Overview of ResPubliQA 2010: Question Answering Evaluation over European Legislation

Anselmo Peñas[1], Pamela Forner[2], Álvaro Rodrigo[3], Richard Sutcliffe[4], Corina Forăscu[5], Cristina Mota[6]

[1] NLP&IR group, UNED, Spain (anselmo@lsi.uned.es)
[2] CELCT, Italy (forner@celct.it)
[3] NLP&IR group, UNED, Spain (alvarory@lsi.uned.es)
[4] University of Limerick, Ireland (richard.sutcliffe@ul.ie)
[5] UAIC and RACAI, Romania (corinfor@info.uaic.ro)
[6] SINTEF ICT, Norway (cmota@ist.utl.pt)

Abstract. This paper describes the second round of ResPubliQA, a Question Answering (QA) evaluation task over European legislation, a LAB of CLEF 2010. Two tasks have been proposed this year: Paragraph Selection (PS) and Answer Selection (AS). The PS task consisted of extracting a relevant paragraph of text that satisfies completely the information need expressed by a natural language question. In the AS task, the exercise was to demarcate the shorter string of text corresponding to the exact answer supported by the entire paragraph.
The general aims of this exercise are (i) to move towards a domain of potential users; (ii) to propose a setting which allows the direct comparison of performance across languages; (iii) to allow QA technologies to be evaluated against IR approaches; (iv) to promote validation technologies to reduce the amount of incorrect answers by leaving some questions unanswered. These goals are achieved through the use of parallel aligned document collections (JRC-Acquis and EUROPARL) and the possibility to return two different types of answers, either passages or exact strings. The paper describes the task in more detail, presenting the different types of questions, the methodology for the creation of the test sets and the evaluation measure, and analyzing the results obtained by systems and the more successful approaches. Thirteen groups participated in both PS and AS tasks submitting 49 runs in total.

1. INTRODUCTION

The ResPubliQA 2010 exercise is aimed at retrieving answers to a set of 200 questions over EUROPARL and ACQUIS collections. Questions were offered in 8 different languages: Basque (EU), English (EN), French (FR), German (DE), Italian (IT), Portuguese (PT), Romanian (RO) and Spanish (ES). All Monolingual and Cross-language subtasks combinations of questions between the last 7 languages above were activated, including monolingual English (EN). Basque (EU), instead, was included exclusively as a source language, as there is no Basque translation of the document collection, which means that no monolingual EU-EU sub-task could be enacted.

The design of the ResPubliQA 2010 evaluation campaign was to a large extent the repetition of the previous year's exercise [1] with the addition of a number of refinements. Thus, the main goals of the lab this year are basically the same: Moving towards a domain of potential users; Moving to an evaluation setting able to compare systems working in different languages; Comparing current QA technologies with pure Information Retrieval (IR) approaches; Allowing more types of questions; Introducing in QA systems the Answer Validation technologies developed in the past campaigns [2,3,5].

As a difference with the previous campaign, this year participants had the opportunity to return both paragraph and exact answers as system output. Another novelty this year is the addition of a portion of the EUROPARL collection[1] in the languages involved in the task. The subject of EUROPARL's parliamentary domains is different in style and content from ACQUIS while being fully compatible with it. This has given participants the opportunity to adapt their systems in a way which widens their coverage in compatible domains; and for the organizers it has represented the opportunity to widen the scope of the questions (through the introduction of new types of question, as for example opinion).

[1] http://www.europarl.europa.eu/

The paper is organized as follows: Section 2 illustrates the document collection; Section 3 gives an overview of the different types of question developed; Section 4 addresses the various steps to create the ResPubliQA data set; Section 5 provides an explanation of the evaluation measure and of how systems have been evaluated; Section 6 gives some details about participation in this year evaluation campaign; Section 7 presents and discusses the results achieved by participating systems and across the different languages; Section 8 shows the approaches used by participating systems; and Sections 9 draws some conclusions.

2. DOCUMENT COLLECTION

Two sets of documents have been included in ResPubliQA 2010 collection: a subset of the JRC-ACQUIS Multilingual Parallel Corpus[2] and a small portion of the EUROPARL collection. Both are multilingual parallel collections. JRC-ACQUIS[3] is a freely available parallel corpus containing the total body of European Union (EU) documents, mostly of legal nature. It comprises contents, principles and political objectives of the EU treaties; the EU legislation; declarations and resolutions; international agreements; and acts and common objectives. Texts cover various subject domains, including economy, health, information technology, law, agriculture, food, politics and more. This collection of legislative documents currently includes selected texts written between 1950 and 2006 with parallel translations in 22 languages. The corpus is encoded in XML, according to the TEI guidelines.

The subset used in ResPubliQA consists of 10,700 parallel and aligned documents per language (Bulgarian, English, French, German, Italian, Portuguese, Romanian and Spanish). The documents are grouped by language, and inside each language directory, documents are grouped by year. All documents have a numerical identifier called the CELEX code, which helps to find the same text in the various languages. Each document contains a header (giving for instance the download URL and the EUROVOC codes) and a text (which consists of a title and a series of paragraphs).

EUROPARL[4] is a collection of the Proceedings of the European Parliament dating back to 1996. European legislation is a topic of great relevance to a large number of potential users from citizens to lawyers, government agencies politicians and many others. EUROPARL comprises texts in each of the 11 official languages of the European Union (Danish, German, Greek, English, Spanish, Finnish, French, Italian, Dutch, Portuguese and Swedish). With the enlargement of the European Union to 25 member countries in May 2004, the European Union has begun to translate texts into even more languages. However, translations into Bulgarian and Romanian start from January 2009 and for this reason we only compiled documents from the European Parliament site (http://www.europarl.europa.eu/) starting from that date. In this way, we ensured a parallel collection for 9 languages (Bulgarian, Dutch, English, French, German, Italian, Portuguese, Romanian and Spanish).

3. TYPES OF QUESTIONS

Beside the question types used last year (Factoid, Definition, Procedure) two additional question categories were added in the 2010 campaign: Opinion and a miscellanea called Other. Moreover, Reason and Purpose categories were merged into a single one as the distinction between them was a little blurred in the past edition. The following are examples of these types of questions:

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

Q: *What percentage of people in Italy relies on television for information?*
P: In Italy, 80% of the people get their daily information from television. If that television is not broadcasting all voices, then people do not get the chance to make their own decisions. That is fundamental to democracy.
A: 80%

Q: *What is the maximum efficiency index for a ten-place dishwasher?*

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

[3] http://wt.jrc.it/lt/Acquis/

[4] http://www.europarl.europa.eu/

P: (a) Dishwashers with 10 or more place settings shall have an energy efficiency index lower than 0,58 as defined in Annex IV to Commission Directive 97/17/EC of 16 April 1997 implementing Council Directive 92/75/EEC with regard to energy labelling of household dishwashers(1), using the same test method EN 50242 and programme cycle as chosen for Directive 97/17/EC.
A: 0,58

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

Q: *What is avian influenza?*
P: (1) Avian influenza is an infectious viral disease in poultry and birds, causing mortality and disturbances which can quickly take epizootic proportions liable to present a serious threat to animal health and to reduce sharply the profitability of poultry farming. Under certain circumstances the disease may also pose a risk to human health. There is a risk that the disease agent might be spread to other holdings, to wild birds and from one Member State to other Member States and third countries through the international trade in live birds or their products.
A: an infectious viral disease in poultry and birds, causing mortality and disturbances which can quickly take epizootic proportions liable to present a serious threat to animal health and to reduce sharply the profitability of poultry farming.

Q: *What does MFF signify in a financial context?*
P: 1. Recalls that its political priorities and its assessment of the budgetary framework for the year 2010 were set out in its resolution of 10 March 2009, where Parliament was highly critical of the tight margins available in most of the headings of the Multiannual Financial Framework (MFF);
A: Multiannual Financial Framework

Reason_Purpose. Reason_Purpose questions ask for the reasons/goals for something happening. For example:

Q: *Why was Perwiz Kambakhsh sentenced to death?*
P: I. whereas the 23 year-old Afghan journalist Perwiz Kambakhsh was sentenced to death for circulating an article about women's rights under Islam, and whereas, after strong international protests, that sentence was commuted to 20 years" imprisonment,
A: for circulating an article about women's rights under Islam

Q: *What were the objectives of the 2001 Doha Round?*
P: A. whereas the Doha Round was launched in 2001 with the objectives of creating new trading opportunities, strengthening multilateral trade rules, addressing current imbalances in the trading system and putting trade at the service of sustainable development, with an emphasis on the economic integration of developing countries, especially the least developed countries (LDCs), arising from the conviction that a multilateral system, based on more just and equitable rules, can contribute to fair and free trade at the service of the development of all continents,
A: creating new trading opportunities, strengthening multilateral trade rules, addressing current imbalances in the trading system and putting trade at the service of sustainable development, with an emphasis on the economic integration of developing countries, especially the least developed countries (LDCs), arising from the conviction that a multilateral system, based on more just and equitable rules, can contribute to fair and free trade at the service of the development of all continents,

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

Q: *How do you calculate the monthly gross occupancy rate of bed places?*
P: The gross occupancy rate of bed places in one month is obtained by dividing total overnight stays by the product of the bed places and the number of days in the corresponding month (sometimes termed bed-nights) for the same group of establishments, multiplying the quotient by 100 to express the result as a percentage.
A: by dividing total overnight stays by the product of the bed places and the number of days in the corresponding month (sometimes termed bed-nights) for the same group of establishments, multiplying the quotient by 100 to express the result as a percentage

Q: *How do you make a blank test?*
P: 7.1. A blank test shall be made regularly using an ashless filter paper (5.8) moistened with a mixture of 90 ml (4.1) sodium citrate solution, 1 ml saturated solution of calcium chloride (4.2), 0,5 ml of liquid rennet (4.5), and washed with 3 x 15 ml of distilled water before mineralisation by the Kjeldahl method as described at IDF standard 20A 1986.
A: using an ashless filter paper (5.8) moistened with a mixture of 90 ml (4.1) sodium citrate solution, 1 ml saturated solution of calcium chloride (4.2), 0,5 ml of liquid rennet (4.5), and washed with 3 x 15 ml of distilled water before mineralisation by the Kjeldahl method as described at IDF standard 20A 1986

Opinion. Opinion questions ask for the opinions/feelings/ideas about people, topics, events. For example:

Q: *What did the Council think about the terrorist attacks on London?*
P: (10) On 13 July 2005, the Council reaffirmed in its declaration condemning the terrorist attacks on London the need to adopt common measures on the retention of telecommunications data as soon as possible.
A: condemning the terrorist attacks on London

Q: *What is the Socialist Group position with respect to the case of Manuel Rosales?*
P: – Madam President, concerning the next vote, on ‘Venezuela: the case of Manuel Rosales’, the Socialist Group, of course, has withdrawn its signature from the compromise resolution. We have not taken part in the debate and we will not take part in the vote.
A: has withdrawn its signature from the compromise resolution. We have not taken part in the debate and we will not take part in the vote.

Other. It is used for any reasonable questions which do not fall into the other categories. For example:

Q: *What is the e-Content program about?*
P: A multiannual programme "European digital content for the global networks" (hereinafter referred to as "eContent") is hereby adopted.
A: European digital content for the global networks

Q: *By whom was the Treaty of Lisbon rejected?*
P: The Treaty of Lisbon, which is 96 per cent identical to the draft Constitutional Treaty, was rejected in the referendum in Ireland. Prior to that, the draft Constitutional Treaty was rejected in referendums in France and the Netherlands.
A: was rejected in the referendum in Ireland

Q: *Which ideals are central to the EU?*
P: (1) Security incidents resulting from terrorism are among the greatest threats to the ideals of democracy, freedom and peace, which are the very essence of the European Union.
A: democracy, freedom and peace

4. TEST SET PREPARATION

Three hundred questions were initially formulated, manually verified against the document collection, translated into English and collected in a common XML format using a web interface specifically designed for this purpose. To avoid a bias towards a language, the 300 questions were developed by 4 different annotators originally in 4 different languages (75 each). All questions had at least one answer in the target corpus of that language. Then, a second translation from English back into all the nine languages of the track was performed. Translators checked whether a question initially created for a particular language had an answer or not in all other languages.

Beside the paragraph containing the answer, annotators were also required to demarcate the shorter string of text that responses to a question in all different languages. Pinpointing the precise extent of an answer is a more difficult problem than finding a paragraph that contains an answer. The purposes of demarcating exact responses are (i) to show to the evaluators what the question creators considered to be the exact answer, and (ii) to create a GoldStandard which has been used to *automatically* compare the responses retrieved by the systems against those manually collected by the annotator. Nevertheless, the exact answer returned by a system was judged by human assessors besides the automatic evaluation.

The final pool of 200 questions was selected out of the 300 produced, attempting to balance the question set according to the different question types (factoid, definition, reason/purpose, procedure, opinion and others). The distribution of the different questions types in the collection is shown in Table 1. 130 questions had an answer in JRC-ACQUIS and 70 in EUROPARL. All the questions were formulated in such a way that they have an answer in all the collections, that is, there were no NIL questions.

Table 1: Distribution of question types

Question type	**Total number of questions**
DEFINITION	32
FACTOID	35
REASON/PURPOSE	33
PROCEDURE	33
OPINION	33
OTHER	34
Total	200

All language dependent tasks (question creation, translation and assessments of runs) have been performed by native speakers in a distributed setting. For this reason, a complete set of guidelines for each of these tasks have been shared among annotators and central coordination has been maintained in order to ensure consistency.

5. EVALUATION METHODOLOGY

Systems were allowed to participate in one or both tasks (PS and/or AS) which operated simultaneously on the same input questions. A maximum of two runs in total per participant could be submitted, i.e. two PS runs, two AS runs or one PS plus one AS run. Participants were allowed to submit just one response per question.

As in the previous campaign, systems had two options as output for each question:

1. To give an answer (which could be one full paragraph for the PS task; or the shortest possible string of text which contains an exact answer to the question, for the AS task)
2. To choose not to answer the particular question (if the system considers that it is not able to find a correct answer). This option is called NoA answer.

5.1 Evaluation Measure

NoA answers should be used when a system is not confident about the correctness of its answer to a particular question. The goal is to reduce the amount of incorrect responses, keeping the number of correct ones, by leaving some questions unanswered. Systems should ensure that only the portion of wrong answers is reduced, maintaining as high as possible the number of correct answers. Otherwise, the reduction in the number of correct answers is punished by the evaluation measure for both the answered and unanswered questions. We used c@1 as a measure to make account of this behaviour.

c@1, which was introduced in ResPubliQA 2009, was used also this year as the main evaluation measure for both PS and AS tasks. The formulation of c@1 is given in:

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

where

n_R: number of questions correctly answered.
n_U: number of questions unanswered.
n: total number of questions (200 in this edition)

Regarding the evaluation of **exact answers**, we also provide a measure of Answer Extraction performance, that is, the proportion of exact answers correctly extracted from correctly selected paragraphs.

Optionally, a system can also give the discarded candidate answer when responding NoA. These candidate answers were also assessed by evaluators in order to give a feedback to the participants about the validation performance of their systems, even though they are not considered in the main evaluation measure.

5.2 Assessment for Paragraph Selection

Each returned paragraph had a binary assessment: Right (R) or Wrong (W). Questions which were left unanswered were automatically filtered and marked as U (Unanswered). However, the discarded candidate answers given to these questions were also evaluated. Human assessors didn't know that these answers belong to unanswered questions.

The evaluators were guided by the initial "gold" paragraph, which contained the answers. This "gold" paragraph was only a hint, since there could be other responsive paragraphs in the same or different documents.

5.3 Assessment for Answer Selection

In order to judge the exact answer strings (**AS task**), assessors had to take into account also the paragraph as it provided the context and a justification to the exactness of the answer. Each paragraph/answer couple was manually judged and assessed considering one of the following judgments:

- **R (Right)**: the answer-string consists of an exact and correct answer, supported by the returned paragraph;

- **X (ineXact)**: the answer-string contains either part of a correct answer present in the returned paragraph or it contains all the correct answer plus unnecessary additional text; this option allowed the judge to indicate the fact that the answer was only partially correct (for example, because of missing information, or because the answer was more general/specific than required by the question, etc.)

- **M (Missed)**: the answer-string does not contain a correct answer even in part but the returned paragraph in fact does contain a correct answer. In other words, the answer was there but the system missed it completely (i.e. the system did not extract it correctly);

- **W (Wrong)**: the answer-string does not contain a correct answer and moreover the returned paragraph does not contain it either; or it contains an unsupported answer

5.4 Automatic Assessments

As human assessment is a time consuming and resource expensive task, this year it was decided to make some experiment with automatic evaluation in order to reduce the amount of work for human evaluators. The evaluation was performed in two steps:

1. Each run for both the PS and AS tasks were firstly automatically compared against the Gold Standard manually produced.
2. Non-matching paragraphs/ answers were judged by human assessors

The automatic script filtered out the answers that exactly match with the GoldStandard, assigning them correct values (R), leaving to human assessors only the evaluation of non-matching paragraphs/answers. The parameters which allow determining the correctness of a response are based on the exact match of Document identifier, Paragraph identifier, and the text retrieved by the system with respect to those in the GoldStandard.

Almost 31% of the answers (91% of them for Paragraph Selection and 9% for Answer selection) did match the GoldStandard and so it was possible to automatically mark them as correct.

The rest of the paragraphs and answers returned by systems were manually evaluated by native speaker assessors who considered if the system output was responsive or not. Answers were evaluated anonymously and simultaneously for the same question to ensure that the same criteria are being applied to all systems.

5.5 Tools and Infrastructure

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

- First of all, the need to develop a proper and coherent tool for the management of the data produced during the campaign, to store it and to make it re-usable, as well as to facilitate the analysis and comparison of results.
- Secondly, the necessity of assisting the different organizing groups in the various tasks of the data set creation and to facilitate the process of collection and translation of questions and their assessment.
- Finally, the possibility for the participants to directly access the data, submit their own runs (this also implied some syntax checks of the format), and later, get the detailed viewing of the results and statistics.

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

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

1. Edition of the data already entered at any time.
2. Revision of the data by the users themselves.
3. Consistency propagation ensuring that modifications automatically re-model the output in which they are involved.
4. Statistics and evaluation measures are calculated and updated in real time.

In particular, ensuring the consistency of data is a key feature in data management. For example, if a typo is corrected in the Translation Interface, the modification is automatically updated also in the GoldStandard files, in the Test Set files, etc.

6. PARTICIPANTS

Out of the 24 groups who had previously registered showing interest in the task, a total of 13 groups participated in the ResPubliQA 2010 tasks in 8 different languages (German, English, Spanish, Basque, French, Italian, Portuguese and Romanian), as shown in Table 2. The list of participating systems, teams and the reference to their reports are shown in Table 2.

Table 2: Systems and teams with the reference to their reports

System	Team	Reference
bpac	SZTAKI, HUNGARY	Nemeskey
dict	Dhirubhai Ambani Institute of Information and Communication Technology, INDIA	Sabnani et al
elix	University of Basque Country, SPAIN	Agirre et al
icia	RACAI, ROMANIA	Ion et al
iles	LIMSI-CNRS, FRANCE	Tannier et al
ju_c	Jadavpur University, INDIA	Pakray et al
loga	University Koblenz, GERMANY	Glöckner and Pelzer
nlel	U. politecnica Valencia, SPAIN	Correa et al
prib	Priberam, PORTUGAL	-
uaic	Al.I.Cuza\" University of Iasi, Faculty of Computer Science, ROMANIA	Iftene et al
uc3m	Universidad Carlos III de Madrid, SPAIN	Vicente-Díez et al
uiir	University of Indonesia, INDONESIA	Toba et al
uned	UNED, SPAIN	Rodrigo et al

A total of 49 runs were officially submitted considering both the PS and AS tasks. Specifically, 42 submissions in the PS task and only 7 in the AS task. It is quite encouraging that compared to last year, both the number of participating teams and the number of submissions have increased. Table 3 shows the runs submitted in each language as well as the distribution among PS and AS runs.

Table 3: Tasks and corresponding numbers of submitted runs. In brackets, the number of PS and AS runs

		Target languages (corpus and answer)							
		DE	EN	ES	FR	IT	PT	RO	**Total**
Source languages (questions)	DE	4 (4,0)							**4 (4,0)**
	EN		19 (16,3)					2 (2,0)	**21 (18,3)**
	ES			7 (6,1)					**7 (6,1)**
	EU		2 (2,0)						**2 (2,0)**
	FR				7 (5,2)				**7 (5,2)**
	IT					3 (2,1)			**3 (2,1)**
	PT						1 (1,0)		**1 (1,0)**
	RO							4 (4,0)	**4 (4,0)**
	Total	**4 (4,0)**	**21 (18,3)**	**7 (6,1)**	**7 (5,2)**	**3 (2,1)**	**1 (1,0)**	**6 (6,0)**	**49 (42,7)**

As usual, the most popular language was English (with 21 submitted runs), with Spanish and French as second (with 7 submissions each). Almost all runs were monolingual; only two participating teams attempted a cross-language task (EU-EN and EN-RO) that produced 4 runs.

7. RESULTS

7.1 Overall Results for Paragraph Selection

The use of the same set of questions in all the languages allows, as in last year, a general comparison among different languages. Table 4 shows the c@1 value for all systems. Systems were able to find answers for more than 70% of questions in all languages (combination row in Table 4) except Portuguese where there was only one participant.

Considering all languages, 99% of questions received at least one correct answer by at least one system. This is an indication that the task is feasible for current systems. It also suggests that multi-stream systems might obtain good results. One way of obtaining this improvement could be the inclusion of the validation step to choose among the different systems (streams).

Some IR based baselines were developed last year in order to compare the performance of pure IR approaches against more sophisticated QA technologies. These baselines were produced using the Okapi-BM25 ranking function [5] and are described in more detail in [4]. In this edition, the UNED group sent two similar baselines for English and Spanish and these are described in [16]. Therefore, these runs can be used for comparing QA technologies with pure IR systems in this edition.

Table 4: c@1 in participating systems in the PS task according to the language

System	DE	EN	ES	FR	IT	PT	RO
Combination	0.75	0.94	0.82	0.74	0.73	0.56	0.70
uiir101		0.73					
dict102		0.68					
bpac102		0.68					
loga102	0.62						
loga101	0.59						
prib101						0.56	
nlel101	0.49	0.65	0.56	0.55	0.63		
bpac101		0.65					
elix101		0.65					
IR baseline (uned)		0.65	0.54				

uned102			0.54				
uc3m102			0.52				
uc3m101			0.51				
dict101		0.64					
uiir102		0.64					
uned101		0.63					
elix102		0.62					
nlel102	0.59	0.62	0.20	0.55	0.53		
ju_c101		0.50					
iles102		0.48		0.36			
uaic102		0.46		0.24			0.55
uaic101		0.43		0.30			0.52
icia102							0.49
icia101							0.47
elix102euen		0.36					
elix101euen		0.33					
icia101enro							0.29
icia102enro							0.29

Although we cannot compare these results directly with those of last year, there seems to be a certain improvement in performance. Whereas the best result this year is a little higher than last year's one (from c@1 of 0.73 in English compared to 0.68) there has been a considerable improvement in the average results, with an increase from 0.39 to 0.54 in c@1 in the monolingual PS task.

EUROPARL turned out to be an easier collection than ACQUIS: 84% of all the answers by all systems over EUROPARL were correct whereas only 20% were over ACQUIS.

Table 4 shows the proportion of correct answers given by all systems to each different question type. Surprisingly, Definition questions turned out to be more difficult and Reason/Purpose slightly easier than the rest of the question types. These results contradict the performance obtained in past campaigns of QA@CLEF, where a very good performance was usually obtained in Definition questions. However, in ResPubliQA, Definition questions tend to be considerably more complex than those which appeared in earlier campaigns based on newspaper articles.

Table 4: Correct answers according to different question type

Question type	**% of correct answers**
DEFINITION	28.64%
FACTOID	46.53%
REASON_PURPOSE	53.18%
PROCEDURE	41.62%
OPINION	42.80%
OTHER	44.00%

Finally, considering the UNED baselines runs we can see that once again they performed extremely well. For English only three of the seventeen runs where better than the baselines. For Spanish, only one of the five runs was better.

7.2 Results per Language in the Paragraph Selection task

Tables 5-12 show the individual results by target language of each participant system at the PS task. Moreover, a combination of systems in each language is also given in these Tables. This combination represents the number of questions correctly answered by at least one system. All the results are ranked by c@1 values. These tables contain the following columns:

- c@1: official measure at ResPubliQA 2010.
- #R: number of questions correctly answered.
- #W: number of questions wrongly answered.
- #NoA: number of questions left unanswered.
- #NoA R: number of questions unanswered where the candidate answer was Right. In this case, the system took the bad decision of leaving the question unanswered.
- #NoA W: number of questions unanswered where the candidate answer was Wrong. In this case, the system took a good decision leaving the question unanswered.
- #NoA empty: number of questions unanswered where no candidate answer was given. Since all the questions had an answer, these cases were considered as if the candidate answer were wrong for *accuracy* calculation purposes.

Overall general statistics, together with test set questions and adjudicated runs are available at the RespubliQA website http://celct.isti.cnr.it/ResPubliQA/ under Past Campaigns.

The best results for German were obtained by the systems that include a validation step. These systems showed a very good performance validating answers (more than 75% of the rejected answers were actually incorrect). This means that these systems are able to improve their performance in the future by trying to answer the questions they left unanswered.

Table 5: Results for German in the PS task

System	c@1	#R	#W	#NoA	#NoA R	#NoA W	#NoA empty
combination	0.75	150	50	0	0	0	0
loga102PSdede	0.62	105	59	36	2	29	5
loga101PSdede	0.59	101	65	34	2	27	5
nlel101PSdede	0.49	90	93	17	2	15	0
nlel102PSdede	0.44	88	112	0	0	0	0

The combination of English systems shows that 94% of questions were correctly answered by at least one system, which means that the task is feasible for current technologies. There are still some systems that perform worse than the IR baseline. As already discussed in the last edition, participant should care more about the correct tuning of the IR engine.

Most of the systems that left some questions unanswered didn't provide the candidate answer, so the organizers couldn't provide feedback about the actual state of validation technologies in English. There is some evidence that more efforts should be applied to the validation step in English for improving overall results as has been shown in German.

Table 6: Results for English in the PS task

System	c@1	#R	#W	#NoA	#NoA R	#NoA W	#NoA empty
combination	0.94	188	12	0	0	0	0
uiir101PSenen	0.73	143	54	3	0	3	0
bpac102PSenen	0.68	136	64	0	0	0	0
dict102PSenen	0.68	117	52	31	17	14	0
bpac101PSenen	0.65	129	71	0	0	0	0
elix101PSenen	0.65	130	70	0	0	0	0
nlel101PSenen	0.65	128	68	4	2	2	0
IR baseline (uned)	0.65	129	71	0	0	0	0
dict101PSenen	0.64	127	73	0	0	0	0
uiir102PSenen	0.64	127	73	0	0	0	0
uned101PSenen	0.63	117	66	17	13	4	0
nlel102PSenen	0.62	122	76	2	0	2	0
elix102PSenen	0.62	123	77	0	0	0	0
ju_c101PSenen	0.50	73	52	75	0	0	75

iles102PSenen	0.48	89	95	16	0	0	16
uaic102PSenen	0.46	85	98	17	0	0	17
uaic101PSenen	0.43	78	99	23	0	0	23
elix102PSeuen	0.36	72	128	0	0	0	0
elix101PSeuen	0.33	66	134	0	0	0	0

With respect to Spanish, 80% of questions were correctly answered by at least one system. However, this combination is almost a 50% higher than the best system. Only one system (*nlel*101PSeses) performed better than the IR baseline. This system was able to reduce the number of incorrect answers while maintaining the same number of correct answers as the baseline. This is what allowed it to obtain a better performance according to c@1.

Table 7: Results for Spanish in the PS task

System	c@1	#R	#W	#NoA	#NoA R	#NoA W	#NoA empty
combination	0.82	165	35	0	0	0	0
nlel101PSeses	0.56	108	86	6	1	5	0
IR baseline (uned)	0.54	108	92	0	0	0	0
uned101PSeses	0.54	92	73	35	22	13	0
uc3m102PSeses	0.52	104	96	0	0	0	0
uc3m101PSeses	0.51	101	99	0	0	0	0
nlel102PSeses	0.20	39	161	0	0	0	0

Similar results were obtained for French and Romanian as target, where the difference between the combination row and the best system is relevant. Again, the system *nlel* shows that accurate validation technologies have been developed.

Table 8: Results for French in the PS task

System	c@1	#R	#W	#NoA	#NoA R	#NoA W	#NoA empty
combination	0.74	148	52	0	0	0	0
nlel101PSfrfr	0.55	105	86	9	2	7	0
nlel102PSfrfr	0.55	109	88	3	0	3	0
iles102PSfrfr	0.36	62	105	33	0	0	33
uaic101PSfrfr	0.30	54	124	22	0	0	22
uaic102PSfrfr	0.24	47	153	0	0	0	0

Table 9: Results for Romanian in the PS task

System	c@1	#R	#W	#NoA	#NoA R	#NoA W	#NoA empty
combination	0.70	140	60	0	0	0	0
UAIC102PSroro	0.55	95	74	31	0	0	31
UAIC101PSroro	0.52	102	93	5	0	0	5
icia102PSroro	0.49	63	29	108	0	0	108
icia101PSroro	0.47	93	107	0	0	0	0
icia102PSenro	0.29	56	137	7	0	0	7
icia101PSenro	0.29	58	139	3	0	0	3

Table 10: Results for Italian in the PS task

System	c@1	#R	#W	#NoA	#NoA R	#NoA W	#NoA empty
combination	0.73	146	54	0	0	0	0
nlel101PSitit	0.63	124	72	4	2	2	0
nlel102PSitit	0.53	105	94	1	0	1	0

Table 11: Results for Portuguese in the PS task

System	c@1	#R	#W	#NoA	#NoA R	#NoA W	#NoA empty
prib101PSptpt	0.56	111	88	1	0	0	1

7.3 Results in the Answer Selection Task

Tables 12-14 show the results by language of participant systems at the AS task. The results for all the languages are given in Table 12. These tables contain similar information to the PS tables plus the following additional information:

- #M: number of questions where the paragraph contained a correct answer, but the answer string given was wrong
- #X: number of questions where the answer string given was judged as inexact.

All runs were monolingual. Three groups (*iles*, *ju_c* and *nlel*) submitted seven runs for the Answer Selection (AS) task. Each of these groups submitted one EN run. In addition, *iles* submitted one FR run and *nlel* submitted one ES, one FR and one IT run. Thus there were three attempts at EN and two at FR, allowing some comparison. For ES and IT there was only one run each.

Table 12: General Results in the AS task

System	c@1	#R	#W	#M	#X	#NoA	#NoA R	#NoA W	#NoA M	#NoA X	#NoA empty
combination	0.30	60	140	0	0	0	0	0	0	0	0
ju_c101ASenen	0.26	31	12	10	8	115	0	40	24	0	75
iles101ASenen	0.09	17	124	6	44	9	0	0	0	0	9
iles101ASfrfr	0.08	14	128	7	36	15	0	0	0	0	15
nlel101ASenen	0.07	10	97	20	6	67	0	0	0	0	67
nlel101ASeses	0.06	12	138	21	1	28	0	0	0	0	28
nlel101ASitit	0.03	6	139	18	7	30	0	0	0	0	30
nlel101ASfrfr	0.02	4	132	13	11	40	0	0	0	0	40

Considering EN first, the best system by c@1 was *ju_c* with a score of 0.26. Interestingly, while *iles* scored only 0.09, it had a high number of X answers (44). Thus, *iles* was identifying the vicinity of answers better than *ju_c* but was not demarcating them exactly right. Of course, the demarcation in cases of question types like reason is not beyond debate. Finally, the third system *nlel* scored 0.07.

Table 13: Results for English in the AS task

System	c@1	#R	#W	#M	#X	#NoA	#NoA R	#NoA W	#NoA M	#NoA X	#NoA empty
combination	0.24	49	151	0	0	0	0	0	0	0	0
ju_c101ASenen	0.26	31	12	10	8	115	0	40	24	0	75
iles101ASenen	0.09	17	124	6	44	9	0	0	0	0	9
nlel101ASenen	0.07	10	97	20	6	67	0	0	0	0	67

Now, turning to FR, *iles* scored 0.08 and nlel scored 0.02. Notice once again the large number of inexact answers for *iles*.

Table 14: Results for French in the AS task

System	c@1	#R	#W	#M	#X	#NoA	#NoA R	#NoA W	#NoA M	#NoA X	#NoA empty
combination	0.8	17	183	0	0	0	0	0	0	0	0
iles101ASfrfr	0.08	14	128	7	36	15	0	0	0	0	15
nlel101ASfrfr	0.02	4	132	13	11	40	0	0	0	0	40

Notice that these figures are all very low compared to traditional factoid QA where figures of 0.8 can be obtained. We can attribute this to the inclusion of difficult question types which go beyond the factoid concept with its dependence on the Named Entity concept. Recall that the breakdown of questions was 40 factoids and 32 definitions, with 32 each of opinion, procedure, reason-purpose and "other" questions. Thus 64% of questions fell into the latter four "difficult" types.

Another consideration is the effect of allowing systems to mark questions as unanswered even though they had in fact answered them. Only in the case of EN and the *ju_c* run was there any loss of score incurred by not answering. For *ju_c*, 24 unanswered questions had a missed answer, i.e. *ju_c* identified the correct paragraph containing the exact answer, but was not able to demarcate it. For all the other EN runs (and indeed all the other runs), unanswered questions had an empty answer, so nothing can be said about how close these other systems were to getting the right answer in the case of unanswered questions.

8. SYSTEM DESCRIPTIONS

A summary of the techniques reported by participants is shown in Table 15. Most of the systems that analyze the questions use manually built patterns. Regarding the IR model, BM25 has been applied by almost half of the participants that reported the retrieval model used. The other reported models were the standard ones supplied by Lucene.

Table 15: Methods used by participating systems

System name	**Question Analyses**				**Retrieval Model**	**Linguistic Unit which is indexed**				
	No Question Analyses	**Manually done Patterns**	**automatically acquired patterns**	**Other**		**Words**	**Lemmas**	**Stems**	**N-grams**	**Chunks/ phrases**
bpac				Lemmatization, POS tagging and *very* minimal pattern usage	Okapi BM25	x		x		
dict		X						x		
elix				lemmatization, part of speech tagging	BM25			x		
icia	X				Lucene Boolean Search Engine		x			
iles		X					x			
loga		X			standard lucene model; word senses are indexed		x			
nlel		X			Distance Density N-gram Model , BM25			x	x	
ju_c			X		Apache Lucene	x	x	x	x	
prib		X				x	x	x		x
uaic		X			LUCENE	x				
uc3m		X			Passage IR		x			
uiir		X					x			
uned		X			BM25			x		

A summary of Answer Extraction techniques in the AS task is given in Table 16. The most common processing was the use of named entities, numeric and temporal expressions, while some systems relied on syntactic processing by means of chunking, dependency parsing or syntactic transformations.

Table 16: Methods used by systems for extracting answers

Answer Extraction – Further processing											
System name	**Chunking**	**n-grams**	**Named Entity Recognition**	**Temporal expressions**	**Numerical expressions**	**Dependency analysis**	**Syntactic transformations**	**Logic representation**	**Theorem prover**	**Other**	**None**
bpac											x
dict	x	x	x	x	x						
elix										x	
icia	x									x	
iles			x		x	x	x				
loga				x	x			x	x		
nlel											x
ju_c	x										
prib			x	x	x						
uaic			x	x	x		x				
uc3m				x							
uiir							x				
uned			x	x	x						

The validation of answers was applied by 9 of the 13 participants. According to Table 17, which shows the different validation techniques applied by participants, the most common processing was to measure the lexical overlapping between questions and candidate answers (it was performed by 5 participants). On the other hand, more complex techniques such as syntactic similarity or theorem proving were applied by very few participants.

These observations are different from the ones obtained last year, where participants applied more techniques and performed more complex analysis like semantic similarity or the combination of different classifiers. That is, participants at ResPubliQA 2010 relied on naive techniques for performing validation. However, the experience during the last years shows that the validation step can improve results if it is performed carefully. Otherwise, the effect will be the opposite, the harming results.

Table 17: Techniques used for the Answer Validation component

System name	**No answer validation**	**Machine Learning**	**Redundancies in the collection**	**Lexical similarity (term overlapping)**	**Syntactic similarity**	**Theorem prooving or similar**	**Other**
bpac	x						
dict				x			
elix	x						
icia	x						
iles			x		x		
loga		x		x		x	

nlel							x
ju_c				x			x
prib	x						
uaic				x	x		
uc3m				x			
uiir			x				
uned							x

9. CONCLUSIONS

An important result demonstrated in 2009 was that a good IR system can be better than a QA system if the IR parameters are carefully tuned to the requirements of the domain. A relevant portion of participants already moved towards better IR models, although there are still many systems that don't outperform IR baselines.

While the Paragraph Selection task is just paragraph retrieval, the main difference from pure IR systems is to add the decision of leaving the question unanswered, that is, the validation step. Best performing systems in German, Spanish and French have accurate validation steps.

The PS task allows the inclusion of more complex questions, as well as their evaluation in a simple and natural way. However, when the aim is to extract an exact answer (as in the AS task), it turns out to be very difficult for systems to perform well, except were answers are named entities. This is because NE is a well-studied and largely solved problem. On the other hand, "exact" answer demarcation for more complex queries against documents such as those used in ResPubliQA needs further study by both system designers and evaluation task organizers.

ACKNOWLEDGMENTS

Our thanks are due to all the annotators who took care of the translations of the questions in the different languages and of the evaluation of the respective submitted runs:

- Anna Kampchen, Julia Kramme (University of Hagen, Hagen, Germany), and Stefan Rigo (FBK, Trento, Italy) for the German language
- Iñaki Alegria (University of Basque Country, Spain) for the Basque language
- Patricia Hernandez Llodra (UNED, Madrid, Spain) for the French language
- Rosário Silva (Linguateca/FCCN, Portugal) for the Portuguese language

Special thanks are also due Giovanni Moretti (CELCT, Trento, Italy) for the technical support in the management of all data of the campaign.

This work has been partially supported by the Regional Government of Madrid under the Research Network MA2VICMR (S-2009/TIC-1542), the Education Council of the Regional Government of Madrid and the European Social Fund.

REFERENCES

1. Anselmo Peñas, Pamela Forner, Richard Sutcliffe, Álvaro Rodrigo, Corina Forascu, Iñaki Alegria, Danilo Giampiccolo, Nicolas Moreau, Petya Osenova. Overview of ResPubliQA 2009: Question Answering Evaluation over European Legislation. In C. Peters, G. di Nunzio, M. Kurimo, Th. Mandl, D. Mostefa, A. Peñas, G. Roda (Eds.), Multilingual Information Access Evaluation Vol. I Text Retrieval Experiments, Workshop of the Cross-Language Evaluation Forum, CLEF 2009, Corfu, Greece, 30 September - 2 October. Revised Selected Papers. (to be published)

2. Anselmo Peñas, Álvaro Rodrigo, Felisa Verdejo. Overview of the Answer Validation Exercise 2007. In C. Peters, V. Jijkoun, Th. Mandl, H. Müller, D.W. Oard, A. Peñas, V. Petras, and D. Santos, (Eds.): Advances in Multilingual and Multimodal Information Retrieval, LNCS 5152, September 2008.

3. Anselmo Peñas, Álvaro Rodrigo, Valentín Sama, Felisa Verdejo. Overview of the Answer Validation Exercise 2006. In C. Peters, P. Clough, F. C. Gey, J. Karlgren, B. Magnini, D. W. Oard, M. de Rijke, M. Stempfhuber (Eds.): Evaluation of Multilingual and Multi-modal Information Retrieval, 7th Workshop of the Cross-Language Evaluation Forum, CLEF 2006, Alicante, Spain, September 20-22, 2006, Revised Selected Papers.

4. Joaquín Pérez, Guillermo Garrido, Álvaro Rodrigo, Lourdes Araujo and Anselmo Peñas. Information Retrieval Baselines for the ResPubliQA Task. CLEF 2009, LNCS 6241.

5. Álvaro Rodrigo, Anselmo Peñas, Felisa Verdejo. Overview of the Answer Validation Exercise 2008. In C. Peters, Th. Mandl, V. Petras, A. Peñas, H. Müller, D. Oard, V. Jijkoun, D. Santos (Eds), Evaluating Systems for Multilingual and Multimodal Information Access, 9th Workshop of the Cross-Language Evaluation Forum, CLEF 2008, Aarhus, Denmark, September 17-19, 2008, Revised Selected Papers.

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

7. Xavier Tannier and Véronique Moriceau. FIDJI @ ResPubliQA 2010. Notebook Paper for the CLEF 2010 LABs Workshop, 22-23 September, Padua, Italy

8. Hitesh Sabnani and Prasenjit Majumder. Question Answering System: Retrieving relevant passages. Notebook Paper for the CLEF 2010 LABs Workshop, 22-23 September, Padua, Italy

9. Partha Pakray, Pinaki Bhaskar, Santanu Pal, Dipankar Das, Sivaji Bandyopadhyay and Alexander Gelbukh. JU_CSE_TE: System Description QA@CLEF 2010. Notebook Paper for the CLEF 2010 LABs Workshop, 22-23 September, Padua, Italy

10. Eneko Agirre, Olatz Ansa, Xabier Arregi, Maddalen Lopez de Lacalle, Arantxa Otegi and Xabier Saralegi. Document Expansion for Cross-Lingual Passage Retrieval. Notebook Paper for the CLEF 2010 LABs Workshop, 22-23 September, Padua, Italy

11. Hapnes Toba and Mirna Adriani. Contextual Approach for Paragraph Selection in Question Answering Task. Notebook Paper for the CLEF 2010 LABs Workshop, 22-23 September, Padua, Italy

12. Adrian Iftene, Diana TRANDABAT, Alex Moruz and Maria HUSARCIUC. Question Answering on Romanian, English and French Languages. Notebook Paper for the CLEF 2010 LABs Workshop, 22-23 September, Padua, Italy

13. David M. Nemeskey. SZTAKI @ ResPubliQA 2010. Notebook Paper for the CLEF 2010 LABs Workshop, 22-23 September, Padua, Italy

14. María Teresa Vicente-Díez, Julián Moreno-Schneider and Paloma Martínez. Temporal information needs in ResPubliQA: an attempt to improve accuracy. The UC3M Participation at CLEF 2010. Notebook Paper for the CLEF 2010 LABs Workshop, 22-23 September, Padua, Italy

15. Ingo Glöckner and Björn Pelzer. The LogAnswer Project at ResPubliQA 2010. Notebook Paper for the CLEF 2010 LABs Workshop, 22-23 September, Padua, Italy

16. Álvaro Rodrigo, Joaquin Perez-Iglesias, Anselmo Peñas, Guillermo Garrido and Lourdes Araujo. A Question Answering System based on Information Retrieval and Validation. Notebook Paper for the CLEF 2010 LABs Workshop, 22-23 September, Padua, Italy

17. Radu Ion, Alexandru Ceausu, Dan Ştefănescu, Dan Tufis, Elena Irimia and Verginica Barbu Mititelu. Monolingual and Multilingual Question Answering on European Legislation. Notebook Paper for the CLEF 2010 LABs Workshop, 22-23 September, Padua, Italy

18. Santiago Correa, Davide Buscaldi and Paolo Rosso. NLEL at RespubliQA 2010. Notebook Paper for the CLEF 2010 LABs Workshop, 22-23 September, Padua, Italy