OVERVIEW OF THE CLEF 2008 MULTILINGUAL QUESTION ANSWERING TRACK

Pamela Forner[1], Anselmo Peñas[2], Iñaki Alegria[3], Corina Forăscu[4], Nicolas Moreau[5], Petya Osenova[6], Prokopis Prokopidis[7], Paulo Rocha[8], Bogdan Sacaleanu[9], Richard Sutcliffe[10], and Erik Tjong Kim Sang [11]

[1] CELCT, Trento, Italy (forner@celct.it)
[2] Departamento de Lenguajes y Sistemas Informáticos, UNED, Madrid, Spain (anselmo@lsi.uned.es)
[3] University of Basque Country, Spain (i.alegria@ehu.es)
[4] Al. I. Cuza, University of Iasi, Romania Institute for Computer Science, Romania (corinfor@info.uaic.ro)
[5] ELDA/ELRA, Paris, France (moreau@elda.org)
[6] BTB, Bulgaria, (petya@bultreebank.org)
[7] ILSP Greece, Athena Research Center (prokopis@ilsp.gr)
[8] Linguateca, DEI UC, Portugal, (Paulo.Rocha@di.uminho.pt)
[9] DFKI, Germany, (bogdan@dfki.de)
[10] DLTG, University of Limerick, Ireland (richard.sutcliffe@ul.ie)
[11] University of Groningen (e.f.tjong.kim.sang@rug.nl)

Abstract The QA campaign at CLEF [1], was manly the same as that proposed last year. The results and the analyses reported by last year's participants suggested that the changes introduced in the previous campaign had led to a drop in systems' performance. So for this year's competition it has been decided to practically replicate last year's exercise.
Following last year's experience some QA pairs were grouped in clusters. Every cluster was characterized by a topic (not given to participants). The questions from a cluster contained co-references between one of them and the others. Moreover, as last year, the systems were given the possibility to search for answers in Wikipedia[1] as document corpus beside the usual newswire collection.
In addition to the main task, three additional exercises were offered, namely the Answer Validation Exercise (AVE), the Question Answering on Speech Transcriptions (QAST), which continued last year's successful pilot, and Word Sense Disambiguation for Question Answering (QA-WSD).
As general remark, it must be said that the task still proved to be very challenging for participating systems. In comparison with last year's results the Best Overall Accuracy dropped significantly from 41,75% to 19% in the multi-lingual subtasks,

[1] `http://wikipedia.org`

while instead it increased a little in the monolingual sub-tasks, going from 54% to 63,5%.

1 Introduction

QA@CLEF 2008 was carried out according to the spirit of the campaign, consolidated in previous years. Beside the classical main task, three additional exercises were proposed:

- the *main* task: several monolingual and cross-language sub-tasks, were offered: Bulgarian, English, French, German, Italian, Portuguese, Romanian, Greek, Basque and Spanish were proposed as both query and target languages.
- the *Answer Validation Exercise* (AVE) [2]: in its third round was aimed at evaluating answer validation systems based on textual entailment recognition. In this task, systems were required to emulate human assessment of QA responses and decide whether an *Answer* to a *Question* is correct or not according to a given *Text*. Results were evaluated against the QA human assessments.
- the *Question Answering on Speech Transcripts* (QAST) [3,14]: which continued last year's successful pilot task, aimed at providing a framework in which QA systems could be evaluated when the answers to factual and definition questions must be extracted from spontaneous speech transcriptions.
- the *Word Sense Disambiguation for Question Answering* (QA- WSD) [4], a pilot task which provided the questions and collections with already disambiguated Word Senses in order to study their contribution to QA performance.

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

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

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

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

This paper describes the preparation process and presents the results of the QA track at CLEF 2008. In section 2, the tasks of the track are described in detail. The results are reported in section 3. In section 4, some final analysis about this campaign is given.

2 Task Description

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

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

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

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

The subtasks were both:

- monolingual, where the language of the question (Source language) and the language of the news collection (Target language) were the same;

- cross-lingual, where the questions were formulated in a language different from that of the news collection.

Two new languages have been added, i.e. Basque and Greek both as source and target languages. In total eleven source languages were considered, namely, Basque, Bulgarian, Dutch, English, French, German, Greek, Italian, Portuguese, Romanian and Spanish. All these languages were also considered as target languages.

Table **1.** Tasks activated in 2008 (coloured cells)

		TARGET LANGUAGES (corpus and answers)										
		BG	**DE**	**EL**	**EN**	**ES**	**EU**	**FR**	**IT**	**NL**	**PT**	**RO**
SOURCE LANGUAGES (questions)	**BG**	■										
	DE		■			■						
	EL			■								
	EN	■	■	■		■	■	■	■	■	■	■
	ES		■		■	■	■	■	■		■	
	EU				■	■	■		■			
	FR				■	■		■			■	
	IT				■	■		■	■			
	NL				■	■				■		
	PT				■	■		■	■		■	
	RO				■							■

As shown in Table 1, 43 tasks were proposed:

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

tered, some target language question sets were translated into the combined source languages).

Anyway, as Table 2 shows, not all the proposed tasks were then carried out by the participants.

Table 2. Tasks chosen by at least 1 participant in QA@CLEF campaigns

	MONOLINGUAL	CROSS-LINGUAL
CLEF-2004	6	13
CLEF-2005	8	15
CLEF-2006	7	17
CLEF-2007	7	11
CLEF-2008	**8**	**12**

As long-established, the monolingual English (EN) task was not available as it seems to have been already thoroughly investigated in TREC campaigns. English was still both source and target language in the cross-language tasks.

2.1 Questions Grouped by Topic

The procedure followed to prepare the test set was the same as that used in the 2007 campaign. First of all, each organizing group, responsible for a target language, freely chose a number of topics. For each topic, one to four questions were generated. Topics could be not only named entities or events, but also other categories such as objects, natural phenomena, etc. (e.g. George W. Bush; Olympic Games; notebooks; hurricanes; etc.). The set of ordered questions were related to the topic as follows:

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

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

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

The requirement for questions related to a same topic necessarily implies that the questions refer to common concepts and entities within the domain. The most common form is pronominal anaphoric reference to the topic declared in the first question, e.g.:

Q4: *What is a polygraph?*; Q5: *When was **it** invented?*

However, other forms of co-reference occurred in the questions. Here is an example:

Q6: *Who wrote the song "Dancing Queen"?*; Q7: *How many people were in* ***the group****?*

Here *the group* refers to an entity expressed not in the question but only in the answer. However the QA system does not know this and has to infer it, a task which can be very complex, especially if the topic is not provided in the test set.

2.2 Document collections

Beside the data collections composed of news articles provided by ELRA/ELDA (see Table 3), also Wikipedia was considered.

The Wikipedia pages in the target languages, as found in the version of November 2006, could be used. Romanian had Wikipedia[2] as the only document collection, because there was no newswire Romanian corpus. The "snapshots" of Wikipedia were made available for download both in XML and HTML versions. The answers to the questions had to be taken from actual entries or articles of Wikipedia pages. Other types of data such as images, discussions, categories, templates, revision histories, as well as any files with user information and meta-information pages, had to be excluded.

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

2.3 Types of Questions

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

1. *Factoid questions*, fact-based questions, asking for the name of a person, a location, the extent of something, the day on which something happened, etc. We consider the following 8 answer types for factoids:
   - PERSON, e.g.: Q8: *Who was called the "Iron-Chancellor"?* A8: *Otto von Bismarck.*

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

- TIME, e.g.: Q9: *What year was Martin Luther King murdered?* A9: *1968*.
- LOCATION, e.g.: Q10: *Which town was Wolfgang Amadeus Mozart born in?* A10: *Salzburg*.
- ORGANIZATION, e.g.: Q11: *What party does Tony Blair belong to?*: A11: *Labour Party.*
- MEASURE, e.g.: Q12: *How high is Kanchenjunga?* A12: *8598m.*
- COUNT, e.g.: Q13: *How many people died during the Terror of PoPot?* A13: *1 million.*
- OBJECT, e.g.: Q14: *What does magma consist of?* A14: *Molten rock.*
- OTHER, i.e. everything that does not fit into the other categories above, e.g.: Q15: *Which treaty was signed in 1979?* A15: *Israel-Egyptian peace treaty.*

Table 3. Document collections used in QA@CLEF 2008

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

2. *Definition questions*, questions such as “What/Who is X?”, and are divided into the following subtypes:

   - PERSON, i.e., questions asking for the role/job/important information about someone, e.g.: Q16*: Who is Robert Altmann?* A16*: Film maker*
   - ORGANIZATION, i.e., questions asking for the mission/full name/important information about an organization, e.g.: Q17: *What is the Knesset?* A17: *Parliament of Israel.*
   - OBJECT, i.e., questions asking for the description/function of objects, e.g.: Q18: *What is Atlantis?* A18: *Space Shuttle.*
   - OTHER, i.e., question asking for the description of natural phenomena, technologies, legal procedures etc., e.g.: Q19: *What is Eurovision?* A19: *Song contest.*

3. closed list questions: i.e., questions that require one answer containing a determined number of items, e.g.: Q20: *Name all the airports in London, England.* A20: *Gatwick, Stansted, Heathrow, Luton and City.*

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

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

> Q21: *Who was the Chancellor of Germany from 1974 to 1982?*
> A21: *Helmut Schmidt.*
>
> Q22: *Which book was published by George Orwell in 1945?*
> A22: *Animal Farm.*
>
> Q23: *Which organization did Shimon Perez chair after Isaac Rabin's death?*
> A23: *Labour Party Central Committee.*

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

The distribution of the questions among these categories is described in Table 4. Each question set was then translated into English, which worked as interlanguage during the translation of the datasets into the other tongues for the activated cross-lingual subtasks.

Table 4. Test set breakdown according to question type, number of participants and number of runs

	F	D	L	T	NIL	# Participants	# Runs
BG	159	24	17	28	9	1	1
DE	160	30	10	9	13	3	12
EL	163	29	8	31	0	0	0
EN	160	30	10	12	0	4	5
ES	161	19	20	42	10	4	10
EU	145	39	16	23	17	1	4
FR	135	30	35	66	10	1	3
IT	157	31	12	13	10	0	0
NL	151	39	10	13	10	1	4
PT	162	28	10	16	11	6	9
RO	162	28	10	47	11	2	4

2.4 Formats

As the format is concerned, also this year both input and output files were formatted as an XML file. For example, the first four questions in the EN-FR test set, i.e. English questions that hit a French document collection - were represented as follows:

```
 <input>
  <q target_lang="FR" source_lang="EN" q_id="0001"
     q_group_id="1600">Which is the largest bird in Africa?</q>
  <q target_lang="FR" source_lang="EN" q_id="0002"
     q_group_id="1600">How many species of ostriches are there?</q>
  <q target_lang="FR" source_lang="EN" q_id="0003"
     q_group_id="1601">Who served as a UNICEF goodwill ambassador be-
     tween 1988 and 1992?</q>
  <q target_lang="FR" source_lang="EN" q_id="0004"
     q_group_id="1601">What languages did she speak?</q>
...
 </input>
```

An example of system output which answered the above questions was the following:

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE output SYSTEM "QA-CLEF-OUTPUT.dtd">
<output>
```

```
<a q_id="0001" q_group_id="1600" run_id="syna081enfr" score="0.000">
<answer>version</answer>
<docid>Afrique des Grands Lacs</docid>
<support>
<s_id>Afrique des Grands Lacs</s_id>
<s_string>Comprendre la crise de l'Afrique des grands lacs - dossier
     RFI (version archivée par Internet Archive).</s_string>
</support>
</a>
<a q_id="0002" q_group_id="1600" run_id="syna081enfr" score="0.000">
<answer>500 000</answer>
<docid>ATS.940202.0138</docid>
<support>
<s_id>ATS.940202.0138</s_id>
<s_string>Avec une superficie de seulement 51 000 km2, le Costa Rica
     abrite quelque 500 000 espèces végétales et animales. Il compte
     plus d'espèces d'oiseaux et d'arbres qu'il n'y en a sur
     l'ensemble du territoire des Etats-Unis. </s_string>
</support>
</a>
<a q_id="0003" q_group_id="1601" run_id="syna081enfr" score="0.000">
<answer>NIL</answer>
<docid/>
<support>
<s_id/>
<s_string/>
</support>
</a>
<a q_id="0004" q_group_id="1601" run_id="syna081enfr" score="0.000">
<answer>NIL</answer>
<docid/>
<support>
<s_id/>
<s_string/>
</support>
</a>
...
</output>
```

2.5 Evaluation

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

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

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

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

In addition most assessor groups computed the following measures:

- Confident Weighted Score (CWS). Answers are in a decreasing order of confidence and CWS rewards systems that give correct answers at the top of the ranking [16]
- the Mean Reciprocal Rank (MRR) over N assessed answers per question (to consider the three answers). That is, the mean of the reciprocal of the rank of the first correct label over all questions. If the first correct label is ranked as the 3rd label, then the reciprocal rank (RR) is 1/3. If none of the first N responses contains a correct label, RR is 0. RR is 1 if the highest ranked label matches the correct label.

3 Results

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

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

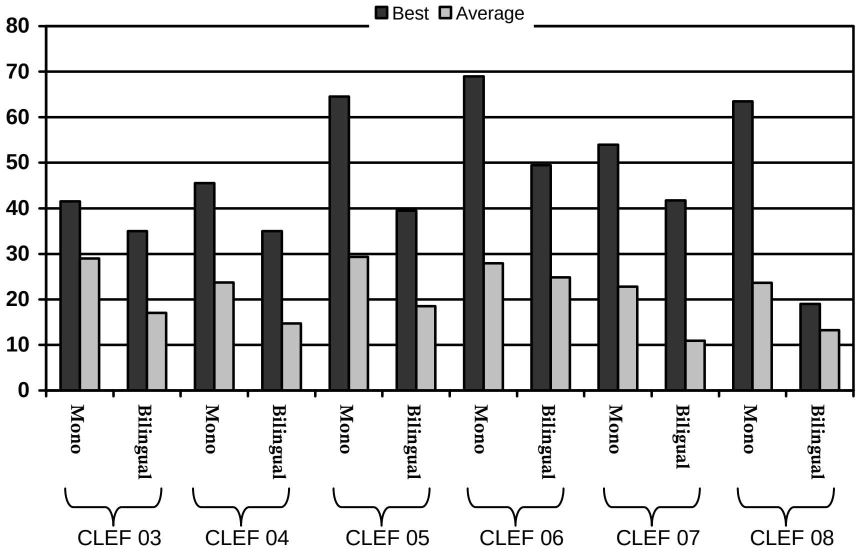

Figure 1. Best and average scores in QA@CLEF campaigns

On the contrary, Best accuracy over the bilingual tasks, decreased considerably. This is also true for average performances. This year a small increase was recorded in the bilingual tasks but it seems that the high level of difficulty of the question sets particularly impacted the bilingual tasks and the task proved to be still difficult also for veterans.

3.1 Participation

Table 5. Number of participants in QA@CLEF

	America	Europe	Asia	Australia	TOTAL
CLEF 2003	3	5	0	0	8
CLEF 2004	1	17	0	0	18
CLEF 2005	1	22	1	0	24
CLEF 2006	4	24	2	0	30
CLEF 2007	3	16	1	1	21
CLEF 2008	**1**	**20**	**0**	**0**	**21**

The number of participants has remained almost the same as in 2007 (see Table 5). As noticed, this is probably the consequence of the new challenges introduced last year in the exercise.

Also the geographical distribution remained almost unchanged, even though there was no participation from Australia and Asia. No runs were submitted neither for Italian or Greek tasks.
Anyway, the number of submitted runs, increased from a total of 37 registered last year to 51 (see Table 6). The breakdown of participants and runs, according to language, is shown in Table 4 (Section 2.3). As in previous campaigns, more participants chose the monolingual tasks, which once again demonstrated to be more approachable.

Table 6. Number of submitted runs

	Submitted runs	Monolingual	Cross-lingual
CLEF 2003	17	6	11
CLEF 2004	48	20	28
CLEF 2005	67	43	24
CLEF 2006	77	42	35
CLEF 2007	37	23	14
CLEF 2008	**51**	**31**	**20**

In the following subsections a more detailed analysis of the results in each language follows, giving specific information on the performances of the participating systems in the single sub-tasks and on the different types of questions, providing the relevant statistics and comments.

3.2 Basque as target

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

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

The news were from the Egunkaria newspaper during 2000, 2001 and 2002 years and the information from Wikipedia was the exportation corresponding to the 2006 year.

Table 7 shows the evaluation results for the four submitted runs (one monolingual and three cross-lingual). The table shows the number of Right, Wrong, ineXact and Unsupported answers, as well as the percentage of correctly answered Factoids, Temporally restricted questions, Definition and List questions.

Table 7. Evaluation results for the four submitted runs.

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

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

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

3.3 Bulgarian as Target

Table 8. Results for the submitted run for Bulgarian

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

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

3.4 Dutch as Target

The questions for the Dutch subtask of CLEF-QA 2008 were written by four native speakers. They selected random articles from either Wikipedia or the news collection and composed questions based on the topics of the articles.

Table 9. Properties of the 200 Dutch questions (134 topics) in the test set

Question types	
Definition	39
Factoid	151
List	
Answer source	
News	20
None (NIL answer)	5
Wikipedia	175
Definition answer types	
Location	3
Object	6
Organization	8
Other	12
Person	10

Factoid answer types	
Count	20
Location	18
Measure	20
Object	19
Organization	18
Other	17
Person	19
Time	20
List answer types	
Location	6
Other	1
Person	2
Time	1

Temporal restriction	
No	187
Yes	13
Question per topic	
1 question	100
2 questions	15
3 questions	6
4 questions	13
Topic types	
Location	15
Object	23
Organization	14
Other	50
Person	32

The quartet produced a total of 222 question-answer pairs from which they selected a set of 200 that satisfied the type distribution requirements of the task organizers. An overview of the question types and answer types can be found in Table 9.

This year, only one team took part in the question answering task with Dutch as target language: the University of Groningen. The team submitted two monolingual runs and two cross-lingual runs (English to Dutch). All runs were assessed twice by a single assessor. This resulted in a total of eight conflicts (1%). These were corrected. The results of the assessment can be found in Table 10.

Table 10. Assessment results for the four submitted runs for Dutch.

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

The two cross-lingual runs gron081ennl andron082ennl produced exactly the same answers.

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

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

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

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

3.5 English as Target

Table 11. Evaluation results for the English submitted runs.

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

* Total number in the test set.

Creation of Questions. The task this year was exactly the same as in 2007 and moreover the three collections were the same: Glasgow Herald, LA Times and Wikipedia. However, given the considerable interest in the Wikipedia which has been shown by Question Answering groups generally, it was decided to increase the number of questions drawn from it to 75% overall, with just 25% coming from the two newspaper collections. This means that 40 of the 160 Factoids came from the newspapers, together with seven of the 30 Definitions and two of the ten Lists. These questions were divided equally between the Glasgow Herald and LA Times. All the remainder we drawn from the Wikipedia.

Considerable care was taken in the selection of the questions. The distribution by answer type was controlled exactly as in previous years. As requested by the organisers there were exactly twenty each of Factoid target type PERSON, TIME, LOCATION, MEASURE, COUNT, ORGANIZATION, OBJECT and OTHER. Similarly for Definitions there were eight PERSON, seven ORGANIZATION, seven OBJECT and eight OTHER. For Lists there were four OTHER, two each of PERSON and ORGANIZATION, and one each of LOCATION and OBJECT.

In addition to the above distribution, we also controlled the distribution of topics for the question groups, something which was made practicable by the use of the Wikipedia. Questions were drawn from a number of predefined subject fields:

countries towns, roads and bridges, shops, politicians and politics, sports and sports people, foods and vegetables, cars, classical music including instruments, popular music, literature poetry and drama, philosophy, films, architecture, languages, science, consumer goods, and finally organisations. Questions were distributed among these topics. The maximum in any topic was twenty (sports) and the minimum was two (shops). For the majority there were between four and six question groups. For each such topic, one or more questions were set depending on what information the texts contained. As a change from last year, the organisers asked us to include 100 singleton topics. This effectively meant that half the questions in the overall set of 200 were simple "one-off" queries as were set in CLEF prior to 2007 and for the earlier TREC campaigns.

Questions were entered via a web interface developed by the organisers last year. However, this year they improved it considerably, for example allowing modifications to be made to existing entries. This was a great help and a commendable effort on their part.

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

Assessment Procedure. Last year we used the excellent Web-based assessment system developed originally for the QiQA task by University of Amsterdam. However, we were asked not to use this in 2008 because it only allows one answer per question per system to be assessed and it was required to assess multiple answers per question per system. For this reason we used a Web-based tool developed by UNED in Madrid.

All answers were double-judged. Where the assessors differed, the case was reviewed and a decision taken. There were 63 judgement differences in total. Three of the runs contained multiple answers to individual questions in certain cases, and these were all assessed, as per the requirement of the organisers. If we assume that the number of judgements was in fact 200 questions * five runs, i.e. 1,000, we can compute a lower bound for the agreement level. This gives a figure of (1,000-63)/1,000, i.e. 93.7%. The equivalent figure for 2007 (called Agreement Level 2 in the Working Notes for last year) was 97.6%. Given that we have computed a lower bound this year (and not therefore the exact figure) this seems acceptable.

Results Analysis. Of the five runs with English as target, wlvs081roen was the best with an accuracy of 19.00% overall. They also did very will on the definitions, scoring 66.67%. The only source language for which there was more than one run was German, for which there were three submissions from two groups. dfki081 scored the best with 14.00% and this was followed by dcun081deen with

8.00% and dcun082deen with 0.50%. dfki also did very well on definitions with an accuracy of 60.00. Interestingly, none of the systems answered any of the list questions correctly. Only dcun082deen answered one list question inexactly.

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

3.6 QA-WSD subtask

The QA-WSD task brings semantic and retrieval evaluation together. The participants were offered the same queries and document collections as for the main QA exercise, but with the addition of word sense tags as provided by two automatic word sense disambiguation (WSD) systems. Contrary to the main QA task, Wikipedia articles are not included, and thus systems need to reply to the questions that have an answer in the news document collection. The goal of the task is to test whether WSD can be used beneficially for Question Answering, and is closely related to the Robust-WSD subtask of the ad-hoc track in CLEF 2008.

The exercise scenario is event-targeted QA on a news document collection. In the QA-WSD track only English monolingual and Spanish to English bilingual tasks are offered, i.e. English is the only target language, and queries are available on both English and Spanish. The queries were the same as for the main QA exercise, and the participation followed the same process, except for the use of the sense-annotated data.

The goal of this task is to evaluate whether word sense information can help in certain queries. For this reason, participants were required to send two runs for each of the monolingual/bilingual tasks where they participate: one which does not use sense annotations and another one which does use sense annotations. Whenever possible, the only difference between the two runs should be solely the use or not of the sense information. Participants which send a single run would be discarded from the evaluation.

The WSD data is based on WordNet version 1.6 and was supplemented with freely available data from the English and Spanish WordNets in order to test different expansion strategies. Two leading WSD experts run their systems [17][18], and provided those WSD results for the participants to use.

The task website [4] provides additional information on data formats and resources.

Results

From the 200 questions provided to participants, only 49 queries had a correct answer in the news collection. The table below provides the results for the participant on those 49 questions.

Table 12. Results of the EN2EN QA-WSD runs on the 49 queries which had replies in the news collections

Run	R	W	X	U	%F	%T	%D	L%	NIL		CWS	Over-all accu-racy
	#	#	#	#	[40]	[5]	[7]	[2]	0	% [0]		
nlel08 1enen	8	41	0	0	17.5	0	14.2	0	0	0	0.03	16.32
nlel08 2enen	7	42	0	0	15.0	0	0	0	0	0	0.02	14.29

The first run does not use WSD, while the second uses the sense tags returned by the NUS WSD system. The WSD tags where used in the passage retrieval module. The use of WSD does not provide any improvement, and causes one more error. For the sake of completeness we also include below the results on all 200 queries. Surprisingly the participant managed to find two (one in the WSD run) correct answer for the Wikipedia questions in the news collection.

Table 13. Results of the EN2EN QA-WSD runs on all 200 queries, just for the sake of comparison

Run	R	W	X	U	%F	%T	%D	L%	NIL		CWS	Over-all accu-racy
	#	#	#	#	[160]	[5]	[7]	[10]	0	% [0]		
nlel08 1enen	10	188	0	2	5.6	0	3.3	0	0	0	0.00	5.00
nlel08 2enen	8	189	0	3	4.4	0	3.3	0	0	0	0.00	4.00

3.7 French as Target

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

Synapse submitted three runs in total:

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

In the following, these will be referred to as:

- syn08frfr (for FR-to-FR),
- syn08enfr (for EN-to-FR),
- syn08ptfr (for PT-to-FR).

As last year, three types of questions were proposed: factual, definition and closed list questions. Participants could return one exact answer per question and up to two runs. Some questions (10%) had no answer in the document collection, and in this case the exact answer is "NIL".

The French test set consists of 200 questions:

- 135 Factual (F),
- 30 Definition (D),
- 35 closed List questions (L).

Among these 200 questions, 66 were temporally restricted questions (T) and 12 were NIL questions (i.e. a "NIL" answer was expected, meaning that there is no valid answer for this question in the document collection).

Table 14. Results of the monolingual and bilingual French runs.

Run	Assessed Answers (#)	R #	W #	X #	U #	%F [135]	%T [66]	%D [30]	L% [35]	NIL Answers		CWS	Overall accuracy
										#	% [12]		
syn08frfr	200	131	77	9	1	54.8	51.5	86.7	37.1	20	50.0	0.30937	56.5
syn08enfr	200	36	157	6	1	15.6	15.1	50.0	0.0	60	8.3	0.02646	18.0
syn08ptfr	200	33	163	4	0	14.1	13.6	43.3	2.9	67	11.9	0.02387	16.5

Table 14 shows the final results of the assessment of the 3 runs submitted by Synapse. For each run, the following statistics are provided:

- The number of correct (R), wrong (W), inexact (X) and unsupported answers (U),
- The accuracy calculated within each of the categories of questions: F, D, T and L questions,
- The number of NIL answers and the proportion of correct ones (i.e. corresponding to a NIL questions),
- The Confidence Weighted Score (CWS) measure.
- The accuracy calculated over all answers.

Figure 2 shows the best scores for systems using French as target in the last five CLEF QA campaigns.

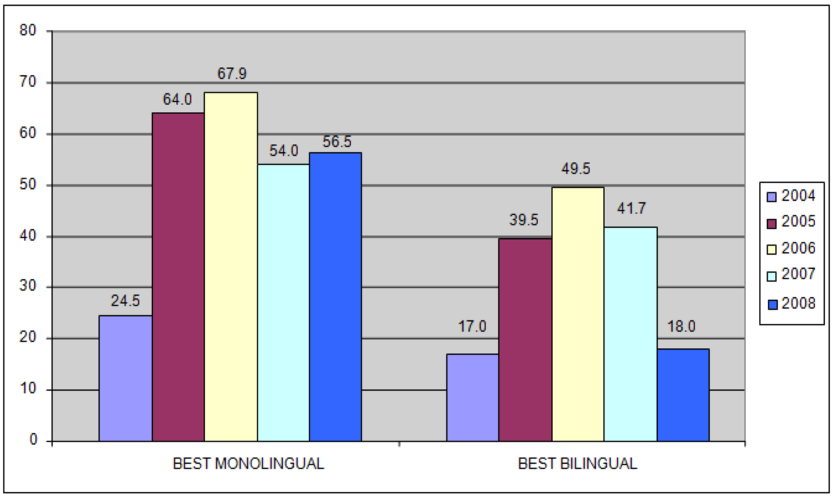

Figure 2: Best scores for systems using French as target in CLEF QA campaigns

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

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

It is even more obvious when looking at the bilingual runs. If the systems performed pretty well on the definition questions (50.0% and 43.3% for EN-to-FR and PT-to-FR respectively), they could not cope with the closed list questions. The

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

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

It is also interesting to look at the results when categorizing questions by the size of the topic they belong to. This year, topics could contain from 1 single question to 4 questions. The CLEF 2008 set consists of:

- 52 single question topics,
- 33 topics with 2 questions (66 questions in total),
- 18 topics with 3 questions (54 questions in total),
- 7 topics with 4 questions (28 questions in total).

Table 15, Table 16 and Table 17 give the results of each run according to the size of the topics.

Table 15. Results per topic size (FR-to-FR)

Run	**Size of topic**	**Assessed Answers #**	**Overall accuracy (%)**
syn08frfr	1	52	55.8
syn08frfr	2	66	50.0
syn08frfr	3	24	66.7
syn08frfr	4	28	53.6

Table 16. Results per topic size (EN-to-FR)

Run	**Size of topic**	**Assessed Answers #**	**Overall accuracy (%)**
syn08enfr	1	52	21.2
syn08enfr	2	66	22.7
syn08enfr	3	24	13.0
syn08enfr	4	28	10.7

Table 17. Results per topic size (PT-to-FR)

Run	Size of topic	Assessed Answers #	Overall accuracy (%)
syn08ptfr	1	52	25.0
syn08ptfr	2	66	18.2
syn08ptfr	3	24	9.3
syn08ptfr	4	28	10.7

The monolingual system (Table 15) is not sensitive to the size of the topic question set. On the opposite, the performances of the bilingual systems (Table 16 and Table 17) decrease by a half, when comparing the 1- and 2-question sets to the 3- and 4-question sets. A possible explanation is that the bilingual systems perform poorly with questions containing anaphoric references (which are more likely to occur in the 3- and 4-question sets).

In conclusion, there was unfortunately only one participant this year. In particular; it would have been interesting to see how the LCC group, which submitted a bilingual run last year, would have performed this year.

This decrease in participation can be explained by the discouragement of some participants. Some have complained that the task is each year harder (e.g. this year, there were more closed list questions and anaphoric references than last year) that can result in a decrease in the systems performances.

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

3.8 German as Target

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

[redacted]

Figure 3. Results evolution

Compared to the previous editions of the evaluation forum, this year an increase in the accuracy of the best performing system and of an aggregated virtual system for monolingual and a decrease in the accuracy of the best performing system and of an aggregated virtual system for cross-language tasks was registered.

Table 18. Topic distribution over data collections

Topic Size	**# Topics / CLEF**	**# Topics / WIKI**	**# Topics**
1	39	35	74
2	10	14	24
3	5	5	10
4	3	9	12
Total	**57**	**63**	**120**

Table 19. Topic type breakdown over data collections

	CLEF					**WIKI**				
Topic Type	**Topic Size**				**Total**	**Topic Size**				**Total**
	1	2	3	4		1	2	3	4	
PERSON	5	2	1	1	9	0	1	0	2	3
OBJECT	7	1	0	0	8	16	3	0	2	21
ORGANIZATION	9	1	2	1	13	7	2	1	1	11
LOCATION	8	2	2	1	13	1	3	2	2	8
EVENT	0	0	0	0	0	0	2	0	0	2
OTHER	9	4	0	1	14	11	3	2	2	18
					57					63

The number of topics covered by the test set questions was of 120 distributed as it follows: 74 topics consisting of 1 question, 24 topics of 2 related questions, 10 topics of 3 related questions, and 12 topics of 4 related questions. The distribution of the topics over the document collections (CLEF vs. Wikipedia) is presented in Table 18.

Table 20. Question EAType breakdown over data collections

EAType	CLEF	WIKI	Total
PERSON	15	15	30
LOCATION	13	12	25
TIME	13	8	21
COUNT	13	7	20
OBJECT	7	18	25
MEASURE	12	8	20
ORGANIZATION	15	13	28
OTHER	9	22	31
Total	97	103	200

The details of systems' results can be seen in Table 21.

Table 21. System Performance – Details

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

According to Table 19 the most frequent topic types were OTHER (32), OBJECT (29) and ORGANIZATION (24), with first two types more present for the Wikipedia collection of documents (WIKI).

As regards the source of the answers, 97 questions from 57 topics asked for information out of the CLEF document collection and the rest of 103 from 63 topics for information from Wikipedia. Table 20 shows a breakdown of the test set questions by the expected answer type (EAType) for each collection of data.

3.9 Portuguese as Target

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

In this fourth year of Portuguese participation, Priberam repeated the top place of its previous years, with University of Évora behind. Again we added the classification the classification X-, meaning incomplete, keeping the classification X+ for answers with extra text or other kinds of inexactness. In Table 22 we present the overall results (all tables in these notes refer exclusively to the first answer by each system).

Table 22: Results of the runs with Portuguese as target: all 200 questions (first answers only)

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

To provide a more direct comparison with pre-2006 results, in Table 23 we present the results both for first question of each topic (which we believe is more readily comparable to such results) and for the linked questions.

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

Table 23. Results of the runs with Portuguese as target: answers to linked and unlinked questions

Run Name	First questions (# 151)						Linked questions (# 49)	
	R (#)	**W** (#)	**X+** (#)	**X-** (#)	**U** (#)	**Accuracy** (%)	**R** (#)	**Accuracy** (%)
diue081	82	59	6	3	1	54.3	11	22.4
esfi081	42	92	5	7	5	27.3	7	14.3
esfi082	33	97	6	9	6	21.9	8	16.3
feup081	29	116	2	2	2	19.2	3	6.1
feup082	25	120	3	1	2	16.6	3	6.1
idsa081	54	85	6		6	35.8	11	22.4
ines081	35	106	2	3	5	23.2	8	16.3
ines082	35	106	2	3	5	23.2	8	16.3
prib081	105	32	9	4	1	69.5	22	44.9

Table 24. Results of the assessment of the monolingual Portuguese runs: definitions

Run	loc	obj	org	oth	per	TOT	%
	1	6	6	8	6	27	
diue081		5	6	8	5	24	89%
esfi081		1	2	4	2	9	33%
esfi082				1	1	2	7%
feup081		1	1	1	1	4	15%
feup082		1	1	1	1	4	15%
idsa081	1	5	1	5	5	17	63%
ines081	1	5	1	7	3	17	63%
ines082	1	5	1	7	3	17	63%
prib081		5	5	6	2	18	67%
combination	1	6	6	8	6	27	100%

Unlike last year , the results over linked questions are significatively different (and below) from those over not-linked. Question 180 was wrongly redacted, referring to Aida's opera *Verdi* instead of the other way around, which also affected two linked questions. Therefore, we accepted both NIL answers to those questions, as well as correct ones.

Table 24 shows the results for each answer type of definition questions, while Table 25 shows the results for each answer type of factoid questions (including list questions). As it can be seen, four out of six systems perform clearly better when it comes to definitions than to factoids. Particularly Senso has a high accuracy regarding definitions.

Table 25. Results of the assessment of the Portuguese runs: factoids, including lists.

Run	**cou**	**loc**	**mea**	**obj**	**org**	**oth**	**per**	**tim**	**TOT**	**%**
	17	38	16	2	10	33	33	24	173	
diue081	6	17	8	1	5	13	8	11	69	35%
esfi081	8	8	2		2	2	14	4	40	20%
esfi082	8	8	2		2	2	13	4	39	20%
feup081	5	4	4		1	2	8	4	28	14%
feup082	5	3	4		1	2	6	3	24	12%
idsa081	9	9	9			6	8	7	48	24%
ines081	4	9	2			1	4	6	26	13%
ines082	4	9	2			1	4	6	26	13%
prib081	11	21	13	1	7	18	22	16	109	55%
combination	16	31	15	1	7	23	27	21	141	82%

We included in both Table 24 and Table 25 a virtual run, called combination, in which one question is considered correct if at least one participating system found a valid answer. The objective of this combination run is to show the potential achievement when combining the capacities of all the participants. The combination run can be considered, somehow, state-of-the-art in monolingual Portuguese question answering. All definition questions were answered by at least one system.

Table 26. Average size of answers (values in number of words)

Run name	**Non-NIL Answers (#)**	**Average answer size**	**Average answer size (R only)**	**Average snippet size**	**Average snippet size (R only)**
diue081	179	2.8	3.6	25.9	26.1
esfi081	180	2.6	3.0	78.4	62.5
esfi082	180	1.8	1.7	78.2	62.4
feup081	58	1.8	3.4	64.2	51.6
feup081	51	1.8	3.7	63.3	51.4
idsa081	188	5.0	10.0	28.6	34.4
ines081	77	3.0	7.4	79.6	36.6
ines082	77	3.0	7.4	79.6	36.6
prib081	192	3.2	3.4	27.6	25.1

The system with best results, Priberam, answered correctly 64.8% the questions with at least one correct answer. In all, 130 questions were answered by more than one system.

In Table 26, we present some values concerning answer and snippet size.

Temporally restricted questions: Table 27 presents the results of the 17 temporally restricted questions. As in previous years, the effectiveness of the systems to answer those questions is visibly lower than for non-TRQ questions.

Table 27. Accuracy of temporally restricted questions.

Run name	Correct answers (#)	T.R.Q correctness (%)	Non-T.R.Q correctness (%)	Total correctness (%)
diue081	4	23.5	48..6	46.5
esfi081	3	17.6	24.0	23.5
esfi082	3	17.6	19.7	19.5
feup081	1	5.9	15.3	14.5
feup082	1	5.9	13.1	12.5
Idsa081	2	11.8	34.4	32.5
ines081	1	5.9	21.3	20.0
ines082	1	5.9	21.3	20.0
prib081	8	47.1	65.0	63.5

List questions: ten questions were defined as list questions all closed list factoids with two to five each[3]. The results haven't improved with UE getting two correct answers. Priberam three and all other system zero. There were however seven cases of incomplete answers (i.e.. answering some elements of the list only) although only two of them with than one element of the answer.

Table 28. Answers by source and their correctness

Run	News		Wikipedia		NIL	
	#	% correct	#	% correct	#	% correct
Selection	34	-	144	-	10	-
diue081	35	40%	144	53%	21	10%
esfi081	85	21%	95	28%	20	10%
esfi082	81	17%	99	24%	20	5%
feup081	10	40%	48	33%	142	6%
feup082	9	44%	42	29%	149	6%
idsa081	50	28%	138	36%	12	17%
ines081	31	23%	46	52%	123	7%
ines082	31	23%	46	52%	123	7%
prib081	46	63%	146	66%	8	13%

[3] There were some open list questions as well, but they were classified and evaluated as ordinary factoids.

Answer source: Table 28 presents the distribution of questions by source during their selection. The distribution of sources used by the different runs and their correctness.

3.10 Romanian as Target

In the third year of Romanian participation in QA@CLEF, and the second one with Romanian addressed as a target language, the question generation was based on the collection of Wikipedia Romanian pages frozen in November 2006[4]- the same corpus as in the previous edition[5].

Creation of Questions. The questions were generated starting from the corpus and based on the Guidelines for Question Generation[6], the Guidelines for Participants[7] and the final decisions taken after email discussions between the organizers. The 200 questions are distributed according to Table 29, where for each type of question and expected answer we indicate also the temporally restricted questions out of the total number of questions. Without counting the NIL questions, 100% of the questions has the answer in Wikipedia collection.

Table 29. Question & Answer types distribution in Romanian (in brackets the number of temporally restricted questions)

Q type /expected A type	PER SON	TIM E	LOC.	ORG.	MEAS URE	COU NT	OBJE CT	OTH ER	TOTAL
FACTOID	20 (9)	23 (5)	26 (4)	20 (10)	17 (3)	22 (5)	18 (4)	16 (4)	**162 (44)**
DEF.	8		1	6 (2)			6	7	**28 (2)**
LIST	3		1 (1)	1			2 (1)	3	**10 (2)**
NIL									**8**

[4] http://static.wikipedia.org/downloads/November_2006/ro/

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

[6] http://celct.isti.cnr.it/ClefQA/QA@CLEF08_Question_Generation_Guidelines.pdf

[7] http://nlp.uned.es/clef-qa/QA@CLEF08_Guidelines-for-Participants.pdf

As the Guidelines for Question Generation did not change since the previous edition, there were no major difficulties in creating the Romanian gold standard for the 2008 QA@CLEF. The working version of the GS was uploaded on the question generation interface developed at CELCT (Italy), by filling all the required fields.

For the topic-related questions (clusters of up to four questions, related to one same topic) we kept about the same number as in the previous edition: in 2007 we had 122 topics and now there are 119 topics. The percentage of topic-linked questions is illustrated in Table 30, showing that 127 questions were grouped under 46 topics, hence 63.5% out of the total 200 questions were linked in topics with more than one question.

Table 30. Topic-related questions

# of questions / Topic type	PERSO N	LOC.	ORG.	EVEN T	OBJE CT	OTHE R	**Total topics**	**Total ques-tions**
4 Qs	5	1	1			5	**12**	**48**
3 Qs	5	1		1	1	3	**11**	**33**
2 Qs	5	3	4		2	9	**23**	**46**
1 Q	13	6	19		17	18	**73**	**73**
TOTAL	**28**	**11**	**24**	**1**	**20**	**35**	**119**	**200**

In fact the questions contain not 127, but only 51 anaphoric elements of various types, so that 25.5% of the questions are linked through coreferential relations. The personal, possessive or demonstrative pronouns were used in most of the cases to create anaphoric relations. The antecedents are mainly the focus of the previous question, or the previous answer. Few such questions require inference in order to be correctly answered. For example in order to correctly answer the F-Time question *When was the first Esperanto dictionary for Romanian published?* and then the L-Other *Name all the grammatical cases of this artificial language.*, one needs to correctly link the anaphor "artificial language" to its antecedent which is "Esperanto" and not "Romanian" (also a language but not artificial); this is possible by establishing, based on a text snippet, that Esperanto is an artificial language.

The 8 NIL questions, even though they seem somehow unnatural, were created by including questions about facts impossible from a human perception; for example the question *In which year did Paul Kline publish his work about the natural phenomena called hail?* has no answer in any of the articles about the psychologist. Another type of NIL questions are those based on inference – the question *How many bicameral Parliaments are there in Cuba?* is a NIL question because in all wiki articles one can find that Cuba has a unicameral parliament. Another type of NIL questions (with answer in English, but not in Romanian) we have created cannot be good items neither in a cross-lingual evaluation where the answers are to be find in any language, nor in an evaluation based on an open text collection such as the web. The question *What is a micron?* has no answer in the Romanian wiki articles from 2006, but it can have an answer in other Romanian webpages, and, moreover, in the English wiki articles it has more than a correct answer depending on the domain where the term is used (in the metric system or in vacuum engineering).

For the LIST type we created only questions whose answers are to be found in one same text section. The 2007 evaluation for Romanian showed that "open list" questions (with answers in various sections of an article or even in various articles) are difficult to handle, therefore we made the LIST questions easier.

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

The statistics includes a system, named *combined*, obtained through the combination of the 4 participating RO-RO systems. Because at the evaluation time we observed that there are correct answers not only in the first position, but also on the second or the third, the *combined* system considers that an answer is R if there exists at least one R answer among all the answers returned by the four systems. If there is no R answer, the same strategy is applied to X, U and finally W answers. This "ideal" system permits to calculate the percentage of the questions (and their type), answered by at least one of the four systems in any of the maximum 3 answers returned for a question.

All three systems crashed on the LIST questions. The best results were obtained by ICIA for DEFINITION questions, whereas UAIC performed best with the FACTOID questions. The *combined* system suggests that a joint system, developed by both groups, would improve substantially the general results for Romanian.

Using in a first stage the web interface for assessing the QA runs, developed at UNED in Spain, the assessment took into consideration one question with all its answers at the time, assuring that the same evaluation criteria are applied to all answers. The judgment of the answers was based on the same Guidelines as in 2007, therefore we kept the same criteria as in 2007, in order to assure consistency inside the Romanian language, which gives also the possibility to evaluate the systems in their evolution from one year to another. For example, one could easily see that the UAIC systems had most of the answers for the DEFINITION questions evaluated as ineXact, because the answers were judged as being "longer than the minimum amount of information required" and hence "unnecessary pieces of information were penalized". Since all the 2007 and 2008 answers were evaluated this way, we considered it is more important to have uniformly applied rules inside one language than to change the evaluation in order to be consistent across languages. On the other hand the ICIA answers judged as ineXact are due to answers that are too long, snippets shortened as such as they do not contain the answer, or because the answer and the snippet has no connections.

Tables 31. Results in the monolingual task, Romanian as target language

Run	R	W		U	F	T	D	L	NIL		CWS	MRR	Overall accuracy
	#	#		#	[162]	[47]	[28]	[10]	#	% [8]			
icia081roro	10	179	1	0	4.938	8.511	7.143	0.0	15	6.667	0.00812	0.08217	5.0
icia082roro	21	168	1	0	6.173	8.511	39.286	0.0	15	6.667	0.02191	0.14319	10.5
uaic081roro	41	128	7	3	24.691	25.532	3.571	0.0	65	7.692	0.03679	0.34324	20.5
uaic082roro	45	125	6	4	26.543	27.660	3.571	10.0	64	9.375	0.04892	0.36799	22.5

Run	FACTOID QUESTIONS					LIST QUESTIONS					DEFINITION QUESTION				
	R	W	X	U	ACC	R	W	X	U	ACC	R	W	X	U	ACC
Combined	*72*	*75*	*12*	*3*	*44.444*	*1*	*9*	*0*	*0*	*10.000*	*14*	*5*	*10*	*0*	*50.000*
icia081roro	8	144	10	0	4.938	0	10	0	0	0.000	2	25	1	0	7.143
icia082roro	10	143	9	0	6.173	0	10	0	0	0.000	11	15	2	0	39.286
uaic081roro	40	113	6	3	24.691	0	9	1	0	0.000	1	6	21	0	3.571
uaic082roro	43	110	5	4	26.543	1	9	0	0	10.000	1	6	21	0	3.571

The evaluation was made more difficult because two of the submitted runs contain the answers in a totally arbitrary order, with topic-related questions having their answers in various parts of the submitted file. If in the first stage the UNED interface was of a great help, after the xml file was generated with all the evaluations, the corrections needed a thorough manual inspection. Anyway it was nice to find out that the answer to the question *Which terrorist organization does Osama bin Laden belong to?* is *Pentagon.*

3.11 Spanish as Target

The participation at the Spanish as Target subtask has decreased from 5 groups in 2007 to 4 groups this year. 6 runs were monolingual and 3 runs were crosslingual. Table 32 shows the summary of systems results with the number of Right (R), Wrong (W), Inexact (X) and Unsupported (U) answers. The table shows also the accuracy (in percentage) of factoids (F), factoids with temporal restriction (T), definitions (D) and list questions (L). Best values are marked in bold face.

Table 32. Results for Spanish as target

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

Table 33. Results for self-contained and linked questions, compared with overall accuracy

Run	% Accuracy over Self-contained questions [139]	% Accuracy over Linked questions [61]	% Overall Accuracy [200]
prib081eses	53,24	18,03	42,50
inao082eses	25,18	13,11	22,00
inao081eses	25,18	9,84	21,00
qaua082eses	22,30	13,11	19,50
mira081eses	21,58	3,28	16,00
mira082eses	21,58	3,28	14,50
qaua081enes	17,27	-	12,50
qaua082enes	12,23	1,64	9,00
mira081fres	6,47	1,64	5,00

Table 33 shows that the first question of the topic group is answered much more easily than the rest of the questions which need to solve some references to previous questions and answers.

Regarding NIL questions, Table 34 shows the harmonic mean (F) of precision and recall for self-contained questions, linked questions and all questions, taking into account only the first answer. In most of the systems, NIL is not given as second or third candidate answer.

Table 34. Results for Spanish as target for NIL questions

	F-measure (Self-contained@1)	**F-measure (@1)**	**Precision (@1)**	**Recall (@1)**
prib081eses	0,26	0,17	0.12	0.30
inao082eses	0,14	0.10	0.06	0.40
inao081eses	0,19	0.13	0.08	0.30
qaua082eses	0,27	0.15	0.09	0.60
mira081eses	0,27	0.21	0.17	0.30
mira082eses	0,29	0.23	0.19	0.30
qaua081enes	0,26	0.19	0.11	0.80
qaua082enes	0,20	0.15	0.09	0.60
mira081fres	0,15	0.12	0.07	0.30

The correlation coefficient r between the self-score and the correctness of the answers (shown in Table 34) has been similar to the obtained last year, being not good enough yet, and explaining the low results in CWS and K1 [6] measures.

Table 35. Answer extraction and correlation coefficient (r) for Spanish as target

Run	**%Answer Extraction**	**r**
prib081eses	90,53	0,4006
mira082eses	80,56	0,0771
inao082eses	80,00	0,1593
mira081eses	80,00	0,0713
qaua082eses	73,58	0,2466
inao081eses	67,74	0,1625
qaua081enes	75,76	0,0944
qaua082enes	58,06	0,0061
mira081fres	55,56	0,0552

Since a supporting snippet is requested in order to assess the correctness of the answer, we have evaluated the systems capability to extract the answer when the snippet contains it. The first column of Table 35 shows the percentage of cases where the correct answer was present in the snippet and correctly extracted. This information is very useful to diagnose if the lack of performance is due to the passage retrieval or to the answer extraction process. As shown in the table, the best systems are also better in the task of answer extraction. In general, all systems have improved their performance in Answer Extraction compared with previous editions.

With respect to the source of the answers, Table 36 shows that in this second year of using Wikipedia, this collection is now the main source of correct answers for most of the systems (with the exception of U. of Alicante).

Table 36. Results for questions with answer in Wikipedia and EFE

Run	% Of correct answers found in EFE	% Of Correct Answers found in Wikipedia	% Of Correct answers found NIL
prib081eses	36,97	60,50	2,52
inao082eses	24,14	68,97	6,90
inao081eses	25	70	5
qaua082eses	48,53	42,65	8,82
mira081eses	23,26	69,77	6,98
mira082eses	21,62	70,27	8,11
qaua081enes	52,27	29,55	18,18
qaua082enes	48,57	34,29	17,14
mira081fres	33,33	41,67	25

4 Conclusions

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

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

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

In six years of QA experimentation, a lot of resources and know-how have been accumulated, nevertheless systems do not show a brilliant overall performance, even those that have participated to most QA campaigns, and still seem not to manage suitably the different challenges proposed.

In conclusion, it is clear that a redefinition of the task should be thought in the next campaign. This new definition of the task should permit the evaluation and comparison of systems even working in different languages. The new setting should also take as reference a real user scenario, perhaps in a new document collection.

Acknowledgements. A special thank to Danilo Giampiccolo (CELCT, Trento, Italy), who has given his precious advise and valuable support at many levels for the preparation and realization of the QA track at CLEF 2008.

Jesús Herrera has been partially supported by the Spanish Ministry of Education and Science (TIN2006-14433-C02-01 project).

Anselmo Peñas has been partially supported by the Spanish Ministry of Science and Technology within the Text-Mess-INES project (TIN2006-15265-C06-02).

Paulo Rocha was supported by the Linguateca project, jointly funded by the Portuguese Government and the European Union (FEDER and FSE), under contract ref. POSC/339/1.3/C/NAC

References

1. QA@CLEF Website: http://clef-qa.itc.it/
2. AVE Website: http://nlp.uned.es/QA/ave/.
3. QAST Website: http://www.lsi.upc.edu/~qast/
4. QA-WSD Website: http://ixa2.si.ehu.es/qawsd/
5. QA@CLEF 2007 Organizing Committee. Guidelines 2007. http://clef-qa.itc.it/2007/download/QA@CLEF07_Guidelines-for-Participants.pdf
6. Hartrumpf, S., Glöckner, I., Leveling, J.: University of Hagen at QA@CLEF 2007: Coreference Resolution for Questions and Answer Merging. This volume.
7. Herrera, J., Peñas A., Verdejo, F.: Question Answering Pilot Task at CLEF 2004. In: Peters, C., Clough, P., Gonzalo, J., Jones, Gareth J.F., Kluck, M., Magnini, B. (eds.): Multilingual Information Access for Text, Speech and Images. Lecture Notes in Computer Science, Vol. 3491. Springer-Verlag, Berlin Hidelberg New York (2005) 581–590
8. Ion, R.: Word Sense Disambiguation Methods Applied to English and Romanian. PhD thesis, Romanian Academy, Bucharest (2007).
9. Ion, R., Mititelu, V.B.: Constrained Lexical Attraction Models. In: Nineteenth International Florida Artificial Intelligence Research Society Conference, pp. 297-302. AAAI Press, Menlo Park, California, USA (2006).
10. Landis, J. R. and Koch, G. G.: The measurements of observer agreement for categorical data. Biometrics, 33 (1997) 159–174.
11. Laurent, D., Séguéla, P., Nêgre S.: Cross Lingual Question Answering using QRISTAL for CLEF 2007. This volume.
12. Magnini, B., Giampiccolo, D., Forner, P., Ayache, C., Jijkoun, V., Osenova, P., Peñas, A., Rocha, P., Sacaleanu , B., and Sutcliffe, R.: Overview of the CLEF 2006 Multilingual Question Answering Track. In: Evaluation of Multilingual and Multi-modal Information Retrieval. Lecture Notes in Computer Science, Vol. 4730. Springer-Verlag, Berlin Heidelberg New York (2007) 223-256.
13. Peñas, A., Rodrigo, Á., Verdejo, F.: Overview of the Answer Validation Exercise 2007. This volume.
14. Turmo, J., Comas, P., Ayache, C, Mostefa, D., Rosset, S., Lamel, L.: Overview of QAST 2007.
15. Vallin, A., Magnini, B., Giampiccolo, D., Aunimo, L., Ayache, C.,Osenova, P., Peñas, A., de Rijke, M., Sacaleanu , B., Santos, D. and Sutcliffe, R. .: Overview of the CLEF 2005 Multilingual Question Answering Track. In: Accessing Multilingual Information Repositories. Lecture Notes in Computer Science, Vol. 4022. Springer-Verlag, Berlin Heidelberg New York (2006) 307-331.
16. Voorhees, E.: Overview of the TREC 2002 Question Answering Track. In NIST Special Publication 500-251: The Eleventh Text REtrieval Conference (TREC 2002). National Institute of Standards and Technology, USA (2002).
17. Agirre, Eneko & Lopez de Lacalle, Oier (2007). UBC-ALM: Combining k-NN with SVD for WSD. Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval 2007), pp. 341-345. Prague, Czech Republic.
18. Chan, Yee Seng, & Ng, Hwee Tou, & Zhong, Zhi (2007). NUS-PT: Exploiting Parallel Texts for Word Sense Disambiguation in the English All-Words Tasks. Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval 2007), pp. 253-256. Prague, Czech Republic.