Semantic knowledge in Question-Answering systems

Vincent Barbier, Brigitte Grau, Anne-Laure Ligozat, Isabelle Robba, Anne Vilnat

Semantic Knowledge in Question Answering Systems

Vincent Barbier, Brigitte Grau, Anne-Laure Ligozat, Isabelle Robba and Anne Vilnat
LIMSI-CNRS, BP 133
91403 Orsay cedex
France
e-mail : FirstName.Name@limsi.fr

Abstract

QA systems need semantic knowledge to find in documents variations of the question terms. They benefit from the use of knowledge resources such as synonym dictionaries or ontologies like WordNet. Our goal here is to study to which extent variations are needed and to determine what kinds of variations are useful or necessary for these systems. This study is based on different corpora in which we analyze semantic term variations, based on reference sets of possible variations.

1 Introduction

Most QA systems are composed of three main components. First, question analysis extracts terms from the question and finds the expected type of the answer. Then, a search engine searches the collection for documents. To this end, one or more successive requests are built with the question terms, and possibly with term variations. Finally, answers are selected following relevance criteria taking into account syntax and semantics. To find relevant documents, QA systems have to identify variations of question terms in these documents.

Our goal in this paper is to study to which extent variations are needed and to determine what kinds of variations are useful or necessary for these systems.

To this end, we present on the one hand an evaluation of our own strategy. Our QA system, working both on French and English languages, takes into account semantic variations of simple or composed terms of the question in order to cut down the set of documents retrieved by the search engine; this paper presents an evaluation of how relevant this strategy is, focusing on the French system.

On the other hand, we show to which point a QA system is able to find answers without requiring any semantic resource and to which extent results would be enhanced by such resources. This study is based on different corpora, in which we study semantic term variations. Two corpora come from the evaluation on French: one is made of all the correct answers given by the participants, the other is our set of answers. The last corpus is an automatically built corpus of correct and incorrect passages from TREC-11 questions.

After a state of the art on QA systems, we present our system FRASQUES, then we describe the studies we made on different corpora. We also describe the dedicated corpus we constituted to prevent us from the bias introduced by the use of the participants' results; finally we detail the variations present in this corpus, thanks to Wordnet ontology.

2 Semantic knowledge for selecting documents in QA systems

In order to improve document selection in QA systems, several strategies can be conceived. They consist in using semantic knowledge, present in thesauri or lexicons, at different stages of the system: i) for elaborating the query given to the search engine; ii) on the results of the search engine, for selecting the best documents or extracting small passages. QA systems generally make use of thesauri by selecting words close to question words according to semantic or lexical relations, such as synonymy, hyperonymy and hyponymy.

In the first strategy, namely query elaboration, a first problem consists in choosing the right keywords in the question; then a second problem is raised by the search of related words. Keyword selection is often based on the morpho-syntactic category of question words. They can also be weighted or considered differently in the query according to pre-established rules or to their weights in a reference corpus.

In addition to keyword selection, it can be interesting to consider their linguistic variations in order to take into account some lexical distances between questions and answer-sentences. [Moldovan *et al.*, 2003] generate morphologic, lexical and semantic variations of question keywords from WordNet ([Fellbaum, 1998]), and introduce them progressively in the queries when their system do not return enough answers. [Yang and Chua, 2002]'s system merges two kinds of knowledge sources, the Web and WordNet, for extending queries: after questioning the Web, they keep those words that are the most correlated to question words and consider them as query terms since they seem relevant in the question context. They also add some related words found in WordNet.

[Ittycheriah *et al.*, 2001] have tested different document retrieval techniques, with and without query expansion. Applying expansion mechanisms as filtering criteria for selecting answers in retrieved documents gives better results than applying them for expanding requests. We also chose this solution in our systems, FRASQUES (for French language) and QALC (for English language).

3 FRASQUES System

FRASQUES roughly follows the same principles than QALC, our English QA system, even if they slightly differ in their realisation. Both are made of four main modules, colored with gray in Figure 1.

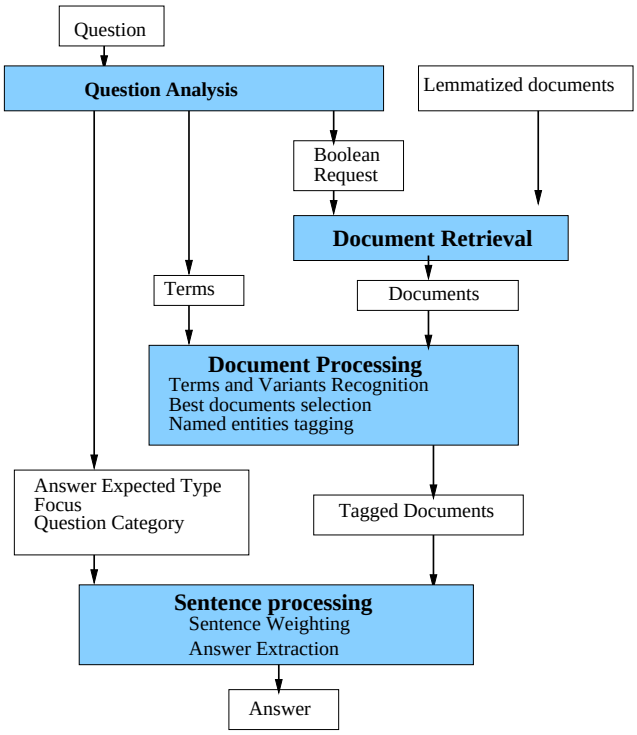

Figure 1: FRASQUES system

Question analysis proceeds in two steps. First, some information is determined such as the expected type of the answer, when this type belongs to our named entity list. Second, the lists of synonyms for non empty words of the question are built (this point is detailed in section 4.1).

The search engine, Lucene[1], is a boolean engine. To interrogate the French collection, Lucene is given a set of requests built from the non empty words of the questions. If no documents are retrieved (or a number smaller than a given threshold), the collection is searched again with fewer terms (see details section 4.3).

The documents retrieved by Lucene are re-indexed by Fastr ([Jacquemin, 1999]) in order to recognize morphological, syntactic or semantic variants of simple or composed terms of the question. These terms are weighted, and thus documents are weighted in turn. Documents are then re-ordered and a sub-set of them is considered. The named entities tagging module is then applied to these documents. The final module is in charge of extracting the answer from weighted sentences. The process differs depending on the fact that the question expects an answer which is or is not a named entity.

[1]http://jakarta.apache.org/lucene/docs/index.html

4 Analysis of the FRASQUES system

The corpora we analyze in this section come from the EQueR evaluation campaign. They are composed of the correct passages (250 characters maximum) returned by the participants, plus the results of FRASQUES.

4.1 The questions

For each question, FRASQUES question analysis module determines several kinds of information, among which three sets are more thoroughly studied in this article: i) the set of non-empty words of the question, ii) the set of their synonyms extracted from Fastr and iii) the set of their synonyms extracted from EuroWordNet.

In EQueR, the main task consisted of 500 questions. Among these, 33 did not have any Fastr synonym, and 73 did not have any EuroWordNet synonym. The average number of words per question was 5.6 while the average of Fastr synonyms was 12.8, which is relatively high, especially since among the 500 questions, there were 592 proper names, which rarely accept synonyms. The average of EuroWordNet synonyms per question was 7.1. Thus, there are nearly twice as many Fastr synonyms as EuroWordNet synonyms, which can be explained by the low coverage of EuroWordNet.

4.2 Quantitative analysis of participants correct passages

The correct passages given by the participants constitute an interesting corpus to analyze. To gauge the benefits brought by knowledge sources such as synonyms, we calculated the presence rate of synonyms in correct passages.

The corpus is composed of 2213 passages, that is an average of 4.7 passages per question (only 30 questions have not been answered). Among these passages, 82% do not contain any Fastr synonym, and 88% do not contain any EuroWordNet synonym. Only the words of the question obtain an significant rate as shown Table 1, containing the average rate of question words or synonyms per question.

Question words	60.4
Question words as Fastr synonyms	3.6
Question words as EuroWordNet synonyms	2.7

Table 1: Question words and synonyms in correct passages

Several reasons explain those rather low rates of synonyms in the corpus. First, the synonym bases are not the ones the other participants use, moreover few of them take into account such knowledge. Second, in EQueR, a lot of correct answers could be found with the words of the question. It seems (it is also true in TREC campaigns) that there is often at least one formulation close to the question, which is probably due to the large amount of documents (1.5 gigabytes).

4.3 FRASQUES answers

The set of documents returned by FRASQUES is also interesting to exploit. This corpus can be divided into two parts: the documents returned by the search engine Lucene, and the documents selected and ordered after indexation by Fastr. On

the basis of these two corpora, we investigated the influence of our use of semantic knowledge on the passage selection at the different stages of the question answering process.

When querying the search engine, we favour documents containing all words contained in the question. If no such document exists, or if there are too few of them, the constraints on the query are relaxed by omitting some of the words in the question. First, a query composed of all the non-empty words of the question if formed. If a threshold number of documents (fixed at 200), is not reached, a new query is constructed which contains the focus of the question, its main verb and its proper names. Then relaxation consists in suppressing the verb, and constructing different queries for each proper name. When we take off words, their variants may still be found in the returned documents.

For the EQueR campaign, we ran the system twice, in order to test different document selection strategies. For the first run, all proper names were used without considering the threshold of 200 documents; for the second run, we checked the number of documents after each query.

To evaluate our strategy, we listed all the various short answers, in order to have a set of admissible template answers as large as possible. We then evaluated our corpus of documents and the passages returned with respect to these answers. For each run of our system, we counted the number of occurrences of template answers in our corpus, after the first two steps of our question answering process, namely document selection by the search engine, and selection by Fastr.

The search engine returns documents containing a template answer for only 73 to 76 % of the questions. This can be explained by several factors : imprecision in the choice of the keywords of the question, which are selected only due to their morpho-syntactic tagging, errors of lemmatization, problems of anaphor and so on.

The selection of 50 documents after indexing by Fastr does not entail a decrease in the number of correct documents. The first run makes use of synonyms only for multiterms, while the second run also searches synonyms of monoterms. The second run could be expected to have a better recall, but this is not the case. This can be explained by the high degree of similarity between the questions and some of their correct sentences, and also by the noise introduced by searching "incorrect" synonyms.

Multiterm semantic variants have been found by Fastr in 40 questions (9% of the questions). These variations enabled the system to link for example the phrases "transfert d'animal" and "transport des animaux", or "avocat de M." and "défenseurs de M.". The synonyms used in these cases seem more relevant than those used for recognition of monoterm variation, which can be explained by the fact that monoterms lack a context which could enable to choose between all the possible synonyms. Multiterm variants here prove their interest, and it could be useful to favour them in other steps of the question answering process, in order to reiterate document retrieval with found synonyms.

We also carried out evaluation experiments to determine how relevant it is to use the words of the question when looking for the answer. This is summarized in Table 2. For each sentence returned by our system, we counted the number of words of the question present, as well as the number of Fastr and EuroWordNet synonyms. Then we made similar experiments where we took only in consideration those sentences which were judged correct: those have higher scores in terms of occurrences of words of the question and of synonyms, which justifies our current approach of passage selection.

	Correct passages	All passages
Words of the question	69.7%	57.3%
Fastr synonyms	4.9%	4.3%
EuroWordNet synonyms	4.0%	3.2%

Table 2: Question words and their synonyms in the extracts returned by FRASQUES

The rates of synonyms in the sentences we returned correspond to those found in the corpus of answers of all participants. Like the corpus of participants, our corpus contains a high number of sentences containing no synonym : between 78 and 80%, depending on the origin of the synonyms and of the sentences. This very low occurence rate of synonyms is probably due to the lexical proximity between the questions and the answering sentences, as it was foreseeable.

5 Extension to other variations

In order to study the reliability of more extended variations, such as those given by WordNet relations, we have built semi-automatically a corpus of answers. Questions are taken from TREC11 QA evaluation and answers are extracted from the TREC11 Aquaint collection. The corpus is composed of 123 questions and 1066 pertinent answer passages (corresponding to a paragraph or 3 sentences), which makes a mean of 8.7 passages per question.

For each question, we collected a set of pertinent and non-pertinent paragraphs. At first, passage pertinence is automatically evaluated thanks to an answer pattern, which is a regular expression. But this method is quite noisy: among the passages considered as relevant, only one of three passages is really pertinent. This method reduces dramatically the human work, but a manual validation is still required. Now, we detail the method used to constitute the corpus.

5.1 Request and Corpus Filtering

In most system, query variation is limited to synonym or morphological variations which makes difficult to study more complex variations. Our aim is to study the various possible variations of each term of a question with as little bias as possible. In this purpose we decided to collect the answers by building one specific query for each studied question term.

For each query, we omit one term of the question, which will be the studied term. This term is not represented, neither by its actual form nor by a variation. Thus a variation of this term should not be favoured over another one.

The query is a conjunction of disjunction of terms. Each disjunction of terms represents a term of the question and is a set of variations of this term obtained thanks to WordNet.

The allowed variations are synonyms, plus the most frequent word of all synsets at a distance of two WordNet rela-

tions or less. The distance of the variations is reduced in order to prevent the generated query to bring too noise, which is already important. Last, the less significant terms of the question are not used in the queries.

Named Entities are considered to be better filters than common nouns and other grammatical categories. For the other terms, the term significantness is estimated by human judgment. This human ranking permits the system to automatically build more judicious queries. For example, in order to study the variations of the term "destroy", the query will be :

Pompeii & expansion(volcano) & expansion(ancient)
with: *expansion(volcano)=(mountain|mount|crater|volcano)*

Once the documents have been fetched, pertinent and non-pertinent documents are separated according to the answer pattern. Note that thanks to this method, the pertinent and non-pertinent documents are fetched with the same query, which is once more aimed at limiting bias possibilities.The last step is the manual validation of the corpus.

5.2 Study of the Variations

We searched the kinds of variations between the terms of the question and words in the retrieved passages. A term variant exists if a path of WordNet relations links the synset containing the word of the question to a synset containing the word of the passage. The links taken into account are any combination of WorNet relations, except for glosses and morphological derivations. The passages are about 180 words long.

As to measure the frequency of kinds of variants, we aggregated them into a small number of classes. We chose to classify them according to the WordNet path length. Classes are synonyms (lemmas are different but words belong to a same synset) and words when they are distant from n relations (with n from 1 to 4).

We counted how many links each passage, either pertinent or not, contains. We only considered the less distant variation of each term of the question, if exists. Figure2 shows the average frequency of classes of links in both pertinent and non-pertinent passages, and the ratio between those two numbers.

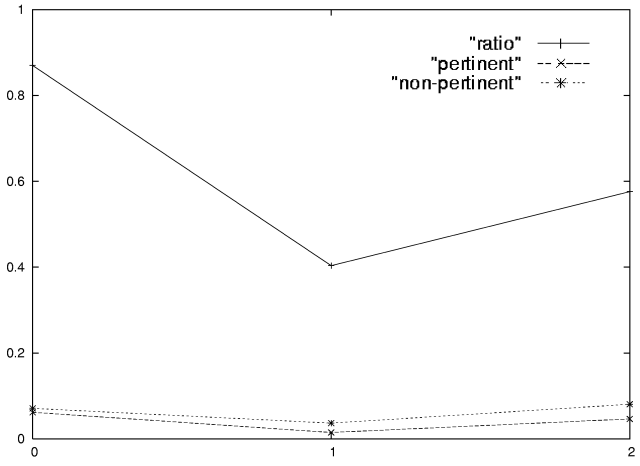

Figure 2: Precision of extended variations

We notice that the frequency of synonyms (point 0 on the x-axis in Figure2) is similar to the results we obtained in our preceding study. It appears that links compound of two relations are more frequent than 1-relation links. Moreover, they achieve a better precision. These 2-relation links often consist in hypernymy relations followed by hyponymy relations. For instance, *town*, a correct variant for *city* is given by the following path: *city* (hypernym) *municipality* (hyponym) *town*.

A question containing the word *wife* can be answered thanks to the word *husband*, because they are related through: *wife* (hypernym) *spouse* (hyponym) *husband*.

These observations show that term expansion can benefit of the use of compound relations and that variations should not be limited to synonymy or to one link hyponymy relations.

6 Conclusion

In TREC as in EQueR campaigns, systems that obtain the best results make use of semantic knowledge sources. Intuitively, one can assume that such information is necessary in open domain question answering. Nevertheless, few of these robust systems have evaluated how their results are enhanced by using this kind of information, because doing this evaluation is sometimes complex or even impossible because of the system architecture. In this paper we propose a solution which consists in exploring result corpora aiming to find what kind of knowledge was really used.

This evaluation allowed us to check that uncontrolled use of synonyms gives very few improvements in the system results; on the contrary, multiterms provide a context that makes possible the discrimination of synonyms and the selection of relevant variations.

Most importantly, this study allows us to evaluate to about 85% the rate of correct answers that may be found without quite any semantic knowledge. This rate is rather high, but not sufficient, and obtaining best results necessarily means a better use of semantic knowledge.

References

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

[Ittycheriah *et al.*, 2001] A. Ittycheriah, M. Franz, and S. Roukos. Ibm's statistical question answering system - trec-10. In *Proceedings of the Tenth Text retrieval conference*, Gaithersburg, MD, 2001. NIST.

[Jacquemin, 1999] C. Jacquemin. Syntagmatic and paradigmatic representations of term variation. In *Proceedings, ACL'99*, pages 341–348, University of Maryland, 1999.

[Moldovan *et al.*, 2003] D. Moldovan, M. Paşca, S. Harabagiu, and M. Surdeanu. Performance issues and error analysis in an open-domain question answering system. *ACM Transactions on Information Systems*, 21(2):133–154, 2003.

[Yang and Chua, 2002] H. Yang and T.-S. Chua. The integration of lexical knowledge and external resources for question answering. *Proceedings od The Eleventh Text Retrieval Conference*, 2002.