsing the portal and to send it back by email. Log data were automatically recorded by the Web server of The European Library, with the aim of studying user preferences implicitly and validate the results obtained by questionnaires. All students, who were not previously aware of the existence of The European Library Web portal, were told to act freely, browse the Web site without any restriction, and not feel obliged to fill-in all the parts of the questionnaire. 216 students from three different faculties of the University of Padua, that is the faculties of Humanities, Statistics, and Psychology, participated in this user study.

Many different analyses were carried out, among the others: length of sessions, number of queries, and search actions. Here we briefly summarize some points related to the use of languages in the search of library catalogues:

- the first thing a user does when entering The European Library portal in 50% of the cases is to change the language of the interface to their mother tongue. This is an important aspect to bear in

[30] See `http://www.theeuropeanlibrary.org/telplus/index.php`

mind when designing multilingual information access tools, users may feel insecure when using a different language; this was true even for the group of students who studied languages and participated in this study;
- the same people who change the language of the interface perform actions which involve the use of different languages: browsing collections of documents of National Libraries of countries different from the country they are from, querying and searching for documents in different languages;
- there is other evidence of this fact which can be observed indirectly from another point of view: pictures and images are very appealing for users even though these documents are described in a variety of languages. Both the "treasures" section and the "exhibition" section were deeply browsed by users even before trying any query to the portal.

This brief summary of the results of the analysis carried out on log files together with questionnaires showed that in a real context, like searching digital library catalogues in The European Library, there is indeed the need for multilingual information access tools; therefore, it is important to think about advanced techniques of interaction with the user which make the search easier, for example classification of queries and query suggestions in one or more languages, clustering of the results, and so on.

For this reason, after the experience gained with The European Library portal, there has been an intense and fruitful collaboration between The European Library staff and CLEF. In particular, the Ad-Hoc TEL track of CLEF 2008 was designed in order to test multilingual information access tools on library catalogues (details of this track are presented in Section 3.1) and it will be offered again in 2009 in order to deeply analyse the interaction between semi-structured catalog records and multilinguality.

In 2009 a new track of CLEF, called LogCLEF, has been designed in order to continue and improve the analysis of log files of The European Library, opening to the research community the opportunity to have real data to work on multilingual personalization problems. LogCLEF will deal with the analysis of queries as expression of user behaviour and the goal is the analysis and classification of queries in order to improve search systems. There will be a particular task for this track called Log Analysis for Digital Societies which intends to analyze user behaviour with a focus on multilingual search. This task is expected to be successful since it may reach different potential targets like query reformulation, multilingual search behaviour and community identification. The evolution and results of this new track are going to be reported and documented during the CLEF 2009 campaign in the LogCLEF 2009 Web site[31].

7 Final Remarks

We have presented the activities and the results achieved in four main tracks of the CLEF 2008 campaign: Ad Hoc, ImageCLEF, iCLEF, and QA@CLEF.

We have shown how a considerable effort has been made to design and implement tasks within these tracks that match the objectives of TrebleCLEF, emulating real-world scenarios as far as possible and promoting the production of experimental results that can be transferred to relevant application and developer communities:

- The Ad hoc track offered the TEL tasks in the context of an operational European Digital Library with the aim of investigating the most efficient and effective algorithms to search sparse, multilingual documents;
- The ImageCLEF track promoted both the development of systems which provide diversity of results in the visual scenario, an important theme for both research and the industrial search engine developers; and also tasks aimed at an important application community – that involved in medical image processing;

[31] See `http://www.uni-hildesheim.de/logclef/`

- The iCLEF track offered an online multilingual search game based on the Flickr database of photos which attracted a large number of players and enable the acquisition of a large volume of log data for the study of user search behavior in a multilingual environment, according to language competence;
- The QA@CLEF track attempted to address an important research question – how to handle series of questions often containing anaphora – again very much representing a real-world situation, even though this problem proved to be hard for current cross-language question answering systems.

A significant set of results have been obtained from these tasks and have been analysed as reported in this document; this data will however be made available to the scientific community for further studies. However, we feel that the community process that led to these results is in itself an important achievement. Each track comported the active involvement of diverse research communities (from the domains of IR, Image processing, HCI, NLP, Library Science,…) with lively discussions with respect to the objectives, design and outcomes, often producing innovative ideas and/or solutions. In addition, this participation has led to both the re-design of some tasks and the introduction of new ones for the CLEF 2009 campaign.

Acknowledgements

CLEF is organised on a distributed basis with different research groups being responsible for the running of the various tracks. We should like to express out gratitude to all those who have been involved in the coordination of this activity over the years. A complete list of all the organizations involved in the coordination of CLEF can be found on the homepage of the CLEF website at http://www.clef-campaign.org/.

References

[1] Aghazade, Z., Dehghani, N., Farzinvash, L., Rahimi, R., AleAhmad, A., Amiri, H., and Oroumchian, F. (2008). Fusion of Retrieval Models at CLEF 2008 Ad-Hoc Persian Track. In Borri, F., Nardi, A., and Peters, C., editors, *Working Notes for the CLEF 2008 Workshop*. Published online at http://www.clef-campaign.org/2008/working_notes/CLEF2008WN-Contents.html.

[2] Agirre, E., Di Nunzio, G. M., Ferro, N., Mandl, T., and Peters, C. (2008). CLEF 2008: Ad Hoc Track Overview. In Borri, F., Nardi, A., and Peters, C., editors, *Working Notes for the CLEF 2008 Workshop*. http://www.clef-campaign.org/2008/working_notes/adhoc-final.pdf.

[3] Agirre, E. and Lopez de Lacalle, O. (2007). UBC-ALM: Combining k-NN with SVD for WSD. In Agirre, E., Màrquez, L., and Wicentowski, R., editors, *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval 2007)*, pages 341-345. The Association for Computational Linguistics (ACL), Stroudsburg (PA), USA.

[4] AleAhmad, A., Kamalloo, E., Zareh, A.,Rahgozar, M., and Oroumchian, F. (2008). Cross Language Experiments at Persian@CLEF 2008. In Borri, F., Nardi, A., and Peters, C., editors, *Working Notes for the CLEF 2008 Workshop*. Published online at http://www.clef-campaign.org/2008/working_notes/CLEF2008WN-Contents.html.

[5] Arni, T., Clough, P., Sanderson, M., and Grubinger, M. (2008). Overview of the ImageCLEFphoto 2008 Photographic Retrieval Task. In Borri, F., Nardi, A., and Peters, C., editors, *Working Notes for the CLEF 2008 Workshop*. http://www.clef-campaign.org/2008/working_notes/ImageCLEFphoto2008-final.pdf.

[6] Arni, T., Tang, J., Sanderson, M., and Clough, P. (2008). Creating a Test Collection to Evaluate Diversity in Image Retrieval. In Bennett, P.N., Carterette, B., Chapelle, O., and Joachims, T., editors, *Proc. SIGIR 2008 Workshop on Beyond Binary Relevance: Preferences, Diversity, and Set-Level Judgments*, pages 15-21.

[7] Braschler, M. (2004). Combination Approaches for Multilingual Text Retrieval. *Information Retrieval*, 7(1/2):183–204.

[8] Chan, Y.S., Ng, H.T., and Zhong, Z. (2007). NUS-PT: Exploiting Parallel Texts for Word Sense Disambiguation in the English All-Words Tasks. In Agirre, E., Màrquez, L., and Wicentowski, R., editors, *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval 2007)*, pages 253-256. The Association for Computational Linguistics (ACL), Stroudsburg (PA), USA.

[9] Cleverdon, C.W. (1997). The Cranfield Tests on Index Languages Device. In Spärck Jones, K. and Willet, P., editors, *Readings in Information Retrieval*, pp. 47-60. Morgan Kaufmann Publisher, Inc., San Francisco, CA, USA.

[10] Clinchant, S., Renders, J.-M. (2008). XRCE's Participation to CLEF 2008 Ad-Hoc Track. In Borri, F., Nardi, A., and Peters, C., editors, *Working Notes for the CLEF 2008 Workshop*. Published online at http://www.clef-campaign.org/2008/working_notes/CLEF2008WN-Contents.html.

[11] Clough, P. (2007), Large-scale evaluation of cross-language image retrieval systems, ASIST Bulletin, Feb-Mar 2007.

[12] Crivellari, F., Di Nunzio, G. M., and Ferro, N. (2008). A Statistical and Graphical Methodology for Comparing Bilingual to Monolingual Cross-Language Information Retrieval. In Agosti, M., editor, *Information Access through Search Engines and Digital Libraries*, pages 171–188. Springer-Verlag, Heidelberg, Germany.

[13] Crivellari, F., Di Nunzio, G. M., and Ferro, N. (2007). How to Compare Bilingual to Monolingual Cross-Language Information Retrieval. In Amati, G., Carpineto, C., and Romano, G., editors, *Advances in Information Retrieval. Proc. 29th European Conference on IR Research (ECIR 2007)*, pages 533–540. Lecture Notes in Computer Science (LNCS) 4425, Springer, Heidelberg, Germany.

[14] Di Nunzio, G. M. and Ferro, N. (2008). Appendix A: Results of the TEL@CLEF Task. In Borri, F., Nardi, A., and Peters, C., editors, *Working Notes for the CLEF 2008 Workshop*. http://www.clef-campaign.org/2008/working_notes/AppendixA.pdf.

[15] Di Nunzio, G. M. and Ferro, N. (2008). Appendix B: Results of the Persian@CLEF Task. In Borri, F., Nardi, A., and Peters, C., editors, *Working Notes for the CLEF 2008 Workshop*. http://www.clef-campaign.org/2008/working_notes/AppendixB.pdf.

[16] Di Nunzio, G. M. and Ferro, N. (2008). Appendix C: Results of the Robust Task. In Borri, F., Nardi, A., and Peters, C., editors, *Working Notes for the CLEF 2008 Workshop*. http://www.clef-campaign.org/2008/working_notes/AppendixC.pdf.

[17] Dolamic, L., Fautsch, C., and Savoy, J. (2008). UniNE at CLEF2008: TEL, Perisan and Robust IR. In Borri, F., Nardi, A., and Peters, C., editors, *Working Notes for the CLEF 2008 Workshop*. Published online at http://www.clef-campaign.org/2008/working_notes/CLEF2008WN-Contents.html.

[18] Ferro, N. and Peters, C. (2008). From CLEF to TrebleCLEF: the Evolution of the Cross-Language Evaluation Forum. In Kando, N. and Sugimoto, M., editors, *Proc. 7th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-Lingual Information Access*, pages 577-593. National Institute of Informatics, Tokyo, Japan.

[19] Gonzalo, J., Clough, P., and Karlgren, J. (2008). Overview of iCLEF 2008: Search Log Analysis for Multilingual Image Retrieval. In Borri, F., Nardi, A., and Peters, C., editors, *Working Notes for the CLEF 2008 Workshop*. http://www.clef-campaign.org/2008/working_notes/iclef_overview.pdf.

[20] Gonzalo, J., Clough, P., and Vallin, A. (2006). Overview of the CLEF 2005 Interactive Track. In Peters, C., Gey, F.C., Gonzalo, J., Jones, G.J.F., Kluck, M., Magnini, B., Müller, H., and de

Rijke, M., editors, *Accessing Multilingual Information Repositories: Sixth Workshop of the Cross--Language Evaluation Forum (CLEF 2005). Revised Selected Papers*, 251-262. Lecture Notes in Computer Science (LNCS) 4022, Springer, Heidelberg, Germany.

[21] Grubinger, M., Clough, P., Hanbury, A. and Müller, H. (2008). Overview of the ImageCLEFphoto 2007 Photographic Retrieval Task . In Peters, C., Jijkoun, V., Mandl, T., Müller, H., Oard, D. W., Peñas, A., Petras, V., and Santos, D., editors, *Advances in Multilingual and Multimodal Information Retrieval: Eighth Workshop of the Cross-Language Evaluation Forum (CLEF 2007). Revised Selected Papers*, 433-444. Lecture Notes in Computer Science (LNCS) 5152, Springer, Heidelberg, Germany.

[22] Grubinger, M., Clough, P., Müller, H. and Deselaers, T. (2006) The IAPR TC-12 Benchmark: A New Evaluation Resource for Visual Information Systems, In Grefenstette, G., Sanderson, M., and Preteux, F., editors, *Proceedings of International Workshop OntoImage'2006 Language Resources for Content-Based Image Retrieval, held in conjunction with LREC'06*, pages 13-23.

[23] Hartrumpf, S., Glöckner, I., and Leveling, J. (2008). University of Hagen at QA@CLEF 2007: Coreference Resolution for Questions and Answer Merging by Validation. In Peters, C., Jijkoun, V., Mandl, T., Müller, H., Oard, D. W., Peñas, A., Petras, V., and Santos, D., editors, *Advances in Multilingual and Multimodal Information Retrieval: Eighth Workshop of the Cross-Language Evaluation Forum (CLEF 2007). Revised Selected Papers*, 269-272. Lecture Notes in Computer Science (LNCS) 5152, Springer, Heidelberg, Germany.

[24] Herrera, J., Peñas A., Verdejo, F. (2005). Question Answering Pilot Task at CLEF 2004. In: Peters, C., Clough, P., Gonzalo, J., Jones, Gareth G.J.F., Kluck, M., and Magnini, B. editors. *Multilingual Information Access for Text, Speech and Images: Fifth Workshop of the Cross--Language Evaluation Forum (CLEF 2004) Revised Selected Papers*, pages 581–590. Lecture Notes in Computer Science (LNCS) 3491, Springer, Heidelberg, Germany.

[25] Ion, R. (2007) Word Sense Disambiguation Methods Applied to English and Romanian. PhD thesis, Romanian Academy, Bucharest.

[26] Ion, R., Mititelu, V.B. (2006). Constrained Lexical Attraction Models. In Sutcliffe, G. and Goebel, R., editors. *Nineteenth International Florida Artificial Intelligence Research Society Conference*, pages 297-302. AAAI Press, Menlo Park (CA), USA.

[27] Jadidinejad, A.H:, Mohtarami, M., and Amiri, H. (2008). Investigation on Application of Local Cluster Analysis and Part of Speech Tagging on Persian Text. In Borri, F., Nardi, A., and Peters, C., editors, *Working Notes for the CLEF 2008 Workshop*. Published online at http://www.clef-campaign.org/2008/working_notes/CLEF2008WN-Contents.html.

[28] Karimpour, R., Ghorbani, A., Pishdad, A., Mohtarami, M., AleAhmad, A., Amiri, H., and Oroumchian, F. (2008). Using Part of Speech tagging in Persian Information Retrieval. In Borri, F., Nardi, A., and Peters, C., editors, *Working Notes for the CLEF 2008 Workshop*. Published online at http://www.clef-campaign.org/2008/working_notes/CLEF2008WN-Contents.html.

[29] Kuersten, J., Wilhelm, T., and Eibl, M. (2008). CLEF 2008 Ad-Hoc Track: On-line Processing Experiments with Xtrieval. In Borri, F., Nardi, A., and Peters, C., editors, *Working Notes for the CLEF 2008 Workshop*. Published online at http://www.clef-campaign.org/2008/working_notes/CLEF2008WN-Contents.html.

[30] Landis, J.R., and Koch, G.G. (1997). The measurements of observer agreement for categorical data. *Biometrics*, 33:159–174.

[31] Larson, R. (2008). Logistic Regression for Metadata: Cheshire takes on Adhoc-TEL. In Borri, F., Nardi, A., and Peters, C., editors, *Working Notes for the CLEF 2008 Workshop*. Published online at http://www.clef-campaign.org/2008/working_notes/CLEF2008WN-Contents.html.

[32] Laurent, D., Séguéla, P., and Nêgre S. (2008). Cross Lingual Question Answering using QRISTAL for CLEF 2007. In Nardi, A., and Peters, C., editors, *Working Notes for the CLEF*

2007 Workshop. Published online at http://www.clef-campaign.org/2007/working_notes/CLEF2007WN-Contents.html.

[33] López-Ostenero, F., and Gonzalo, J., and Verdejo, F. (2005). Noun phrases as building blocks for cross-language Search Assistance. *Information Processing & Management.* 41(3): 549-568

[34] López-Ostenero, F., Peinado, V., Gonzalo, J., and Verdejo, F. (2008). Interactive question answering: Is Cross-Language harder than monolingual searching? *Information Processing & Management*. 44(1): 66-81.

[35] Machado, J., Martins, B., and Borbinha, J. (2008). Technical University of Lisbon CLEF 2008 Submission (TEL@CLEF Monolingual Task). In Borri, F., Nardi, A., and Peters, C., editors, *Working Notes for the CLEF 2008 Workshop*. Published online at http://www.clef-campaign.org/2008/working_notes/CLEF2008WN-Contents.html.

[36] MacNamee, P. (2008). JHU Ad Hoc Experiments at CLEF 2008. In Borri, F., Nardi, A., and Peters, C., editors, *Working Notes for the CLEF 2008 Workshop*. Published online at http://www.clef-campaign.org/2008/working_notes/CLEF2008WN-Contents.html.

[37] Magnini, B., Giampiccolo, D., Forner, P., Ayache, C., Jijkoun, V., Osenova, P., Peñas, A., Rocha, P., Sacaleanu , B., and Sutcliffe, R. (2007). Overview of the CLEF 2006 Multilingual Question Answering Track. In Peters, C., Clough, P., Gey, F.C., Karlgren, J., Magnini, B., Oard, D.W., de Rijke, M., and Stempfhuber, M., editors, Evaluation of Multilingual and Multi-modal Information Retrieval : Seventh Workshop of the Cross--Language Evaluation Forum (CLEF 2006). Revised Selected Papers, pages 223-256. Lecture Notes in Computer Science (LNCS) 4730, Springer, Heidelberg, Germany.

[38] Oard, D. W., Gonzalo, J., Sanderson, M., López-Ostenero, F., and Wang, J. (2004). Interactive Cross-Language Document Selection. *Information Retrieval*. 7(1-2): 205-228.

[39] Paskin, N., ed. (2006), *The DOI Handbook – Edition 4.4.1*. International DOI Foundation (IDF).

[40] Peinado, V., Artiles, J., Gonzalo, J., Barker, E., and López-Ostenero, F. (2008). FlickLing: a Multilingual Search Interface for Flickr. In Borri, F., Nardi, A., and Peters, C., editors, *Working Notes for the CLEF 2008 Workshop*. http://www.clef-campaign.org/2008/working_notes/peinado-paper_OC_CLE2008.pdf.

[41] Peñas, A., Rodrigo, Á., and Verdejo, F. (2008). Overview of the Answer Validation Exercise 2007. In Peters, C., Jijkoun, V., Mandl, T., Müller, H., Oard, D. W., Peñas, A., Petras, V., and Santos, D., editors, *Advances in Multilingual and Multimodal Information Retrieval: Eighth Workshop of the Cross-Language Evaluation Forum (CLEF 2007). Revised Selected Papers*, 237-248. Lecture Notes in Computer Science (LNCS) 5152, Springer, Heidelberg, Germany.

[42] Peters, C. (2001). Introduction. In Peters, C., editor, *Cross-Language Information Retrieval and Evaluation: Workshop of Cross-Language Evaluation Forum (CLEF 2000)*, pages 1-6. Lecture Notes in Computer Science (LNCS) 2069, Springer, Heidelberg, Germany.

[43] Sanderson, M., Tang, J., Arni, T., and Clough, P. (2009). Search Results Need to be Diverse. In *Proc. 32nd European Conference on IR Research (ECIR 2009)*. Lecture Notes in Computer Science (LNCS), Springer, Heidelberg, Germany (in print).

[44] Tomlinson, S. (2008). German, French, English and Persian Retrieval Experiments at CLEF 2008. In Borri, F., Nardi, A., and Peters, C., editors, *Working Notes for the CLEF 2008 Workshop*. Published online at http://www.clef-campaign.org/2008/working_notes/CLEF2008WN-Contents.html.

[45] Turmo, J., Comas, P., Ayache, C, Mostefa, D., Rosset, S., and Lamel, L. (2008). Overview of QAST 2007. In Peters, C., Jijkoun, V., Mandl, T., Müller, H., Oard, D. W., Peñas, A., Petras, V., and Santos, D., editors, *Advances in Multilingual and Multimodal Information Retrieval: Eighth Workshop of the Cross-Language Evaluation Forum (CLEF 2007). Revised Selected Papers*, 249-256. Lecture Notes in Computer Science (LNCS) 5152, Springer, Heidelberg, Germany.

[46] Vallin, A., Magnini, B., Giampiccolo, D., Aunimo, L., Ayache, C., Osenova, P., Peñas, A., de Rijke, M., Sacaleanu, B., Santos, D. and Sutcliffe, R. (2006). Overview of the CLEF 2005 Multilingual Question Answering Track. In Peters, C., Gey, F.C., Gonzalo, J., Jones, G.J.F., Kluck, M., Magnini, B., Müller, H., and de Rijke, M., editors, *Accessing Multilingual Information Repositories: Sixth Workshop of the Cross--Language Evaluation Forum (CLEF 2005). Revised Selected Papers*, pages 307-331. Lecture Notes in Computer Science (LNCS) 4022, Springer, Heidelberg, Germany.

[47] Voorhees, E.M. (2002). Overview of the TREC 2002 Question Answering Track. In Voorhees, E.M. and Buckland, L.P., editors, *The Eleventh Text REtrieval Conference (TREC 2002)*. National Institute of Standards and Technology (NIST), Special Pubblication 500-251, Washington, USA. Published online at http://trec.nist.gov/pubs/trec11/t11_proceedings.html.

[48] Westerveld, T. and van Zwol, R. (2007). The INEX 2006 Multimedia Track. In N. Fuhr, M. Lalmas, and A. Trotman (Eds), *Advances in XML Information Retrieval:Fifth International Workshop of the Initiative for the Evaluation of XML Retrieval (INEX 2006)*, Lecture Notes in Computer Science/Lecture Notes in Artificial Intelligence (LNCS/LNAI). Springer-Verlag, Berlin-Heidelberg, Germany.