EQueR: the French Evaluation campaign of Question-Answering Systems

Christelle Ayache (1), Brigitte Grau (2), Anne Vilnat (2)

(1) ELDA
55/57, rue Brillat Savarin 75013 Paris
ayache@elda.org
(2) LIMSI-CNRS
Bât 508 Université Paris XI 91403 Orsay Cedex
grau@limsi.fr, vilnat@limsi.fr

Abstract

This paper describes the EQueR-EVALDA Evaluation Campaign, the French evaluation campaign of Question-Answering (QA) systems. The EQueR Evaluation Campaign included two tasks of automatic answer retrieval: the first one was a QA task over a heterogeneous collection of texts - mainly newspaper articles, and the second one a specialised one in the Medical field over a corpus of medical texts. In total, seven groups participated in the General task and five groups participated in the Medical task. For the General task, the best system obtained 81.46% of correct answers during the evalaution of the passages, while it obtained 67.24% during the evaluation of the short answers. We describe herein the specifications, the corpora, the evaluation, the phase of judgment of results, the scoring phase and the results for the two different types of evaluation.

1. Introduction

The EQueR Evaluation Campaign (Ayache, 2005) (Ayache et al., 2005) is part of the EVALDA project in the Technolangue program supported by the French Ministries in charge of Research, Industry and Culture. The EQueR Evaluation Campaign provides a general evaluation framework for Question-Answering systems for the French language. It aims at assessing the state of the arrt of this research activity in France and provide an up-to-date evaluation framework.

The EQueR Evaluation Campaign was divided into six main phases. The first phase aimed at specifying and producing the linguistic resources necessary for the evaluation campaign. The second phase aimed at setting up the scientific and technical environment essential to the evaluation test. The third phase consisted in having the participants carry out the tests. The fourth phase consisted of the collection and analysis of the results. The fifth was the organisation of the closing workshop. The sixth and final phase compiled the results for the production of a validated corpus and created a final evaluation package (containing all the data provided to the participants during the campaign).

The EQueR Evaluation Campaign included two tasks of automatic answer retrieval: the first one was a General task over a heterogeneous collection of texts and the second one a specialised one in the field of Medicine over a corpus of medical texts.

In this article, we describe the constitution of the test set, the second part explains the evaluation of the data and the last part describes the systems' results.

2. Test set

2.1. Collections of data

ELDA, the Evaluations and Language resources Distribution Agency, provided the collection of data to the participants three months before the evaluation test. The data is in three different formats: a source version (contains raw data), a clean version (contains simple tags: document identifier, title and paragraph) split according to the collection of data, and a clean version (without tags) split in document (a file = an article). Each document is ISO-Latin-1 coded (ISO-8859-1).

Below is an extract of the data.

```
<DOC>
<DOCID>LEMONDE95000001</DOCID>
<TITLE>Un commerce mondial mieux réglementé
</TITLE>
<P> AVEC l'année 1995, une nouvelle institution
voit le jour, qui devrait être porteuse de plus
de justice économique : l'Organisation mondiale
du commerce(OMC).<P>
</DOC>
```

Two collections of data have been created: the first one for the "General task", the other one for the "Medical task".

We decided to organise a Question-Answering evaluation in a specialised domain because it represents a different kind of real application. Even if questions are not all factual, a subset of them are and it was worth giving the opportunity to participants to test their systems on two kinds of tasks. This way, they could provide a comparison between the two kinds of corpora and their linguistic specificity.

2.1.1. General data (collection of texts)

The generic data (1.5 GB) comprises many years worth of newspaper articles from Le Monde and Le Monde Diplomatique, French Swiss news agency releases (SDA, Schweizerischen Depeschenagentur) and the French Senate's reports on various issues.

The whole corpus contains about 560000 documents ; about 460000 documents from Le Monde 1992-2000, 7800 from Le Monde Diplomatique 1992-2000, 65800 from SDA 1994-1995, and 570 documents from the French Senate's reports.

The size of documents varies according to the source data. The French Senate's reports in particular are very "long" in comparison to other collections. Some documents can have up to 20 times as many words as other documents.
The collection size is half the size of the TREC collection (3Gb) and at least four times bigger than CLEF specific collections.

2.1.2. Medical data (collection of texts)

The corpus of medical texts (approx. 140 MB) is composed of scientific articles and various references to "good medical practice". The texts were chosen by the CISMef team (Catalogue et Index des Sites Médicaux Francophones, http://www.cismef.org) of the Centre Hospitalier Universitaire de Rouen.
The initial formats of the Medical data are pdf and html files. The Medical data has been provided to the participants in the form of a single file with simple tags (document identifier, title and paragraph).

2.2. Questions

Five types of questions were given to participating systems. These types of questions differ according to the type of expected answers: "simple Factual", "Definition", "List", and "Yes/No".
"Simple factual" questions were divided into 7 sub-categories: person ("Who is the President of Chile?"), organization, date ("When was the Avignon festival?"), place, measure, manner and object/other. Some of these questions have been reformulated several times.
"Definition" questions must be answered with a definition and are divided into two sub-categories: person ("Who is Salvador Dali?") and organization ("What is NATO?").
"List" questions must be answered with a list containing as many items as requested in the question : "Which are the four main religions practiced in Hungary?", must be answered with 4 items.
"Yes/No" questions are answered only with Yes or No, and not with a corpus extract, but they must be justified by a corpus extract : "Is there a TGV railway line from Paris to Valencia?".
Questions without any possible answer in the collection of documents have been added to the "general" questions corpus. For this type of question, the system should provide the answer "NIL".
There are various sources and ways of creating questions. Some questions were created using key words from the newspaper articles and press releases whereas the rest was created by a group of potential users who were familiar with NLP. At least one correct answer in the corpus had been verified manually for each question given to participants.
For the general task, ELDA worked on a corpus of 500 questions that were grouped as follows: 407 "simple Factual" questions, 32 "Definition" questions, 31 "List" questions and 30 "Yes/No" questions.
For the specialised task, the CISMef team worked on a corpus of 200 questions that were divided as follows: 81 "Factual" questions ("What is the gene involved in aniridia?"), 70 "Definition" questions ("What is a mental illness?", 25 "List" questions ("What are the four major symptoms of ovarian cancer?"), and 24 "Yes/No" questions ("Is it possible for a child to be schizophrenic?").

2.3. Answers

For each question, the systems can provide either (a) a brief but precise answer, (b) a passage from a document (under 250 continuous characters extracted from a document in the corpus), and (c) the document ID that states which passage supports the answer, *or*, at least, (b) a passage and (c) the document ID.
For each type of question (except the «list» questions) the systems can provide up to five answers. Up to 20 answers for the «list» questions are allowed. We chose to limit the number of possibilities because giving more answers would be a paradox as Question-Answering systems aim at limiting the information provided by a search engine.
The answers should be arranged in the same order as the questions. For the «Yes/No» answers, the systems must be able to provide the passage that justifies them being either positive or negative.
We chose to give all these possibilities for enlarging the evaluation, as it was the first time such an evaluation had been done on the French language. We also wanted to provide a maximum of reusable data. This choice does not prevent comparisons with other evaluation campaigns, as it includes the same kinds of answers and the protocol for evaluating them remain the same (see Grau, 2004 for a presentation of the different evaluation campaigns in Question-Answering), with one exact answer per question justified by a document. Therefore, we have maintained a link to the existing evaluations in Question-Anwering.

3. Evaluation

The evaluation of each system took place at the participants' sites from the 16th to the 23rd of July 2004.

3.1. Exact answers and passages

Most of the participating systems provided a passage and a brief "exact" answer (only one group chose not to have their "exact" answers evaluated.) The two types of answers were evaluated separately.
The participants agreed on two types of assessments; one based on "short" answers and the other on passages. There were four assessments for the "short" answers; the answers were either "correct" (accurate and as precise as possible), "inexact" (accurate answer but not precise enough), "incorrect" (inaccurate answer) or "not justified" (accurate and precise answer but not supported by a document).
There are only two types of assessments for the evaluation of passages; the passage is "correct" (it contains the answer to the question and the document justifies the answer) or "incorrect."
There is no difference between the evaluation of short answers and passages. For the evaluation of «Yes/No» questions, the answer is "correct" if, and only if, the passage justifies the answer. Thus the passage is also judged "correct".
When a system provides the answer «NIL» the answer must be evaluated in the same way as a passage. First we must check to see if the correct response is in fact "NIL";

if this is so then the answer is "correct" and if not then it is deemed "incorrect."

3.2. Metrics used

The measure used for the "factual", "definition" and "Yes/No" questions was the Mean Reciprocal Rank (MRR) as in (Vorhees 2000). This method takes into account the number of times a certain accurate answer is found. Even if we find the same answer in many places we count it as one answer.

$$MRR = \frac{1}{\#questions} \sum_{i=1}^{\#questions} \frac{1}{answer_i\ rank}$$

The measure used for the "List" questions is the "Non Interpolated Average Precision" metric (NIAP, a rank-based metric used in TREC) (Vorhees 2000).
This criterion takes into account both the "Recall"[1] and "Precision"[2] (standard metrics as well as the order (rank) of the correct answers in the list.
The NIAP metric is defined as follows:
Starting from the highest ranked document, the actual relevant documents are counted. If the *i*th relevant document has rank ri, then:

$$niap = \frac{\sum \frac{i}{r_i}}{T}$$

where T is the total number of relevant documents in the test collection. For example, let us assume that there are three relevant documents in the test collection and a system assigns the ranks 2, 4, and 6 to these documents, then:

$$niap = \frac{1/2 + 2/4 + 3/6}{3} = 0.5.$$

4. Presentation of the results

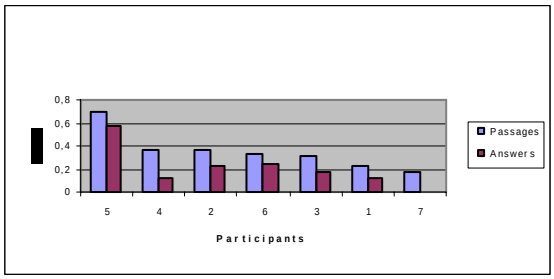
4.1. General task

Seven groups participated in the general task for EQueR. The four laboratories were: LIMSI, University of Neuchâtel, LIA (Laboration Informatique d'Avignon) in collaboration with iSmart and CEA-LIST/LIC2M. The three private institutions were: France Telecom R&D, Synapse Développement and Sinequa. In total, twelve runs were evaluated. Two judges assessed the results. Various discussions between judges led to a clearer idea of how to assess the runs.
Only 5 of the 500 questions in the corpus posed problems. We decided to omit those 5 questions from the corpus as well as from every submission file.

[1] Recall: the percentage of correct answers found in the list divided by the total number of correct responses that exist)

[2] Precision: the percentage of correct answers found divided by all the answers found)

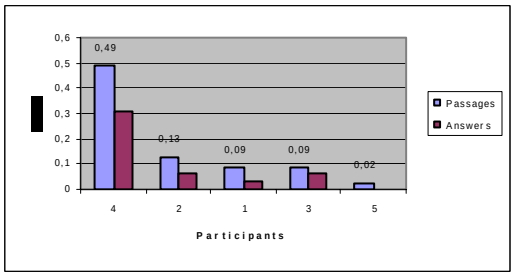
The remaining 495 questions were calculated; 400 «factual», 33 «definition», 31»Yes/No» and 32 «list» questions.
The three QA systems that obtained the best results for the general task in the EQueR/EVALDA 2004 project are the following:
For the passages: Synapse Développement (participant #5) ; Sinequa (participant #4) ; LIMSI (participant #2).
For the short answers: Synapse Développement ; LIA (partipant #6), and, then, LIMSI.
The results were provided to the participants on October 1st, 2004, as a set of recapitulative tables and graph.
Below is the graph of results for the general task for both the passages and short answers.



Graph 1: Results for the General task

4.2. Medical task

Five groups participated in the specialised medical task. There were three laboratories: University of Neuchâtel, CEA-LIST/LIC2M and AP/HP teaming with Paris XIII. There were two private institutions: France Telecom R&D and Synapse Développement.
In total, seven submission files were evaluated.
A specialised judge from the CISMeF team of the Medicine University of Rouen evaluated the results.
The scores were calculated for 200 questions divided as follows: 81 "Factual", 70 "Definition", 24 "Yes/No" and 25 "List" questions.
The three best systems for the specialised medical task were those of:
For the passages: Synapse Développement (participant #4) in 1st position ; the University of Neuchâtel (participant #2) in 2nd position, and, then, at the same level, AP/HP-Paris XIII (participant #3) and France Télécom R&D (participant #1);
For the short answers: Synapse Développement ; both AP/HP-Paris XIII and the University of Neuchâtel in 2nd position, and, in 3rd position, France Télécom R&D (participant #1).
Below is the graph of results for the general task.



Graph 2: Results for the Medical task

4.3. Results analysis

We observed that the systems obtained better results for the General task than for the Medical task.
The results for the General task range from 0.18 (for the system showing the worst results) to 0.7 (for the best system) according to the metric used, MRR (see paragraph 3.2) whereas the results for the Medical task vary from 0,02 (for the system with the worst results) to 0,49 (for the best system).
These results are probably due to the specific vocabulary required in the Medical domain.

Furthermore, each system obtained better results for the evaluation of passages than for the evaluation of short answers. It seems to be more difficult for a QA system to extract an exact short answer than it is for the passage (which is much longer) where there is a higher chance of finding the expected answer.

If we compare each participating QA system, we observe that each of them more or less used Natural Language Processing components.
Regarding the results, there was a considerable difference between the best system and the second best one. This applies to both tasks.
For the General task, we found it interesting to present to the participants the results according to the type of answers (person, organization, time…). Therefore, the participants would know which type of answers their system had difficulties with during the evaluation.
For all systems, the best results obtained were for the "Definition" questions, then the "Factual" questions, and then the "Yes/No" questions. Systems produced the worst results for the "List" questions.
More specifically, for the "Definition" questions, systems obtained the best results when the answer was an organization rather than a person.
For the simple "Factual" questions, systems obtained better results when the answer was a "Location", "Organization", "Person" or "Date" rather than "Manner", "Measure" or "Object".
For the General task, during the evaluation of passages, the best system obtained 81.46% correct answers compared to 51.07% for the second system.
During the evaluation of short answers, the results were slightly worse, with 67.24% correct answers for the best system and only 29.95% for the second system.
For the specialised task, the results regress even more. The best systems, during the evaluation of passages, obtained 62.85% correct answers and the second best system achieved 15.42%.
Finally, during the evaluation of short answers, the best systems obtained 40.57% correct answers and the next best system obtained 7.42%.
These results show a huge gap between the best system and the rest.

5. Conclusion and prospects

This paper details the principal aspects of the first evaluation of Question-Answering systems in France: EQueR.
There appears to be major interest in the project on the part of various academics and of "maestros" of this domain (there were 7 French participants and 1 Swiss participant). Some participants had never taken part in this kind of evaluation and certainly never the evaluation of a question-answer system.

EQueR has come up with an innovative type of question, the "Yes/No" question, which has sparked much interest in the participants.

EQueR is one of the few projects to draw upon the medical field for its question-answering tasks. This, in turn, has attracted many new participants who would like to be involved in specialised domains.

EQueR is linked to CLEF, a larger European project, which for two years has provided a specialised task for the evaluation of question-answering systems in Europe.
We compared the results of the best systems of both EQueR and CLEF QA task in 2004 (Valin et al., 2004) and found them to be consistent.

At the beginning of the project, we decided, along with the participants, to create a final evaluation package. This package will contain all the data provided to the participants during the campaign (campaign's guidelines, data (text corpora and question corpora) and tools). This package will allow anyone in the QA field to evaluate their system under the same conditions as those in EQueR. Users of this package will also be able to compare their results with the official EQueR results.
The first version of the EQueR package will be distributed soon by ELDA.

6. References

Ayache, C. (2005). Rapport final de la campagne EQueR-EVALDA, Evaluation en Question-Réponse. Site web Technolangue, http://www.technolangue.net/article61.html.

Ayache, C., Grau, B., Vilnat, A., (2005). Campagne d'évaluation EQueR-EVALDA : Evaluation en Question-Réponse. Actes de l'Atelier EQueR-EASY de TALN'05, Dourdan, France.

Grau B., (2004), Evaluation des systèmes de question-réponse, chap 3. dans Evaluation des systèmes de traitement de l'information, dir. S. Chaudiron, Hermes, pp. 77-98

Peters, C., Brashler, M., Gonzalo, J., Kluck, M., (2001). Evaluation of Cross-Language Information Retrieval Systems. number 2406 in LNCS.

Valin, A., Magnini B., et AL. (2004). Overview of the CLEF 2004 Multilingual Question-Anwering Track. In *Working Notes for the CLEF 2004 Workshop*, Bath, UK, p.281.

Vorhees, E., (2000). Overview of TREC-9 Question-Anwering task. In *Proceedings of Text REtrieval Conference 9.*