Quantum, a French/English Cross-language Question Answering System

Luc Plamondon George Foster
RALI, Université de Montréal
C.P. 6128, Succ. Centre-Ville
Montréal, Québec, Canada H3C 3J7
{*plamondl, foster*}*@iro.umontreal.ca*

Abstract

We describe a method for modifying a monolingual English question answering system to allow it to accept French questions. Our method relies on a statistical translation engine to translate keywords, and a set of manually written rules for analyzing French questions. The additional steps performed by the cross-language system lower its performance by 28% compared to the original system.

1 Introduction

A question answering (QA) system can be described as a particular type of search engine that allows a user to ask a question using natural language instead of an artificial query language. Moreover, a QA system pinpoints the exact answer in the document, while a classical search engine returns entire documents that have to be skimmed by the user.

Clarke [Clarke *et al.*, 2002] has shown that, for document collections smaller than 500 GB (100 billion words), the bigger the size of the collection, the better the performance of their QA system. If we suppose that an English speaker has access to about 10 times more digital documents — webpages, encyclopaedias on CDs, etc. — than a French speaker (estimation based on the number of pages on the web, see Fig. 1), there is no doubt that a QA system designed for French speakers but able to search English documents would open new possibilities both in terms of the quantity of topics covered and the quality of the answers.

We had previously developped the Quantum QA system [Plamondon *et al.*, 2002] for the TREC evaluation campaigns. This system operates in English only: the question must be asked in English, the document collection is in English and the answer extraction is performed in English. For the purpose of a pilot project conducted with the National Research Council of Canada [Plamondon and Foster, 2003], we transformed Quantum into a bilingual system to allow French speakers to ask their questions in French and to get answers in French as well, but using an English document collection. We entered the cross-language QA track at CLEF 2003 with this bilingual system without further modifications.

2 Monolingual English System

Quantum was developped primarily for the TREC evaluation campaigns. It was designed to answer simple, syntactically well-formed, short and factual English questions such as *What is pilates? Who was the architect of Central Park? How wide is the Atlantic Ocean? At what speed does the Earth revolve around the sun? Where is the French consulate in New York?* The document collection from which the answers are extracted are news from major newswires. For

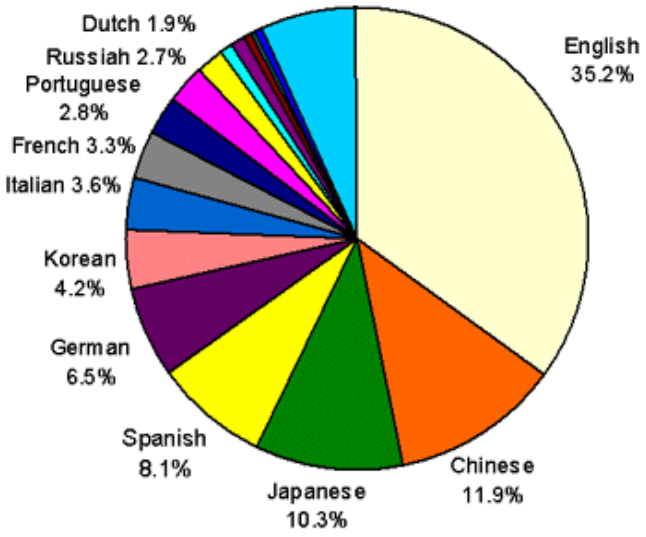

Figure 1: Online language populations (March 2003), on a total of 640 million webpages. Source: http://www.glreach.com/globstats

more details on the track and the system requirements, see the description of the TREC-11 QA track [Voorhees, 2002].

The architecture of the Quantum monolingual system is shown on Fig. 2, along with a sample question from the CLEF set. In the following sections, we describe only the elements that are relevant to the modifications we made in order to make the system cross-lingual (Sect. 3).

2.1 Question Analysis

The goal of the question analysis phase is to determine the expected type of the answer to a particular question, and thus to determine the answer extraction mechanism — or extraction function — to use. Some of the extraction functions require an additional parameter called the question's *focus*. The focus is a word or group of words that appears in the question and that is closely related to the answer. For instance, the answer to *With what radioactive substance was Eda Charlton injected in 1945?* should be an hyponym of the question's focus *substance*. The answer to *In how many countries does Greenpeace have offices?* should contain a number followed by a repetition of the focus *countries*. Some types of questions such as *When was the Bombay Symphony Orchestra established?* do not require the identification of a focus because, in this case, we look for the time named entity that *when* stands for. All the words of the question, whether they are part of the focus or not, play a role in the process of finding the answer, at least through the information retrieval score (Sect. 2.2). We stress that our classification of questions, the interpretation of the question's focus and whether an extraction function requires a focus or not are all motivated by technical considerations specific to Quantum. A more rigorous study of questions based on psycho-linguistic criteria has been made by Graesser [Graesser *et al.*, 1992].

Before Quantum can analyze a question, it must undergo several operations: tokenization, POS-tagging and NP-chunking. The analysis itself is performed via a set of 60 patterns and rules based on words, part-of-speech tags and noun-phrase tags. For example, Quantum uses the following pattern and rule to analyze the question in Fig. 2:

$$\texttt{how many}\ \langle\text{noun-phrase NP}_1\rangle\ \rightarrow\ \text{type} = \textit{cardinality},\ \ \text{focus} = \text{NP}_1$$

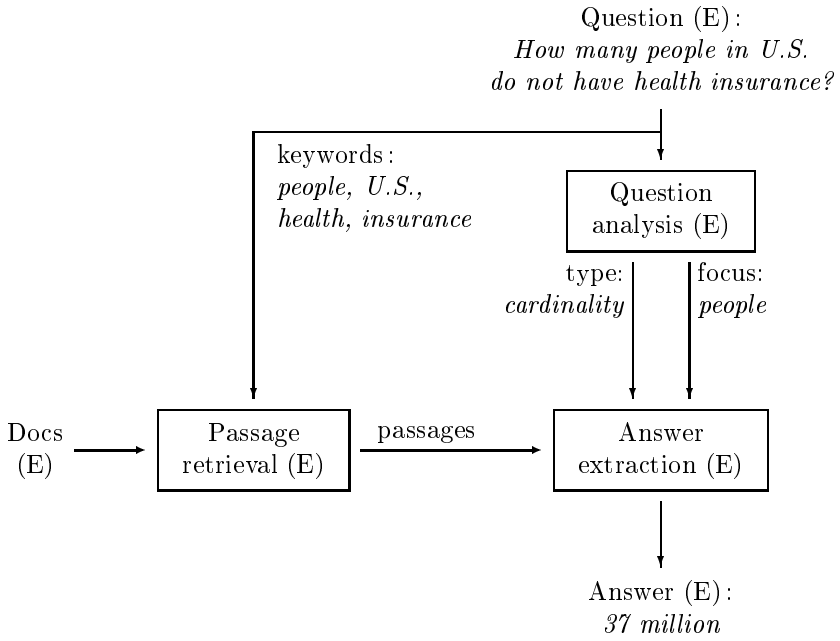

Figure 2: Architecture of the monolingual version of Quantum. The question, the documents and the answer are all in English (E).

2.2 Retrieval of relevant passages

The answer extraction mechanisms are too complex to be performed on the entire document collection. For this reason, we employ a classical search engine to retrieve only the most relevant passages before we proceed with answer extraction. We use Okapi [Robertson and Walker, 1998] because it allows for the retrieval of paragraphs instead of complete documents. We query it with the whole question and we let it stem the words and discard the stopwords. As a result, the query of the sample question in Fig. 2 would be a best match of *people, U.S., health* and *insurance*. We keep the 20 most relevant paragraphs along with their retrieval score as computed by Okapi.

2.3 Answer extraction

The extraction function selected during question analysis (optionally parameterized with the focus) is applied on the most relevant paragraphs. Three techniques or tools are used, depending on the extraction function: regular expressions, WordNet (for hypernyms/hyponyms relations) and the Annie named entity extractor from the GATE suite [Cunningham *et al.*, 2002]. For example, we would use WordNet to verify that *37 million Americans* can be an answer to the sample question in Fig. 2 because *Americans* is an hyponym of the question's focus *people* (or its singular form *person*).

Each noun phrase in the relevant paragraphs is assigned an extraction score when it satisfies the extraction function criteria. This extraction score is combined with the retrieval score of the source paragraph to take into account the density of the question keywords surrounding the extracted noun phrase. The three best-scoring noun phrases are retained as answer candidates. We decided to consider noun phrases as base units for answers because we found that only 2% of the questions from the past TREC campaigns could not be answered with a single noun phrase.

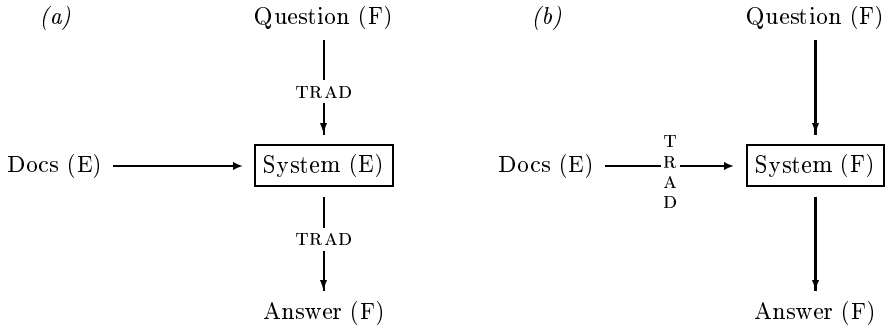

Figure 3: Two approaches for the transformation of an English (E) monolingual system into a cross-language system for French (F) questions. In *(a)*, the system's core remains unmodified and better English linguistic resources can be used. In *(b)*, the core is transposed to French, new resources in French need to be found and whole documents need to be translated.

3 Making the system cross-lingual

For Quantum as well as many other QA systems, the answer extraction phase is the most complex. Therefore, it was the impact on this phase which was decisive in selecting among strategies to transform our monolingual system into a cross-language bilingual system. Two factors were preponderant: the availability of linguistic resources for answer extraction and the amount of work needed to transform the system.

Both factors argued in favour of an unmodified English answer extraction module (Fig. 3*a*) and the addition of a translation module for the question and the documents, instead of the creation of a new answer extraction module in French. On one hand, the quality and availability of linguistic resources is usually better for English than French. In fact, many good quality English resources are free, as it is the case for WordNet and the named entity extractor Annie used by Quantum. Furthermore, by retaining the answer extraction module in its original language, fewer modifications are required to transform the monolingual system. Indeed, in order to write a new answer extraction module in the same language as the questions (Fig. 3*b*), we would have to find linguistic resources for that language, adapt the system to the new resources' interfaces and then translate whole documents prior to extracting the answers, which is currently a time-consuming and error-prone process. On the other hand, it is more efficient to translate only the question and the extracted answer. We will show that a full syntactically correct translation of the question is not mandatory and that the translation of the answer is facilitated by the particular context of QA.

In Fig. 3, we assume that the translation of the question and documents is perfect so that it is completely external to the blackbox system. Unfortunately, machine translation has not yet reached such a level of reliability. It is currently more efficient to *open* the system in order to make the translation steps easier. In our case, this allows us to avoid having to produce a complete and syntactically correct translation of the question. It also allows us to use different translation models depending on the task.

We first replaced the question analysis module by a new French version (Fig. 4) because the statistical techniques we use to translate the question are not reliable enough to produce syntactically correct sentences. Hence, our analysis patterns would seldom apply. Once the question is analyzed directly in French, the selected extraction function can be passed to the answer extraction module along with the question's focus, if any. However, the focus must be translated into English, seeing that we have retained the original English answer extraction module (among other things, the focus has to be known by WordNet). As for the passage retrieval module, we still

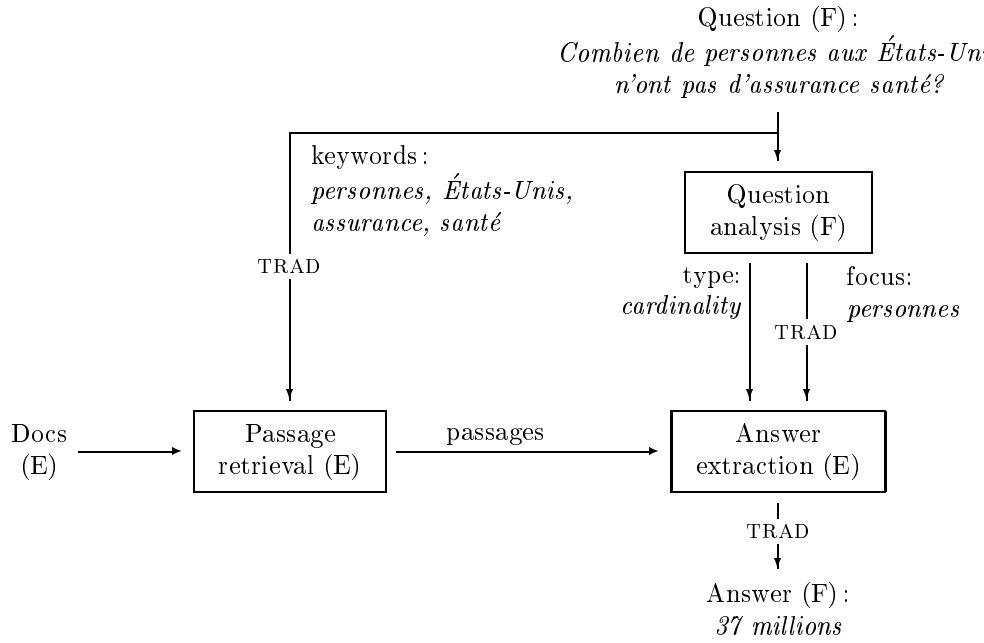

Figure 4: Architecture of the bilingual version of Quantum (to be compared with the monolingual version in Fig. 2). The question and the answer are in French (F), while the documents are in English (E). The question analysis module operates in French and the other modules remain in English. Translation is required at three points: for the keywords, the focus and the answer.

use Okapi on the English document collection, which therefore requires translating the question keywords from French to English. Finally, the answer extraction module does not require any modification. Let us now examine each of the modified modules in more detail.

3.1 Converting the question analysis module and translating the question's focus

We use regular expressions that combine words and POS tags to analyze a question. The original English module uses around 60 analysis patterns. We wrote about the same number of patterns for French.

We found that French questions were more difficult to analyze because of the greater flexibility in the formulation of questions. For example, *How much does one ton of cement cost* can be formulated in two ways in French: *Combien coûte une tonne de ciment* or *Combien une tonne de ciment coûte-t-elle*. In addition, English question words — the base of the analysis — do not always map to a single equivalent in French: this is the case of *what*, which can be mapped to *qu'est-ce que* in *What is leukemia / Qu'est-ce que la leucémie*, to *que* in *What does "kain ayinhore" mean / Que signifie "kain ayinhore"*, to *quoi* in *Italy is the largest producer of what / L'Italie est le plus grand producteur de quoi* and to *quel* in *What party did Occhetto lead / Quel parti Occhetto dirigeait-il*. Among other difficulties there are the masculine/feminine and singular/plural agreements, the addition of an euphonic *t* in the interrogative form of certain verbs in the 3rd person singular (in *Combien une tonne coûte-t-elle* but not in *Combien deux tonnes coûtent-elles*), elisions (*Qu'appelle-t-on*) and two forms of the past tense (*Quand le mur de Berlin a-t-il été construit / Quand le mur de Berlin fut-il construit*, while the only appropriate form in English is *When was the Berlin Wall built*).

At the same time that the analysis rules select an extraction function, they also identify the

question's focus. The focus semantics sometimes has an impact on the expected answer type. For instance, in *What is the longest river in Norway*, the focus *river* indicates that the answer is the name of a location. We use WordNet to make such links. This means that the focus from the French question has to be translated into English before the expected answer type is definitely known. To do so, we use an IBM2 statistical translation model trained on a set of documents composed of debates of the Canadian Parliament, news releases from *Europa - The European Union On-Line* and a sample of TREC questions. The IBM2 model is the simplest of the IBM series that takes into account the word's position in the source sentence. We need this feature because we want a translation that is the most probable given a particular word of the source sentence and, to a lesser degree, given all the other words of the source sentence. We keep only the best translation that is a noun.

We conducted an experiment on a sample set of TREC questions to measure the variation of performance between the original English question analysis module and the new French module [Plamondon and Foster, 2003]. Tested on a set of 789 questions from TREC, the regular expressions (used in conjunction with the semantic network of WordNet) of the English module select the correct extraction function for 96% of the questions. These questions were manually translated into French[1] and we found that the new French module selects the correct extraction function for 77% of them. This drop is due to two factors: the narrower coverage of the regular expressions and the incorrect translation of the focus (the focus is correctly translated half of the times). Most of these translation errors are due to the absence of the word in the training corpus, because many questions contain rare words, especially in definitions: *What is thalassemia, amoxicillin, a shaman,* etc. The translation of the focus is crucial to the question answering process. For example, it is almost impossible to determine that *37 million Americans* can properly answer the sample question in Fig. 4 if the focus *people* is wrongly translated into *flower*.

3.2 Translating the keywords for passage retrieval

Cross-language information retrieval has been widely addressed outside the QA domain. State-of-the-art retrieval engines combine the translation model with the retrieval model [Kraaij *et al.*, 2003]. However, since the search engine we use does not allow modifications to its retrieval model, we chose a simpler approach: we use an IBM1 translation model to get the best translations given the question and then we proceed as usual with Okapi. The selected target words are unordered and we retain only the nouns and verbs. Every word of the source sentence contributes equally to the selection of the best translations because the IBM1 model does not take the position of words into account, as the IBM2 model does. Hence, our method is slightly different from translating question keywords one by one. Our experiments showed that the best results are obtained when the query has as many non-stopwords as there are in the original question (5 on average for the CLEF questions).

We tested the cross-language passage retrieval module on the same TREC test set as for the question analysis module. We obtained an average precision of 0.570 with the original English module and an average precision of 0.467 with the cross-language module. Unlike in the question analysis module, a translation error does not compromise the location of the answer, as long as the query includes other keywords.

3.3 Translating the answer

Even though it was not required at CLEF to translate the extracted answer back into the same language as the question, our pilot project included this step in order to make the QA process transparent to a French speaker. However, due to a lack of time, we were unable to complete the answer translation module. Nevertheless, we believe that the particular context of QA should make things easier than in typical machine translation. For one thing, a lot of answers are named entities that do not require a translation. On a random set of 200 questions from TREC, 25% have

[1] A French/English set of almost 2000 TREC questions is freely available on our website at `http://www-rali.iro.umontreal.ca/LUB/qabilingue.en.html`

Run	Strict evaluation (MRR)	Lenient evaluation (MRR)	Inexact answers	Unsupported answers
50-byte	0.213	0.221	0	2
exact	0.140	0.161	11	5

Table 1: Statistics on the runs produced by Quantum on the French-to-English QA task at CLEF. *MRR* stands for *Mean Reciprocal Rank*. Strict evaluation considers only the *right* answers while lenient evaluation also considers *inexact* (too long or too short) answers and answers that are *unsupported* by the source document. *Inexact* and *unsupported* answers are not the total of inexact and unsupported answers in the whole run but the number of questions missed because the correct answer was inexact or unsupported.

an answer that is a person or location name which is identical in both languages, or a number, a date, a company name or a title that does not require a translation. To translate other types of answers, it would be worth exploring the use of the question to help disambiguation.

3.4 Performance of the complete system

We submitted two runs at CLEF: one with 50-byte answers and one with exact answers. The underlying QA process is the same in both, apart from additional checks performed on the 50-byte snippets to avoid submitting an answer a second time if it is already encompassed in a better ranking string. Statistics on the runs submitted at CLEF are listed in Table 1. As expected, the 50-byte run performed better than the exact run, but the gap is wider than we anticipated: we estimated that only 2% of TREC questions could not be answered suitably by a single noun phrase but it appears that this number is higher for CLEF, given the number of inexact answers we obtained. As for the number of unsupported answers, they remain a lesser concern.

We wanted to compare the cross-language version of Quantum with the monolingual English version. We ran the monolingual version on the English CLEF questions and we obtained a MRR of 0.223 (exact answers, lenient evaluation). At CLEF, the cross-language version of our system obtained a MRR of 0.161 on the French questions. As mentioned above, the principal reasons for this 28% performance drop include the different French question analysis patterns, the focus translation and the keyword translation.

We also measured a drop of 44% after a similar experiment conducted on TREC data [Plamondon and Foster, 2003]. We believe that CLEF questions were easier to process because they included no definition questions, thus there were less focus words to translate. We have also tried to translate our TREC question set with Babelfish[2] and then to use the original English system, but with this approach, performance dropped even more (53%).

4 Conclusion

We have shown how it is possible to transform an English QA system into a cross-language system that can answer questions asked in a different language than the document collection. In theory, it is possible to translate only the system's input/output (with Babelfish, for example) and to make no modification to the English system itself. In practice, as long as machine translation will not produce perfect translations, it is more efficient to decompose the task and to plug in translation at different points in the system. For our QA system Quantum, we use an IBM1 translation model to get English keywords from the French question for passage retrieval. We then use a new set of French question analysis patterns to analyze the question, because the English patterns would hardly match a badly structured question translated automatically. The question's focus is the only part that needs to be translated. We use an IBM2 translation model for that purpose.

[2] `http://world.altavista.com`

Overall, on the CLEF questions, the performance of our cross-language system is 28% lower than the monolingual system.

We hope the cross-language QA systems that entered the CLEF campaign will give French, Dutch, German, Italian and Spanish speakers access to a greater amount of information sources. For French speakers in particular, we have measured that it is better to use a cross-language system (even one in a development stage) than to limit oneself to a monolingual French QA system on French documents and therefore to be confined to one tenth the amount of information available to English speakers.

Acknowledgements

This project was financially supported by the Bell University Laboratories, the Natural Science and Engineering Council of Canada and the National Research Council of Canada. We would also like to thank Elliott Macklovitch and Guy Lapalme from the RALI, and Joel Martin from the NRC, for their help in conducting this research.

References

[Clarke *et al.*, 2002] C. L. A. Clarke, G. V. Cormack, M. Laszlo, T. R. Lynam, and E. L. Terra. The Impact of Corpus Size on Question Answering Performance. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research Information Retrieval (SIGIR 02)*, Tampere, Finland, 2002.

[Cunningham *et al.*, 2002] H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan. GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL 2002)*, Philadelphia, Pennsylvania, July 2002.

[Graesser *et al.*, 1992] Arthur Graesser, Natalie Person, and John Huber. *Mechanisms that Generate Questions*. Lawrence Erlbaum Associates, Hillsdale, New Jersey, 1992.

[Kraaij *et al.*, 2003] W. Kraaij, J.-Y. Nie, and M. Simard. Embedding Web-based Statistical Translation Models in CLIR. *Computational Linguistics*, 29(2), 2003. To appear.

[Plamondon and Foster, 2003] L. Plamondon and G. Foster. Multilinguisme et question-réponse: adaptation d'un système monolingue. In *Actes de la 10e conférence sur le traitement automatique des langues naturelles (TALN 2003)*, volume 2, Batz-sur-Mer, France, 2003.

[Plamondon *et al.*, 2002] L. Plamondon, G. Lapalme, and L. Kosseim. The QUANTUM Question Answering System at TREC-11. In *Notebook Proceedings of the Eleventh Text Retrieval Conference (TREC-11)*, Gaithersburg, Maryland, 2002.

[Robertson and Walker, 1998] S.E. Robertson and S. Walker. Okapi/Keenbow at TREC-8. In *Proceedings of the Eighth Text Retrieval Conference (TREC-8)*, Gaithersburg, Maryland, 1998.

[Voorhees, 2002] Ellen M. Voorhees. Overview of the TREC 2002 Question Answering Track. In *Notebook Proceedings of the Eleventh Text Retrieval Conference (TREC-11)*, Gaithersburg, Maryland, 2002.