Evaluation and Improvement of Cross-Lingual Question Answering Strategies

Anne-Laure Ligozat and **Brigitte Grau** and **Isabelle Robba** and **Anne Vilnat**
LIMSI-CNRS
91403 Orsay Cedex, France
firstname.lastname@limsi.fr

Abstract

This article presents a bilingual question answering system, which is able to process questions and documents both in French and in English. Two cross-lingual strategies are described and evaluated. First, we study the contribution of biterms translation, and the influence of the completion of the translation dictionaries. Then, we propose a strategy for transferring the question analysis from one language to the other, and we study its influence on the performance of our system.

1 Introduction

When a question is asked in a certain language on the Web, it can be interesting to look for the answer to the question in documents written in other languages in order to increase the number of documents returned. The CLEF evaluation campaign for cross-language question answering systems addresses this issue by encouraging the development of such systems.

The objective of question answering systems is to return precise answers to natural-language questions, instead of the list of documents usually returned by a search engine. The opening to multilingualism of question answering systems raises issues both for the Information Retrieval and the Information Extraction points of view.

This article presents a cross-language question answering system able to treat questions and documents either in French or in English. Two different strategies for shifting language are evaluated, and several possibilities of evolution are presented.

2 Presentation of our question answering system

Our bilingual question answering system has participated in the CLEF 2005 evaluation campaign [1]. The CLEF QA task aims at evaluating different question answering systems on a given set of questions, and a given corpus of documents, the questions and the documents being either in the same language (except English) or in two differents languages. Last year, our system participated in the French to English task, for which the questions are in French and the documents to search in English.

This system is composed of several modules that are presented Figure 1. The first module analyses the questions, and tries to detect a few of their characteristics, that will enable us to find the answers in the documents. Then the collection is processed thanks to MG search engine [2]. The documents returned are reindexed according to the presence of the question terms, and more precisely to the number and type of these terms ; next, a module recognizes the named entities, and the sentences from the documents are weighted according to the information on the question. Finally, different processes are applied depending on the expected answer type, in order to extract answers from the sentences.

3 Cross-language strategies for question answering systems

Two main approaches are possible to deal with multilingualism in question answering systems :

[1] Multilingual Question Answering task at the Cross Language Evaluation Forum, http ://clef-qa.itc.it/

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

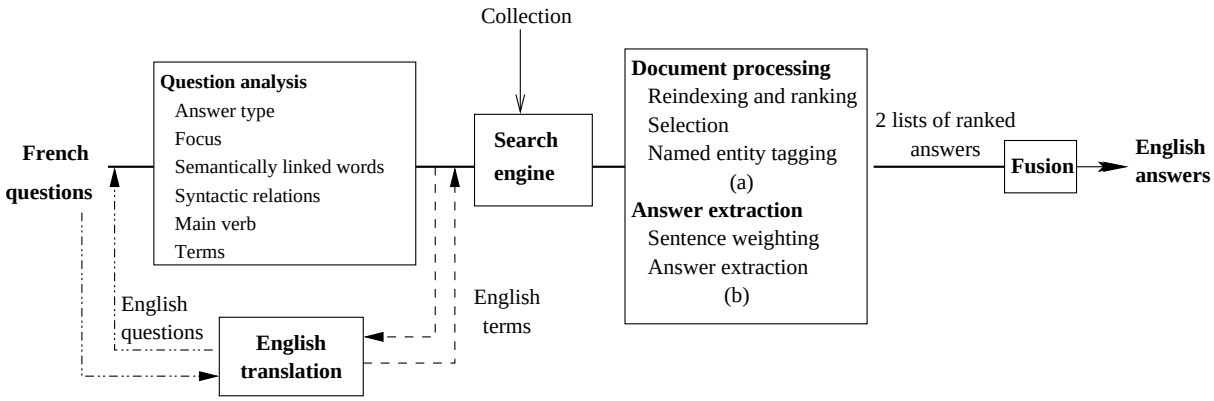

FIG. 1 – Architecture of our cross-language question answering system

question translation and term-by-term translation. These approaches have been implemented and evaluated by many systems in the CLEF evaluations, which gives a wide state-of-the-art of this domain and of the possible cross-language strategies.

The first approach consists in translating the whole question into the target language, and then processing the question analysis in this target language. This approach is the most widely used, and has for example been chosen by the following systems : (Perret, 2004), (Jijkoun et al., 2004), (Neumann and Sacaleanu, 2005), (de Pablo-Sánchez et al., 2005), (Tanev et al., 2005). Among these systems, several have measured the performance loss between their monolingual and their bilingual systems. Thus, the English-French version of (Perret, 2004) has a 11 % performance loss (in terms of absolute loss), dropping from 24.5% to 13.5% of correct answers. The English-Dutch version of (Jijkoun et al., 2004)'s system has an approximative 10% performance loss of correct answers : the percentage of correct answers drops from 45.5% to 35%. As for (de Pablo-Sánchez et al., 2005), they lose 6% of correct answers between their Spanish monolingual system and their English-Spanish bilingual system. (Hartrumpf, 2005) also conducted an experiment by translating the questions from English to German, and reports a drop from about 50% of performance.

For their cross-language system, (Neumann and Sacaleanu, 2004) chose to use several machine translation tools, and to gather the different translations into a "bag of words" that is used to expand queries. Synonyms are also added to the "bag of words" and EuroWordNet [3] is used to disambiguate. They lose quite few correct answers between their German monolingual system and their German-English bilingual system, with which they obtain respectively 25 and 23.5% of correct answers.

Translating the question raises two main problems : syntactically incorrect questions may be produced, and the resolution of translation ambiguities may be wrong. Moreover, the unknown words such as some proper names are not or incorrectly translated. We will describe later several possibilities to deal with these problems, as well as our own solution.

Other systems such as (Sutcliffe et al., 2005) or (Tanev et al., 2004) use a term-by-term translation. In this approach, the question is analyzed in the source language and then the information returned by the question analysis is translated into the target language. (Tanev et al., 2004), who participated in the Bulgarian-English and Italian-English tasks in 2004, translate the question keywords by using bilingual dictionaries and MultiWordNet [4]. In order to limit the noise stemming from the different translations and to have a better cohesion, they validate the translations in two large corpora, AQUAINT and TIPSTER. This system got a score of 22.5% of correct answers in the bilingual task, and 28% in the monolingual task in 2004. (Sutcliffe et al., 2005) combine two translation tools and a dictionary to translate phrases. Eventually, (Laurent et al., 2005) also translate words or idioms, by using English as a pivot language. The performance of this system is of 64% of correct answers for the French monolingual task, and

[3] Multilingual database with wordnets for several European languages, http ://www.illc.uva.nl/EuroWordNet/

[4] Multilingual lexical database in which the Italian WordNet is strictly aligned with Princeton WordNet, http ://multiwordnet.itc.it

39.5% for the English-French bilingual task.

4 Adopted approach

In order to deal with the conversion from French to English in our system, two strategies are applied in parallel. They differ on what is translated to treat the question asked in French. The first subsystem called MUSQAT proceeds to the question analysis in French, and then translates the question terms extracted by this question analysis module, following the - - - arrows in Figure 1. The second sub-system makes use of a machine translation tool (Reverso [5]) to obtain translations of the questions and then our English monolingual system called QALC is applied, following the ..-.. arrows in Figure 1 . These strategies will be detailed later in the article.

If they represent the most common strategies for this kind of task, an original feature of our system is the implementation of both strategies, which enables us to merge the results obtained by following these strategies, in order to improve the global performance of our system.

In Table 1, we present an analysis of the results we obtained for the CLEF evaluation campaign. We evaluate the results obtained at two different points of the question-answering process, i.e. after the sentence selection (point (a) in Figure 1), and after the answer extraction (point (b) in Figure 1). At point (a), we count how many questions (among the global evaluation set of 200 questions) have an appropriate answer in the first five sentences. At point (b), we distinguish the answers the analysis process labels as named entities (NE), from the others, since the corresponding answering processes are different. We also detail how many answers are ranked first, or in the first five ranks, as we take into account the first five answers.

As illustrated in Table 1, the two strategies for dealing with multilingualism give quite different results, which can be explained by each strategy characteristics.

MUSQAT proceeds to the question analysis with French questions correctly expressed, and which analysis is therefore more reliable. Yet, the terms translations are then obtained from every possible translation of each term, and thus without taking account any context ; moreover, they depend on the quality of the dictionaries used, and introduce noise because of the erroneous translations.

		MUSQAT %	Reverso +QALC %
(a) : Sentences with an answer	first 5 ranks	41	46
(b) : Correct NE answers	rank 1	18	14
	first 5 ranks	26	17
(b) : Correct other answers	rank 1	16	13
	first 5 ranks	23	20
(b) : Total (NE + non NE)	rank 1	17	13
	first 5 ranks	24	19
Final result (fusion of both strategies)		19	

TAB. 1 – Performance of our system in CLEF 2005

In MUSQAT, we do not only translate monoterms (i.e. terms composed of single word) : the biterms (composed of two words) of the French questions are also extracted by the question analysis. Every sequence of two terms which are tagged as *adjective/common noun* or *proper noun/proper noun*... constitutes a biterm. Each word of the biterm is translated, and then the existence of the corresponding biterm built in English is checked in the corpus. The biterms thus obtained are then used by the further modules of the system. Taking biterms into account is useful since they provide a minimal context to the words forming them, as well for the translation as for the re-indexing and re-ranking of the documents (see Figure 1), as explained in (Ferret et al., 2002). Moreover, the presence of the biterm translations in the corpus is a kind of validation of the monoterms translations.

As for translating the question, which is implemented by Reverso+QALC, it presents the advantage of giving a unique translation of the question terms, which is quite reliable. But the grammaticality or realism of the question are not assured, and thus the question analysis, based on regular expression patterns, can be disturbed.

In this work, we tried to evaluate each strategy

[5] http ://www.reverso.net/

and to bypass their drawbacks : on the one hand (Section 5), by examining how the biterm translation in MUSQAT could be more reliable, and on the other hand (Section 6) by improving the question analysis, by relying on the French questions, for QALC.

5 Biterm translation

The translation of terms and biterms present in the question is achieved using two dictionaries. The first of them, which was used last year for our participation to CLEF is Magic-Dic [6]. It is a dictionary under GPL licence, which was retained for its capacity to evolve. Indeed users can submit new translations which are controlled before being integrated. Yet, it is quite incomplete. This year we used FreeDict as well (FreeDict is also under GPL licence), to fill in the gaps of Magic-Dic. FreeDict added 424 translations to the 690 terms already obtained. By mixing both sets of translations we obtained 463 additional biterms, making a total of 777 biterms.

Nevertheless, whatever the quality and the size of the dictionaries are, the problem of biterm translation remains the same : since biterms are not in the dictionaries, the only way for us to get their translation is to combine all the different term translations. The main drawback of this approach is the generated noise, for none of the terms constituting the biterm is disambiguated. For example, three different translations are found for the biterm *Conseil de défense* : *defense council*, *defense advice* and *defense counsel* ; but only the first of those should be finally retained by our system.

To reduce this noise, an interesting possibility is to validate the obtained biterms by searching them or their variants in the complete collection of documents. (Grefenstette, 1999) reports a quite similar experiment in the context of a machine translation task : he uses the Web in order to order the possible translations of noun phrases, and in particular noun biterms. Fastr (Jacquemin, 1996) is a parser which takes as input a corpus and a list of terms (multi or monoterms) and outputs the indexed corpus in which terms and their variants are recognized. Hence, Fastr is quite adequate for biterms validation : it tags all the biterms present in the collection, whether in their original form or in a variant that can be semantic or syntactic.

In order to validate the biterms, the complete collection of the CLEF campaign (500 Mbyte) was first tagged using the TreeTagger, then Fastr was applied. The results are presented Table 2 : 39.5% of the 777 biterms were found in the collection, in a total of 63,404 occurrences. Thus there is an average of 206 occurrences for each biterm. If we do not take into account the biterm which is the most represented (*last year* with 30,981 occurrences), this average falls to 105. The 52 biterms which are found in their original form only are most of the time names of persons. Lastly, biterms that are never found in their original form, are often constituted of one term badly translated, for example the biterm *oil importation* is not present in the collection but its variant *import of oil* is found 28 times. Then, it may be interesting to replace these biterms by the most represented of their variants.

Whenever a biterm is thus validated (found in the collection beyond a chosen threshold), the translation of its terms is itself validated, other translations being discarded. Thus, biterm validation enables us to validate monoterm translations. Then, the following step will be to evaluate how this new set of terms and biterms improves the results of MUSQAT.

After CLEF 2005 evaluation, we had at our disposal the set of questions in their English original version (this set was provided by the organizers). We had also the English translation (far less correct) provided by the automatic translator Reverso.

As we can see it Table 3, for each set of questions the number of terms and biterms is nearly the same. In the set of translations given by Reverso, we manually examined how many biterms were false and found that here again the figures were close to those of the original version. There are two main reasons for which a biterm may be false :

- in two thirds of cases, the association itself is false : the two terms should not have been associated ; it is the case for example of *many country* from the question *How many countries joined the international coalition to restore the democratic government in Haiti ?* [7]
- in one third of cases, one of the terms is not translated or translated with an erroneous term, like *movement zapatiste* coming from the question *What carry the courtiers of the movement zapatiste in Mexico ?* [8]

[6] http ://magic-dic.homeunix.net

[7] This sentence is an example of very good translation given by Reverso

[8] This sentence is an example of bad translation given by

Total Number of biterms	777
Number of biterms found in the collection	307 - 39.5%
Number of biterms found in their original form only	52 - 17%
Number of biterms found with semantic variations only	150 - 54%

TAB. 2 – Magic-Dic and FreeDict biterms validated by Fastr

	Questions in French	Questions translated in English by Reverso	Questions in English (original version)
Terms	1180	1122	1163
Biterms	272	204	261
False Biterms	33	38	27
Common Biterms	-	106	

TAB. 3 – Biterms in the different sets of questions

However, we calculated that among the 204 biterms given by Reverso, 106 are also present in the original set of questions in English. Among the 98 remaining biterms, 38 are false (for the reasons given above). Then, there are 60 biterms which are neither erroneous nor present in the original version. Some of them contain a term which has been translated using a different word, but that is nevertheless correct ; yet, most of these 60 biterms have a different syntax from those constructed from the original version, which is due to the syntax of the questions translated by Reverso.

This leads us to conclude that even if Reverso produces syntactically erroneous questions, the vocabulary it chooses is most of the time adequate. Yet, it is still interesting to use also the biterms constructed from the dictionaries since they are much more numerous and provide variants of the biterms returned by Reverso.

6 Multilingual question analysis

We have developed for the evaluations a question analysis in both languages. It is based on the morpho-syntactic tagging and the syntactic analysis of the questions. Then different elements are detected from both analyses : recognition of the expected answer type, of the question category, of the temporal context...

There are of course lexicons and patterns which are specific to each language, but the core of the module is independent from the language. This module was evaluated on corpora of similar questions in French and in English, and its results on both languages are quite close (around 90% of recall and precision for the expected answer type for example ; for more details, see (Ligozat et al., 2006)).

As presented above, our system relies on two distinct strategies to answer to a cross-language question :

– Either the question is analyzed in the original language, and next translated term-by-term. The question analysis is then more reliable since it processes a grammatically correct question ; yet, the translation of terms has no context to rely on.
– Or the question is first translated into the target language before being analyzed. Although this strategy improves the translation, its main inconvenient is that each translation error has strong consequences on the question analysis. We will now try to evaluate to which extent the translation errors actually influence our question analysis and to find solutions to avoid minimize this influence in the Reverso+QALC system.

An error in the question translation can lead to wrong terms or an incorrect English construction. Thus, the translation of the question “Combien y a-t-il d’habitants en France ?” (“How many inhabitants are there in France ?”) is “How much is there of inhabitants in France ?”.

In order to evaluate our second strategy, Reverso+QALC, using question translation and then a monolingual system, it is interesting to estimate

Reverso, which should have produced *What do supporters of the Zapatistas in Mexico wear ?*

the influence of a such a coarse translation on the results of our system.

In order to avoid these translating problems, it is possible to adapt either the input or the output of the translating module. (Ahn et al., 2004) present an example of a system processing pre- and post-corrections thanks to surface reformulation rules. However, this type of correction is highly dependent on the kind of questions to process, as well as on the errors of the translation tool that is used.

We suggest to use another kind of processing, which makes the most of the cross-lingual character of the task, in order to improve the analysis of the translated questions and to take into account the possibilities of errors in these questions.

Our present system already takes into account some of the most frequent translation errors, by allowing the question analysis module to loosen some of its rules in case the question be translated. Thus, a definition question such as "Qu'est-ce que l'UNITA ?", translated "What UNITA ?" by our translating tool, instead of "What is the UNITA ?", will nevertheless be correctly analyzed by our rules : indeed, the pattern *WhatGN* will be considered as corresponding to a definition question, while on a non-translated question, only the pattern *WhatBeGN* will be allowed.

In order to try and improve our processing of approximations in the translated questions, the solution we suggest here consists in making the question analysis in both the source and the target languages, and in reporting the information (or at least part of it) returned by the source analysis into the target analysis. This is possible first because our system treats both the languages in a parallel way, and second, some of the information returned by the question analysis module use the same terms in English and in French, like for example the question category or the expected Named Entity type.

More precisely, we propose, in the task with French questions and English documents, to analyse the French questions, and their English translations, and then to report the question category and the expected answer type of the French questions into the English question analysis. The information found in the source language should be more reliable since obtained on a real question.

For example, for the question "Combien de communautés Di Mambro a-t-il crée ?" ("How many communities has Di Mambro created ?"), Reverso's translation is "How many Di Mambro communities has he create ?" which prevents the question analysis module to analyze it correctly. The French analysis is thus used, which provides the question category *combien* (*how many*) and the expected named entity type *NUMBER*. This information is reported in the English analysis file.

These characteristics of the question are used at two different steps of the question answering process : when selecting the candidate sentences and when extracting the answers. Improving their reliability should then enable us to increase the number of correct answers after these two steps.

In order to test this strategy, we conducted an experiment based on the CLEF 2005 FR-EN task, and the 200 corresponding French questions. We launched the question answering system on three question files :

- The first question file (here called English file) contained the original English questions (provided by the CLEF organizers). This file will be considered as a test file, since the results of our system on this file represent those that would be reached without translation errors.
- The second file (called Translated file) contained the translated questions analysis.
- The last file (called Improved file) contained the same analysis, but for which the question category and the expected answer type were replaced by those of the French analysis.

Then we searched for the number of correct answers for each input question file after the sentence selection and after the answer extraction. The results obtained by our system on each file are presented on Figure 2, Figure 3 and Figure 4. These figures present the number of questions expecting a named entity answer, expecting another kind of answer, and the total number of questions, as well as the results of our system on each type of question : the number of correct questions are given at the first five ranks, and at the first rank, first for the sentences ("long answers") and then for the short answers.

These results show that the information transfer from the source language to the target language significantly improves the system's results ; the number of correct answers increases in every case. It increases from 34 on the translated questions file to 36 on the improved file, and from 52

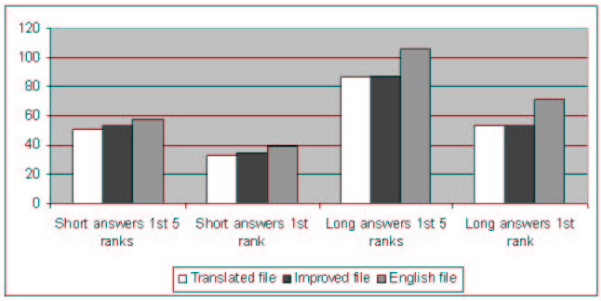

FIG. 2 – QALC's results (i.e. number of correct answers) on the 200 questions

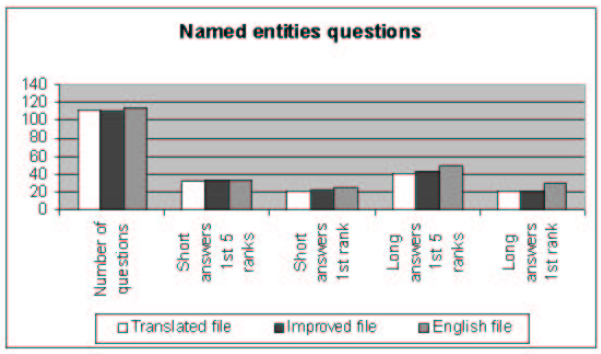

FIG. 3 – Results on the named entities questions

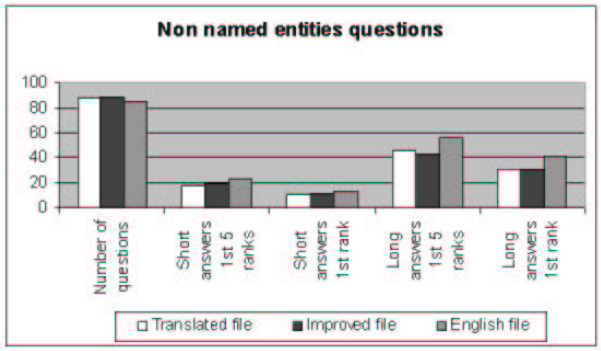

FIG. 4 – Results on the non named entities questions

to 55 for the first 5 ranks. These results are closer to those of the monolingual system, which returns 41 correct answers at the first rank, and 59 on the first 5 ranks.

It is interesting to see that the difference between the monolingual and the bilingual systems is less noticeable after the sentence selection step than after the answer extraction step, which tends to prove that the last step of our process is more sensitive to translation errors. Moreover, this experiment shows that this step can be improved thanks to an information transfer between the source and the target languages. In order to extend this strategy, we could also match each French question term to its English equivalent, in order to translate all the information given by the French analysis into English. Thus, the question analysis errors would be minimized.

7 Conclusion

The originality of our cross-language question answering system is to use in parallel the two most widely used strategies for shifting language, which enables us to benefit from the advantages of each strategy. Yet, each method presents drawbacks, that we tried to evaluate in this article, and to bypass.

For the term-by-term translation, we make the most of the question biterms in order to restrict the possible translation ambiguities. By validating the biterms in the document collection, we have improved the quality of both the biterms and the monoterms translations. We hope this improvement will lead to a better selection of the candidate sentences from the documents.

For the question translation, we use the information deduced from the source language to avoid the problems coming from a bad or approximative translation. This strategy enables us to solve some of the problems coming from non-grammatical translations ; matching each term of the French question with its English equivalent would enable us to transfer all the information of the French analysis. But the disambiguation errors of the translation remain.

References

Kisuh Ahn, Beatrix Alex, Johan Bos, Tiphaine Delmas, Jochen L. Leidner, and Matthew B. Smillie. 2004. Cross-lingual question answering with QED.

In *Working Notes, CLEF Cross-Language Evaluation Forum*, pages 335–342, Bath, UK.

César de Pablo-Sánchez, Ana González-Ledesma, José Luis Martínez-Fernández, José Maria Guirao, Paloma Martinez, and Antonio Moreno. 2005. MIRACLE's 2005 approach to cross-lingual question answering. In *Working Notes, CLEF Cross-Language Evaluation Forum*, Vienna, Austria.

Olivier Ferret, Brigitte Grau, Martine Hurault-Plantet, Gabriel Illouz, Christian Jacquemin, Laura Monceaux, Isabelle Robba, and Anne Vilnat. 2002. How NLP can improve question answering. *Knowledge Organization*, 29(3-4).

Gregory Grefenstette. 1999. The world wide web as a resource for example-based machine translation tasks. In *ASLIB Conference on Translating and the Computer*, volume 21, London, UK.

Sven Hartrumpf. 2005. University of Hagen at QA@CLEF 2005 : Extending knowledge and deepening linguistic processing for question answering. In *Working Notes, CLEF Cross-Language Evaluation Forum*, Vienna, Austria.

Christian Jacquemin. 1996. A symbolic and surgical acquisition of terms through variation. *Connectionist, Statistical and Symbolic Approaches to Learning for Natural Language Processing*, pages 425–438.

Valentin Jijkoun, Gilad Mishne, Maarten de Rijke, Stefan Schlobach, David Ahn, and Karin Muller. 2004. The University of Amsterdam at QA@CLEF2004. In *Working Notes, CLEF Cross-Language Evaluation Forum*, pages 321–325, Bath, UK.

Dominique Laurent, Patrick Séguéla, and Sophie Nègre. 2005. Cross lingual question answering using QRISTAL for CLEF 2005. In *Working Notes, CLEF Cross-Language Evaluation Forum*, Vienna, Austria.

Anne-Laure Ligozat, Brigitte Grau, Isabelle Robba, and Anne Vilnat. 2006. L'extraction des réponses dans un système de question-réponse. In *Traitement Automatique des Langues Naturelles (TALN 2006)*, Leuven, Belgium.

Günter Neumann and Bogdan Sacaleanu. 2004. Experiments on robust NL question interpretation and multi-layered doument annotation for a cross-language question / answering system. In *Working Notes, CLEF Cross-Language Evaluation Forum*, pages 311–320, Bath, UK.

Günter Neumann and Bogdan Sacaleanu. 2005. DFKI's LT-lab at the CLEF 2005 multiple language question answering track. In *Working Notes, CLEF Cross-Language Evaluation Forum*, Vienna, Austria.

Laura Perret. 2004. Question answering system for the French language. In *Working Notes, CLEF Cross-Language Evaluation Forum*, pages 295–305, Bath, UK.

Richard F.E. Sutcliffe, Michael Mulcahy, Igal Gabbay, Aoife O'Gorman, Kieran White, and Darina Slattery. 2005. Cross-language French-English question answering using the DLT system at CLEF 2005. In *Working Notes, CLEF Cross-Language Evaluation Forum*, Vienna, Austria.

Hristo Tanev, Matteo Negri, Bernardo Magnini, and Milen Kouylekov. 2004. The DIOGENE question answering system at CLEF-2004. In *Working Notes, CLEF Cross-Language Evaluation Forum*, pages 325–333, Bath UK.

Hristo Tanev, Milen Kouylekov, Bernardo Magnini, Matteo Negri, and Kiril Simov. 2005. Exploiting linguistic indices and syntactic structures for multilingual question answering : ITC-irst at CLEF 2005. In *Working Notes, CLEF Cross-Language Evaluation Forum*, Vienna, Austria.