CLEF 2009 Ad Hoc Track Overview: Robust-WSD Task

Eneko Agirre[1], Giorgio Maria Di Nunzio[2], Thomas Mandl[3], and Arantxa Otegi[1]

[1] Computer Science Department, University of the Basque Country, Spain
{e.agirre,arantza.otegi}@ehu.es
[2] Department of Information Engineering, University of Padua, Italy
dinunzio@dei.unipd.it
[3] Information Science, University of Hildesheim, Germany
mandl@uni-hildesheim.de

Abstract. The Robust-WSD at CLEF 2009 aims at exploring the contribution of Word Sense Disambiguation to monolingual and multilingual Information Retrieval. The organizers of the task provide documents and topics which have been automatically tagged with Word Senses from WordNet using several state-of-the-art Word Sense Disambiguation systems. The Robust-WSD exercise follows the same design as in 2008. It uses two languages often used in previous CLEF campaigns (English, Spanish). Documents were in English, and topics in both English and Spanish. The document collections are based on the widely used LA94 and GH95 news collections. All instructions and datasets required to replicate the experiment are available from the organizers website (http://ixa2.si.ehu.es/clirwsd/). The results show that some top-scoring systems improve their IR and CLIR results with the use of WSD tags, but the best scoring runs do not use WSD.

1 Introduction

The Robust-WSD task at CLEF 2009 aims at exploring the contribution of Word Sense Disambiguation to monolingual and multilingual Information Retrieval. The organizers of the task provide documents and topics which have been automatically tagged with Word Senses from WordNet using several state-of-the-art Word Sense Disambiguation systems. The task follows the same design as at CLEF 2008.

The robust task ran for the fourth time at CLEF 2009. 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 [17,18]. 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.

For the second year, the robust task also incorporated word sense disambiguation information provided by the organizers to the participants. The task follows

C. Peters et al. (Eds.): CLEF 2009 Workshop, Part I, LNCS 6241, pp. 36–49, 2010.

the 2007 joint SemEval-CLEF task [2] and the 2008 Robust-WSD exercise [3], and has the aim of exploring the contribution of word sense disambiguation to monolingual and cross-language information retrieval. 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 [1] and NUS [8]. These systems provided weighted word sense tags for each of the nouns, verbs, adjectives and adverbs that they could disambiguate. These systems participated in the Semeval 2007 task on Word Sense Disambiguation[16], with similar results. NUS ranked 4th in the all-words task, with an accuracy of 57.4, and UBC ranked 5th, with an accuracy of 54.4. In the all-words task, the output of both systems was compared with the gold stadard sense tags on a sample of three documents.

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.

The datasets used in this task can be used in the future to run further experiments. Check `http://ixa2.si.ehu.es/clirwsd` for information of how to access the datasets. Topics and relevance judgements are freely available. The document collection can be obtained from ELDA purchasing the CLEF Test Suite for the CLEF 2000-2003 Campaigns – Evaluation Package. As an alternative, the website offers the unordered set of words in each document, that is, the full set of documents where the positional information has been eliminated to avoid replications of the originals. Lucene indexes for the later are also available from the website.

In this paper, we first present the task setup, the evaluation methodology and the participation in the different tasks (Section 2). We then describe the main features of each task and show the results (Sections 3 - 5). The final section provides a brief summing up. For information on the various approaches and resources used by the groups participating in this task and the issues they focused on, we refer the reader to the rest of the papers in the Robust-WSD part of the Ad Hoc section of these Proceedings.

2 Task Setup

The Ad Hoc task 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 [9]. 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.

2.1 Test Collections

The Documents. The robust task used existing CLEF news collections but with word sense disambiguation (WSD) information added. The word sense disambiguation data was automatically added by systems from two leading research laboratories, UBC [1] and NUS [8]. 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[1]:

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

The Topics. Topics 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 the relevance assessment criteria. Topics are prepared in xml format and identified by means of a Digital Object Identifier (DOI)[2] of the experiment [14] which allows us to reference and cite them.

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

- CLEF 2001; Topics 10.2452/41-AH – 10.2452/90-AH; LA Times 94
- CLEF 2002; Topics 10.2452/91-AH – 10.2452/140-AH; LA Times 94
- CLEF 2003; Topics 10.2452/141-AH – 10.2452/200-AH; LA Times 94, Glasgow Herald 95
- CLEF 2004; Topics 10.2452/201-AH – 10.2452/250-AH; Glasgow Herald 95
- CLEF 2005; Topics 10.2452/251-AH – 10.2452/300-AH; LA Times 94, Glasgow Herald 95
- CLEF 2006; Topics 10.2452/301-AH – 10.2452/350-AH; LA Times 94, Glasgow Herald 95

[1] A sample document and dtd are available at `http://ixa2.si.ehu.es/clirwsd/`
[2] `http://www.doi.org/`

```
<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. 1. Example of Robust WSD topic: topic `10.2452/141-WSD-AH`

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 [1] and NUS [8] systems, yielding word senses from WordNet version 1.6. A large-scale disambiguation system for Spanish was not available, so we used the first-sense heuristic, yielding senses from the Spanish wordnet, which is tightly aligned to the English WordNet version 1.6 (i.e., they share synset numbers or sense codes). An excerpt from a topic is shown in Figure 1, where each term in the topic is followed by its senses with their respective scores as assigned buy the automatic WSD system[3].

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

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 robust WSD task used existing relevance assessments from previous years. The relevance assessments regarding the training topics were provided to participants before competition time.

The total number of assessments was 66,441 documents of which 4,327 were relevant. The distribution of the pool according to each year was the following:

- CLEF 2003: 23,674 documents, 1,006 relevant;
- CLEF 2005: 19,790 document, 2,063 relevant;
- CLEF 2006: 21,247 document, 1,258 relevant;

Seven topics had no relevant documents at all: 10.2452/149-AH, 10.2452/161-AH, 10.2452/166-AH, 10.2452/186-AH, 10.2452/191-AH, 10.2452/195-AH, 10.2452/321-AH. Each topic had an average of about 28 relevant documents and a standard deviation of 34, a minimum of 1 relevant document and a maximum of 229 relevant documents per topic.

2.2 Result Calculation

Evaluation campaigns such as TREC and CLEF are based on the belief that the effectiveness of *Information Retrieval Systems (IRSs)* can be objectively evaluated by an analysis of a representative set of sample search results. For this, effectiveness measures are calculated based on the results submitted by the participants and the relevance assessments. Popular measures usually adopted for exercises of this type are Recall and Precision. Details on how they are calculated for CLEF are given in [7].

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 [17,18]. This makes especially sense in multilingual retrieval where results can differ from results based on MAP [15].

The individual results for all official Ad Hoc experiments in CLEF 2009 are given in the one of the Appendices of the CLEF 2009 Working Notes [11].

2.3 Participants and Experiments

As shown in Table 1, 10 groups submitted 89 runs for the Robust tasks:

- 8 groups submitted monolingual non-WSD runs (25 runs out of 89);
- 5 groups also submitted bilingual non-WSD runs (13 runs out of 89).

All groups submitted WSD runs (51 out of 89 runs):

- 10 groups submitted monolingual WSD runs (33 out of 89 runs)
- 5 groups submitted bilingual WSD runs (18 out of 89 runs)

Table 1. CLEF 2009 Ad Hoc Robust participants. See text in Section 2.3. for comments on participants with *.

participant	task	No. experiments
alicante*	AH-ROBUST-WSD-MONO-EN-TEST-CLEF2009	3
darmstadt	AH-ROBUST-MONO-EN-TEST-CLEF2009	5
darmstadt	AH-ROBUST-WSD-MONO-EN-TEST-CLEF2009	5
geneva*	AH-ROBUST-MONO-EN-TEST-CLEF2009	5
geneva*	AH-ROBUST-WSD-BILI-X2EN-TEST-CLEF2009	1
geneva*	AH-ROBUST-WSD-MONO-EN-TEST-CLEF2009	2
ixa	AH-ROBUST-BILI-X2EN-TEST-CLEF2009	1
ixa	AH-ROBUST-MONO-EN-TEST-CLEF2009	1
ixa	AH-ROBUST-WSD-BILI-X2EN-TEST-CLEF2009	4
ixa	AH-ROBUST-WSD-MONO-EN-TEST-CLEF2009	3
jaen*	AH-ROBUST-WSD-MONO-EN-TEST-CLEF2009	2
know-center	AH-ROBUST-BILI-X2EN-TEST-CLEF2009	3
know-center	AH-ROBUST-MONO-EN-TEST-CLEF2009	3
know-center	AH-ROBUST-WSD-BILI-X2EN-TEST-CLEF2009	3
know-center	AH-ROBUST-WSD-MONO-EN-TEST-CLEF2009	3
reina*	AH-ROBUST-BILI-X2EN-TEST-CLEF2009	5
reina*	AH-ROBUST-MONO-EN-TEST-CLEF2009	5
reina*	AH-ROBUST-WSD-BILI-X2EN-TEST-CLEF2009	5
reina*	AH-ROBUST-WSD-MONO-EN-TEST-CLEF2009	5
ufrgs	AH-ROBUST-BILI-X2EN-TEST-CLEF2009	1
ufrgs	AH-ROBUST-MONO-EN-TEST-CLEF2009	1
ufrgs	AH-ROBUST-WSD-MONO-EN-TEST-CLEF2009	1
uniba	AH-ROBUST-BILI-X2EN-TEST-CLEF2009	3
uniba	AH-ROBUST-MONO-EN-TEST-CLEF2009	3
uniba	AH-ROBUST-WSD-BILI-X2EN-TEST-CLEF2009	5
uniba	AH-ROBUST-WSD-MONO-EN-TEST-CLEF2009	5
valencia	AH-ROBUST-MONO-EN-TEST-CLEF2009	2
valencia	AH-ROBUST-WSD-MONO-EN-TEST-CLEF2009	4

Table 2 provides a breakdown of the number of participants and submitted runs by task. Note that jaen submitted a monolingual non-WSD run as if it was a WSD run, and that alicante missed to send their non-WSD run on time. Although REINA submitted some runs under WSD, they did not use WSD information [20], only lemma and PoS. Geneva did not submit a paper describing their systems. The figures used in this paper are the official figures at the time of the task.

3 Results

Table 3 shows the best results for the monolingual runs, and Table 4 shows the best results for the bilingual runs. In the following pages, Figure 2 shows the performances of the best systems in terms of average precision of the top participants of the Robust Monolingual and Monolingual WSD, and Figure 3

Table 2. Number of runs per track

Track	# Part.	# Runs
Robust Mono English Test	8	25
Robust Mono English Test WSD	10	33
Robust Biling. English Test	5	13
Robust Biling. English Test WSD	5	18

Table 3. Best entries for the robust monolingual task, including both WSD and non-WSD runs. The **Q** columns shows the information used to build the query.

	Rank	Participant	Q	Experiment DOI	MAP	GMAP
Non-WSD	**1st**	darmstadt	TD	10.2415/AH-ROBUST-MONO-EN-TEST-CLEF2009.DARMSTADT.DA_4	45.09%	20.42%
	2nd	reina	TDN	10.2415/AH-ROBUST-MONO-EN-TEST-CLEF2009.REINA.ROB2	44.52%	21.18%
	3rd	uniba	TDN	10.2415/AH-ROBUST-MONO-EN-TEST-CLEF2009.UNIBA.UNIBAKRF	42.50%	17.93%
	4th	geneva	TDN	10.2415/AH-ROBUST-MONO-EN-TEST-CLEF2009.GENEVA.ISIENNATTDN	41.71%	17.88%
	5th	know-center	TD	10.2415/AH-ROBUST-MONO-EN-TEST-CLEF2009.KNOW-CENTER.ASSO	41.70%	18.64%
WSD	**1st**	darmstadt	TD	10.2415/AH-ROBUST-WSD-MONO-EN-TEST-CLEF2009.DARMSTADT.DA_WSD_4	45.00%	20.49%
	2nd	uniba	TDN	10.2415/AH-ROBUST-WSD-MONO-EN-TEST-CLEF2009.UNIBA.UNIBAKEYSYNRF	43.46%	19.60%
	3rd	know-center	TD	10.2415/AH-ROBUST-WSD-MONO-EN-TEST-CLEF2009.KNOW-CENTER.ASSOWSD	42.22%	19.47%
	4th	geneva	TDN	10.2415/AH-ROBUST-WSD-MONO-EN-TEST-CLEF2009.GENEVA.ISINUSLWTDN	38.11%	16.26%
	5th	ixa	TD	10.2415/AH-ROBUST-WSD-MONO-EN-TEST-CLEF2009.IXA.ENENBESTSENSE500DOCS	38.05%	16.57%

Table 4. Best entries for the robust ES-EN bilingual task, including both WSD and non-WSD runs. The **Q** columns shows the information used to build the query.

	Rank	Participant	Q	Experiment DOI	MAP	GMAP
Non-WSD	**1st**	reina	TDN	10.2415/AH-ROBUST-BILI-X2EN-TEST-CLEF2009.REINA.BILI2	38.42%	15.11%
	2nd	uniba	TDN	10.2415/AH-ROBUST-BILI-X2EN-TEST-CLEF2009.UNIBA.UNIBACROSSKEYRF	38.09%	13.11%
	3rd	know-center	TD	10.2415/AH-ROBUST-BILI-X2EN-TEST-CLEF2009.KNOW-CENTER.BILIASSO	28.98%	06.79%
	4th	ufrgs	TD	10.2415/AH-ROBUST-BILI-X2EN-TEST-CLEF2009.UFRGS.BILINGUAL	27.65%	07.37%
	5th	ixa	TD	10.2415/AH-ROBUST-BILI-X2EN-TEST-CLEF2009.IXA.ESENNOWSD	18.05%	01.90%
WSD	**1st**	uniba	TDN	10.2415/AH-ROBUST-WSD-BILI-X2EN-TEST-CLEF2009.UNIBA.UNIBACROSSKEYSYNRF	37.53%	13.82%
	2nd	geneva	TD	10.2415/AH-ROBUST-WSD-BILI-X2EN-TEST-CLEF2009.GENEVA.ISINUSWSDTD	36.63%	16.02%
	3rd	know-center	TD	10.2415/AH-ROBUST-WSD-BILI-X2EN-TEST-CLEF2009.KNOW-CENTER.BILIASSOWSD	29.64%	07.05%
	4rd	ixa	TD	10.2415/AH-ROBUST-WSD-BILI-X2EN-TEST-CLEF2009.IXA.ESEN1STTOPSBESTSENSE500DOCS	18.38%	01.98%

shows the performances of the best participants of the Robust Bilingual and Bilingual WSD. Some teams used the Title and Description fields to construct the query (TD), while others also used the narrative (TDN).

The comparison of the bilingual runs with respect to the monolingual results yield the following:

- ES → EN: 85.2% of best monolingual English IR system (MAP);
- ES → EN WSD: 83.3% of best monolingual English IR system (MAP);

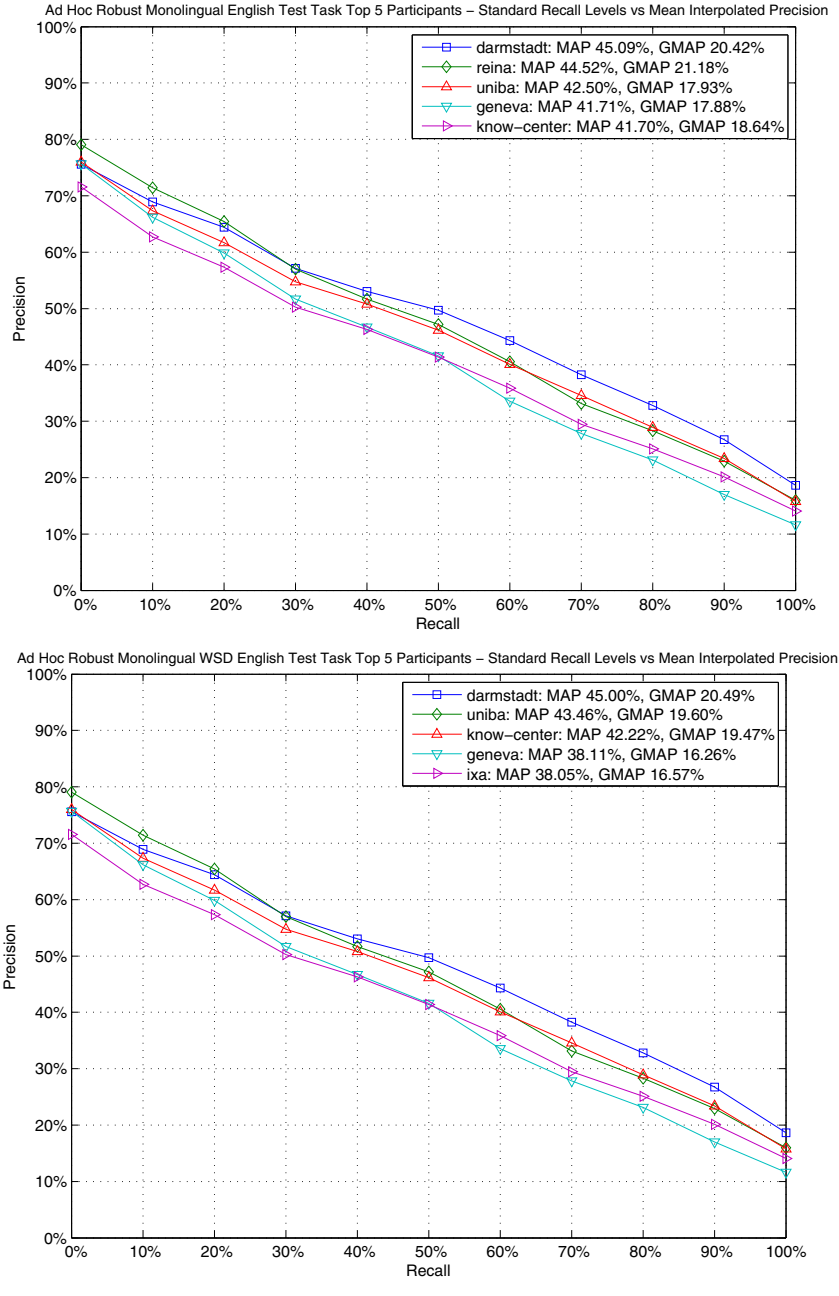

Fig. 2. Mean average precision of the top 5 participants of the Robust Monolingual English Task (top graph) and Robust WSD Monolingual English Task (bottom).

Table 5. Statistical tests comparison between non-WSD and WSD runs. Differences or equalities are statistically significant with alpha = 5%

	Monolingual	Bilingual
Task	Non-WSD > WSD	Non-WSD > WSD
Set of best runs	Non-WSD > WSD	Non-WSD = WSD
Single best run	Non-WSD = WSD	Non-WSD = WSD

3.1 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 whether the results of the runs of a task are significantly different from the results of other tasks. In particular, we want to test whether there is any difference between applying WSD techniques or not. 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) [18].

We used the MATLAB Statistics Toolbox, which provides the necessary functionality plus some additional functions and utilities. Following the approach presented by [10], the first step is to verify whether the distributions of performances are normal, the second step is to analyze whether the variances of the distributions are equal, and finally to test whether the means of the distributions are the same. Three different pairs of distributions were analyzed to verify the differences between WSD and non-WSD experiments:

- Task:
  - Robust Monolingual vs Robust WSD Monolingual;
  - Robust Bilingual vs Robust WSD Bilingual.
- Set of best experiments:
  - Best performers of Robust Monolingual vs Best Robust WSD Monolingual (experiments of Table 3);
  - Best performers of Robust Bilingual vs Best performers Robust WSD Bilingual (experiments of Table 4).
- Single best experiment:
  - Best Robust Monolingual vs Best Robust WSD Monolingual;
  - Best Robust Bilingual vs Best Robust WSD Bilingual.

Results are summarized in Table 3.1, showing that overall, systems not using WSD perform better than those using WSD. When we take the best systems, results are not statistically different. These comparisons are done without taking into account who is producing which runs. Another alternative is to analize each participant system separately, as we will do in the next section.

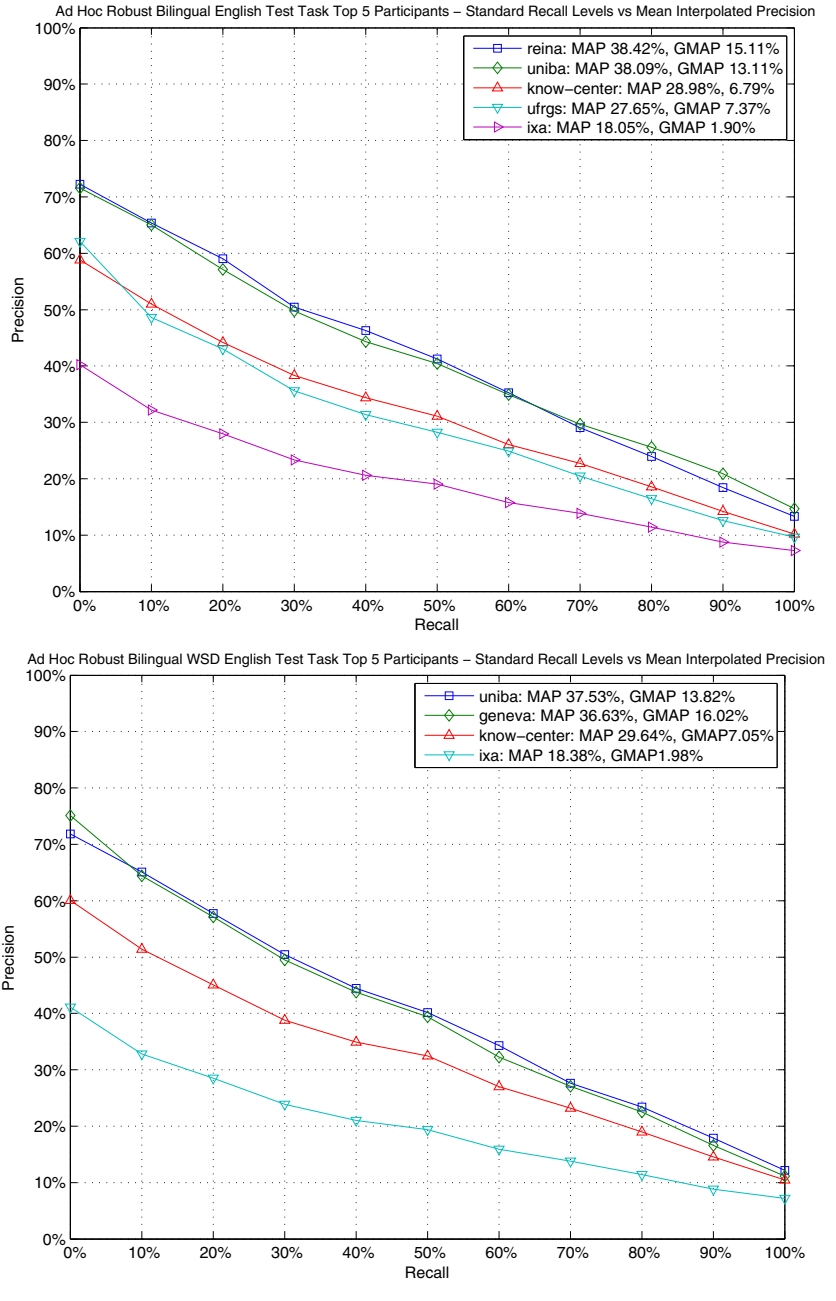

Fig. 3. Mean average precision of the top 5 participants of the Robust Bilingual English Task (top graph) and Robust WSD Bilingual English Task (bottom)

Table 6. Statistical tests comparison between non-WSD and WSD runs of best participants. Differences or equalities are statistically significant with alpha = 5%.

System	Monolingual	Bilingual
darmstadt	Non-WSD = WSD	n/a
uniba	Non-WSD < WSD	Non-WSD = WSD
geneva	Non-WSD > WSD	n/a
know-center	Non-WSD = WSD	Non-WSD = WSD

3.2 Analysis

In this section we focus on the comparison between WSD and non-WSD runs of each participant. Overall, the best MAP and GMAP results in the monolingual system were for two distinct runs which did not use WSD information, but several participants were able to obtain their best MAP and GMAP scores using WSD information. In the bilingual experiments, the best results in MAP were for non-WSD runs, but two participants were able to profit from the WSD annotations. As it is difficult to summarize the behavior of all participants, we will only mention the performance of the best teams, as given in Tables 3 and 4. In addition, Table 3.2 summarizes whether the best WSD and non-WSD runs for participants that submitted both runs are statistically significant. The interested reader is directed to the papers of each participant in this volume for additional details.

In the **monolingual experiments**, cf. Table 3, the best results overall in MAP were for **darmstadt**. Their WSD runs scored very similar to the non-WSD runs, with a slight decrease of MAP (0.09 percentage points, with no statistical difference) and a slight increase of GMAP (0.07 percentage points) [19]. The method to include WSD information was to create additional indexes for word senses, and then combine them with other indexes using weights as optimized from training. The retrieval system used the BM25 model, with an additional monolingual translation-based model.

The second best MAP score and best GMAP was attained by **reina** [20] without WSD. Unfortunately they did not submit any run using WSD. Systems such as this introduce noise in the comparisons in the previous section.

The third best MAP and second GMAP where obtained by **uniba** [5] using WSD. This team showed a 0.94 statistically different increase in MAP and 1.67 increase in GMAP with respect to their best non-WSD run. They constructed an additional index with synset numbers, and then combined the synsets using the N-levels model.

geneva [12] also attained good results, but their WSD system had a statistically significant drop in both MAP and GMAP. Unfortunately, the authors did not explain how they integrated WSD information in their system.

Finally, **know-center** [13] attained 0.52 improvements in MAP using WSD (not statistically significant difference) and 0.83 increase in GMAP with the use of WSD. They added synsets and synonyms to the index, and used an axiomatic retrieval approach.

In the **bilingual experiments**, cf. Table 4, the best results overall in MAP were for **reina** with a system which did not use WSD annotations [20], and again, they did not submit runs using WSD.

The best GMAP was for **geneva** using WSD [12], but unfortunately, they did not submit any non-WSD run.

uniba [5] got the second best MAP, with better MAP for the non-WSD run and better GMAP for the WSD run. The differences were small in both cases (0.56 in MAP, 0.71 in GMAP).

Those three teams had the highest results, well over 35% MAP, and the rest got more modest performances. **know-center** [13] reported better results using WSD information (0.66 MAP, 0.26 GMAP). **Ufrgs** [6] only submitted the WSD result. Finally **ixa** [4] got low results, with small improvements using WSD information (0.33 MAP, 0.08 GMAP).

All in all, the exercise showed that some teams did improve results using WSD (close to 1 MAP point and more than 1 GMAP point in monolingual, and below 1 MAP/GMAP point in bilingual), but the best results for both monolingual and bilingual tasks were for systems which did not use WSD.

4 Conclusions

This new edition of the robust WSD exercise has measured to what extent IR systems could profit from automatic word sense disambiguation information. The conclusions on the monolingual subtask are similar to the conclusions of 2008. The evidence for using WSD in monolingual IR is mixed. Some top scoring groups report improvements in MAP and GMAP, with significant improvements in the case of **uniba**, which attained the third best results. Still, the best overall scores are for two systems not using WSD. Regarding the cross-lingual task, the situation is very similar, but the reported improvements using WSD are smaller.

The lower performance of some groups when using WSD seems to indicate that using WSD for IR is not straightforward, and can lead to worse results if not done with care. From another perspective, the results of groups which do attest significant improvements are very relevant, as they show that a careful system design can render WSD information effective. We thus think that, overall, the results of the 2008 and 2009 campaigns are promising, and that there is still room for improvement.

The instructions and all datasets to replicate the results (including Lucene indexes) are available from `http://ixa.si.ehu.es/clirwsd`. Topics and relevance judgements are freely available. The document collection can be obtained from ELDA purchasing the CLEF Test Suite for the CLEF 2000-2003 Campaigns – Evaluation Package. As an alternative, the website offers the unordered set of words in each document, that is, the full set of documents where the positional information has been eliminated to avoid replications of the originals. Lucene indexes for the later are also available from the website. Given the availability of these resources, interested parties can now evaluate their own systems, and we thus felt that there is no need to organize another edition of the competition.

Acknowledgements

The robust task was partially funded by the Ministry of Education (project KNOW2 TIN2009-14715) and the European Commission (project KYOTO ICT-2007-211423). We want to thank Oier Lopez de Lacalle, who run 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 organize this exercise.

References

1. Agirre, E., Lopez de Lacalle, O.: UBC-ALM: Combining k-NN with SVD for WSD. In: In Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval 2007), Czech Republic, Prague, pp. 341–345 (2007)
2. 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: Carol, P., et al. (eds.) CLEF 2007. LNCS, vol. 5152, pp. 908–917. Springer, Heidelberg (2008)
3. Agirre, E., Di Nunzio, G.M., Ferro, N., Peters, C., Mandl, T.: CLEF 2008: Ad Hoc Track Overview. In: Carol, P., et al. (eds.) CLEF 2008. LNCS, vol. 5706, pp. 118–125. Springer, Heidelberg (2009)
4. Agirre, E., Otegi, A., Zaragoza, H.: Using Semantic Relatedness and Word Sense Disambiguation for (CL)IR. In: Peters, C., et al. (eds.) CLEF 2009 Workshop, Part I. LNCS, vol. 6241, pp. 166–173. Springer, Heidelberg (2010)
5. Basile, P., Caputo, A., Semeraro, G.: UNIBA-SENSE at CLEF 2009: Robust WSD task. In: Peters, C., et al. (eds.) CLEF 2009 Workshop, Part I. LNCS, vol. 6241, pp. 150–157. Springer, Heidelberg (2010)
6. Borges, T.B., Moreira, V.P.: UFRGS@CLEF2009: Retrieval by Numbers. In: Peters, C., et al. (eds.) CLEF 2009 Workshop, Part I. LNCS, vol. 6241, pp. 135–141. Springer, Heidelberg (2010)
7. Braschler, M., Peters, C.: CLEF 2003 Methodology and Metrics. In: Peters, C., Gonzalo, J., Braschler, M., Kluck, M. (eds.) CLEF 2003. LNCS, vol. 3237, pp. 7–20. Springer, Heidelberg (2004)
8. Chan, Y.S., Ng, H.T., Zhong, Z.: NUS-PT: Exploiting Parallel Texts for Word Sense Disambiguation in the English All-Words Tasks. In: Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval 2007), Prague, Czech Republic, pp. 253–256 (2007)
9. Cleverdon, C.: The Cranfield Tests on Index Language Devices. In: Sparck Jones, K., Willett, P. (eds.) Readings in Information Retrieval, pp. 47–59. Morgan Kaufmann Publisher, Inc., San Francisco (1997)
10. Crivellari, F., Di Nunzio, G.M., Ferro, N.: A statistical and graphical methodology for comparing bilingual to monolingual cross-language information retrieval. In: Agosti, M. (ed.) Information Access through Search Engines and Digital Libraries. Information Retrieval Series, vol. 22, pp. 171–188. Springer, Heidelberg (2008)
11. 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 2009 Workshop (2009), `http://www.clef-campaign.org/`

12. Guyot, J., Falquet, G., Radhouani, S.: UniGe at CLEF 2009 Robust WSD Task. In: Borri, F., Nardi, A., Peters, C. (eds.) Working Notes for the CLEF 2009 Workshop (2009), http://www.clef-campaign.org/
13. Kern, R., Juffinger, A., Granitzer, M.: Application of Axiomatic Approaches to Crosslanguage Retrieval. In: Peters, C., et al. (eds.) CLEF 2009 Workshop, Part I. LNCS, vol. 6241, pp. 142–149. Springer, Heidelberg (2010)
14. Paskin, N. (ed.): The DOI Handbook – Edition 4.4.1. International DOI Foundation, IDF (2006), http://dx.doi.org/10.1000/186
15. Mandl, T., Womser-Hacker, C., Di Nunzio, G.M., Ferro, N.: How Robust are Multilingual Information Retrieval Systems? In: Proceedings ACM SAC Symposium on Applied Computing (SAC), Fortaleza, Brazil, March 16-20, pp. 1132–1136 (2008)
16. Pradhan, S., Loper, E., Dligach, D., Palmer, M.: SemEval Task-17: English Lexical Sample, SRL and All Words. In: Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007), Prague, Czech Republic, pp. 87–92. Association for Computational Linguistics (2007)
17. Robertson, S.: On GMAP: and Other Transformations. In: Yu, P.S., Tsotras, V., Fox, E.A., Liu, C.B. (eds.) Proc. 15th International Conference on Information and Knowledge Management (CIKM 2006), pp. 78–83. ACM Press, New York (2006)
18. Voorhees, E.M.: The TREC Robust Retrieval Track. SIGIR Forum 39, 11–20 (2005)
19. Wolf, E., Bernhard, D., Gurevych, I.: Combining Probabilistic and Translation-Based Models for Information Retrieval based on Word Sense Annotations Information Retrieval. In: Peters, C., et al. (eds.) CLEF 2009 Workshop, Part I. LNCS, vol. 6241, pp. 120–127. Springer, Heidelberg (2010)
20. Zazo, A., Figuerola, C.G., Alonso Berrocal, J.L., Gomez, R.: REINA at CLEF 2009 Robust-WSD Task: Partial Use of WSD Information for Retrieval. In: Borri, F., Nardi, A., Peters, C. (eds.) Working Notes for the CLEF 2009 Workshop (2009), http://www.clef-campaign.org/