Cross Lingual Question Answering using QRISTAL for CLEF 2005

Dominique Laurent, Patrick Séguéla, Sophie Nègre

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

Abstract

QRISTAL [8] is a question answering system making intensive use of natural language processing both for indexing documents and extracting answers. It recently ranked first in the EQueR evaluation campaign (Evalda, Technolangue [3]). This article proposes a functional description of the system. Then, it presents our results for the CLEF 2005 campaign and a critical description of the system. QRISTAL is possibly the first Question Answering system available on the consumer market. That fact generates drastic constraints and explains the technical choices we detail here.

1 Introduction

QRISTAL (French acronym for "Question Answering Integrating Natural Language Processing Techniques") is a cross lingual question answering system for French, English, Italian, Portuguese and Polish. 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.). To our knowledge, this system is the first to be marketed for general public. We are now integrating Qristal in professional applications. For example, Qristal is currently used in the M-CAST European project of E-content (22249, Multilingual Content Aggregation System based on TRUST Search Engine). Everybody can try the Qristal technology for French at www.qristal.fr. Note that the testing corpus for the testing web page is the grammar handbook proposed at http://www.synapse-fr.com.

For each language, a linguistic module analyzes questions and searches for potential answers. For CLEF 2005, the French, English, Portuguese and Italian modules were used for question analysis. Only the French module was used for answers extraction. The linguistic modules were developed by different companies. They share however a common architecture and similar resources (general taxonomy, typology of questions and answers and terminological fields).

For French, our system is based on the Cordial technology. It massively uses NLP tools, such as syntactic analysis, semantic disambiguation, anaphora resolution, metaphor detection, handling of converses, named entities extraction as well as conceptual and domain recognition. As the product is being marketed, it required a constant optimization of the various modules so that the software remains extremely fast. Users are now accustomed to obtain something that looks like an answer within a very short time, not exceeding three seconds.

2 Architecture

The architecture of the Qristal system is described on figure 1 :

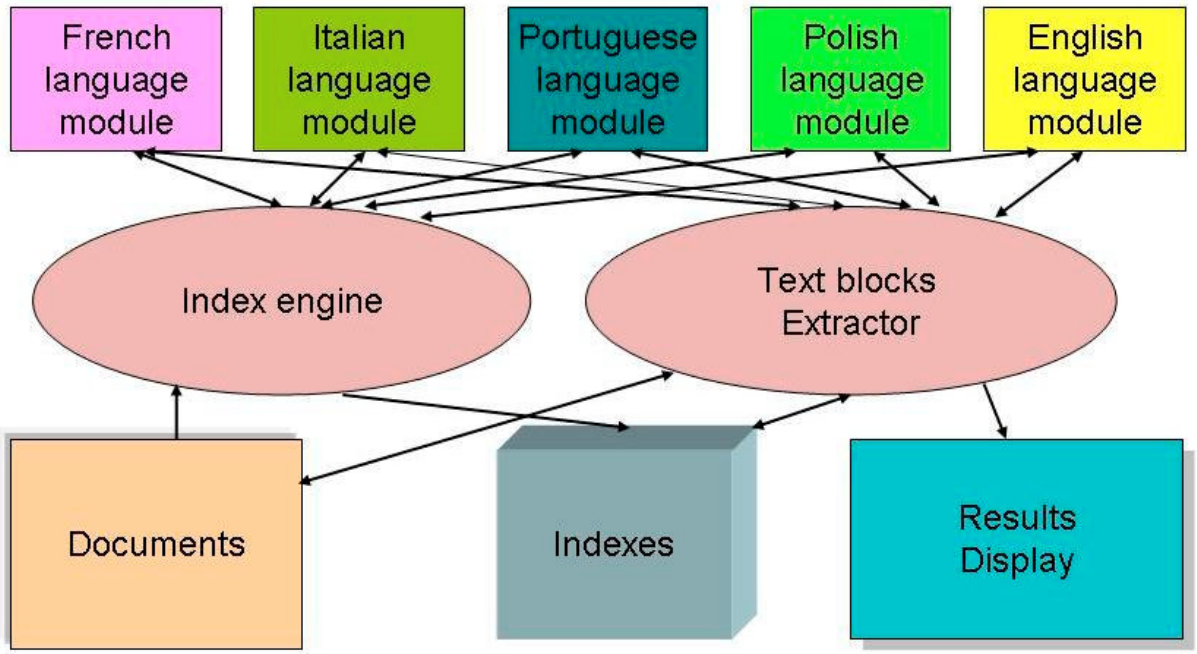

Figure 1. Architecture of the system

Qristal is a complete engine for indexation and answers extraction. However, it doesn't index the Web. Indexing is processed only for documents based on disks. Web search uses a meta-search engine we have implemented.

Our company is responsible for the indexing process of Qristal. Moreover, it ensures the integration and interoperability between all linguistic modules. Both English and Italian modules were developed by Expert System Company. The Portuguese module was developed by the Priberam Company which also takes part in CLEF 2005 for Portuguese monolingual. The Polish module was developed by the TiP Company. These modules were developed within the European project TRUST [9] (Text Retrieval Using Semantic Technologies). Note that currently, for another European project (M-CAST, Multilingual Content Aggregation System based on Trust Search Engine), the same partners develop a client-server version of this system in order to exploit digital resources of libraries.

2.1 Multicriteria indexing

While indexing documents, the technology automatically identifies the document language of and the system calls the corresponding language module. There are as many indexes as languages identified in the corpus. Documents are treated per blocks. The size of each block is approximately 1 kilobyte. Block limits are settled on the end of sentences or paragraphs. This size of block (1 kb) appeared to be optimal during our tests. Some indexes relate to blocks like fields or taxonomy whereas other relate to words, like idioms or named entities (see figure 2).

Each linguistic module processes a syntactic and semantic analysis for each block to be indexed. It fills a complete structure of data for each sentence. This structure is passed to the general processor that uses it to increment the various indexes. Figure 2 describes the linguistic processes launched while indexing, question analyzing and answer extracting. This description is accurate for the French module. Other language modules are very close to that framework but don't always include all its elements. For example, English and Italian modules do not include an indexing based on heads of derivation.

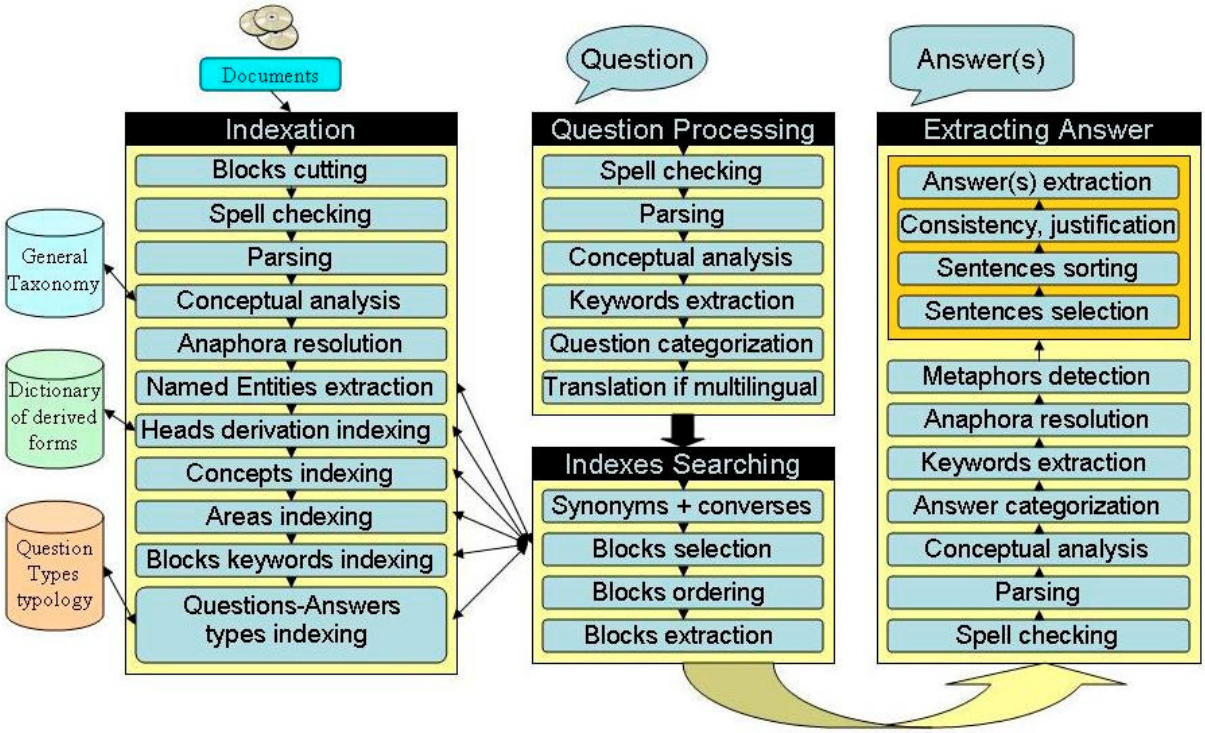

Figure 2. Linguistic analysis process for indexation

Texts are converted into Unicode. Then, they are divided into one kilobyte blocks. This reduces the index size as only the number of occurrences per block is stored for a given lemma. This number of occurrences is used to infer the relevance of each block while searching a given lemma in the index. In fact we here use lemmas but the system stores heads of derivation and not lemmas. For example, *symmetric*, *symmetrical*, *asymmetry*, *dissymmetrical* or *symmetrize* will be indexed in the same entry : *symmetry*.

Each text block is analyzed syntactically and semantically. Considering results of this analysis, 8 different indexes are built for:

- heads of derivation. A head of derivation can be a sense for a word. In French, the verb *voler* has 2 different meanings (*to steal* or *to fly)*. The meaning "dérober" (*to steal*) will lead to *vol* (*robbery*), *voleur* (*thief*) or *voleuse* (*female thief*). The second meaning, "se mouvoir dans l'air" (*to fly*), will lead to *vol* (*flight*), *volant* (*flying* as an adjective), *voleter* (*to flutter*) or *envol* (*taking flight)* and all its forms.
- proper names. If they appear in our dictionaries.
- idioms. Those idioms are listed in our idioms dictionaries. They encompass approximately 50 000 entries, like *word processing, fly blind* or *as good as your word*.
- named entities. Named entities are extracted from texts. *George W. Bush* or *Defense Advanced Research Project Agency* are named entities.
- concepts. Concepts are nodes of our general taxonomy. 2 levels of concepts are indexed. The first level lists 256 categories, like "visibility". The second level, actually the leaves of our taxonomy, lists 3387 subcategories, like "lighting" or "transparency",
- fields. 186 fields, like "aeronautics", "agriculture", etc.,
- question and answer types for categories like "distance", "speed", "definition", "causality", etc.,
- keywords of the text.

For each language, the indexing process is similar. Extracted data are the same. Thus, the handling of those data is independent of their original language. This is particularly important for cross language question answering.

For the French language, the rate of correct grammatical disambiguation (distinction between name-verb-adjective-adverb) is higher than 99%. The rate of semantic disambiguation is approximately 90% for 9 000 polysemous words and approximately 30 000 senses for these words. Note that this number of senses is markedly inferior to the Larousse one (Larousse is one of the most famous French dictionaries). Note however that our idioms dictionary covers a large number of the senses mentioned in this kind of dictionaries. The indexing speed varies between 200 and 400 Mo per hour with a Pentium 3 GHz, according to the size and number of indexed files.

Indexing question types is undoubtedly one of the most original aspects of our system. While the analysis of the blocks is being made, possible answers are located. For example, a name of function for a person (like *baker*, *minister*, *director of public prosecutions)*, a date of birth (like *born on April 28, 1958*), a causality (like *due to snow drift* or *because of freezing*), a consequence (like *leading to serious disruption* or *facilitating the management of the traffic*). This caused the block to be indexed like being able to provide an answer for a given question type.

Presently, our question typology includes 86 types of questions. Those types are divided into two subcategories: factual types and non factual types. Factual types are dimension, surface, weight, speed, percentage, temperature, price, number of inhabitants or work of art. Nonfactual types are form, possession, judgement, goal, causality, opinion, comparison or classification. For the EQueR evaluation [3] [8], 492 questions out of 500 were classified according to this typology with only 6 errors. For CLEF 2005, results were as follows:

	French	English	Italian	Portuguese
Good choice	95.5 %	87.0 %	74.5 %	91.5 %

Figure 3. Success rate for question type analysis

Building a keyword index for each text is also peculiar to our system. Dividing text into blocks made it compulsory. Isolated blocks cannot explicitly mention main subjects of the original text although sentences of these blocks relate to these subjects. The keyword index makes it possible to add contextual information about the main subjects of the text for blocks. Keywords can be a concept, a person, an event, etc.

1.2 Answer extraction

After the user typed its question, it is syntactically and semantically analyzed by the system. Question type is inferred. We would like here to draw the attention to the fact that questions are shorter than texts. This lack of context makes the semantic analysis of the question more dubious. That's why the semantic analysis processed on the question is more comprehensive than the analysis processed on texts. Moreover, users have the possibility to interactively force a sense. This possibility, however, was not used for CLEF as the entire process was automatic. The result of the semantic analysis of the question is a weight for each sense of each word recognized as a pivot. For example, sense 1 is recognized with 20%, sense 2 with 65% and sense 3 with 15%. This weight, together with synonyms, question and answer types or concepts, is considered while searching the index. Thus all senses of a word are taken into account during the index search. This prevents from dramatic consequences due to errors in the semantic disambiguation while making the most of good analysis.

After question analysis, all indexes are searched and the best ranked blocks are analyzed again. As one can notice on figure 2, the analysis of the selected blocks is close to the analysis processed while indexing or question analyzing. On top of this "classic" analysis, a weight for each sentence is inferred. This weight is based on the number of words, synonyms and named entities found in this sentence, the presence of an answer corresponding to the question type and a correspondence between the fields and domain.

After this analysis, sentences are ranked. Then, an additional analysis is processed to extract named entities, idioms or lists that match the answer. This extraction relies on the syntactic characteristics of those groups.

For a question on a corpus located on a hard disk, the response time is approximately 3 seconds with a Pentium 3 GHz. On the Web, first answers are provided after 2 seconds. Then the system computes a progressive refining during ten seconds, according to user's parameters like the number of words, the number of analyzed pages, etc.

We tested many answer justification modules, mostly implemented from Web [4], [7] or [13]. Our technology enables, as an option, to use such a module of justification. It consists in searching the web with the words of the question looking for potential answers the system inferred. However this process is seldom selected by users as it increases the response time of a few seconds. It was not used in CLEF 2005 either. The only justification module we used was an internal module which makes the most of the semantic information for proper names enclosed in our dictionaries. For more than 40 000 proper names, we possess information about the country of origin, the year of birth and death, the function for people, country, the area and population for a city, etc. We think this justification module is at the origin of some "unjustified" answers. As a matter of fact, it caused the system to rank first a text including the answer even if the system did not find any clear justification of that answer in the text.

For cross language question answering, English is used as pivot language. The fact that most users are only interested in documents in their own language and English motivated that choice. Thus, for cross language answering, the system processes generally only one translation. For this evaluation, both Portuguese to French and Italian to French runs required two translations: from source language to English and then from English to French. QRISTAL does not use any Web Services for translation because of response time. Only words or idioms recognized as pivots are translated.

3 Results for CLEF 2005

QRISTAL was evaluated for CLEF 2005 for French to French, English to French, Portuguese to French and Italian to French. That is 1 monolingual and 3 multilingual campaigns. For each one of these tasks, we processed only one run. Note that results obtained in CLEF 2005 could have been obtained with the commercial version of our Qristal software, in the version of May 2005.

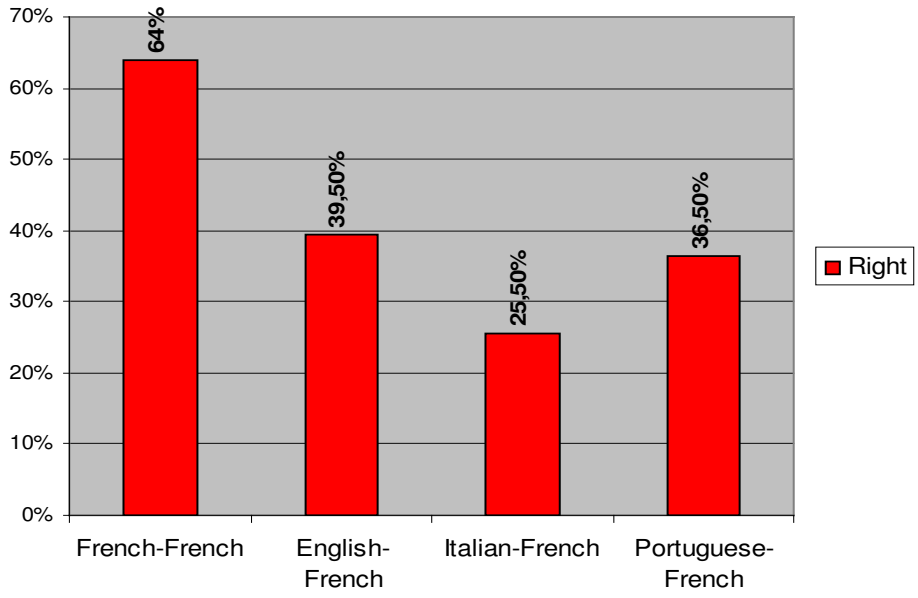

Figure 3. Results of the general task

For French to French, these results are better than those we obtained for the EQueR campaign. That is 64% for CLEF 2005 and 52% for EQueR 2004. Compared to EQueR, CLEF proposed additional difficulties as an exact answer is required. Only a passage of 50 maximum characters was asked for EQueR. On the contrary it seems the level of difficulty of the questions of CLEF is lower than those of EQueR which included for example research of lists as answers. This overall improvement of our performances validates the developments we did last year and particularly for the extraction of definitions.

Results per category are as follows:

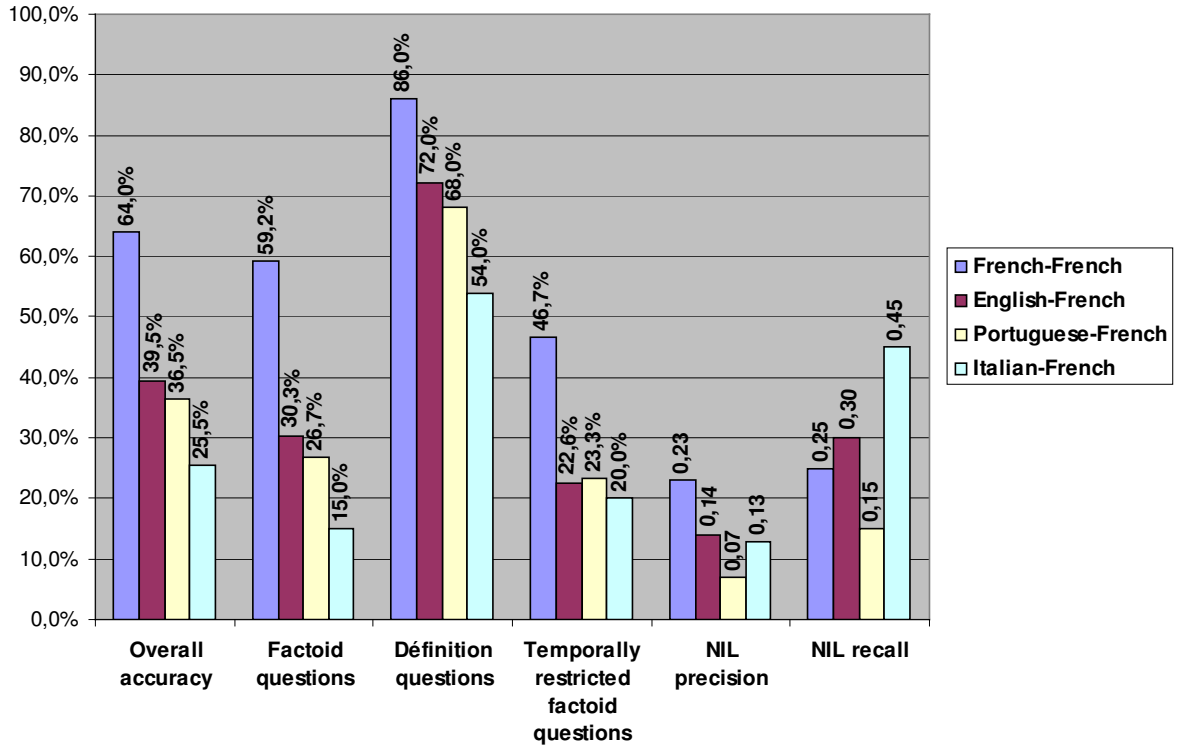

Figure 4 : Results of our 4 runs for each question type

One can notice the quality of answers is lower for cross lingual answering. In fact, the English to French run finds approximately 60% of the answers found in the French to French run. The Portuguese to French run finds 57% of the French to French run. Because of a bug we didn't manage to solve during the CLEF campaign, the Italian module did not translate pivot words to English. Thus, for the Italian to French run, results were computed only using words common in both languages. Fortunately, most proper names are identical between French and Italian. Anyway, taxonomy, question types and fields were used for this Italian to French run.

Precision and recall for NIL questions are quite poor. Nil questions are questions without any existing answer in the corpus. Our search engine was designed to find answers (!) and strategies dedicated to the detection of questions without any answer are quite unsatisfactory. Actually, the main routine in this process compares the possible date of the question and the date of text to remove texts anterior to the date of the question.

Figure 4 presents statistics for answers evaluated as 'R' that stands for right. But CLEF proposed two other qualifications for answers that is 'U' for unjustified and 'X' for inexact. We think 'U' and 'X' answers would be often accepted by users, even 'X' answers if they are presented with their context. For question 57 *Qui est Flavio Briatore ?* (*Who is Flavio Briatore?*), the answer provided by our system was *directeur général de Benetton Formula* (*general manager of Benetton Formula*), whereas the awaited answer was *directeur général de Benetton Formula 1* (*general manager of Benetton Formula 1*). Likewise, for question 96 *A quel parti politique Jacques Santer appartient-il ?* (*Which political party does Jacques Santer belong to ?*), the provided answer by Qristal was *Parti chrétien-social dès 1966* (*Christian Social Party since 1966*) whereas the awaited answer was *Parti chrétien-social* (*Christian Social party*). This lead us to consider statistics for all answers considered as "not wrong", that is right (R), unjustified (U) or inaccurate (X):

	French-French	English-French	Portuguese-French	Italian-French
Not wrong (R+U+X)	138 (69.0%)	92 (46.0%)	85 (42.5%)	64 (32.0 %)

The difference between Italian and both English and Portuguese was explained before. It is due to the fact that no translation is processed from Italian to English. For the other language pairs, we tried to determine the reasons of that gap. Thus, we scanned all translations and list words that had inadequate or inexistent translations. Results of this work are presented in the following table.

	English-French	Portuguese-French	Italian-French
Total of pivot words	623	701	769
Not translated from source language to English		47	451
Not translated from English to French	4	6	
Badly translated from English to French	55	63	
Total for translation mistakes	59 (9.5%)	116 (16.5 %)	451 (58.6 %)

This table underlines the relation between errors in translation and the performance of the multilingual system. Portuguese module is rather more precise than the English module but it is penalized by the double translation, from Portuguese to English and then from English to French. For example, for question 11 (*Which French Prime Minister committed suicide?*) *Prime* and *minister* are translated into French as *prime* and *ministre* quite far from the correct translation that is *premier ministre* !

Then we had a closer look to questions where the monolingual process finds the answer but the cross language does not. This leads us to the following remark. Questions are defined by reading the corpus and, deliberately or not, people formulating questions tend to reuse words or expressions mentioned in the text of the identified answer. On one hand, this influences the capacity of the system and the importance of each module in the overall process. For example, the use of synonyms is not that important for CLEF as it normally is. On the other hand, for cross language question answering, translations can be fuzzy and potentially quite far from the targeted word or expression especially when one uses English as an intermediate language. This way, translated words are quite often different from the terms mentioned both the question and the answer.

For question 10 *Quel professeur de Bonn a reçu le prix Nobel d'économie ?* (*Which professor from Bonn received the Nobel Prize for Economics?*), the French module extracts the noun phrase *prix Nobel d'économie* (*Nobel Prize in economy*). In English, translation provides the words *professeur*, *Bonn*, *recevoir*, *Nobel*, *Prix* and *économie* but neither *prix Nobel* nor *prix Nobel d'économie*.

Consider question 55 *Quel poste tenait Silvio Berlusconi avant qu'il ne démissionne*? for French, *What minister was Silvio Berlusconi prior to his resignation* for English and *Que ministério ocupava Silvio Berlusconi antes da sua demissão?* for Portuguese. The word *minister* in English and *ministério* translated into *minister* by the Portuguese module are translated into French as *ministère* (*ministry*). But the answer provided by the French module is *président du Conseil italien* extracted from "Décidant finalement de renoncer à demander un vote de confiance au Parlement, le président du conseil italien, Silvio Berlusconi, a remis, jeudi 22 décembre dans l'après-midi, sa démission au président de la République, Oscar Luigi Scalfaro." One can notice that the answer has nothing to do with a ministry. Moreover, the English word *resignation* correctly translated into French as *resignation* is quite far from the word *démission* (*abdication*).

To determine the various roles of every part of our system, we disconnected some modules and measured the overall performance. With the 200 questions of CLEF, the most important module was the question and answer type extractor. Disconnecting it causes an 11.5% drop for the monolingual campaign. This module is used to index, to search the index, and to extract exact answers from blocks. Other modules have secondary roles. We noted a drop of 2% by disconnecting synonyms dictionaries and a 3% drop by giving the same weight to all pivots. Note however that main components such as the part of speech tagger are almost impossible to test separately as all the system relies on it.

Priberam, the company responsible for the Portuguese module in our engine, participated in CLEF 2005 in the Portuguese to Portuguese evaluation track. It is interesting to note that they obtained results very similar to our results for the monolingual run [1] [2]. This seems to objectively validate our common choices and our resulting

technology that is very close to the best systems available for English which have participated for many years now in this type of evaluation via the TREC evaluation campaigns [6] [14].

4 Outlines

For 4 years now we sell out our question answering system on the French market. According to user's reactions and remarks we established a list of compulsory elements for such a system not to be directly rejected by users:

- Response time must not exceed 3 seconds, and preferably 1 or 2 seconds. By any means, a first answer must be displayed within this period.
- Success rate must approach 100%. A system providing only one answer out of 3 or less is acceptable only by a reduced number of users.
- Questions like those used for TREC or CLEF represent only a part of user's questions. Users often types "why" or "how" questions. Such a system has to handle it correctly.

An 80% of correct answers for monolingual search and 60% for multilingual is a minimum to convince a large audience of the interest of question answering systems. To reach such a level of performance, while reducing response time, very few approaches are available. Unfortunately, it is not possible to use strategies based on the Web redundancy [10] [12]. Actually, user's search often relates to specific corpus not redundant on the Web. Moreover, parallel research strategies or parallel justification using the Web take too long and are therefore inadequate.

However, we identified the following approaches to improve our technology:

- A general improvement of our resources. Particularly the semantic disambiguation process and the translation dictionaries.
- A refinement of the typology of questions and a precise definition of named entities expected in the answer for various question types would be of higher interest.
- An improvement of the answer delimitation process would reduce the number of "inaccurate" answers.
- Handling the presentation of documents. Documents are considered as rough text, without taking into account possible tags, titles, paragraphs, bolded parts, etc.
- The use of databases for question answering. The best translation systems use translation memory. Question answering systems could use memory as well. This process would imply the construction of a database storing factual predicates. This construction could be based on an automatic analysis of the Web. We could imagine storing questions and related answers users ask the system as well. A major interest of such databases is that accessing them is very quick. Thus, by using them, the system can process an answer instantly and then search for justification - or invalidation - in documents of the corpus.
- A specific handling for questions that cannot have an answer on the Web or in the corpus. Questions like *What is the weather forecast for tomorrow?* cannot be addressed by question answering systems based on corpus analysis. For those questions, you would rather send a well formatted request to weather forecast dedicated web sites than to general search engines. Same remarks concerning questions about itineraries or prices for which specialized sites are more likely to generate an appropriate answer, if there is at least one answer to the question, of course !.

As treating non factual question is of higher interest we encourage evaluation campaigns such as CLEF to take them into account [5]. Note that the evaluation of those questions and answer pairs is more difficult.

Analyzing the questions and user's reactions in front of provided answers underlines the gap between the words used in the question and those mentioned in the expected answer. This especially occurs for questions like *what do clients think about the Large hotel in Berlin?* or *which guarantees offer my blue card?* Lexical (synonyms, analogies) or thematic assistances could help the user to better find expected answers.

More technically, the M-CAST European project (http://www.m-cast.infovide.pl) in which our company as well as our Italian, Portuguese and Polish partners take part will enable us to test our system in a professional environment.

5 Conclusion

QRISTAL is the first question answering system marketed for general public and professionals. Results obtained for the EQUER evaluation, largely confirmed by those we obtained during CLEF evaluation, show that the intensive use of NLP technologies for analyzing the question, indexing texts and extracting answers leads to a good outcome. The fact that Priberam, which uses the same framework, obtained very close results for Portuguese encourages this assertion.

Our results, to be compared to the best international prototypes, can however be regarded as insufficient, particularly when it comes to Web searching. As we use a meta-search engine, that is an engine using the results of other search engines of the Web, we have no control over the indexation process. Moreover, downloading pages provided by Web search engines is time consuming and therefore limited. Thus, Web searching with our present technology is not as accurate as corpus searching.

One can assume that present boolean search engines will be replaced by NLP systems in several years. Nevertheless, demonstrating the advantages of such systems and revealing present weaknesses of boolean search engines will take quite a long time, partly because experts in this domain often are Boolean search experts!

Acknowledgments

The authors thank Bruno Wieckowski and all engineers and linguists that took part in the development of QRISTAL. They also thank the Italian company Expert System and the Portuguese company Priberam for allowing them to use their modules for question analysis in English, Italian and Portuguese. They finally thank the European Commission which supported and still supports our development efforts through TRUST and M-CAST projects.

Last but not least, authors thank Carol Peters, Alessandro Vallin, Danilo Giampiccolo and Christelle Ayache for the remarkable organization of CLEF.

References

[1] AMARAL C., LAURENT D., MARTINS A., MENDES A., PINTO C. (2004), Design & Implementation of a Semantic Search Engine for Portuguese, *Proceedings of the Fourth Conference on Language Resources and Evaluation*.

[2] AMARAL C., FIGUEIRA H., MARTINS A., MENDES A., MENDES P., PINTO C. (2005), Priberam's question answering system for Portuguese, *Working Notes for the CLEF 2005 Workshop,* 21-23 September, Wien, Austria.

[3] 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.

[4] CLARKE C. L. A., CORMACK G. V., LYNAM T. R. (2001), Exploiting Redundancy in Question Answering, *Proceedings of 24th Annual International ACM SIGIR Conference (SIGIR 2001)*, p. 358-365.

[5] GRAU B.. (2004), L'évaluation des systèmes de question-réponse, *Évaluation des systèmes de traitement de l'information,* TSTI, p. 77-98, éd. Lavoisier.

[6] HARABAGIU S., MOLDOVAN D., CLARK C., BOWDEN M., WILLIAMS J., BENSLEY J. (2002), Answer Mining by Combining Extraction Techniques with Abductive Reasoning, *Proceedings of The Twelfth Text Retrieval Conference (TREC 2003)*.

[7] JIJKUN V., MISHNE G., DE RIJKE M., SCHLOBACH S., AHN D., MÜLLER K. (2004), The University of Amsterdam at QALEF 2004, *Working Notes of the Workshop of CLEF 2004*, Bath, 15-17 september 2004.

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

[9] LAURENT D., VARONE M., AMARAL C., FUGLEWICZ P. (2004), Multilingual Semantic and Cognitive Search Engine for Text Retrieval Using Semantic Technologies, *First International Workshop on Proofing Tools and Language Technologies*, Patras, Grèce.

[10] MAGNINI B., NEGRI M., PREVETE R., TANEV H. (2002), Is It the Right Answer? Exploiting Web Redundancy for Answer Validation, *Proceedings of the 40th Annual Meeting of the ACL*, p. 425-432

[11] MAGNINI B., VALLIN A., AYACHE C., ERBACH G., PEÑAS A., DE RIJKE M., ROCHA P., SIMOV K., SUTCLIFFE R. (2004), Overview of the CLEF 2004 Multilingual Question Answering Track, *Working Notes of the Workshop of CLEF 2004*, Bath, 15-17 september 2004.

[12] MONZ C. (2003), From Document Retrieval to Question Answering, *ILLC Dissertation Series 2003-4*, ILLC, Amsterdam.

[13] NEUMANN G., SACALEANU B. (2004), Experiments on Robust NL Question Interpretation and Multi-layered Document Annotation for a Cross-Language Question/Answeringg System, *Working Notes of the Workshop of CLEF 2004*, Bath, 15-17 september 2004.

[14] VOORHEES E. M.. (2003), Overview of the TREC 2003 Question Answering Track, NIST, 54-68 (http://trec.nist.gov/pubs/trec12/t12_proceedings.html).