Overview of ResPubliQA 2009: Question Answering Evaluation over European Legislation

Anselmo Peñas[1], Pamela Forner[2], Richard Sutcliffe[3], Álvaro Rodrigo[4], Corina Forăscu[5], Iñaki Alegria[6], Danilo Giampiccolo[7], Nicolas Moreau[8], Petya Osenova[9]

[1] UNED, Spain (anselmo@lsi.uned.es)
[2] CELCT, Italy (forner@celct.it)
[3] DLTG, University of Limerick, Ireland (richard.sutcliffe@ul.ie)
[4] UNED, Spain (alvarory@lsi.uned.es)
[5] UAIC and RACAI, Romania (corinfor@info.uaic.ro)
[6] University of Basque Country, Spain (i.alegria@ehu.es)
[7] CELCT, Italy (giampiccolo@celct.it)
[8] ELDA/ELRA, France (moreau@elda.org)
[9] BTB, Bulgaria, (petya@bultreebank.org)

Abstract. This paper describes the first round of ResPubliQA, a Question Answering (QA) evaluation task over European legislation, proposed at the Cross Language Evaluation Forum (CLEF) 2009. The exercise consists of extracting a relevant paragraph of text that satisfies completely the information need expressed by a natural language question. The general goals of this exercise are (i) to study if the current QA technologies tuned for newswire collections and Wikipedia can be adapted to a new domain (law in this case); (ii) to move to a more realistic scenario, considering people close to law as users, and paragraphs as system output; (iii) to compare current QA technologies with pure Information Retrieval (IR) approaches; and (iv) to introduce in QA systems the Answer Validation technologies developed in the past three years. 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 new evaluation measure, and analyzing the results obtained by systems and the more successful approaches. Eleven groups participated with 28 runs. In addition, we evaluated 16 baseline runs (2 per language) based only in pure IR approach, for comparison purposes. Considering accuracy, scores were generally higher than in previous QA campaigns.

1. INTRODUCTION

This year, the Multilingual Question Answering Track proposed three separate and independent exercises:

1. *QAST*: The aim of the third QAST exercise is to evaluate QA technology in a real multilingual speech scenario in which written and oral questions (factual and definitional) in different languages are formulated against a set of audio recordings related to speech events in those languages. The scenario is the European Parliament sessions in English, Spanish and French.
2. *GikiCLEF*: Following the previous GikiP pilot at GeoCLEF 2008, the task focuses on open list questions over Wikipedia that require geographic reasoning, complex information extraction, and cross-lingual processing, at least for Dutch, English, German, Norwegian, Portuguese and Romanian.
3. *ResPubliQA*: Given a pool of 500 independent questions in natural language, systems must return the passage - not the exact answer - that answers each question. The document collection is JRC-Acquis about EU documentation[1]. Both questions and documents are translated into and aligned for a subset of official European languages, i.e. Bulgarian, English, French, German, Italian, Portuguese, Romanian and Spanish.

This overview is dedicated only to the ResPubliQA exercise. For more details about QAST and GikiCLEF see the respective overviews in this volume.

The ResPubliQA 2009 exercise is aimed at retrieving answers to a set of 500 questions. The answer of a question is a paragraph of the test collection. The hypothetical user considered for this exercise is a person interested in making inquiries in the law domain, specifically on the European legislation. The ResPubliQA document collection is a subset of JRC-Acquis[1], a corpus of European legislation that has parallel translations aligned at document level in many European languages.

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

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

The paper is organized as follows: Section 2 gives an explanation of the task objectives; Section 3 illustrates the document collection; Section 4 gives an overview of the different types of question developed; Section 5 addresses the various steps to create the ResPubliQA data set; Section 6 shows the format of the test set and of the submissions that systems have returned; Section 7 provides an explanation of the evaluation measure and of how systems have been evaluated; Section 8 gives some details about participation in this year evaluation campaign; Section 9 presents and discusses the results achieved by participating systems and across the different languages; Section 10 shows the methodologies and technique used by participating systems and Section 11 and 12 draws some conclusions highlighting the challenges which are still to be addressed.

2. TASK OBJECTIVES

The general objectives of the exercise are:

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

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

In confirmation of the increasing importance of this issue, a Legal Track [13], aimed at advancing computer technologies for searching electronic legal records, was also introduced in 2006 as part of the yearly TREC conferences sponsored by the National Institute of Standards and Technology (NIST).[2] The task of the Legal Track is to retrieve all the relevant documents for a specific query and to compare the performances of systems operating in a setting which reflects the way lawyers carry out their inquiries.

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

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

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

4. Comparing current QA technologies with pure Information Retrieval (IR) approaches. Returning a complete paragraph instead of an exact answer allows the comparison between pure IR approaches and current QA technologies. In this way, a nice benchmark for evaluating IR systems oriented to high precision, where only one paragraph is needed, has been also created. The documents are nicely divided into xml paragraph marks solving the technical issues for paragraph retrieval. Furthermore, a paragraph is presumably a more realistic output for the users of the new collection domain.

5. Allowing more types of questions. Returning one paragraph allows new types of questions with the only restriction that they must be answered by a single paragraph.

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

6. Introducing in QA systems the Answer Validation technologies developed in the past campaigns. During the last campaigns we wanted to stick to the easiest and most comprehensible evaluation of systems, that is, requesting only one answer per question and counting the proportion of questions correctly answered (namely accuracy). In this campaign, we wanted to introduce a more discriminative measure, allowing systems to leave some questions unanswered. Given two systems that answer correctly the same proportion of questions, the one that returns less incorrect answers (leaving some questions unanswered) will score better. Thus, systems can add a final module to decide whether they found enough evidence or not to return their best answer.

This is a classification problem that takes advantage of more sophisticated Answer Validation technologies developed during the last years [7,8,11].

3. DOCUMENT COLLECTION

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

The ResPubliQA collection in 8 of the languages involved in the track - Bulgarian, English, French, German, Italian, Portuguese, Romanian and Spanish - consists of roughly 10,700 parallel and aligned documents per language. The documents are grouped by language, and inside each language directory, documents are grouped by year. All documents have a numerical identifier called the CELEX code, which helps to find the same text in the various languages. Each document contains a header (giving for instance the download URL and the EUROVOC codes) and a text (which consists of a title and a series of paragraphs).

4. TYPES OF QUESTIONS

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

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

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

Q: *In how many languages is the Official Journal of the Community published?*
A: The Official Journal of the Community shall be published in the four official languages.

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

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

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

3 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.

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

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

Q: *Why did a Commission expert conduct an inspection visit to Uruguay?*
A: A Commission expert has conducted an inspection visit to Uruguay to verify the conditions under which fishery products are produced, stored and dispatched to the Community.

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

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

Q: *What is the overall objective of the eco-label?*
A: The overall objective of the eco-label is to promote products which have the potential to reduce negative environmental impacts, as compared with the other products in the same product group, thus contributing to the efficient use of resources and a high level of environmental protection. In doing so it contributes to making consumption more sustainable, and to the policy objectives set out in the Community's sustainable development strategy (for example in the fields of climate change, resource efficiency and eco-toxicity), the sixth environmental action programme and the forthcoming White Paper on Integrated Product Policy Strategy.

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

Q: *How are stable conditions in the natural rubber trade achieved?*
A: To achieve stable conditions in natural rubber trade through avoiding excessive natural rubber price fluctuations, which adversely affect the long-term interests of both producers and consumers, and stabilizing these prices without distorting long-term market trends, in the interests of producers and consumers;
Q: *What is the procedure for calling an extraordinary meeting?*
A: 2. Extraordinary meetings shall be convened by the Chairman if so requested by a delegation.

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

5. TEST SET PREPARATION

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

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

During the creation of the 100 questions in a source language and their “mapping to English” the question creator was supposed not only to translate the questions into English, but also to look for the corresponding answer at least in the English corpus. After the selection of the final 500 questions, during their translation from

English into the other source language, checking the availability of answers for all the questions in all the languages of the parallel corpus ensured that there is no NIL question, as in the previous QA@CLEF editions. The most frequent problematic situations were due to the misalignments between documents at the paragraph level:

- Entire paragraphs missing from one language, but, of course, existing in other(s); for example jrc31982D0886-ro contains only 25 paragraphs, but the English document contains 162 paragraphs, with the text containing an EC Convention, absent from the Romanian version.
- Different paragraph segmentation into different languages of the parallel corpus; for example the document jrc31985L0205-en contains one single paragraph (n="106") corresponding to 685 Romanian paragraphs (n="106_790"). From the point of view of our track, this means that one question having the answer in the (only one) English paragraph had to be removed, since the answer in Romanian is supposed to be found in exactly one paragraph.
- Missing information (parts of the text) in one paragraph; for example a question like "What should be understood by "living plants"?" had answer in English document jrc31968R0234-en paragraph number 8 "Whereas the production of live trees and other plants, bulbs, roots and the like, cut flowers and ornamental foliage (hereinafter where appropriate called 'live plants')". However, the corresponding Romanian paragraph number 9, does not include the list of the live plants.
- Contradictory information in corresponding paragraphs; for example the corresponding paragraphs that answers the question "How much does cotton increase in weight after treatment with formic acid?" indicate a loss of 3% in the Romanian version, whereas in English the loss is 4%.

6. FORMAT

6.1 Test set

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

```
source_lang    target_lang    q_id    q_string
```

where:

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

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

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

6.2 Submission format

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

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

where:

- `q_id` is the question number as given in the test set (of the form 0001 to 0500) Passages must be returned in the same ascending (increasing) order in which questions appear in the test set;

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

i.e.

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

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

7. EVALUATION

7.1 Responses

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

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

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

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

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

7.2 Assessments

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

One of the following judgements was given to each question-answer by human assessors during the evaluation:

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

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

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

7.3 Evaluation Measure

The use of Machine Learning-based techniques able to decide if a candidate answer is finally acceptable or not was introduced by the Answer Validation Exercise[4] during the past campaigns. This is an important achievement, as an improvement in the accuracy of such decision-making process leads to more powerful QA architectures with new feedback loops. One of the goals of the ResPubliQA exercise is to effectively introduce these techniques in current QA systems.

For this reason, the unique measure considered in this evaluation campaign was the following:

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

where:

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

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

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

7.4 Tools and Infrastructure

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

- First of all the need to develop a proper and coherent tool for the management of the data produced during the campaign, to store it and to make it re-usable, as well as to facilitate the analysis and comparison of the results.
- Secondly, the necessity of assisting the different organizing groups in the variuos 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.

[4] http://nlp.uned.es/clef-qa/ave

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

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

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

8. PARTICIPANTS

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

Table 1: Tasks and corresponding numbers of submitted runs

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

The most chosen language appeared to be English with 12 submitted runs, followed by Spanish with 6 submissions. No runs were submitted either in Bulgarian or Portuguese. Participants came above all from Europe, except two different groups from India. Table 1 shows the run distribution in the different languages.
The list of participating systems, teams and the reference to their reports are shown in Table 2.

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

System	Team	Reference
elix	ELHUYAR-IXA, SPAIN	Agirre et al., [1]
icia	RACAI, ROMANIA	Ion et al., [5]
iiit	Search & Info Extraction Lab, INDIA	Bharadwaj et al., [15]
iles	LIMSI-CNRS-2, FRANCE	Moriceau et al., [6]
isik	ISI-Kolkata, INDIA	-
loga	U.Koblenz-Landau, GERMAN	Gloeckner and Pelzer, [3]
mira	MIRACLE, SPAIN	Vicente-Díez et al., [14]
nlel	U. politecnica Valencia, SPAIN	Correa et al., [2]
syna	Synapse Developpment, FRANCE	-
uaic	AI.I.Cuza U. of IASI, ROMANIA	Iftene et al., [4]
uned	UNED, SPAIN	Rodrigo et al., [12]

9. RESULTS

9.1 IR Baselines

Since there were a parallel collection and one set of questions for all languages, the only variable that did not permit strict comparison between systems was the language itself. Running exactly the same IR system in all languages did not permit to fix this variable but at least we have some evidence about the starting difficulty in each language.

Two baseline runs per language, based on pure Information Retrieval, were prepared and assessed with two objectives:

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

These baselines were produced in the following way:

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

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

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

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

The same parameters in all runs for all languages were used. For more details about the preparation of these baselines see [9].

9.2 Results per language

Tables 3-8 show systems performance divided by language. The content of the columns is as follows:

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

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

where N is the number of questions (500).

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

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

Table 3: Results for German

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

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

Table 4: Results for English

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

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

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

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

Something very interesting in the English runs is that the two best teams (see uned092enen, nlel091enen runs) produced paragraph rankings considering matching n-grams between question and paragraph [2]. This retrieval approach seems to be promising, since combined with paragraph validation filters it achieved the best score [12] in English.

These two approaches obtained the best score also in Spanish (uned091eses, nlel091eses). Additionally, [2] performed second experiment (nlel092eses) that achieved the best result considering the whole parallel collection to obtain a list of answers in different languages (Spanish, English, Italian and French).

Table 5: Results for Spanish

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

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

Table 6: Results for French

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

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

Table 7: Results for Italian

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

With respect to Italian (Table 7), the only participant obtained better results than the baselines.

Table 8: Results for Romanian

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

The best system in Romanian [5] showed a very good performance compared to the rest of runs, as Table 8 shows. This is a system that uses a sophisticated similarity based model for paragraph ranking, question analysis, classification and regeneration of the question, classification of paragraphs and consideration of the EUROVOC terms associated to each document.

9.3 Comparison of results across languages

Strict comparison between systems across languages is not possible without ignoring the language variable. However, this is the first time that systems working in different languages were evaluated with the same questions over the same document collection manually translated into different languages. So, extracting information about which approaches are more promising should be possible.

For this purpose, we considered both the systems participating in more than one language and the baseline IR runs for all languages.

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

Table 9 : c@1 in participating systems according to the language

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

In the first table (Table 9) systems are ordered by c@1 values. Reading column by column, systems are correctly ordered in each language, except some swaps with respect to the baseline IR suns. Systems icia092, uned and nlel seem to have the more powerful approaches.

In the next table (Table 10) we tried to partially fix the language variable, dividing c@1 values by the score of the best IR baseline system. Values over 1 indicate better performance than the baseline, and values under 1 indicate worse performance than the baseline.

In Table 10, the ranking of systems change, showing that also system loga proposes a promising approach, whereas nlel091 system appear more aligned with the baselines than loga. Of course, this evidence is affected by another variable that must be taken into account before making strong claims, i.e. the baseline itself, which perhaps is not the best approach for all languages (specially agglutinative languages such as German).

Table 10: C@1/Best IR baseline

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

Table 11 shows that the majority of questions have been answered by systems in many different languages. For example, 74 questions have been answered in all languages, whereas only 6 questions remained unanswered considering all languages. Notice that 99% of questions have been answered by at least one system in at least one language.

Table11: Number of questions answered by systems in different languages

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

10. SYSTEM DESCRIPTION

Tables 12 and 13 summarise the characteristics of the participant systems. As can be seen, some systems did not analyse the questions at all. Among those that did, the most popular technique was the use of manually created query patterns (e.g. “Where is...” could indicate a location question). As regards retrieval models, two systems used Boolean methods while the rest mainly used Okapi or a VSM-type model.

Table 12: Methods used by participating systems

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

Table 13 shows the type of processing techniques which were used on document fragments returned by the information retrieval components. As would be expected, Named Entity recognition and Numerical Expression recognition were widely used approaches.

Table 13: 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	Functions (sub, obj,)	Syntactic transformations	Semantic parsing	Semantic role labeling	Logic representation	Theorem prover	None
SYNA			x	x	x	x	x		x	x			
ICIA			x		x								
ISIK													x
NLEL													x
UAIC			x	x	x								
MIRA			x	x	x								
ILES	x		x		x	x	x	x					
IIIT	x		x		x								
UNED		x	x	x	x								
ELIX													
LOGA				x	x				x	x	x	x	

Table 14 shows the types of technique used for the answer validation component. Some systems did not have such a component, but for those that did, lexical similarity and syntactic similarity were the most widely used approaches.

Table 14: Technique used for the Answer Validation component

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

11. OPEN ISSUES

Whereas in previous years, almost all responses were double-blind evaluated to check inter-evaluator agreement, this year it was not possible. A measure of the inter-annotator agreement would have provided us an idea of the complexity and ambiguity of both questions and their supporting passages.

Moreover, this was the first year of using the JRC-Acquis collection which claims to be parallel in all languages. The supposed advantage of this was that all systems answer the same questions against the same document collections. Only the language of the questions and documents vary as otherwise the text is supposed to mean exactly the same. However, we found that in fact the texts are not parallel, being many passages left out or translated in a completely different way. The result was that many questions were not supported in all languages and could not therefore be used. This problem resulted in a huge amount of extra work for the organisers. Furthermore, the character of the document collection necessitated changes to the type of the questions. In most cases the questions became more verbose in order to deal with the vagueness and ambiguity of texts.

The idea of introducing new question types Reason, Purpose and Procedure was good in principle, but it did not seem to work as expected. Reason and Purpose questions resulted to be understood as more or less the same and the way in which these reasons and purposes are stated in the documents sometimes is meaningless. A typical type of reason is “to ensure smooth running of the EU” and a typical purpose is “to implement such and such a law”. With respect to procedures there were also some non informative responses similar to the idea “the procedure to apply the law is to put it into practice”.

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

12. CONCLUSIONS

494 questions (99%) were answered by at least one system in at least one language; nevertheless the systems that gave more correct answers only answered 288. This shows that the task is feasible and systems still have room to improve and solve it in a short time.

One of the main issues is the retrieval model. Many systems must pay more attention to it since they performed under the baselines based on just IR. From this perspective, paragraph ranking approaches based on n-grams seems promising.

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

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

ACKNOWLEDGMENTS

This work has been partially supported by TrebleCLEF project, and the Education Council of the Regional Government of Madrid and the European Social Fund.
Special thanks are due to: Fernando Luis Costa, and two German students (Anna Kampchen and Julia Kramme) for taking care of the translations of the questions and the evaluation of the submitted runs for the Portuguese and German languages respectively.
Special thanks are also due to Cosmina Croitoru, a bright Romanian student whose help in the answers evaluation permitted to detect about 5 evaluation errors and some unevaluated answers in the RO-RO runs.
Our appreciation also to the advisory board: Donna Harman (NIST, USA), Maarten de Rijke (University of Amsterdam, The Netherlands), Dominique Laurent (Synapse Développement, France.)

REFERENCES

1. Eneko Agirre, Olatz Ansa, Xabier Arregi, Maddalen Lopez de Lacalle, Arantxa Otegi, Xabier Saralegi and Hugo Zaragoza. Elhuyar-IXA: Semantic Relatedness and Cross-lingual Passage Retrieval. Working Notes for the CLEF 2009 Workshop, 30 September-2 October, Corfu, Greece.

2. Santiago Correa, Davide Buscaldi and Paolo Rosso. NLEL-MAAT at CLEF-ResPubliQA. Working Notes for the CLEF 2009 Workshop, 30 September-2 October, Corfu, Greece.

3. Ingo Gloeckner and Bjoern Pelzer. The LogAnswer Project at CLEF 2009. Working Notes for the CLEF 2009 Workshop, 30 September-2 October, Corfu, Greece.

4. Adrian Iftene, Diana Trandabăţ1, Ionuţ Pistol, Alex-Mihai Moruz1, Maria Husarciuc1, Mihai Sterpu and Călin Turliuc. Question Answering on English and Romanian Languages. Working Notes for the CLEF 2009 Workshop, 30 September-2 October, Corfu, Greece.

5. Radu Ion, Dan Ştefănescu, Alexandru Ceauşu, Dan Tufiş, Elena Irimia and Verginica Barbu-Mititelu. A Trainable Multi-factored QA System. Working Notes for the CLEF 2009 Workshop, 30 September-2 October, Corfu, Greece.

6. Véronique Moriceau and Xavier Tannier. FIDJI in ResPubliQA 2009. Working Notes for the CLEF 2009 Workshop, 30 September-2 October, Corfu, Greece.

7. 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.

8. Anselmo Peñas, Alvaro 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.

9. Joaquín Pérez, Guillermo Garrido, Álvaro Rodrigo, Lourdes Araujo and Anselmo Peñas. Information Retrieval Baselines for the ResPubliQA Task. Working Notes for the CLEF 2009 Workshop, 30 September-2 October, Corfu, Greece.

10. Stephen E. Robertson and SteveWalker. 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.

11. 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. (to be published).

12. Álvaro Rodrigo, Joaquín Pérez, Anselmo Peñas, Guillermo Garrido and Lourdes Araujo Approaching Question Answering by means of Paragraph Validation. Working Notes for the CLEF 2009 Workshop, 30 September-2 October, Corfu, Greece.

13. Stephen Tomlinson, Douglas W. Oard, Jason R. Baron, Paul Thompson. Overview of the TREC 2007 Legal Track. Proceedings of The Sixteenth Text REtrieval Conference, TREC 2007, Gaithersburg, Maryland, USA, November 5-9, 2007.

14. María Teresa Vicente-Díez, César de Pablo-Sánchez, Paloma Martínez, Julián Moreno Schneider and Marta Garrote Salazar. Are Passages Enough? The MIRACLE Team Participation at QA@CLEF2009. Working Notes for the CLEF 2009 Workshop, 30 September-2 October, Corfu, Greece.

15. Rohit Bharadwaj, Surya Ganesh, Vasudeva Varma. A Naïve Approach for Monolingual Question Answering. Working Notes for the CLEF 2009 Workshop, 30 September-2 October, Corfu, Greece.