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 2.3.2
Analysis of Evaluation Campaign Results: CLEF 2009

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

Organisation Name of Lead Contractor for this Deliverable: UNIPD

Version 1.00, December 2009 [final]

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

Document Information

Deliverable number:	2.3.2
Deliverable title:	Analysis of Evaluation Campaign Results – CLEF 2009
Due date of deliverable:	31/12/2009
Actual date of deliverable:	31/12/2009
Author(s):	Maristella Agosti, Martin Braschler, Paul Clough, Franco Crivellari, Giorgio Maria Di Nunzio, Nicola Ferro, Julio Gonzalo, Djamel Mostefa, Anselmo Peñas, Carol Peters, Mark Sanderson
Participant(s):	ELDA, 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	24/11/2009	Draft	UNIPD	First draft and outline circulated to all partners
0.20	16/12/2009	Draft	All	Contributions from all partners integrated
1.00	21/12/2009	Final	UNIPD	Final version with integrated comments from all partners

Abstract

This document reports and analyses the results of the CLEF 2009 campaign and, specifically, deals with the results and findings of the main tracks of CLEF which are organized by TrebleCLEF partners: 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; the QA@CLEF track, which covers multiple language question answering; the INFILE track, which concentrates on multilingual information filtering; the LogCLEF track, which copes with log analysis from search engine and digital library logs; and, the Grid@CLEF track, which runs systematic multilingual experiments. Moreover, the document discusses these tracks and their achievements in the light of the main activities conducted during the second year of the TrebleCLEF project.

Table of Contents

Executive Summary

The Cross-Language Evaluation Forum (CLEF) has been running for ten years now. The intention of this document is to provide a panorama of the CLEF 2009 results, focusing attention on the analysis and discussion of the results in the following tracks:

- *Multilingual Textual Document Retrieval (Ad Hoc)*: The Ad Hoc track is considered our core track. It is the one track that has been offered each year, from 2000 through 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 are 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.
- *Information, Filtering, Evaluation (INFILE)*: is a cross-language adaptive filtering evaluation campaign. The goal of the INFILE track is to organize evaluation campaigns for monolingual and multilingual information filtering systems based on close-to-real-usage conditions for intelligence applications.
- *Log File Analysis (LogCLEF)*: LogCLEF was a new track at CLEF 2009; the main purpose of the track has been to stimulate research on user behavior in multilingual environments and to develop standard evaluation collections able to support long-term research in log analysis. Two tasks were addressed: Log Analysis and Geographic Query Identification (LAGI) and Log Analysis for Digital Societies (LADS).
- *Systematic Multilingual Experimentation (Grid@CLEF)*: This tracks has been launched as a pilot effort in CLEF 2009 to create a framework and run as much experiment as possible to systematically investigate the interaction among languages, IR system components', and tasks.

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 2009 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.

The overall outcome of CLEF 2009 is the involvement of a ever increasing research community and the organization of tracks which are getting closer and close to the needs and requirements emerging from user and developer communities.

1 Introduction

As we stated in Del 2.3.1, when we launched the Cross Language Evaluation Forum (CLEF) in 2000, our focus was on text and document retrieval. Over the years our scope has gradually expanded to include different kinds of text retrieval across languages and different kinds of media. 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 both 2008 and 2009 – the TrebleCLEF years[1]. 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.

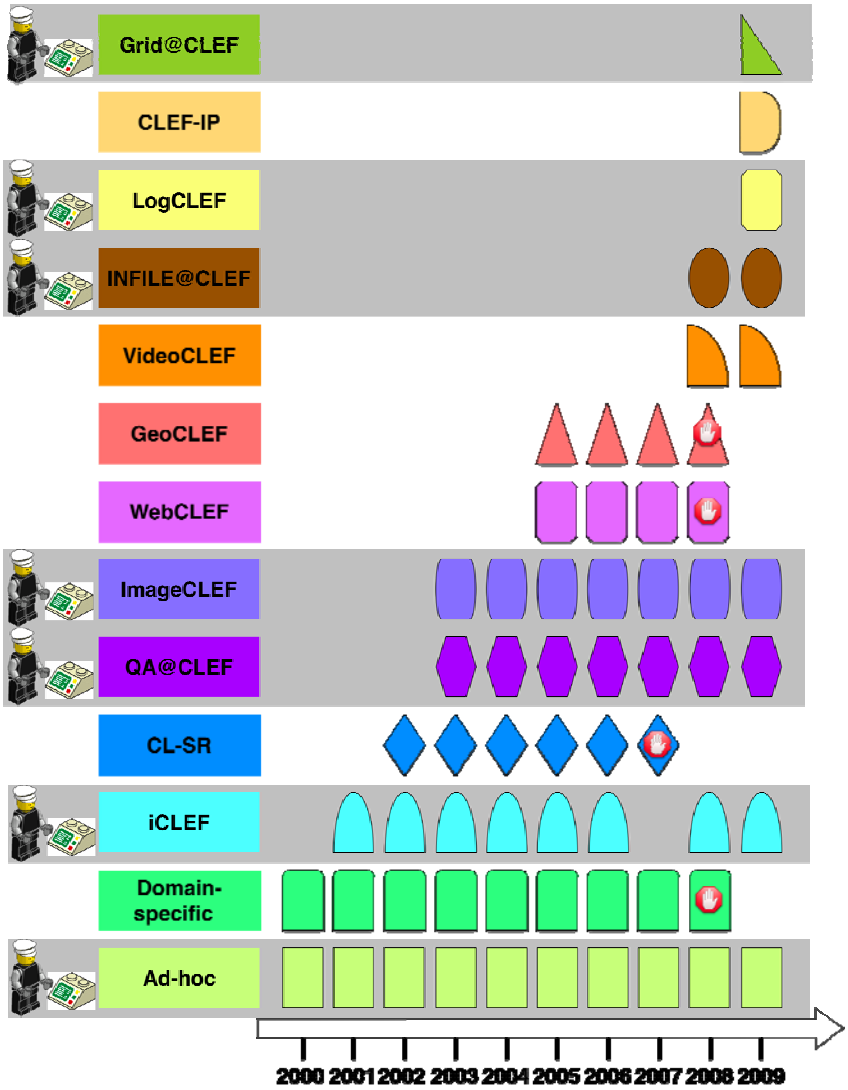


Figure 1: CLEF 2000-2009 tracks with tracks under analysis in this report highlighted.

[1] Throughout its lifetime, the central coordination of CLEF has been partially supported by the European Commission under the 5th, 6th and 7th Framework Programmes. For many years it was organised as part of the DELOS Network of Excellence. CLEF 2008 and 2009 have received funding from the TrebleCLEF Coordination Action.

Figure 1 shows clearly the differences between the CLEF 2008 and CLEF 2009 campaigns. In 2009, three of the tracks that had been hosted in 2008 were discontinued: GeoCLEF, WebCLEF and Domain-specific. The reasons for terminating these tracks were diverse.

- The domain-specific track had run since CLEF began in 2000. Although the participation was never very high (on average between 5 – 8 groups each year), over the years the track has made it possible to investigate exhaustively the particular features of domain-specific retrieval on semi structured data records and a good test collection has been built. We are now working on making this publicly available (see Del 5.3). In 2008 the coordinating group of this track decided that its purpose had been served and that it was time to terminate this activity.
- The GeoCLEF track has been an interesting track and has involved a lot of discussion among the participants and coordinators with respect to the most appropriate task design. At the end of the 2008 campaign, it was decided to discontinue this track in CLEF with the coordination of the main task moving to NTCIR[2] and with one of the subtasks (the GikiP pilot) moving to the Question Answering track as GikiCLEF, focusing on open list questions over Wikipedia that require geographic reasoning, complex information extraction, and cross-lingual processing.
- Basically, the WebCLEF track was dropped due to lack of strong interest by participants; this is somewhat surprising as the Web would seem the most natural context for cross-language retrieval experimentation. The reason for the low participation remains unclear.

In CLEF2009, we introduced three new tracks: GridCLEF, LogCLEF and CLEF-IP. The aim was to further strengthen one of the stated objectives of TrebleCLEF:

> *"support the annual CLEF system evaluation campaigns with tracks designed to meet the specific requirements of the user and application communities, and particular focus on user modeling, language-specific experimentation, and results presentation[3]"*

Thus the LogCLEF track joined the existing interactive track (iCLEF) in the task of investigating user behaviour in a multilingual context via the analysis of log data, the CLEF-IP track studied multilingual access and retrieval in an important application area, that of patent retrieval, and the Grid@CLEF track made a first attempt at organising a set of large-scale, systematic grid experiments aimed at improving our comprehension of MLIA systems and gaining an exhaustive picture of their behaviour with respect to languages.

The organisation of the CLEF 2009 campaign has already been described in Deliverable 2.1.2. In this Introduction, we briefly summarise the main details necessary to provide background information for the rest of this report. The aim is to provide an in-depth analysis of the results for a selection of tasks and tracks: Ad Hoc – TEL and Persian tasks; ImageCLEF; iCLEF, Question Answering – ResPubliQA task, LogCLEF, INFILE and Grid@CLEF.

These tracks and/or tasks 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 is their strategic importance in the general multilingual information access paradigm and within the context of TrebleCLEF.

Although all these tracks (and their subtasks) are of strong interest for the R&D multilingual system development community, they can be divided into three groups: more user-oriented; more-application-oriented; more-research-oriented as follows:

- More user-oriented: iCLEF and LogCLEF
- More application oriented: ImageCLEF and TEL tasks
- More research-oriented: Persian task; ResPubliQA, INFILE and GridCLEF

This deliverable is organized as follows: in the rest of this section we briefly introduce the tracks and/or tasks under examination in this report; Sections 2 to 8 then provide a detailed analysis of each

[2] NTCIR: Evaluation of Information Access Technologies, see `http://research.nii.ac.jp/ntcir/`
[3] See TrebleCLEF Annex 1

of them; Section 9 discusses CLEF 2009 from the TrebleCLEF perspective and provides some concluding remarks.

2 The Ad-hoc Track

The Ad Hoc track is considered as our core track [Ferro and Peters 2009b]. It is the one track that has been offered all through the ten years of CLEF. The track has been primarily research-oriented with the aim of studying and promoting 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. In the first eight years of the Ad Hoc track, monolingual, bilingual and multilingual document retrieval tasks were offered for target collections of comparable news documents in thirteen European languages[4], with bilingual tasks often being proposed for unusual pairs of languages, such as Finish to German, or French to Dutch. This activity has resulted in a unique set of test collections which can be used for benchmarking multilingual IR systems. We aim to make these available with Del 5.3.

In 2008 there was a big change in focus in the 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 a more application-oriented task, offering monolingual and cross-language search on library catalog records. It was organized in collaboration with The European Library (TEL)[5] 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. 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. Persian was chosen 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. 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.

The 2009 Ad Hoc track was to a large extent a repetition of last year's track, with the same three tasks: Tel@CLEF, Persian@CLEF, and Robust-WSD. The track was again coordinated jointly by ISTI-CNR and Padua University, Italy; the University of the Basque Country, Spain; with the collaboration of the Database Research Group, University of Tehran, Iran. In this deliverable we focus on analysing the results of the first two tasks: TEL and Persian.

[4] Over the years, CLEF has built up two multilingual corpora of more than 3 million news documents, in 14 European languages. This corpus is divided into two comparable collections: 1994-1995 - Dutch, English, Finnish, French, German, Italian, Portuguese, Russian, Spanish, Swedish; 2000-2002 - Basque, Bulgarian, Czech, English, Hungarian.

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

	Monolingual	Bilingual
CLEF 2008	fa TEL de;en;fr Robust WSD en	en→fa TEL x→de;en;fr Robust WSD es→en
CLEF 2009	fa TEL de;en;fr Robust WSD en	en→fa TEL x→de;en;fr Robust WSD es→en

Table 1: Ad Hoc 2008–2009 Tasks. The following ISO 639-1 language codes have been used: de=German; en=English; es=Spanish; fa=Farsi; fr=French.

A detailed description of the Ad hoc track and all the experimental results can be found in [Ferro and Peters 2009a; Di Nunzio and Ferro 2009a,b,c]. We used the DIRECT (Distributed Information Retrieval Evaluation Campaign Tool)[6] system [Agosti and Ferro 2009; Dussin and Ferro 2009; Ferro 2008] to manage the different aspects of the track, i.e. topic creation, experiment submission, relevance assessment, and performance measures computation.

2.1 TEL Task

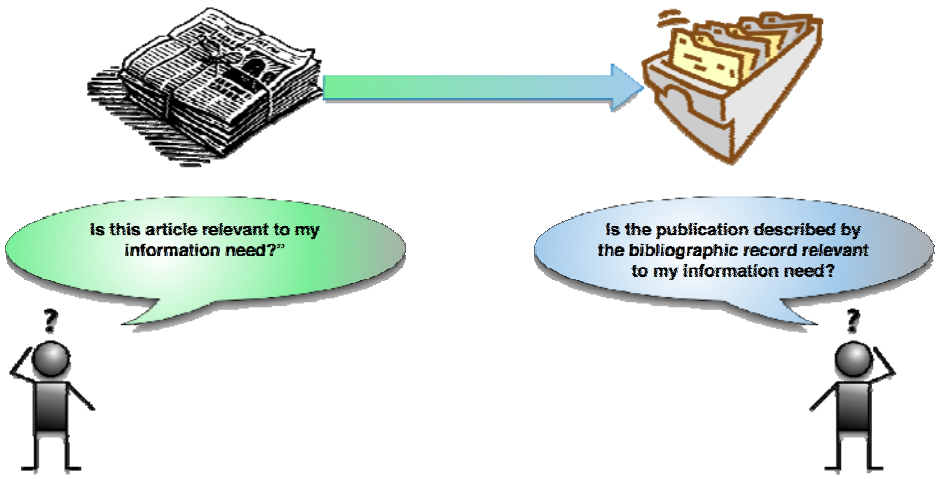


Figure 2: Shift from searching for documents of interest to searching for surrogates that represent documents of interest.

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?".

[6] See http://direct.dei.unipd.it/about.html

Two subtasks were offered: Monolingual and Bilingual. In both tasks, the aim was to retrieve documents relevant to the query. In CLEF 2009, the activity we simulated was that of users who have a working knowledge of English, French and German 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 from 11 countries submitted 211 runs for the TEL task: 116 runs out of 211 were monolingual; 95 runs out of 211 were bilingual.

2.1.1 Documents

The TEL task used three collections:

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

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

The TEL data is very different from the newspaper articles and news agency dispatches previously used in the CLEF ad hoc track. The data tends to be very sparse. Many records contain only title, author and subject heading information; other records provide more detail. The title and (if existing) an abstract or description may be in a different language to that understood as the language of the collection. The subject heading information is normally in the main language of the collection.

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 Chinese, Greek, and Italian in response to demand. Only the Title and Description fields were released to the participants. The narrative was employed to provide information for the assessors on how the topics should be judged. The topic sets were prepared on the basis of the contents of the collections.

In ad hoc, when a task uses data collections in more than one language, we consider it important to be able to use versions of the same core topic set to query all collections. This makes it easier to compare results over different collections and also facilitates the preparation of extra topic sets in additional languages. However, it is never easy to find topics that are effective for several different collections and the topic preparation stage requires considerable discussion between the coordinators for each 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 3, can be expected to find far more relevant documents in the BL collection than in BNF or ONB.

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

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

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

Figure 3: Example of TEL topic in all five languages: topic 10.2452/711-AH.

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.

The box plot of Figure 4 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 French and German distributions appear similar and are slightly asymmetric towards topics with a greater number of relevant documents while the TEL English distribution is slightly asymmetric towards topics with a lower number of relevant documents. All the distributions show some upper outliers, i.e. topics with a greater number of relevant document with respect to the behaviour of the other topics in the distribution. These outliers are probably due to the fact that CLEF topics have to be able to retrieve relevant documents in all the collections; therefore, they may be considerably broader in one collection compared with others depending on the contents of the separate datasets.

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

Table 2: Summary information about TEL pools.

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

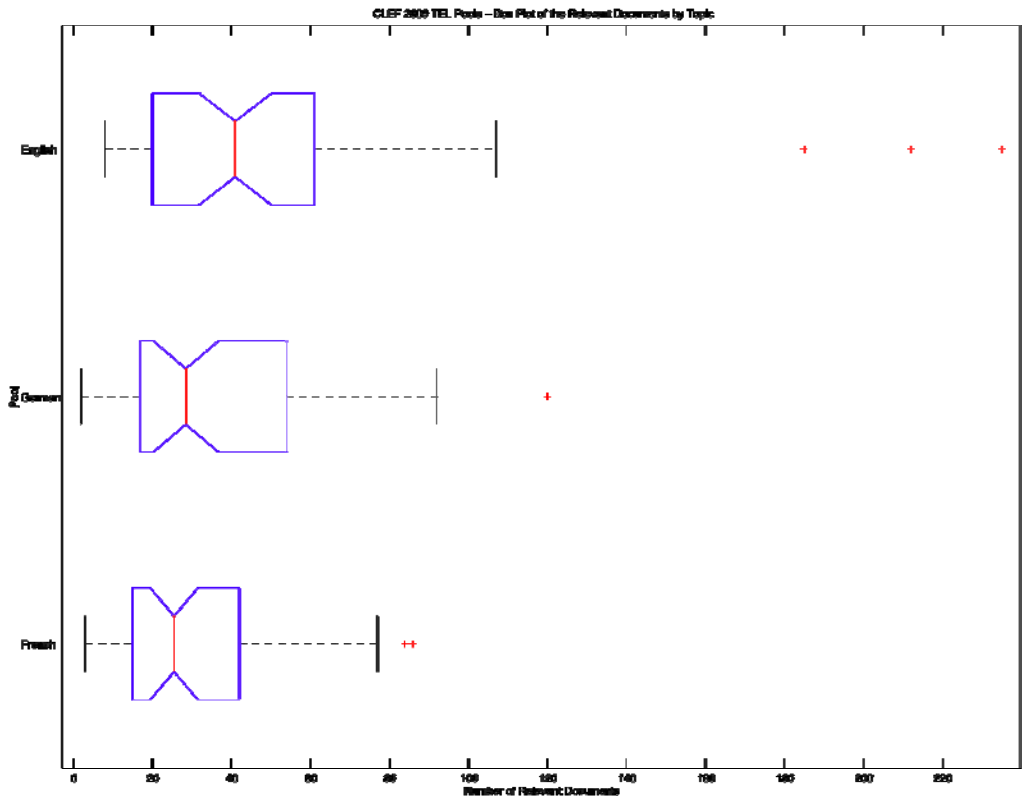


Figure 4: Distribution of the relevant documents in the TEL pools.

2.1.4 Monolingual Results

The individual results for all official Ad-hoc experiments in CLEF 2009 are given in the Appendices of the CLEF 2009 Working Notes [Di Nunzio and Ferro 2009a]. You can also access them online at:

- Monolingual English:
  `http://direct.dei.unipd.it/DOIResolver.do?type=task&id=AH-TEL-MONO-EN-CLEF2009`
- Monolingual French:
  `http://direct.dei.unipd.it/DOIResolver.do?type=task&id=AH-TEL-MONO-FR-CLEF2009`
- Monolingual German:
  `http://direct.dei.unipd.it/DOIResolver.do?type=task&id=AH-TEL-MONO-DE-CLEF2009`

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

Table 3: Best entries for the monolingual TEL tasks.

Figure 5 compares the performances of the top participants in the TEL Monolingual tasks.

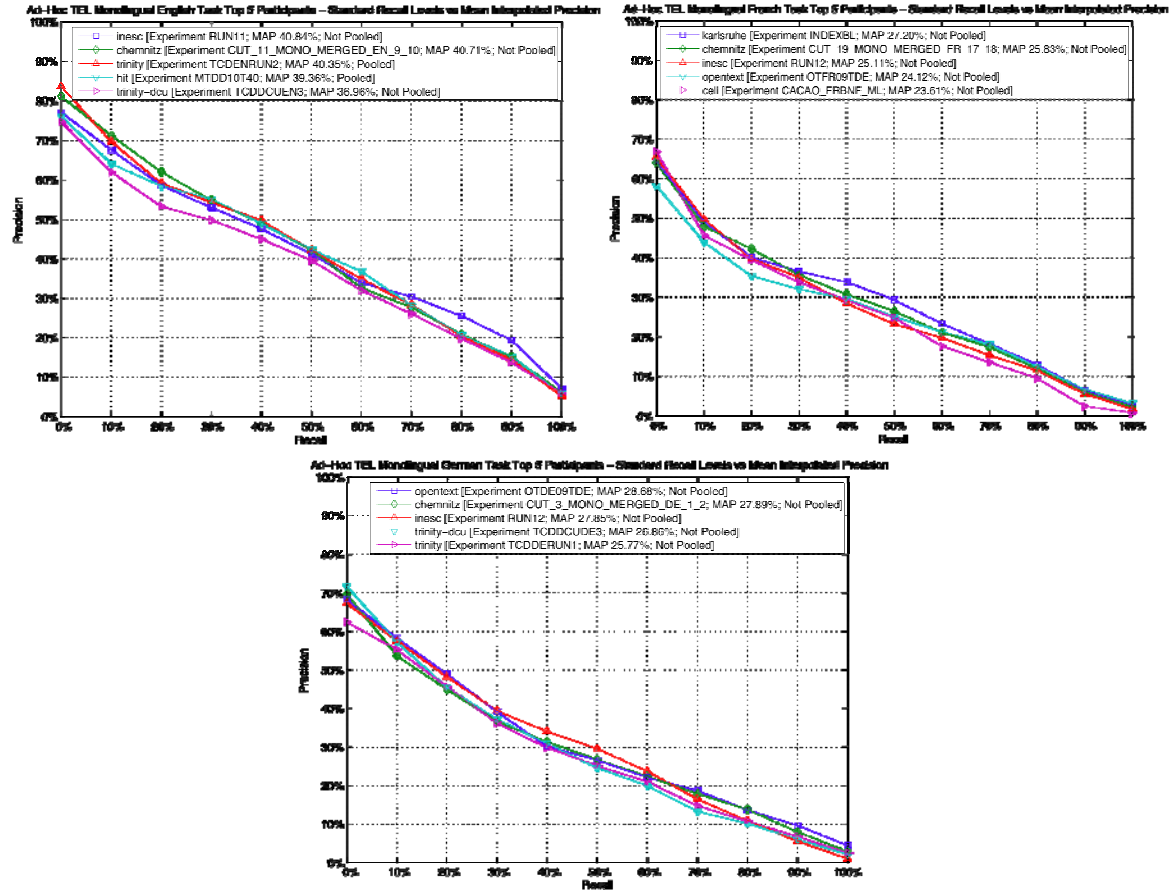


Figure 5: Comparison of the performances of the top participants in the monolingual TEL tasks.

2.1.5 Bilingual Results

The individual results for all official Ad-hoc experiments in CLEF 2009 are given in the Appendices of the CLEF 2009 Working Notes [Di Nunzio and Ferro 2009a]. You can also access them online at:

- Bilingual English:
  `http://direct.dei.unipd.it/DOIResolver.do?type=task&id=AH-TEL-BILI-X2EN-CLEF2009`
- Bilingual French:
  `http://direct.dei.unipd.it/DOIResolver.do?type=task&id=AH-TEL-BILI-X2FR-CLEF2009`
- Bilingual German:
  `http://direct.dei.unipd.it/DOIResolver.do?type=task&id=AH-TEL-BILI-X2DE-CLEF2009`

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.

Track	Rank	Participant	Experiment DOI	MAP
English	1st	chemnitz	10.2415/AH-TEL-BILI-X2EN-CLEF2009.CHEMNITZ.CUT.13.BILI.MERGED.DE2EN.0.10	40.46%
	2nd	hit	10.2415/AH-TEL-BILI-X2EN-CLEF2009.HIT.XTUD10T40	35.27%
	3rd	trinity	10.2415/AH-TEL-BILI-X2EN-CLEF2009.TRINITY.TCDEDEMOE3	35.05%
	4th	trinity-dcu	10.2415/AH-TEL-BILI-X2EN-CLEF2009.TRINITY-DCU.TCDDCUDE01	33.33%
	5th	karlsruhe	10.2415/AH-TEL-BILI-X2EN-CLEF2009.KARLSRUHE.DE_INDEXEL	32.70%
	Difference			23.73%
French	1st	chemnitz	10.2415/AH-TEL-BILI-X2FR-CLEF2009.CHEMNITZ.CUT.24.BILI.EN2FR.MERGED.LAEG.SPEC.REF.CUT.17	25.57%
	2nd	karlsruhe	10.2415/AH-TEL-BILI-X2FR-CLEF2009.KARLSRUHE.EN_INDEXEL	24.62%
	3rd	chesire	10.2415/AH-TEL-BILI-X2FR-CLEF2009.CHESHIRE.BIEN2FRT2FB	16.77%
	4th	trinity	10.2415/AH-TEL-BILI-X2FR-CLEF2009.TRINITY.TCDEDEMOE2	16.33%
	5th	weimar	10.2415/AH-TEL-BILI-X2FR-CLEF2009.WEIMAR.CLEM1602MERGEDTOP5	14.51%
	Difference			69.67%
German	1st	chemnitz	10.2415/AH-TEL-BILI-X2DE-CLEF2009.CHEMNITZ.CUT.5.BILI.MERGED.EN2DE.1.2	25.83%
	2nd	trinity	10.2415/AH-TEL-BILI-X2DE-CLEF2009.TRINITY.TCDEDEMOE3	19.35%
	3rd	karlsruhe	10.2415/AH-TEL-BILI-X2DE-CLEF2009.KARLSRUHE.EN_INDEXEL	16.46%
	4th	weimar	10.2415/AH-TEL-BILI-X2DE-CLEF2009.WEIMAR.CUMULEDEFRDE	15.75%
	5th	chesire	10.2415/AH-TEL-BILI-X2DE-CLEF2009.CHESHIRE.BIEN2DET2FBT	11.50%
	Difference			124.60%

Table 4: Best entries for the bilingual TEL tasks.

Figure 6 compares the performances of the top participants in the TEL Bilingual tasks.

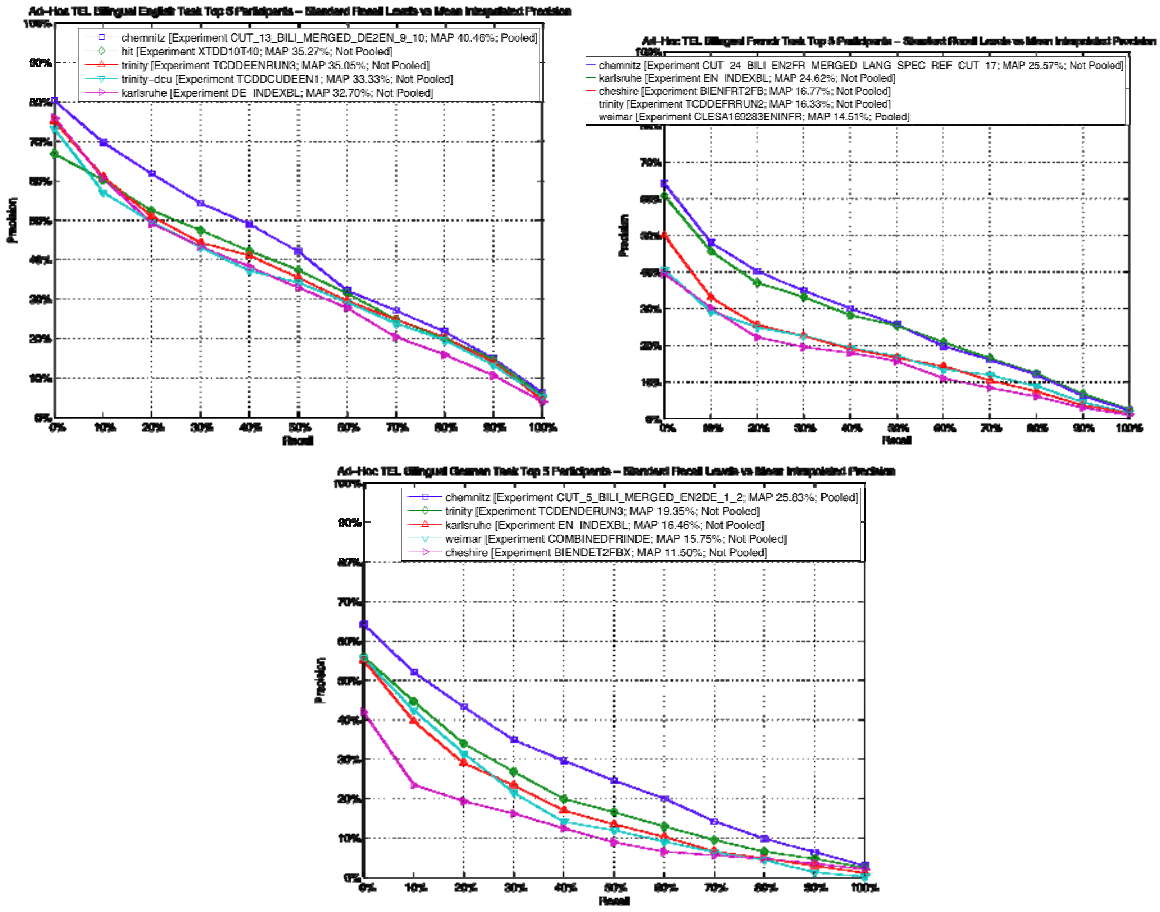


Figure 6: Comparison of the performances of the top participants in the bilingual TEL tasks.

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

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

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

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

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

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

2.1.6 Approaches and Discussion

As stated in the introduction, the TEL task this year is a repetition of the task set last year. A main reason for this was to create a good reusable test collection with a sufficient number of topics; another reason was to see whether the experience gained and reported in the literature last year, and the opportunity to use last year's test collection as training data, would lead to differences in approaches

and/or improvements in performance this year. Although we have exactly the same number of participants this year as last year, only five of the thirteen 2009 participants also participated in 2008. These are the groups tagged as Chemnitz, Cheshire, Karlsruhe, INESC-ID and Opentext. The last two of these groups only tackled monolingual tasks. These groups all tend to appear in the top five for the various tasks. In the following we attempt to examine briefly the approaches adopted this year, focusing mainly on the cross-language experiments.

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

Five of the ten groups that attempted cross-language tasks used the Google Translate functionality, while a sixth used the LEC Power Translator [Larson 2009a]. Another group also used an MT system combining it with concept-based techniques but did not disclose the name of the MT system used [Sorg et al. 2009]. The remaining three groups used a bilingual term list [Katsiouli and Kalamboukis 2009], a combination of resources including on-line and in house developed dictionaries [Bosca and Dini 2009], and Wikipedia translation links [Jadidinejad and Mahmoudi 2009]. It is important to note that four out of the five groups in the bilingual to English and bilingual to French tasks and three out of five for the bilingual to German task used Google Translate, either on its own or in combination with another technique. One group noted that topic translation using a statistical MT system resulted in about 70% of the mean average precision (MAP) achieved when using Google Translate [Leveling et al. 2009]. Another group [Anderka et al. 2009] found that the results obtained by simply translating the query into all the target languages via Google gave results that were comparable to a far more complex strategy known as Cross-Language Explicit Semantic Analysis, CL-ESA, where the library catalog records and the queries are represented in a multilingual concept space that is spanned by aligned Wikipedia articles. As this year's results were significantly better than last year's, can we take this as meaning that Google is going to solve the cross-language translation resource quandary?

Taking a closer look at three groups that did consistently well in the cross-language tasks we find the following. The group that had the top result for each of the three tasks was Chemnitz [Kürsten 2009]. They also had consistently good monolingual results. Not surprisingly, they appear to have a very strong IR engine, which uses various retrieval models and combines the results. They used Snowball stemmers for English and French and an n-gram stemmer for German. They were one of the few groups that tried to address the multilinguality of the target collections. They used the Google service to translate the topic from the source language to the four most common languages in the target collections, queried the four indexes and combined the results in a multilingual result set. They found that their approach combining multiple indexed collections worked quite well for French and German but was disappointing for English.

Another group with good performance, Karlsruhe [Sorg et al. 2009], also attempted to tackle the multilinguality of the collections. Their approach was again based on multiple indexes for different languages with rank aggregation to combine the different partial results. They ran language detectors on the collections to identify the different languages contained and translated the topics to the languages recognized. They used Snowball stemmers to stem terms in ten main languages, fields in other languages were not pre-processed. Disappointingly, a baseline consisting of a single index without language classification and a topic translated only to the index language achieved similar or even better results. For the translation step, they combined MT with a concept-based retrieval strategy based on Explicit Semantic Analysis and using the Wikipedia database in English, French and German as concept space.

A third group that had quite good cross-language results for all three collections was Trinity [Zhou and Wade 2009]. However, their monolingual results were not so strong. They used a language modelling

retrieval paradigm together with a document re-ranking method which they tried experimentally in the cross-language context. Significantly, they also used Google Translate. Judging from the fact that they did not do so well in the monolingual tasks, this seems to be the probable secret of their success for cross-language.

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.

Four groups submitted 20 runs for the Persian task: 17 runs out of 20 were monolingual; 3 runs out of 20 were bilingual.

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[7].

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. An example of topic is shown in Figure 7.

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

    <title lang="en">Pollution in the Persian Gulf</title>
    <title lang="fa">وضعیت آلودگی دریای خلیج فارس</title>

    <description lang="en">
        Find information about pollution in the Persian Gulf and the causes.
    </description>
    <description lang="fa">
        بررسی وضعیت دریای خلیج فارس از نظر آلودگی و عوامل آن
    </description>

    <narrative lang="en">
        Find information about conditions of the Persian Gulf with respect to
        pollution; also of interest is information on the causes of pollution
        and comparisons of the level of pollution in this sea against that of
        other seas.
    </narrative>
    <narrative lang="fa">
        یافتن اطلاعاتی در مورد وضعیت آلودگی دریای خلیج فارس و بررسی عوامل ایجاد آل
        ودگی در این دریا و اطلاعاتی نظیر مقایسه آن با سایر دریاها
    </narrative>
</topic>
```

Figure 7: Example of Persian topic: topic 10.2452/641-AH.

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

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/AR-PERSIAN-CLEF2009)	
Pool size	23,536 pooled documents • 19,072 not relevant documents • 4,464 relevant documents 50 topics
Pooled Experiments	20 out of 20 submitted experiments • monolingual: 17 out of 17 submitted experiments • bilingual: 3 out of 3 submitted experiments
Assessors	23 assessors

Table 5: Summary information about the Persian pool.

As shown in the box plot of Figure 8, 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.

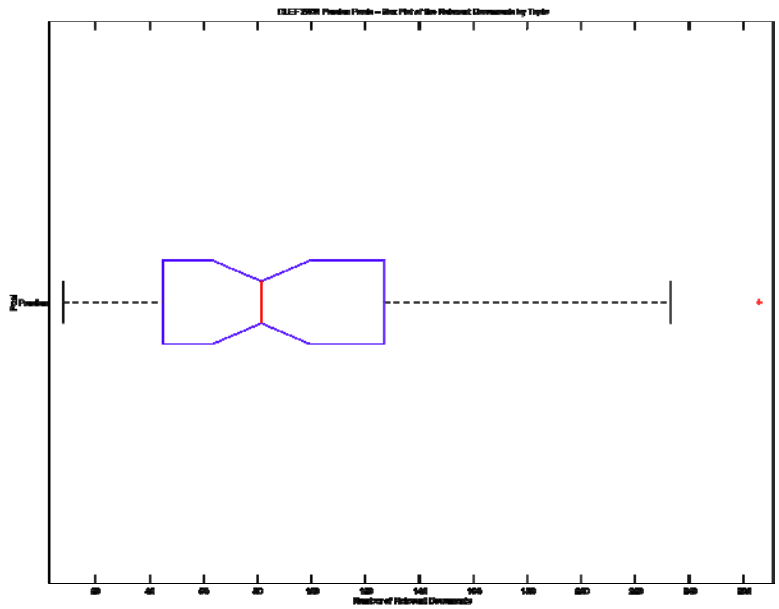


Figure 8: Distribution of the relevant documents in the Persian pool.

2.2.4 Monolingual Results

The individual results for all official Ad-hoc experiments in CLEF 2009 are given in the Appendices of the CLEF 2009 Working Notes [Di Nunzio and Ferro 2009ab]. You can also access them online at:

- Monolingual Farsi:
  `http://direct.dei.unipd.it/DOIResolver.do?type=task&id=AH-PERSIAN-MONO-FA-CLEF2009`

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	jhu-apl	10.2415/AH-PERSIAN-MONO-FA-CLEF2009.JHU-APL.JHUFASK41R400TD	49.38%
	2nd	unine	10.2415/AH-PERSIAN-MONO-FA-CLEF2009.UNINE.UNINEPE4	49.37%
	3rd	opentext	10.2415/AH-PERSIAN-MONO-FA-CLEF2009.OPENTEXT.OTFA09TDE	39.53%
	4th	qazviniau	10.2415/AH-PERSIAN-MONO-FA-CLEF2009.QAZVINIAU.IAUPERFA3	37.62%
	5th	–	–	–%
	Difference			31.25%

Table 6: Best entries for the monolingual Persian task.

Figure 9 compares the performances of the top participants in the Persian task.

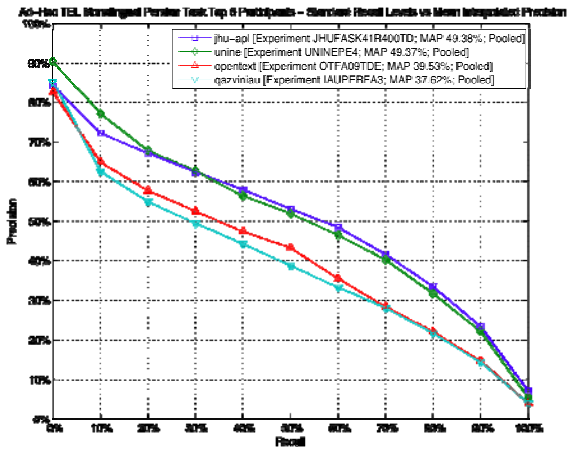


Figure 9: Comparison of the performances of the top participants in the monolingual Persian task.

2.2.5 Bilingual Results

The individual results for all official Ad-hoc experiments in CLEF 2009 are given in the Appendices of the CLEF 2009 Working Notes [Di Nunzio and Ferro 2009a, b]. You can also access them online at:

- Bilingual Farsi:
  `http://direct.dei.unipd.it/DOIResolver.do?type=task&id=AH-PERSIAN-BILI-X2FA-CLEF2009`

We had only one participant in the bilingual task (qazviniau) whose best run (10.2415/AH-PERSIAN-BILI-X2FA-CLEF2009.QAZVINIAU.IAUPEREN3) achieved a MAP of 2.72%. Figure 10 shows its performances.

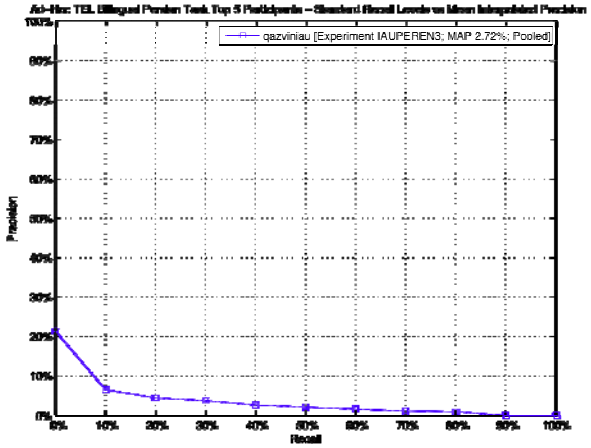


Figure 10: Comparison of the performances of the top participants in the bilingual Persian task.

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

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

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

- EN → FA: 92.26% of best monolingual Farsi IR system.

2.2.6 Approaches and Discussion

We were very disappointed this year that despite the fact that 14 groups registered for the Persian task, only four actually submitted results. And only one of these groups was from Iran. We suspect that one of the reasons for this was that the date for submission of results was not very convenient for the Iranian groups. Furthermore, only one group [Jadidinejad and Mahmoudi 2009] attempted the bilingual task with the very poor results cited above. The technique they used was the same as that adopted for their bilingual to English experiments, exploiting Wikipedia translation links, and the reason they give for the very poor performance here is that the coverage of Farsi in Wikipedia is still very scarce compared to that of many other languages.

In the monolingual Persian task, the top two groups had very similar performance figures. [Dolamić et al. 2009] found they had best results using a light suffix-stripping algorithm and by combining different indexing and searching strategies. Interestingly, their results this year do not confirm their findings for the same task last year when the use of stemming did not prove very effective. The other group [McNamee 2009] tested variants of character n-gram tokenization; 4-grams, 5-grams, and skipgrams all provided about a 10% relative gain over plain words.

3 ImageCLEF

ImageCLEF[8] is the CLEF cross-language image retrieval track and has already seen participation from both academic and commercial research groups worldwide from communities including: Cross-Language Information Retrieval (CLIR), Content-Based Image Retrieval (CBIR) and user interaction. The overall focus, similar to previous years, has been on investigating methods for exploiting and combining textual and visual features for image retrieval and automated image annotation. In 2009, ImageCLEF organised six different tasks. In addition to running the CLEF tasks, a workshop was organised and sponsored by THESEUS[9]. The Workshop on Visual Information Retrieval Evaluation had contributions from a number of researchers in the following fields:

[8] See `http://imageclef.org/`
[9] See `http://theseus-programm.de`

- Evaluation of visual information retrieval and annotation methods
- Image retrieval/image annotation with application to photos, medical data and robotic vision
- Multi-label image annotation supported by knowledge structures (ontologies)

In total, 84 groups registered for ImageCLEF 2009 with 62 groups actually submitting results (medical annotation: 7 groups; medical retrieval: 17 groups; photo annotation: 19 groups; photo retrieval: 19 groups; robot vision: 7 groups; wikipediaMM: 8 groups). The tasks for 2009 included 3 retrieval tasks and 3 purely visual tasks, with the emphasis on language independence. Most of the collections used are English with queries (for the retrieval tasks) typically in multiple languages. The main changes in 2009 included a new task (the robot vision task) and the addition of new databases to almost all tasks. A detailed description of all tasks can be found in the CLEF working notes for 2009 [Borri et al. 2009].

3.1 Photo retrieval task

The ImageCLEF Photo Retrieval Task 2009 focused on image retrieval and diversity. A new collection was utilized in this task consisting of approximately half a million images with English annotations. Queries were based on analyzing search query logs and two different types were released: one containing information about image clusters; the other without. A total of 19 participants submitted 84 runs. Evaluation, based on Precision at rank 10 and Cluster Recall at rank 10, showed that participants were able to generate runs of high diversity and relevance. Findings show that submissions based on using mixed modalities performed best compared to those using only concept-based or content-based retrieval methods. The selection of query fields was also shown to affect retrieval performance. Submissions not using the cluster information performed worse with respect to diversity than those using this information.
Given a set of information needs (topics), participants were tasked with finding not only relevant images, but also generating ranked lists that promote diversity. To make the task harder, two types of queries were released: the first type of query included written information about the specific requirement for diversity (represented as clusters); queries of the second type contained a more conventional title and example relevant images. In the former type of query participants were required to retrieve diverse results with some indication of what types of clusters were being sought; in the latter type of query little evidence was given for what kind of diversity was required. Evaluation gave more credence to runs that presented diverse results without sacrificing precision than those exhibiting less diversity.
The Photo Retrieval task in 2009 aimed to study diversity further than in 2008. Using resources from Belga, a Belgian news agency, a much larger collection was provide than in 2008: containing just under half a million images compared to 20,000 images provided in 2008. Statistics on popular queries submitted to the Belga website in 2008[10] were obtained and exploited to create representative queries for this diversity task. Different ways of specifying the need for diversity were experimented with and this year decided to release half of the queries without any indication of diversity required or expected.

3.1.1 Document collection

The collection consists of 498,920 images with English-only annotations (i.e. captions) describing the content of the image. However, different to the structured annotations of 2008, the annotations in this collection are presented in an unstructured way (see Table 8). This increases the challenge for participants as they must automatically extract information about the location, date, photographic source, etc of the image as a part of the indexing and retrieval process. The photos cover a wide-ranging time period, and there are many cases where pictures have not been orientated correctly, thereby increasing the challenge for content-based retrieval methods.

[10] This analysis of queries from a 2008 Belga query log was carried out by Theodora Tsikrika (in 2009).

	Annotation: 20090126 - DENDERMONDE, BELGIUM: Lots of people pictured during a commemoration for the victims of the knife attack in Sint-Gilles, Dendermonde, Belgium, on Monday 26 January 2009. Last Friday 20-Year old Kim De Gelder killed three people, one adult and two childs, in a knife attack at the children's day care center "Fabeltjesland" in Dendermonde. BELGA PHOTO BENOIT DOPPAGNE

Table 7. Example image and caption

3.1.2 Topics

In an attempt to generate more realistic topics for ImageCLEFPhoto 2009, queries were derived from analyzing query logs from the search engine used by Belga to access the document collection already provided. Similar to 2008, the aim of this year's task was to investigate promoting diversity in the search results. Unlike the 2008 task, however, the queries were chosen to reflect more realistic kinds of diversity one might expect in practice for further information about creating the topics) [Lestrari Paramita et al. 2009a]. The idea is that the results for certain kinds of query (e.g. ambiguous ones) will cluster into groups (e.g. different senses of an ambiguous query; different aspects of a query etc.), which can be observed in user's query reformulations. In total, 50 topics were generated; from these 25 were randomly selected containing titles for the clusters (identified by the organisers in advance from analyzing query reformulations in the query logs) and an example image for each cluster (called Query Part 1), and 25 did with no cluster title but just an example image (called Query Part 2). On average, across the topics, there were 3.96 clusters per topic and an average of 208 relevant documents per cluster per query.

3.1.3 Relevance assessments

Relevance assessments were performed using the DIRECT [Agosti and Ferro 2009; Dussin and Ferro 2009; Ferro 2008] to enable assessors to work in a collaborative environment. A total of 25 assessors were hired to be involved in this process and assessments were divided into 2 phases: in the first phase, assessors were asked to identify images relevant to a given query. Information about all relevant clusters to the topic was given to assessors to ensure they were aware of the scope of relevant images for a query. After a set of relevant images were found, for the second stage different assessors were asked to find images relevant to each cluster (some images could belong to multiple clusters). Since topics varied widely in content and diversity, the number of relevant images varied from 1 to 1,266 for each cluster. Initially, there were 206 clusters created for the 50 queries, but this number dropped to 198 as there were 8 clusters with no relevant images which had to be deleted. There are an average number of 208 relevant documents for each cluster, with a standard deviation of 280.59.

3.1.4 Results

A total of 44 groups registered to participate in this task, with 19 groups finally submitting runs (limited to 5 runs per group). A total of 84 runs were submitted and evaluated (some groups submitted < 5 runs). A full breakdown of the runs can be found in the track overview paper [Lestrari Paramita et al. 2009b]. Of the 84 runs, 36 runs used both textual and visual features for retrieval and promoting diversity, 41 used textual features only and 7 runs used image features only. The evaluation measures used for 2009 were Precision at rank 10 (P@10), Cluster Recall at rank 10 (CR@10) and a combined measure of precision and recall, the harmonic mean F_1.

You can also access the result online at:

- http://direct.dei.unipd.it/DOIResolver.do?type=task&id=IC-PR-MONO-EN-CLEF2009

No	Group	Run Name	Query	Modality	P@10	CR@10	F1
1	XEROX-SAS	XRCEXKNND	T-CT-I	TXT-IMG	0.794	0.824	0.809
2	XEROX-SAS	XRCECLUST	T-CT-I	TXT-IMG	0.772	0.818	0.794
3	XEROX-SAS	KNND	T-CT-I	TXT-IMG	0.8	0.727	0.762
4	INRIA	LEAR5_TI_TXTIMG	T-I	TXT-IMG	0.798	0.729	0.762
5	INRIA	LEAR1_TI_TXTIMG	T-I	TXT-IMG	0.776	0.741	0.758
6	InfoComm	LRI2R_TI_TXT	T-I	TXT	0.848	0.671	0.749
7	XEROX-SAS	XRCE1	T-CT-I	TXT-IMG	0.78	0.711	0.744
8	INRIA	LEAR2_TI_TXTIMG	T-I	TXT-IMG	0.772	0.706	0.737
9	Southampton	SOTON2_T_CT_TXT	T-CT	TXT	0.8240	0.654	0.729
10	Southampton	SOTON2_T_CT_TXT_IMG	T-CT	TXT-IMG	0.746	0.71	0.727

Table 8. Overall results for top 10 runs computed across all 50 queries ranked in descending order of F_1 score (the Query column represents the information used from the query: T=Title; I=Image; CT=Cluster Title).

Table 10 shows the average results (computed across all queries) for the two types of query: Query Part 1 and Query Part 2. The results indicate that the Cluster Title field has an important role in identifying diversity. When Cluster Title is not being used, the F_1 scores of both Query Part 1 and Query Part 2 do not differ significantly. Using a two-tailed paired t-test, the scores between Queries Part 1 and Queries Part 2 were found to be significantly different (p=0.02). There is also a significant correlation between the scores: the Pearson correlation coefficient equals 0.691.

Queries	P@10 Mean	P@10 SD	CR@10 Mean	CR@10 SD	F1 Mean	F1 SD
All Queries	0.655	0.209	0.547	0.137	0.585	0.166
Query Part 1	0.677	0.221	0.558	0.164	0.6	0.182
- Query Part 1 with CT	0.685	0.2	0.594	0.159	0.625	0.17
- Query Part 1 without CT	0.664	0.254	0.5	0.157	0.558	0.196
Query Part 2	0.632	0.219	0.542	0.133	0.569	0.173

Table 9. Results across all runs (mean values) for different query categories (Query Part 1 and Query Part 2).

Table 11 shows the results across all runs broken down by modality and, as shown in previous years, the best performance is obtained using a combination of textual and visual features. The mean of the runs using image content only (IMG) is drastically lower based on the P@10 score; however the gap decreases when considering only the CR@10 score. Further research should be carried out to improve runs using content-based approaches only, as the best run using this approach had the lowest F_1 score (0.218) compared to TXT (0.351) and TXT-IMG (0.297).

Modality	Number of Runs	P@10 Mean	P@10 SD	CR@10 Mean	CR@10 SD	F1 Mean	F1 SD
TXT-IMG	36	**0.713**	0.116	**0.612**	0.107	**0.656**	0.102
TXT	41	0.698	0.142	0.539	0.094	0.598	0.096
IMG	7	0.103	0.027	0.254	0.079	0.146	0.04

Table 10. Results across all runs (mean values) for different modalities.

A more detailed analysis of results can be found in the track overview paper [Lestrari Paramita et al. 2009b]. For details about approaches used by participating groups, see papers contributed by groups to the ImageCLEFPhoto workshop proceedings. Overall the results show that cluster information is highly beneficial in providing diversity in search results (this cluster information could be obtained from analyzing query reformulations in a query log). When cluster information is not available in verbal form (e.g. there is no prior search history), example images for each cluster can help produce diverse results. The best performance at this task is obtained with using a combination of all query fields and, as shown in previous years of ImageCLEF across a range of tasks, a combination of visual and textual features (i.e. mixed modality) will produce the highest system effectiveness.

3.2 Medical retrieval task

One of the longest running tasks is the ImageCLEF medical retrieval task, which ran for the 6th time in 2009 (the longest running being the general photo task). In total, 38 groups registered for the task with 17 groups submitting runs. The document collection in 2009 was similar to the one used in 2008, containing scientific articles from two radiology journals, Radiology and Radiographics. The size of the database was increased to a total of 74,902 images. For each image, captions and access to the full text article through the Medline PMID (PubMed Identifier) were provided. An article's PMID could be used to obtain the officially assigned MeSH (Medical Subject Headings) terms. The collection was entirely in English. However, the topics were, as in previous years, supplied in German, French, and English. A total of 25 image–based topics were provided, of which ten each were visual and mixed and five were textual. In addition, for the first time, 5 case–based topics were provided as an exploratory task. Here the unit of retrieval was intended to be *the article* and not the image. Case–based topics are designed to be a step closer to the clinical workflow: clinicians often seek information about patient cases with incomplete information consisting of symptoms, findings, and a set of images. Supplying cases to a clinician from the scientific literature that are similar to the case (s)he is treating can be an important application of image retrieval in the future.
As in previous years, most groups concentrated on fully automatic retrieval. However, four groups submitted a total of seven manual or interactive runs. The interactive runs submitted this year performed quite well compared to previous years, but did not show a substantial increase in performance over the automatic approaches. In previous years multimodal combinations were the most frequent submissions, however in 2009 about half as many mixed runs as purely textual runs were submitted. Very few fully visual runs were submitted, and again, the ones submitted performed poorly. The best mean average precisions (MAP) were obtained using automatic textual methods. There were mixed feedback runs that had high MAP. The best early precision was also obtained using automatic textual methods, with a few mixed automatic runs also doing well. For further information about this task, see the task overview paper by [Müller et al. 2009].

3.2.1 Document collection

The database in 2009 was, again, made accessible by the Radiological Society of North America (RSNA)[11]. The database contained a total of 74,902 images, all images taken from the journals Radiology and Radiographics of the RSNA. A similar database is also available via the Goldminer[12] interface. This collection constitutes an important body of medical knowledge from the peer–reviewed scientific literature including high quality images with annotations. Images are associated with journal articles and can be part of a figure. This creates high–quality textual annotations enabling textual searching in addition to content–based retrieval. As the PubMed IDs were also made available, participants could access the MeSH (Medical Subject Headings) terms created by the National Library of Medicine for PubMed.

[11] See `http://www.rsna.org/`
[12] See `http://goldminer.arrs.org/`

3.2.2 Topics

Realistic search topics were identified by surveying actual user needs: a Google grant funded a qualitative user study conducted at Oregon Health & Science University (OHSU) during early 2009 with 37 medical practitioners. Participants performed a total of 95 searches using textual queries written in English. From these, 25 queries were randomly selected as candidate queries from which to create topics for ImageCLEFMed 2009. For each query, 2-4 sample images from the previous collections of ImageCLEFMed were added as visual exemplars. In addition the textual descriptions for each topic were translated into French and German. The resulting set of 25 topics was categorized into three groups: 10 visual topics, 10 mixed topics, and 5 semantic topics. The entire set of topics was finally approved by a physician. For 2009, 5 case-based topics were generated for participants to help move the retrieval task closer to clinical routine as found in practice. The topics were created based on cases from the teaching file CasImage. This task is harder than the standard ad hoc task as multiple images, and other sources of data, can belong to a single case and the goal of the task is to find cases similar to the query.

3.2.3 Relevance assessments

Relevance judgments were performed with the same on–line system as in 2008 for the image–based topics. The system was adapted for the case–based topics showing the article title and several images appearing in the. A ternary judgment scheme was used: each image was judged to be "relevant", "partly relevant" or "non–relevant". Judges were instructed in the criteria and results were manually verified during the judgment process.

3.2.4 Results

In total, 17 groups participated in 2009, submitting a total of 124 runs for the ad hoc task (of which 13 runs were interactive/manual). There were only 16 "visual–only" runs, 59 "text–only" runs and 30 mixed runs (using a combination of textual and visual features). For the case-based task, a total of 6 groups participated, submitting at total of 18 runs. The results were quite promising with one group achieving a relatively high MAP of 0.33. Table 12 shows the results from the ad hoc retrieval task for the top performing groups (ranked by descending order of MAP score) for the runs involving combined visual and textual features. It is interesting to note that the best performing run overall for this task was using textual features only (MAP of 0.43), however upon further analysis it is found that mixed modality runs often improve early precision.

Run	Run Type	MAP	bpref	P5	P10	P30	rel_ret
deu_imaged_vsm	Mixed Automatic	0.37	0.39	0.63	0.54	0.48	1754
york.BO1.EdgeHistogram0.2	Mixed Automatic	0.36	0.37	0.58	0.58	0.54	1724
york.BO1.Tamura0.2	Mixed Automatic	0.35	0.37	0.62	0.57	0.51	1722
BM25b=0.75k_1=1.2k_3=8.0_ICLEFPProcess_3	Mixed Automatic	0.35	0.37	0.60	0.59	0.50	1763
York.BO1.colorHistogram0.2	Mixed Automatic	0.34	0.36	0.59	0.57	0.50	1719
BM25b=0.75k_1=1.2k_3=8.0_ICLEFPProcess_1	Mixed Automatic	0.33	0.35	0.59	0.57	0.47	1757
medGIFT0.3_withNegImg_EN	Mixed Automatic	0.29	0.32	0.63	0.60	0.52	1176
Multimodal_Text_Rerank	Mixed Automatic	0.27	0.40	0.49	0.52	0.45	1553
UNTMixed Automatic1	Mixed Automatic	0.24	0.28	0.46	0.40	0.37	1659
medGIFT0.5_EN	Mixed Automatic	0.21	0.25	0.70	0.59	0.43	848
UNTMixed Automaticrf1	Mixed Automatic	0.19	0.24	0.50	0.42	0.37	1197
ohsu_j_umls	Mixed Automatic	0.18	0.21	0.71	0.66	0.42	591
ohsu_j_mod1	Mixed Automatic	0.17	0.22	0.59	0.55	0.38	943
OHSU_SR6	Mixed Automatic	0.16	0.20	0.68	0.61	0.43	543
OHSU_SR2	Mixed Automatic	0.16	0.21	0.62	0.54	0.39	801
OHSU_SR3	Mixed Automatic	0.15	0.20	0.61	0.52	0.37	801
clef2009	Mixed Automatic	0.15	0.21	0.38	0.33	0.26	1381
medGIFT_mix_0.5vis_withNegImg	Mixed Automatic	0.14	0.17	0.56	0.49	0.33	547
Alicante-Run4	Mixed Automatic	0.13	0.17	0.31	0.34	0.33	992
uwmTextAndModality	Mixed Automatic	0.13	0.17	0.49	0.46	0.38	521
Alicante-Run5	Mixed Automatic	0.13	0.16	0.33	0.35	0.33	982
OHSU_SR4	Mixed Automatic	0.11	0.15	0.60	0.48	0.31	381
OHSU_SR5	Mixed Automatic	0.11	0.15	0.58	0.52	0.31	514
uwmTextAndImageDistance	Mixed Automatic	0.07	0.09	0.44	0.40	0.27	204
medGIFT_sum_withNegImg	Mixed	0.01	0.03	0.06	0.04	0.05	210

Table 11. Results of average scores from submitted runs for medical ad hoc task.

3.3 Lung nodule detection task (new for 2009)

In 2009, a new task was introduced into ImageCLEF called the lung nodule detection task in 2009. This was part of the automatic image annotation task: 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. The specific goal of the lung nodule task was to compare the performance of lung nodule detection techniques with a gold standard of manually identified nodules. This task used the CT slices from the Lung Imaging Data Consortium (LIDC)[13] which included ground truth in the form of manual annotations. The goal of the task was to create algorithms to automatically detect lung nodules. Although there was initially significant interest in the task as evidenced by the substantial number of registrations (>25 groups), only two groups submitted results with proprietary software from a industrial participant achieving impressive results.

3.4 Automatic medical image annotation task

In 2005, the medical image annotation task was introduced in the ImageCLEF1 challenge. Its main contribution was to provide a resource for benchmarking content-based image classification systems focusing on medical images. Hospitals collect hundreds of imaging data every day, and automatic image annotation can be an important step when searching for images in huge databases. Automatic techniques able to identify acquisition modality, body orientation, body region, and biological system examined could be used for multilingual image annotations as well as for DICOM header corrections in medical image acquisition routine. As the 5th medical image annotation task, the aim for 2009 was defined as a survey of prior task experience. Seven groups participated in the challenge submitting 19 runs. They were asked to train their algorithms on 12,677 images, labeled according to four different settings representing the yearly annotation tasks, and to classify 1,733 images in the four annotation frameworks. The aim was to understand how each strategy answers to the increasing number and unbalancing of classes. Further details can be found in [Tommasi et al. 2009].

3.4.1 Data

As in the past challenge editions, the annotation task was designed on the basis of the IRMA project[14]. For 2009, a database of 12,677 fully classified radiographs, taken randomly from medical routine, was made available as the training set. The test data consisted of 1,733 images. Participants in the medical annotation task were asked to classify the test images according to four different labeling schemes, as used over the past 4 years of the task (see Figure 13). Each group is allowed to submit different runs, but each of them were required to be based only on one algorithm, optimized to tackle the four different classification problems. The aim was to understand how each algorithm answered to the increasing number of classes and to the unbalanced nature of various classification schemes. The classification results were considered per year and the error rate score summed to produce a performance ranking for the submitted runs.

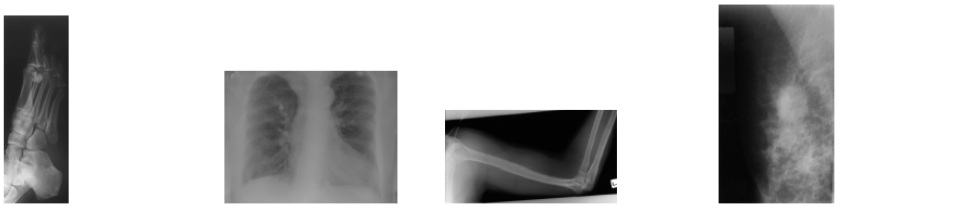


Figure 11: Examples of data from the 2005-2008 medical image annotation data and the label settings.

[13] See http://imaging.cancer.gov/programsandresources/InformationSystems/LIDC
[14] See http://ganymed.imib.rwth-aachen.de/irma/index en.php

3.4.2 Results

In 2009, a total of 7 groups from five nations of two continents participated in the medical annotation task, submitting 19 runs. Table 13 shows the results for the medical image annotation task across the 4 years of the task (2005-2008), sorted by error score sum over the four year label setting (score TOT).

		best				
	# runs	score 05	score 06	score 07	score 08	score TOT
TAU	1	365	263	64.30	169.50	852.80
Idiap	4	393	260	67.23	178.93	899.16
FEITIJS	1	549	433	128.10	242.46	1352.56
VPA	5	578	462	155.05	261.16	1456.21
MedGIFT	3	618	507	190.73	317.53	1633.26
IRMA	1	790	638	207.55	359.29	1994.84
DEU	4	1368	1183	487.50	642.50	3681.00

Table 12. Results from the medical image annotation task.

Overall, the results from this task showed that: the top performing runs do not consider the hierarchical structure of the task, local features always outperform global features, discriminative Support Vector Machine (SVM) classification methods outperform other approaches, in the 2005-06 data there is a decrease in error score (57 classes difficult to model), in 2007-08 there is an increase in the error score which is most likely due to the increasing number of classes and the unbalanced nature of the classification scheme.

3.5 Robot vision task (new for 2009)

The robot vision task was proposed to the ImageCLEF participants for the first time in 2009. The task attracted a considerable attention, with 19 research groups signing up for the task, 7 groups eventually participating and a total of 27 submitted runs. The task addressed the problem of visual place recognition applied to robot topological localization. Specifically, participants were asked to classify rooms on the basis of image sequences, captured by a perspective camera mounted on a mobile robot. The sequences were acquired in an office environment, under varying illumination conditions and across a time span of almost two years (see Figure 14).

Figure 12: Examples of data from the 2009 robot vision task.

The training and validation set consisted of a subset of the IDOL2 database. The test set consisted of sequences similar to those in the training and validation set, but acquired 20 months later and imaging also additional rooms. Participants were asked to build a system able to answer the question "where are you?" (I am in the kitchen, in the corridor, etc) when presented with a test sequence imaging rooms seen during training, or additional rooms that were not imaged in the training sequence. The system had to assign each test image to one of the rooms present in the training sequence, or indicate that the image came from a new room. All participants were asked to solve the problem separately for each test image (obligatory task). Additionally, results could also be reported for algorithms exploiting the temporal continuity of the image sequences (optional task). Of the 27 runs, 21 were submitted to the obligatory task, and 6 to the optional task. The best result in the obligatory task was obtained by the Multimedia Information Retrieval Group of the University of Glasgow, UK with an approach based on local feature matching. The best result in the optional task was obtained by the Intelligent Systems and Data Mining Group (SIMD) of the University of Castilla-La Mancha, Albacete, Spain, with an

approach based on local features and a particle filter. See the task overview paper for further details [Caputo et al. 2009].

3.6 Large-scale visual concept detection and annotation task

The large-scale visual concept detection and annotation task (LS-VCDT) in ImageCLEF 2009 aimed to detect (and annotate) up to 53 depicted visual concepts from consumer photos. The concepts were additionally structured into ontological form, implying a hierarchical ordering which could be utilised during the training phase for automatic annotation and classification of photos. The dataset consisted of 18,000 Flickr photos (5,000 photos were used for training and 13,000 for testing), which were manually annotated with 53 concepts. A total of 19 research groups participated in the task and submitted 73 runs. Two evaluation measures were applied to submissions to this task: evaluation per concept and evaluation per photo. The evaluation per concept was performed by calculating the Equal Error Rate (EER) and the Area Under Curve (AUC). For evaluation per photo, a recently proposed hierarchical measure was utilized that takes the hierarchy and the relations of the ontology into account and calculates a score per photo. For the concepts, an average AUC of 84% could be achieved, including concepts with an AUC of 95%. The classification performance for each photo ranged between 69% and 100% with an average score of 90%. Further details can be found in the task overview paper [Nowak and Dunker 2009].

3.6.1 Data

The dataset used for this task was 18,000 photos from Flickr, a large-scale photo sharing website. A classification scheme for tagging photos was applied manually to the 25,000 photos to create a training set of 5,000 photos used for training and 13,000 for testing (see Figure 15 for an example annotated photo). For the annotation stage, 43 people were used with 3 people annotating the same photo to ensure consistency in tagging and enable computation of inter-annotator agreement measures.

Figure 13: Example annotated photo from the large-scale image annotation dataset.

3.6.2 Results

A total of 19 groups submitted results to this task and two evaluation measures were computed from these submissions: evaluation per concept and evaluation per photo. The evaluation per concept was performed by calculating the Equal Error Rate (EER) and the Area Under Curve (AUC). For evaluation per photo, a recently proposed hierarchical measure was utilized that takes the hierarchy and the relations of the ontology into account and calculates a score per photo. Table 14 shows the results from the large-scale concept detection and annotation task for 2009.

		Best Run			Average Runs		
Team	Runs	Rank	EER	AUC	Rank	EER	AUC
ISIS	5	1	0.234	0.839	3.2	0.240	0.833
LEAR	5	5	0.249	0.823	13.2	0.268	0.798
CVIUI2R	2	7	0.253	0.814	9.0	0.255	0.813
FIRST	4	8	0.254	0.817	10.5	0.258	0.803
XRCE	1	14	0.267	0.803	14.0	0.267	0.803
bpacad	5	17	0.292	0.773	20.6	0.312	0.746
MMIS	5	21	0.312	0.744	27.8	0.345	0.699
IAM Southampton	3	23	0.330	0.715	24.7	0.335	0.709
LSIS	5	24	0.331	0.721	42.2	0.418	0.602
LIP6	5	33	0.372	0.673	42.0	0.414	0.554
MRIM	4	34	0.384	0.643	38.0	0.415	0.584
AVEIR	4	41	0.441	0.551	49.8	0.461	0.548
Wroclaw University	5	43	0.446	0.221	45.4	0.449	0.200
KameyamaLab	5	47	0.452	0.164	53.4	0.466	0.133
UAIC	1	54	0.479	0.106	54.0	0.479	0.106
apexlab	3	56	0.483	0.070	60.3	0.487	0.078
INAOE TIA	5	57	0.485	0.099	61.0	0.489	0.080
Random	1	-	0.500	0.499	-	0.500	0.499
CEA LIST	4	68	0.500	0.469	69.5	0.502	0.463
TELECOM ParisTech	2	72	0.526	0.459	72.5	0.527	0.459

Table 13. Summary of the results for the evaluation per concept. The table shows the EER and AUC for the best run per group and the averaged EER and AUC for all runs of one group.

The results show that in average the task could be solved reasonably well with the best system achieving an AUC of 84% for all photos. Four other groups got an AUC score over or equal to 80%. Evaluated on the concept basis, the concepts could be annotated in average with an AUC of 84%. In terms of HS, the best system annotated all photos with an average annotation rate of 83%. Three other systems were very close to these results with 83%, 82% and 81%. Part of the groups used the ontology for post-processing or to learn correlations of concepts. No participant integrated the ontology in a reasoning system and tried to apply this system for the classification task. The large number of concepts and photos posed no problem to the classification systems.

3.7 Wikipedia multimedia task (WikipediaMM)

The WikipediaMM task provides a testbed for system-oriented evaluation of multimedia information retrieval from a collection of Wikipedia images. The task aims to investigate retrieval approaches in the context of a large and heterogeneous collection of images (similar to those encountered on the Web) that are commonly searched for by users with diverse information needs. The evaluation scenario is thereby similar to the classic TREC ad-hoc retrieval task and the ImageCLEF photo retrieval task: simulation of the situation in which a system knows the set of documents to be searched, but cannot anticipate the particular topic that will be investigated (i.e. topics are not known to the system in advance). Given a multimedia query that consists of a title and one or more sample images describing a user's multimedia information need, the aim is to find as many relevant images as possible from the (INEX MM) Wikipedia image collection. A multi-modal retrieval approach in that case should be able to combine the relevance of different media types into a single ranking that is presented to the user. For further information see the task overview paper [Tsikrika and Kludas 2009].

3.7.1 Data

The dataset for this task consists of 151,590 JPEG and PNG images from Wikipedia (the INEX MM Wikipedia image collection). Each image is associated with user-generated alphanumeric, semi-structured metadata in English. These metadata usually contain a brief caption or description of the image, the Wikipedia user who uploaded the image and the copyright information. These descriptions are highly heterogeneous and of varying length.

A total of 45 topics were created based on: (i) analysis from an image search engine, and (ii) suggested by participants in the 2009 task. Topics range from easy (e.g. "bikes") to difficult topics which are highly semantic (e.g. "aerial photos of non-artificial landscapes"), which provide a challenge for state-of-the-art retrieval algorithms. The topics consist of a title, narrative and example images (Figure 16) with an average of 1.85 images/topic, 36 relevant documents per topic and 37.96 words per relevant

document. Participants were asked to assist with generating topics and carrying out relevance assessments.

<topic>

<number>108 </number>

<title> bird nest </title>

<image> http://chiotsrun.com/.../birds-nest-at-moms.jpg </image>

<image> http://farm1.static.flickr.com/...21a4.jpg </image>

<narrative> We search for photos of birds and their nests, empty nests are also relevant. Food that has 'nest' in their name is not relevant, neither is the Beijing Olympic stadium. </narrative>

</topic>

Figure 14: Example topic from the WikipediaMM task.

3.7.2 Results

A total of 32 groups registered for the task with 8 groups finally submitting a total of 57 runs (CEA (LIC2M-CEA, Centre CEA de Saclay, France), DCU (Dublin City University, School of Computing, Ireland), DEU (Dokuz Eylul University, Department of Computer Engineering, Turkey), IIIT-Hyderabad (Search and Info Extraction Lab, India), LaHC (Laboratoire Hubert Curien, UMR CNRS, France), SZTAKI (Hungarian Academy of Science, Hungary), SINAI (Intelligent Systems, University of Jaen, Spain) and UALICANTE (Software and Computer Systems, University of Alicante, Spain). Table 15 shows the results of the task ordered by descending order of MAP. The highest scoring group obtained an average MAP score of 0.2397 using text-only retrieval and some form of query expansion (QE). This is perhaps surprising given that highly semantic multimedia topics were developed with the aim to encourage and show the potential of multi-modal approaches. A total of 29/57 groups combined textual and visual features for retrieval. It is worth noting though that all of the participants that submitted both mono-media and multi-modal runs achieved their best results with their multimodal runs.

	Participant	Run	Modality	Feedback/Expansion	MAP	P@10	P@20	R-prec.	Bpref
1	deuceng	deuwiki2009_205	TXT	QE	0.2397	0.4000	0.3133	0.2683	0.2191
2	deuceng	deuwiki2009_204	TXT	QE	0.2375	0.4000	0.3111	0.2692	0.2170
3	deuceng	deuwiki2009_202	TXT	QE	0.2358	0.3933	0.3189	0.2708	0.2217
4	lach	run_TXTIMG_100_3_1_5_meanstd	TXTIMG	NOFB	0.2178	0.3378	0.2811	0.2538	0.2006
5	lach	run_TXTIMG_50_3_1_5_meanstd	TXTIMG	NOFB	0.2148	0.3356	0.2867	0.2536	0.2023
6	cea	cealateblock	TXTIMG	QE	0.2051	0.3622	0.2744	0.2388	0.1938
7	cea	ceaearlyblock	TXTIMG	QE	0.2046	0.3556	0.2833	0.2439	0.2014
8	cea	ceabofblock	TXTIMG	QE	0.1975	0.3689	0.2789	0.2342	0.1886
9	cea	ceatlepblock	TXTIMG	QE	0.1959	0.3467	0.2733	0.2236	0.1847
10	cea	ceabofblockres	TXTIMG	QE	0.1949	0.3689	0.2789	0.2357	0.1890
11	cea	ceatlepblockres	TXTIMG	QE	0.1934	0.3467	0.2733	0.2236	0.1847
12	lach	run_TXTIMG_Siftdense$_{0}$.084	TXTIMG	NOFB	0.1903	0.3111	0.2700	0.2324	0.1828
13	lach	run_TXT_100_3_1_5	TXT	NOFB	0.1890	0.2956	0.2544	0.2179	0.1687
14	lach	run_TXT_50_3_1_5	TXT	NOFB	0.1880	0.3000	0.2489	0.2145	0.1715
15	ualicante	Alicante-MMLCA	TXTIMG	FB	0.1878	0.2733	0.2478	0.2138	0.1734

Table 14. Summary of the results for the WikipediaMM 2009 task.

4 iCLEF

iCLEF is the interactive track of CLEF (Cross-Language Evaluation Forum), an annual evaluation exercise for Multilingual Information Access systems. 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.

Since 2006, iCLEF has moved away from news collections (a standard for text retrieval experiments) in order to explore user behavior in scenarios where the necessity for cross-language search arises more naturally for the average user. We chose Flickr[15], a large-scale, web-based image database based on a large 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, question answering or text-based image retrieval. 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, 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.

The main novelty of the iCLEF 2008 shared experience, which has been kept in 2009, was to focus on the shared analysis of a large search log from a single search interface provided by the iCLEF organizers. The focus is, therefore, on search log analysis rather than on system design. The idea is to study the behavior of users in an (almost) naturalistic search scenario, having a much larger data set than in previous iCLEF campaigns. The search interface provided by iCLEF organizers is a basic cross-language retrieval system to access images in Flickr, presented as an online game: the user is given an image, and she must find it again 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.

The iCLEF 2009 task is the same as in 2008, the only difference being the approach to select the target images (the topics for our task). In 2008 a large log was harvested, but in over half of the search sessions the user had active language skills in the target language, and the situations where the user has only passive or no abilities in the target language were underrepresented. The reason was that many images in the target set had annotations in English (plus other languages in many cases), and the set of users (over 200 active searchers) tend to have English as a native or at least as a well-known language. Therefore, this year we explicitly avoided images annotated in English to increase the chances of having search sessions in unknown languages.

4.1 Task guidelines

The task is the same as in 2008, and the differences lie in the search log collected (target images, set of registered users, etc.)

4.1.1 Search task definition

First of all, the decision to use Flickr as the target collection is based on (i) the inherent multilingual nature of the database, provided by tagging and commenting features utilized by a worldwide network of users, (ii) although it is in constant evolution, which may affect reproducibility of results, the Flickr search API allows the specification of timeframes (e.g. search in images uploaded between 2004 and 2007), which permits defining a more stable dataset for experiments; and (iii) the Flickr search API provides a stable service which supports full Boolean queries, something which is essential to perform cross-language searches without direct access to the index.
For 2008, our primary goal was harvesting a large search log of users performing multilingual searches on the Flickr database. Rather than recruiting users (which inevitably leads to small populations), we wanted to publicize the task and attract as many users as possible from all around the world, and engage them with search. To reach this goal, we needed to observe some restrictions:

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

- The search task should be clear and simple, requiring no a-priori training or reading for the casual user.
- The search task should be engaging and addictive. Making it an online game - with a rank of users - helps achieve that, with the rank providing a clear indication of success.
- There should be no need for manual judgments in order to establish the success of a search session, in order to avoid discouraging delays in the online game rankings.
- It should have an adaptive level of difficulty to prevent novice users from being discouraged, and to prevent advanced users from being unchallenged.
- The task should be naturally multilingual.

We decided to adopt a known-item retrieval search task: the user is given a raw (un-anotated) image and the goal is to find the image again in the Flickr database, using a multilingual search interface provided by iCLEF organizers. The user does not know in advance in which languages the image is annotated; therefore searching in multiple languages is essential to get optimal results. Although the task is probably not the most natural one (thematic-based searches are probably more common than "stuff I've seen before" search needs), it has the definitive advantage of not requiring manual judgments, and that makes possible to keep an instantly updated user ranking.
Indeed the task is organized as an online game: the more images found, the higher a user is ranked. In case of ties, the ranking will also depend on precision (number of images found /number of images attempted). At any time the user can see the “Hall of Fame" with a rank of all registered users.
Depending on the image, the source and target languages, this can be a very challenging task. To have an adaptive level of difficulty, we implemented a hints mechanism. At any time whilst searching, the user is allowed to quit the search (skip to next image) or ask for a hint. The first hint is always the target language (and therefore the search becomes mono or bilingual as opposed to multilingual). The rest of the hints are keywords used to annotate the image. Each image found scores 25 points, but for every hint requested, there is a penalty of 5 points. The hint mechanism proved essential to engage users in 2008 and even more in 2009 (for reasons explained later).
Initially a five minute time limit per image was considered, but initial testing indicated that such a limitation was not natural and had a deep impact on users' search behavior. Therefore we decided to remove time restrictions from the task definition.

4.1.2 Search interface

We designed the so-called Flickling interface to provide a basic cross-language search front-end to Flickr. Its basic functionalities are:

- User registration, which records the user's native language and language skills in each of the six European languages considered (EN, ES, IT, DE, NL, FR).
- Localization of the interface in all six languages.
- Two search modes: mono and multilingual. The latter takes the query in one language and returns search results in up to six languages, by launching a full Boolean query to the Flickr search API.
- Cross-language search is performed via term-to-term translations between six languages using free dictionaries (taken from: http://xdxf.revdanica.com/down).
- A term-to-term automatic translation facility which selects the best target translations according to (i) string similarity between the source and target words; (ii) presence of the candidate translation in the suggested terms offered by Flickr for the whole query; and (iii) user translation preferences.
- A query translation assistant that allows users to pick/remove translations, and add their own translations (which go into a personal dictionary). We did not provide back-translations to support this process, in order to study correlations between target language abilities (active, passive, none) and selection of translations.

- A query refinement assistant that allows users to refine or modify their query with terms suggested by Flickr and terms extracted from the image rank. When the term is in a foreign language, the assistant tries to display translations into the user's preferred language to facilitate feedback.
- Control of the game-like features of the task: user registration and user profiles, groups, ordering of images, recording of session logs and access to the hall of fame.
- Post-search questionnaires (launched after each image is found or failed) and final questionnaires (launched after the user has searched fifteen images, not necessarily at the end of the experience).

Figure 17 shows a snapshot of the search interface, localized in Spanish.

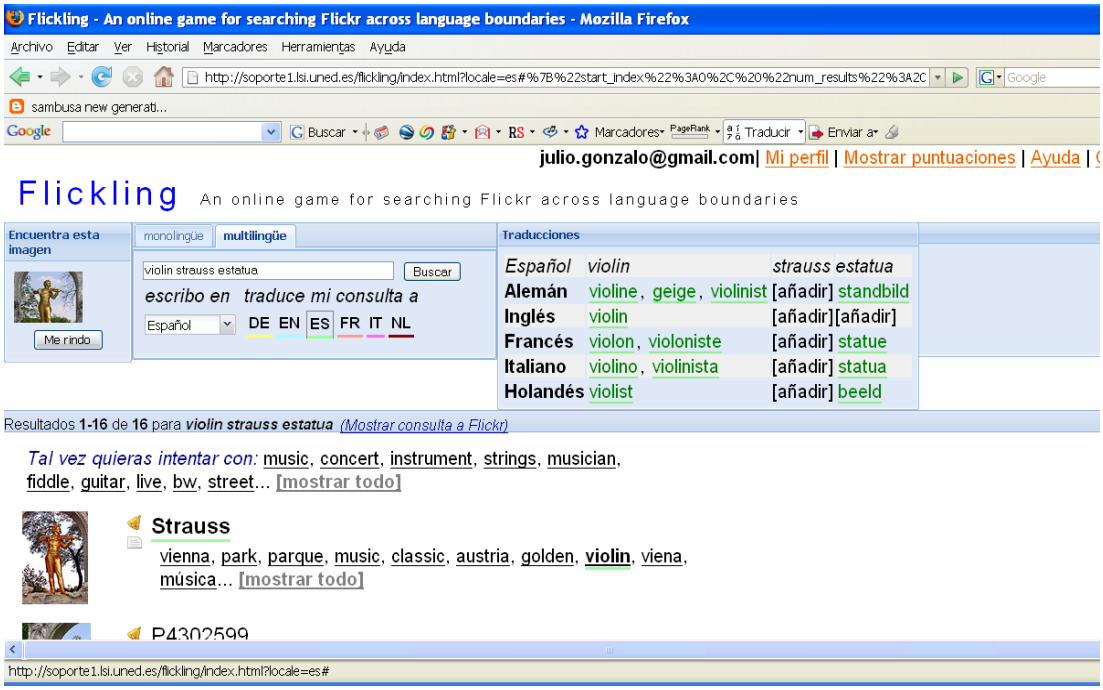


Figure 17: A snapshot of the Flickling Multilingual Search front end for Flickr

4.1.3 Participation in the track

As in 2008, iCLEF 2009 participants can essentially adopt two types of methodology: (1) analyze log files based on all participating users (which is the default option) and, (2) perform their own interactive experiments with the interface provided by the organizers. CLEF individuals registered in the interface as part of a team, so that a ranking of teams is produced in addition to a ranking of individuals.

4.1.4 Generation of search logs

Participants can mine data from the search session logs, for example looking for differences in search behavior according to language skills, correlations between search success and search strategies, etc.

4.1.5 Interactive experiments

Participants can recruit their own users and conduct their own experiments with the interface. For instance, they could recruit a set of users with passive language abilities and another with active abilities in certain languages and, besides studying the search logs, they could perform observational studies on how they search, conduct interviews, etc. iCLEF organizers provided assistance with defining appropriate user groups and image lists, for example, within the common search interface. Besides these two options, and given the community spirit of iCLEF, we were open to groups having

their own plans (e.g. testing their own interface designs or using a specific set of images) as long as they did not change the overall shared search task (known-item search on Flickr).

4.2 Dataset: Flickling search logs

Search logs were harvested from the Flickling search interface between May and June 2009. In order to entice a large set of users, the "CLEF Flickr Challenge" was publicized in Information Access forums (e.g. the SIG-IR and CLEF lists), Flickr blogs and general photographic blogs. As in 2008, we made a special effort to engage the CLEF community in the experience, with the goal of getting researchers closer to the CLIR problem from a user's perspective. To achieve this goal, CLEF organizers agreed to award two prizes consisting of free registrations for the workshop: one for the best individual searcher and one for the best scoring CLEF group.
Overall, 130 users registered for the task, for a total of 2527 search sessions, many of them ending in success (2149). There were 19 native languages in our user set, with this distribution: 46 Spanish, 38 Romanian, 10 English, 9 Italian, 4 Persian/Farsi, 4 German, 3 Chinese, 2 Finnish, 2 Catalan, 2 Basque, 2 Arabic, 1 Danish, 1 Vietnamese, 1 Malay, 1 Russian, 1 Greek and 1 Belarusian.
Apart from general users, the group affiliation revealed two dominant user profiles: university researchers and students (most of them in Computer Science) and photography fans.
The 2008 search log was skewed towards "active" search sessions (where users had active skills in some of the languages used to annotate the image). Therefore this year we changed the methodology to select the target images, excluding those which had annotations in English, and reducing the number of images annotated in Spanish (because it was a well represented native language in our user base). The strategy was – too – successful: we harvested 1585 search sessions where the target language was unknown to the user, 18 where the user had passive abilities (i.e. could read results but not write queries), and none where the user had active skills in the target language. That makes this search log an excellent tool to study the behavior of users searching in foreign language, but it can hardly be used to compare the three profiles. We also found that the combination of users and images is so different from the 2008 experience that merging the two search logs, even if the task is the same, is not advisable.
Overall, it has been possible to collect a large controlled multilingual search log, which includes both search behavior (interactions with the system) and users' subjective impressions of the system (via questionnaires). This offers a rich source of information for helping to understand multilingual search characteristics from a user's perspective.

4.3 Participation and findings

Six sites submitted results for this year's interactive track: two newcomers (University of North Texas and Alexandru Ioan Cuza University, UAIC, in Romania) and four groups with previous experience in iCLEF: Universidad Nacional de Educación a Distancia (UNED), the Swedish Institute of Computer Science (SICS), Manchester Metropolitan University (MMU), and the University of Alicante.

University of Alicante [Navarro et al. 2009] investigated whether there is a correlation between lexical ambiguity in queries and search success and, if so, whether explicit Word Sense Disambiguation can potentially solve the problem. To do so, they mined data from the search log distributed by the iCLEF organization, and found that less ambiguous queries lead to better search results and coarse-grained Word Sense Disambiguation might be helpful in the process.

UAIC [Cristea et al. 2009] tried to find correlations between different search parameters using a subset of the search log consisting of searchers performed by a set of 31 users recruited for the task (which were very active, performing almost 46% of all queries in the general search log). They did not find a clear connection between the results of over-achieving users and their particular actions, and they found hints of a possible (light) collaboration between them, which eventually makes our search log less reliable than initially thought.

Manchester Metropolitan University [Vassilakaki et al. 2009] tried to demonstrate the value in focusing on user's trust and confidence in the exploration of seeking behavior to reveal users' perception of the tasks involved when searching across languages. Instead of focusing on log analysis, MMU recruited their own set of 24 users selected a specific set of three images (in Dutch, German and Spanish) and performed a qualitative and quantitative analysis including questionnaires, observational study of the search sessions, retrospective thinking aloud and interviews. Among other things, they found that variations in perceptions of searching and approach to using translations which is unrelated to the amount or type of help or guidance given. They also found that, in general, users only think about languages after asking for the first hint (i.e. the target language), and facing cross-linguality only when it is inevitable.

UNED [Peinado et al. 2009] tried to establish differences between users with active/passive/no knowledge of the target language, including search success and cognitive effort, and compared the results using search logs from 2008 and 2009. Unfortunately the skewed distribution of language profiles in 2009 did not permit direct comparisons and made results from the merged logs unreliable. UNED then worked on establishing successful search strategies when searching in foreign, unknown language. They found that the usage of cross-language search assistance features has an impact on search success, and that such features are highly appreciated by users.

University of North Texas [Ruiz and Chin 2009] aimed at understanding the challenges that users face when searching for images that have multilingual annotations, and how they cope with these challenges to find the information they need. Similarly to MMU, instead of using the search log this group recruited their own set of six North American students and studied their search behavior and subjective impressions using questionnaires, training, interviews and observational analysis. They found that users have strong difficulties using Flickr tags, particularly when doing cross-language search, and that their typical session requires two hints: the target language and a keyword.

SICS has continued to investigate methods for how to study confidence and satisfaction of users. In previous years' studies, results have been somewhat equivocal; this year, some preliminary studies of the number of reformulations versus success rate have been performed. The SICS team found that the length of query sequences which eventually were successful were longer, indicating persistence when a search appears to be in the right direction. The number of query reformulations also correlates well with success: successful query sequences are a result of active exploration of the query space. But for users who persist in working with monolingual searches (search calls), the SICS team found that queries, firstly tended to be vastly less often reformulated to begin with, and that the successful sequences were more parsimonious than the failed ones (conversely from the clsearch calls): instead the number of scroll actions were much more frequent. This would seem to indicate that if users are fairly confident of a well put query, they will persist by scrolling through result lists.

4.4 Overall Comments about iCLEF

iCLEF 2009 has continued to run a large-scale interactive experiment as an online game to generate log files for further study. A default multilingual information access system developed by the organizers was provided to participants to lower the cost of entry and generate search logs recording user's interaction with the system and qualitative feedback about the search tasks and system (through online questionnaires). In addition, two groups have decided to replace (or extend) log analysis by recruiting their own set of users and employ the usual methodology (training, questionnaires, interviews, retrospective thinking aloud, observational studies) on them.
The search logs generated by the iCLEF track in 2008 and 2009 together are a reusable resource for future user-orientated studies of cross-language search behavior, and we hope to see new outcomes in the near future coming from in-depth analysis of our logs. Researchers interested in this resource might contact the iCLEF organization (see http://nlp.uned.es/iCLEF) for details.

5 ResPubliQA 2009: Multilingual Question Answering over European Legislation

5.1 Introduction

The ResPubliQA 2009 exercise is aimed at retrieving answers to a set of 500 questions. The answer of a question is a paragraph of the test collection. The hypothetical user considered for this exercise is a person interested in making inquiries in the law domain, specifically on the European legislation. The ResPubliQA document collection is a subset of JRC-Acquis, a corpus of European legislation that has parallel translations aligned at document level in many European languages.

In the ResPubliQA 2009 exercise, participating systems could perform the task in any of the following languages: Basque (EU), Bulgarian (BG), English (EN), French (FR), German (DE), Italian (IT), Portuguese (PT), Romanian (RO) and Spanish (ES). All the monolingual and bilingual combinations of questions between the languages above were activated, including the monolingual English (EN) task – usually not proposed in the QA track at CLEF. Basque (EU) was included exclusively as a source language, as there is no Basque collection available - which means that no monolingual EU-EU sub-task could be enacted.

5.2 Task Objectives

The general objectives of the exercise are:

1. **Moving towards a domain of potential users.** While looking for a suitable context, improving the efficacy of legal searches in the real world seemed an approachable field of study. The retrieval of information from legal texts is an issue of increasing importance given the vast amount of data which has become available in electronic form over the last few years.

   Moreover, the legal community has showed much interest in IR technologies as it has increasingly faced the necessity of searching and retrieving more and more accurate information from large heterogeneous electronic data collections with a minimum of wasted effort.

   In confirmation of the increasing importance of this issue, a Legal Track [Tomlinson et al. 2007], aimed at advancing computer technologies for searching electronic legal records, was also introduced in 2006 as part of the yearly TREC conferences sponsored by the National Institute of Standards and Technology (NIST)[16]. The task of the Legal Track is to retrieve all the relevant documents for a specific query and to compare the performances of systems operating in a setting which reflects the way lawyers carry out their inquiries.

2. **Studying if current QA technologies tuned for newswire collections and Wikipedia can be easily adapted to a new domain (law domain in this case).** It is not clear if systems with good performance in newswire collections, after many years spent adapting the system to the same collections, perform well in a new domain. In this sense, the task is a new challenge for both, seniors and newcomers.

3. **Moving to an evaluation setting able to compare systems working in different languages.** Apart from the issue of domain, a shortcoming of previous QA campaigns at CLEF was that each target language used a different document collection. This meant that the questions for each language had to be different and as a consequence the performance of systems was not directly comparable unless they happened to work with the same target language.

   In the current campaign, this issue was addressed by adopting a document collection which is parallel at the document level in all the supported languages. This meant that for the first time, all participating systems were answering the same set of questions even though they might be using different languages.

[16] It may be interesting to know that in 2008 the TREC QA Track moved to the Text Analysis Conference (TAC). In 2009 no QA Track has been proposed at any conferences sponsored by NIST.

4. **Comparing current QA technologies with pure Information Retrieval (IR) approaches.** Returning a complete paragraph instead of an exact answer allows the comparison between pure IR approaches and current QA technologies. In this way, a nice benchmark for evaluating IR systems oriented to high precision, where only one paragraph is needed, has been also created. The documents are nicely divided into XML paragraph marks solving the technical issues for paragraph retrieval. Furthermore, a paragraph is presumably a more realistic output for the users of the new collection domain.
5. **Allowing more types of questions.** Returning one paragraph allows new types of questions with the only restriction that they must be answered by a single paragraph.
6. **Introducing in QA systems the Answer Validation technologies developed in the past campaigns.** During the last campaigns we wanted to stick to the easiest and most comprehensible evaluation of systems, that is, requesting only one answer per question and counting the proportion of questions correctly answered (namely accuracy). In this campaign, we wanted to introduce a more discriminative measure, allowing systems to leave some questions unanswered. Given two systems that answer correctly the same proportion of questions, the one that returns less incorrect answers (leaving some questions unanswered) will score better. Thus, systems can add a final module to decide whether they found enough evidence or not to return their best answer.

   This is a classification problem that takes advantage of more sophisticated Answer Validation technologies developed during the last years [Peñas et al. 2007, 2008; Pérez et al. 2009].

5.3 Document Collection

The ResPubliQA collection is a subset of the JRC-ACQUIS Multilingual Parallel Corpus[17]. JRC-Acquis is a freely available parallel corpus containing the total body of European Union (EU) documents, mostly of legal nature. It comprises contents, principles and political objectives of the EU treaties; the EU legislation; declarations and resolutions; international agreements; and acts and common objectives. Texts cover various subject domains, including economy, health, information technology, law, agriculture, food, politics and more. This collection of legislative documents currently includes selected texts written between 1950 and 2006 with parallel translations in 22 languages. The corpus is encoded in XML, according to the TEI guidelines.
The ResPubliQA collection in 8 of the languages involved in the track - Bulgarian, English, French, German, Italian, Portuguese, Romanian and Spanish - consists of roughly 10,700 parallel and aligned documents per language. The documents are grouped by language, and inside each language directory, documents are grouped by year. All documents have a numerical identifier called the CELEX code, which helps to find the same text in the various languages. Each document contains a header (giving for instance the download URL and the EUROVOC codes) and a text (which consists of a title and a series of paragraphs).

5.4 Types of Questions

The questions fall into the following categories: Factoid, Definition, Reason, Purpose, Procedure.

Factoid. Factoid questions are fact-based questions, asking for the name of a person, a location, the extent of something, the day on which something happened, etc. For example:

Q: *When must animals undergo ante mortem inspection?*
A: 9. Animals must undergo ante mortem inspection on the day of their arrival at the slaughterhouse. The inspection must be repeated immediately before slaughter if the animal has been in the lairage for more than twenty-four hours.

[17] Please note that it cannot be guaranteed that a document available on-line exactly reproduces an officially adopted text. Only European Union legislation published in paper editions of the Official Journal of the European Union is deemed authentic.

Q: *In how many languages is the Official Journal of the Community published?*

A: The Official Journal of the Community shall be published in the four official languages.

Definition. Definition questions are questions such as "What/Who is X?", i.e. questions asking for the role/job/important information about someone, or questions asking for the mission/full name/important information about an organization. For example:

Q: *What is meant by "whole milk"?*

A: 3. For the purposes of this Regulation, 'whole milk' means the product which is obtained by milking one or more cows and whose composition has not been modified since milking.

Q: *What does IPP denote in the context of environmental policies?*

A: Since then, new policy approaches on sustainable goods and services have been developed. These endeavours undertaken at all political levels have culminated in the Green Paper on Integrated Product Policy(1) (IPP). This document proposes a new strategy to strengthen and refocus product-related environmental policies and develop the market for greener products, which will also be one of the key innovative elements of the sixth environmental action programme - Environment 2010: "Our future, our choice".

Reason. Reason questions ask for the reasons/motives/motivations for something happening. For example:

Q: *Why should the Regulation (EC) 1254 from 1999 be codified?*

A: (1) Commission Regulation (EC) No 562/2000 of 15 March 2000 laying down detailed rules for the application of Council Regulation (EC) No 1254/1999 as regards the buying-in of beef [2] has been substantially amended several times [3]. In the interests of clarity and rationality the said Regulation should be codified.

Q: *Why did a Commission expert conduct an inspection visit to Uruguay?*

A: A Commission expert has conducted an inspection visit to Uruguay to verify the conditions under which fishery products are produced, stored and dispatched to the Community.

Purpose. Purpose questions ask for the aim/goal/objective of something. For example:

Q: *What is the purpose of the Agreement of Luxembourg?*

A: RECALLING the object and purpose of the Agreement of Luxembourg to preserve the existing regime between the five Nordic States pursuant to the Convention on the Abolition of Passport Controls at Intra-Nordic borders signed in Copenhagen on 12 July 1957, establishing the Nordic Passport Union, once those of the Nordic States which are Members of the European Union take part in the regime on the abolition of checks on persons at internal borders set out in the Schengen agreements;"

Q: *What is the overall objective of the eco-label?*

A: The overall objective of the eco-label is to promote products which have the potential to reduce negative environmental impacts, as compared with the other products in the same product group, thus contributing to the efficient use of resources and a high level of environmental protection. In doing so it contributes to making consumption more sustainable, and to the policy objectives set out in the Community's sustainable development strategy (for example in the fields of climate change, resource

efficiency and eco-toxicity), the sixth environmental action programme and the forthcoming White Paper on Integrated Product Policy Strategy.

Procedure. Procedure questions ask for a set of actions which is the official or accepted way of doing something. For example:

Q: *How are stable conditions in the natural rubber trade achieved?*

A: To achieve stable conditions in natural rubber trade through avoiding excessive natural rubber price fluctuations, which adversely affect the long-term interests of both producers and consumers, and stabilizing these prices without distorting long-term market trends, in the interests of producers and consumers;

Q: *What is the procedure for calling an extraordinary meeting?*

A: 2. Extraordinary meetings shall be convened by the Chairman if so requested by a delegation.

Q: *What is the common practice with shoots when packing them?*

A: (2) It is common practice in the sector to put white asparagus shoots into iced water before packing in order to avoid them becoming pink."

5.5 Test Set Preparation

Six hundred questions were initially formulated, manually verified against the document collection, translated into English and collected in a common xml format using a web interface specifically designed for this purpose. To avoid a bias towards a language, the 600 questions were developed by 6 different annotators originally in 6 different languages (100 each). All questions had at least one answer in the target corpus of that language.

In order to share them in a multilingual scenario, a second translation into all nine languages of the track was necessary. Native speakers from each language group with a good command of English were recruited and were asked to translate the questions from English back into all the languages of the task. The final pool of 500 questions was selected by the track-coordinators out of the 600 produced, attempting to balance the question set according to the different question types (factoid, definition, reason, purpose and procedure). The need to select questions which had a supported answer in all the collections implied a great deal of extra work for the track coordinators, as a question collected in a language was not guaranteed to have an answer in all other collections.

During the creation of the 100 questions in a source language and their "mapping to English" the question creator was supposed not only to translate the questions into English, but also to look for the corresponding answer at least in the English corpus. After the selection of the final 500 questions, during their translation from English into the other source language, checking the availability of answers for all the questions in all the languages of the parallel corpus ensured that there is no NIL question, as in the previous QA@CLEF editions. The most frequent problematic situations were due to the misalignments between documents at the paragraph level:

- Entire paragraphs missing from one language, but, of course, existing in other(s); for example jrc31982D0886-ro contains only 25 paragraphs, but the English document contains 162 paragraphs, with the text containing an EC Convention, absent from the Romanian version.
- Different paragraph segmentation into different languages of the parallel corpus; for example the document jrc31985L0205-en contains one single paragraph (n="106") corresponding to 685 Romanian paragraphs (n="106_790"). From the point of view of our track, this means that one question having the answer in the (only one) English paragraph had to be removed, since the answer in Romanian is supposed to be found in exactly one paragraph.
- Missing information (parts of the text) in one paragraph; for example a question like "What should be understood by "living plants"?" had answer in English document jrc31968R0234-en paragraph

number 8 “Whereas the production of live trees and other plants, bulbs, roots and the like, cut flowers and ornamental foliage (hereinafter where appropriate called ‘live plants’)”. However, the corresponding Romanian paragraph number 9, does not include the list of the live plants.
- Contradictory information in corresponding paragraphs; for example the corresponding paragraphs that answers the question “How much does cotton increase in weight after treatment with formic acid?” indicate a loss of 3% in the Romanian version, whereas in English the loss is 4%.

5.6 Format

5.6.1 Test set

Test sets for each source language took the form of a UTF-8 xml file containing the following:

```
source_lang  target_lang  q_id  q_string
```

where:

- `source_lang` is the source language
- `target_lang` is the target language
- `q_id` is the question number (4 digits – 0001 to 0500)
- `q_string` is the question (UTF-8 encoded) string

Here are four questions in a hypothetical EN-EN set:

```
<?xml version="1.0" encoding="UTF-8" ?>
<input>
  <q q_id="0001" source_lang="EN" target_lang="EN"> What should the driver of a Croatian heavy goods vehicle carry with him or her?</q>
  <q q_id="0002" source_lang="EN" target_lang="EN"> What will the Commission create under Regulation (EC) No 2422/2001 create? </q>
  <q q_id="0003" source_lang="EN" target_lang="EN"> What convention was done at Brussels on 15 December 1950? </q>
  <q q_id="0004" source_lang="EN" target_lang="EN"> What is another name for ‘rights of transit’?</q>
</input>
```

5.6.2 Submission format

A run submission file for the ResPubliQA task was also an xml file of the form:

```
q_id run_id answered passage-string p_id docid
```

where:

- `q_id` is the question number as given in the test set (of the form 0001 to 0500) Passages must be returned in the same ascending (increasing) order in which questions appear in the test set;
- `run_id` is the run ID an alphanumeric string which identifies the runs of each participant. It should be the concatenation of the following elements: the team ID (sequence of four lower case ASCII characters), the current year (09 stands for 2009), the number of the run (1 for the first one, or 2 for the second one), the task identifier (including both source and target languages, as in the test set).
- `answered` indicates if question has been answered or not. If the value for the attribute "answered" is NO, then the passage string will be ignored;
- `passage_string` is a text string; the entire paragraph which encloses the answer to the question
- `p_id` is the number of the paragraph from which the passage_string has been extracted

- `docid` is the ID of the document

i.e.

```
<?xml version="1.0" encoding="UTF-8" ?>
<output>
<a q_id="0001-0500" run_id="XXXX091XXXX" answered="YES|NO">
<passage_string p_id="11" docid "jrc31960D051-
en.xml">xyz</passage_string>
</a>
</output>
```

As it can be seen, systems were not required to answer all questions. See later for further discussion.

5.7 Evaluation

5.7.1 Responses

In this year's evaluation campaign, participants could consider questions and target collections in any language. Participants were allowed to submit just one response per question and up to two runs per task. Each question had to receive one of the following system responses:

1. A paragraph with the candidate answer. Paragraphs are marked and identified in the documents by the corresponding XML marks.
2. The string `NOA` to indicate that the system preferred not to answer the question.

Optionally, systems that preferred to leave some questions unanswered, could decide to submit also the candidate paragraph. If so, systems were evaluated for the responses they returned also in the cases in which they opted not to answer. This second option was used to additionally evaluate the validation performance.

One of the principles that inspired the evaluation exercise is that leaving a question unanswered has more value than giving a wrong answer. In this way, systems able to reduce the number of wrong answers, by deciding not to respond to some questions are rewarded by the evaluation measure.

However, a system choosing to leave some questions unanswered, returning NOA as a response, must ensure that only the portion of wrong answers is reduced, maintaining as high as possible the number of correct answers. Otherwise, the reduction in the number of correct answers is punished by the evaluation measure.

5.7.2 Assessments

Each run was manually judged by one human assessor for each language group, who considered if the paragraph was responsive or not. Answers were evaluated anonymously and simultaneously for the same question to ensure that the same criteria are being applied to all systems. This year, no second annotation was possible, so no data about the inter-annotator agreement are available.
One of the following judgements was given to each question-answer by human assessors during the evaluation:

- R : the question is answered correctly
- W: the question is answered incorrectly
- U : the question is unanswered

The evaluators were guided by the initial "gold" paragraph, which contained the answers. This "gold" paragraph was only a hint, since there were many cases when:

- correct answers did not exactly correspond to the "gold" paragraph, but the correct information was found in another paragraph of the same document as the "gold" one
- correct answers corresponded to the "gold" paragraph, but were found in another JRC document
- answers were evaluated as correct, even if the paragraphs returned contained more or less information than the "gold" paragraph
- answers from different runs were evaluated as correct, even if they contained different but correct information; for example the question 44 (Which country wishes to export gastropods to the Community?) had Jamaica as the "gold" answer; but in the six runs evaluated, all the answers indicated Chile and Republic of Korea, which were also correct.

5.7.3 Evaluation Measure

The use of Machine Learning-based techniques able to decide if a candidate answer is finally acceptable or not was introduced by the Answer Validation Exercise[18] during the past campaigns. This is an important achievement, as an improvement in the accuracy of such decision-making process leads to more powerful QA architectures with new feedback loops. One of the goals of the ResPubliQA exercise is to effectively introduce these techniques in current QA systems.
For this reason, the unique measure considered in this evaluation campaign was the following:

$$c@1 = \frac{1}{n}(n_R + n_U \frac{n_R}{n})$$

where:

- n_R: is the number of correctly answered questions
- n_U: number of unanswered questions
- n: the total number of questions

Notice that this measure is parallel to the traditional accuracy used in past editions. The interpretation of the measure is the following:

1. A system that gives an answer to all the questions receives a score equal to the accuracy measure used in the previous QA@CLEF main task: in fact, since in this case $n_U = 0$ then $c@1 = n_R/n$;
2. The unanswered questions add value to c@1 only if they do not reduce much the accuracy (i.e. n_R/n) that the system would achieve responding to all questions. This can be thought as a hypothetical second chance in which the system would be able to replace some NoA answers by the corrects one. How many, the same proportion the showed before (i.e. n_R/n).
3. A system that does not respond any question (i.e. returns only NOA as answer) receives a score equal to 0, as $n_R=0$ in both addends.

5.7.4 Tools and Infrastructure

This year, CELCT has developed a series of infrastructures to help the management of the ResPubliQA exercise. We had to deal with many processes and requirements:

- First of all the need to develop a proper and coherent tool for the management of the data produced during the campaign, to store it and to make it re-usable, as well as to facilitate the analysis and comparison of the results.
- Secondly, the necessity of assisting the different organizing groups in the various tasks of the data set creation and to facilitate the process of collection and translation of questions and their assessment.

[18] See `http://nlp.uned.es/clef-qa/ave`

- Finally, the possibility for the participants to directly access the data, submit their own runs (this also implied some syntax checks of the format), and later, get the detailed viewing of the results and statistics.

A series of automatic web interfaces were specifically designed for each of these purposes, with the aim of facilitating the data processing and, at the same time, showing the users only what is important for the task they had to accomplish. So, the main characteristics of these interfaces are the flexibility of the system specifically centred on the user's requirements.

While designing the interfaces for question collection and translation one of the first issues which was to be dealt with, was the fact of having many assessors, a big amount of data, and a long process. So tools must ensure an efficient and consistent management of the data, allowing:

1. Edition of the data already entered at any time.
2. Revision of the data by the users themselves.
3. Consistency propagation ensuring that modifications automatically re-model the output in which they are involved. For example, if a typo is corrected in the Translation Interface, the modification is automatically updated also in the GoldStandard files, in the Test Set files and so on.
4. Statistics and evaluation measures are calculated and updated in real time.

5.8 Participants

11 groups participated with 28 runs. In addition, we evaluated 16 baseline runs (2 per language) based only in pure IR approach, for comparison purposes. All runs were monolingual except two runs Basque-English (EU-EN).

		Target languages (corpus and answer)							
Source languages (questions)		BG	DE	EN	ES	FR	IT	PT	RO
	BG								
	DE		2						
	EN			10					
	ES				6				
	EU			2					
	FR					3			
	IT						1		
	PT								
	RO								4

Table 15. Tasks and corresponding numbers of submitted runs.

The most chosen language appeared to be English with 12 submitted runs, followed by Spanish with 6 submissions. No runs were submitted either in Bulgarian or Portuguese. Participants came above all from Europe, except two different groups from India. Table 1 shows the run distribution in the different languages.

The list of participating systems, teams and the reference to their reports are shown in Table 17.

System	Team	Reference
elix	ELHUYAR-IXA, SPAIN	[Agirre et al. 2009b]
icia	RACAI, ROMANIA	[Ion et al. 2009]
iiit	Search & Info Extraction Lab, INDIA	[Bharadwaj et al. 2009]
iles	LIMSI-CNRS-2, FRANCE	[Moriceau et al. 2009]
isik	ISI-Kolkata, INDIA	-
loga	U.Koblenz-Landau, GERMAN	[Gloeckner and Pelzer 2009]
mira	MIRACLE, SPAIN	[Vicente-Díez et al. 2009]
nlel	U. politecnica Valencia, SPAIN	[Correa et al. 2009]
syna	Synapse Developpment, FRANCE	-
uaic	AI.I.Cuza U. of IASI, ROMANIA	[Iftene et al. 2009]
uned	UNED, SPAIN	[Rodrigo et al. 2009]

Table 16. Systems and teams with the reference to their reports.

5.9 Analysis of Results

5.9.1 IR Baselines

Since there were a parallel collection and one set of questions for all languages, the only variable that did not permit strict comparison between systems was the language itself. Running exactly the same IR system in all languages did not permit to fix this variable but at least we have some evidence about the starting difficulty in each language.
Two baseline runs per language, based on pure Information Retrieval, were prepared and assessed with two objectives:

1. to test how well can a pure Information Retrieval system perform on this task.
2. to compare the performance of more sophisticated QA technologies against a simple IR approach.

These baselines were produced in the following way:

1. Indexing the document collection at the paragraph level. Stopwords were deleted in all cases and the difference between the two runs is the application or not of stemming techniques.
2. Querying with the exact text of each question as a query.
3. Returning the paragraph retrieved in the first position of the ranking as the answer to the question.

The selection of an adequate retrieval model that fits the specific characteristic of the supplied data was a core part of the task. Applying an inadequate retrieval function would return a subset of paragraphs where the answer could not appear, and thus the subsequent techniques applied in order to detect the answer within the subset of candidates paragraphs would fail. For example, we found that simple models as the Vector Space Model or the default model of Lucene are not appropriate for this collection. For this reason, the baselines were produced using the Okapi-BM25 ranking function [10].

Using Okapi-BM25 the selection of the appropriate values for its parameters is crucial for a good retrieval. The parameters were fixed to:

1. b: 0.6. Those paragraphs with a length over the average obtain a slightly higher score.
2. k1: 0.1. The effect of term frequency over final score is minimised.

The same parameters in all runs for all languages were used. For more details about the preparation of these baselines see [Pérez et al. 2009].

5.9.2 Results per language

Tables 18 show systems performance divided by language. The content of the columns is as follows:

- **#R**: Number of questions answered correctly.
- **#W**: Number of questions answered wrongly.
- **#NoA**: Number of questions unanswered.
- **#NoA R**: Number of questions unanswered in which the candidate answer was Right. In this case, the system took the bad decision of leaving the question unanswered.
- **#NoA W**: Number of questions unanswered in which the candidate answer was Wrong. In this case, the system took a good decision leaving the question unanswered.
- **#NoA empty**: Number of questions unanswered in which no candidate answer was given. Since all questions had an answer, these cases were counted as if the candidate answer were wrong for accuracy calculation purpose.
- **c@1**: Official measure as it was explained in the previous section.
- **Accuracy**: The proportion of correct answers considering also the candidate answers of unanswered questions. That is:

$$accuracy = \frac{R + NoA\_R}{N}$$

where N is the number of questions (500).

Besides systems, there are three additional rows in each table:

- **Combination**: is the proportion of questions answered by at least one system or, in other words, the score of a hypothetical system doing the perfect combination of the runs.
- **Base091**: IR baseline as explained above, without stemming.
- **Base092**: IR baseline with stemming.

System	c@1	Accuracy	#R	#W	#NoA	#NoA R	#NoA W	#NoA empty
combination	0.56	0.56	278	222	0	0	0	0
loga091dede	0.44	0.4	186	221	93	16	68	9
loga092dede	0.44	0.4	187	230	83	12	62	9
base092dede	0.38	0.38	189	311	0	0	0	0
base091dede	0.35	0.35	174	326	0	0	0	0

Table 17. Results for German.

The system participating in the German task performed better than the baseline, showing a very good behaviour detecting the questions it could not answer. In 73% of unanswered questions (83% if we consider empty answers) the candidate answer was in fact incorrect. This shows the possibility of system improvement in a short time, adding further processing to the answering of questions predicted as unanswerable.

System	c@1	Accuracy	#R	#W	#NoA	#NoA R	#NoA W	#NoA empty
combination	0.9	0.9	451	49	0	0	0	0
uned092enen	0.61	0.61	288	184	28	15	12	1
uned091enen	0.6	0.59	282	190	28	15	13	0
nlel091enen	0.58	0.57	287	211	2	0	0	2
uaic092enen	0.54	0.52	243	204	53	18	35	0
base092enen	0.53	0.53	263	236	1	1	0	0
base091enen	0.51	0.51	256	243	1	0	1	0
elix092enen	0.48	0.48	240	260	0	0	0	0
uaic091enen	0.44	0.42	200	253	47	11	36	0
elix091enen	0.42	0.42	211	289	0	0	0	0
syna091enen	0.28	0.28	141	359	0	0	0	0
isik091enen	0.25	0.25	126	374	0	0	0	0
iiit091enen	0.2	0.11	54	37	409	0	11	398
elix092euen	0.18	0.18	91	409	0	0	0	0
elix091euen	0.16	0.16	78	422	0	0	0	0

Table 18. Results for English.

The first noticeable result in English is that 90% of questions received a correct answer by at least one system. However, this perfect combination is 50% higher than the best system result. This shows that the task is feasible but the systems still have room for improvement. Nevertheless, 0.6 of c@1 and accuracy is a result aligned with the best results obtained in other tasks of QA in the past campaigns of CLEF.

English results are indicative of the difference between c@1 and Accuracy values. The system uaic092 answered correctly 20 questions less than the baselines. However, this system was able to reduce the number of incorrect answers in a significant way, returning 32 incorrect answers less than the baselines. This behaviour is rewarded by c@1, producing a swap in the rankings (with respect to accuracy) between these two systems.

Another example is given by systems uaic091 and elix091, where the reduction of incorrect answers by uaic091 is significant in the case of with respect to elix091.

Something very interesting in the English runs is that the two best teams (see uned092enen, nlel091enen runs) produced paragraph rankings considering matching n-grams between question and paragraph [Correa et al. 2009]. This retrieval approach seems to be promising, since combined with paragraph validation filters it achieved the best score [Rodrigo et al. 2009] in English.

These two approaches obtained the best score also in Spanish (uned091eses, nlel091eses). Additionally, [Correa et al. 2009] performed second experiment (nlel092eses) that achieved the best result considering the whole parallel collection to obtain a list of answers in different languages (Spanish, English, Italian and French).

System	c@1	Accuracy	#R	#W	#NoA	#NoA R	#NoA W	#NoA empty
combination	0.71	0.71	355	145	0	0	0	0
nlel092eses	0.47	0.44	218	248	34	0	0	34
uned091eses	0.41	0.42	195	275	30	13	17	0
uned092eses	0.41	0.41	195	277	28	12	16	0
base092eses	0.4	0.4	199	301	0	0	0	0
nlel091eses	0.35	0.35	173	322	5	0	0	5
base091eses	0.33	0.33	166	334	0	0	0	0
mira091eses	0.32	0.32	161	339	0	0	0	0
mira092eses	0.29	0.29	147	352	1	0	0	1

Table 19. Results for Spanish

The experiment consisted in searching the questions in all languages, first selecting the paragraph with the highest similarity and then, returning the corresponding paragraph aligned in Spanish. This experiment obtained the best score in Spanish, opening the door to exploit the multilingual and parallel condition of the document collection.

In the case of French, baseline runs obtained the best results. Unexpectedly, Synapse (syna091frfr) usually obtaining the best scores in the news domain, did not perform well in this exercise. This proves that there are difficulties in moving from one domain into another.

System	c@1	Accuracy	#R	#W	#NoA	#NoA R	#NoA W	#NoA empty
combination	0.69	0.69	343	157	0	0	0	0
base092frfr	0.45	0.45	223	277	0	0	0	0
base091frfr	0.39	0.39	196	302	2	2	0	0
nlel091frfr	0.35	0.35	173	316	11	0	0	11
iles091frfr	0.28	0.28	138	362	0	0	0	0
syna091frfr	0.23	0.23	114	385	1	0	0	1

Table 20. Results for French.

System	c@1	Accuracy	#R	#W	#NoA	#NoA R	#NoA W	#NoA empty
combination	0.61	0.61	307	193	0	0	0	0
nlel091itit	0.52	0.51	256	237	7	0	5	2
base092itit	0.42	0.42	212	288	0	0	0	0
base091itit	0.39	0.39	195	305	0	0	0	0

Table 21. Results for Italian.

With respect to Italian (Table 22), the only participant obtained better results than the baselines.

System	c@1	Accuracy	#R	#W	#NoA	#NoA R	#NoA W	#NoA empty
combination	0.76	0.76	381	119	0	0	0	0
icia092roro	0.68	0.52	260	84	156	0	0	156
icia091roro	0.58	0.47	237	156	107	0	0	107
UAIC092roro	0.47	0.47	236	264	0	0	0	0
UAIC091roro	0.45	0.45	227	273	0	0	0	0
base092roro	0.44	0.44	220	280	0	0	0	0
base091roro	0.37	0.37	185	315	0	0	0	0

Table 22. Results for Romanian.

The best system in Romanian [Ion et al. 2009] showed a very good performance compared to the rest of runs, as Table 8 shows. This is a system that uses a sophisticated similarity based model for paragraph ranking, question analysis, classification and regeneration of the question, classification of paragraphs and consideration of the EUROVOC terms associated to each document.

5.9.3 Comparison of results across languages

System	BG	DE	EN	ES	FR	IT	PT	RO
icia092								0.68
nlel092				0.47				
uned092			0.61	0.41				
uned091			0.6	0.41				
icia091								0.58
nlel091			0.58	0.35	0.35	0.52		
uaic092			0.54					0.47
loga091		0.44						
loga092		0.44						
base092	0.38	0.38	0.53	0.4	0.45	0.42	0.49	0.44
base091	0.38	0.35	0.51	0.33	0.39	0.39	0.46	0.37
elix092			0.48					
uaic091			0.44					0.45
elix091			0.42					
mira091				0.32				
mira092				0.29				
iles091					0.28			
syna091			0.28		0.23			
isik091			0.25					
iiit091			0.2					
elix092euen			0.18					
elix091euen			0.16					

Table 23. c@1 in participating systems according to the language.

Strict comparison between systems across languages is not possible without ignoring the language variable. However, this is the first time that systems working in different languages were evaluated with the same questions over the same document collection manually translated into different languages. So, extracting information about which approaches are more promising should be possible. For this purpose, we considered both the systems participating in more than one language and the baseline IR runs for all languages.

Furthermore, the organization did not impose special restrictions to make use of a specific language or a combination of more languages. At the end, it can be said that the system that gave more correct answers and less incorrect ones is the best one, regardless of the language. However, the purpose is to compare approaches and follow the more promising one. Tables 9 and 10 mix all systems in all languages and rank them together in two dimensions, the value of c@1, and the target language.

In the first table (Table 24) systems are ordered by c@1 values. Reading column by column, systems are correctly ordered in each language, except some swaps with respect to the baseline IR suns. Systems icia092, uned and nlel seem to have the more powerful approaches.

In the next table (Table 25) we tried to partially fix the language variable, dividing c@1 values by the score of the best IR baseline system. Values over 1 indicate better performance than the baseline, and values under 1 indicate worse performance than the baseline.

In Table 25, the ranking of systems change, showing that also system loga proposes a promising approach, whereas nlel091 system appears more aligned with the baselines than loga. Of course, this evidence is affected by another variable that must be taken into account before making strong claims, i.e. the baseline itself, which perhaps is not the best approach for all languages (especially agglutinative languages such as German).

System	DE	EN	ES	FR	IT	RO
icia092						1.55
icia091						1.32
nlel092			1.175			
loga091	1.158					
loga092	1.158					
uned092		1.151	1.025			
uned091		1.132	1.025			
nlel091		1.094	0.875	0.78	1.24	
uaic092		1.019				1.07
elix092		0.906				
uaic091		0.83				1.02
mira091			0.8			
elix091		0.792				
mira092			0.725			
iles091				0.62		
syna091		0.528		0.51		
isik091		0.472				
iiit091		0.377				
elix092euen		0.34				
elix091euen		0.302				

Table 24. c@1/Best IR baseline.

Table 26 shows that the majority of questions have been answered by systems in many different languages. For example, 74 questions have been answered in all languages, whereas only 6 questions remained unanswered considering all languages. Notice that 99% of questions have been answered by at least one system in at least one language.

Languages	Questions
0	6
1	20
2	45
3	52
4	55
5	76
6	76
7	96
8	74

Table 25. Number of questions answered by systems in different languages.

5.10 Systems description

Tables 27 and 28 summarise the characteristics of the participant systems. As can be seen, some systems did not analyse the questions at all. Among those that did, the most popular technique was the use of manually created query patterns (e.g. "Where is..." could indicate a location question). Two systems used Boolean retrieval while the rest mainly used Okapi or a VSM-type model.

System name	Question Analyses			Retrieval Model	Linguistic Unit which is indexed		
	No Question Analysis	Manually done Patterns	Other		Words	Lemmas	Stems
SYNA		x		question category		x	
ICIA			MaxEnt question classification, automatic query generation using POS tagging and chunking	Boolean search engine	x	x	
ISIK	x			DFR	x		
NLEL	x			Clustered Keywords Positional Distance model			
UAIC		x			x	x	
MIRA		x		Vector			x
ILES		x				x	
IIIT		x	statistical method	boolean model	x		x
UNED		x	Question classification	Okapi BM25			x
ELIX			Basque lemmatizer	BM25			x
LOGA		x	classification rules applied to question parse	Lucene, sentence segmentation. Also indexes contained answer types of a sentence		x	

Table 26. Methods used by participating systems.

Table 27 shows the type of processing techniques which were used on document fragments returned by the information retrieval components. As would be expected, Named Entity recognition and Numerical Expression recognition were widely used approaches.

Table 28 shows the types of technique used for the answer validation component. Some systems did not have such a component, but for those that did, lexical similarity and syntactic similarity were the most widely used approaches.

Answer Validation								
System name	**No answer validation**	**Machine Learning is used to validate answers**	**Combined classifiers, Minimum Error Rate Training**	**Redundancies in the collection**	**Lexical similarity (term overlapping)**	**Sintactic similarity**	**Semantic similarity**	**Theorem proof or similar**
SYNA				x				
ICIA		x	x		x	x	x	
ISIK	x							
NLEL	x							
UAIC					x	x		
MIRA	x							
ILES				x	x	x		
IIIT					x	x		
UNED								
ELIX	x							
LOGA		x	x	x	x			x

Table 27. Technique used for the Answer Validation component.

5.11 Analysis and discussion about the task

Whereas in previous years, almost all responses were double-blind evaluated to check inter-evaluator agreement, this year it was not possible. A measure of the inter-annotator agreement would have provided us an idea of the complexity and ambiguity of both questions and their supporting passages. Moreover, this was the first year of using the JRC-Acquis collection which claims to be parallel in all languages. The supposed advantage of this was that all systems answer the same questions against the same document collections. Only the language of the questions and documents vary as otherwise the text is supposed to mean exactly the same. However, we found that in fact the texts are not parallel, being many passages left out or translated in a completely different way. The result was that many questions were not supported in all languages and could not therefore be used. This problem resulted in a huge amount of extra work for the organisers. Furthermore, the character of the document collection necessitated changes to the type of the questions. In most cases the questions became more verbose in order to deal with the vagueness and ambiguity of texts.

The idea of introducing new question types Reason, Purpose and Procedure was good in principle, but it did not seem to work as expected. Reason and Purpose questions resulted to be understood as more or less the same and the way in which these reasons and purposes are stated in the documents sometimes is meaningless. A typical type of reason is “to ensure smooth running of the EU” and a

typical purpose is “to implement such and such a law”. With respect to procedures there were also some non informative responses similar to the idea “the procedure to apply the law is to put it into practice”.

Finally, the user model is still unclear, even after checking the kind of questions and answers that were feasible with the current setting: neither lawyers or ordinary people would not ask the kind of questions proposed in the exercise. Once more, the problem is to find the trade-off between research and a user centred development.

494 questions (99%) were answered by at least one system in at least one language; nevertheless the systems that gave more correct answers only answered 288. This shows that the task is feasible and systems still have room to improve and solve it in a short time.
One of the main issues is the retrieval model. Many systems must pay more attention to it since they performed under the baselines based on just IR. From this perspective, paragraph ranking approaches based on n-grams seems promising.

Some systems are able to reduce the number of incorrect answers maintaining a similar level in the number of correct answers, just leaving some questions unanswered. We expect this to be a first step towards the improvement of systems. This ability has been rewarded by the c@1 measure.

Finally, moving to a new domain has raised new questions and challenges for both organizers and participants.

6 InFile@CLEF

The purpose of the INFILE (INformation FILtering Evaluation) track is to evaluate cross-language adaptive filtering systems, i.e. the ability of automated systems to successfully separate relevant and non-relevant documents in an incoming stream of textual information with respect to a given profile, the document and profile being possibly written in different languages.

The INFILE track was first run as a pilot track in CLEF 2008 campaign [Besançon et al 2008].

Information filtering in this track is considered in the framework of competitive intelligence (CI): the evaluation protocol of the campaign has been designed with a particular attention to the context of use of filtering systems by real professional users. Even if the campaign is mainly a technological oriented evaluation process, the protocol and the metrics have been adapted as closely as possible to how a normal user would proceed, including some interaction and adaptation of his system.

The INFILE campaign can mainly be seen as a cross-lingual follow-on to the TREC 2002 Adaptive Filtering task [Robertson and Soboroff 2002] (adaptive filtering track was run from 2000 to 2002), with a particular interest in matching the protocol to the ground truth of CI professionals. In keeping with this objective, we asked CI professionals to create topics according to their experience in the domain.

In addition to the adaptive filtering task already proposed in 2008 [Besançon et al, 2008], we introduced in 2009 the possibility to test batch filtering systems.

For both tasks, the document collection consists of a set of newswire articles provided by the Agence France Presse (AFP) and covering recent years. The topic set is composed of two different kinds of profiles, one concerning general news and events, and a second one on scientific and technological subjects.

The filtering process may be cross lingual: English, French and Arabic are available for the documents and topics, and participants may be evaluated on monolingual runs, bilingual runs, or multilingual runs (with several target languages).

The purpose of the information filtering process is to associate documents in an incoming stream to zero, one or several topics: Filtering systems must provide a Boolean decision for each document with respect to each topic.

For the batch filtering task, participants are provided with the whole document collection and must return the list of relevant documents for each topic (since the filtering process supposes a binary decision for each document, the document list does no need to be ranked).

For the adaptive filtering task, the evaluation is performed using an automatic interactive process, with a simulated user feedback: systems are allowed for each document considered relevant to a topic to ask for a feedback on this decision (i.e. ask if the document was indeed relevant for the topic or not), and can modify their behaviour according to the answer. The feedback is allowed only on kept documents, there is no relevance feedback possible on discarded documents. In order to simulate the limited patience of the user, a limited number of feedbacks is allowed: this number has been fixed in 2009 to 200 feedbacks (it was 50 in 2008; but most of the participants considered this insufficient). The adaptive filtering task uses an interactive client-server protocol, that is described in more details in [Besançon et al. 2008].

6.1 Test collections

6.1.1 The topics

A set of 50 profiles has been prepared, for two different categories: the first group (30 topics) deals with general news and events concerning national and international affairs, sports, politics etc, the second one (20 topics) with scientific and technological subjects.

<top> **<num>**147**</num>** **<title>**Care management of Alzheimer disease**</title>** **<desc>**News in the care management of Alzheimer disease by families, society and politics**</desc>** **<narr>**Relevant documents will highlight different aspects of Alzheimer disease management: - human involvement of carers : families, health workers - financial means: nursing facilities, diverse grants to carers - political decisions leading to guidelines for optimal management of this great public health problem **</narr>** **<keywords>** **<keyword>**Alzheimer disease**</keyword>** **<keyword>**Dementia **</keyword>** **<keyword>**Care management **</keyword>** **<keyword>**Family support **</keyword>** **<keyword>**Public health**</keyword>** **</keywords>** **<sample>**The AAMR/IASSID practice guidelines, developed by an international workgroup, provide guidance for stage–related care management of Alzheimer's disease, and suggestions for the training and education of carers, peers, clinicians and programme staff. The guidelines suggest a three-step intervention activity process, that includes: (1) recognizing changes; (2) conducting...**</sample>** **</top>**	**<top>** **<num>**147**</num>** **<title>**Prise en charge de la maladie d'Alzheimer**</title>** **<desc>**Actualités dans le domaine de la prise en charge de la maladie d'Alzheimer, tant au niveau des familles, de la société qu'au niveau des choix politiques**</desc>** **<narr>**Les documents pertinents présenteront les divers aspects de la prise en charge de la maladie d'Alzheimer : - moyens humains mis en jeu : familles, personnels de santé - moyens financiers : structures d'accueil, aides diverses aux malades et aux aidants - décisions politiques avec établissement de recommandations permettant d'encadrer de façon optimale ce problème majeur de santé publique **</narr>** **<keywords>** **<keyword>**Maladie d'Alzheimer**</keyword>** **<keyword>**Démence **</keyword>** **<keyword>**Prise en charge **</keyword>** **<keyword>**Aide aux familles **</keyword>** **<keyword>**Santé publique **</keyword>** **</keywords>** **<sample>**Un an après l'entrée en vigueur du plan ministériel, un rapport de l'OPEPS rendu public le 12 juillet 2005 dresse un bilan assez sévère de la prise en charge de la maladie d'Alzheimer et des maladies apparentées. Selon l'OPEPS*, la politique de prévention des facteurs de risque est insuffisante, ... **</sample>** **</top>**	**<top>** **<num>**147**</num>** **<title>**بمرض الزهايمر العناية**</title>** **<desc>**المتعلقة بالعناية بمرض الزهايمر، الأحداث مستوى الأسر والمجتمع وأيضا على مستوى على الاختيارات السياسية.**</desc>** **<narr>**التي تتعلق بالعناية بمرض الزهايمر الوثائق البشرية المستخدمة الإمكانات - : مختلف الجوانب من بنيات : المالية الموارد - موضفو الصحة، الأسر، : والمساعدين، الإستقبال، المساعدات المختلفة للمرضى الصادرة من أجل التعليمات : السياسية القرارات - الكبير في الصحة وضع إطار أمثل لهذا المشكل العمومية.**</narr>** **<keywords>** **<keyword>**العمومية الصحة**</keyword>** **<keyword>**الأسر مساعدة**</keyword>** **<keyword>**عناية**</keyword>** **<keyword>**الجنون**</keyword>** **<keyword>**الزهايمر مرض**</keyword>** **</keywords>** **<sample>**... عبر الهاتف كلما اقتضت الوضع دراسة سابقة قد كشفت أن عدد وكانت .الحاجة ذلك الزهايمر سيتضاعف أربع مرات المصابين بمرض الأربعة المقبلة، ويصيب واحداً من أصل خلال العقود الدراسة أن وأكدت.على وجه الأرض شخصاً 85 كل بشكل رئيسي بارتفاع هذه الإحصائية المخيفة مرتبطة مختلف دول العالم، الناجم عن عدد كبار السن في الصحية، وقدرت أنه بحلول العام تحسن الأنظمة 62.8 أعداد أولئك المرضى ستقفز إلى فإن 2050 شخص مليون. بحسب الـCNN.**</sample>** **</top>**

Table 28 An example of topic for the INFILE track, in the three languages

The scientific topics were developed by CI professionals from INIST[19], ARIST Nord Pas de Calais[20], Digiport[21] and OTO Research[22]. The topics were developed in both English and French. The Arabic version has been translated from English and French by native speakers.

Topics are defined with the following elements: a unique identifier, a title (6 words max.), describing the topic in a few words, a description (20 words max.), corresponding to a sentence-long description, a narrative (60 words max.), corresponding to the description of what should be considered a relevant document and possibly what should not, keywords (up to 5) and an example of relevant text (120 words max.), taken from a document that is not in the collection (typically from the web).

Each record of the structure in the different languages corresponds to translations, except for the samples which need to be extracted from real documents. An example of topic in the three languages is presented in Table 10.

6.1.2 The document collection

The INFILE corpus is provided by the Agence France Presse (AFP) for research purpose. We used newswire articles in 3 languages: Arabic, English and French and a 3 years period (2004-2006) which represents a collection of about one and half million newswires for around 10 GB, from which 100,000 documents of each language have been selected to be used for the INFILE filtering test. News articles are encoded in XML format and follow the News Markup Language (NewsML) specifications[23]. All fields are available to the systems and can be used in the filtering process (including keywords, categorization...).

The method used to build the collection of documents with the knowledge of the relevant documents is presented in details in [Besançon et al.,2008]. A summary of this method is given here.

We used a set of 4 search engines (Lucene1, Indri2, Zettair3 and the search engine developed at CEA-LIST) to index the complete collection of 1.4 million documents. Each search engine has been queried using different fields of the topics, which provides us with a pool of runs. We first selected the first 10 retrieved documents of each run, and these documents were assessed manually. We then iterate using a Mixture of Experts model, computing a score for each run according to the current assessment and using this score to weight the choice of the next documents to assess. The final document collection is then built by taking all documents that are relevant to at least one topic (core relevant corpus), all documents that have been assessed and judged not relevant (difficult corpus: documents are not relevant, but share something in common with at least one topic, since they have been retrieved by at least one search engine), and a set of documents taken randomly in the rest of the collection (filler corpus, with documents that have not been retrieved by any search engines for any topic, which should limit the number of relevant documents in the corpus that have not been assessed).

Statistics on the number of assessed documents and relevant documents is presented in Table 29. The repartition of relevant documents across topics is presented in Figure 30.

	eng	fre	ara
number of documents assessed	7312	7886	5124
number of relevant documents	1597	2421	1195
avg number of relevant docs / topic	31,94	48,42	23,9
std deviation on number of relevant docs / topic	28,45	47,82	23,08
[min,max] number of relevant docs / topics	[0,107]	[0,202]	[0,101]

Table 29 Statistics on the number of assessed documents and the number of relevant documents, in each language

[19] The French Institute for Scientific and Technical Information Center, http://international.inist.fr/
[20] Agence Régionale d'Information Stratégique et Technologique, http://www.aristnpdc.org/
[21] See http://www.digiport.org
[22] See http://www.otoresearch.fr/
[23] NewsML is an XML standard designed to provide a media-independent, structural framework for multi-media news. NewsML was developed by the International Press Telecommunications Council. see http://www.newsml.org/

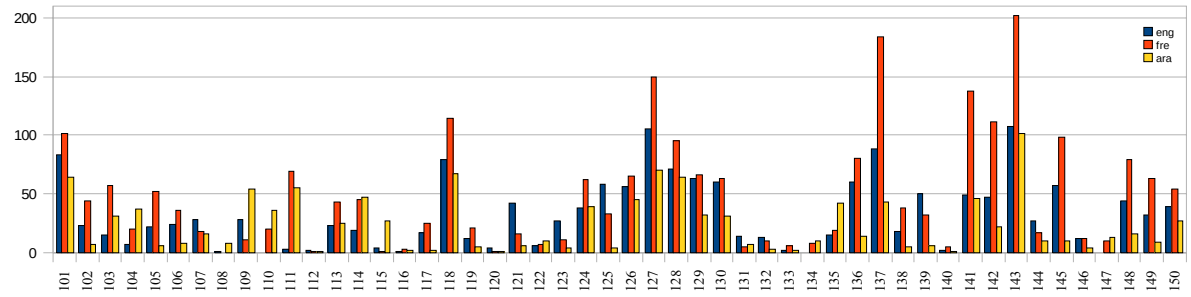


Figure 30 Number of relevant documents for each topic, in each language

6.2 Metrics

The results returned by the participants are binary decisions on the association of a document with a profile. On these data a set of standard evaluation measures is computed:

- Precision
- Recall
- F-Measure
- The linear utility as defined in [Hull and Robertson, 1999, Robertson and Soboroff, 2002]

Additionally, we use the two following measures, introduced last year in INFILE: the first one is an originality measure, defined as a comparative measure corresponding to the number of relevant documents the system uniquely retrieves (among participants). It gives more importance to systems that use innovative and promising technologies that retrieve "difficult" documents.

The second one is an anticipation measure, designed to give more interest to systems that can find the first document in a given profile. This measure is motivated in CI by the interest of being at the cutting edge of a domain, and not missing the first information to be reactive. It is measured by the inverse rank of the first relevant document detected (in the list of the documents), averaged on all profiles. The measure is similar to the mean reciprocal rank (MRR) used for instance in Question Answering Evaluation [Voorhees, 1999], but is not computed on the ranked list of retrieved documents but on the chronological list of the relevant documents.

6.3 Results

On the 9 participants registered for both tasks, 5 submitted results : 3 participants submitted results for the batch filtering task (a total of 9 runs), 2 for the interactive filtering task (3 runs). Participants were different for the two tasks. Table 31 presents the participant list.

team name	institute	country
IMAG	Institut Informatique et Mathématiques Appliquées de Grenoble	France
SINAI	University of Jaen	Spain
UAIC	Universitatea Alexandru Ioan Cuza of IASI	Romania
HossurTech	société CADEGE	France
UOWD	University of Wollongong (Comp.Sci & Engineering)	Dubai

Table 31 Participant list

6 runs out of 9 are monolingual English for the batch filtering task, 3 are multilingual from English to French/English. For the interactive task, one run is monolingual English, one is monolingual French, and one is bilingual French to English. Table 32 summarizes the total number of runs for each language pair. No participant submitted runs with Arabic as source or target language.

nb runs			
	english	French	Arabic
English	10	3	0
French	1	1	0
Arabic	0	0	0

Table 32 #runs per source and target languages

Evaluation scores for the batch filtering task are presented in Table 33, gathered by the target language (multilingual runs appears in several groups, in order to present the individual scores on each target language). Best result is obtained on monolingual English, but for the only participant that tried multilingual runs, the results obtained for the different target languages (English and French) are comparable.

monolingual english

team	run	num_rel	num_rel_ret	precision	recall	F-score	Utility	anticipation
IMAG	IMAG_1	1597	413	0,26	0,30	0,21	0,21	0,43
UAIC	uaic_4	1597	1267	0,09	0,66	0,13	0,05	0,73
UAIC	uaic_1	1597	1331	0,06	0,69	0,09	0,03	0,75
UAIC	uaic_2	1597	1331	0,06	0,69	0,09	0,03	0,75
UAIC	uaic_3	1597	1507	0,06	0,82	0,09	0,03	0,86
IMAG	IMAG_2	1597	109	0,13	0,09	0,07	0,16	0,22
IMAG	IMAG_3	1597	66	0,16	0,06	0,07	0,22	0,14
SINAI	topics_1	1597	940	0,02	0,50	0,04	0,00	0,57
SINAI	googlenews_2	1597	196	0,01	0,08	0,01	0,13	0,10

crosslingual english # french

team	run	num_rel	num_rel_ret	precision	recall	F-score	Utility	anticipation
UAIC	uaic_4	2421	1120	0,09	0,44	0,12	0,05	0,58
UAIC	uaic_3	2421	1905	0,06	0,75	0,10	0,03	0,83
UAIC	uaic_2	2421	1614	0,06	0,67	0,09	0,02	0,76

multilingual english # english/french

team	run	num_rel	num_rel_ret	precision	recall	F-score	Utility	anticipation
UAIC	uaic_4	4018	2387	0,07	0,56	0,11	0,02	0,72
UAIC	uaic_3	4018	3412	0,05	0,81	0,08	0,02	0,85
UAIC	uaic_2	4018	2945	0,05	0,70	0,07	0,02	0,80

Table 33 Scores for batch filtering runs, sorted by F-score

Scores for the adaptive filtering task are presented in Table 34. The scores are worse than the ones obtained on the batch filtering task, but the language pairs and the participants are not the same. We also note than both batch and adaptive results for the INFILE 2009 campaign are worse than the results obtained for the adaptive task in the INFILE 2008 edition.

monolingual english

team	run	num_rel	num_rel_ret	precision	recall	F-score	Utility	anticipation
UOWD	base	1597	20	0,00	0,01	0,01	0,03	0,05

monolingual french

team	run	num_rel	num_rel_ret	precision	recall	F-score	Utility	anticipation
HossurTech	hossur-tech-004	2421	790	0,05	0,31	0,06	0,05	0,53

crosslingual french # english

team	run	num_rel	num_rel_ret	precision	recall	F-score	Utility	anticipation
HossurTech	hossur-tech-001	1597	819	0,10	0,45	0,10	0,07	0,59

Table 34 Scores for adaptive filtering runs

Results for originality measure are presented in Table 35. The upper part of the table present originality scores for every run that has the same target language (i.e. the number of relevant documents that this particular run uniquely retrieves). Since this global comparison may not be fair for participants who submitted several runs, which are presumably variants of the same technique and will share most of the relevant retrieved documents, we present in the lower part of the table the originality scores using only one run for each participant (we chose the run with the best recall score). We see here that participant with lower F-scores can have a better originality score. However, due to the small number of participants, the relevance of the originality score is arguable in this context, since it seems to be strongly linked to the difference of the recall score.

originality on all runs

target lang=eng			target lang=fre		
team	**run**	**originality**	**team**	**run**	**originality**
UAIC	uaic_3	39	HossurTech	hossur-tech-004	177
HossurTech	hossur-tech-001	18	UAIC	uaic_3	82
SINAI	googlenews_2	15	UAIC	uaic_2	0
SINAI	topics_1	9	UAIC	uaic_4	0
UAIC	uaic_4	4			
IMAG	IMAG_1	1			
UAIC	uaic_1	0			
IMAG	IMAG_3	0			
UOWD	base	0			
UAIC	uaic_2	0			
IMAG	IMAG_2	0			

originality on best run

target lang=eng			target lang=fre		
team	**run**	**originality**	**team**	**run**	**originality**
UAIC	uaic_3	267	UAIC	hossur-tech-004	1292
HossurTech	hossur-tech-001	20	HossurTech	uaic_3	177
SINAI	topics_1	9			
IMAG	IMAG_1	4			
UOWD	base	0			

Table 35 Originality scores

6.4 Approaches and discussion

The INFILE campaign has been organized for the second time this year in CLEF, to evaluate adaptive filtering systems in a cross-language environment. The document and topic collection were the same as the 2008 edition of the INFILE@CLEF track. Two tasks have been proposed: a batch filtering task and an adaptive filtering task. The adaptive task used an original setup to simulate the incoming of newswires documents, and the interaction of a user through a simulated feedback. We had this year more participants than last year and more results to analyze. However, the innovative cross lingual aspect of the task has still not really been explored, since most runs were monolingual English and no participant used the Arabic topics or documents. The lack of participation for the adaptive task is also disappointing since it does not provide enough data to compare batch techniques to adaptive techniques and does not allow to conclude on the interest of the use of the used feedback on the documents.

7 LogCLEF

Log is a concept commonly used in computer science; in fact, log data are collected by an operating system to make a permanent record of events during the usage of the operating system itself. This is done to better support its operations, and in particular its recovery procedures. Due to the experience

gained in the management of operating systems and many application systems that manage permanent data, log procedures are commonly put in place to collect and store data on the usage of the application system by its users. Initially, these data were mainly used to manage recovery procedures of the application system, but over time it became apparent that they could also be used to study the usage of the application by its users, and to better adapt the system to the objectives the users were expecting to reach [Agosti, 2008].

Log data can be collected during the use of a search engine to monitor its functioning and usage by final and specialized users, this means that log data constitute a relevant aspect in the evaluation process of the quality of a search engine and the quality of a multilingual search services. There are a number of issues regarding log analysis which were raised in recent workshops about query log analysis [TrebleCLEF Workshop 2009; UIIR SIGIR Workshop 2009] among the others: the lack of recent and long-term period log data, the repeatability of experiments for comparing different evaluations, the production of logs for research with anonymized data.

LogCLEF [Mandl et al. 2009] is a new track at CLEF which wants to stimulate research on user behavior in multilingual environments and develop standard evaluation collections which support long-term research. LogCLEF started in 2009[24] as an evaluation initiative for the analysis of queries and other logged activities as expression of user behavior. The main goal was the analysis and classification of queries in order to understand search behavior in multilingual contexts and ultimately to improve search systems. Two tasks were defined in 2009:

1. Log Analysis and Geographic Query Identification (LAGI) and
2. Log Analysis for Digital Societies (LADS).

LAGI required the identification of geographical queries within logs from the Tumba! Search engine and The European Library multilingual information system. LADS intended to analyze the user behavior in the multilingual information system of The European Library (TEL). The distribution and usage of the TEL search logs are regulated by the CLEF end-user agreement which has to be signed by all the institutions who want to have access to the data. About 20 institutions from all over the world registered to the LogCLEF track to get the data for the analysis, but only 4 groups submitted results for the LADS task.

7.1 Goals

Participants of the LADS task were required to run their log analyses according to some suggested potential targets: query reformulation, multilingual search behaviour, and community identification. This task was open to diverse approaches, in particular data mining techniques in order to extract knowledge from the data and find interesting user patterns.

Suggested analyses of the log data were:

1. user session reconstruction, where this step needs to be considered as a prerequisite to the following ones;
2. user interaction with the portal at query time; e.g. how users interact with the search interface, what kind of search they perform (simple or advanced), and how many users feel satisfied/unsatisfied with the first search and how many of them reformulate queries, browse results, leave the portal to follow the search in a national library;
3. multilinguality and query reformulation; e.g. what are the collections that are selected the most by users, how the language (country/portal interface) of the user is correlated to the collections selected during the search, how the users reformulate the query in one language or in a different language;

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

4. user context and user profile; e.g. how the study of the actions in the log can identify user profiles, how the implicit feedback information recorded in the logs can be exploited to create the context in which the user operates and how this context evolves.

Participants were required to:

- apply some algorithm to process the complete logs;
- analyze if the resources created based on the logs (e.g. annotations of a small subsets) need to be made publicly available;
- find out interesting issues about the user behavior as exhibited in the logs;
- submit results in a structured file.

7.2 Data and Participants

The data used for the LADS task are the action logs of The European Library portal which cover the period from 1st January 2007 until 30th June 2008. The data were released as a table of a relational database, where each table record represents a user action. The most significant columns of the table are:

- A numeric id, for identifying registered users or "guest" otherwise;
- Anonymous user's IP address;
- An automatically generated alphanumeric, identifying sequential actions of the same user (sessions) ;
- Query contents;
- Name of the action that a user performed;
- The corresponding collection's alphanumeric id;
- Date and time of the action's occurrence.

A total of 4 groups submitted results for the LADS task:

- University of Sunderland, England;
- CELI – Language and Information Technology, Italy;
- University of Hildesheim, Germany;
- Trinity College Dublin – Dublin City University, Ireland.

7.3 Results

The 4 groups which participated in the CLEF 2009 Workshop submitted very diversified results which covered:

- Identification of list of pairs of queries in two languages combined with session information;
- Correlation between language of the interface and language of the query;
- Activities at query time to study different user backgrounds.

7.3.1 University of Sunderland, England

The approach followed by this participant [Oakes and Xu 2009] was to design a search engine which retrieves documents based on previous users' choices. In particular, the log-based search engine indexes not the content of the document but the query itself, as found in the search logs, in order to suggest new queries to the user. The new queries are matched against the old query terms in the indexes, and documents are ranked by the degree of match between their index terms and the new query.

They also studied the language of the queries using an automatic language identification approach based on trigrams which produce a probability for a query of belonging to a language. The purpose of

this study was to use the search logs annotated with the most likely language of the query to find whether users tended to stick with one language throughout a search session, or whether they tended to change languages in mid-session as part of the query reformulation process. The result of the analysis was a table of occurrences which summarizes the probability of switching from a language to another one within the same session, or the probability of starting a new session.

7.3.2 CELI – Language and Information Technology, Italy

The experiments carried out by [Bosca and Dini 2009b] focused on the possibility to exploit The European Library logs as a source for inferring new translations and thus enriching already existing translation resources for dictionary based cross language access to digital libraries. The methodology is based on the assumption that when users are aware of consulting a multilingual digital collection, they are likely to repeat the same query several times, in several languages.
The participants used the CACAO system, which is part of the CACAO[25] project, in order to obtain a set of translation candidates for each query in the log, starting from the language of the query and searching into one of the supported languages: English, French, German, Polish, Hungarian and Italian. By adopting the proposed algorithm, it is possible to discover translationally equivalent queries in logs produced by monitoring user queries.

7.3.3 University of Hildesheim

[Lamm et al 2009] propose indicators to suggest whether a search session was successful or not. For example, actions which indicate that the user came across an interesting document; one possible indicator is when the user chose the Available at Library link to view the record in a particular national library interface. There are three level defined for each search session: success, when the user performs an action which indicates clearly that an interesting record was found (like available_at, see_online, option_print and so on); failure, if none of the actions of the success was performed; strong failure then it is a failure and the user do not perform any view_full action. They used the Flare Data Visualization tool[26] to analyse the sequence of actions of a search session to enable a more qualitative human assessment.

During the qualitative analysis of the user path information, they observed some differences between users from different countries; there seem to exist two prevailing search patterns: British, Dutch, Italian, Polish and Spanish users seems to examine more documents after the first query; German, French and US users seems to rephrase their queries more often.

7.3.4 Trinity College Dublin – Dublin City University

[Ghorab et al. 2009] performed a series of deep log analysis in order to investigate the following hypothesis: users from different linguistic or cultural backgrounds behave differently in search; there are patterns in user actions which could be useful for stereotypical grouping of users; user queries reflect the mental model or prior knowledge of a user about a search system.

They started the analysis with the computation of general statistics about search sessions and query reformulation; the latter showed that the majority of users have little knowledge of the search system, as they include stop-words and even change them. On the other hand, a small group of users used advanced query operators such as wildcards in their queries, which may correspond to experienced users.

The frequency distribution of the six main actions across each of the five interface languages was studied. It was found that for some languages the ratio between the number of simple search and advanced search actions was lower than the average, and the cause for this may be that a greater number of queries submitted under some language were not satisfied through simple search, and users had to reformulate their queries through advanced search. This finding may support the hypothesis that users from different linguistic or cultural backgrounds behave differently in search.

[25] http://www.cacaoproject.eu/
[26] http://flare.prefuse.org/

In advanced search, the percentages of queries made up of three or more terms surpass those of simple search. This may suggest that users are encouraged to enter more search terms by the availability of multiple input fields.

At the end of the analysis they suggested a number of improvements for the TEL interface such as: integrating a query adaptation process into TEL; offering focused online help if a user spends an uncharacteristically long time between some actions; highlighting elements in the TEL GUI as a default action or a typical next action; identifying the type of user for the sake of search personalisation.

7.4 Future of LogCLEF

Studies on log files are essential for personalization purposes, since they implicitly capture user intentions and preferences in a particular instant of time. There is an emerging research activity about log analysis which tackles cross-lingual issues: extending the notion of query suggestion to cross-lingual query suggestion studying search query logs; leveraging click-through data to extract query translation pairs. LogCLEF has provided an evaluation resource with log files of user activities in multilingual search environments.

At the end of the session at CLEF 2009, a breakout session was organized to discuss how to continue LogCLEF and promote research on log analysis. There were about 20 people and the discussion was very positive and fruitful. The group of the University of Amsterdam [Hofmann et al. 2009] presented a sort of position paper to suggest possible scenarios for the future of LogCLEF, in particular the need of a task that goes beyond basic statistical patterns, for example the enrichment of queries with a broader range of semantic information seems more suitable.

All the participants agreed on the fact that access to the content, even partial, of the catalogue record selected by the user is mandatory to perform a deeper log analysis and to design advanced functionalities which support the users during the search session.

There is indeed a clear intention to go on with this initiative and grow this community on log analysis, and there is an increasing number of institutions and research groups around the world which consider the availability of TEL log data as a very important step in sharing resources for research in this area and filling an important gap which has been the lack of recent and long-term data for verifiability and repeatability of experiments.

For this reason, the organizers of LogCLEF 2009 have proposed another edition of the effort in 2010 as a Lab at CLEF 2010 and they will ask for the usage of both the search logs and HTTP logs.

8 Grid@CLEF

The Grid@CLEF[27] track [Ferro and Harman 2009] is envisaged as a long term activity which aims at running a series of systematic experiments in order to improve the comprehension of MLIA systems and gain an exhaustive picture of their behaviour with respect to languages.

Grid@CLEF 2009 has been a pilot track that has made the first steps in this direction by giving the participants the possibility of gaining experienced with a new way of carrying out the type of experimentation that is needed in Grid@CLEF in order to test all the different combinations of IR components and languages. Grid@CLEF 2009 has provided us with an opportunity to begin to set up a suitable framework in order to carry out a first set of experiments which allows us to acquire an initial set of measurements and to start to explore the interaction among IR components and languages. This initial knowledge will allow us to tune the overall protocol and framework, to understand what directions are more promising, and to scale the experiments up to a finer-grain comprehension of the behaviour of IR components across languages.

Individual researchers or small groups do not usually have the possibility of running large-scale and systematic experiments over a large set of experimental collections and resources. Figure 15 and Figure 16 depict the performances, e.g. mean average precision, of the composition of different IR

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

components across a set of languages as a kind of surface area which we intend to explore with our experiment. The average CLEF participants, shown in Figure 15, may only be able to sample a few points on this surface since, for example, they usually test just a few variations of their own or customary IR model with a stemmer for two or three languages. Instead, the expert CLEF participant, represented in Figure 16, may have the expertise and competence to test all the possible variations of a given component across a set of languages, as [Savoy 2002] does for stemmers, thus investigating a good slice of the surface area.

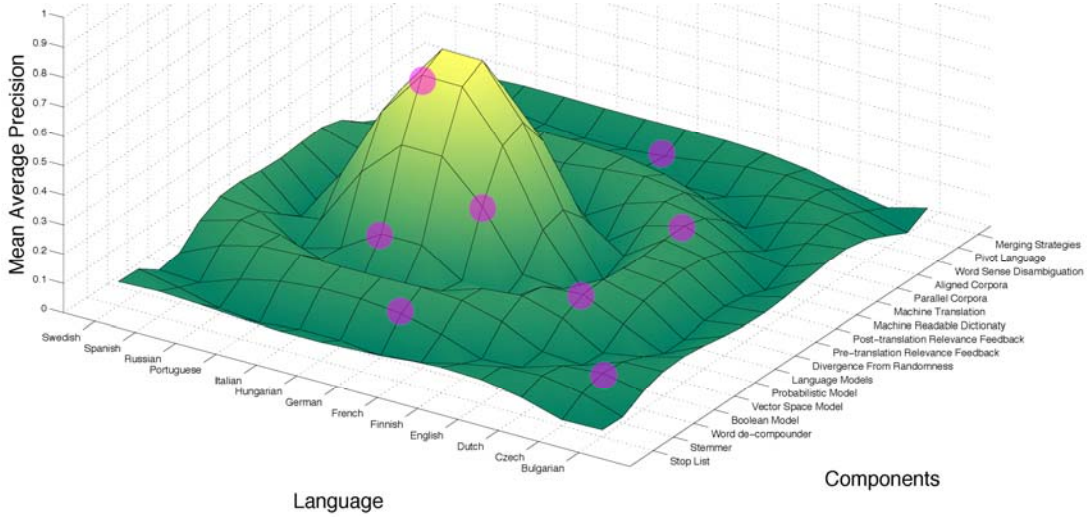


Figure 15: Average CLEF Participant.

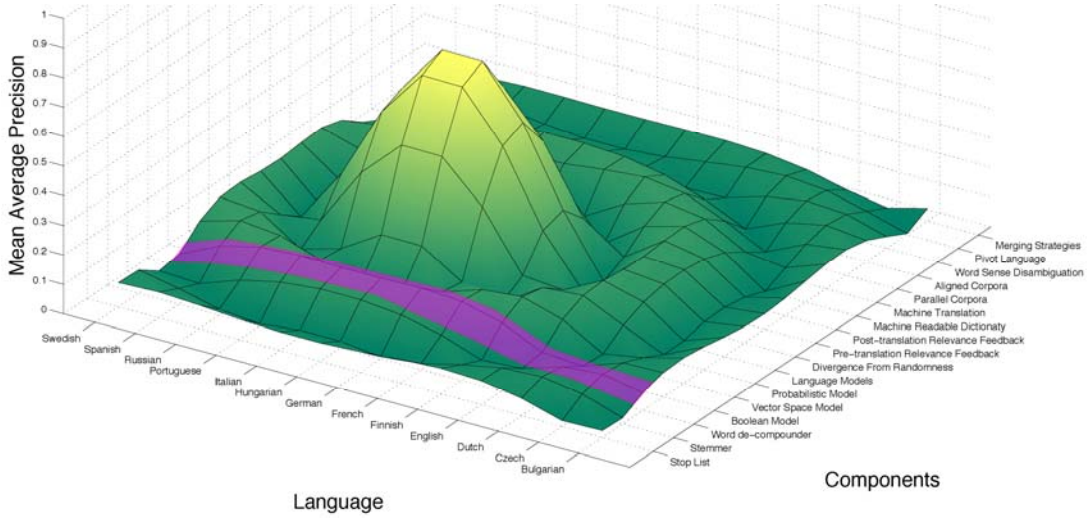


Figure 16: Expert CLEF Participant.

However, even though each of these cases produces valuable research results and contributes to the advancement of the discipline, they are both still far removed from a clear and complete comprehension of the features and properties of the surface. A far deeper sampling would be needed for this, as shown in Figure 17: in this sense, Grid@CLEF will create a fine-grained grid of points over this surface and, hence, the name of the track comes.

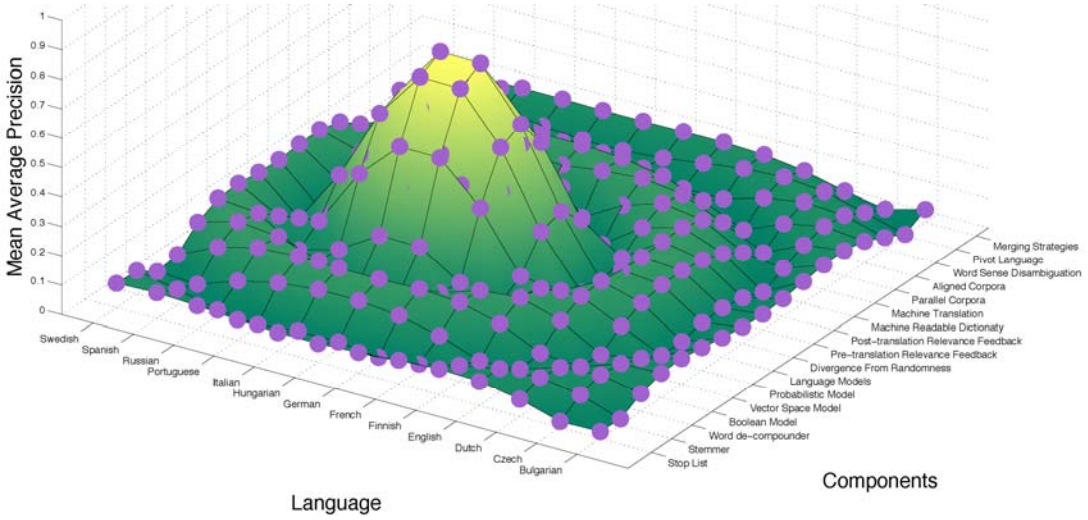


Figure 17: Systematic investigation of the interaction among components and languages.

It is our hypothesis that a series of systematic experiments can re-use and exploit the valuable resources and experimental collections made available by CLEF in order to gain more insights about the effectiveness of, for example, the various weighting schemes and retrieval techniques with respect to the languages. In order to do this, we must deal with the interaction of three main entities:

- **Component**: in charge of carrying out one of the steps of the IR process;
- **Language**: will affect the performance and behaviour of the different components of an Information Retrieval System (IRS) depending on its specific features, e.g. alphabet, morphology, syntax, and so on.
- **Task**: will impact on the performances of IRS components according to its distinctive characteristics;

We assume that the contributions of these three main entities to retrieval performance tend to overlap; nevertheless, at present, we do not have enough knowledge about this process to say whether, how, and to what extent these entities interact and/or overlap – and how their contributions can be combined, e.g. in a linear fashion or according to some more complex relation.

The above issue is in direct relationship with another long-standing problem in the IR experimentation: the impossibility of testing a single component independently of a complete IRS. [Robertson 1981, p. 12] points out that "if we want to decide between alternative indexing strategies for example, we must use these strategies as part of a complete information retrieval system, and examine its overall performance (with each of the alternatives) directly". This means that we have to proceed by changing only one component at time and keeping all the others fixed, in order to identify the impact of that component on retrieval effectiveness; this also calls for the identification of suitable baselines with respect to which comparisons can be made.

8.1 The CIRCO Framework.

In order to run these grid experiments, we need to set up a framework in which participants can exchange the intermediate output of the components of their systems and create a run by using the output of the components of other participants.

For example, if the expertise of participant A is in building stemmers and decompounders while participant B's expertise is in developing probabilistic IR models, we would like to make it possible for participant A to apply his stemmer to a document collection, pass the output to participant B, who tests his probabilistic IR model, thus obtaining a final run which represents the test of participant A' stemmer + participant B probabilistic IR model.

To this end, the objective of the *Coordinated Information Retrieval Components Orchestration* (CIRCO) framework [Ferro 2009] is to allow for a *distributed*, *loosely–coupled*, and *asynchronous* experimental evaluation of information retrieval systems where:

- *distributed* highlights that different stakeholders can take part to the experimentation each one providing one or more components of the whole IR system to be evaluated;
- *loosely-coupled* points out that minimal integration among the different components is required to carry out the experimentation;
- *asynchronous* underlines that no synchronization among the different components is required to carry out the experimentation.

The CIRCO framework allows different research groups and industrial parties, each one with their own areas of expertise, to take part in the creation of collaborative experiments. This is a radical departure from today's IR evaluation practice where each stakeholder has to develop (or integrate components to build) an entire IR system to be able to run a single experiment.

The base idea – and assumption – behind CIRCO to streamline the architecture of an IR system and represent it as a pipeline of components chained together. The processing proceeds by passing the results of the computations of a component as input to the next component in the pipeline without branches, i.e. no alternative paths are allowed in the chain.

To get an intuitive idea of the overall approach adopted in CIRCO, consider the example pipeline shown in Figure 18.

The example IR system is constituted by the following components:

- *tokenizer*: breaks the input documents into a sequence of tokens;
- *stop word remover*: removes stop words from the sequence of tokens;
- *stemmer*: stems the tokens;
- *indexer*: weights the tokens and stores them and the related information in an index.

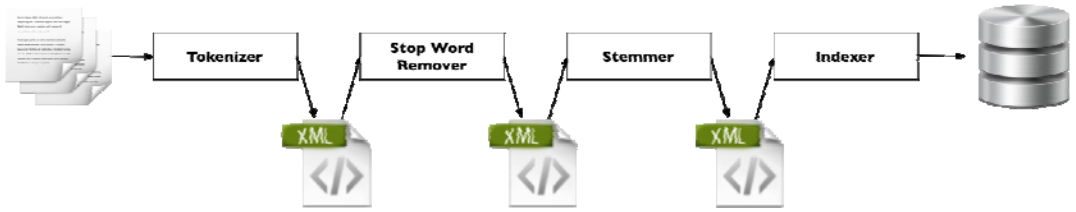


Figure 18: Example of CIRCO pipeline for an IR system.

Instead of directly feeding the next component as usually happens in an IR system, CIRCO operates by requiring each component to input and output from/to eXtensible Markup Language (XML) [W3C 2008] files in a well-defined format, as shown in Figure 18.

These XML files can then be exchanged among the different stakeholders that are involved in the evaluation. In this way, we can meet the requirements stated above by allowing for an experimentation that is:

- *distributed* since different stakeholders can take part in the same experiment, each one providing his own component(s);
- *loosely-coupled* since the different components do not need to be integrated into a whole and running IR system but only need to communicate by means of a well-defined XML format;
- *asynchronous* since the different components do not need to operate all at the same time or immediately after the previous one but can exchange and process the XML files at different rates.

In order to allow this way of conducting experiments, the CIRCO framework consists of:

- CIRCO Schema: an XML Schema [W3C 2004a,b] model which precisely defines the format of the XML files exchanged among stakeholders' components;

- CIRCO Java[28]: an implementation of CIRCO based on the Java programming language to facilitate its adoption and portability.

The choice of using an XML-based exchange format is due to the fact that the main other possibility, i.e. to develop a common *Application Program Interface* (API) IR systems have to comply with, presents some issues:

- the experimentation would not be *loosely-coupled*, since all the IR systems would have to be coded with respect to the same API;
- much more complicated solutions would be required for allowing the *distributed* and *asynchronous* running of the experiments, since you would need some kind of middleware for process orchestration and message delivery;
- multiple versions of the API in different languages should be provided to take into account the different technologies used to develop IR system;
- the integration with legacy code could be problematic and require a lot of effort;
- overall, stakeholders would be distracted from their main objective, which is running an experiment and evaluating a system.

8.2 Track Setup

The Grid@CLEF tracks offers a traditional ad-hoc task which makes use of experimental collections developed according to the Cranfield paradigm [Cleverdon 1997]. This first year task focuses on monolingual retrieval, i.e. querying topics against documents in the same language of the topics, in five European languages: Dutch, English, French, German, and Italian.

The selected languages allow participants to test both romance and Germanic languages, as well as languages with word compounding issues. These languages have been extensively studied in the MultiLingual Information Access (MLIA) field and, therefore, it will be possible to compare and assess the outcomes of the first year experiments with respect to the existing literature.

This first year track has a twofold goal:

1. to prepare participants' systems to work according to CIRCO framework;
2. to conduct as many experiments as possible, i.e. to put as many dots as possible on the grid.

We used the DIRECT system [Agosti and Ferro 2009; Dussin and Ferro 2009; Ferro 2008] to manage the different aspects of the track.

8.2.1 Test Collections

Grid@CLEF 2009 used the test collection originally developed for the CLEF 2001 and 2002 campaigns [Braschler 2002, 2003].

Table 8 reports the document collections that have been used for each of the languages offered for the track.

[28] The documentation is available at the following address:
`http://ims.dei.unipd.it/software/circo/apidoc/`.
The source code and the binary code are available at the following address:
`http://ims.dei.unipd.it/software/circo/jar/`.

Language	Collection	Documents	Size (approx.)
Dutch	NRC Handelsblad 1994/95	84,121	291 Mbyte
	Algemeen Dagblad 1994/95	106,484	235 Mbyte
		190,605	**526 Mbyte**
English	Los Angeles Times 1994	**113,005**	**420 Mbyte**
French	Le Monde 1994	44,013	154 Mbyte
	French SDA 1994	43,178	82 Mbyte
		87,191	**236 Mbyte**
German	Frankfurter Rundschau 1994	139,715	319 Mbyte
	Der Spiegel 1994/95	13,979	61 Mbyte
	German SDA 1994	71,677	140 Mbyte
		225,371	**520 Mbyte**
Italian	La Stampa 1994	58,051	189 Mbyte
	Italian SDA 1994	50,527	81 Mbyte
		108,578	**270 Mbyte**

Table 36: Grid@CLEF 2009 document collections.

In Grid@CLEF 2009, we used 84 out of 100 topics in the set `10.2452/41-AH-10.2452/140-AH` originally developed for CLEF 2001 and 2002 since they have relevant documents in all the used collections.

The same relevance assessment developed for CLEF 2001 and 2002 have been used.

8.2.2 Result Calculation

Evaluation campaigns such as TREC and CLEF are based on the belief that the effectiveness of IRSs can be objectively evaluated by an analysis of a representative set of sample search results. For this, effectiveness measures are calculated based on the results submitted by the participants and the relevance assessments. Popular measures usually adopted for exercises of this type are Recall and Precision. Details on how they are calculated for CLEF are given in [Braschler and Peters 2004]. We used `trec_eval`[29] 8.0 to compute the performance measures.

The individual results for all official Grid@CLEF experiments in CLEF 2009 are given in the Appendices of the CLEF 2009 Working Notes [Di Nunzio and Ferro 2009c]. You can also access them online at:

- Monolingual English:
  `http://direct.dei.unipd.it/DOIResolver.do?type=task&id=GRIDCLEF-MONO-EN-CLEF2009`
- Monolingual French:
  `http://direct.dei.unipd.it/DOIResolver.do?type=task&id=GRIDCLEF-MONO-FR-CLEF2009`
- Monolingual German:
  `http://direct.dei.unipd.it/DOIResolver.do?type=task&id=GRIDCLEF-MONO-DE-CLEF2009`

[29] See `http://trec.nist.gov/trec_eval`

8.3 Track Outcomes

8.3.1 Participants and Experiments

A total of 2 groups from 2 different countries submitted official results for one or more of the Grid@CLEF 2009 tasks. Participants were required to submit at least one title+description ("TD") run per task in order to increase comparability between experiments: all the 18 submitted runs used this combination of topic fields.

The participation in this first year was especially challenging because of the need of modifying existing systems to implement the CIRCO framework. Moreover, it has been challenging also from the computational point of view since, for each component in a IR pipeline, CIRCO could produce XML files that are 50-60 times the size of the original collection; this greatly increased the indexing time and the time needed to submit runs and deliver the corresponding XML files.

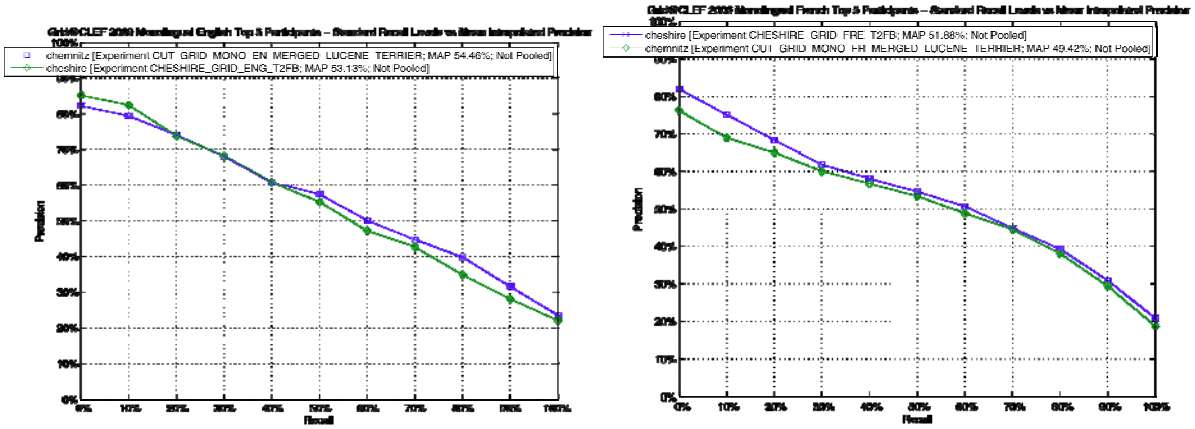
8.3.2 Results

Table 9 shows the top runs 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.

Figure 21 compares the performances of the top participants of the Grid@CLEF monolingual tasks.

Track	Rank	Participant	Experiment DOI	MAP
English	1st	chemnitz	10.2415/GRIDCLEF-MONO-EN-CLEF2009.CHEMNITZ.CUT.GRID.MONO.EN.MERGED.LUCENE.TERRIER	54.45%
	2nd	chesire	10.2415/GRIDCLEF-MONO-EN-CLEF2009.CHESHIRE.CHESHIRE.GRID.ENG.T2FB	53.13%
	Difference			2.48%
French	1st	chesire	10.2415/GRIDCLEF-MONO-FR-CLEF2009.CHESHIRE.CHESHIRE.GRID.FRE.T2FB	51.88%
	2nd	chemnitz	10.2415/GRIDCLEF-MONO-FR-CLEF2009.CHEMNITZ.CUT.GRID.MONO.FR.MERGED.LUCENE.TERRIER	49.42%
	Difference			4.97%
German	1st	chemnitz	10.2415/GRIDCLEF-MONO-DE-CLEF2009.CHEMNITZ.CUT.GRID.MONO.DE.MERGED.LUCENE.TERRIER	48.64%
	2nd	chesire	10.2415/GRIDCLEF-MONO-DE-CLEF2009.CHESHIRE.CHESHIRE.GRID.GER.T2FB	40.02%
	Difference			21.53%

Table 37: Best entries for the monolingual Grid@CLEF tasks.



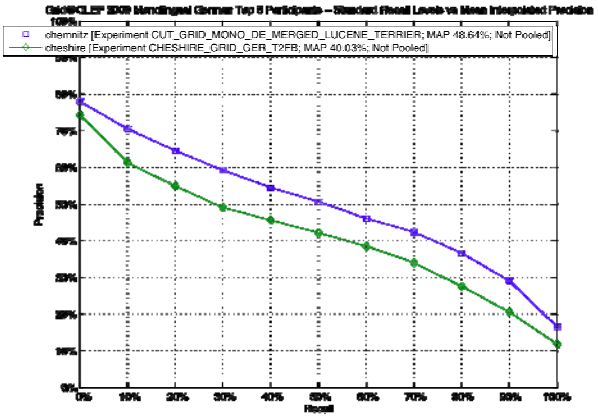


Figure 19: Comparison of the performances of the top participants in the monolingual Grid@CLEF tasks.

8.3.3 Approaches and Discussion

Chemnitz [Eibl and Kürsten 2009] approached the participation in Grid@CLEF into the wider context of the creation of an archive of audiovisual media which can be jointly used by German TV stations, stores both raw material as well as produced and broadcasted material and needs to be described as comprehensively as possible in order to be easily searchable. In this context, they have developed the Xtrieval system, which aims to be flexible and easily configurable in order to be adjusted to different corpora, multimedia search tasks, and annotation kinds. Chemnitz tested both the vector space model, as implemented by Lucene[30] and BM25, as implemented by Terrier[31], in combination with Snowball and Savoy's stemmers. They found out that the impact of retrieval techniques are highly depending on the corpus and quite unpredictable and that, even if over they years they have learned how to guess reasonable configurations for their system in order to get good results, there is still the need of "strong rules which let us predict the retrieval quality . . . [and] enable us to automatically configure a retrieval engine in accordance to the corpus". This was for them motivation to participate in Grid@CLEF 2009, which represented a first attempt that will allow them to go also in this direction.

Cheshire [Larson 2009b] participated in Grid@CLEF with their Cheshire II[32] system based on logistic regression and their interest was in understanding what happens when you try to separate the processing elements of IR systems and look at their intermediate output, taking this as an opportunity to re-analyse and improve their system, and, possibly, finding a way to incorporate into Cheshire II components of other IR systems for subtasks in which they currently cannot do or cannot do effectively, such as decompounding German words. They also found that "the same algorithms and processing systems can have radically different performance on different collections and query sets". Finally, the participation in Grid@CLEF actually allowed Cheshire to improve their system and to point out some suggestions for the next Grid@CLEF, concerning the support for the creation of multiple indexes according to the structure of a document and specific indexing tasks related to the geographic information retrieval, such as geographic names extraction and geo-referencing

9 CLEF 2009 in the TrebleCLEF Context

The previous sections discussed and analysed in detail the achievements and the results of the tracks at CLEF 2009 and presented some future directions for these tracks, as gained from the experience with the campaigns conducted within the confines of the TrebleCLEF coordination action. Some of the tracks in CLEF 2009 have used input from various TrebleCLEF activities conducted in 2008. Deliverable 2.3.1 outlines some of the plans we had for incorporating results from a series of

[30] See `http://lucene.apache.org/`
[31] See `http://ir.dcs.gla.ac.uk/terrier/`
[32] See `http://cheshire.berkeley.edu/`

workshops (Workshop on Novel Methodologies for Evaluation in IR, System Developer's Workshop, User Communities Workshop) into track design for 2009.

Notably, participants at the System Developer's workshop had asked for a more component-oriented approach to CLEF experiments, which was addressed by the Grid@CLEF track. As outlined in section 8, this track successfully implemented a pipeline-oriented framework, where individual components in the document indexing stage can be studied. This work, when more widely adopted, has strong potential to complement the work conducted in the context of WP3 of TrebleCLEF, where best practice recommendations for MLIA system developer's were derived from an analysis of the experiment descriptions by CLEF participants.

The output of all previous CLEF campaigns (2000-2008) was one of the principal inputs for the work on the best practice guidelines on system and user oriented multilingual information access (deliverable 3.3) The implementation of a fully operational multilingual or cross-language information retrieval system remains complex, and there are few "packaged" MLIA/CLIR offerings in the marketplace. For a practitioner wanting to enter the MLIA/CLIR field, the volume of academic output of CLEF can be daunting, with the relatively unstructured nature of the experiment descriptions (participants are essentially free in how they describe their experiments) adding to the difficulty in quickly locating relevant, generalisable approaches. The best practice recommendations contained in deliverable 3.3 are thus the result of an attempt to digest the CLEF academic papers, to unify the conclusions contained in them, and to present the outcome in a way which is more directly suitable for an implementation-oriented reader.

In order to support the analysis ("digestion") of the corpus of CLEF academic output, the experiment descriptions were indexed, lists with the technical terminology contained in them were compiled. These lists have shown that there are indeed a number of approaches and building blocks that are used and re-used with great frequency. This observation supports the earlier analysis contained in [Braschler & Peters, 2004], which has also shown that CLEF participants eagerly adopt well-performing approaches and innovate on them for subsequent campaigns. It also underscores the importance of continuity of tracks in CLEF, in that a track should run at least for two, and probably better for three years to allow for this kind of "ripening" of approaches (this continuity also usually ensures that the test collections are built up to a suitable size). For the recommendations that have emerged from this analysis, we refer the reader to deliverable 3.3.

It is important to consider that when designing the experiments that were ultimately studied for this analysis, the individual participants made a number of design assumptions which are often not explicitly stated in the experiment descriptions. Thus, while it has proven to be possible to find patterns in these descriptions that point to approaches that generalize well across different systems and tracks, there is no guarantee that all combinations of important components were tested by the participants at some point during the campaigns. Grid@CLEF thus provides the kind of exhaustive batch combination that fills this gap, and potentially verifies the recommendations given in deliverable 3.3 in the future.

Further reflecting on the work of the work conducted for the creation of the deliverable, it has become clear that many readers would benefit from the ability to quickly put individual experiment descriptions into a more general context, e.g. by relating different components and approaches described to an overall MLIA "flow". The Grid@CLEF pipeline is a good start at a subset of such a "flow".

This contribution by Grid@CLEF is especially important as the distributed nature of the organization of CLEF evaluation tracks makes it difficult to introduce much more structure into the experiment descriptions. In the future, a desirable addition may be a glossary of MLIA/CLIR terms as used in CLEF, which would help participants to describe their experiments in a consistent way, and practitioners to more easily locate all relevant descriptions.

There has also been a considerable effort been made to design and implement tasks within a number of tracks, especially ad-hoc, CLEF-IP and LogCLEF, to 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. 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.

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 our 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

[Agirre et al. 2009a] Agirre, E., Di Nunzio, G. M., Mandl, T., and Otegi, A. (2009). CLEF 2009 Ad Hoc Track Overview: Robust – WSD Task. In [Borri et al. 2009]. Published online at: http://www.clef-campaign.org/2009/working_notes/agirre-robustWSDtask-paperCLEF2009.pdf

[Agirre et al. 2009b] Agirre, E., Ansa, O., Arregi, X., Lopez de Lacalle, M., Otegi, A., Saralegi, X., and Zaragoza, H. (2009). Elhuyar-IXA: Semantic Relatedness and Cross-lingual Passage Retrieval. In [Borri et al. 2009]. Published online at: http://www.clef-campaign.org/2009/working_notes/agirre-paperCLEF2009.pdf

[Agosti 2008] Agosti, M. (2008). Log Data in Digital Libraries. In: Agosti, M., Esposito, F., and Thanos, C. editors. *Post-proceedings of the Fourth Italian Research Conference on Digital Library Systems (IRCDL 2008)*, pages 115–121. DELOS: an Association for Digital Libraries.

[Agosti and Ferro 2009] Agosti, M. and Ferro, N. (2009). Towards an Evaluation Infrastructure for DL Performance Evaluation. In Tsakonas, G. and Papatheodorou, C., editors, *Evaluation of Digital Libraries: An insight into useful applications and methods*, pages 93–120. Chandos Publishing, Oxford, UK.

[Anderka et al. 2009] Anderka, M., Lipka, N., and Stein, B. (2009). Evaluating Cross-Language Explicit Semantic Analysis and Cross Querying at TEL@CLEF 2009. In [Borri et al. 2009]. Published online at: http://www.clef-campaign.org/2009/working_notes/anderka-paperCLEF2009.pdf

[Besançon et al. 2008] Besançon R., Chaudiron S., Mostefa D., Hamon O., Timimi I. and Choukri K. (2008) Overview of CLEF 2008 INFILE Pilot Track. In [Borri et al. 2009]. Published online at: http://www.clef-campaign.org/2009/working_notes/CLEF_INFILE_report_2009_v3.pdf

[Bharadwaj et al. 2009] Bharadwaj, R., Ganesh, S., and Varma, V. (2009). A Naïve Approach for Monolingual Question Answering. In [Borri et al. 2009]. Published online at: http://www.clef-campaign.org/2009/working_notes/Bharadwaj.pdf

[Borri et al. 2009] Borri, F., Nardi, A., and Peters, C., editors (2009). *Working Notes for the CLEF 2009 Workshop*. Published Online at http://www.clef-campaign.org/2009/working_notes/CLEF2009WN-Contents.html.

[Bosca and Dini 2009a] Bosca, A. and Dini, L. (2009). CACAO Project at the TEL@CLEF 2009 Task. In [Borri et al. 2009]. Published online at: http://www.clef-campaign.org/2009/working_notes/bosca-paperCLEF2009.pdf

[Bosca and Dini 2009b] Bosca, A. and Dini, L. (2009). CACAO Project at the LogCLEF Track. In [Borri et al. 2009]. Published online at: http://www.clef-campaign.org/2009/working_notes/celi-paperLogCLEF2009.pdf

[Braschler 2002] Braschler, M. (2002). Report on CLEF 2001 – Overview of Results. In Peters, C., Braschler, M., Gonzalo, J., and Kluck, M., editors, *Evaluation of Cross-Language Information Retrieval Systems: Second Workshop of the Cross–Language Evaluation Fo- rum (CLEF 2001) Revised Papers*, pages 9–26. Lecture Notes in Computer Science (LNCS) 2406, Springer, Heidelberg, Germany.

[Braschler 2003] Braschler, M. (2003). CLEF 2002 – Overview of Results. In Peters, C., Braschler, M., Gonzalo, J., and Kluck, M., editors, *Advances in Cross- Language Information Retrieval: Third Workshop of the Cross–Language Evaluation Forum (CLEF 2002) Revised Papers*, pages 9–27. Lecture Notes in Computer Science (LNCS) 2785, Springer, Heidelberg, Germany.

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

[Braschler and Peters 2004] Braschler, M. and Peters, C. (2004). CLEF 2003 Methodology and Metrics. In Peters, C., Braschler, M., Gonzalo, J., and Kluck, M., editors, *Comparative Evaluation of Multilingual Information Access Systems: Fourth Workshop of the Cross–Language Evaluation Forum (CLEF 2003) Revised Selected Papers*, pages 7–20. Lecture Notes in Computer Science (LNCS) 3237, Springer, Heidelberg, Germany.

[Caputo et al. 2009] Caputo, B., Pronobis, A., and Jensfelt, P. (2009). Overview of the CLEF 2009 Robot Vision Track. In [Borri et al. 2009]. Published online at: http://www.clef-campaign.org/2009/working_notes/CLEF2009_RobVisOverview.pdf

[Cleverdon 1967] Cleverdon, C. W. (1997). The Cranfield Tests on Index Languages Devices. In *Readings in Information Retrieval*, pages 47–60. Morgan Kaufmann Publisher.

[Correa et al. 2009] Correa, S., Buscaldi, D., and Rosso, P. (2009). NLEL-MAAT at CLEF-ResPubliQA. In [Borri et al. 2009]. Published online at: http://www.clef-campaign.org/2009/working_notes/Correa_QA.pdf

[Cristea et al. 2009] Cristea, F.-T., Alexa, V., and Iftene, A. (2009). UAIC at iCLEF 2009: Analysis of Logs of Multilingual Image Searches in Flickr .In [Borri et al. 2009]. Published online at: http://www.clef-campaign.org/2009/working_notes/cristeaCLEF2009_iCLEF.pdf

[Crivellari et al. 2007] Crivellari, F., Di Nunzio, G. M., and Ferro, N. (2007). How to Compare Bilingual to Monolingual Cross-Language Information Retrieval. In Am- ati, 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.

[Crivellari et al. 2008] 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.

[Di Nunzio and Ferro 2009a] Di Nunzio, G. M. and Ferro, N. (2009). Appendix A: Results of the TEL@CLEF Task. In [Borri et al. 2009]. Published online at http://www.clef-campaign.org/2009/working_notes/AppendixA.pdf

[Di Nunzio and Ferro 2009b] Di Nunzio, G. M. and Ferro, N. (2009). Appendix B: Results of the Persian@CLEF Task. In [Borri et al. 2009]. Published online at http://www.clef-campaign.org/2009/working_notes/AppendixB.pdf

[Di Nunzio and Ferro 2009c] Di Nunzio, G. M. and Ferro, N. (2009). Appendix C: Results of the Robust Task. In [Borri et al. 2009]. Published online at http://www.clef-campaign.org/2009/working_notes/AppendixC.pdf

[Dolamić et al. 2009] Dolamić, L., Fautsch, C., and Savoy, J. (2009). UniNE at CLEF 2009: Persian Ad Hoc Retrieval and IP. In [Borri et al. 2009]. Published online at: http://www.clef-campaign.org/2009/working_notes/Dolamic-paper.pdf

[Dussin and Ferro 2009] Dussin, M. and Ferro, N. (2009). Managing the Knowledge Creation Process of Large-Scale Evaluation Campaigns. In Agosti, M., Borbinha, J., Kapidakis, S., Papatheodorou, C., and Tsakonas, G., editors, *Proc. 13th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2009)*, pages 63–74. Lecture Notes in Computer Science (LNCS) 5714 , Springer, Heidelberg, Germany.

[Eibl and Kürsten 2009] Eibl, M. and Kürsten, J. (2009). The Importance of being Grid – Chemnitz University of Technology at Grid@CLEF. In [Borri et al. 2009]. Published online at: http://www.clef-campaign.org/2009/working_notes/Eibl_paper_CLEF2009_Grid.pdf

[Ferro 2008] Ferro, N. (2008). *Deliverable 2.2 – Operational Scientific Digital Library*. Treble-CLEF: Evaluation, Best Practices and Collaboration for Multilingual Information Access. Availabe online at:
http://www.trebleclef.eu/getfile.php?id=75

[Ferro 2009] Ferro, N. (2009). Specification of the CIRCO Framework, Version 0.10. Technical Report IMS.2009.CIRCO.0.10, Department of Information Engineering, University of Padua, Italy.

[Ferro and Harman 2009] Ferro, N. and Harman, D. (2009). CLEF 2009: Grid@CLEF Pilot Track Overview. In [Borri et al. 2009]. Published online at: http://www.clef-campaign.org/2009/working_notes/CLEF2009-gridCLEF-overview.pdf

[Ferro and Peters 2009a] Ferro, N. and Peters, C. (2009). CLEF 2009 Ad Hoc Track Overview: TEL & Persian Tasks. In [Borri et al. 2009]. Published online at: http://www.clef-campaign.org/2009/working_notes/CLEF2009-adhoc-final2.pdf

[Ferro and Peters 2009b] Ferro, N. and Peters, C. (2009). CLEF Ad-hoc: A Perspective on the Evolution of the Cross-Language Evaluation Forum. In Agosti, M., Esposito, F., and Thanos, C., editors, *Post-proceedings of the Fifth Italian Research Conference on Digital Library Systems (IRCDL 2009)*, pages 72–79. DELOS Association and Department of Information Engineering of the University of Padua.

[Fiscus and Wheatley 2004] Fiscus, J. and Wheatley, B. (2004). Overview of the TDT 2004 evaluation and results. In TDT'02. NIST.

[Ghorab et al. 2009] Ghorab, M.R., Leveling, J., Zhou, D., Jones, G.J.F., and Wade, V. (2009). TCD-DCU at LogCLEF 2009: An Analysis of Queries, Actions, and Interface Languages. In [Borri et al. 2009]. Published online at: http://www.clef-campaign.org/2009/working_notes/ghorab-paperCLEF2009.pdf

[Gloeckner and Pelzer 2009] Gloeckner, I. and Pelzer, B. (2009). The LogAnswer Project at CLEF 2009. In [Borri et al. 2009]. Published online at: http://www.clef-campaign.org/2009/working_notes/glockner-paperCLEF2009.pdf

[Hofman et al. 2009] Hofmann, K., de Rijke, M., Huurnink, B., and Meij, E: (2009). A Semantic Perspective on Query Log Analysis. In [Borri et al. 2009]. Published online at: http://www.clef-campaign.org/2009/working_notes/hofmann_et al-clef2009-querylog.pdf

[Hull and Roberston 1999] Hull, D. and Roberston, S. (1999). The trec-8 filtering track final report. In Proceedings of the Eighth Text REtrieval Conference (TREC-8). NIST.

[Ifene et al. 2009] Iftene, A., Trandabăţ, D., Pistol, I., Moruz, A.-M., Husarciuc, M., Sterpu, M., and Turliuc, C. (2009). Question Answering on English and Romanian Languages. In [Borri et al. 2009]. Published online at: http://www.clef-campaign.org/2009/working_notes/ifteneCLEF 2009_QA_ResPubliQA.pdf

[Ion et al. 2009] Ion, R., Ştefănescu, D., Ceauşu, A., Tufiş, D., Irimia, E., and Barbu-Mititelu, V. (2009). A Trainable Multi-factored QA System. In [Borri et al. 2009]. Published online at: http://www.clef-campaign.org/2009/working_notes/ion-paperCLEF2009.pdf

[Jadidinejad and Mahmoudi 2009] Jadidinejad, A. H. and Mahmoudi, F. (2009). Query Wikification: Min- ing Structured Queries From Unstructured Information Needs using Wikipedia-based Semantic Analysis. In [Borri et al. 2009]. Published online at: http://www.clef-campaign.org/2009/working_notes/Jadidineja d-paperCLEF2009.pdf

[Katsiouli and Kalamboukis 2009] Katsiouli, P. and Kalamboukis, T. (2009). An Evaluation of Greek-English Cross Language Retrieval within the CLEF Ad-Hoc Bilingual Task. In [Borri et al. 2009]. Published online at: http://www.clef-campaign.org/2009/working_notes/Katsiouli-paperCLEF2009.pdf

[Kürsten 2009] Kürsten, J. (2009). Chemnitz at CLEF 2009 Ad-Hoc TEL Task: Combining Different Retrieval Models and Addressing the Multilinguality. In [Borri et al. 2009]. Published online at: http://www.clef-campaign.org/2009/working_notes/kuersten-ah-paperCLEF2009.pdf

[Lamm et al. 2009] Lamm, K., Mandl, T., and Kölle, R. (2009). Search Path Visualization and Session Performance Evaluation with Log Files from The European Library (TEL). In [Borri et al. 2009]. Published online at: http://www.clef-campaign.org/2009/working_notes/lamm-paperCLEF2009.pdf

[Larson 2009a] Larson, R. R. (2009). Multilingual Query Expansion for CLEF Adhoc-TEL. In [Borri et al. 2009]. Published online at: http://www.clef-campaign.org/2009/working_notes/larson-Adhoc-TEL-CLEF2009.pdf

[Larson 2009b] Larson, R. R. (2009a). Decomposing Text Processing for Retrieval: Cheshire tries GRID@CLEF. In [Borri et al. 2009]. Published online at: http://www.clef-campaign.org/2009/working_notes/larson-GRIDCLEF2009.pdf

[Lestrari Paramita et al. 2009a] Lestari Paramita, M., Sanderson, M. and Clough, P. (2009) Developing a Test Collection to Support Diversity Analysis, In *Proceedings of Redundancy, Diversity, and Interdependence Document Relevance workshop held at ACM SIGIR*, pp. 39-45.

[Lestrari Paramita et al. 2009b] Lestari Paramita, M., Sanderson, M. and Clough, P. (2009). Diversity in Photo Retrieval: Overview of the ImageCLEFPhoto Task 2009. In [Borri et al. 2009]. Published online at: http://www.clef-campaign.org/2009/working_notes/overview_ImageCLEFPhoto.pdf

[Leveling et al. 2009] Leveling, J., Zhou, D., Jones, G. J. F., and Wade, V. (2009). TCD-DCU at TEL@CLEF 2009: Document Expansion, Query Translation and Lan- guage Modeling. In [Borri et al. 2009]. Published online at: http://www.clef-campaign.org/2009/working_notes/leveling-paperCLEF2009.pdf

[Mandl et al. 2009] Mandl, T., Agosti, M., Di Nunzio, G.M., Yeh, A., Mani, I., Doran, C., and Schulz, J.M. (2009). LogCLEF 2009: the Multilingual Logfile Analysis Track Overview. In [Borri Published online at: http://campaign.org/2009/working_notes/LogCLEF-200Overview-Working-Notes-2009-09-14.pdf

[McNamee 2009] McNamee, P. (2009). JHU Experiments in Monolingual Farsi Document Retrieval at CLEF 2009. In [Borri et al. 2009]. Published online at: http://www.clef-campaign.org/2009/working_notes/mcnamee-paperCLEF2009.pdf

[Moriceau et al. 2009] Moriceau, V. and Tannier, X. (2009). FIDJI in ResPubliQA 2009. In [Borri et al. 2009]. Published online at: http://www.clef-campaign.org/2009/working_notes/tannier-paperCLEF2009.pdf

[Müller et al. 2009] Müller, H., Kalpathy–Cramer, J., Eggel, I., Bedrick, S., Radhouani, S., Bakke, B., Kahn Jr., C. E., and Hersh. W. (2009). Overview of the CLEF 2009 Medical Image Retrieval Track. In [Borri et al. 2009]. Published online at: http://www.clef-campaign.org/2009/working_notes/CLEF2009_medOverview_jkc.pdf

[Navarro-Colorado et al. 2009] Navarro-Colorado, B., Puchol-Blasco, M., Terol, R. M., Vázquez, S., and Lloret, E. (2009). Lexical Ambiguity in Cross-language Image Retrieval: a Preliminary Analysis.In [Borri et al. 2009]. Published online at: http://www.clef-campaign.org/2009/working_notes/bnavarro-paperCLEF09.pdf

[Nowak and Dunker 2009] Nowak, S. and Dunker, P. (2009). Overview of the CLEF 2009 Large Scale - Visual Concept Detection and Annotation Task. In [Borri et al. 2009]. Published online at: http://www.clef-campaign.org/2009/working_notes/Overview_VCDT.pdf

[Oakes and Xu 2009] Oakes, M. and Xu, Y. (2009). A Search Engine based on Query Logs, and Search Log Analysis at the University of Sunderland. In [Borri et al. 2009]. Published online at: http://www.clef-campaign.org/2009/working_notes/oakes-paperCLEF2009.pdf

[Peinado et al. 2009] Peinado, V., López-Ostenero, F., and Gonzalo, J. (2009). UNED at iCLEF 2009: Analysis of Multilingual Image Search Sessions.In [Borri et al. 2009]. Published online at: http://www.clef-campaign.org/2009/working_notes/peinado-paperCLEF2009.pdf

[Peñas et al. 2007] Peñas, A., Rodrigo, A., Sama, V., and Verdejo (2007). F. Overview of the Answer Validation Exercise 2006. 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 257-264. Lecture Notes in Computer Science (LNCS) 4730, Springer, Heidelberg, Germany.

[Peñas et al. 2008] 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, pages 237-248. Lecture Notes in Computer Science (LNCS) 5152, Springer, Heidelberg, Germany.

[Pérez et al. 2009] Pérez, J., Garrido, G., Rodrigo, Á., Araujo, L., and Peñas. A. (2009). Information Retrieval Baselines for the ResPubliQA Task. In [Borri et al. 2009]. Published online at: http://www.clef-campaign.org/2009/working_notes/ResQABaselines.pdf

[Robertson 1981] Robertson, S. E. (1981). The methodology of information retrieval experiment. In Spärck Jones, K., editor, *Information Retrieval Experiment*, pages 9–31. Butterworths, London, United Kingdom.

[Robertson and Soboroff 2002] Robertson, S. and Soboroff, I. (2002). The TREC 2002 filtering track report. In Proceedings of The Eleventh Text Retrieval Conference (TREC 2002). NIST.

[Robertson and Walker 1994] Robertson, S.E. and Walker, S. (1994). Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval. In SIGIR '94: Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval, pages. 232-241.

[Rodrigo et al. 2009a] Rodrigo, Á., Peñas, A., and Verdejo, F. (2009). Overview of the Answer Validation Exercise 2008. In Peters, C., Deselaers, T., Ferro, N., Gonzalo, J., Jones, G. J. F., Kurimo, M., Mandl, T., Peñas, A., editors, Evaluating Systems for Multilingual and Multimodal Information Access: Ninth Workshop of the Cross--Language Evaluation Forum (CLEF 2008). Revised Selected Papers, pages 296-313. Lecture Notes in Computer Science (LNCS) 5706, Springer, Heidelberg, Germany

[Rodrigo et al. 2009b] Rodrigo, Á., Pérez, J., Peñas, A., Garrido, G., and Araujo, L. (2009). Approaching Question Answering by means of Paragraph Validation. In [Borri et al. 2009]. Published online at: http://www.clef-campaign.org/2009/working_notes/rodrigo-paperCLEF2009.pdf

[Ruiz and Chin 2009] Ruiz, M. E. and Chin, P. (2009). Users' Image Seeking Behavior in a Multilingual Tag Environment .In [Borri et al. 2009]. Published online at: http://www.clef-campaign.org/2009/working_notes/Ruiz-paper-interactive-CLEF2009-v2.pdf

[Savoy 2002] Savoy, J. (2002). Report on CLEF-2001 Experiments: Effective Combined Query-Translation Approach. In Peters, C., Braschler, M., Gonzalo, J., and Kluck, M., editors, *Evaluation of Cross-Language Information Re- trieval Systems: Second Workshop of the Cross–Language Evaluation Fo- rum (CLEF 2001) Revised Papers*, pages 27–43. Lecture Notes in Computer Science (LNCS) 2406, Springer, Heidelberg, Germany.

[Tomlinson et al. 2007] Tomlinson, S., Oard, D.W., Baron, J.R., and Thompson, P. (2007). Overview of the TREC 2007 Legal Track. Proceedings of The Sixteenth Text REtrieval Conference, TREC 2007.

[Tommasi et al. 2009] Tommasi, T., Caputo, B., Welter, P., Güld, M. O., and Deserno, T. M. (2009). Overview of the CLEF 2009 Medical Image Annotation Track. In [Borri et al. 2009]. Published online at: http://www.clef-campaign.org/2009/working_notes/CLEF2009_MedAnnot_Overview-TD.pdf

[Tsikrika and Kludas 2009] Tsikrika, T. and Kludas, J. (2009). Overview of the WikipediaMM Task at ImageCLEF 2009. In [Borri et al. 2009]. Published online at: http://www.clef-campaign.org/2009/working_notes/wikiMMoverview-paperCLEF2009.pdf

[Vassilakaki et al. 2009] Vassilakaki, E., Johnson, F., Hartley, R.J., and Randall, D. (2009). Users' Perceptions of Searching in FlickLing. In [Borri et al. 2009]. Published online at: http://www.clef-campaign.org/2009/working_notes/vassilaki.pdf

[Vicente-Díez et al. 2009] Vicente-Díez, M.T., de Pablo-Sánchez, C., Martínez, P., Moreno Schneider, J., and Garrote Salazar, M. (2009). Are Passages Enough? The MIRACLE Team Participation at QA@CLEF2009. In [Borri et al. 2009]. Published online at: http://www.clef-campaign.org/2009/working_notes/vicente-diez-paperResPubliQA2009.pdf

[Voorhees 1999] Voorhees, E. (1999). The trec-8 question answering track report. In Proceedings of the Eighth Text REtrieval Conference (TREC-8). NIST.

[Yang et al. 2005] Yang, Y., Yoo, S., Zhang, J., and Kisiel, B. (2005). Robustness of adaptive filtering methods in a cross-benchmark evaluation. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in inormation retrieval*, pages 98–105, Salvador, Brazil.