Evaluation of Question Answering Systems over European Legislation

Pamela Forner
Center for the Evaluation of Language and Communication Technologies (CELCT)
Via Alla Cascata, 56/c
38100 Povo –Trento, Italy
+(39) 0461 314 804
forner@celct.it

Anselmo Peñas
Spanish Distance Learning University, (UNED)
c/ Juan del Rosal, 16
28040 Madrid
(+34) 91 3987750
anselmo@lsi.uned.es

Danilo Giampiccolo
Center for the Evaluation of Language and Communication Technologies (CELCT)
Via Alla Cascata, 56/c
38100 Povo –Trento, Italy
+(39) 0461 314 874
giampiccolo@celct.it

ABSTRACT
This paper describes the first attempt at a new Question Answering (QA) evaluation track proposed at the Cross Language Evaluation Forum (CLEF) 2009, called ResPubliQA. Started in 2000, CLEF [2] has been mainly devoted to the organization of a series of evaluation tracks to test different aspects of cross-language information retrieval system development. Since 2003, also a Question Answering track has been carried out. This year, the ResPubliQA track focuses on the evaluation of the performances of QA systems dealing with the law domain. The exercise consists in extracting a relevant paragraph of text containing the answer to a given question from a set of legal-EU documents, i.e. a subset of the JRC-Acquis collection. The ResPubliQA track is meant to represent a first step inside CLEF towards the implementation of Information Access systems specifically dedicated to address the needs of the legal community.

Keywords
Question Answering, Evaluation, Legal Domain.

1. INTRODUCTION
The motivating goal of the ResPubliQA track [1] at CLEF is to make an initial step towards a new direction in QA research, starting from the point of view of real 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 domain is an issue of increasing importance given the vast amount of data which has become available in electronic form over the last few years. In fact, as stated in the Sedona Conference Best Practices Commentary on the Use of Search and Information Retrieval Methods in E-Discovery, "discovery has changed. In just a few years, the review process needed to identify and produce information has evolved from one largely involving the manual review of paper documents to one involving vastly greater volumes of electronically stored information." [14]

Moreover, the legal community has showed much interest in IR technologies as it has been 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 [12], 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). The task of the Legal Track is to retrieve all the relevant documents for a specific query and compare the performances of systems operating in a setting which reflects the way lawyers carry out their inquiries. This track has managed to attract IR researchers to the legal domain, and has gained more and more visibility over the years: the number of participants raised from 6 in 2006 to a total of 15 in the 2008 evaluation campaign.

The purpose of the ResPubliQA task is to introduce QA technologies into the field of e-discovery and to draw QA researchers' attention to the problems related to the retrieval of information in the legal domain.

A new approach to this research filed is indeed advisable because according to the Sedona best practices "the use of search and information retrieval tools does not guarantee that all responsive documents will be identified in large data collections, due to characteristics of human language. Moreover, differing search methods may produce differing results, subject to a measure of statistical variation inherent in the science of information retrieval." [14]

Question Answering evaluation has been a framework for testing text processing systems beyond document retrieval such as information extraction, knowledge acquisition, question analysis,

question expansion, answer extraction and validation, etc. These technologies are components of current QA systems that can be reconfigured to meet other user needs and other final applications. For example, QA systems search for very precise pieces of information that can be formulated in the corpus differently from the way the question is posed. If systems do not bridge this gap, no answers can be found. The use of dictionaries, ontologies, thesauri, or automatic terminology extraction techniques from the partial searching results, are examples of resources and techniques used in QA that can be useful also for e-Discovery. As pointed out in the Commentary to Principle 11 of the Sedona Principles, the "selective use of keyword searches can be a reasonable approach when dealing with large amounts of electronic data," and QA technologies can help find such keywords for querying the big amounts of material that lawyers have to manage [16].

The paper, which describes the preparation of the ResPubliQA campaign, is organized as follows: Section 2 gives a brief description of the task; Section 3 presents the different types of question developed for the creation of the ResPubliQA dataset; Section 4 shows the type of responses that systems are expected to return; Section 5 gives a brief explanation of how systems will be evaluated; and Section 6 highlights the challenges which are still to be addressed and some perspectives for future campaigns.

2. TASK DESCRIPTION

The purpose of the ResPubliQA track is to foster research on systems that answer a set of questions retrieving paragraphs of text rather than document lists as in the TREC Legal Track. The basic idea is that real users usually prefer to be given an answer in a context, so that they can have an evidence of its relevance, but do not like having to find the answer themselves in a list of documents.

In defining the task, the initial assumption is that the questioner is a lawyer or an ordinary person interested in making inquiries 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 paragraph level in Bulgarian, Dutch, English, French, German, Italian, Portuguese, Romanian and Spanish. This subset is available at the ResPubliQA website[2]. The fact that this corpus has parallel aligned translations in all languages gives the advantage of allowing the comparability of results. Anyway, as the alignment is not always perfect, it took a lot of extra work to ensure that each question had at least one answer in all languages.

Systems are evaluated against a pool of 500 independent questions that can be answered by one single paragraph in the collection.

Participating systems can perform the task in any of the following languages: Basque (EU), Bulgarian (BG), Dutch (NL), English (EN), French (FR), German (DE), Italian (IT), Portuguese (PT), Romanian (RO) and Spanish (ES).

When queries, target document collection, and responses are formulated in the same language, the sub-task is monolingual, meanwhile in the cross-language sub-tasks the document collection and the queries are expressed in two different languages, and the responses are in the language of the target corpus. In the ResPubliQA exercise, all monolingual and bilingual combinations of questions between the languages above are activated, including the monolingual English (EN) task – usually not proposed in the QA track at CLEF. Basque (EU) has been 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.

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 the contents, principles and political objectives of the EU treaties; EU legislation; declarations and resolutions, international agreements; acts and common objectives. Texts cover various subject domains, including economy, health, information technology, law, agriculture, food, politics and more.

This collection of legislative text 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 all the 9 languages involved in the track - Bulgarian, Dutch, 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. QUESTIONS

The test set is made up of a pool of 500 questions in all the languages involved in the task. These questions fall into the following categories:

1. Factoid
2. Definition
3. Reason
4. Purpose
5. Procedure

4.1 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?*

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

2 http://celct.isti.cnr.it/ResPubliQA/Downloads

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

P**: 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?*

P: The Official Journal of the Community shall be published in the four official languages.

4.2 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"?*

P: 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 the environmental policies?*

P: 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”.

4.3 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?*

P(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?*

P: 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.

4.4 Purpose

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

Q: *What is the purpose of the Agreement of Luxembourg?*

P 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?*

P: 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.

4.5 Procedure

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

Q: *How are the stable conditions in natural rubber trade achieved**?***

P: 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?*

P: 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?*

P: (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. RESPONSES

Participants can consider questions and target collections in any language. Each question must receive one of the following responses:

1. A paragraph with the candidate answer, or
2. The string `NOA` to indicate that the system prefers not to answer the question.

Each paragraph returned by the system is required to be an extract from a document in the parallel corpus.

Each paragraph is supposed to contain the answer to the question. The selected paragraph must provide enough context to make it clear for the human assessors whether the answer is indeed responsive or not.

6. EVALUATION

One of the principles that will guide the evaluation of the task is that leaving a question unanswered has more value than giving a wrong answer. In this way, the systems able to reduce the number of wrong answers, by deciding not to respond to the questions they are not sure of, will be rewarded by the evaluation measure. However, if a system chooses to leave some questions unanswered, returning `NOA` as a response, it must ensure that only the portion of wrong answers is reduced, maintaining the total number of correct answers it would return if it responded to all questions. A reduction in the number of correct answers will be punished by the evaluation measure.

The Answer Validation Exercise[4] [9-10-11] opened the development of the Machine Learning-based techniques able to decide if a candidate answer is finally acceptable or not. An improvement in the accuracy of this decision will lead 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.

One of the following judgements will be given to each answer by human assessors during the evaluation:

1. **AC**: the question is answered correctly
2. **AW**: the question is answered incorrectly
3. **U**: the question is unanswered

The unique measure considered in this evaluation campaign is the following:

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

where

n_{AC}: is the number of correctly answered questions

n_U: number of unanswered questions

n: the total number of questions

The interpretation of the measure is the following:

1. A system that gives an answer to all the questions will receive a score equal to the accuracy measure used in the previous QA@CLEF main task [3]: in fact, since in this case n_U is 0, $c@1 = n_{AC}/n$;
2. The unanswered questions will add value to c@1 only if they do not reduce the accuracy (i.e. n_{AC}/n) that the system would achieve responding to all questions. This can be thought as a hypothetical second chance in which the system is able to replace wrong answers with `NOA` leaving the percentage of correct answers unchanged.
3. A system that does not respond to any questions (i.e. returns only `NOA` as an answer) will receive a score equal to 0, as $n_{AC}=0$ in both addends.

7. STATUS OF THE CAMPAIGN

At the time this report was submitted, a set of 600 questions tailored on this subject had been created by human annotators. Then 500 questions had been selected by the track coordinators to be used in the real competition. Most of the questions had been written in natural language and then translated into all the other languages involved.

A small set of questions was made available for system training and development in December 2008, meanwhile the test question set is to be released in May 2009.
Runs submitted by participating systems will be assessed by human annotators during June. The results of the evaluation are expected by the middle of July.

8. FUTURE CHALLENGES

A number of issues are still to be addressed, namely (i) the size and heterogeneity of data sets; (ii) the pervasiveness of specialized language used in the legal collections which increases the difficulty of retrieving information; (iii) the necessity of better defining the nature of a legal search task; and (iv) the study of the modalities in which professional users concerned with legal matters search for information relevant to their work.

Moreover, the work done so far has helped us detect some challenges very difficult to sort out without the contribution of the legal community. In fact, some issues emerged during the preparation of the ResPubliQA evaluation campaign ask for the involvement of legal professionals in the definition of the task and in the development of benchmarks. The system developers participating in the ResPubliQA Track, are adapting the QA modules to the legal domain and by inviting professionals to participate in the definition of the evaluation task, we intend to better adjust the QA components in order to meet the information needs of the legal community, even if the final results are different applications from pure QA systems.

9. CONCLUSIONS

Apart from providing an innovative way of consulting large legal collections, one important achievement of the ResPubliQA track is the development of new interesting kinds of questions available in nine major European languages and, for the first time in QA@CLEF campaigns, presented together in a single dataset.

Most questions are related to the information need that a lawyer, rather than an ordinary person, would have, since legal experts have a better knowledge of legislation and a greater awareness of what they can expect as a result of their searches.

This confirms the necessity for the ResPubliQA track of having professional users from the legal community as advisors in the development of this task. In fact, an increased understanding of the specific needs of search related to legal issues, will allow (i) to find the appropriate user profiles to reflect its real nature and scope, and (ii) to create richer data sets and more realistic queries.

Besides, one of the final goals of the QA community at CLEF is to facilitate the technology transfer from research to industry, and also the ResPubliQA track aims at evaluating systems that may be made available for public use in the real world. In fact, some of the groups participating in the QA competitions are industrial companies which have already successfully developed commercial products for Information Access. A successful outcome of the ResPubliQA track will hopefully stimulate researchers to implement QA systems specifically dedicated to the needs of the legal community.

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

10. ACKNOWLEDGMENTS

ResPubliQA is a joint effort of several institutions and people beside UNED and CELCT: Corina Forascu (UAIC and RACAI, Romania), Nicolas Moreau (ELDA/ELRA, France), Petya Osenova (BTB, Bulgaria), Richard Sutcliffe (University of Limerick, Ireland), Iñaki Alegria (UBC, University of Basque Country, Spain), Álvaro Rodrigo (UNED, Spain).

Special thanks are due to the advisory board: Donna Harman (*NIST*, USA), Maarten de Rijke (University of Amsterdam, The Netherlands), Dominique Laurent (*Synapse* Développement, France.)

11. REFERENCES

1. http://celct.isti.cnr.it/ResPubliQA/

2. http://www.clef-campaign.org/

3. Overview of the CLEF 2008 Multilingual Question Answering Track. P. Forner, A. Peñas, I. Alegria, C. Forăscu, N. Moreau, P. Osenova, P. Prokopidis, P. Rocha, B. Sacaleanu, R. Sutcliffe, E. Tjong Kim Sang. 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)

4. Overview of the CLEF 2007 Multilingual Question Answering Track. D. Giampiccolo, P. Forner, J. Herrera, A. Peñas, C. Ayache, C. Forascu, V. Jijkoun, P. Osenova, P. Rocha, B. Sacaleanu, and R. Sutcliffe. In: C. Peters, V. Jijkoun, Th. Mandl, H. Müller, D.W. Oard, A. Peñas, and D. Santos, editors, *Advances in Multilingual and Multimodal Information Retrieval*, 8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007, Budapest, Hungary, September 19-21, 2007, Revised Selected Papers.

5. Overview of the CLEF 2006 Multilingual Question Answering Track. B. Magnini, D. Giampiccolo, P. Forner, C. Ayache, V. Jijkoun, P. Osenova, A. Peñas, P. Rocha, B. Sacaleanu, and R. Sutcliffe. 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.

6. Overview of the CLEF 2005 Multilingual Question Answering Track. A. Vallin, B. Magnini, D. Giampiccolo, L. Aunimo, C. Ayache, P. Osenova, A. Peñas, M. de Rijke, B. Sacaleanu, D. Santos, R. Sutcliffe. In: C. Peters, F. C. Gey, J. Gonzalo, H. Müller, G.J.F. Jones, M. Kluck, B. Magnini, M. de Rijke (Eds.): *Accessing Multilingual Information Repositories*, 6th Workshop of the Cross-Language Evaluation Forum, CLEF 2005, Vienna, Austria, 21-23 September, 2005, Revised Selected Papers.

7. Overview of the CLEF 2004 Multilingual Question Answering Track. B. Magnini, A. Vallin, C. Ayache, G. Erbach, A. Peñas, M. de Rijke, Paulo Rocha, K. Simov, and R. Sutcliffe. In C. Peters, P. Clough, J. Gonzalo, G. J. F. Jones, M. Kluck, B. Magnini (Eds.): *Multilingual Information Access for Text, Speech and Images*, 5th Workshop of the Cross-Language Evaluation Forum, CLEF 2004, Bath, UK, September 15-17, 2004, Revised Selected Papers.

8. The Multiple Language Question Answering Track at CLEF 2003. B. Magnini, S. Romagnoli, A. Vallin, J. Herrera, A. Peñas, V. Peinado, F. Verdejo, M. de Rijke. In C. Peters, J. Gonzalo, M. Braschler, M. Kluck (Eds.): *Comparative Evaluation of Multilingual Information Access Systems*, 4th Workshop of the Cross-Language Evaluation Forum, CLEF 2003, Trondheim, Norway, August 21-22, 2003, Revised Selected Papers

9. Overview of the Answer Validation Exercise 2008. Á. Rodrigo, A. Peñas, F. Verdejo. In 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)

10. Overview of the Answer Validation Exercise 2007. A. Peñas, Á. Rodrigo, F. Verdejo. 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.

11. Overview of the Answer Validation Exercise 2006. A. Peñas, A. Rodrigo, V. Sama, F. Verdejo. 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.

12. The TREC Legal Track: http://trec-legal.umiacs.umd.edu/

13. Overview of the TREC 2007 Legal Track. S. Tomlinson, D. W. Oard, J. R. Baron, P. Thompson, available at http://trec.nist.gov/pubs/trec16/t16_proceedings.html

14. The Sedona Conference Best Practices Commentary on the Use of Search & Information Retrieval Methods in E-Discovery. The Sedona Conference Journal Vol. 8, Fall 2007. pp.189-223

15. The Sedona Canada Principles Addressing Electronic Discovery (January 2008)

16. The Sedona Principles, Second Edition: Best Practices Recommendations & Principles for Addressing Electronic Document Production (The Sedona Conference® Working Group Series, 2007) ("The Sedona Principles, Second Edition, 2007"), available at www.thesedonaconference.org.

17. The Sedona Conference Glossary: E-Discovery & Digital Information Management (Second Edition) December 2007 available at www.thesedonaconference.org.

18. The Sedona Guidelines: Best Practice Guidelines & Commentary for Managing Information & Records in the Electronic Age (November 2007) available at www.thesedonaconference.org.