Answering French Questions in English by Exploiting Results from Several Sources of Information

Brigitte Grau[1], Gabriel Illouz[1], Laura Monceaux[3],
Isabelle Robba[1], Anne Vilnat[1]
Guillaume Bourdil[1], Faïza Elkateb-Gara[1], Olivier Ferret[2] and Benoît Mathieu[2]

[1] LIR group, LIMSI-CNRS, BP 133, 91403 Orsay Cedex
firstName.name@limsi.fr
[2] LIC2M, CEA-LIST, BP 6, 92265 Fontenay-aux-Roses
firstName.name@cea.fr
[3] LINA, 2 rue de la Houssinière, BP 92208, 44332 Nantes Cedex 3
firstName.name@lina.univ-nantes.fr

Abstract. Our bilingual QA system MUSCLEF, is based on QALC, the monolingual system with which we have participated in the previous TREC[1], where our best results were obtained when we combined the results of several searches. First, QALC searched a reliable document collection for answers, and second the WEB. We kept this strategy for CLEF, returning two runs. In the first one, we modified QALC so as to handle multilinguality by translating the terms identified in the question. In the second run, we combined the results of the first run with those obtained by first translating the question, then applying the full QALC strategy i.e. searching both the collection and the WEB. The final evaluation confirms the fact that the best results are obtained by combining 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 whole documents. When these engines also return snippets, as Google[2], they aim at providing justifications of documents rather than just giving an answer. One challenge in this field consists in finding only one answer in which we are sufficiently confident. The approach we developed in QALC, our monolingual English 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 for its elaboration. However, we found that providing just an endogenous estimation with respect to the collection was not sufficient. Thus,

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

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 propositions common to both sources over unique ones, even if the latter had a high score. Because such reasoning applies better if the sources of knowledge are different enough, we chose the Web as second source. Moreover, the diversity and redundancy of the Web lead to find a lot of answers, as we can see in [9], [10], [5] and [3].

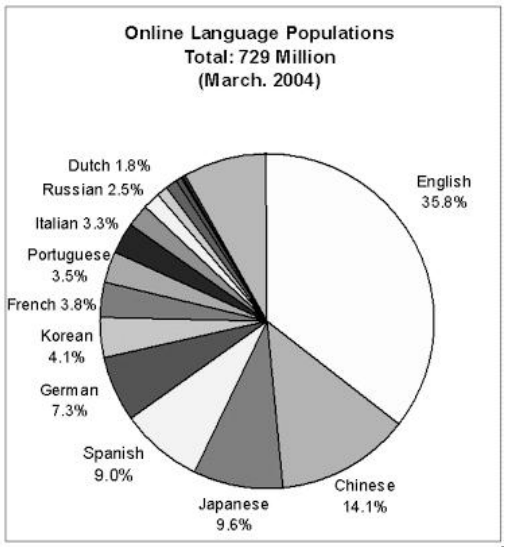

Fig. 1. The languages on the Web

In CLEF evaluation, the problem is to adapt this strategy in a multilingual context. Moreover, we must take into account the fact that the interest in using the Web (i.e. its redundancy), is only effective in English, as proved[3] by Figure 1. Thus, searching the French Web would not give as significant results as searching the English Web. For CLEF evaluation, we developed MUSCLEF (Multilingual System for CLEF) which uses two strategies. The first one consists in analyzing the French question, translating "interesting parts", and then using these translated terms to search the reference collection. MUSQAT, which is our multilingual module, follows this strategy. The second strategy consists in translating the question in English using a professional version of Systran (which was possible thanks to CEA), then, applying QALC, our existing monolingual system, including the Web search. The first strategy corresponds to our first run,

[3] This figure is extracted from the Centre for Public Policy of the University of Melbourne: www.public-policy.unimelb.edu.au/egovernance/papers/33_Skidmore.pdf

while we did the second run with the combination of the multilingual and the two monolingual results (browsing both the collection and the Web).

After introducing our approach to multilingualism, we present the global architecture of MUSCLEF (MUSQAT and QALC), and then detail MUSQAT, the multilingual module.

2 Multilingualism: Different Approaches

In Question Answering, several solutions exist for dealing with multilingualism. The first one consists in using machine translation for the question. In this first case, the advantage for us is that our monolingual system may then be applied without any modification. The major problem is that automatic translation does not deal correctly with disambiguation problems in open-domain question-answering systems. Machine translation is the solution we adopted as a basic line for the CLEF 04 campaign. The second solution consists in translating the complete collection of documents. In this second case, the translation may be guided by the context, and the system does not have to be changed. A major drawback is that the collection size is n times its initial length for n languages! Another difficult problem is posed by the impossibility to translate the whole Web! The last solution is to proceed to the analysis of the question in the source language (French in our case), and then to translate only the information produced by the analysis. In this solution, we do not try to obtain a complete translation: only the terms that are considered important by the analysis are translated. It is the solution we adopted for our multilingual module, which we detail in the following paragraphs, after presenting an overview of the MUSCLEF system.

3 Overview of MUSCLEF

The global architecture of MUSCLEF is illustrated in Figure 2. First, its question analysis module aims at deducing characteristics which may help to find possible answers in selected passages and to reformulate questions in a declarative form dedicated to the Web search engine (Google). These characteristics are the question focus, the main verb and syntactic relations for modifiers. We focused our translation efforts on these elements, as explained in the next section. For CLEF 04 campaign, we developed a new version of this module for questions in French. The analysis is based on the results of the French version of XIP, the robust syntactic parser of Xerox [2]. For the analysis of the translated question, we use IFSP, another robust syntactic parser of Xerox[1]. We made an evaluation of the French question analysis module for questions whose answer type is a named entity. For these 119 questions, recall is 95% and precision is 97%.

Queries are not the same for the Web search and for the CLEF collection search. In the last case, we use MG[4] for retrieving passages. For querying the

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

Web, we choose to send a nearly exact formulation of the answer assuming that the Web redundancy would always provide documents.

Retrieved documents are then processed. They are re-indexed by the question terms and their linguistic variants, reordered according to the number and the kind of terms found in them, so as to select a subset of them. 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 weighted answers. For our second run, the final step consists in comparing, for the translated questions, (a)the results issued from the collection, (b)the results issued from the Web and, for the translated terms, (c) 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.

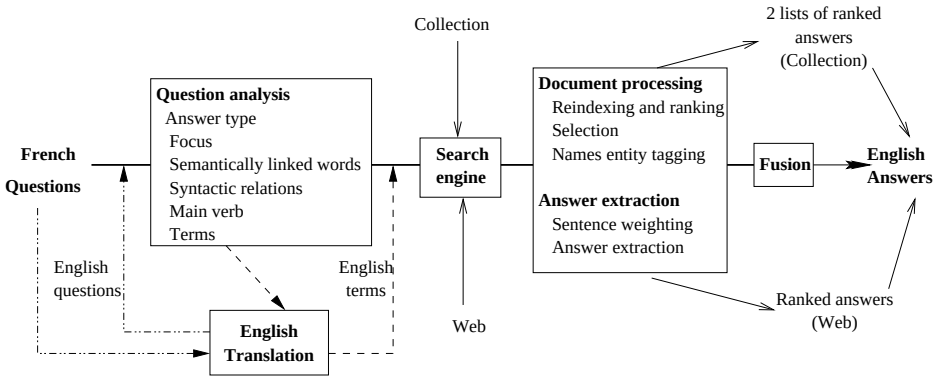

Fig. 2. MUSCLEF architecture

4 Question Processing

As mentioned in section 1, two solutions were tested for building a representation of questions that can be matched with documents. The first one makes use of an automatic translator for translating questions from French to English and then, performs the analysis of the translated questions. The second solution consists in analyzing questions in the source language, which is French in our case, and translating in the target language, which is English, those terms that are considered as the most important ones.

4.1 Automatic Translation of Questions

Thus, the first solution we tested for solving the language mismatch between questions and documents relies on the automatic translation of questions. In our case, this automatic translation was performed by the SYSTRANLinks online interface provided by Systran[5]. No additional dictionary was used. As most of the questions of the CLEF evaluation are not very complex from a syntactical point of view and address general subjects, their translations can often be considered as reliable, as illustrated by Figure 3.

0009 - Quand est apparu pour la première fois le virus Ebola ?
0009 - When did the Ebola virus appear for the first time?

0166 - Où se trouve Halifax ?
0166 - Where is Halifax?

Fig. 3. Examples of correct question translation

However, translation mistakes may also occur for simple questions, as we can see it in Figure 4. These mistakes may concern syntax: in question 175 for instance, "Quel est" should be translated as "Who is" and not as "Which is" and in question 165, the phrase "Qu'est-ce que", that is specific to questions, is only partially translated[6]. But these mistakes also may concern semantics: in question 175 again, "réalisateur" is translated as "realizer" while an answer is more likely to be found if it is translated as "director" or "film director". Finally, question 165 also illustrates the problem of the uncompleteness of dictionaries, which is impossible to circumvent fully in an open-domain system, especially for acronyms: "OMC" (Organisation Mondiale du Commerce) should be translated as "WTO" (World Trade Organization), just as "OTAN" is translated as "NATO" in question 143.

4.2 Term Translation

Different methods can be used to achieve term translation. Results may be obtained by a translation based on bilingual ontologies; but as mentioned in the previous section, the required tools do not really exist in open-domain. Among the other translation possibilities, we considered the easiest one, which consists in using a bilingual dictionary to translate the terms from the source language to

[5] We would like to thank Systran for the access they give to us to this service in the context of the ALMA project.

[6] By the way, this observation shows that as for part-of-speech taggers or syntactic analyzers, questions should be specifically taken into account by machine translation systems while they are generally not.

0175 - Quel est le réalisateur de “Nikita” ?
0175 - Which is the realizer of “Nikita”?
0165 - Qu'est-ce que l'OMC ?
0165 - What OMC?
0143 - En quelle année a été créée l'OTAN ?
0143 - In which year was creates NATO?

Fig. 4. Examples of mistakes in question translation

the target language. This simple method presents two drawbacks: it is impossible to directly disambiguate the various meanings of the words to be translated, and the two languages must be of equivalent lexical richness. Since this last constraint is verified for the couple English/French, we used this method. To give an idea of the ambiguities we may encounter in a QA context, we studied the corpus of 1893 questions in English of TREC. After analysis, we kept 9000 of the 15624 words used in this corpus. The average of the number of meanings was 7.35 in WordNet. The extrema were 1 (example: *neurological*) and 59 (example: *break*). Around the average value, we found common words such as *prize, blood, organization*. Hence, we could not consider a dictionary giving only one meaning for a word, moreover we needed to define a measure of the value of a translation in our QA context.

With these constraints, we studied the different dictionaries we could use: the online dictionaries (such as Reverso[7], Systran[8], Google[9], Dictionnaire Terminologique[10] or FreeTranslation [11]), and the dictionaries under GPL licences (such as Magic-Dic[12] or Unidic). The online dictionaries are generally complete. But they resolve the ambiguity and they only give one translation per word. Another limitation was the fact that we could not modify these dictionaries, and that we had to deal with some technical constraints such as the limited number of requests we may adress and the access time. Concerning the GPL dictionaries, they are obviously less complete, but they can be modified, they are very fast and for most of all, they give several translations for a request, as classical bilingual dictionaries. Among the GPL dictionaries, we chose Magic-dic, because of its evolutivity: terms can be added by any user, but they are 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 excerpt):

- porte bagages - luggagerack, luggage rack
- porte cigarette - cigarette holder
- porte clefs - key-ring

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

– 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 better results than a unique one, we intend to enrich it with the Google dictionary.

4.3 The Multilingual Module

We will illutrate the strategy defined in MUSQAT on the following example: *"Quel est le nom de la principale compagnie aérienne allemande?"*, which is translated in English *"What is the name of the main German airline company?"*.

The first step is the parsing of the French question that provides a list of the uni-terms and all the bi-terms (such as *adjective/common noun*) which were in the question, and eliminates the stop words. The biterms are useful, because they (indirectly) disambiguate by giving a (small) context to a word. In our example, the biterms (in their lemmatized form) are: *principal compagnie, compagnie aérien, aérien allemand*; and the uniterms: *nom, principal, compagnie, aérien, allemand.*

With the help of the Magic-dic dictionnary, we attempted to translate the biterms (when they exist), and the uniterms. All the proposed translations were taken into account. All the terms were grammatically tagged. If a bi-term could not be directly translated, it was 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.* When a word does not exist in the dictionnary, we keep it as it whithout any diacritic.

These terms plus their categories (given by the Tree Tagger) instead of the original words were then given as input to the other modules of MUSQAT, instead of the original words. The translation module did not try to solve the ambiguity between the different translations: the MG request is made from the union of all the translations and the disambiguisation takes place during document selection. If the different terms are synonyms, pertinent documents are then retrieved with these synonyms, thanks to a larger search. If the word is incoherent within the context, we suppose its influence is not sufficient to generate noise.

We made an evaluation of the translation given by MUSQUAT. The 200 questions in French contained 731 words, corresponding to 1091 English words, and 932 terms (uni-terms + bi-terms) corresponding to 1464 terms in English. Studying this translation, we observed that:

– 59% of the translated terms were correct, (but for 12,63% of terms the translation may be enhanced)
– 8% of the translated terms were correct, but identical to the terms in the source language

– 33% of the translated terms were incorrect

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

Another evaluation concerns the biterms, that we presented as very important to disambiguate ambiguous uniterms. To accomplish this goal, we determined the document 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. According to this study, 47.5% of the biterms were found in the corpus. An interesting approach could be to validate the translations, by scoring them following their frequency both in a bilingual corpus, and in a monolingual corpus (target language).

We also noticed that an important work had to be done on proper nouns, especially geographic names, organization names and acronyms. We then need to develop bilingual lists for the most frequent nouns.

5 Fusion of Several Sources of Information

As it was said in section 3 (overview of MUSCLEF), our second run is obtained by comparing three sets of results: the first is given by MUSQAT, the second by QALC searching on the Web and the third by QALC searching on the CLEF collection. The Web provides our system with a knowledge source obviously much larger than the CLEF collection. Using such source gives to 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 found in any CLEF document. So, it is to be noticed that Web answers must be present in CLEF collection.

Each of the three results sets contains for each question a set of answers which are ordered according to a confidence score. This score is updated all along the different steps of the answer extraction. Before describing the algorithm we wrote for the final selection, we will describe the way the confidence score is attributed to each candidate answer.

5.1 Answer Weighting

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

- question lemmas, weighted by their specificity degree[13],
- variants of question lemmas,
- exact words of the question (only in the "all english" version),
- 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 further 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. To order the answers, MUSCLEF 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 the answer in the top ten sentences.

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 regular expressions with the focus noun as pivot. More detail can be found in [7].

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

5.2 Final Selection Algorithm

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 correct answers to be assigned the first rank. Table 1 contains an example of these sets corresponding to the question: *"En quelle année Thomas Mann a-t-il obtenu le Prix Nobel ?"*, translated in English *"In what year did Thomas Mann win the Nobel Prize?"*.

The three sets of results are compared two by two using an algorithm written for TREC. This algorithm examines each couple ($answer_i$, $answer_j$), i and j being the answer positions in their own set. When both answers are equal or included one in the other, the algorithm attributes a bonus to the best score of the couple. This bonus is calculated according to both positions i and j: $(10 - (i + j)) * 100$. The additional bonus was chosen in order to place the

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

Table 1. Answer set example

QALC + Web		MUSQAT		QALC + Collection	
Answer	Score	Answer	Score	Answer	Score
0) **in 1929**	**1082**	0) *in 1976*	721	0) October 11 , 1994	878
1) 1875-1955	1005	1) *in 1976*	721	1) **in 1929**	**853**
2) 08th March 1879	903	2) *in 1929*	664	2) *in 1976*	798
3) in 1903	877	3) 2	640	3) October 12 , 1994	703
4) *in 1929*	849	4) 1964	561	4) in 1979	696

confirmed answers before the unconfirmed ones. Thus the algorithm builds a set of answer couples ordered according to their new score. Since in CLEF we had to compare three sets of results, we applied the algorithm on each couple of answer sets (three times), the answer finally returned belongs to the couple which obtains the best score.

Looking at Table 1, we see that two dates appear in the three sets: *in 1929* and *in 1976*. The couple which appears in bold font in Table 1 receives 900 as a bonus. So the answer *in 1929* obtains the best final score (1082 + 900) and is then returned.

This algorithm which compares answer sets two by two is thus easy to apply on more than two sets. Nevertheless, we observed that a comparison made directly between the three answer sets would give different results. Indeed making the comparisons two by two, we do not take into account in the same way the answers appearing in the three sets.

6 Results

Table 2. Comparative evaluation of the different strategies

		MUSQAT	QALC + Collection	QALC + Web	Fusion (Official results)
Sentences	5 first ranks	56	65	61	
NE answers	Rank 1	17	26	24	
	5 first ranks	33	37	43	
Non NE answers	Rank 1	7	3	0	
	5 first ranks	12	8	0	
Total	Rank 1	24 (12%)	29 (14.5%)	24(12%)	39 (19,5%)
	5 first ranks	44	45	43	

Table 2 presents a comparative evaluation between MUSQAT and QALC. The evaluation was made by an automatic process that looks for the answer

patterns in the system answers, applying regular expressions. These results were computed with 178 answer patterns that we built for the 200 questions of CLEF.

The first line indicates the number of correct answers found in the 5 first sentences given by MUSQAT (using term translation) and both applications of QALC (collection and Web search). The second line, "NE answers", gives the number of correct answers on questions that had a Named Entity as answer, the third line, "non NE answers", concerns the other questions. Results are presented when the system just gives one answer and when it gives 5 answers. The last column indicates the best official result of our system on the 200 questions. The official score of MUSQAT was 22 (11%), thus we can observe that merging answers obtained by different strategies enables a gain of 17 answers (77%).

At TREC 2002, the systems had to provide a unique answer for each question. On 500 questions, QALC alone found 128 right answers (25%). When making the fusion between Web answers and TREC collection answers, QALC found 148 answers (29,68%). These results have to be compared with the numbers on the last line (Table 2) for correct answers at rank 1: the multilingual problem entails a reduction of performance of 10%.

We can also notice that the three strategies are equivalent, and that a weak point of our system remains the extraction of answers from selected sentences for non Named Entity questions[14].

7 Conclusion

Even if its first results are encouraging, MUSCLEF, our first multilingual system, can yet be enhanced. However, its architecture, organized into several independent modules, was chosen to be able to easily make these enhancements. Moreover, we observed that both strategies that we adopted (term translation and question translation) were relevant and should be maintained together in further experiences.

Obviously, better multilingual resources will be necessary, but since complete resources are not available, it could be interesting to search the Web to control the obtained translations.

References

1. Aït-Mokhtar, S., Chanod, J.-P.: Incremental finite-state parsing. Proceedings of the 5th Conference on Applied Natural Language Processing (ANLP-97), Washington, DC, USA (1997)
2. Aït-Mokhtar, S., Chanod, J.-P., Roux, C.: Robustness beyond shallowness: incremental deep parsing. Natural Language Engineering Vol. 8 (2/3), (2002)121–144
3. Brill, E., Lin, J., Banko, M., Dumais, S., Ng, A.: Data-Intensive Question Answering. TREC 10 Notebook, Gaithersburg (2001)

[14] Even if we discovered, but too late, a bug on the Web run.

4. Chu-Carroll, J., Prager, J., Welty, C., Czuba, K., Ferruci, D.: A Multi-Strategy and multi-source Approach to Question Answering. TREC 11 Notebook, Gaithersburg, USA (2002)124–133
5. Clarke,C.L., Cormack, G.V., Lynam, T.R., Li, C.M., McLearn,G.L.: Web Reinforced Question Answering (MultiText Experiments for Trec 2001), TREC 10 Notebook, Gaithersburg, USA (2001)
6. Fellbaum, C.: WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press (1998)
7. Ferret, O., Grau, B., Hurault-Plantet, M., Illouz, G., Jacquemin, C., Monceaux, L., Robba, I., Vilnat, A.: How NLP Can Improve Question Answering Knowledge Organization, Vol. 29, N3-4 (2002)135–155
8. Hermjakob, U., Echihabi, A., Marcu, D.: Natural Language Based Reformulation Resource and Web Exploitation for Question Answering, TREC 11 Notebook, Gaithersburg, USA (2002)
9. Magnini, B., Negri, M., Prevete, R., Tanev, H.: Is It the Right Answer? Exploiting Web redundancy for Answer Validation, Proceedings of the 40 th ACL (2002)425–432
10. Magnini, B., Negri, M., Prevete, R., Tanev, H.: Mining Knowledge from Repeated Co-occurrences: DIOGENE at TREC-2002, TREC 11 Notebook, Gaithersburg, USA (2002)
11. Moldovan, D., Harabagiu, S., Girju, R., Morarescu, P., Lacatusu, F., Novischi, A., Badalescu, A., Bolohan, O.: LCC Tools for Question Answering, TREC 11 Notebook, Gaithersburg, USA (2002)