

CLEF 2008: Ad Hoc Track Overview

Eneko Agirre¹, Giorgio M. Di Nunzio², Nicola Ferro², Thomas Mandl³, and Carol Peters⁴

¹ Computer Science Department, University of the Basque Country, Spain
e.agirre@ehu.es

² Department of Information Engineering, University of Padua, Italy
{dinunzio, ferro}@dei.unipd.it

³ Information Science, University of Hildesheim, Germany
mandl@uni-hildesheim.de

⁴ ISTI-CNR, Area di Ricerca, Pisa, Italy
carol.peters@isti.cnr.it

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

Categories and Subject Descriptors

H.3 [Information Storage and Retrieval]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval; H.3.4 [Systems and Software]: Performance evaluation.

General Terms

Experimentation, Performance, Measurement, Algorithms.

Additional Keywords and Phrases

Multilingual Information Access, Cross-Language Information Retrieval, Word Sense Disambiguation

1 Introduction

The ad hoc retrieval track is generally considered to be the core track in the *Cross-Language Evaluation Forum (CLEF)*. The aim of this track is to promote the development of monolingual and cross-language textual document retrieval

systems. From 2000 - 2007, the track used exclusively collections of European newspaper and news agency documents¹. This year the focus of the track was considerably widened: we introduced very different document collections, a non-European target language, and an information retrieval (IR) task designed to attract participation from groups interested in natural language processing (NLP). The track was thus structured in three distinct streams:

- TEL@CLEF
- Persian@CLEF
- Robust WSD

The first task offered monolingual and cross-language search on library catalog records and was organized in collaboration with The European Library (TEL)². The second task resembled the ad hoc retrieval tasks of previous years but this time the target collection was a Persian newspaper corpus.

The third task was the robust activity which this year used word sense disambiguated (WSD) data, and involved English documents and monolingual and cross-language search in Spanish.

In this paper we first present the track setup, the evaluation methodology and the participation in the different tasks (Section 2). We then describe the main features of each task and show the results (Sections 3 - 5). Statistical testing is discussed in Section 6 and the final section provides a brief summing up.

For information on the various approaches and resources used by the groups participating in this track and the issues they focused on, we refer the reader to the other papers in the Ad Hoc section of these Working Notes.

2 Track Setup

The ad hoc track in CLEF adopts a corpus-based, automatic scoring method for the assessment of system performance, based on ideas first introduced in the Cranfield experiments in the late 1960s [13]. The **tasks** offered are studied in order to effectively measure textual document retrieval under specific conditions. The test collections are made up of **documents**, **topics** and **relevance assessments**. The topics consist of a set of statements simulating information needs from which the systems derive the queries to search the document collections. Evaluation of system performance is then done by judging the documents retrieved in response to a topic with respect to their relevance, and computing the recall and precision measures. The distinguishing feature of CLEF is that it applies this evaluation paradigm in a multilingual setting. This means that the criteria normally adopted to create a test collection, consisting of suitable documents, sample queries and relevance assessments, have been adapted to

¹ Over the years, this track has built up test collections for monolingual and cross-language system evaluation in 13 European languages (see the Introduction to this volume for more details)

² See <http://www.theeuropeanlibrary.org/>

satisfy the particular requirements of the multilingual context. All language dependent tasks such as topic creation and relevance judgment are performed in a distributed setting by native speakers. Rules are established and a tight central coordination is maintained in order to ensure consistency and coherency of topic and relevance judgment sets over the different collections, languages and tracks.

2.1 The Documents

Each of the three ad hoc tasks this year used a different set of 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. About 66% of the documents in the English and German collection have textual subject headings, in the French collection only 37%. Dewey Classification (DDC) is not available in the French collection; negligible (0.3%) in the German collection; but occurs in about half of the English documents (456,408 docs to be exact).

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

The Persian task used the Hamshahri corpus of 1996-2002 newspapers as the target collection. This corpus was made available to CLEF by the Data Base Research Group (DBRG) of the University of Tehran. Hamshahri is one of the most popular daily newspapers in Iran. The Hamshahri corpus 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 size³.

³ For more information, see <http://ece.ut.ac.ir/dbrg/hamshahri/>

The robust task used existing CLEF news collections but with word sense disambiguation (WSD) added. The word sense disambiguation data was automatically added by systems from two leading research laboratories, UBC [2] and NUS [12]. Both systems returned word senses from the English WordNet, version 1.6.

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

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

An excerpt for a document⁴ is shown in Figure 1, where each term in the document is followed by its senses with their respective scores as assigned by the automatic WSD system.

2.2 Topics

Topics in the CLEF ad hoc track are structured statements representing information needs. Each topic typically consists of three parts: a brief “title” statement; a one-sentence “description”; a more complex “narrative” specifying the relevance assessment criteria. Topics are prepared in xml format and identified by means of a Digital Object Identifier (DOI)⁵ of the experiment [35] which allows us to reference and cite them.

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

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

⁴ Full sample and dtd are available at <http://ixa2.si.ehu.es/clirwsd/>

⁵ <http://www.doi.org/>

```

<DOC>
<DOCNO>GH950102-000000</DOCNO>
<DOCID>GH950102-000000</DOCID>

<HEADLINE>
<TERM ID="GH950102-000000-1" LEMA="alien" POS="JJ">
  <WF>Alien</WF>
  <SYNSET SCORE="0.6" CODE="01295935-a"/>
  <SYNSET SCORE="0.4" CODE="00984080-a"/>
</TERM>

<TERM ID="GH950102-000000-2" LEMA="treatment" POS="NN">
  <WF>treatment</WF>
  <SYNSET SCORE="0.827904118008605" CODE="00735486-n"/>
  <SYNSET SCORE="0" CODE="03857483-n"/>
  <SYNSET SCORE="0.172095881991395" CODE="00430183-n"/>
  <SYNSET SCORE="0" CODE="05340429-n"/>
</TERM>

<TERM ID="GH950102-000000-3" LEMA="be" POS="VBZ">
  <WF>is</WF>
  <SYNSET SCORE="0.0113384126222329" CODE="01787769-v"/>
  <SYNSET SCORE="0.181174635551023" CODE="01784339-v"/>
  <SYNSET SCORE="0.644489771431999" CODE="01775973-v"/>
  <SYNSET SCORE="0.00515927770112184" CODE="01666138-v"/>
  <SYNSET SCORE="0.0420541124242606" CODE="01775163-v"/>
  <SYNSET SCORE="0.00347951286819845" CODE="01840295-v"/>
  <SYNSET SCORE="0.0540524326594277" CODE="01811792-v"/>
  <SYNSET SCORE="0" CODE="01843641-v"/>
  <SYNSET SCORE="0.000119983202351671" CODE="01552250-v"/>
  <SYNSET SCORE="0.0418741376207331" CODE="01781222-v"/>
  <SYNSET SCORE="5.99916011758354e-05" CODE="01782836-v"/>
  <SYNSET SCORE="0.0161977323174756" CODE="01817610-v"/>
</TERM>

...
</HEADLINE>

...
</DOC>

```

Fig. 1. Example of Robust WSD document.

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

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

For example, Iran commonly uses a different calendar from Europe and reference was often made in the Persian topics to events that are well known to Iranian society but not often discussed in English. This is shown in the example of Figure 3, where the rather awkward English rendering evidences the uncertainty of the translator.

```

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

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

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

```

Fig. 2. Example of TEL topic in all five languages: topic 10.2452/451-AH.

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

- CLEF 2001; Topics 41-90; LA Times 94
- CLEF 2002; Topics 91-140; LA Times 94
- CLEF 2003; Topics 141-200; LA Times 94, Glasgow Herald 95
- CLEF 2004; Topics 201-250; Glasgow Herald 95
- CLEF 2005; Topics 251-300; LA Times 94, Glasgow Herald 95
- CLEF 2006; Topics 301-350; LA Times 94, Glasgow Herald 95

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

All topics were offered both with and without WSD. Topics in English were disambiguated by both UBC [2] and NUS [12] systems, yielding word senses from WordNet version 1.6. A large-scale disambiguation system for Spanish was not available, so we used the first-sense heuristic, yielding senses from the Spanish wordnet, which is tightly aligned to the English WordNet version 1.6 (i.e., they share synset numbers or sense codes). An excerpt for a topic⁶ is shown in Figure 4, where each term in the topic is followed by its senses with their respective scores as assigned by the automatic WSD system.

2.3 Relevance Assessment

The number of documents in large test collections such as CLEF makes it impractical to judge every document for relevance. Instead approximate recall values are calculated using pooling techniques. The results submitted by the groups

⁶ Full sample and dtd are available at <http://ixa2.si.ehu.es/clirwsd/>

```

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

  <description lang="en">Find documents that include information about the 2nd of Khordad
  presidential elections.</description>
  <description lang="fa">سندھایی را پیدا کن کہ شامل اطلاعاتی در مورد انتخابات دوم خرداد ماه سال 76
  هستند</description>

  <narrative lang="en">Any information about candidates and their sayings, Khatami's unexpected
  winning in the 2nd of Khordad 1376 presidential election is relevant.</narrative>
  <narrative lang="fa">سندھای مربوط شامل اطلاعاتی در مورد نامزدها و گفته های آنها، پیروزی
  غیرمنتظره خاتمی در انتخابات ریاست جمهوری در دوم خرداد ماه سال 76 است</narrative>
</topic>

```

Fig. 3. Example of Persian topic: topic 10.2452/599-AH.

participating in the ad hoc tasks are used to form a pool of documents for each topic and language by collecting the highly ranked documents from selected runs according to a set of predefined criteria. Traditionally, the top 100 ranked documents from each of the runs selected are included in the pool; in such a case we say that the pool is of depth 100. This pool is then used for subsequent relevance judgments. After calculating the effectiveness measures, the results are analyzed and run statistics produced and distributed.

The stability of pools constructed in this way and their reliability for post-campaign experiments is discussed in [9] with respect to the CLEF 2003 pools. New pools were formed in CLEF 2008 for the runs submitted for the TEL and the Persian mono- and bilingual tasks. Instead, the robust tasks used the original pools and relevance assessments from previous CLEF campaigns.

The main criteria used when constructing the pools were:

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

One important limitation when forming the pools is the number of documents to be assessed. Last year, for collections of newspaper documents, we estimated that assessors could judge from 60 to 100 documents per hour, providing binary

```

<top>
<num>10.2452/141-WSD-AH</num>
<EN-title>
  <TERM ID="10.2452/141-WSD-AH-1" LEMA="letter" POS="NNP">
    <WF>Letter</WF>
    <SYNSET SCORE="0" CODE="05115901-n"/>
    <SYNSET SCORE="0" CODE="05362432-n"/>
    <SYNSET SCORE="0" CODE="05029514-n"/>
    <SYNSET SCORE="1" CODE="04968965-n"/>
  </TERM>
  <TERM ID="10.2452/141-WSD-AH-2" LEMA="bomb" POS="NNP">
    <WF>Bomb</WF>
    <SYNSET SCORE="0.888888888888889" CODE="02310834-n"/>
    <SYNSET SCORE="0" CODE="05484679-n"/>
    <SYNSET SCORE="0.111111111111111" CODE="02311368-n"/>
  </TERM>
  <TERM ID="10.2452/141-WSD-AH-3" LEMA="for" POS="IN">
    <WF>for</WF>
  </TERM>
  ...
</EN-title>
<EN-desc>
  <TERM ID="10.2452/141-WSD-AH-5" LEMA="find" POS="VBP">
    <WF>Find</WF>
    <SYNSET SCORE="0" CODE="00658116-v"/>
    ...
  </TERM>
  ...
</EN-desc>
<EN-narr>
  ...
</EN-narr>
</top>

```

Fig. 4. Example of Robust WSD topic: topic 10.2452/141-WSD-AH.

judgments: relevant / not relevant. Our estimate this year for the TEL catalog records was higher as these records are much shorter than the average newspaper article (100 to 120 documents per hour). In both cases, it can be seen what a time-consuming and resource expensive task human relevance assessment is. This limitation impacts strongly on the application of the criteria above - and implies that we are obliged to be flexible in the number of documents judged per selected run for individual pools.

This meant that this year, in order to create pools of more-or-less equivalent size (approx. 25,000 documents), the depth of the TEL English, French, and German pools and of the Persian pool was 60⁷.

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

Table 1 reports summary information on the 2008 ad hoc pools used to calculate the results for the main monolingual and bilingual experiments. In particular, for each pool, we show the number of topics, the number of runs submitted, the number of runs included in the pool, the number of documents in the pool (relevant and non-relevant), and the number of assessors.

The box plot of Figure 5 compares the distributions of the relevant documents across the topics of each pool for the different ad hoc pools; the boxes are ordered by decreasing mean number of relevant documents per topic.

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

For the TEL documents, we judged for relevance only those documents that are written totally or partially in English, French and German (and Spanish for searches on the English collection as we expected this language to be used only for ES to EN runs), e.g. a catalog record written entirely in Hungarian was counted as not relevant as it was of no use to our hypothetical user; however, a catalog record with perhaps the title and a brief description in Hungarian, but with subject descriptors in French, German or English was judged for relevance as it could be potentially useful. Our assessors had no additional knowledge of the documents referred to by the catalog records (or surrogates) contained in the collection. They judged for relevance on the information contained in the records made available to the systems. This was a non trivial task due to the lack of information present in the documents. During the relevance assessment activity there was much consultation between the assessors for the three TEL collections in order to ensure that the same assessment criteria were adopted by everyone.

As shown in the box plot of Figure 5, the Persian distribution presents a greater number of relevant documents per topic with respect to the other distributions 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 1, it has been possible to sample all the experiments submitted for the Persian tasks. This means that there were fewer unique documents per run and this fact, together with the greater number of relevant documents per topic suggests either that all the systems were using similar approaches and retrieval algorithms (however this is not so - see Section 4 below) or that the systems found the Persian topics quite easy.

Table 1. Summary information about CLEF 2008 pools.

TEL English Pool (DOI 10.2454/AH-TEL-ENGLISH-CLEF2008)	
Pool size	28,104 pooled documents – 25,571 not relevant documents – 2,533 relevant documents 50 topics
Pooled Experiments	21 out of 61 submitted experiments – monolingual: 13 out of 37 submitted experiments – bilingual: 8 out of 24 submitted experiments
Assessors	3 assessors
TEL French Pool (DOI 10.2454/AH-TEL-FRENCH-CLEF2008)	
Pool size	24,530 pooled documents – 23,191 not relevant documents – 1,339 relevant documents 50 topics
Pooled Experiments	14 out of 45 submitted experiments – monolingual: 9 out of 29 submitted experiments – bilingual: 5 out of 16 submitted experiments
Assessors	3 assessors
TEL German Pool (DOI 10.2454/AH-TEL-GERMAN-CLEF2008)	
Pool size	28,734 pooled documents – 27,097 not relevant documents – 1,637 relevant documents 50 topics
Pooled Experiments	16 out of 47 submitted experiments – monolingual: 10 out of 30 submitted experiments – bilingual: 6 out of 17 submitted experiments
Assessors	4 assessors
Persian Pool (DOI 10.2454/AH-PERSIAN-CLEF2008)	
Pool size	26,814 pooled documents – 21,653 not relevant documents – 5,161 relevant documents 50 topics
Pooled Experiments	66 out of 66 submitted experiments – monolingual: 53 out of 53 submitted experiments – bilingual: 13 out of 13 submitted experiments
Assessors	22 assessors

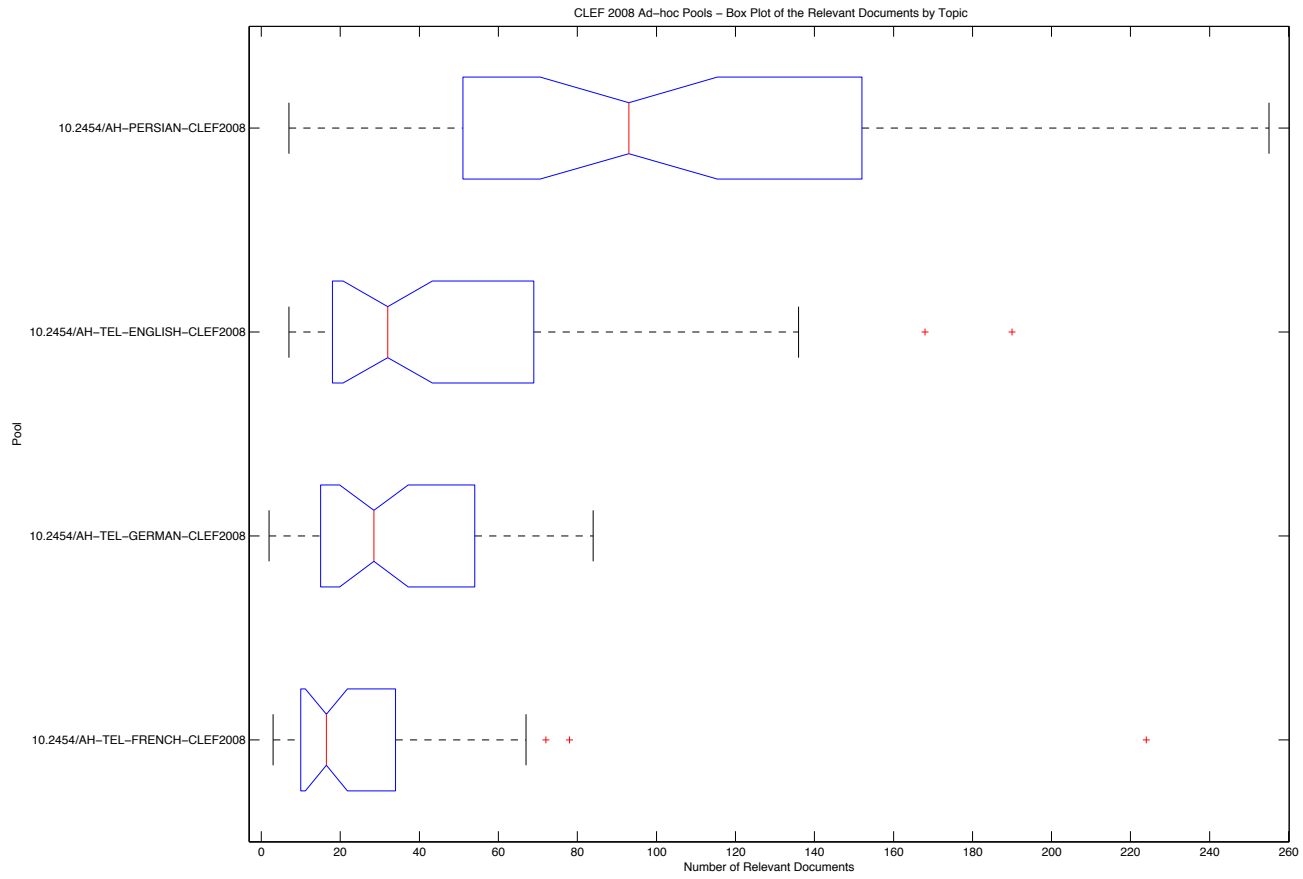


Fig. 5. Distribution of the relevant documents across the ad-hoc pools.

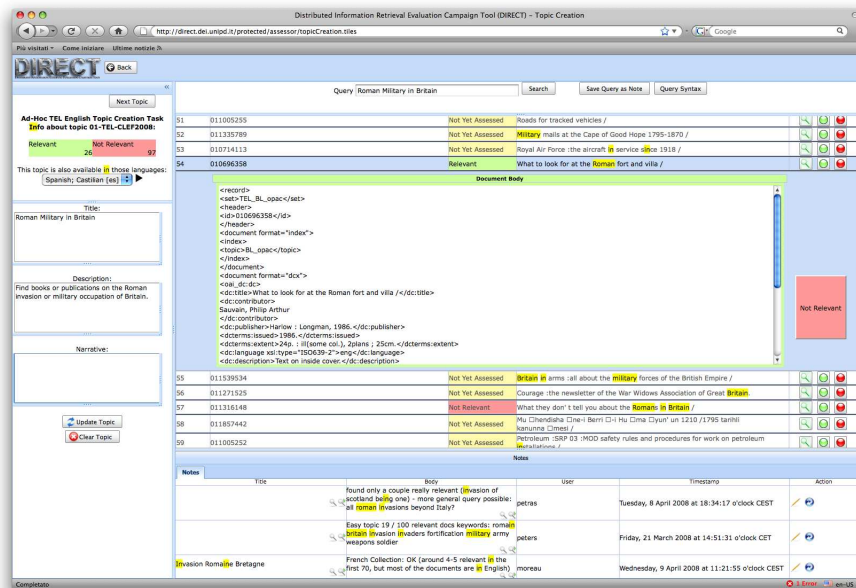


Fig. 6. Topic creation interface of the DIRECT system.

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.

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

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

Similarly to last year, in his paper, Stephen Tomlinson, has reported some sampling experiments aimed at estimating the judging coverage for the CLEF 2008 test collections. He finds that this tends to be lower than the estimates he produced for the CLEF 2007 collections. With respect to the TEL collections, the implication is that at best 50% to 70% of the relevant documents are included in the pools - and that most of the unjudged relevant documents are for the 10 or more queries that have the most known answers [40]. For Persian the coverage

seems to be lower; this could be a result of the fact that all the Persian topics tend to be relatively broad. It is our intention to look more closely into the question of coverage of these pools by performing some post-workshop stability tests. The results will be reported in our Proceedings paper.

2.4 Result Calculation

Evaluation campaigns such as TREC and CLEF are based on the belief that the effectiveness of *Information Retrieval Systems (IRSs)* can be objectively evaluated by an analysis of a representative set of sample search results. For this, effectiveness measures are calculated based on the results submitted by the participants and the relevance assessments. Popular measures usually adopted for exercises of this type are Recall and Precision. Details on how they are calculated for CLEF are given in [10]. For the robust task, we used additional measures, see Section 5.

The individual results for all official ad hoc experiments in CLEF 2008 are given in the Appendices at the end of these Working Notes [17,18,19].

2.5 Participants and Experiments

As shown in Table 2, a total of 24 groups from 14 different countries submitted official results for one or more of the ad hoc tasks - a slight increase on the 22 participants of last year. Table 3 provides a breakdown of the number of participants by country⁸

A total of 289 runs were submitted with an increase of about 22% on the 235 runs of 2007. The average number of submitted runs per participant also increased: from 10.6 runs/participant of 2007 to 12.0 runs/participant of this year.

Participants were required to submit at least one title+description (“TD”) run per task in order to increase comparability between experiments. The large majority of runs (215 out of 289, 74.40%) used this combination of topic fields, 27 (9.34%) used all fields⁹, 47 (16.26%) used the title field. The majority of experiments were conducted using automatic query construction (273 out of 289, 94.47%) and only in a small fraction of the experiments (16 out 289, 5.53%) were queries been manually constructed from topics. A breakdown into the separate tasks is shown in Table 4(a).

Seven different topic languages were used in the ad hoc experiments. As always, the most popular language for queries was English, with Farsi second. The number of runs per topic language is shown in Table 4(b).

⁸ Two additional Spanish groups presented results after the deadline for the robust tasks; their results were thus not reported in the official list but their papers are included in this volume [30], [32].

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

Table 2. CLEF 2008 ad hoc participants

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

Table 3. CLEF 2008 ad hoc participants by country.

Country	# Participants
Austria	1
Brazil	1
Canada	1
France	2
Germany	2
India	1
Iran	5
Italy	2
Mexico	1
Netherlands	1
Portugal	1
Spain	2
Switzerland	2
United States	2
Total	24

Table 4. Breakdown of experiments into tracks and topic languages.

(a) Number of experiments per track, participant.			(b) List of experiments by topic language.	
Track	# Part.	# Runs	Topic Lang.	# Runs
TEL Mono English	13	37	English	120
TEL Mono French	9	29	Farsi	51
TEL Mono German	10	30	German	44
TEL Bili English	8	24	French	44
TEL Bili French	5	16	Spanish	26
TEL Bili German	6	17	Dutch	3
Mono Persian	8	53	Portuguese	1
Bili Persian	3	13	Total	289
Robust Mono English Test	8	20		
Robust Mono English Training	1	2		
Robust Bili English Test	4	8		
Robust Mono English Test WSD	7	25		
Robust Mono English Training WSD	1	5		
Robust Bili English Test WSD	4	10		
Total		289		

3 TEL@CLEF

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

3.1 Tasks

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

In CLEF 2008, the activity we simulated was that of users who have a working knowledge of English, French and German (plus wrt the English collection also Spanish) and who want to discover the existence of relevant documents that can be useful for them in one of our three target collections. One of our suppositions was that, knowing that these collections are to some extent multilingual, some systems may attempt to use specific tools to discover this. For example, a system trying the cross-language English to French task on the BNF target collection

Table 5. Best entries for the monolingual TEL tasks.

Track	Rank	Participant	Experiment DOI	MAP
English	1st	unine	10.2415/AH-TEL-MONO-EN-CLEF2008.UNINE.UNINEEN3	37.53%
	2nd	inesc	10.2415/AH-TEL-MONO-EN-CLEF2008.INESC.RUN3	36.23%
	3rd	chemnitz	10.2415/AH-TEL-MONO-EN-CLEF2008.CHEMNITZ.CUT_SIMPLE	35.61%
	4th	jhu-apl	10.2415/AH-TEL-MONO-EN-CLEF2008.JHU-APL.JHUMOEN4RF	35.31%
	5th	cheshire	10.2415/AH-TEL-MONO-EN-CLEF2008.CHESHIRE.BKAHTELMFRTDT2F	34.66%
	Difference			
French	1st	unine	10.2415/AH-TEL-MONO-FR-CLEF2008.UNINE.UNINEFR3	33.27%
	2nd	xerox	10.2415/AH-TEL-MONO-FR-CLEF2008.XEROX.J1	30.88%
	3rd	jhu-apl	10.2415/AH-TEL-MONO-FR-CLEF2008.JHU-APL.JHUMOFR4	29.50%
	4th	opentext	10.2415/AH-TEL-MONO-FR-CLEF2008.OPENTEXT.OTFR08TD	25.23%
	5th	cheshire	10.2415/AH-TEL-MONO-FR-CLEF2008.CHESHIRE.BKAHTELMFRTDT2FB	24.37%
	Difference			
German	1st	opentext	10.2415/AH-TEL-MONO-DE-CLEF2008.OPENTEXT.OTDE08TDE	35.71%
	2nd	jhu-apl	10.2415/AH-TEL-MONO-DE-CLEF2008.JHU-APL.JHUMODE4	33.77%
	3rd	unine	10.2415/AH-TEL-MONO-DE-CLEF2008.UNINE.UNINEDE1	30.12%
	4th	xerox	10.2415/AH-TEL-MONO-DE-CLEF2008.XEROX.T1	27.36%
	5th	inesc	10.2415/AH-TEL-MONO-DE-CLEF2008.INESC.RUN3	22.97%
	Difference			

but knowing that documents retrieved in English and German will also be judged for relevance might choose to employ an English-German as well as the probable English-French dictionary. Groups attempting anything of this type were asked to declare such runs with a ++ indication.

3.2 Participants

13 groups submitted 153 runs for the TEL task: all groups submitted monolingual runs (96 runs out of 153); 8 groups also submitted bilingual runs (57 runs out of 153). Table 4(a) provides a breakdown of the number of participants and submitted runs by task.

3.3 Results.

Monolingual Results

Table 5 shows the top five groups for each target collection, ordered by mean average precision. The table reports: the short name of the participating group; the mean average precision achieved by the experiment; the DOI of the experiment; and the performance difference between the first and the last participant.

Figures 7, 9, and 11 compare the performances of the top participants of the TEL Monolingual tasks.

Table 6. Best entries for the bilingual TEL tasks.

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

Bilingual Results

Table 6 shows the top five groups for each target collection, ordered by mean average precision. The table reports: the short name of the participating group; the mean average precision achieved by the experiment; the DOI of the experiment; and the performance difference between the first and the last participant.

Figures 8, 10, and 12 compare the performances of the top participants of the TEL Bilingual tasks.

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

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

While the best result for English, obtained with German topics, is very good and can be considered as state-of-the-art for a good cross-language system running on well-tested languages with reliable processing tools and resources such as English and German, the results for the other two target collections are fairly disappointing. We have no explanation for this at the present.

3.4 Approaches

In the TEL experiments, all the traditional approaches to monolingual and cross-language retrieval were attempted by the different groups. Retrieval algorithms

Ad-Hoc TEL Monolingual English Task Top 5 Participants – Standard Recall Levels vs Mean Interpolated Precision

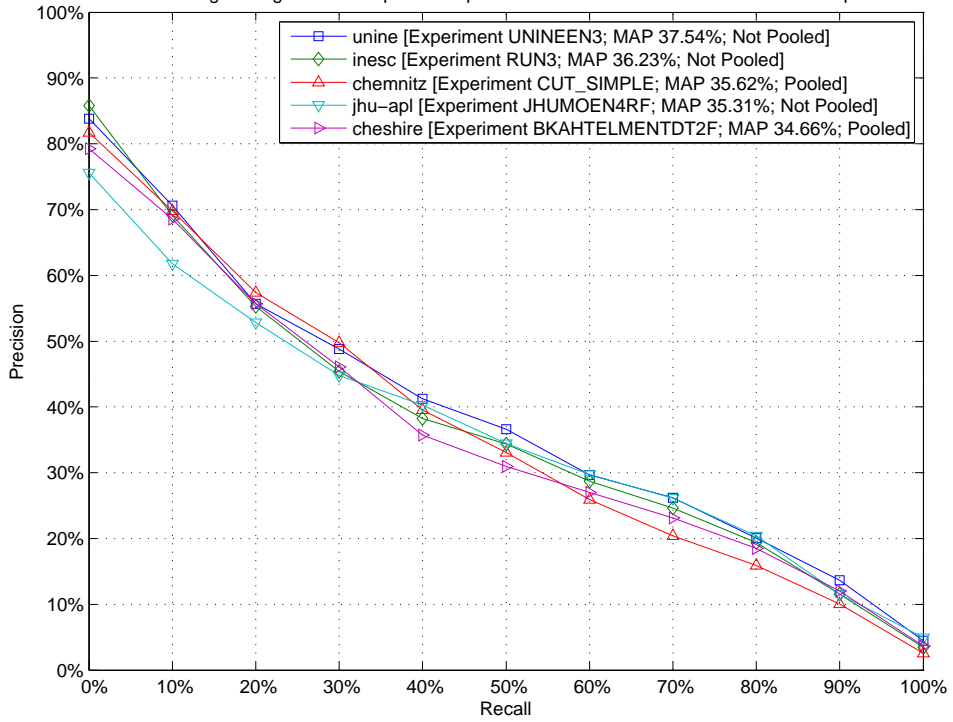


Fig. 7. Monolingual English

Ad-Hoc TEL Bilingual English Task Top 5 Participants – Standard Recall Levels vs Mean Interpolated Precision

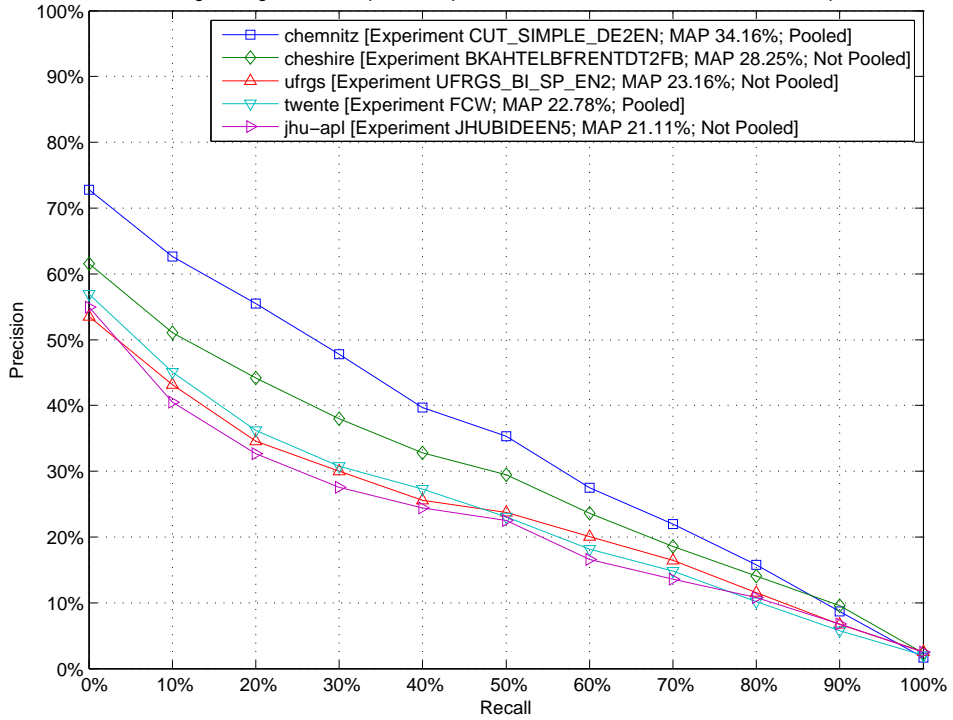


Fig. 8. Bilingual English

Ad-Hoc TEL Monolingual French Task Top 5 Participants – Standard Recall Levels vs Mean Interpolated Precision

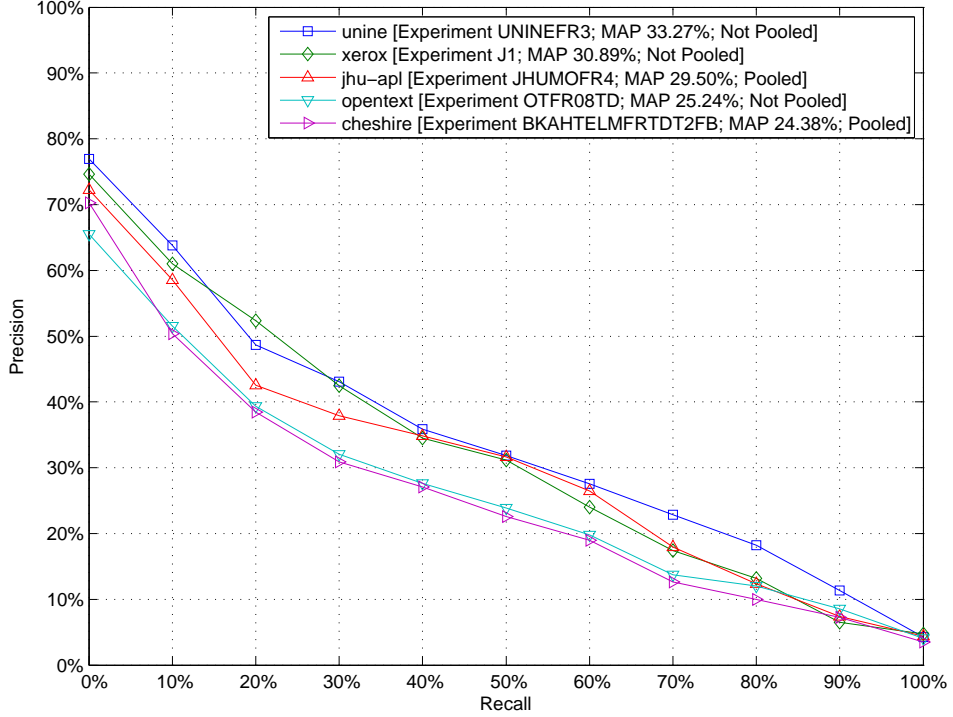


Fig. 9. Monolingual French

Ad-Hoc TEL Bilingual French Task Top 5 Participants – Standard Recall Levels vs Mean Interpolated Precision

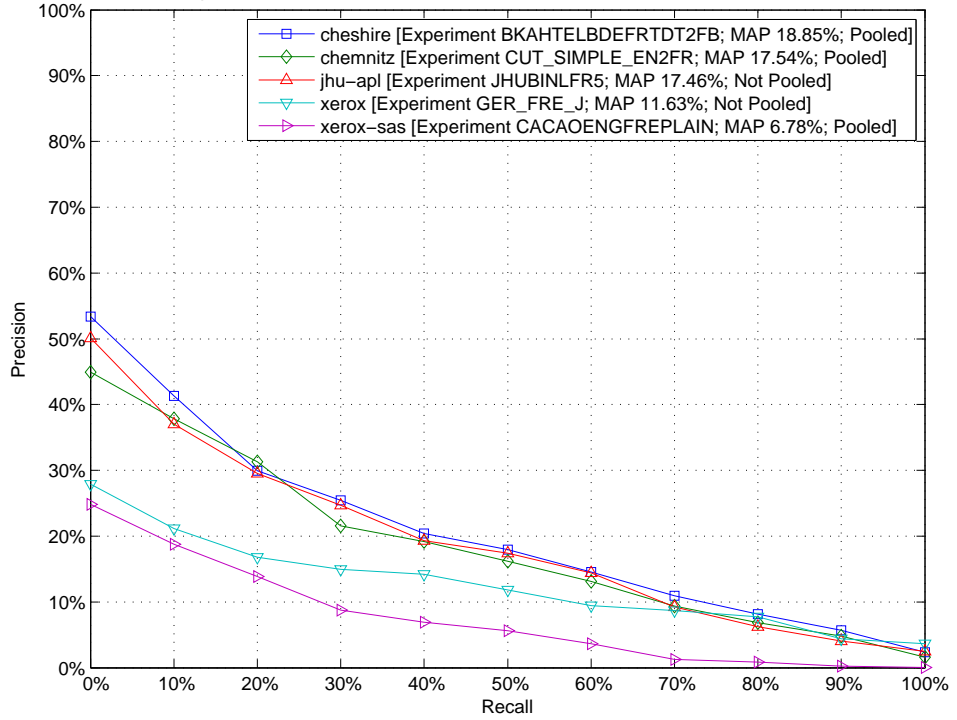


Fig. 10. Bilingual French

Ad-Hoc TEL Monolingual German Task Top 5 Participants – Standard Recall Levels vs Mean Interpolated Precision

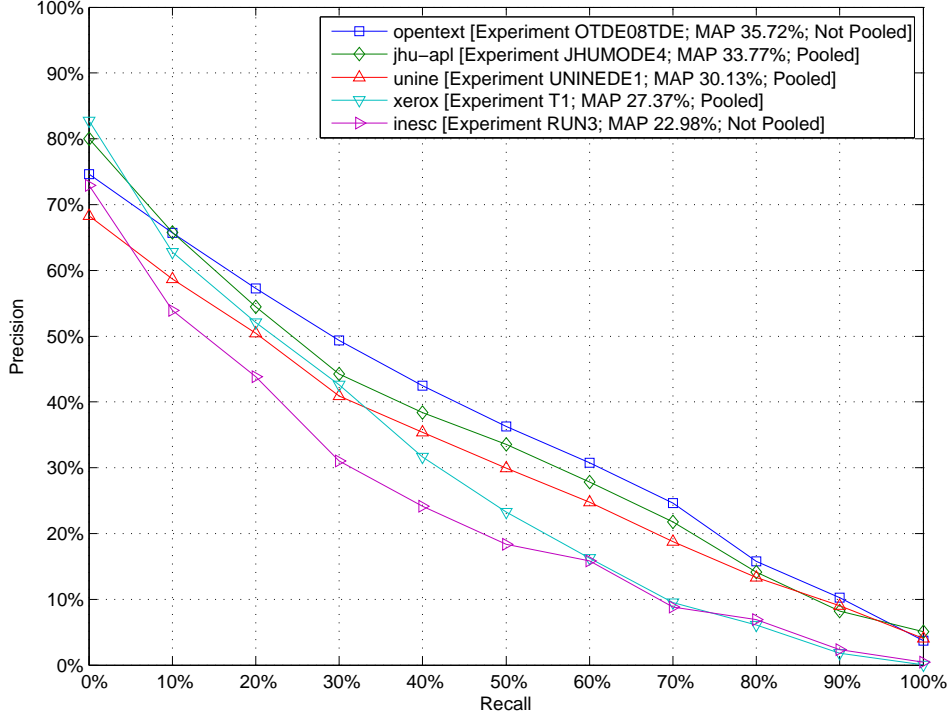


Fig. 11. Monolingual German

Ad-Hoc TEL Bilingual German Task Top 5 Participants – Standard Recall Levels vs Mean Interpolated Precision

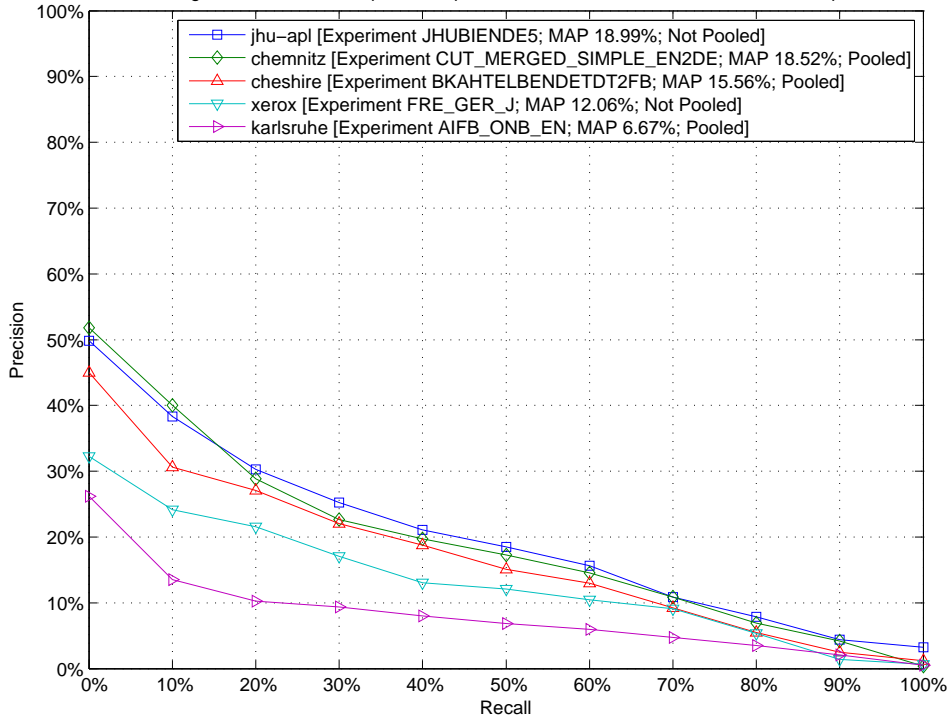


Fig. 12. Bilingual German

included language models, vector-space and probabilistic approaches, and translation resources ranged from bilingual dictionaries, parallel and comparable corpora, to on-line MT systems and Wikipedia. Groups often used a combination of more than one resource.

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

An example of this is the group from the Technical University of Chemnitz, which had overall the best results in the bilingual tasks (1st for XtoEN; 2nd for XtoFR and DE) although they did not do so well in the monolingual tasks. This group attempted to tackle the multilinguality of the collections in several ways. First, they tried to identify the language of each record in the collections using a language detector. Unfortunately, due to an error, they were unable to use the indices created in this way¹⁰. Second, in both their monolingual and cross-language experiments they implemented a retrieval algorithm which translated the query into the top 10 (in terms of occurrence) languages and merged these multilingual terms into a single query. They ran experiments weighting the query in different ways on the basis of estimated distribution of language content in the collections. In the monolingual experiments, rather disappointingly, the results showed that their purely monolingual baseline always outperformed experiments with query translations and language weights. This finding was confirmed with the bilingual experiments where again the better results were achieved with the baseline configurations. They attributed their good overall results for bilingual to the superiority of the Google online translation service [27].

Another group that attempted to tackle the multilinguality of the target collections was Xerox. This group built a single index containing all languages (according to the expected languages which they identified as just English, French and German although as stated the collections actually contain documents in other languages as well). This, of course, meant that the queries also had to be issued in all three languages. They built a multilingual probabilistic dictionary and for each target collection gave more weight to the official language of the collection [14]. Although their results for both monolingual and bilingual experiments for the French and German collections were always within the top five; they were not quite so successful with the English collection.

However, most groups appear to have ignored the multilinguality of the single collections in their experiments. Good examples of this are three veteran CLEF groups, UniNE which had, overall the best monolingual results, JHU which appeared in the top five for all bilingual tasks, and Berkeley which figured in the top five for all experiments except for monolingual German. UniNe appeared to

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

focus on testing different IR models and combination approaches whereas the major interest of JHU was on the most efficient methods for indexing. Berkeley tested a version the Logistic Regression (LR) algorithm that has been used very successfully in cross-language IR by Berkeley researchers for a number of years together with blind relevance feedback [20], [31], [28].

As was mentioned in Section 2.1, the TEL data is structured data; participants were told that they could use all fields. Some groups attempted to exploit this by weighting the contents of different fields differently. See, for example [29]

To sum up, from a preliminary scanning of the results of this task, it appears that the majority of groups took it as a traditional ad hoc retrieval task and applied traditional methods. However, it is far too early to confirm whether this is really the best approach to retrieval on library catalog cards. We expect that this issue will be discussed at the workshop.

4 Persian@CLEF

This activity was coordinated in collaboration with the Data Base Research Group (DBRG) of Tehran University. It was the first time that CLEF offered a non-European language target collection. Persian is an Indo-European language spoken in Iran, Afghanistan and Tajikistan. It is also known as Farsi. However, the Academy of Persian Language and Literature has declared in an official pronouncement that the name "Persian" is more appropriate, as it has the longer tradition in the western languages and better expresses the role of the language as a mark of cultural and national continuity.

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

4.1 Tasks

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

4.2 Participants

Eight groups submitted 66 runs for the Persian task: all eight submitted monolingual runs (53 runs out of 66); 3 groups also submitted bilingual runs (13 runs

Table 7. Best entries for the Persian tasks.

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

out of 66). Five of the groups were formed of Persian native speakers, mostly from the University of Tehran; they were all first time CLEF participants. Unfortunately, at the time of writing we just have reports from four of these groups. The other three groups were CLEF veterans with much experience in the CLEF ad hoc track.

Table 4(a) provides a breakdown of the number of participants and submitted runs by task.

4.3 Results

Table 7 shows the top five groups for each target collection, ordered by mean average precision. The table reports: the short name of the participating group; the mean average precision achieved by the experiment; the DOI of the experiment; and the performance difference between the first and the last participant.

Figures 13 and 14 compare the performances of the top participants of the Persian tasks.

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

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

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

4.4 Approaches

As was to be expected a common theme in a number of the papers was the most effective way to handle the Persian morphology. The group with the best results in the monolingual task tested three approaches; no stemming, a light stemmer developed in-house, and a 4-gram indexing approach. Their best results were

Ad-Hoc Persian Monolingual Persian Task Top 5 Participants – Standard Recall Levels vs Mean Interpolated Precision

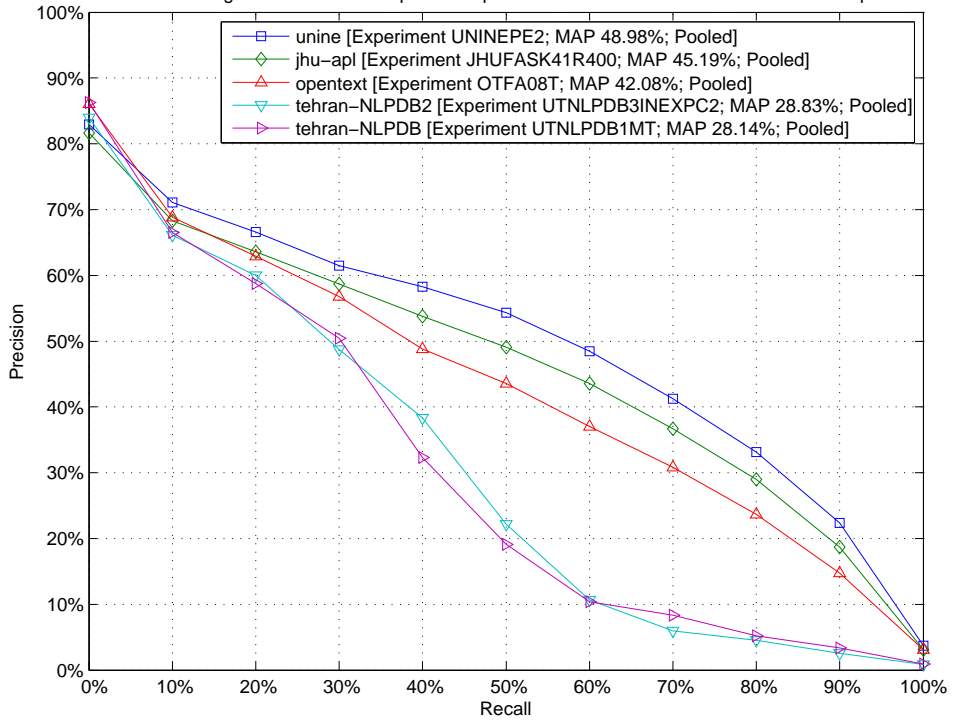


Fig. 13. Monolingual Persian

Ad-Hoc Persian Bilingual Persian Task Top 5 Participants – Standard Recall Levels vs Mean Interpolated Precision

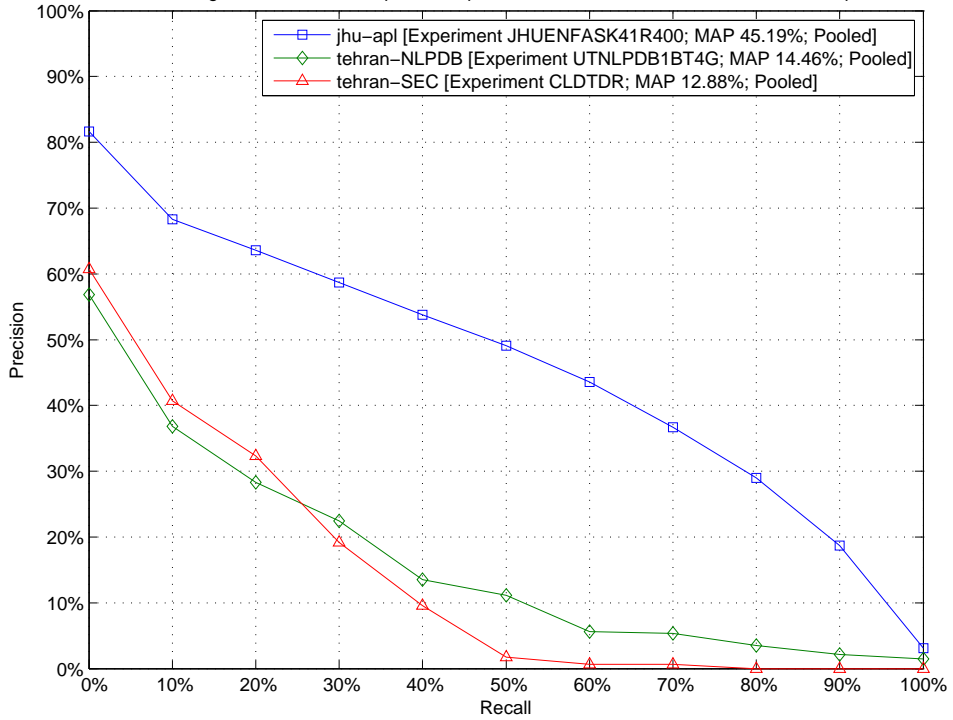


Fig. 14. Bilingual Persian

achieved using their light stemmer which has been made freely available on their website. However, they commented that the loss in performance with the no stemming approach was not very great. This group also tested three probabilistic models: Okapi, DFR and statistical language model (LM). The best results were obtained with the latter two [20]. The group with the second best results compared several different forms of textual normalization: character n-grams, n-gram stems, ordinary words, words automatically segmented into morphemes, and a novel form of n-gram indexing based on n-grams with character skips. They found that that character 5-grams and skipgrams performed the best [31]. The findings of [20] were confirmed by [40]. This group also tested runs with no stemming, with the UniNE stemmer and with n-grams. Similarly, they reported that stemming had relatively little impact.

Somewhat surprisingly, most of the papers from the Iran-based groups do not provide much information wrt morphological analysis or stemming in their papers. One mentions the application of a light Porter-like stemmer but reported that the algorithm adopted was too simple and results did not improve [7]. Only one of these groups provides some detailed discussion of the impact of stemming. This group used a simple stemmer (PERSTEM¹¹ and reported that in most cases stemming did improve performance but noted that this was in contrast with experiments conducted by other groups at the University of Tehran on on the same collection. They suggest that further experiments with different types of stemmers and stemming techniques are required in order to clarify the role of stemming in Persian text processing [26]. Two of the Persian groups also decided to annotate the corpus with part-of-speech tags in order to evaluate the impact of such information on the performance of the retrieval algorithms [24], [26]. The results reported do not appear to show any great boost in performance.

Other experiments by the groups from Iran included an investigation into the effect of fusion of different retrieval technique. Two approaches were tested: combining the results of nine distinct retrieval methods; combining the results of the same method but with different types of tokens. The second strategy applied a vector space model and ran it with three different types of tokens namely 4-grams, stemmed single terms and unstemmed single terms. This approach gave better results [1].

For the cross-language task, the English topics were translated into Persian. As remarked above, the task of the translators was not easy as it was both a cross-language and also a cross-cultural task. The best result - again by a CLEF veteran participant - obtained 92% of the best monolingual performance. This is well in line with state-of-the-art performance for good cross-language retrieval systems. This group used an online machine translation system applied to the queries¹² [31].

The other two submissions for the cross-language task were from Iran-based groups. We have received a report from just one of them. This group applied both query and document translation. For query translation they used a method

¹¹ <http://sourceforge.net/projects/perstem>

¹² <http://www.parstranlator.net/eng/translate.htm>

based on the estimation of translation probabilities. In the document translation part they used the Shiraz machine translation system to translate the documents into English. They then created a Hybrid CLIR system by score-based merging of the two retrieval system results. The best performance was obtained with the hybrid system, confirming the reports of other researchers in previous CLEF campaigns, and elsewhere.

5 Robust – WSD Experiments

The robust task ran for the third time at CLEF 2008. It is an ad-hoc retrieval task based on data of previous CLEF campaigns. The robust task emphasizes the difficult topics by a non-linear integration of the results of individual topics into one result for a system, using the geometric mean of the average precision for all topics (GMAP) as an additional evaluation measure [37,41]. Given the difficulty of the task, training data including topics and relevance assessments was provided for the participants to tune their systems to the collection.

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

The experiment involved both monolingual (topics and documents in English) and bilingual experiments (topics in Spanish and documents in English). In addition to the original documents and topics, the organizers of the task provided both documents and topics which had been automatically tagged with word senses from WordNet version 1.6 using two state-of-the-art word sense disambiguation systems, UBC [2] and NUS [12]. These systems provided weighted word sense tags for each of the nouns, verbs, adjectives and adverbs that they could disambiguate.

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

5.1 Participants

Eight groups submitted 63 runs for the Robust tasks: all groups submitted monolingual runs (45 runs out of 63); 4 groups also submitted bilingual runs (18 runs out of 63). Moreover, 7 groups participated in the WSD tasks, submitting 40 out

Table 8. Best entries for the robust monolingual task.

Track	Rank	Participant	Experiment DOI	MAP	GMAP
English	1st	unine	10.2415/AH-ROBUST-MONO-EN-TEST-CLEF2008.UNINE.UNINEROBUST4	45.14%	21.17%
	2nd	geneva	10.2415/AH-ROBUST-MONO-EN-TEST-CLEF2008.GENEVA.ISILEMTDN	39.17%	16.53%
	3rd	ucm	10.2415/AH-ROBUST-MONO-EN-TEST-CLEF2008.UCM.BM25_B01	38.34%	15.28%
	4th	ixa	10.2415/AH-ROBUST-MONO-EN-TEST-CLEF2008.IXA.EN2ENNOWSDPSREL	38.10%	15.72%
	5th	ufrgs	10.2415/AH-ROBUST-MONO-EN-TEST-CLEF2008.UFRGS.UFRGS_R_MONO2_TEST	33.94%	13.96%
	Difference			33.03%	51.64%
English WSD	1st	unine	10.2415/AH-ROBUST-WSD-MONO-EN-TEST-CLEF2008.UNINE.UNINEROBUST6	44.98%	21.54%
	2nd	ucm	10.2415/AH-ROBUST-WSD-MONO-EN-TEST-CLEF2008.UCM.BM25_B01_CLAUSES_09	39.57%	16.17%
	3rd	ixa	10.2415/AH-ROBUST-WSD-MONO-EN-TEST-CLEF2008.IXA.EN2ENUBCDCSPSREL	38.99%	15.52%
	4th	geneva	10.2415/AH-ROBUST-WSD-MONO-EN-TEST-CLEF2008.GENEVA.ISINUSLWTDN	38.13%	16.25%
	5th	ufrgs	10.2415/AH-ROBUST-WSD-MONO-EN-TEST-CLEF2008.UFRGS.UFRGS_R_MONO_WSD5_TEST	34.64%	14.17%
	Difference			29.84%	52.01%

of 63 runs, 30 monolingual and 10 bilingual. Table 4(a) provides a breakdown of the number of participants and submitted runs by task. Two further groups were late, so they are not included in the official results but they do have working notes papers.

5.2 Results

Monolingual Results

Table 8 shows the best results for this task. The performance difference between the best and the last (up to 5) placed group is given (in terms of average precision).

Figures 15 and 16 compare the performances of the top participants of the Robust Monolingual and Monolingual WSD.

Bilingual Results

Table 9 shows the best results for this task. The performance difference between the best and the last (up to 5) placed group is given (in terms of average precision). All the experiments were from English to French.

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

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

Figures 17 and 18 compares the performances of the top participants of the Robust Bilingual and Bilingual WSD tasks.

Ad-Hoc Robust Monolingual English Test Task Top 5 Participants – Standard Recall Levels vs Mean Interpolated Precision

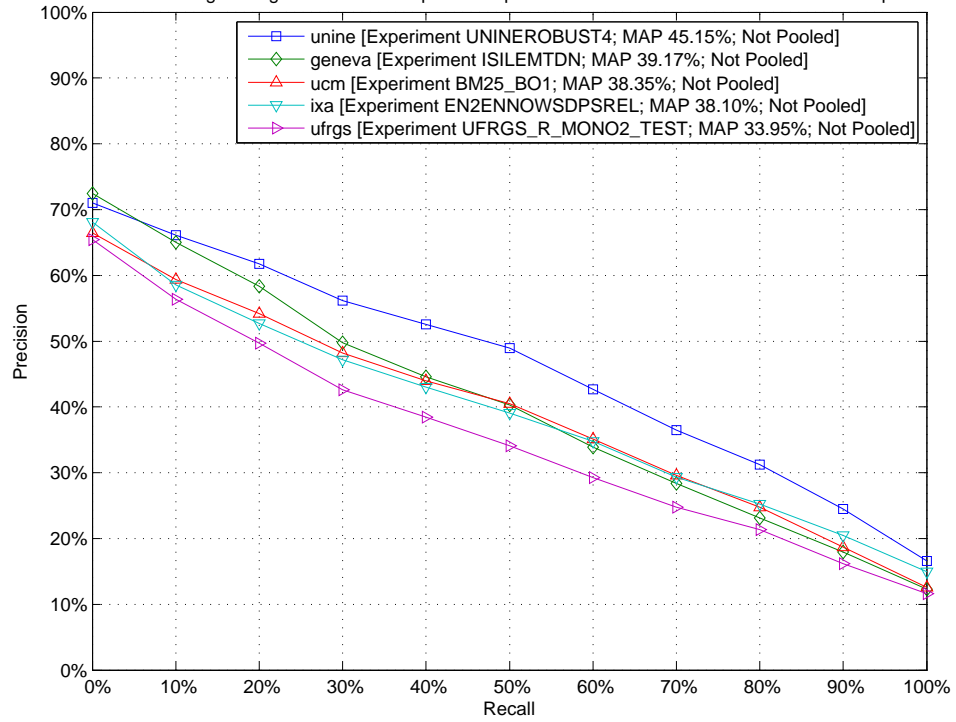


Fig. 15. Robust Monolingual English.

Robust Word Sense Disambiguation Monolingual English Test Task Top 5 Participants – Standard Recall Levels vs Mean Interpolated Precision

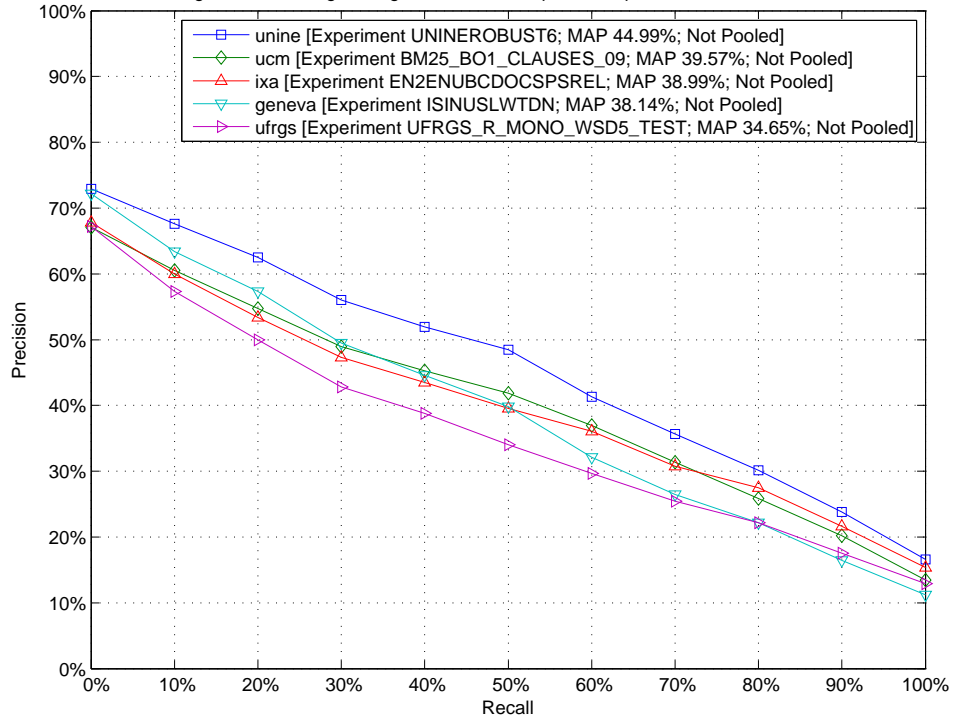


Fig. 16. Robust Monolingual English WSD.

Ad-Hoc Robust Bilingual English Test Task Top 5 Participants – Standard Recall Levels vs Mean Interpolated Precision

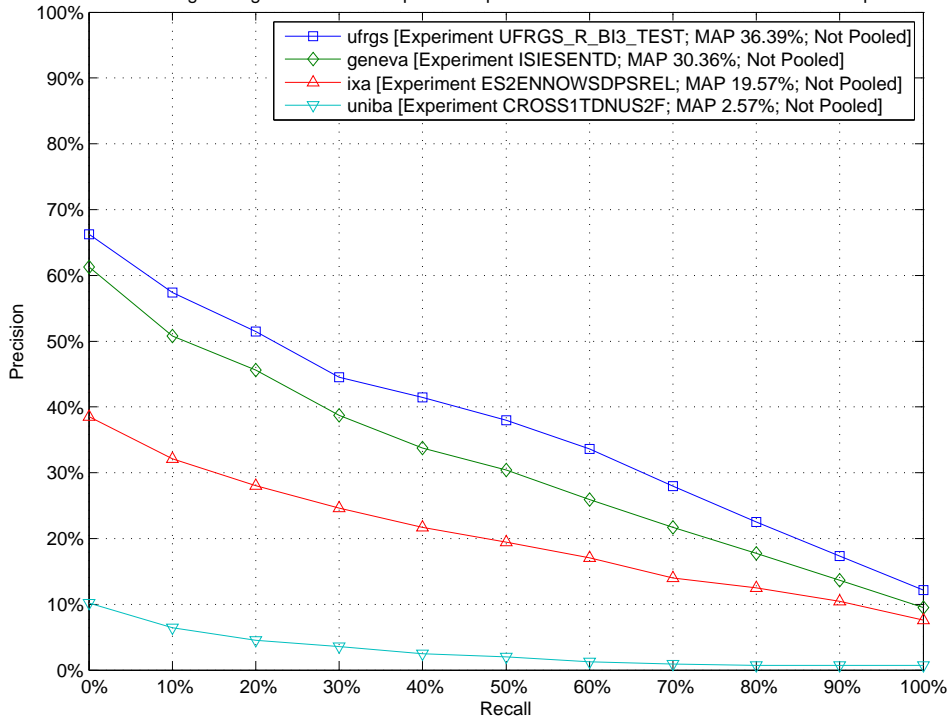


Fig. 17. Robust Bilingual English.

ic Robust Word Sense Disambiguation Bilingual English Test Task Top 5 Participants – Standard Recall Levels vs Mean Interpolated Precision

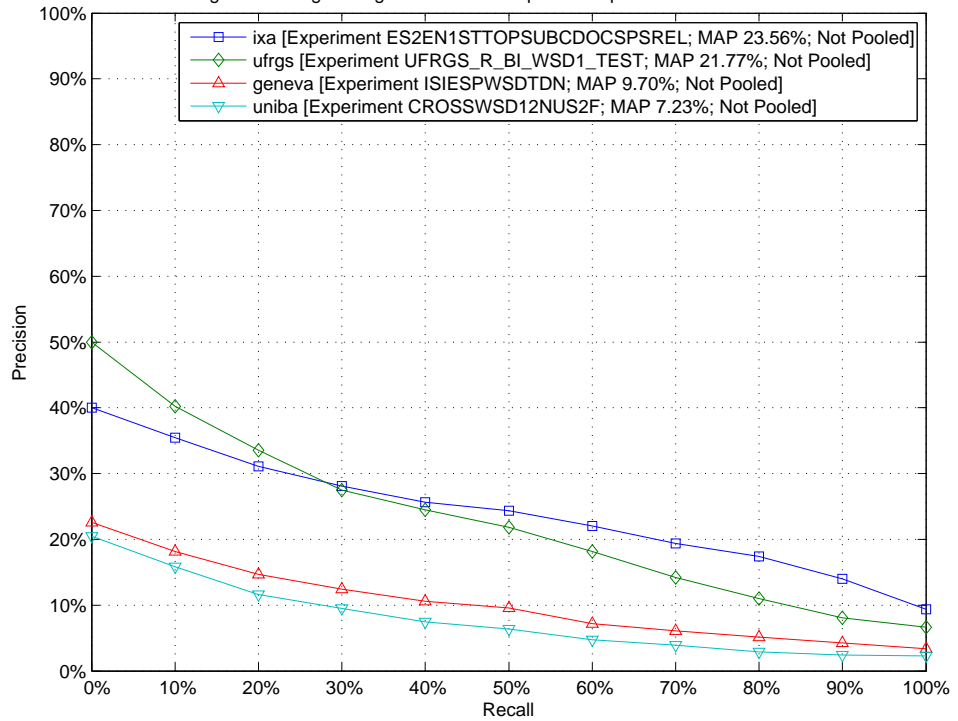


Fig. 18. Robust Bilingual English WSD.

Table 9. Best entries for the robust bilingual task.

Track	Rank	Participant	Experiment DOI	MAP	GMAP
English	1st	ufrgs	10.2415/AH-ROBUST-BILI-X2EN-TEST-CLEF2008.UFRGS.UFRGS_R_BI3_TEST	36.38%	13.00%
	2nd	geneva	10.2415/AH-ROBUST-BILI-X2EN-TEST-CLEF2008.GENEVA.ISIESENTD	30.36%	10.96%
	3rd	ixa	10.2415/AH-ROBUST-BILI-X2EN-TEST-CLEF2008.IXA.ES2ENNOWSDPSREL	19.57%	1.62%
	4th	uniba	10.2415/AH-ROBUST-BILI-X2EN-TEST-CLEF2008.UNIBA.CROSS1TDNUS2F	2.56%	0.04%
	5th	–	–	–	–
	Difference				1,321.09%
English WSD	1st	ixa	10.2415/AH-ROBUST-WSD-BILI-X2EN-TEST-CLEF2008.IXA.ES2EN1STTOPSUBCDCCSPSREL	23.56%	1.71%
	2nd	ufrgs	10.2415/AH-ROBUST-WSD-BILI-X2EN-TEST-CLEF2008.UFRGS.UFRGS_R_BI_WSD1_TEST	21.77%	5.14%
	3rd	geneva	10.2415/AH-ROBUST-WSD-BILI-X2EN-TEST-CLEF2008.GENEVA.ISIESPWSDTDN	9.70%	0.37%
	4th	geneva	10.2415/AH-ROBUST-WSD-BILI-X2EN-TEST-CLEF2008.UNIBA.CROSSWSD12NUS2F	7.23%	0.16%
	5th	–	–	–	–
	Difference				225.86%

Analysis

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

In the monolingual experiments, cf. Table 8, the best results overall in both MAP and GMAP were for UNINE. Their WSD runs scored very similar to the non-WSD runs, with a slight decrease of MAP (0.16 percentage points) and a slight increase of GMAP (0.27 percentage points) [20]. The second best scoring team in MAP was UCM, which did attain MAP and GMAP improvements using WSD (from 38.34 MAP – 15.28 GMAP in their best non-WSD run to 39.57 MAP – 16.17 GMAP in their best WSD run) [36]. The third best scoring team in MAP was GENEVA who achieved lower scores on both MAP and GMAP using WSD information [22]. The fourth best team, IXA, obtained better MAP results using WSD information (from 38.10 to 38.99 MAP), but lower GMAP (from 15.72 to 15.52) [34]. Regarding the rest of participants, while UFRGS and UNIBA obtained improvements, KNOW-CENTER did not, and INAOE only submitted non-WSD runs. There were two additional groups (IRn and SINAI) that sent their results late. Both groups had their best scores for non-WSD systems. Please check the respective working notes reports for specific results.

In the bilingual experiments, cf. Table 9, the best results overall in both MAP and GMAP were for UFRGS, with a system which did not use WSD annotations (36.39, compared to 21.77 MAP for their best run using WSD) [16]. The second scoring team, GENEVA also failed to profit from WSD annotations (30.36 compared to 9.70 MAP) [22]. The other two participating groups did

obtain improvements, with IXA attaining 23.56 MAP with WSD (compared to 19.57 without) and UNIBA attaining 7.23 MAP) [34], [8].

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

6 Statistical Testing

When the goal is to validate how well results can be expected to hold beyond a particular set of queries, statistical testing can help to determine what differences between runs appear to be real as opposed to differences that are due to sampling issues. We aim to identify runs with results that are significantly different from the results of other runs. “Significantly different” in this context means that the difference between the performance scores for the runs in question appears greater than what might be expected by pure chance. As with all statistical testing, conclusions will be qualified by an error probability, which was chosen to be 0.05 in the following. We have designed our analysis to follow closely the methodology used by similar analyses carried out for *Text REtrieval Conference (TREC)* [23].

We used the MATLAB Statistics Toolbox, which provides the necessary functionality plus some additional functions and utilities. We use the *ANalysis Of VAriance (ANOVA)* test. ANOVA makes some assumptions concerning the data to be checked. Hull [23] provides details of these; in particular, the scores in question should be approximately normally distributed and their variance has to be approximately the same for all runs. Two tests for goodness of fit to a normal distribution were chosen using the MATLAB statistical toolbox: the Lilliefors test [15] and the Jarque-Bera test [25]. In the case of the CLEF tasks under analysis, both tests indicate that the assumption of normality is violated for most of the data samples (in this case the runs for each participant).

In such cases, a transformation of data should be performed. The transformation for measures that range from 0 to 1 is the arcsin-root transformation:

$$\arcsin(\sqrt{x})$$

which Tague-Sutcliffe [39] recommends for use with precision/recall measures.

Table 10 shows the results of both the Lilliefors and Jarque-Bera tests before and after applying the Tague-Sutcliffe transformation. After the transformation the analysis of the normality of samples distribution improves significantly, with some exceptions. The difficulty to transform the data into normally distributed samples derives from the original distribution of run performances which tend towards zero within the interval [0,1].

Table 10. Lilliefors (LF) and Jarque-Bera (JB) test for each Ad-Hoc track with and without Tague-Sutcliffe (TS) arcsin transformation. Each entry is the number of experiments whose performance distribution can be considered drawn from a Gaussian distribution, with respect to the total number of experiment of the track. The value of alpha for this test was set to 5%.

Track	LF	LF & TS	JB	JB & TS
TEL@CLEF Monolingual English	10	24	16	32
TEL@CLEF Monolingual French	0	23	2	24
TEL@CLEF Monolingual German	4	21	11	24
TEL@CLEF Bilingual English	0	3	2	9
TEL@CLEF Bilingual French	0	5	0	6
TEL@CLEF Bilingual German	0	1	0	1
PERSIAN@CLEF Monolingual Persian	39	43	29	43
PERSIAN@CLEF Bilingual Persian	4	6	4	8
Robust Monolingual English	1	9	0	2
Robust WSD Monolingual English	1	12	0	7
Robust Bilingual English	0	0	0	0
Robust WSD Bilingual English	0	0	0	0

In the following sections, two different graphs are presented to summarize the results of this test. All experiments, regardless of topic language or topic fields, are included. Results are therefore only valid for comparison of individual pairs of runs, and not in terms of absolute performance. Both for the ad-hoc and robust tasks, only runs where significant differences exist are shown; the remainder of the graphs can be found in the Appendices [17,18,19].

The first graph shows participants' runs (y axis) and performance obtained (x axis). The circle indicates the average performance (in terms of precision) while the segment shows the interval in which the difference in performance is not statistically significant.

The second graph shows the overall results where all the runs that are included in the same group do not have a significantly different performance. All runs scoring below a certain group perform significantly worse than at least the top entry of the group. Likewise all the runs scoring above a certain group perform significantly better than at least the bottom entry in that group. To determine all runs that perform significantly worse than a certain run, determine the rightmost group that includes the run, all runs scoring below the bottom entry of that group are significantly worse. Conversely, to determine all runs that perform significantly better than a given run, determine the leftmost group that includes the run. All runs that score better than the top entry of that group perform significantly better.

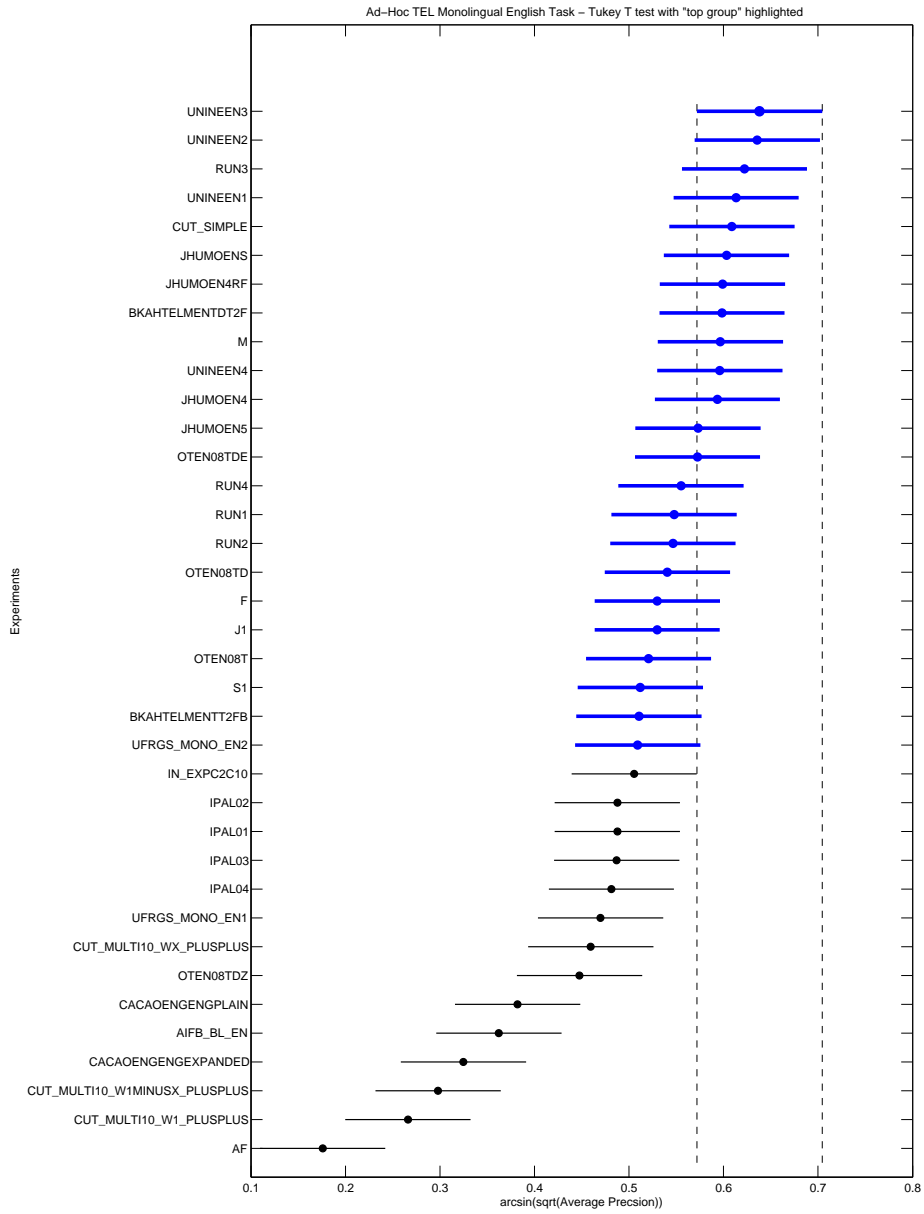
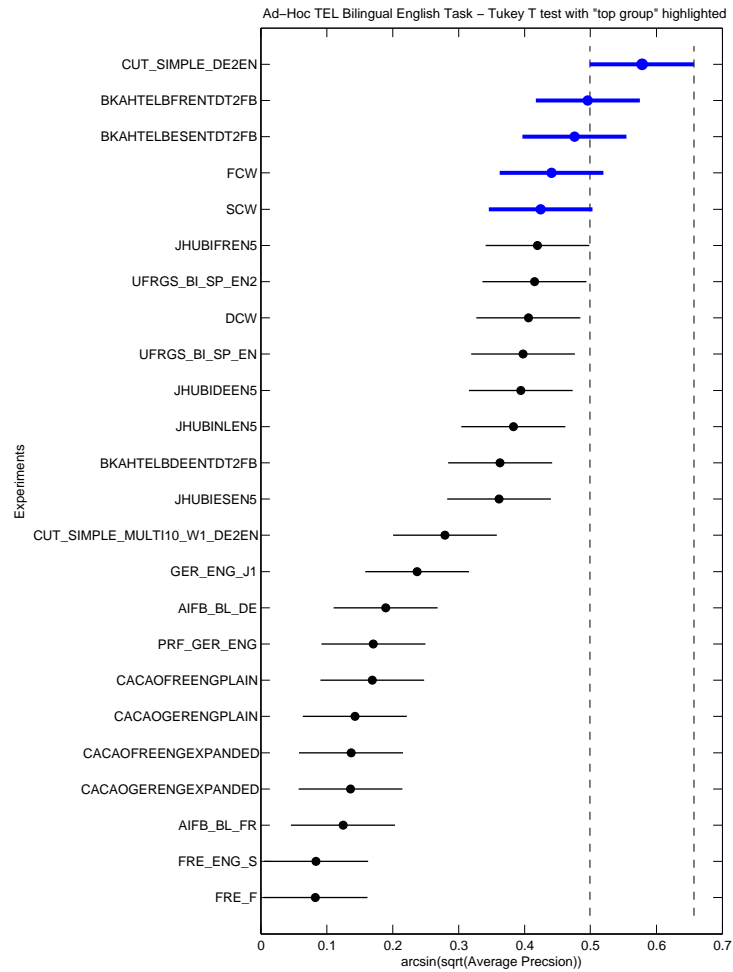


Fig. 19. Ad-Hoc TEL Monolingual English. The figure shows the Tukey T Test.



Experiment DOI	Groups				
10.2415/AH-TEL-BILI-X2EN-CLEF2008.CHEMNITZ.CUT_SIMPLE_DE2EN	X				
10.2415/AH-TEL-BILI-X2EN-CLEF2008.CHESHIRE.BKAHTELBFRENTDT2FB	X	X			
10.2415/AH-TEL-BILI-X2EN-CLEF2008.CHESHIRE.BKAHTELBESENTDT2FB	X	X			
10.2415/AH-TEL-BILI-X2EN-CLEF2008.TWENTE.FCW	X	X			
10.2415/AH-TEL-BILI-X2EN-CLEF2008.TWENTE.SCW	X	X	X		
10.2415/AH-TEL-BILI-X2EN-CLEF2008.JHU-APL.JHUBIFREN5	X	X			
10.2415/AH-TEL-BILI-X2EN-CLEF2008.UFRGS.UFRGS_BI_SP_EN2	X	X			
10.2415/AH-TEL-BILI-X2EN-CLEF2008.TWENTE.DCW	X	X			
10.2415/AH-TEL-BILI-X2EN-CLEF2008.UFRGS.UFRGS_BI_SP_EN	X	X			
10.2415/AH-TEL-BILI-X2EN-CLEF2008.JHU-APL.JHUBIDEEN5	X	X	X		
10.2415/AH-TEL-BILI-X2EN-CLEF2008.JHU-APL.JHUBINLEN5	X	X	X		
10.2415/AH-TEL-BILI-X2EN-CLEF2008.CHESHIRE.BKAHTELBDEENTDT2FB	X	X	X		
10.2415/AH-TEL-BILI-X2EN-CLEF2008.JHU-APL.JHUBIESEN5	X	X	X		
10.2415/AH-TEL-BILI-X2EN-CLEF2008.CHEMNITZ.CUT_SIMPLE_MULTI10_W1_DE2EN	X	X	X		
10.2415/AH-TEL-BILI-X2EN-CLEF2008.XEROX.GER_ENG_J1		X	X	X	
10.2415/AH-TEL-BILI-X2EN-CLEF2008.KARLSRUHE.AIFB_BL_DE			X	X	
10.2415/AH-TEL-BILI-X2EN-CLEF2008.XEROX.PRF_GER_ENG			X	X	
10.2415/AH-TEL-BILI-X2EN-CLEF2008.XEROX-SAS.CACAOFREENGPLAIN				X	X
10.2415/AH-TEL-BILI-X2EN-CLEF2008.XEROX-SAS.CACAOGERENGPLAIN				X	X
10.2415/AH-TEL-BILI-X2EN-CLEF2008.XEROX-SAS.CACAOFREENGEXPANDED				X	X
10.2415/AH-TEL-BILI-X2EN-CLEF2008.XEROX-SAS.CACAOGERENGEXPANDED				X	X
10.2415/AH-TEL-BILI-X2EN-CLEF2008.KARLSRUHE.AIFB_BL_FR				X	X
10.2415/AH-TEL-BILI-X2EN-CLEF2008.XEROX.FRE_ENG_S				X	X
10.2415/AH-TEL-BILI-X2EN-CLEF2008.XEROX.FRE_F				X	X

Fig. 20. Ad-Hoc TEL Bilingual English. Experiments grouped according to the Tukey T Test.

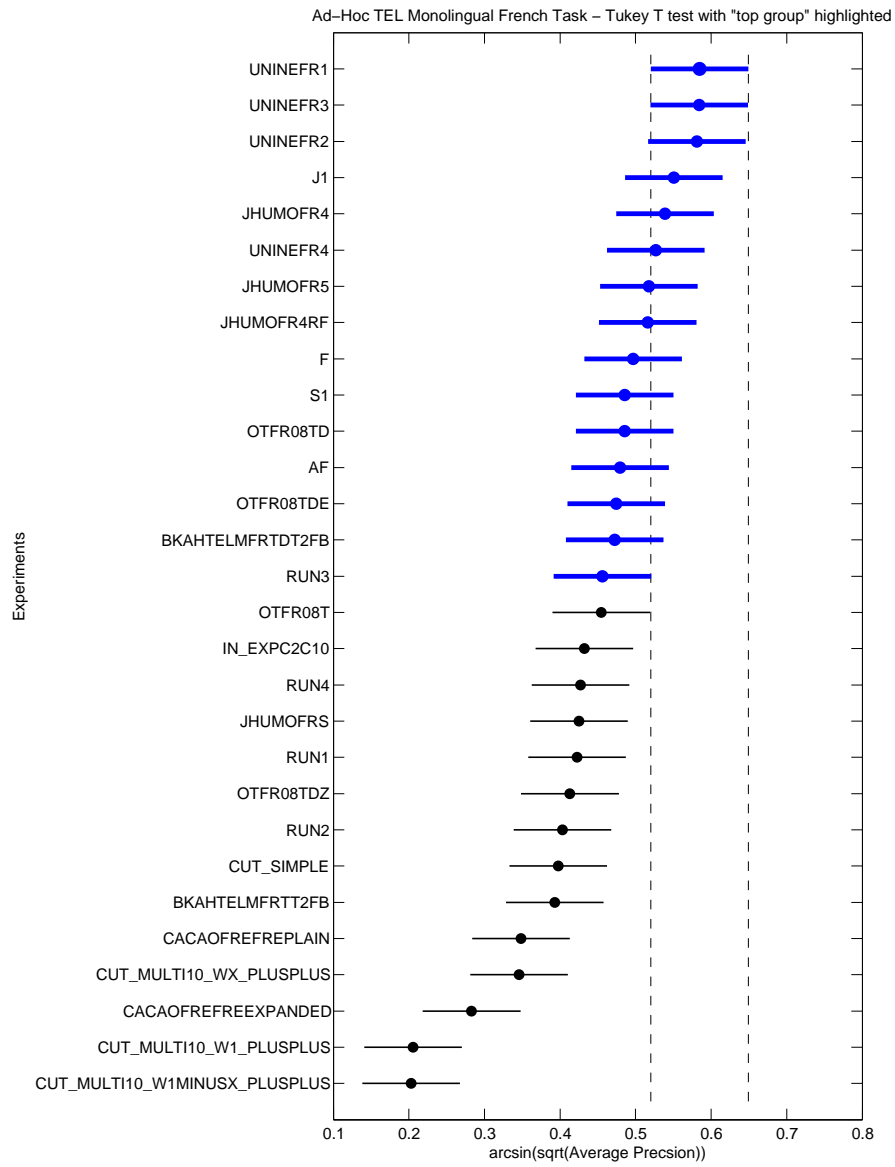
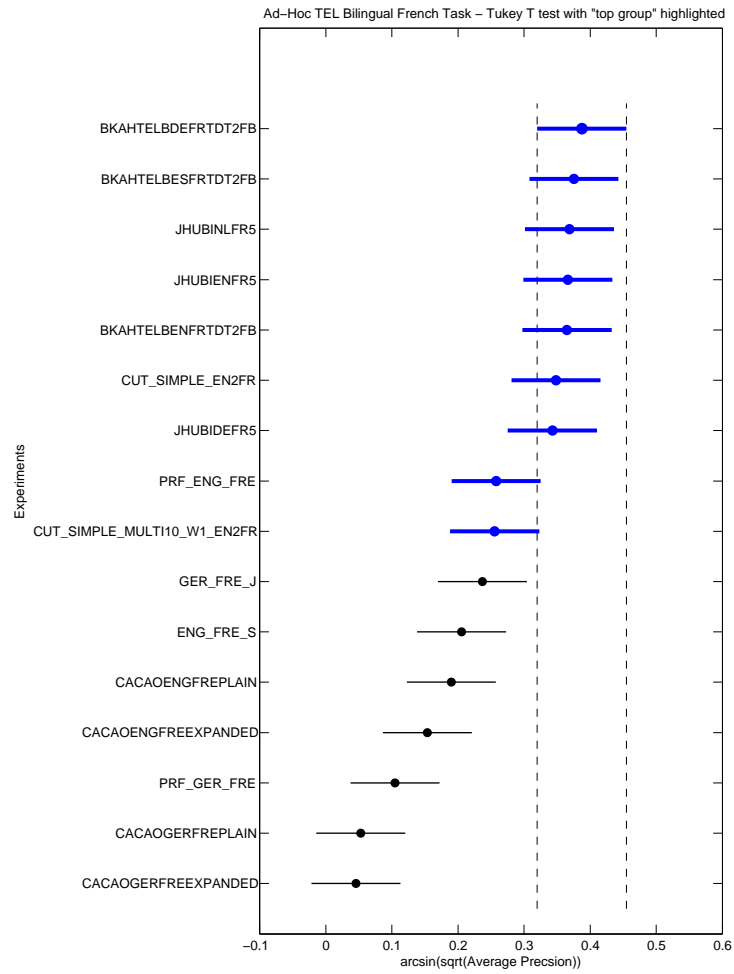


Fig. 21. Ad-Hoc TEL Monolingual French. The figure shows the Tukey T Test.



Experiment DOI	Groups			
10.2415/AH-TEL-BILI-X2FR-CLEF2008.CHESHIRE.BKAHTELBDEFRTDT2FB	X			
10.2415/AH-TEL-BILI-X2FR-CLEF2008.CHESHIRE.BKAHTELBESFRTDT2FB	X	X		
10.2415/AH-TEL-BILI-X2FR-CLEF2008.JHU-APL.JHUBINLFR5	X	X		
10.2415/AH-TEL-BILI-X2FR-CLEF2008.JHU-APL.JHUBIENFR5	X	X		
10.2415/AH-TEL-BILI-X2FR-CLEF2008.CHESHIRE.BKAHTELBENFRTDT2FB	X	X		
10.2415/AH-TEL-BILI-X2FR-CLEF2008.CHEMNITZ.CUT_SIMPLE_EN2FR	X	X		
10.2415/AH-TEL-BILI-X2FR-CLEF2008.JHU-APL.JHUBIDFR5	X	X		
10.2415/AH-TEL-BILI-X2FR-CLEF2008.XEROX.PRF_ENG_FRE	X	X	X	
10.2415/AH-TEL-BILI-X2FR-CLEF2008.CHEMNITZ.CUT_SIMPLE_MULTI10_W1_EN2FR	X	X	X	
10.2415/AH-TEL-BILI-X2FR-CLEF2008.XEROX.GER_FRE_J	X	X	X	
10.2415/AH-TEL-BILI-X2FR-CLEF2008.XEROX.ENG_FRE_S	X	X	X	
10.2415/AH-TEL-BILI-X2FR-CLEF2008.XEROX-SAS.CACAOENGFREPLAIN	X	X	X	
10.2415/AH-TEL-BILI-X2FR-CLEF2008.XEROX-SAS.CACAOENGFREEXPANDED	X	X	X	X
10.2415/AH-TEL-BILI-X2FR-CLEF2008.XEROX.PRF_GER_FRE	X	X	X	X
10.2415/AH-TEL-BILI-X2FR-CLEF2008.XEROX-SAS.CACAOGERFREPLAIN	X	X	X	X
10.2415/AH-TEL-BILI-X2FR-CLEF2008.XEROX-SAS.CACAOGERFREEXPANDED	X	X	X	X

Fig. 22. Ad-Hoc TEL Bilingual French. Experiments grouped according to the Tukey T Test.

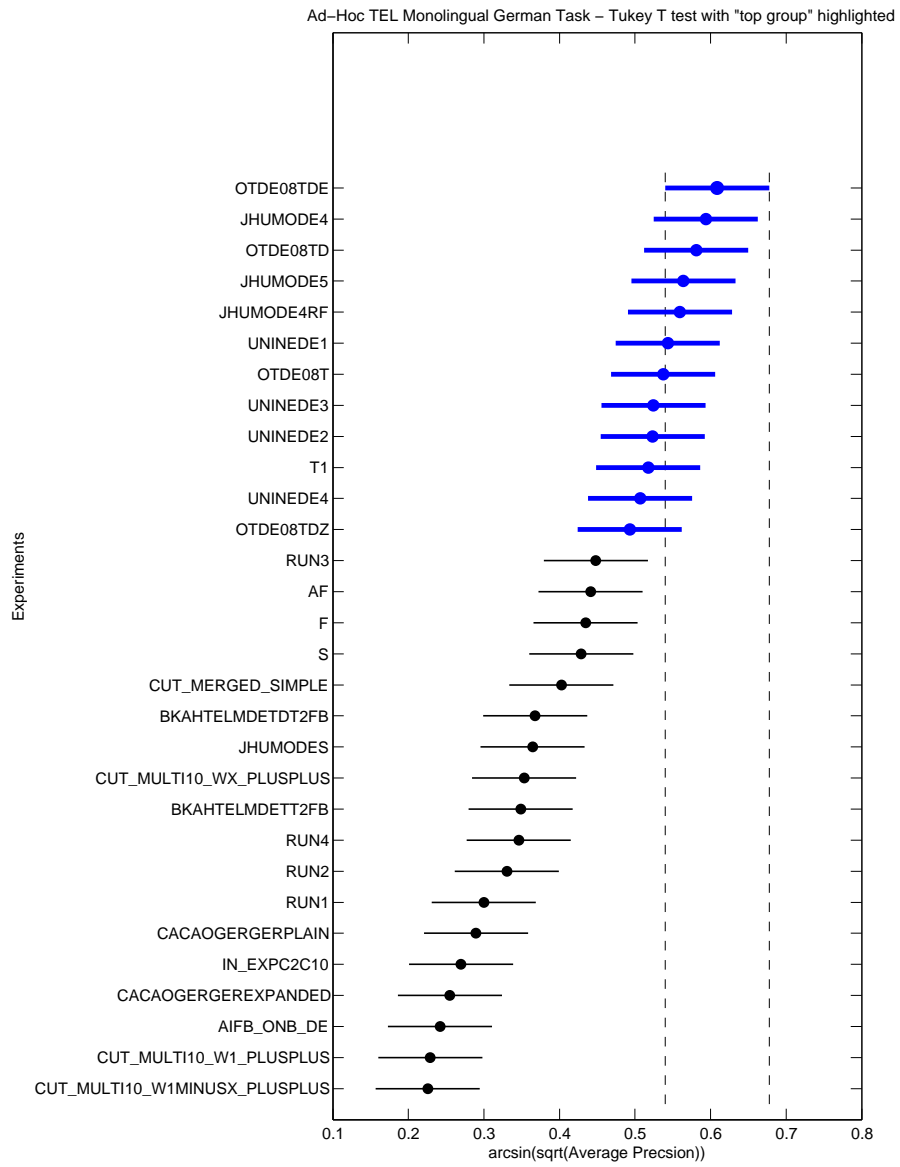
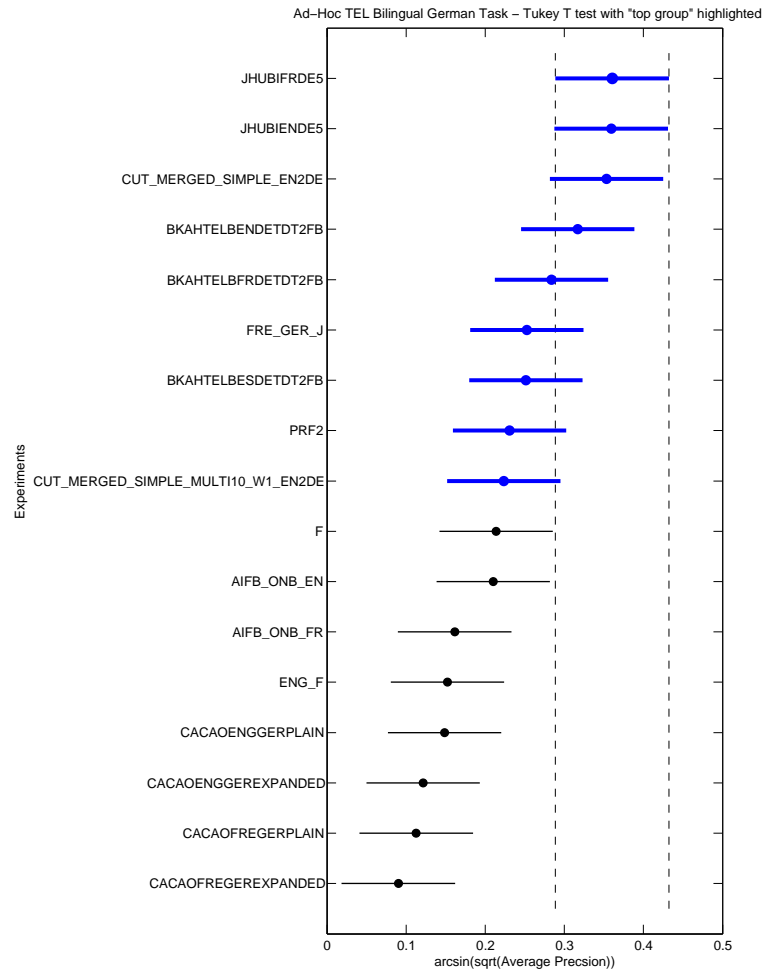


Fig. 23. Ad-Hoc TEL Monolingual German. The figure shows the Tukey T Test.



Experiment DOI	Groups			
10.2415/AH-TEL-BILI-X2DE-CLEF2008.JHU-APL.JHUBIFRDE5	X			
10.2415/AH-TEL-BILI-X2DE-CLEF2008.JHU-APL.JHUBIENDE5	X			
10.2415/AH-TEL-BILI-X2DE-CLEF2008.CHEMNITZ.CUT_MERGED_SIMPLE_EN2DE	X	X		
10.2415/AH-TEL-BILI-X2DE-CLEF2008.CHESHIRE.BKAHTELBENDETD2FB	X	X		
10.2415/AH-TEL-BILI-X2DE-CLEF2008.CHESHIRE.BKAHTELBFRDETD2FB	X	X	X	
10.2415/AH-TEL-BILI-X2DE-CLEF2008.XEROX.FRE_GER_J	X	X	X	X
10.2415/AH-TEL-BILI-X2DE-CLEF2008.CHESHIRE.BKAHTELBESDETD2FB	X	X	X	X
10.2415/AH-TEL-BILI-X2DE-CLEF2008.XEROX.PRF2	X	X	X	X
10.2415/AH-TEL-BILI-X2DE-CLEF2008.CHEMNITZ.CUT_MERGED_SIMPLE_MULT110_W1_EN2DE	X	X	X	X
10.2415/AH-TEL-BILI-X2DE-CLEF2008.XEROX.F	X	X	X	X
10.2415/AH-TEL-BILI-X2DE-CLEF2008.KARLSRUHE.AIFB_ONB_EN	X	X	X	X
10.2415/AH-TEL-BILI-X2DE-CLEF2008.KARLSRUHE.AIFB_ONB_FR	X	X	X	X
10.2415/AH-TEL-BILI-X2DE-CLEF2008.XEROX.ENG_F	X	X	X	X
10.2415/AH-TEL-BILI-X2DE-CLEF2008.XEROX-SAS.CACAOENGGERPLAIN	X	X	X	X
10.2415/AH-TEL-BILI-X2DE-CLEF2008.XEROX-SAS.CACAOENGGEREXPANDED			X	X
10.2415/AH-TEL-BILI-X2DE-CLEF2008.XEROX-SAS.CACAOFREGERPLAIN			X	X
10.2415/AH-TEL-BILI-X2DE-CLEF2008.XEROX-SAS.CACAOFREGEREXPANDED				X

Fig. 24. Ad-Hoc TEL Bilingual German. Experiments grouped according to the Tukey T Test.

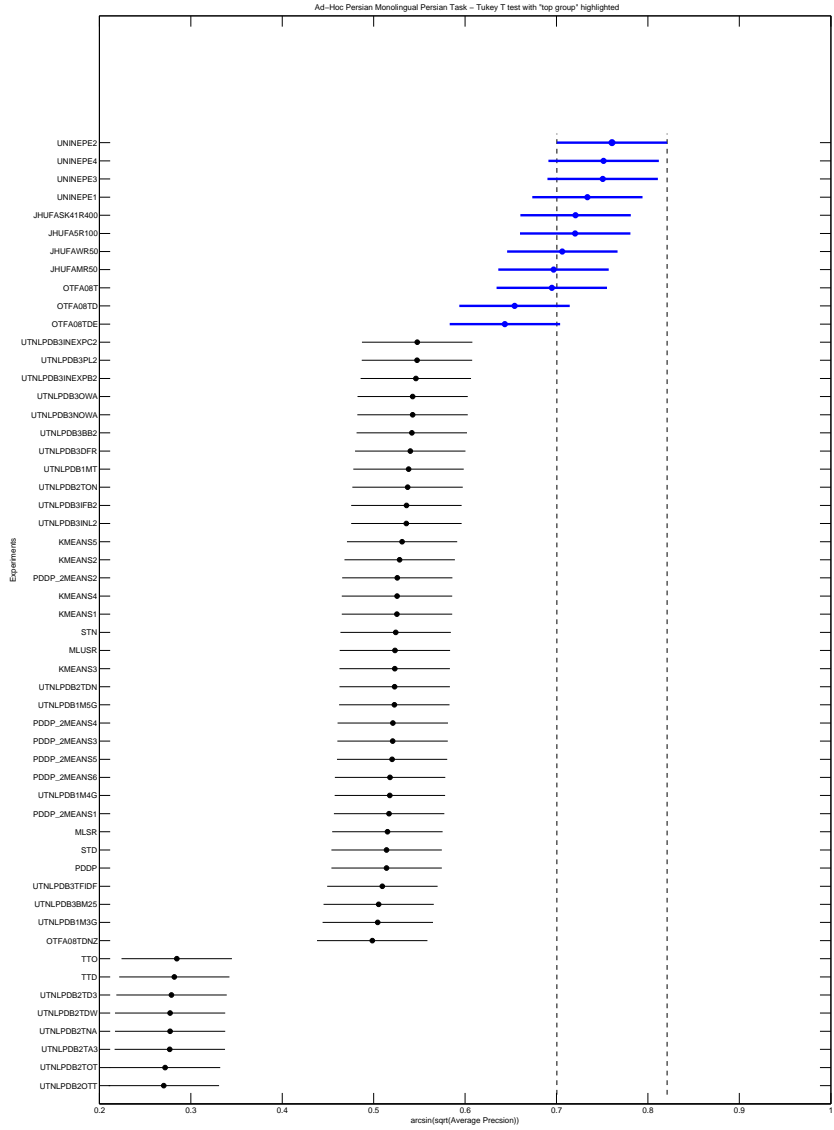
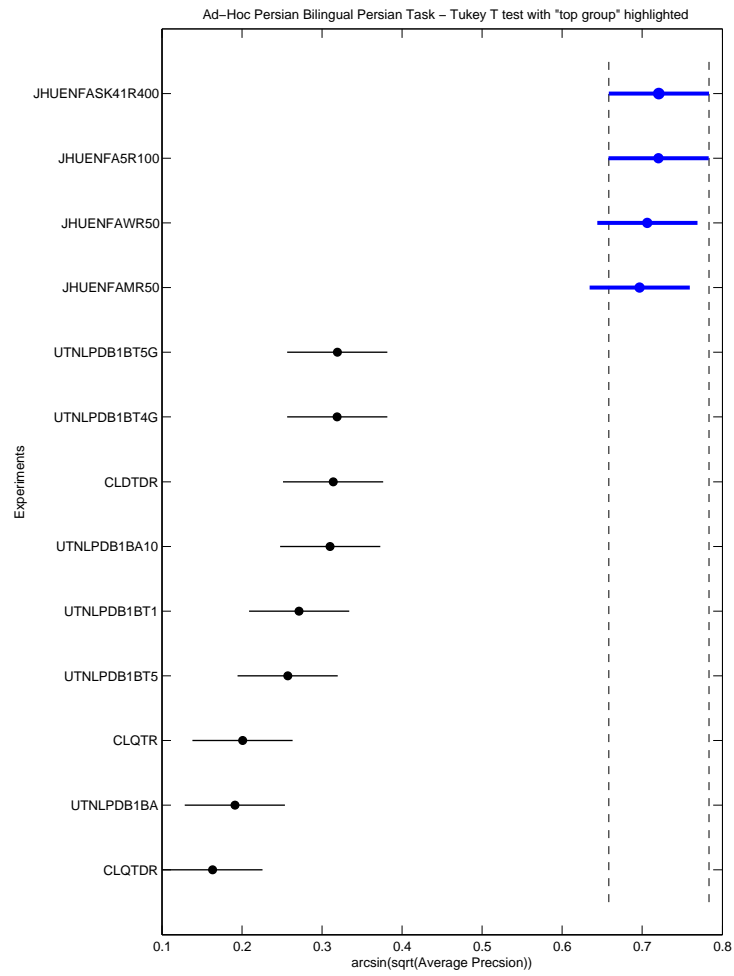


Fig. 25. Ad-Hoc Monolingual Persian. The figure shows the Tukey T Test.

Table 14. Ad-Hoc Monolingual Persian. The table shows the first ten groups of the Tukey T Test.

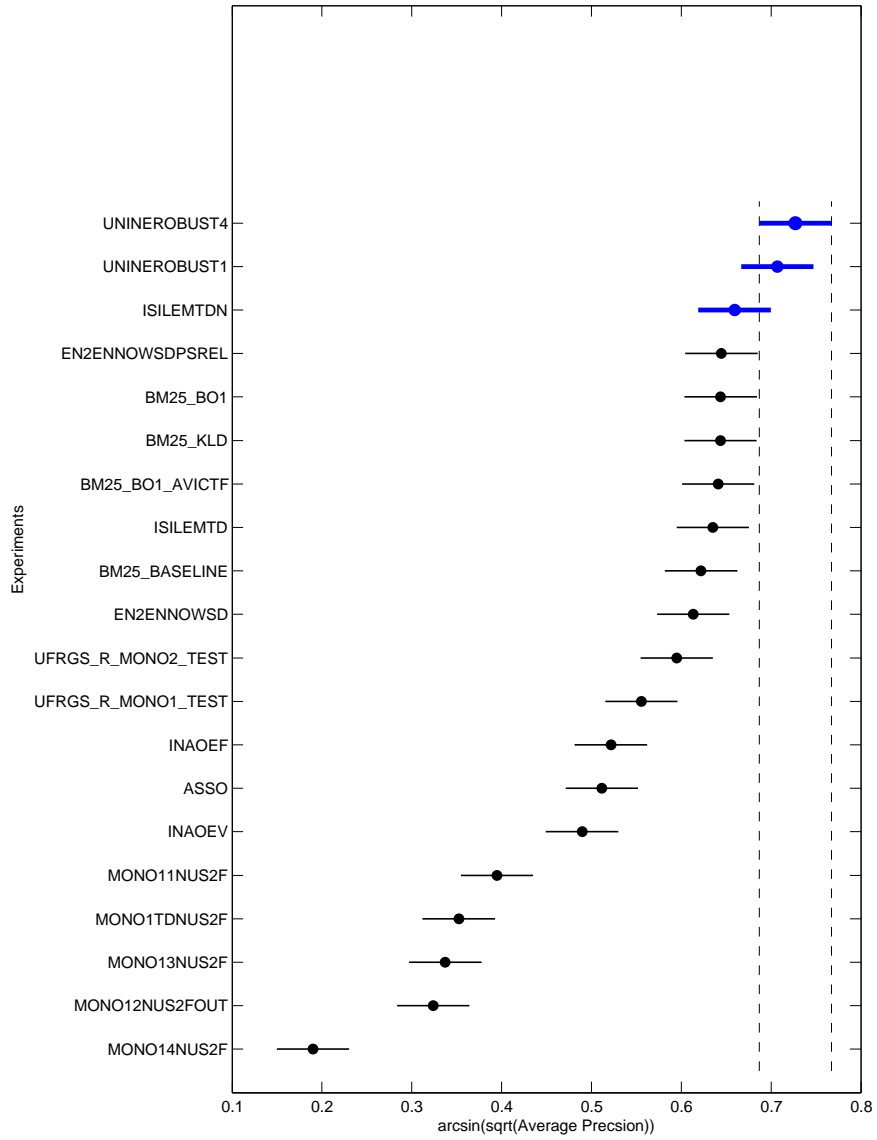
Experiment DOI	Groups			
10.2415/AH-PERSIAN-MONO-FA-CLEF2008.UNINE.UNINEPE2	X			
10.2415/AH-PERSIAN-MONO-FA-CLEF2008.UNINE.UNINEPE4	X			
10.2415/AH-PERSIAN-MONO-FA-CLEF2008.UNINE.UNINEPE3	X			
10.2415/AH-PERSIAN-MONO-FA-CLEF2008.UNINE.UNINEPE1	X			
10.2415/AH-PERSIAN-MONO-FA-CLEF2008.JHU-APL.JHUFASK41R400	X			
10.2415/AH-PERSIAN-MONO-FA-CLEF2008.JHU-APL.JHUFASR100	X			
10.2415/AH-PERSIAN-MONO-FA-CLEF2008.JHU-APL.JHUFAR50	X			
10.2415/AH-PERSIAN-MONO-FA-CLEF2008.JHU-APL.JHUFAMR50	X			
10.2415/AH-PERSIAN-MONO-FA-CLEF2008.OPENTEXT.OTFA08T	X			
10.2415/AH-PERSIAN-MONO-FA-CLEF2008.OPENTEXT.OTFA08TD	X	X		
10.2415/AH-PERSIAN-MONO-FA-CLEF2008.OPENTEXT.OTFA08TDE	X	X	X	
10.2415/AH-PERSIAN-MONO-FA-CLEF2008.TEHRAN-NLPDB2.UTNLPDB3INEXPC2	X	X	X	
10.2415/AH-PERSIAN-MONO-FA-CLEF2008.TEHRAN-NLPDB2.UTNLPDB3PL2	X	X	X	
10.2415/AH-PERSIAN-MONO-FA-CLEF2008.TEHRAN-NLPDB2.UTNLPDB3INEXPB2	X	X	X	
10.2415/AH-PERSIAN-MONO-FA-CLEF2008.TEHRAN-NLPDB2.UTNLPDB3OWA	X	X	X	
10.2415/AH-PERSIAN-MONO-FA-CLEF2008.TEHRAN-NLPDB2.UTNLPDB3NOWA	X	X	X	
10.2415/AH-PERSIAN-MONO-FA-CLEF2008.TEHRAN-NLPDB2.UTNLPDB3BB2	X	X	X	
10.2415/AH-PERSIAN-MONO-FA-CLEF2008.TEHRAN-NLPDB2.UTNLPDB3DFR	X	X	X	
10.2415/AH-PERSIAN-MONO-FA-CLEF2008.TEHRAN-NLPDB.UTNLPDB1MT	X	X	X	
10.2415/AH-PERSIAN-MONO-FA-CLEF2008.TEHRAN-NLP.UTNLPDB2TON	X	X	X	
10.2415/AH-PERSIAN-MONO-FA-CLEF2008.TEHRAN-NLPDB2.UTNLPDB31FB2	X	X	X	
10.2415/AH-PERSIAN-MONO-FA-CLEF2008.TEHRAN-NLPDB2.UTNLPDB31NL2	X	X	X	
10.2415/AH-PERSIAN-MONO-FA-CLEF2008.TEHRAN-IRDB.KMEANS5	X	X		
10.2415/AH-PERSIAN-MONO-FA-CLEF2008.TEHRAN-IRDB.KMEANS2	X	X		
10.2415/AH-PERSIAN-MONO-FA-CLEF2008.TEHRAN-IRDB.PDDP_2MEANS2	X	X		
10.2415/AH-PERSIAN-MONO-FA-CLEF2008.TEHRAN-IRDB.KMEANS4	X	X		
10.2415/AH-PERSIAN-MONO-FA-CLEF2008.TEHRAN-IRDB.KMEANS1	X	X		
10.2415/AH-PERSIAN-MONO-FA-CLEF2008.TEHRAN-NLP.STN	X	X		
10.2415/AH-PERSIAN-MONO-FA-CLEF2008.TEHRAN-SEC.MLUSR	X	X		
10.2415/AH-PERSIAN-MONO-FA-CLEF2008.TEHRAN-IRDB.KMEANS3	X	X		
10.2415/AH-PERSIAN-MONO-FA-CLEF2008.TEHRAN-NLP.UTNLPDB2TDN	X	X		
10.2415/AH-PERSIAN-MONO-FA-CLEF2008.TEHRAN-NLPDB.UTNLPDB1M5G	X			
10.2415/AH-PERSIAN-MONO-FA-CLEF2008.TEHRAN-IRDB.PDDP_2MEANS4	X			
10.2415/AH-PERSIAN-MONO-FA-CLEF2008.TEHRAN-IRDB.PDDP_2MEANS3	X			
10.2415/AH-PERSIAN-MONO-FA-CLEF2008.TEHRAN-IRDB.PDDP_2MEANS5	X			
10.2415/AH-PERSIAN-MONO-FA-CLEF2008.TEHRAN-IRDB.PDDP_2MEANS6	X			
10.2415/AH-PERSIAN-MONO-FA-CLEF2008.TEHRAN-NLPDB.UTNLPDB1M4G	X			
10.2415/AH-PERSIAN-MONO-FA-CLEF2008.TEHRAN-IRDB.PDDP_2MEANS1	X			
10.2415/AH-PERSIAN-MONO-FA-CLEF2008.TEHRAN-SEC.MLSR	X			
10.2415/AH-PERSIAN-MONO-FA-CLEF2008.TEHRAN-NLP.STD	X			
10.2415/AH-PERSIAN-MONO-FA-CLEF2008.TEHRAN-IRDB.PDDP	X			
10.2415/AH-PERSIAN-MONO-FA-CLEF2008.TEHRAN-NLPDB2.UTNLPDB3TFIDF	X			
10.2415/AH-PERSIAN-MONO-FA-CLEF2008.TEHRAN-NLPDB2.UTNLPDB3BM25	X			
10.2415/AH-PERSIAN-MONO-FA-CLEF2008.TEHRAN-NLPDB.UTNLPDB1M3G	X			
10.2415/AH-PERSIAN-MONO-FA-CLEF2008.OPENTEXT.OTFA08TDNZ	X			
10.2415/AH-PERSIAN-MONO-FA-CLEF2008.TEHRAN-NLP.TTO		X		
10.2415/AH-PERSIAN-MONO-FA-CLEF2008.TEHRAN-NLP.TTD		X		
10.2415/AH-PERSIAN-MONO-FA-CLEF2008.TEHRAN-NLP.UTNLPDB2TD3		X		
10.2415/AH-PERSIAN-MONO-FA-CLEF2008.TEHRAN-NLP.UTNLPDB2TDW		X		
10.2415/AH-PERSIAN-MONO-FA-CLEF2008.TEHRAN-NLP.UTNLPDB2TNA		X		
10.2415/AH-PERSIAN-MONO-FA-CLEF2008.TEHRAN-NLP.UTNLPDB2TA3		X		
10.2415/AH-PERSIAN-MONO-FA-CLEF2008.TEHRAN-NLP.UTNLPDB2TOT		X		
10.2415/AH-PERSIAN-MONO-FA-CLEF2008.TEHRAN-NLP.UTNLPDB2OTT		X		



Experiment DOI	Groups		
10.2415/AH-PERSIAN-BILI-X2FA-CLEF2008.JHU-APL.JHUENFASK41R400	X		
10.2415/AH-PERSIAN-BILI-X2FA-CLEF2008.JHU-APL.JHUENFA5R100	X		
10.2415/AH-PERSIAN-BILI-X2FA-CLEF2008.JHU-APL.JHUENFAWR50	X		
10.2415/AH-PERSIAN-BILI-X2FA-CLEF2008.JHU-APL.JHUENFAMR50	X		
10.2415/AH-PERSIAN-BILI-X2FA-CLEF2008.TEHRAN-NLPDB.UTNLPDB1BT5G	X		
10.2415/AH-PERSIAN-BILI-X2FA-CLEF2008.TEHRAN-NLPDB.UTNLPDB1BT4G	X		
10.2415/AH-PERSIAN-BILI-X2FA-CLEF2008.TEHRAN-SEC.CLDTDR	X	X	
10.2415/AH-PERSIAN-BILI-X2FA-CLEF2008.TEHRAN-NLPDB.UTNLPDB1BA10	X	X	
10.2415/AH-PERSIAN-BILI-X2FA-CLEF2008.TEHRAN-NLPDB.UTNLPDB1BT1	X	X	X
10.2415/AH-PERSIAN-BILI-X2FA-CLEF2008.TEHRAN-NLPDB.UTNLPDB1BT5	X	X	X
10.2415/AH-PERSIAN-BILI-X2FA-CLEF2008.TEHRAN-SEC.CLQTR	X	X	X
10.2415/AH-PERSIAN-BILI-X2FA-CLEF2008.TEHRAN-NLPDB.UTNLPDB1BA	X	X	X
10.2415/AH-PERSIAN-BILI-X2FA-CLEF2008.TEHRAN-SEC.CLQTR	X	X	X

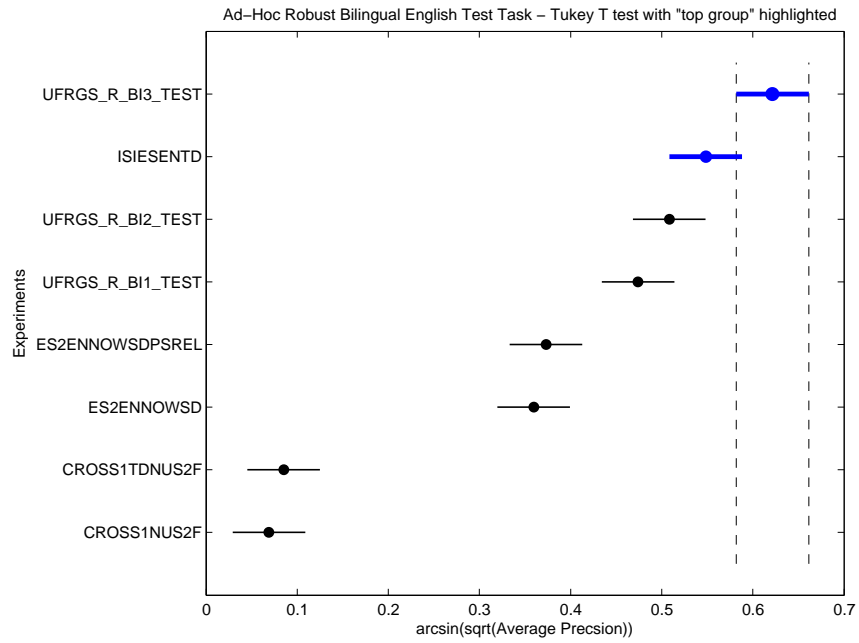
Fig. 26. Ad-Hoc Bilingual Persian. Experiments grouped according to the Tukey T Test.

Ad-Hoc Robust Monolingual English Test Task – Tukey T test with "top group" highlighted



Experiment DOI	Groups					
10.2415/AH-ROBUST-MONO-EN-TEST-CLEF2008.UNINE.UNINEROBUST4	X					
10.2415/AH-ROBUST-MONO-EN-TEST-CLEF2008.UNINE.UNINEROBUST1	X	X				
10.2415/AH-ROBUST-MONO-EN-TEST-CLEF2008.GENEVA.ISILEMTDN	X	X	X			
10.2415/AH-ROBUST-MONO-EN-TEST-CLEF2008.IXA.EN2ENNOWSDPSREL	X	X	X			
10.2415/AH-ROBUST-MONO-EN-TEST-CLEF2008.UCM.BM25_BO1	X	X	X			
10.2415/AH-ROBUST-MONO-EN-TEST-CLEF2008.UCM.BM25_KLD	X	X	X			
10.2415/AH-ROBUST-MONO-EN-TEST-CLEF2008.UCM.BM25_BO1_AVICTF	X	X	X			
10.2415/AH-ROBUST-MONO-EN-TEST-CLEF2008.GENEVA.ISILEMTD	X	X	X	X		
10.2415/AH-ROBUST-MONO-EN-TEST-CLEF2008.UCM.BM25_BASELINE	X	X	X			
10.2415/AH-ROBUST-MONO-EN-TEST-CLEF2008.IXA.EN2ENNOWSD	X	X	X			
10.2415/AH-ROBUST-MONO-EN-TEST-CLEF2008.UFRGS.UFRGS_R_MONO2_TEST	X	X	X	X		
10.2415/AH-ROBUST-MONO-EN-TEST-CLEF2008.UFRGS.UFRGS_R_MONO1_TEST	X	X	X	X	X	
10.2415/AH-ROBUST-MONO-EN-TEST-CLEF2008.INAOE.INAOEF				X	X	
10.2415/AH-ROBUST-MONO-EN-TEST-CLEF2008.KNOW-CENTER.ASSO				X	X	
10.2415/AH-ROBUST-MONO-EN-TEST-CLEF2008.INAOE.INAOEV					X	
10.2415/AH-ROBUST-MONO-EN-TEST-CLEF2008.UNIBA.MONO11NUS2F						X
10.2415/AH-ROBUST-MONO-EN-TEST-CLEF2008.UNIBA.MONO11TDNUS2F						X
10.2415/AH-ROBUST-MONO-EN-TEST-CLEF2008.UNIBA.MONO13NUS2F						X
10.2415/AH-ROBUST-MONO-EN-TEST-CLEF2008.UNIBA.MONO12NUS2FOUT						X

Fig. 27. Robust Monolingual English. Experiments grouped according to the Tukey T Test.



Experiment DOI	Groups		
10.2415/AH-ROBUST-BILI-X2EN-TEST-CLEF2008.UFRGS.UFRGS_R_BI3_TEST	X		
10.2415/AH-ROBUST-BILI-X2EN-TEST-CLEF2008.GENEVA.ISIESENTD	X	X	
10.2415/AH-ROBUST-BILI-X2EN-TEST-CLEF2008.UFRGS.UFRGS_R_BI2_TEST	X	X	
10.2415/AH-ROBUST-BILI-X2EN-TEST-CLEF2008.UFRGS.UFRGS_R_BI1_TEST	X	X	
10.2415/AH-ROBUST-BILI-X2EN-TEST-CLEF2008.IXA.ES2ENNOWSDPSREL			X
10.2415/AH-ROBUST-BILI-X2EN-TEST-CLEF2008.IXA.ES2ENNOWSD			X
10.2415/AH-ROBUST-BILI-X2EN-TEST-CLEF2008.UNIBA.CROSS1TDNUS2F			X
10.2415/AH-ROBUST-BILI-X2EN-TEST-CLEF2008.UNIBA.CROSS1NUS2F			X

Fig. 28. Robust Bilingual English. Experiments grouped according to the Tukey T Test.

Ad-Hoc Robust Word Sense Disambiguation Monolingual English Test Task – Tukey T test with "top group

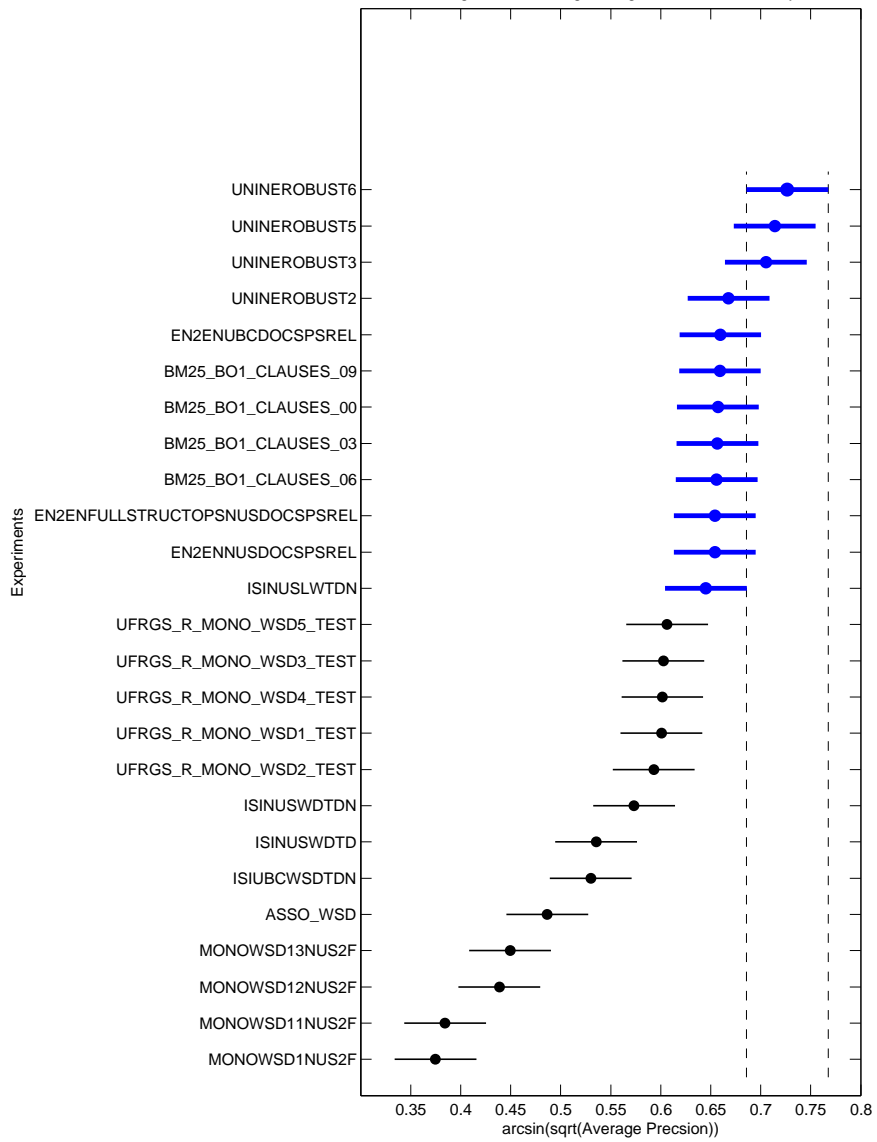
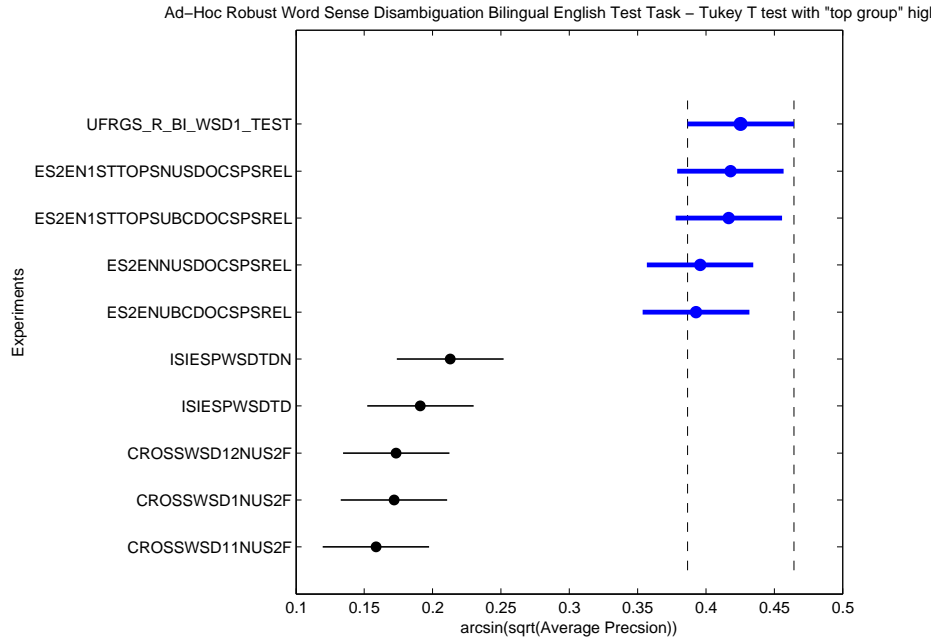


Fig. 29. Robust WSD Monolingual English. The figure shows the Tukey T Test.



Experiment DOI	Groups
10.2415/AH-ROBUST-WSD-BILI-X2EN-TEST-CLEF2008.UFRGS.UFRGS_R_BI_WSD1_TEST	X
10.2415/AH-ROBUST-WSD-BILI-X2EN-TEST-CLEF2008.IXA.ES2EN1STTOPSNUSDOCSPSREL	X
10.2415/AH-ROBUST-WSD-BILI-X2EN-TEST-CLEF2008.IXA.ES2EN1STTOPSUBCDOCSPSREL	X
10.2415/AH-ROBUST-WSD-BILI-X2EN-TEST-CLEF2008.IXA.ES2ENNUSDOCSPSREL	X
10.2415/AH-ROBUST-WSD-BILI-X2EN-TEST-CLEF2008.IXA.ES2ENUBCDOCSPSREL	X
10.2415/AH-ROBUST-WSD-BILI-X2EN-TEST-CLEF2008.GENEVA.ISIESPWSDTDN	X
10.2415/AH-ROBUST-WSD-BILI-X2EN-TEST-CLEF2008.GENEVA.ISIESPWSDTD	X
10.2415/AH-ROBUST-WSD-BILI-X2EN-TEST-CLEF2008.UNIBA.CROSSWSD12NUS2F	X
10.2415/AH-ROBUST-WSD-BILI-X2EN-TEST-CLEF2008.UNIBA.CROSSWSD1NUS2F	X
10.2415/AH-ROBUST-WSD-BILI-X2EN-TEST-CLEF2008.UNIBA.CROSSWSD11NUS2F	X

Fig. 30. Robust WSD Bilingual English. Experiments grouped according to the Tukey T Test.

7 Conclusions

The ad hoc task this year has been almost completely renovated with new collections and new tasks. For all three tasks, we have been very happy with number of participants. However, it is really too soon to be able to provide any deep analysis of the results obtained. This is left to the post-workshop proceedings. In any case, it is our intention to run all three tasks for a second year both in order to provide participants with another chance to test their systems after refinement and tuning on the basis of this year's experiments and also to be able to create useful and consolidated test collections. In particular, for both the TEL and Persian tasks, we intend to perform some experiments on this year's

test collections in order to verify their stability. The results will be reported in the Proceedings.

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

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

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

8 Acknowledgements

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

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

The robust task was partially funded by the Ministry of Education (project KNOW TIN2006-15049) and the European Commission (project KYOTO ICT-2007-211423). We would like to acknowledge the invaluable suggestions and practical assistance of Arantxa Otegi, German Rigau and Piek Vossen. In particular, Arantxa was responsible for the XML coding of the word sense disambiguation

results. We also want to thank Oier Lopez de Lacalle, who runs the UBC WSD system, and Yee Seng Chan, Hwee Tou Ng and Zhi Zhong, who run the NUS WSD system. Their generous contribution was invaluable to run this exercise.

References

1. Aghazade, Z., Dehghani, N., Farzinvas, L., Rahimi, R., AleAhmad, A., Amiri, H., Oroumchian, F.: Fusion of Retrieval Models at CLEF 2008 Ad-Hoc Persian Track. In this volume.
2. Agirre, E., Lopez de Lacalle, O.: UBC-ALM: Combining k-NN with SVD for WSD. In Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval 2007), Prague, Czech Republic (2007) 341–345
3. Agirre, E., Magnini, B., Lopez de Lacalle, O., Otegi, A., Rigau, G., Vossen, P.: SemEval-2007 Task01: Evaluating WSD on Cross-Language Information Retrieval. In Proceedings of CLEF 2007 Workshop, Budapest, Hungary (2007).
4. Agosti, M., Di Nunzio, G.M., Ferro, N.: A Proposal to Extend and Enrich the Scientific Data Curation of Evaluation Campaigns. In Sakay, T., Sanderson, M., Evans, D.K., eds.: Proc. 1st International Workshop on Evaluating Information Access (EVIA 2007), National Institute of Informatics, Tokyo, Japan (2007) 62–73
5. Agosti, M., Di Nunzio, G.M., Ferro, N.: The Importance of Scientific Data Curation for Evaluation Campaigns. In Thanos, C., Borri, F., eds.: DELOS Conference 2007 Working Notes, ISTI-CNR, Gruppo ALI, Pisa, Italy 185–193
6. Agosti, M., Di Nunzio, G.M., Ferro, N.: Scientific Data of an Evaluation Campaign: Do We Properly Deal With Them? In Peters, C., Clough, P., Gey, F.C., Karlgren, J., Magnini, B., Oard, D.W., de Rijke, M., Stempfhuber, M., eds.: Evaluation of Multilingual and Multi-modal Information Retrieval : Seventh Workshop of the Cross-Language Evaluation Forum (CLEF 2006). Revised Selected Papers, Lecture Notes in Computer Science (LNCS) 4730, Springer, Heidelberg, Germany (2007) 11–20
7. AleAhmad, A., Kamaloo, E., Zareh, A., Rahgozar, M., Oroumchian, F.: Cross Language Experiments at Persian@CLEF 2008. In this volume.
8. Basile, P., Caputo, A., Semeraro, G.: UNIBA-SENSE at CLEF 2008: SEMantic N-levels Search Engine. In this volume.
9. Braschler, M.: CLEF 2003 - Overview of results. In Peters, C., Braschler, M., Gonzalo, J., Kluck, M., eds.: Comparative Evaluation of Multilingual Information Access Systems: Fourth Workshop of the Cross-Language Evaluation Forum (CLEF 2003) Revised Selected Papers, Lecture Notes in Computer Science (LNCS) 3237, Springer, Heidelberg, Germany (2004) 44–63
10. Braschler, M., Peters, C.: CLEF 2003 Methodology and Metrics. In Peters, C., Braschler, M., Gonzalo, J., Kluck, M., eds.: Comparative Evaluation of Multilingual Information Access Systems: Fourth Workshop of the Cross-Language Evaluation Forum (CLEF 2003) Revised Selected Papers, Lecture Notes in Computer Science (LNCS) 3237, Springer, Heidelberg, Germany (2004) 7–20
11. Bosca, A., Dini, L.: CACAO Project at the TEL@CLEF 2008 Task. In this volume.
12. Chan, Y. S., Ng, H. T., Zhong, Z.: NUS-PT: Exploiting Parallel Texts for Word Sense Disambiguation in the English All-Words Tasks In Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval 2007), Prague, Czech Republic (2007) 253–256

13. Cleverdon, C.: The Cranfield Tests on Index Language Devices. In Sparck Jones, K., Willett, P., eds.: *Readings in Information Retrieval*, Morgan Kaufmann Publisher, Inc., San Francisco, California, USA (1997) 47–59
14. Clinchant, S., Renders, J.-M.: XRCE’s Participation to CLEF 2008 Ad-Hoc Track. In this volume.
15. Conover, W.J.: *Practical Nonparametric Statistics*. 1st edn. John Wiley and Sons, New York, USA (1971)
16. Costa Acosta, O., Geraldo, A.P., Orengo, V.M., Villavicencio, A.: UFRGS@CLEF2008: Indexing Multiword Expressions for Information Retrieval. In this volume.
17. Di Nunzio, G.M., Ferro, N.: Appendix A: Results of the TEL@CLEF Task. In Borri, F., Nardi, A., Peters, C., eds.: *Working Notes for the CLEF 2008 Workshop*, <http://www.clef-campaign.org/> [last visited 2008, September 5] (2008)
18. Di Nunzio, G.M., Ferro, N.: Appendix B: Results of the Persian Task. In Borri, F., Nardi, A., Peters, C., eds.: *Working Notes for the CLEF 2008 Workshop*, <http://www.clef-campaign.org/> [last visited 2008, September 5] (2008)
19. Di Nunzio, G.M., Ferro, N.: Appendix C: Results of the Robust Task. In Borri, F., Nardi, A., Peters, C., eds.: *Working Notes for the CLEF 2008 Workshop*, <http://www.clef-campaign.org/> [last visited 2008, September 5] (2008)
20. Dolamic, L., Fautsch, C., Savoy, J.: UniNE at CLEF2008: TEL, Perisan and Robust IR. In this volume.
21. Geraldo, A.P., Orengo, V.M.: UFRGS@CLEF2008: Using Association Rules for Cross-Language Information Retrieval. In this volume.
22. Guyot, J., Falquet, G., Radhouani, S., Benzineb, K.: UNIGE Experiments on Robust Word sense Disambiguation. In this volume.
23. Hull, D.: Using Statistical Testing in the Evaluation of Retrieval Experiments. In Korfhage, R., Rasmussen, E., Willett, P., eds.: *Proc. 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1993)*, ACM Press, New York, USA (1993) 329–338
24. Jadidinejad, A.H., Mohtarami, M., Amiri, H.: Investigation on Application of Local Cluster Analysis and Part of Speech Tagging on Persian Text. In this volume.
25. Judge, G.G., Hill, R.C., Griffiths, W.E., Lütkepohl, H., Lee, T.C.: *Introduction to the Theory and Practice of Econometrics*. 2nd edn. John Wiley and Sons, New York, USA (1988)
26. Karimpour, R., Ghorbani, A., Pishdad, A., Mohtarami, M., AleAhmad, A., Amiri, H., Oroumchian, F.: Using Part of Speech tagging in Persian Information Retrieval. In this volume.
27. Kuersten, J., Wilhelm, T., Eibl, M.: CLEF 2008 Ad-Hoc Track: On-line Processing Experiments with Xtrieval. In this volume.
28. Larson, R.: Logistic Regression for Metadata: Cheshire takes on Adhoc-TEL. In this volume.
29. Machado, J., Martins, B., Borbinha, J.: Technical University of Lisbon CLEF 2008 Submission (TEL@CLEF Monolingual Task). In this volume.
30. Martínez-Santiago, F., Perea-Ortega, J.M., García-Cumbreras, M.A.: SINAI at Robust WSD Task @ CLEF 2008: When WSD is a Good Idea for Information Retrieval Tasks. In this volume.
31. MacNamee, P.: JHU Ad Hoc Experiments at CLEF 2008. In this volume.
32. Navarro, S., Llopis, F., Muñoz, R.: IRn in the CLEF Robust WSD Task 2008. In this volume.

33. Nguyen, D., Overwijk, A., Hauff, C., Trieschnigg, R.B., Hiemstra, D., de Jong, F.M.G.: WikiTranslate: Query Translation for Cross-lingual Information Retrieval using only Wikipedia. In this volume.
34. Otegi, A., Agirre, E., Rigau, G. IXA at CLEF 2008 Robust-WSD Task using Word Sense Disambiguation for (Cross Lingual) Information Retrieval. In this volume.
35. Paskin, N., ed.: The DOI Handbook – Edition 4.4.1. International DOI Foundation (IDF). <http://dx.doi.org/10.1000/186> [last visited 2007, August 30] (2006)
36. Pérez-Agüera, J.R., Zaragoza, H.: UCM-Y!R at CLEF 2008 Robust and WSD Tasks. In this volume.
37. Robertson, S.: On GMAP: and Other Transformations. In Yu, P.S., Tsotras, V., Fox, E.A., Liu, C.B., eds.: Proc. 15th International Conference on Information and Knowledge Management (CIKM 2006), ACM Press, New York, USA (2006) 78–83
38. Sorg, P., Cimiano, P.: Cross-lingual Information Retrieval with Explicit Semantic Analysis. In this volume.
39. Tague-Sutcliffe, J.: The Pragmatics of Information Retrieval Experimentation, Revisited. In Sparck Jones, K., Willett, P., eds.: Readings in Information Retrieval, Morgan Kaufmann Publisher, Inc., San Francisco, California, USA (1997) 205–216
40. Tomlinson, S.: German, French, English and Persian Retrieval Experiments at CLEF 2008. In this volume.
41. Voorhees, E.M.: The TREC Robust Retrieval Track. SIGIR Forum **39** (2005) 11–20