# Cross Lingual Question Answering using QRISTAL for CLEF 2007

Dominique Laurent, Patrick Séguéla, Sophie Nègre

Synapse Développement
33 rue Maynard,
31000 Toulouse, France
{dlaurent, p.seguela, sophie.negre }@synapse-fr.com

## Abstract

QRISTAL is a commercial question answering system making intensive use of natural language processing both for indexing documents and extracting answers. It ranked first in the EQueR evaluation campaign (Evalda, Technolangue [1]) and in CLEF 2005 [8] and CLEF 2006 [6] for monolingual task (French-French) and multilingual task (English-French and Portuguese-French). This year Synapse Développement only took part to the French monolingual run with a 54% of overall accuracy. This paper describes the improvements and changes implemented in Synapse Développement QA system since last CLEF 2006 campaign and details new features added to meet the challenges of this year's evaluation, that is sequence of question and the Wikipedia corpus.

## Categories and Subject descriptors

H.3 [Information Storage and Retrieval]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval; H.3.4 Systems and Software; H.3.7 Digital Libraries; H.2 [Database Management]: H.2.3 Languages—Query Languages

## General Terms

Measurement, Performance, Experimentation.

## Keywords

Question answering, French, Questions beyond factoids

## 1. Introduction

Synapse Développement took part in the 2005 [4] and 2006 [5] campaign of CLEF. During last year's campaign, we provided results both for monolingual (French to French) and bilingual tasks (English to French and Portuguese to French). This year, however, we just submitted our run for the monolingual French to French campaign.

The QA system we used for the campaign was an adaptation of our QA commercial product Qristal [3]. Qristal is a cross lingual question answering system for French, English, Italian, Portuguese, Polish and Czech. It was designed to extract answers both from documents stored on a hard disk and from Web pages by using traditional search engines (Google, MSN, AOL, etc.). Qristal is currently used in the M-CAST European project of E-content (22249, Multilingual Content Aggregation System based on TRUST Search Engine). Anyone can assess

Qristal online to ask questions on Wikipedia on : http://synapse.servehttp.com/ASP/ClientAsp_QristalSearch/WebFormWikipedia.aspx

Qristal integrates sophisticated Natural Language Processing components such as embedded ontologies, synonyms dictionaries, semantic relation extraction, coreference resolution, temporal contexts. The main originality of the architecture of our system is that it uses multi-criteria indexes [5]. While indexing, each document is divided into block of 1 kilobyte. Then, each block is analyzed syntactically and semantically, anaphora and metaphor are resolved. As a result, for each block, 8 different indexes are built : heads of derivation for common nouns, proper nouns, idioms, named entities, concepts, fields, QA extraction patterns and keywords.

Compared to last year, the architecture of our QA system remains roughly the same : after the question is submitted, it is categorised according to a question typology and pivots are identified. An automatic search in the index retrieves a set of potentially relevant blocks, containing a list of sentences related with the question. Blocks are sorted with weights and only 100 blocks are linguistically analysed for the selection of best sentences. Sentences are weighted according to their semantic relevance and similarity with the question. Next, through specific answer patterns, these sentences are examined once again and the parts containing possible answers are extracted. Finally, a single answer is chosen among all candidates.

The next section gives an overview of the main improvements of our QA technology over last year. Section 3 addresses the various adjustments we made to adapt to both novelties of this year's QA@CLEF campaign, namely the use of Wikipedia and the handling of topic related questions. Section 4 presents and discusses our results and section 5 concludes with future guidelines.

## 2. Improvement of the technology

### 2.1. Improvement of linguistic resources

As the product is being marketed, the linguistic resources need to be permanently updated.

Proper names dictionary was dramatically increased from 44 000 to 117 000 entries. This work exploited Wikipedia resources to list candidates, mostly works, enterprises and people's name. Thus, candidates were manually validated. We also add to proper names relevant information. For example, for a person, we mentioned its nationality, birthplace, birth date, date of death and main function. For a town, we mentioned its country, its region, its surface, its population and if its a capital or not. Note we only used the link between towns and countries and between persons and nationalities for this QA@CLEF campaign. Actually, a deeper use of those resources could have lead us to provide "unjustified" answers as it encourages the system to rank first a text including some answers even if the system did not find any clear justification of that answer in the text.

Together with our resources, our French syntactic analyser has noticeably been improved as we received the detailed results of the Easy campaign [7] where our analyser was awarded as one of the best tools available for French.

### 2.2. Handling of meta data

Our system now handles meta data. For a given document it stores its title, creation date, author and keywords. Note that dividing texts into blocks made it compulsory to store keywords for a given document. Isolated blocks cannot explicitly mention main subjects of the original text although sentences of these blocks relate to these subjects. Consequently, titles are considered as belonging to all blocks of texts.

We revised our algorithm for index search to take those meta data into account, that is to filter scanned blocks. For example, the analysis of the a question *Qui a reçu le prix Goncourt en 1995?* (*Who was awarded the Goncourt Prize in 1995?*) infers that the question is temporally restricted and its time environment is a date, i.e. the year 1995. Thus, the index search algorithm will give focus on documents created in 1995 while documents created before 1995 will be lowered.

Additionally, other small improvements were made, such as the implementation of new algorithms for searching noun phrases and related words in the index.

## 3. Coping with changes

This year's campaign introduced some new elements participants had to cope with : topic-related questions and articles from Wikipedia considered as an answer source.

### 3.1. Wikipedia searching

First decision was to simplify the Wikipedia corpus in removing XML tags and, on top of suggested exclusion proposed in the guidelines, suppress REDIRECT files. As previous campaigns were based on news collections,0 we here list some elements we identified as specific to Wikipedia.

Texts have very few redundancy. An information is often mentioned once in the overall corpus, the author making extensive use of hyperlinks. Basically, Wikipedia is influenced by the fact that it is meant to be read in a browser in a web environment. Moreover, it is very important to consider titles as the main subject of an article is often referred using anaphora.

Additionally, the document layout, nonexistent in news corpus, is very important in Wikipedia. Lots of information are in tables that are particularly difficult to handle for technology based on the analysis of lists of sentences. Typically, for a town, the population, land area or country will be stated in tables and not referred in the text of the article anymore. For example, table titles are often missing or replaced by dedicated codes (the tag "datedeces" is displayed "date de décès" (*date of death*)).

Finally, there are more spelling mistakes in Wikipedia than in news text. Diacritic signs are particularly neglected. The name of persons stands often in the title and in the very beginning of an article. Unfortunately, all first names of the person are mentioned where only one first name is commonly used. The page "Marivaux" is entitled and begins with "*Pierre Carlet de Chamblain de Marivaux*". The page about "Victor Hugo" begins as follows "Victor-Marie Hugo, né le 26 février 1802 à Besançon, mort..." ("Victor-Marie Hugo, born the 26th of February 1802 in Besançon, died..."). On both those pages our technology will recognise and index the complete named entity making it difficult for the algorithm to find the common usage (Marivaux, Victor Hugo) in this block.

### 3.2. Topic-related questions

Topic-related questions are clusters of questions which are related to the same topic and possibly contain anaphoric references between one question and the other questions. 76 questions out of 200 were topic-related for the French QA@CLEF 2007 run.

We implemented a special treatment to handle the fact that the question are topic-related. For example, last dates and geographic places (country, towns) encountered in a sequence were systematically added in the next question.

But our effort focuses on the detection of anaphora within a topic-related question set. Our overall technology encompass an anaphora resolution technique. The idea was both to organise an interface to treat questions and answers as a flow in a topic-related question set and improve the coverage of our anaphora resolution technique for those data. The anaphora resolution aimed at replacing reference by the word or noun phrase it referred to. This generated question is then treated normally by the QA process.

Qristal already resolved anaphora for pronouns and possessive adjectives. For this campaign, we developed an algorithm to handle demonstrative adjectives. When the syntactic analysis of a question detect a demonstrative adjective, we select the noun phrase it introduces and look at previous questions and answers for this noun phrase or synonyms. If we find one, we collect its extension and replace the demonstrative adjective and the noun phrase of the current question by this extended noun phrase. For example, in the question *Qui a réalisé ce film ?* (*Who directed this movie?*) the demonstrative adjective is "*ce*" (this), the noun phrase introduced is "*film*" (movie). Now let us consider previous answers and questions in this topic-related set : previous answer is *la palme d'or* (*Golden palm*) and previous question is *Quelle récompense le film Pulp Fiction a-t-il reçu lors du festival de Cannes?* (*What award did the film "Pulp Fiction" get at the Cannes Film Festival?*). The noun phrase "*film*" is then found and its extension is "*film Pulp Fiction*". Thus, we replace the demonstrative adjective and the noun phrase of the current question by the extended version to get the following question *Qui a réalisé ce film Pulp fiction?* (*Who directed this film "pulp Fiction"?*).

If the noun phrase and its synonyms are not found in the previous answers and questions set, we take the last noun phrase with the same semantic type. Semantic types can be abstract, concrete, human, animal or animated.

Anyway, either with this additional treatment, some anaphora were judged as too complicated to be handled as *Lequel d'entre eux découvrit l'uranium 235 ?* (*Which one of them discovered uranuim 235?*) coming after *Comment se prénommaient les deux frères Piccard ?* (*What were the first names of the two Piccard brothers?*).

Unfortunately the introduction of topic-related question makes it difficult for us to participate to cross language runs for Portuguese to French. As previously mentioned, we changed the interface of our linguistic modules that analyses the question. and thus made it incompatible with the Portuguese module of our partner Priberam [2]. Making it compatible was not much of a work but we did not managed to do it for the due time. Moreover, to resolve anaphora we had to consider both questions and answers. And thus, translate those answers in a cross language environment for topic-related questions.

## 4. Results

The results of this QA@CLEF campaign for our monolingual French run are as follows :

| | R | W | X | U |
|---|---|---|---|---|
| Total answers | 108 | 82 | 9 | 1 |
| Percentage | 54% | 41% | 4,5% | 0.5% |

Table 1. Results for a monolingual French run.

The general results of Table 1 show that there was a decrease regarding last year's French monolingual task. In 2006 campaign, we achieved 68% of correct answers and 64% in 2005.

As we think this drop could be related to the use of Wikipedia and checked whether questions had answers in the news corpus or in Wikipedia. It was clear that some questions could have answers only in one collection. Thus, we decided to separate questions in 4 sets, NIL questions, questions whose answer can be found in the news corpus, in Wikipedia or both.

Specific developments for the NIL questions had been implemented in CLEF 2006 resulting in a spectacular result of a 100% accuracy for the 9 NIL questions of CLEF 2007. So, NIL questions were not considered and only 191 questions are listed below.

| Corpus of the answer | Answers | R | W | X | U | Overall accuracy |
|---|---|---|---|---|---|---|
| News | 96 | 59 | 30 | 6 | 1 | 61% |
| News + Wikipedia | 21 | 16 | 4 | 1 | 0 | 76% |
| Wikipedia | 74 | 24 | 47 | 3 | 0 | 32% |
| Total | 191 | 99 | 81 | 10 | 1 | 51% |

Table 2. Results for different corpus.

Table 2 shows that redundancy is very important for a QA system based on linguistic analysis of sentences. When the answer appeared at least twice, the system reached an encouraging 76% of correct answer. Results on the news corpus, with a 61% of good results are comforting as well and much closer to last years campaigns. But, as foreseen in a previous section, the Wikipedia encyclopaedia revealed less suited for our QA based on pattern extractions than the news corpus.

Then we decided to check the distribution of errors along the main stages of our QA system.

| Stage | W+X+U | Failure % |
|---|---|---|
| Document retrieval | 11 | 12% |
| Selection of best sentences | 57 | 61% |
| Extraction of the answer | 15 | 17% |
| Anaphora | 9 | 10% |
| Total | 92 | |

Table 3. Distribution of errors along the main stages of our QA system.

Table 3 shows that for 11 questions, The Document Retrieval stage does not extract the blocks where an answer is mentioned. This is related with several issues :

- the number of one kilobyte blocks extracted is limited to 100. Therefore, for a common category of question like the definition category and very common words, numerous blocks get the same score. The question *Quelle est la plus grande banque du Japon* (*Which is the biggest bank in Japan?*), is a typical example of this. Here the category of the question is *definition* and searched pivots are the common noun *bank* and the proper noun *Japan.* Just for Wikipedia, 1166 blocks contain those 3 elements. Those blocks are ranked with meta data, occurrences of pivots and as a result 36 blocks are chosen. But then, the 64 following blocks are somehow randomly selected out of the 1130 remaining.

- some noun phrase or named entities are difficult to extract. This is particularly penalizing as words in the noun phrase or named entity are very common. For the question *Qui remporta le Tournoi des Cinq nations en 1994 ?* (*Who won the 5 nations championship in 1994?*), we do not recognize the named entity *Tournoi des Cinq nations* but the proper nouns *Tournoi* and *Cinq* and the noun *Nation.* Note that here capital letters on *Tournoi* and *Cinq*, very surprising in French indeed, surely make the correct recognition impossible.

- some links between questions and answers are too subtle. If we consider the question *Sur une montre, quelle aiguille représente les heures ?* (*On a watch, which needle represents the hour?*) the answer is "*...l'heure indiquée par la petite aiguille d'une horloge*" (*...the hour indicated by the small needle of a clock*). Our synonym dictionary provides us with the information that *représente* (*represent*) is a synonym of *indiquer* with a proximity of 10%. Unfortunately, *montre* (*watch*) is not mentioned as a synonym of *horloge* (*clock*). As all pivots are very frequent in the corpus, the block containing the answer is not possibly selected here.

As for the Selection of best sentences and the extraction of answer stage, many errors are connected with patterns of extraction that are difficult to fine tuned. Moreover we encounter some difficulties in writing extraction patterns for answers in long sentences or in a particular document layout like tables or enumeration.

Finally, 9 questions failed due to a bad resolution of anaphora resulting in a lack of core pivots making it impossible to find answer blocks. As a matter of fact, out of the 76 questions in sequence, only 43 contained an anaphora.

| | Total | R | W+X+U | Overall Accuracy |
|---|---|---|---|---|
| Anaphora recognised | 33 | 19 | 14 | 56% |
| Anaphora missed | 10 | 1 | 9 | 18% |
| Total | 43 | 20 | 23 | 46% |

Table 4. Results for question with anaphora.

Table 4 points out that the system described previously resolved 74% of the proposed anaphora. For those questions, results were similar to the overall results. Surprisingly, an answer was correctly found for an unrecognised anaphora. This question is *Qui était , avant cette chute, le président du régime communiste en Afghanistan ?* (*Who was the communist president before the fall?*). Here the anaphora is on the *fall* that should be replaced by *Communist regime collapse in Afghanistan*. However, all referred words in the anaphora are mentioned in the question (*Afghanistan, communiste, chute, régime*), thus, making the anaphora resolution not that important for the document retrieval state. And finally, for the Extraction of Answer stage, as the adjunct of time is not necessary here to find the answer.

If our overall results are globally inferior this year, with 54% to be compared to the 68% of 2006, it can be interesting to compare both results on the same category of questions, that is questions out of sequences where answers can be found in the news corpus.

| | Corpus of the answer | Total | R | W+X+U | Overall Accuracy |
|---|---|---|---|---|---|
| Out of sequence | News and<br>News + Wikipedia | 84 | 55 | 29 | 65% |
| | Wikipedia only | 31 | 10 | 21 | 32% |
| | Total | 115 | 65 | 50 | 56% |

| | | | | | |
|---|---|---|---|---|---|
| In sequence | News and<br>News + Wikipedia | 33 | 20 | 13 | 60% |
| | Wikipedia only | 43 | 14 | 29 | 32% |
| | Total | 76 | 34 | 42 | 44% |
| Total | | 191 | 99 | 92 | 51% |

Table 5. Comparing with last year's result.

Table 5 highlights that the decrease observed in our result is due to both the introduction of sequences and the Wikipedia corpus. If we consider the question out of sequences with possible answers on the news corpus, the overall accuracy reaches 65%. Actually, if we add the so called NIL question to those questions, we reach an encouraging 69% (55+9/84+9) of correct answer improving slightly last year's result.

## 5. Conclusions and Future Work

The analysis of the results lead us to identify a few issues that need to be solved.

We think that a measure between words could prevent us to miss some blocks if a noun phrase or a named entity was badly recognised.

In addition, we have to improve our results on the Wikipedia corpus. To begin with, we will try to take the most of the XML enrichment of the corpus we removed somehow too rapidly. For example, REDIRECT links could be used to enlarge our acronym databases and thus resolve automatically this category of questions.

Moreover, sequence treatment should be revised to consider all questions of the sequence as related to the same topic and therefore consider in a first place the blocks found for the previous question in the Document Retrieval stage of a new question. This work will be of particular interest for the introduction of QA treatment in a real (monolingual) oral dialogue with a user often seen as the future of QA systems.

## References

[1] AYACHE C., GRAU B., VILNAT A. (2005), Campagne d'évaluation EQueR-EVALDA : Évaluation en question-réponse, *TALN 2005*, 6-10 juin 2005, Dourdan, France, tome 2. – Ateliers & Tutoriels, p. 63-72.

[2] CASSAN A., FIGUEIRA H., MARTINS A., MENDES A., MENDES P., PINTO C., VIDAL D. (2006), Priberam’s question answering system in a cross-language environment. *CLEF 2006*, 20-22 september 2006, Alicante, Spain.

[3] LAURENT D., SÉGUÉLA P. (2005), QRISTAL, système de Questions-Réponses, *TALN 2005*, 6-10 juin 2005, Dourdan, France, tome 1. –Conférences principales, p. 53-62.

[4] LAURENT D., SÉGUÉLA P, NÈGRE S. (2005), Cross-Lingual Question Answering using QRISTAL for CLEF 2005, *CLEF 2005*, 21-23 september 2005, Wien, Austria. Available at http://clef.isti.cnr.it/2005/working notes/WorkingNotes2005/laurent05.pdf.

[5] LAURENT D., SÉGUÉLA P, NÈGRE S. (2006), Cross-Lingual Question Answering using QRISTAL for CLEF 2006, *CLEF 2006*, 20-22 september 2006, Alicante, Spain.

[6] MAGNINI B., GIAMPICCOLO D., FORNER P., AYACHE C., JIJKOUN V., OSENOVA P., PEÑAS A., ROCHA P., SACALEANU B., SUTCLIFFE R. (2006), Overview of the CLEF 2006 Multilingual Question Answering Track, *Working Notes of the Workshop of CLEF 2006*, 20-22 september 2006, Alicante, Spain.

[7] PAROUBEK, P., VILNAT A., ROBBA I., AYACHE C., Les résultats de la campagne EASY d'évaluation desanalyseurs syntaxiques du français, *Actes de la 14ème conférence annuelle sur le Traitement Automatique des Langues Naturelles (TALN),* juin 2007, Toulouse, France, vol 2, pp 243-252.

[8] VALLIN A., ] MAGNINI B., GIAMPICCOLO D., AUNIMO L., AYACHE C., OSENOVA P., PEÑAS A., DE RIJKE M., SACALEANU B., SANTOS D., SUTCLIFFE R. (2004), Overview of the CLEF 2005 Multilingual Question Answering Track, *Working Notes of the Workshop of CLEF 2005*, 21-23 september 2005, Wien, Austria.