The bilingual system MUSCLEF at QA@ CLEF 2006

Brigitte Grau, Anne-Laure Ligozat, Isabelle Robba, Anne Vilnat, Michael Bagur, Kevin Séjourné

The bilingual system MUSCLEF at QA@CLEF 2006

Brigitte Grau, Anne-Laure Ligozat, Isabelle Robba, Anne Vilnat
Michael Bagur, and Kevin Séjourné

LIR group, LIMSI-CNRS, BP 133 F-91403 Orsay Cedex, France,
firstName.name@limsi.fr
http://www.limsi.fr/Scientifique/lir/

Abstract. This paper presents our bilingual question answering system MUSCLEF. We underline the difficulties encountered when shifting from a mono to a cross-lingual system, then we focus on the evaluation of three modules of MUSCLEF: question analysis, answer extraction and fusion. We finally present how we re-used different modules of MUSCLEF to participate in AVE (Answer Validation Exercise).

Key words: Question answering, evaluation, multiword expressions

1 Introduction

This paper presents our cross-lingual question answering system, called MUSCLEF that participated in the QA@CLEF 2006 French-English cross-language task for which two runs were submitted. As for the past two years, we used two strategies: the first one consists in translating only a set of terms selected by the question analysis module, this strategy being implemented in a system called MUSQAT; the second one consists in translating the whole question and then applying our monolingual system named QALC. In the previous CLEF campaigns, most of the systems used exclusively question translation. This year, several systems used either term translation ([1], [2]) or an hybrid approach ([3], [4]) consisting in question translation plus term translation. Both approaches are interesting since not relying entirely on question translation: translation tools - when they exist - are not efficient for all types of questions or fail to translate some named entities or multiword expressions. The paper is organized according to the following plan: first we describe the architecture of MUSCLEF (section 2), then we underline the difficulties encountered when shifting from a mono to a cross-lingual system (3), next we focus on the evaluation of this system and give the results obtained by three particular modules of MUSCLEF (4). We also give the general results of our participation to CLEF (5). Lastly, we present how we re-used different modules of MUSCLEF to build a first system for the Answer Validation Exercise (6).

2 System overview

QALC, our monolingual system, is composed of four modules described below, the first three of them being classical modules of question answering systems: (a) the first module analyzes the question and detects characteristics that will enable us to finally get the answer: the expected answer type, the focus, the main verb and some syntactic features; (b) the second module is the processing of the collection of documents: a search engine, named MG [1], is applied; then the returned documents are reindexed according to the presence of the question terms, and finally a module recognizes the named entities and each sentence is weighted according to the information extracted from the question; (c) the third module is the answer extraction which applies two different strategies depending on whether the expected answer is a named entity or not; (d) the fourth module is a fusion. Indeed our system QALC is applied on the Web as well as on the closed collection of the CLEF evaluation, then a comparison of both sets of answers is done; this way, we increase the score of answers that are present in both sets.

To build MUSCLEF, our cross-lingual question anwering system, we added several modules to QALC, corresponding to both possible strategies to deal with cross-lingualism: question translation and term-by-term translation. In MUSCLEF, the first strategy uses Reverso [2] to translate the questions then our monolingual system QALC is applied. The second strategy, that we named MUSQAT, uses different dictionnaries to translate the selected terms (a description and an evaluation of this translation are given section 3). Finally, we apply the fusion module to the different sets of answers: a first one corresponds to MUSQAT, a second one corresponds to the application of QALC on the translated questions, both these sets of answers coming from the CLEF collection of documents, and a third one corresponds to the application of QALC on the translated questions using the Web. MUSCLEF is presented Figure 1, where the first line of modules corresponds to our monolingual system QALC and the second line contains the modules necessary to deal with cross-lingualism.

3 Shifting from a monolingual to a cross-lingual system

3.1 Performance comparison

After the CLEF 2005 evaluation, CLEF organizers gave the original set of questions written in *good English* to the participants. Thanks to this new set of questions we could compare the behaviour of our different implementations: monolingual QALC, cross-lingual QALC (using Reverso), and cross-lingual MUSQAT. The results are given in Table 1. The results of document selection and document processing were calculated for 180 questions instead of 200 because of the 20 NIL questions. Each number in this table represents the percentage of questions for which a good document/sentence/answer is returned.

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

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

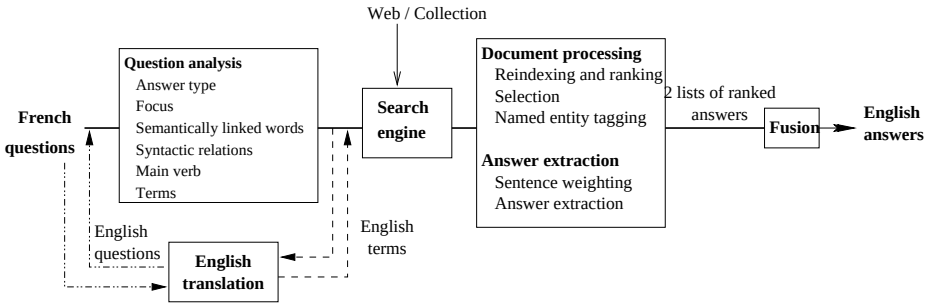

Fig. 1. Architecture of MUSCLEF, our cross-language question answering system

	Monolingual system	Cross-lingual systems	
	QALC	QALC + Reverso	MUSQAT
Document selection	94.4	88.3	84.4
Document processing	93.3	87.7	82.2
Five sentences	67.5	58.5	50.5
Five short answers	40	39.5	36.5
First short answer	28	26	23

Table 1. Comparison of monolingual QALC, cross-lingual QALC and MUSQAT

Concerning the first three lines, we observe a big difference between the monolingual and the cross-lingual systems (from to 6 to 17 %). This difference is due to missing translations: for instance acronyms or proper names (which original alphabet can be different from ours) are often not correctly translated. In the last two lines, the differences are more surprising (and are due to problems in the answer extraction module): the monolingual system lost 40% of good answers during answer extraction, while the best cross-lingual system, QALC+Reverso lost 32.5%, and MUSQAT lost 27.7%. The results of the monolingual and cross-lingual systems on short answers are thus quite close. On the same data of CLEF 2005, [5] made also this kind of comparison: they report a loss of 24.5% of good answers between their monolingual French system QRISTAL (which obtains very high results: 64%) and their English-to-French system.

3.2 Corpus-based translation validation

In this section, we present how in MUSQAT proceeds for the term and multiterm translation and their validation. The translation is done by using two dictionaries, Magic-Dic and FreeDict [3], both being under GPL licence. Thus, the system MUSQAT gets several translations for each French word, which can be either synonyms or different translations when the term is polysemic. The evaluation made last year (reported in [6]) on term translation encouraged us to enhance our approach by the validation of these translations. To proceed to

[3] http://magic-dic.homeunix.net/ and http://freedict.org/en/

this validation, we used Fastr [4] and searched a subset of documents (from 700 to 1000 documents per question) of the CLEF collection for either the bi-terms or syntactic variants of them. When neither a bi-term translation nor a variant was found, we discarded the corresponding translated terms. For example, to the French bi-term *cancer du sein* corresponded the three following translations: *breast cancer*, *chest cancer* and *bosom cancer*. In the documents only the first translation is present, which leads us to discard the terms *chest*, *bosom* and their corresponding bi-term. We hoped this could decrease the noise due to the presence of wrong translations. Unfortunately, this first experience in translation validation was not convincing, since we obtained nearly the same results with or without it. (22% of good answers without the validation, 23.5% with it). Undoubtedly this approach needs to be enhanced but also evaluated on larger corpora. Indeed, we only evaluated it on the corpus of CLEF 2005 questions, on which we obtained the following figures: from the 199 [5] questions, we extracted 998 bi-terms from 167 questions and 1657 non empty mono-terms; only 121 bi-terms were retrieved in documents, which invalidated 121 mono-terms and reduced the number of questions with at least one bi-term to 98. The number of invalidated mono-terms (121) is certainly not high enough in this first experiment to enable MUSQAT to reduce the noise due to wrong translations.

3.3 On-line term translation

Yet, after the translation and its validation, some terms and multiterms remain untranslated. Web resources were then used to help find the missing translations, such as on-line dictionaries like Mediaco, Ultralingua or other dictionaries from Lexilogos [6], the free encyclopedia Wikipedia, and the web site for European languages and cultures Eurocosm. Many of the terms that remain untranslated are multiword terms requiring a special strategy which is composed of three steps. First, all the multiword terms are cut into single words. Then the Web is browsed to get pages from all the on-line resources that contain these words. Each page is mapped into a common format which gives for each term its different translations. Finally, for each term of the original list, all exact matches are searched in the Web pages, and the most frequent translation is chosen. Table 2 shows an example of a mapping for the French term "voiture" and Table 3 the frequency of each of its translations. To avoid incorrect translations, only the translations of the exact term are considered. For the term "voiture", the most frequent translation is "car" and thus this translation is chosen. The results are summed up in Table 4: for each corpus, about 30 % of the originally untranslated terms were translated thanks to this new module.

[4] Fastr (http://www.limsi.fr/Individu/jacquemi/FASTR/), which was developed by Christian Jacquemin, is a transformational shallow parser for the recognition of term occurrences and variants

[5] 199 instead of 200 because one has been thrown out by the process

[6] www.lexilogos.com

French term	Translation
voiture	car
voiture	carriage
voiture d'enfant	baby-carriage
...	...
voiture	car
voiture	automobile
voiture de fonction	company car
...	...
voiture	car
...	...
clé de voiture	car key
voiture	automobile

Table 2. Translations of the French term *voiture*

Translation	# of occurrences
car	3
automobile	2
coach	1
carriage	1

Table 3. Frequency of the different translations of *voiture*

Corpus	CLEF 2005	CLEF 2006
# of translated terms	195 (33%)	408 (32%)
# of untranslated terms	394 (67%)	858 (68%)
Total # of terms	589	1266

Table 4. On-line term translation results

4 Evaluation of MUSCLEF modules

4.1 Question analysis

The question analysis module determines several characteristics of the question among which its category, expected answer type (named entity or not) and focus. We conducted a corpus study in order to validate our choice concerning these characteristics, and the focus in particular, on the corpus of English questions and collection. For the focus, we found that 54% of the correct answers contain the focus of the question, while only 32% of the incorrect answers do (against 20% and 11% for an non-empty word chosen by chance in the question), which tends to validate the choice we made for the focus. The performance of this module was evaluated in [7] which estimated its precision and recall at about 90% in monolingual tasks. The performance is lower on translated questions, since the question words or the structure of the question can be incorrectly translated. For example, the question *Quel montant Selten, Nash et Harsanyi ont-ils reçu pour le Prix Nobel d'Economie ?* (*How much money did Selten, Nash and Harsanyi receive for the Nobel Prize for Economics?*) is translated into *What going up Selten, Nash and Harsanyi did they receive for the Nobel prize of economy?*, which prevents us from determining the right expected answer type: *FINANCIAL_AMOUNT*.

4.2 Answer extraction

In MUSQAT and QALC, the same method was used to extract the final short answer from the candidate sentence. In both systems, this last step of the question answering process entails an important loss of performance. Indeed, in MUSQAT and QALC the percentage of questions for which a candidate sentence containing the correct answer is ranked first is around 35%, and as seen in section 3 the percentage of questions for which a correct short answer is ranked first falls to around 25%. During this step, about one third of good answers is lost. In

[7], we exposed the reasons of the low performances of our answer extraction module. The patterns used to extract the answer when the expected type is not a named entity have been improved for the definition questions. In our last test, indeed, 21 questions among the 48 definition questions of CLEF 2005 were correctly tagged by the patterns. But in other cases, patterns still show a very low efficiency, for here linguistic variations are more important and remain usually difficult to manage.

4.3 Fusion

Since we now have three sets of results to merge, we proceeded in two steps: we first merged the results of QALC+Reverso and MUSQAT, which gave us our first run. And, as a second run, we merged our first run and the set obtained with QALC+web system. On CLEF 2005 data, the fusion step was not really convincing in terms of results: while QALC+Reverso gave 26 % of good answers and MUSQAT 22.5 %, Run1 gave only 25 % and Run2 27 %. Nevertheless, as we can see it section 5, on CLEF 2006 results the second fusion using the web gave better results since we obtained 25 % of good answers with the web and 22 % without. Concerning the first fusion, both systems (QALC+Reverso and MUSQAT) giving similar results, it is not surprising that the fusion does not increase the number of good answers. However, our fusion algorithm (described in [8]) is mainly based on the scores attributed by the different systems to their answers, and does not take into account the performances of the systems themselves, which could be a interesting way to improve it.

5 Results

Table 5 reports the results we obtained at the CLEF 2006 evaluation. As described just above (sub-section 4.3), we remind that the first run is the result of the fusion of two systems: QALC+Reverso and MUSQAT, while the second run is the result of the fusion of this first run and of QALC+Web. Last year, the best of our runs obtained a score of 19%, so the improvements brought to our systems can be considered as encouraging. The difference of results between both runs strengthens the idea that the use of an external source of knowledge is an interesting track to follow. We underline, that at the time of writing this paper, only the first answer of each question has been assessed, so the MRR score does not bring more information than the number of good first short answers. The four first lines of results concern 190 questions, the 10 remaining questions were list questions for which the score is on the last line.

6 Answer validation

In order to build the Answer Validation system, we used our QA system, applied to the hypotheses and justifications rather than to the questions and the

	Run 1	Run 2
First short answer	22.63%	25.26%
Confidence Weighted Score (CWS)	0.08556	0.15447
Mean Reciprocal Rank Score (MRR)	0.2263	0.2526

Table 5. MUSCLEF results at CLEF 2006

collection, and we added a decision module. Our goal was to obtain the information needed to decide whether the answer was entailed by the text proposed to validate it. First the initial corpus file is processed to obtain the input format of our system, then the QA system is used to extract needed information from it, like a tagged hypothesis, a tagged justification snippet or terms extracted from the question. They are written in an xml file passed to the decision algorithm. We also get the answer our QA system would have extracted from the proposed justification, which is used to see if the answer to judge is likely to be true. Then, the decision algorithm proceeds in two main steps. During the first one, we try to detect quite evident mistakes, such as the answers which are completely enclosed in the question, or which are not part of the justification. The second step proceeds to more sophisticated verifications: (a) verifying the adequate type of the expected named entity if there is one; (b) looking the justification for terms judged as important during the question analysis; (c) confirming the decision with an extern-justification module using the latest version of Lucene to execute a number of coupled queries on the collection, like proximity queries (checks if a number of terms can be found close to one another within a text); the top results of each couples queries are compared in order to decide whether the answer is likely to be true or not; (d) comparing the results that our answer-extraction module (part of our QA system) would provide from the justification text. The results obtained by these different verifications are combined to decide if the answer is justified or not and to give a confidence score to this decision. Some errors have been corrected after submitting our results to the AVE campaign (which were rather bad, with very few positive answers). The results are given in Table 6 for the 3064 pairs judged during the evaluation campaign (201 pairs received an *unknown* evaluation). The "YES" (resp. "NO") column corresponds to the pairs which our system judged as justified (resp. not justified). Among the *Correct NO*, those obtained during the first step previously presented, for which our system was sure of the answer (and judged as "NO" with the maximum confidence score 1), were distinguished from the others. The same distinction was established between the *Incorrect NO*. Among the 138 Incorrect "sure" NO , we observed that 63 were *badly* judged validations: the answer is correct, but the text given to justify this answer does not give any validation. For example, the pair with id="5496" is considered as validated, while the document given as a validation *des agresseurs de Naguib Mahfouz* does not contain the answer *EGYPTE* (which is nevertheless present in the hypothesis *Naguib Mahfouz a été poignardé à ÉGYPTE*).

	YES	NO		
Correct	177	2291	sure	1370
			not sure	921
Incorrect	68	528	sure	138
			not sure	390
Precision/Recall	0.72/0.25	0.81/0.97		

Table 6. AVE results at CLEF 2006

7 Conclusion

Our cross-lingual system MUSCLEF presents the particularity to use three strategies in parallel: question translation, term-by-term translation and the use of another source of knowledge (actually limited to the Web). The three sets of answers are finally merged thanks to a fusion algorithm proceeding on two set of answers at the same time. The term-by-term strategy gives lower results than the most widely used strategy consisting in translating the question into the target source then applying a monolingual strategy. Nevertheless, we think it remains interesting from the multilingualism point of view, and we try to improve it by using of different techniques of translation (use of several dictionaries and on-line resources) and validation.

References

1. Laurent, D., Séguéla, P., Nègre, S.: Cross lingual question answering using qristal for clef 2006. In: CLEF Working Notes, Alicante, Spain (2006)
2. Cassan, A., Figueira, H., Martins, A., Mendes, A., Mendes, P., Pinto, C., Vidal, D.: Priberam's question answering system in a cross-language environment. In: CLEF Working Notes, Alicante, Spain (2006)
3. Sutcliffe, R.F.E., White, K., Slattery, D., Gabbay, I., Mulcahy, M.: Cross-language french-english question answering using the dlt system at clef 2006. In: CLEF Working Notes, Alicante, Spain (2006)
4. Bouma, G., Fahmi, I., Mur, J., van Noord, G., van der Plas, L., Tiedemann, J.: The university of groningen at qa@clef 2006: Using syntactic knowledge for qa. In: CLEF Working Notes, Alicante, Spain (2006)
5. Laurent, D., Séguéla, P., Nègre, S.: Cross lingual question answering using qristal for clef 2005. In: CLEF Working Notes, Vienna, Austria (2005)
6. Ligozat, A.L., Grau, B., Robba, I., Vilnat, A.: Evaluation and improvement of cross-lingual question answering strategies. In: Workshop on Multilingual Question Answering, EACL, Trento, Italy (2006)
7. Ligozat, A.L., Grau, B., Robba, I., Vilnat, A.: L'extraction des réponses dans un système de question-réponse. In: TALN Conference, Leuven, Belgium (2006)
8. Berthelin, J.B., de Chalendar, G., Elkateb-Gara, F., Ferret, O., Grau, B., Hurault-Plantet, M., Illouz, G., Monceaux, L., Robba, I., Vilnat, A.: Getting reliable answers by exploiting results from several sources of information. In: CoLogNET-ElsNET Symposium, Question and Answers : Theoretical and Applied Perspectives, Amsterdam, Holland (2003)