How to answer in English to questions asked in French : by exploiting results from several sources of information

Guillaume Bourdil, Faza Elkateb, Brigitte Grau, Gabriel Illouz,
Laura Monceaux, Isabelle Robba and Anne Vilnat
LIR group, LIMSI-CNRS, BP 133 91403 Orsay Cedex
firstName.name@limsi.fr

Abstract

To build our bilingual system, we start from the monolingual QALC system, which has participated to preceding TREC [1] evaluations. Our best results were obtained when we chose to take the advice of several searches. First, we search for answers in a reliable document collection, and second on the Web. We try to maintain this strategy, and develop two runs. The first one is a new version of our system including multilingual aspects. Another approach of the multilinguism problem is to translate the question and then apply our monolingual system, including a search of the Web. Our second run results of a combination of this last results and those from the first run. The results we obtained confirm the fact that the best results are obtained by combination of different sources of information.

1 Introduction

Open-domain Question-Answering (QA) is a growing area of research whose aim is to find precise answers to questions in natural language, unlike search engines that return documents. When those engines also return snippets, as Google[2], they aim at providing a justification of documents rather than just giving the potential answer. One challenge in this field consists in finding only one answer while being sufficiently confident in it. The second possibility, corresponding to the approach we developed in QALC, our monolingual question answering system, consists in estimating the reliability of an answer by scoring it according to the kind of knowledge or the kind of process used. We found that providing just an endogenous estimation was not sufficient. Thus, we decided to apply our system on another source of knowledge in order to confront the results provided by both sources. We chose then to favour common propositions over unique ones, even if these latter had a high score. As such reasoning better applies if the sources of knowledge are different enough, we chose the Web as second source. Moreover, the diversity of the Web and its redundancy both lead to find a lot of answers, as we can see it in [Magnini et al. 2002a], [Magnini et al. 2002b], [Clarke et al. 2001] and [Brill et al. 2001]. The problem is to adapt this strategy in multilinguistic context. What has been said just before, the interest of the Web is its redundancy, but this fact is only true in English, as proved by 1. Thus, the have to search the English Web, to obtain significant results. We decide then to try two strategies : one consists in analyzing the French question, and then translate the "interesting parts" of it to use then this translated terms to search the reference collection of documents. This what we will call further the multilingual module. The second one consists in translating the question in English thanks to the help of CEA where there

[1] TREC evaluations are campaigns organised by the NIST: http://trec.nist.gov
[2] http://www.google.com

Figure 1: The languages on the Web

is an access to a professional version of Systran [3]. Then, we can apply our existing monolingual system, including the Web search. The first strategy gives our first run, the second one is the combination of the multilingual, and the two monolingual results (obtained in the collection, or on the Web).

After introducing our study on the multilinguism, we will present the global architecture of our system, and then detail the multilingual system.

2 Multilinguism : different approaches

To be able to deal with multilinguism, several solutions are possible. The first one consists in translating the question. The advantage is that the monolingual system may then be applied without any modification. The major problem is that this automatic translation cannot solve disambiguation problems in open-domain question-answering systems. It is the solution we adopt as a basic line for the Clef 04 campaign. The second solution is to translate the complete collection of documents. In this case, the translation may be guided by the context, the system doesn't have to be changed. The collection needs n times its initial length for n languages! A more difficult point is posed by the impossibility to translate the Web! The last solution is to proceed to the analysis of the question in the source language (French for us), and to translate the information issued from the analysis. In this solution, we don't try to obtain a complete translation, but only the terms that are considered as important after the analysis are translated. It is the solution we adopt for our multilingual module. It is this module that we detail in the following paragraphs, after presenting an overview of the QALC system.

3 Overview of "multilingual" QALC

The global architecture of the QALC system is illustrated Figure 2. First, its question analysis module aims at deducing characteristics helping to find possible answers in selected passages and to reformulate questions in a declarative form that is given to the search engine (Google). These characteristics are the question focus, the main verb and syntactic relations for modifiers. It is on these elements that we focus our translation efforts, as explained in the next section. For CLEF 04 campaign, we develop two versions of this module : one for question in French, and the other

[3] Special thanks to Olivier Ferret from CEA (French Center for Atomic Energy) who provides us these translations.

for the translated questions in Englis. The analysis is based on the results of the French version of XIP, the robust syntactic parser of [?]. For the analysis of the translated question, we use IFSP, the preceding robust syntactic parser of [?]. Queries are not the same for the Web search and for AQUAINT search (AQUAINT is the reference TREC collection). In the latter case, we use MG[4] for retrieving passages from a query made with AND and OR operators. For querying the Web, we chose to send a nearly exact formulation of the answer assuming that the Web redundancy will always provide documents.

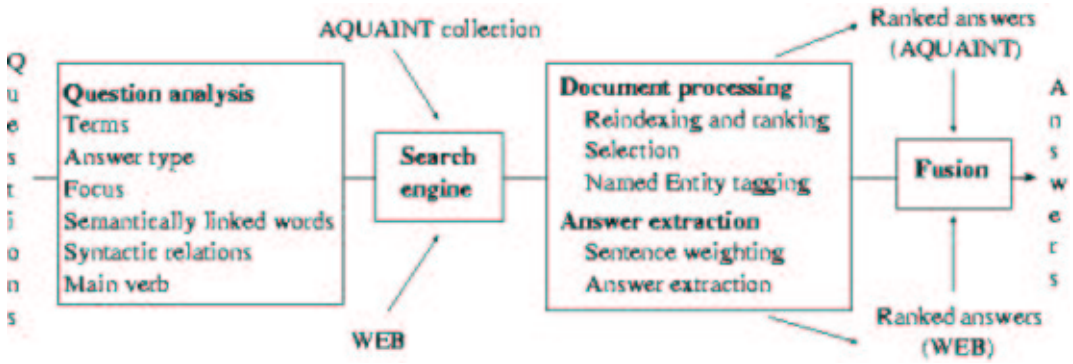

Figure 2: The QALC system

Retrieved documents, 1500 passages with MG (on AQUAINT) and 20 documents from the Web, are then processed. They are re-indexed by the question terms and their variants, reordered according to the kind of terms found in them, so as to select a subset of them in the case of MG, the Web documents being all kept. Named entity recognition processes are then applied. The answer extraction process relies on a weighting scheme of the sentences, followed by the answer extraction itself. We apply different processes according to the kind of expected answer, each of them leading to propose answers with a weight. The final step (for our second run) consists in comparing the results issued from AQUAINT, from the Web (for the translated questions) and the results issued from the multilingual system, and computing a final score. Its principle was to boost an answer if all the chains ranked it in the top 5 propositions, even with relatively low scores.

4 Translation of the important terms

Different methods are possible to proceed to the translation we need. The best results may be obtained by a deep translation, based on semantic grammars, ontologies... but the required tools do not exist in open-domain. Among the other translation possibilities, we consider the easiest one, which consists in using a bilingual dictionary to translate the terms from the source language to the targetted one. This simple method presents two problems : it is impossible to directly disambiguate the various meanings of the words to translate, and the two languages must be of equivalent lexical richness. This last constraint is verified for the 2-uple English/French, we will consider this method. To give an idea of the ambiguities we may encounter in a QA context, we study the corpus of the 1893 TREC questions in English. After analysis, we keep 9000 of the 15624 words used in this corpus. The mean of the number of meanings is 7.35 . The extrema are 1 (example : *neurological*) and 59 (example : *break*). Around the mean value, we find words such as *prize, blood, organization.* Thus, we cannot consider a dictionary giving only one meaning for a word, but we need to define a measure of the value of a translation in our QA context.

With these parameters, we study the different dictionaries we may use : the online dictionaries

[4]MG for Managing Gigabytes http://www.cs.mu.oz.au/mg/

(such as Reverso [5], Systran [6], Google [7], Dictionnaire Terminologique [8] or FreeTranslation [9]), and the dictionaries under GPL licences (such as Magic-Dic [10] or Unidic. The online dictionaries are generally complete. But they resolve the ambiguity : they only give one translation for a word, and thus do not respect our specification. Another limitation is the fact that we cannot modify these dictionaries, and that we have to deal with some technical constraints such as the limited number of requests we may adress and the access delays. Concerning the GPL dictionaries, they are obviously less complete, but they can be modfied, they are very fast and most of all, they give several translations for a request, as bilingual dictionaries. Among the GPL dictionaries, we choose Magic-dic, because of its evolutivity and the presence of terms in the dictionary, those terms being added by any user, but verified before being integrated, which is not the case for Unidic. For example the query for the French word *porte* gives the following results (we only give an extract) : porte bagages -¿ luggagerack, luggage rack porte cigarette -¿ cigarette holder porte clefs -¿ key-ring porte plume -¿ fountain pen porte parole, locuteur -¿ spokesman porte -¿ door, gate

To prevent its uncompleteness, and because it has been proved that the use of several dictionaries gives best results than a unique one, it has been enriched with the Google dictionary.

5 Example of the multilingual module

We will illutrate this multilingual module on the following example :
"Quel est le nom de la principale compagnie aérienne allemande?"
The first module is the parsing of the French question. One of the result of this step, is the list of the uni-terms and all the bi-terms (such as *adjective/common noun*) which were in the question, and the elimination of the stop words. The biterms are useful, because they allow a desambiguisation by giving a (small) context to a word. In our example, the biterms are : *principal compagnie, compagnie aérien, aérien allemand*; and the uniterms : *nom, principal, compagnie, aérien, allemand.*

With the Magic-dic dictionnary, we try to translate the biterms (when they exist), and the uniterms. All the proposed translations are taken into account. All the terms are grammatically tagged. If a bi-term cannot be directly translated, it is recomposed from the uniterms, following the English syntax. For our example, we obtained for the biterms : *principal compagny/main compagny, air compagny, air german*; and for the uniterms :*name/appellation, principal/main, compagny, german.*

These terms are then used by the following modules of multilingual QALC, instead of the original words. The translation doesn't try to solve the ambiguity between the translations : the selection will be made during the search of the documents. If the different terms are synonyms, pertinent documents will then be retrieved with this syninym, thanks to a larger search. If the word is incoherent within the context, its influence will not be sufficient to bring noise.

A first evaluation of this translation multilingual module had been made on the TREC questions, translated in French by RALI.

6 Query formulation

Due to the Web redundancy, we state that it is possible to find documents even with a very specific query, and that a precise query will lead to find relevant documents, i.e. documents containing the searched answer, among the first ones. Thus, our choice was to reformulate the question in an affirmative form with as few variations as possible. For instance, for the question *When was*

[5] http://translation2.paralink.com
[6] http://babel.altavista/translate.dyn
[7] http://www.google.com/language_tools
[8] http://granddictionnaire.com
[9] http://www.freetranslation.com
[10] http://magic-dic.homeunix.net/

Wendy's founded?, we expect to find a document containing the answer in the form: *Wendy's was founded on ....* We first query the Web for strings with exact match in documents as in [Brill et al. 2001], and not for the different words of the query even linked with AND, OR or NEAR operators as in [Magnini et al. 2002a] or [Hermjacob et al. 2002].

7 Fusion of several sources of information

As it was said in section 3 (overview of QALC), we elaborated a strategy based on the comparison of the results of our system from two different sources of knowledge: CLEF collection and Web for translated questions, and multilingual results. The knowledge source explored by the Web search is obviously really much larger than CLEF collection search. Using such source brings our system a relevant way to confirm some of its answers and to reinforce its confidence score. However, among the answers provided by the Web search, some are not corresponding to any document of CLEF. So, it is to be noted that the Web answers also have to be located in CLEF collection.

These three applications of QALC supply for each question a set of answers which are ordered according to the score they received during the answer extraction process. Hence, the role of the final selection is to choose a unique answer between these three ordered sets.

Before describing the algorithms we wrote for the final selection, we will describe the way QALC attributes a confidence score to each potential answer.

7.1 Answer weighting

All the sentences provided by the document processing are examined in order to give them a weight that reflects both the possibility that the sentence contains the answer, and the possibility that the QALC system locates the answer within the sentence. The criteria that we used are closely linked with the basic information extracted from the question. The resulting sentence ranking should not miss obvious answers. Our aim is that the subsequent modules of answer extraction and final answer selection are able to raise a lower weighted answer to an upper rank thanks to added specific criteria. The criteria that we retained use the following features within the candidate sentences:

- question lemmas, weighted by their specificity degree[11],
- variants of question lemmas,
- exact words of the question,
- mutual closeness of the question words,
- presence of the expected named entity type.

First we compute a basic weight of the sentence based on the presence of question lemmas or variants of these lemmas (the two first criteria). The basic weight is relative. We subsequently add an additional weight to this basic weight for each additional criteria that is satisfied. Each additional criteria weight cannot be higher than about 10% of the basic weight.

During answer extraction this weight is still refined. If the expected answer type is a named entity, then selected answers are the words of the sentence that correspond to the expected type. In order to extract the answer, the system first computes additional weights taking into account:

- the precise or generic named entity type of the answer,
- the location of the potential answer with regard to the question words within the sentence,
- the redundancy of an answer in the top ten sentences.

[11] The specificity degree of a lemma depends on the inverse of its relative frequency computed on a large corpus.

When the expected answer type is not a named entity, we use extraction patterns. Each candidate sentence provided by the sentence selection module is analysed using the extraction pattern associated with the question type that has been determined by the question analysis. Extraction patterns are composed of a set of constraint rules on the candidate sentences. Rules are made up of syntactic patterns that are used to locate potential answers within the sentence, and of semantic relations that are used to validate answers. Potential answers are weighted according to the satisfied constraints. More detail can be found in [de Chalendar et al. 2002].

Finally after the extraction and weighting procedure, the five best weighted answers are retained for the final selection module.

7.2 Final selection algorithms

The underlying idea is to compare results obtained from diverse sources of knowledge. Our comparison allows us to reinforce the score of answers belonging to the different result sets, thus allowing a significant number of right answers to get at the first rank. Let us give an example with only two sets of results. Table 1 contains an example of these sets corresponding to the question: *Who defeated the Spanish armada?*

Table 1: Answer set example

Collection answers	Web answers	Final score
0) Queen Elizabeth (score 1205)	0) Elizabeth I (score 1299)	
1) England (score 1202)	1) Elizabeth I (score 1297)	
2) **Francis Drake** (score 982)	2) Philip II (score 1282)	
3) Spain (score 872)	3) **Francis Drake (score 1252)**	**1852**

The algorithm examines each couple ($answer_i$, $answer_j$), i being the position of an answer found in the collection, j being the position of a Web answer, its score being the best of both scores. When both answers of the couple are exactly equal, the algorithm attributes a bonus to the couple score, which is calculated according to both positions: i and j: $(11 - (i + j)) * 100$. The answer that is finally returned belongs to the couple obtaining the best score. The additive bonus was chosen in order to push the confirmed answers before the unconfirmed ones.

Looking at Table 1, we see that the answer at rank two in the collection and the answer at rank three in Web answers are the same. The received bonus is 600, and the answer *Francis Drake* is returned with its final score: 1852.

Indeed, the comparison has been carried out between three strategies rather than two; it could be extended to more than only five answers and it could be more flexible by taking into account answers included one in the other, instead of exactly equal answers.

8 Results

Table 2 presents the results we obtained for the two runs.

Table 2: Results of the runs

	Multilingual	Fusion
Number of right answers on factoid questions (180)	18	36
Number of right answers on definition questions (20)	4	3
Total of right answers (200)	22	39
Confidence weighted score	0.033	0.075

We note that the score of the combined run is increased of 100%, compared to the first one. We proceed to an evaluation of the translation in the multilingual system. We find that :

- 46,50% of the tranlated terms were correct
- 12,63% of the tranlated terms were correct, but may be enhanced (i.e. another translation would have been better)
- 8,33% of the tranlated terms were correct, but identical to the terms in the source language
- 30,10% of the tranlated terms were incorrect
- 2,44% of the source terms were incorrect

It is obvious that the dictionary was not complete enough for this campaign. We will obtain a greater cover by adding manually some obvious errors (no tranlation of the French verb *jouer* in its meaning *to play*, for example). We also begin to add translations by requests to the Google translation module.

Another evaluation concerns the biterms, taht we have presented as very important to disambiguate ambiguous uniterms. To this end, we have determined the documentary frequency of each translation of the different biterms in the CLEF corpus. If the frequency is high, then the biterm may be an adequate translation. From this study, only 43.5% of the biterms are correct. Thus, we need to enhance this translation too. Some ideas to do so are the validation of the translations, by scoring them following their frequency both in a bilingual corpus, and in a corpus in target language.

We also notice that an important work has to be done on proper nouns, specially geographic names and organization names. We then need to develop bilingual lists for the most frequent nouns.

We may conclude on these results that ourmultilingual module needs to be enhanced, but that the solution we adopt allows us to easily proceed to several of these enhancements.

References

[Brill et al. 2001] E. Brill, J. Lin, M. Banko, S. Dumais and A. Ng, 2001. Data-Intensive Question Answering. *TREC 10 Notebook, Gaithersburg, USA*

[de Chalendar et al. 2002] G. de Chalendar, T. Dalmas, F. Elkateb-Gara, O. Ferret, B. Grau, M. Hurault-Plantet, G. Illouz, L. Monceaux, I. Robba, A. Vilnat, 2002, The Question Answering System QALC at LIMSI, Experiments in Using Web and WordNet. *Trec 11, Notebook, Gaithersburg, USA* pp. 457-467

[Chu-Carroll et al. 2002] J. Chu-Carroll, John Prager, Christopher Welty, Krzysztof Czuba and David Ferruci. 2002. A Multi-Strategy and multi-source Approach to Question Answering. *TREC 11 Notebook, Gaithersburg, USA* pp. 124-133

[Clarke et al. 2001] C. L. Clarke, G. V. Cormack, T. R. Lynam, C. M. Li and G. L. McLearn, 2001, Web Reinforced Question Answering (MultiText Experiments for Trec 2001), *TREC 10 Notebook, Gaithersburg, USA*

[Fellbaum 1998] C. Fellbaum. 1998. WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press

[Hermjacob et al. 2002] U. Hermjakob, A. Echihabi and D. Marcu. 2002, Natural Language Based Reformulation Resource and Web Exploitation for Question Answering, *TREC 11 Notebook, Gaithersburg, USA*

[Magnini et al. 2002a] B. Magnini, M. Negri, R. Prevete and H. Tanev. 2002a. Is It the Right Answer? Exploiting Web redundancy for Answer Validation, *Proceedings of the 40 th ACL* pp. 425-432

[Magnini et al. 2002b] B. Magnini, M. Negri, R. Prevete and H. Tanev, 2002b, Mining Knowledge from Repeated Co-occurrences: DIOGENE at TREC-2002, *TREC 11 Notebook, Gaithersburg, USA*

[Moldovan et al. 2002] D. Moldovan, S. Harabagiu, R. Girju, P. Morarescu, F. Lacatusu, A. Novischi, A. Badalescu and O. Bolohan, 2002, LCC Tools for Question Answering, *TREC 11 Notebook, Gaithersburg, USA*