Term Translation Validation by Retrieving Bi-terms

Brigitte Grau, Anne-Laure Ligozat, Isabelle Robba, Madeleine Sialeu and Anne Vilnat
LIR group, LIMSI-CNRS, BP 133 91403 Orsay Cedex
firstName.name@limsi.fr

Abstract

For our second participation to the Question Answering task of CLEF, we kept last year's system named MUSCLEF, which uses two translation strategies implemented in two modules. The multilingual module MUSQAT analyzes the French questions, translates "interesting parts", and then uses these translated terms to search the reference collection. The second strategy consists in translating the question in English and applying QALC our existing English module. Our purpose in this paper is to analyze term translations and propose a mechanism for selecting correct ones. The manual evaluation of bi-terms translation leads us to the conclusion that bi-term translations found in corpus can confirm mono-term translations.

1 Introduction

This paper presents our second participation to the Question Answering task of CLEF evaluation campaign. This year we have participated in two tasks: a monolingual task (in French) for which we submitted one run, and a bilingual task (questions in French, answers in English) for which we submitted two runs. Concerning the bilingual task, we used the same two strategies as last year:

- translation of selected terms issued of the question analysis module, then search in the collection, this first system is called MUSQAT
- question translation thanks to a machine translation system, then application of QALC our monolingual english system

Those strategies are the most commonly adopted, but to our knowledge, any other system except our own implements both. Our whole system is called MUSCLEF; since its architecture is practically the same as last year ([Peters et al. 2005]), we rather chose to present an evaluation of the different translations used last year in MUSCLEF. This manual evaluation was time-consuming thus it has not yet been done for this year's data. It remains nevertheless relevant and lead us to propose a mechanism for selecting correct term translations.

We will first present an overview of our system, then we will focus on our recognition of terms in documents, realized by Fastr, and their translation. We will then present an evaluation of these translations followed by results concerning term validation and our global results at the QA task.

2 Overview of MUSCLEF

The global architecture of MUSCLEF is illustrated Figure 1. First, its question analysis module aims at deducing characteristics which may help to find possible answers in selected passages. These characteristics are: the expected answer type, the question focus, the main verb and some syntactic characteristics. They are deduced from the morpho-syntactic tagging and syntactic analysis of the question. For this campaign, we developped a grammar of question and used the

Cass robust parser[1] to analyze the English questions that were translated using Reverso[2]. As a new type of questions, the temporally restricted questions, was introduced in this year's campaign, we have adjusted question analysis to the category of the question. When a temporal restriction was to be found, we tried to detect it, and to classify it according to the three following types: date, period, and event. The answering strategy was then adapted to the type of temporal constraint.

For querying the CLEF collection and retrieving passages we used MG[3]. Retrieved documents are then processed: they are re-indexed by the question terms and their linguistic variants, re-ordered 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.

The first run we submitted corresponds to the strategy implemented in MUSQAT: translation of selected terms. For the second run, we added a final step consisting in comparing the results issued from both strategies: the translated questions and the translated terms. This module named fusion in Figure 1, computes a final score for each potential answer, its principle is to boost an answer if both chains ranked it in the top 5 propositions, even with relatively low scores.

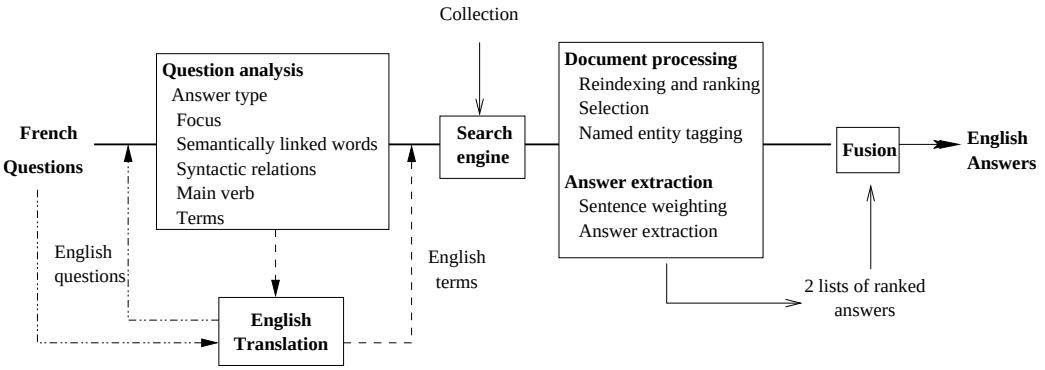

Figure 1: MUSCLEF architecture

3 Searching terms and variants

The automatic indexing of documents is performed by FASTR, a transformational shallow parser for the recognition of term occurrences and variants. Terms are transformed into grammar rules and the single words building these terms are extracted and linked to their morphological and semantic families. The morphological family of a single word w is the set M(w) of terms in the CELEX database ([CELEX 1998]) which have the same root morpheme as w. For instance, the morphological family of the noun *maker* is made of the nouns *maker, make and remake*, and the verbs *to make and to remake*. The semantic family of a single word w is the union S(w) of the synsets of WordNet1.6 ([Fellbaum 1998]) to which w belongs. A synset is a set of words that are synonymous for at least one of their meanings. Thus, the semantic family of a word w is the set of the words w' such that w' is considered as a synonym of one of the meanings of w. The semantic family of *maker*, obtained from WordNet1.6, is composed of three nouns: *maker, manufacturer, shaper* and the semantic family of *car* is *car, auto, automobile, machine, motorcar*. Variant

[1]http://www.vinartus.net/spa/
[2]http://www.reverso.net
[3]MG for Managing Gigabytes http://www.cs.mu.oz.au/mg/

patterns that rely on morphological and semantic families are generated through metarules. They are used to extract terms and variants from the document sentences in the selected documents.

For instance, the following pattern, named NtoSemArg, extracts the occurrence *making many automobiles* as a variant of the term *car maker*:

VM('maker') RP? PREP? ART? (JJ—NN—NP —VBD—VBG)0-3 NS('car')

where RP are particles, PREP prepositions, ART articles, and VBD, VBG verbs. VM('maker') is any verb in the morphological family of the noun *maker* and NS('car') is any noun in the semantic family of *car*.

Relying on the above morphological and semantic families, *auto maker, auto parts maker, car manufacturer, make autos,* and *making many automobiles* are extracted as correct variants of the original term *car maker* through the set of metarules used for the QA-track experiment. Unfortunately, some incorrect variants are extracted as well, such as *make those cuts in auto* produced by the preceding metarule.

4 Term translation

Different methods can be used to achieve term translation and we considered the easiest one, which consists in using a bilingual dictionary to translate the terms from the source language to 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. 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[4], Systran[5], Google[6], Dictionnaire Terminologique[7] or FreeTranslation [8]), and the dictionaries under GPL licences (such as Magic-Dic[9], Unidic or FreeDict[10]). 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, and FreeDict. For example the query for the French word *porte* to Magic-Dic gives the following results (we only give an excerpt):

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

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

- porte parole, locuteur - spokesman
- porte - door, gate

To prevent Magic-dic uncompleteness, and because it has been proved that the use of several dictionaries gives better results than a unique one, we used this year two dictionaries and merged their translations. FreeDict had added 424 different translations of the 690 words. However, these new translations are mainly other synonyms rather than new translations of unknown words.

4.1 The multilingual module MUSQAT

We illutrate the strategy defined in our multilingual module 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 mono-terms and all the bi-terms (such as *adjective/common noun*) which are in the question, and eliminates the stop words. The bi-terms are useful, because they allow a disambiguisation by giving a (small) context to a word. In our example, the bi-terms (in their lemmatized form) are: *principal compagnie, compagnie aérien, aérien allemand*; and the mono-terms: *nom, principal, compagnie, aérien, allemand*.

With the help of the dictionnaries, MUSQAT attempts to translate the bi-terms (when they exist), and the mono-terms. 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 mono-terms, following the English syntax. For our example, we obtained for the bi-terms: *principal company/main company, air company, air german*; and for the mono-terms: *name/appellation, principal/main, company, german*. When a word does not exist in the dictionnaries, we keep it as it whithout any diacritic, which is often relevant for proper nouns. Then, all the words are weighted relative to their existence in a lexicon that contains the vocabulary found in Latimes of the Trec collection, so that each word is weighted according to its specificity within this corpus. If a word is not found in this lexicon, we search with MG if documents contain it (or rather its root because MG indexation was made using stemming). If it is not the case, MUSQAT eliminate it from the list of translated terms. By this way, MUSQAT discarded 72 non-translated words (on 439 non-translated mono-terms, the remaining ones often being proper nouns). As we form boolean requests, it was important not to keep inexisting words.

English terms plus their categories (given by the Tree Tagger) were then given as input to the other modules of the system, instead of the original words. The translation module did not try to solve the ambiguity between the different translations. We account on the document retrieval module to discard irrelevant translations. This module has been improved this year: it always selects passages (the collection was preliminary splitted), but in a very smaller number. It first generates boolean requests, based on proper nouns, numbers and specificity of the words. It aims at retrieving 200 passages maximum, and makes the smaller request with the more specific terms so as to obtain a minimum number of passages, set to 50. Each term of the request is made of the disjonction of the different translations. If the boolean query leads to retrieving too few or too much documents, passage retrieval is made thanks to a ranked research with a query that hold all the terms. If different terms are synonyms, relevant documents are then retrieved with these synonyms. If a word is incoherent within the context, we suppose its influence is not sufficient to generate noise. This hypothesis can only be verified if the question is made of several words.

5 Magic-dic term evaluation

We manually evaluated the bi-term translations for the 200 questions of CLEF04 given by this module. Table 1 presents the results of this evaluation. The system found 375 bi-terms. Among them, 135 are correct translated bi-terms (OK) such as *CERN member*. 24 are bi-terms contextually false i.e. for which one word is not a good translation in the context of this bi-term, such as

accretion hormone instead of *growth hormone* to translate *hormone de croissance*. 74 bi-terms are due to an erroneous bi-term constitution (False Bi-Terms), such as *able animal* in question asking to *Give an animal able to....* Finally, 142 bi-terms are (a) completely erroneous translations (False Translation), such as *overground escort* instead of *main company* (110) or (b) the translation was absent from the dictionnary (Absent Translations), such as *olympique*, where the French word has been kept instead of the English term*olympic* (32).

Table 1: MagicDic terms evaluation

Bi-Terms	#	%
OK	135	36
Contextually False	24	6.4
False	74	19.7
False Transl	110	29.3
Absent Transl	32	8.5
Total False	240	64
Total	375	

It is obvious on this table that a lot of terms are wrong for different reasons. We decided to confirm those that must be kept by considering their presence or absence in the selected documents. To do so, we used FASTR results to evaluate the bi-terms or their variants which are retrieved in the documents. Table 2 shows the results of this evaluation. The second column gives the results obtained by FASTR without considering the semantic variations. The third column includes these semantic variations. The last column indicates the percentage of bi-terms FASTR confirms, taking into account the semantic variations.

Table 2: MagicDic terms validated by Fastr

Bi-terms	#	#retrieved without sem.var.	#retrieved including sem.var.	%
OK	135	61	83	61.5
Context. False	24	4	7	29.2
False	74	11	15	20.3
False Transl	110	7	19	17.3
Absent Transl	32	0	0	0
Total	375	82	120	32

The correct bi-terms are mostly confirmed by FASTR. The contextually false bi-terms obtain a rather high percentage of confirmation due to the semantic variations which lead to recognize correct synonyms of non accurate translated terms. The false bi-terms can be considered as co-occurrences rather than bi-terms. As co-occurrences, they are retrieved by FASTR in the documents and just a few false translations are retrieved.

6 Evaluation of terms extracted from question translations

We also proceeded to a similar evaluation of the terms extracted from the questions translated last year by Systran.

As a first step we proceeded to an evaluation of the question translations themselves. We evaluated the syntactic quality of the translations, and classified them in correct, false, or quite correct. Table 3 recapitulates these results.

Table 3: Questions translations evaluation

Questions	Correct	Quite Correct	False	Total
#	73	12	115	200
%	36.5	6.0	57.5	100

We also evaluated the terms extracted from these translated questions by our monolingual system QALC. We use the same notations than in table 1. Results are given Table 4.

Table 4: Evaluation of terms from translated questions

Bi-Terms	#	%
OK	126	75.4
Contextually False	0	0
False	41	24.6
False Transl	0	0
Absent Transl	0	0
Total False	41	24.6
Total	167	

These results are quite interesting: despite the moderate quality of the translations, QALC is able to identify good terms from these questions. We can also notice that we obtain a smaller number of terms following this procedure because there is only one translation by word.

7 Results

Table 5 gives the results that our system obtained at the CLEF04 and CLEF05 campaigns, with the different strategies: (a) with the translation of the terms (MUSQAT), (b) with QALC applied on the translated questions and searching the collection. 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 CLEF04 and 188 for the CLEF05 questions.

The first line indicates the number of correct answers found in the 5 first sentences given by MUSQAT (using term translation) and QALC. The second line, "NE answers", gives the number of correct answers on questions the system categorized as waiting for a Named Entity (the total is 107 in CLEF04 for MUSQAT and 97 for QALC and 91 in CLEF05 for MUSQAT and 66 for QALC). Our total number of questions of this category is far beyond the real number in CLEF05. The third line, "non NE answers", concerns the other questions (the complement to 178 in CLEF04 and to 188 in CLEF05). Results are presented when the system just gives one answer and when it gives 5 answers. The last line indicates the best official result of our system on the 200 questions. The official score of MUSQAT was 22 (11%) in CLEF04 and 28 (14%) in CLEF05, thus we can observe that merging answers obtained by different strategies enables a significative gain. We also can notice that if our CLEF05 system better selects sentences, it is less performant on extracting the answers, specially on named entity answers.

According to the manual evaluation results of bi-terms translations, we have tested an automatic process for filtering Magic-dic translations on CLEF04 questions. So, if a bi-term or a variant form was found in the selected documents, we kept it as a valid translation and we kept its lemmas as valid mono-term translations. When a validated translation existed for a term, the non-validated translations were taken out. When no translation of a bi-term was found in corpus, we assumed that mono-term translations were wrong and we kept Systran translations. In order to improve the coverage of our translation, we added Systran translation for terms absent from

Table 5: Results at CLEF04 and CLEF05

		MUSQAT04	QALC04	MUSQAT05	QALC05
Sentences	5 first ranks	56 (31 %)	65 (37 %)	78 (41 %)	87(46 %)
NE answers	Rank 1	17	26	16	9
	5 first ranks	33	37	24	11
Non NE answers	Rank 1	7	3	16	16
	5 first ranks	12	8	22	24
Total	Rank 1	24	29	32	25
	%	12%	14.5%	17%	13%
	5 first ranks	44	45	46	39
Fusion (official results)		38 (19 %)		38 (19 %)	

the dictionary. By this way, we took off 253 bi-terms in 112 questions, and added 37 translations, with 12 bi-terms, which concerns 35 questions. The last improvement consisted in adding Systran translations that were different from Magic-dic translations (138 terms in 96 questions) to the filtered terms. This last set of terms was compound of 1311 translations for 836 terms in 200 questions (522 terms with 1 translation, 199 with 2 translations, 81 with 3 translations, 25 with 4 translations, 6 with 5 translations and 3 with 6 translations).

We tested MUSQAT with this new selection. Results are shown Table 6. We see that MUSQAT finds relevant documents for 7 supplementary questions (increase of 4%).

Table 6: MUSQAT new results

		MUSQAT
Sentences	5 first ranks	67
NE answers	Rank 1	25
	5 first ranks	41
Non NE answers	Rank 1	4
	5 first ranks	10
Total	Rank 1	29 (14,5%)
	5 first ranks	51

MUSQAT extracts 7 supplementary correct answers in the top 5 short answers, with 29 answers in rank 1. MUSQAT obtains here slightly better results than QALC with Systran translations, both for short and long answers. We also measured the number of questions for which the selection process based on FASTR indexing provides documents containing the answer pattern. In the original MUSQUAT, it was possible to find the answer for 80% of questions. Term selection allows to improve this value to 85%. These improvements are not significative enough so we had not incorporated them in this year's version, even if we think that this kind of translation validation is worth being tried. So we plan to realize bi-term validation on a larger corpus. Concerning the absence of translations, we began to increase manually our dictionary from lexicons and gazetteers we use for named entities recognition, specially for acronyms and location names, and we plan to use a bilingual aligned corpus.

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*

[CELEX 1998] CELEX, 1998, http:www.ldc.upenn.edureadme_filescelex.readme.html, UPenns, Eds., Actes Consortium for Lexical Resources, (1998)

[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*

CLEF2004,

[Peters et al. 2005] C. Peters, M. Braschler, G. Di Nunzio and N. Ferro, CLEF 2004: Ad Hoc Track Overview and Results Analysis, *Fifth Workshop of the Cross–Language Evaluation Forum (CLEF 2004),Lecture Notes in Computer Science (LNCS), Springer, Heidelberg, Germany (in print)*