Using semantic knowledge to improve answers extraction in an information retrieval system

Tassadit AMGHAR, Styve JAUMOTTE et Bernard LEVRAT

LERIA, 2 boulevard Lavoisier, 49045 Angers cedex 01, France
{amghar, jaumotte, levrat }@info.univ-angers.fr

Abstract - We describe here a system currently in progress aiming at extracting information from a huge corpus of texts. The sought information must be relevant as answers to general questions. Our system proceeds in two steps: - first, a textual Information Retrieval System (henceforth IRS) selects a set of texts likely to contain knowledge relevant as an answer to a given question - second an extraction process determines more accurately information chunks most adapted to constitute a response to the question.

We describe here our system, which is a matter for both IRS and Question/Answering systems, focusing on information reformulation as a way to improve efficiency.

Two implemented reformulation methods are described . The first one is grounded on a categorization of the query in a previously built typology. The second one consists in the reformulation of the questions on the base of semantic information extracted from the thesaurus *Wordnet*. Three kinds of semantic knowledge are distinguished in the thesaurus each of them leading to specific reformulation methods: the definitional knowledge (which results in reformulations based on an understanding of the implied concept), the relational knowledge (which give raise to reformulations guided by existing semantic relations between implied concepts and other ones), and a document like characterization of concepts descriptions (which leads to reformulations using similarity measures between descriptions of implied concepts and others ones. Semantic distances between concepts, are calculated as similarity measures between concept 's entries of Wordnet, considered as documents,)

This paper gives a general overview of the system. Special attention is given to the questions categorization and to the reformulating process.

Keywords - Question Answering System, Reformulation, Conceptual Categorization

1. INTRODUCTION

This paper describes an ongoing work aiming at developing an information retrieval and question answering system where efficiency is increased by introducing NLP technics such as the reformulation of users 's queries. The question answering task is realized by extracting relevant information in a huge corpus[1] of general English newspapers articles.

For example, below are some parts of the documents retrieved from the corpus which could be considered as correct answers to the query *What is caldera?.* It is the result of a preliminary searching process of our system in which texts likely to contain relevant information for a given query are selected.

> AP890825-0113 / Its photos of Triton revealed that Neptune's major moon has large, inactive volcanic craters, called calderas, "filled with ice instead of rock. The lava flow is a flow of ice, not of rock," said Voyager project scientist Ed Stone.
>
> AP890827-0048 / They are huge flat craters, called calderas, similar to the one in which Yellowstone National Park rests.
>
> AP890828-0102 / Scientists announced two of them Friday: huge inactive crater-shaped volcanoes, called calderas, filled with lava made of ice instead of molten rock; and long fault-line valleys filled with oozing ice, a process likened to toothpaste coming out of a tube slit with a razor blade.
>
> AP890829-0015 / Those photographs revealed that giant craters, called calderas, once oozed an icy form of lava that flooded thousands of square miles of lunar terrain.

Our system takes the general structure of competing systems [FGIJM99, IFZRM00] to the "question answering task" of TREC and is divided into the following modules. First, in a query categorization module, the question is analysed to determine relevant information for the general task of searching an answer (what is the question about ? in other terms, what kind of entity types and which attributes of them are involved in the question). Second, using this information, a module collects relevant documents using traditional IR techniques, selecting the best parts of them, if need be. Third, acceptable answers are extracted from the texts or parts of texts, and ranked using some distance measure between found entities and query terms.

[1] 6 Gb of English News Text from various newspapers (New York Times, LA Times, Wall Street Journal, …). Our traning and evaluation corpus is extracted from the one of the *question-answering track* [VT99] proposed by TREC (Text Retrieval Conference). This task offers an opportunity to study and compare different systems aiming at extracting short answers to questions using a representative corpus of newspaper articles and related questions. Several test sets of requests (approximately 1600 questions) are available to develop the system. Each year, TREC supplies a new set of questions in order to compare systems from laboratories which attend the conference.

The main characteristic of our system is to near questions and relevant documents descriptions (index) by using reformulating processes guided by semantic information from WordNet thesaurus. More precisely, semantic information extracted from concept descriptions in Wordnet, give raise to different reformulating processes depending on its type. Three kinds of semantic knowledge are distinguished, each of them leading to particular types of rewriting methods: the definitional knowledge (dictionary like definitions of concepts), the relational knowledge (systemic characterization of concepts through lexical relations between concepts), and the document like characterization of concepts descriptions (where thesaurus terms descriptions are considered as document liable to IRS techniques).
Rewriting based on a definitional kind of knowledge uses a rather deep understanding of the definitions of concepts. Rewriting based on relational information rests mainly on semantic lexical relations between concepts. And documents like characterization of concepts leads us to rewriting where substitution of query terms is governed by using similarity measures between terms viewed as documents.

A general description of our system is given in section 2. Section 3 is devoted to the conceptual categorization of queries. Section 4 presents the different parts played by the thesaurus in the rewriting process. And finally section 5 concludes.

2. GENERAL DESCRIPTION OF THE SYSTEM

Our question-answering system does not only seeks to provide all short answers to user's questions, but also justifications of them in terms of used documents too. Moreover answers are furnished in a list rendering their relevance. Figure 1 below depicts the structure of the system.

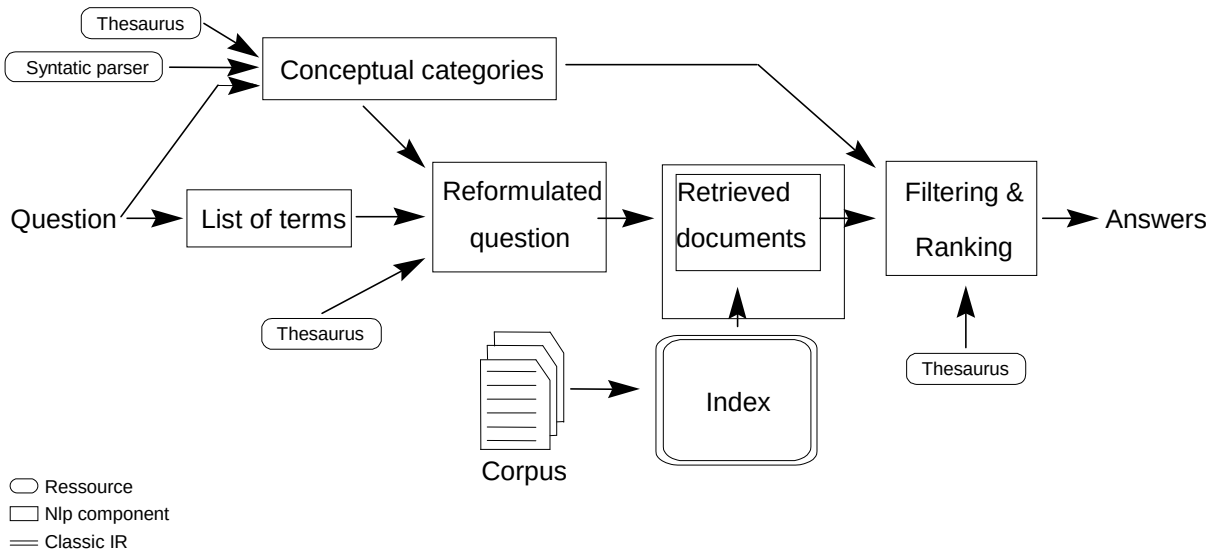

Figure 1 : general structure of the system

The answering process to a question involves the following tasks:

Conceptual analysis of the Question: This task takes as input the results of the syntactic analysis of the question by the *Link Grammar* parser [ST93]. The conceptual type of the question having been determined, the different fields of its conceptual representation are filled.

Reformulation : In this task, the question terms are reformulated using lexical relations of Wordnet [Fel98] depending on the conceptual categorization of the question

Documents Retrieval : The retrieval process, based on the extended Boolean model, selects a set of possibly relevant documents. Logical operators (***And***, ***Or***) are supplemented by a distance operator (***Dist***) which allows to focus on limited part of text.

Filtering and ranking : This process shortens the answers by selecting the best parts, ordered by relevance, of the documents of the preceding set.

The next section describes the conceptual analysis of the question.

3. CONCEPTUAL CATEGORIES OF QUERIES

Conceptual analysis aims at defining a question type in a previously defined typology of questions. Conceptual categories of questions are organized in a tree structure, built up from experiments on query test sets, to reflect hierarchical relations between categories (Figure 2). This hierarchical organization permits the system to vary the

level of the furnished answers. Note that missing categories lead the system to classify some questions as being of an ***Unknow*** type.

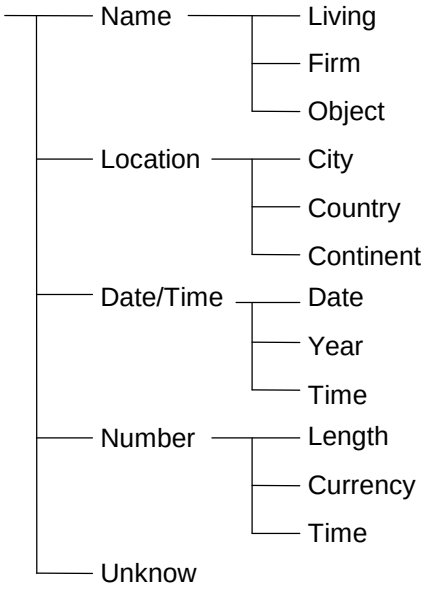

Figure 2 : Part of the categories set

Conceptual categorization of questions is divided into two steps:

1. parsing of the query giving its syntactic and constituent structures
   The syntactic analysis is built up by the Link Grammar parser [ST93]. It has been slightly modified to fit with our corpus of questions. The syntactic structure of a sentence, the link structure, is made up of different kinds of links between couples of words. Moreover, a constituent structure of the sentence is also furnished and gives access to parts of speech occurring in the parsed sentence.

2. Matching of the constituent structure of the question with a set of categorization rules.

The results of the parsing, the interrogative pronoun and the semantic of concepts of the question (for example, their lexical relations with other concepts of the thesaurus) are taken into account in the categorization rules which inform the fields associated with the conceptual category. As an example, figure 3 shows how the query *Name the creator of The Muppets* is associated with the category `Name:Living Entity`.

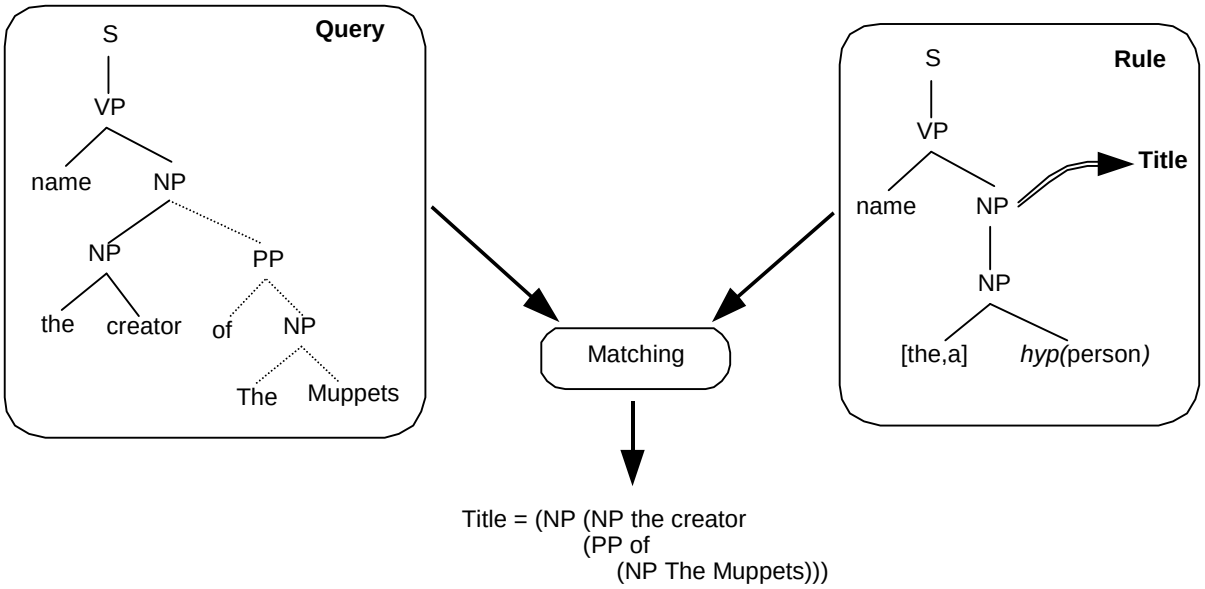

Figure 3 : An example of query categorization

Link Grammar builds a constituent structure of the sentence: `(S (VP name (NP (NP the creator) (PP of (NP The Muppets)))`. The next step of the process is to find the first rule that matches the constituent structure. A correct rule is `Name_Living(s(vp('name',title<-np( np( ['the','a'], hyp('person')))))`. Applying this rule allows the system to extract the **title** property and to associate it with the subtree `(NP (NP the creator) (PP of (NP The Muppets))`. *hyp* designs the lexical relation of hyponymy as mentioned in the thesaurus. In this example, the matching is partial. The dotted subtree of the query is not explicitly defined. It is also possible to define total matching rules. On another hand, different strategies may be used to improve the research results. For example, ordering rules by increasing level of generality results in finer analysis.
The rules application is made according to the order of appearance in the rules file. The system is trying to find an acceptable rule and if it succeeds, extracts the properties defined by the rule.
R0 = `Name_Living (s('who', s(strict(vp(conj('be'), name<-np(*)))))`
R1 = `Name_Living(s('who', s(action<-vp)))`
The first rule R0 categorizes questions of the form “Who”+be+NP as for example the query *Who is Victor Hugo?*. The rule R1 is applies to more general questions such as *“Who”* question like *Who is Tom Cruise married To?* or *Who won the Oscar for the best actor in 1970?*.

4. THREE WAYS TO MAKE USE OF THESAURUS IN THE REWRITING PROCESS

We distinguish three ways to use semantic information of the thesaurus based on the structure of concepts descriptions which are made of two components : - A definition, where a concept is described, often in an Aristotelician way, - and a systemic description which situates the concept among other ones, using lexical relations between concepts. We devote a specific treatment to the semantic information extracted from each of these components in the reformulation of query terms.

4.1. USING DEFINITIONS FOR REWRITING TERMS

The rather well defined forms of definitions in the thesaurus permits us to consider treatments where the process of semantic information extraction could be guided by their forms. It's particularly the case of aristotelician ones or definitions by a componential or functional descriptions. In those cases, the form of the definition authorizes to use filters to extract relevant semantic information to be used for the concept reformulation.

4.2. USING LEXICAL RELATIONS FOR REWRITING TERMS

As a lexical relation, we have experimented the use of the synonymy relation that is an intuitive way of query expansion. The first step consists in changing from natural form of the question to system query.
As an example, consider the natural question *When Golden Gate Bridge get finished?*. Our system will turn it into `(Dist 20 golden (Dist 20 gate (Dist 20 bridg finish )))`. The non-empty words are stemmed and sought in a window of twenty words.
The only match is the following part of text

> NYT19991025.0171 Our houses are like the Golden Gate Bridge. Once we finish everything, it'll be time to put on paint and a roof again.

This text doesn’t contain any date nor temporal references and so couldn’t constitute a correct answer. To find an answer, reformulation is needed. The thesaurus Wordnet gives four synonyms for the verb *to finish*: *to end*, *to complete*, *to terminate* and *to cease*. The verb *to finish* is replaced in the initial query by a conjunction of the synonyms and the system ends up in the query, `(Dist 20 golden (Dist 20 gate (Dist 20 bridg (Ou end (Ou complet (Ou terminate cease)))))`.
Eighteen documents are retrieved with this query. A simple filter (existence of one or more dates or temporal references near the query key words) left only six documents. To get the final answer, a syntactic analysis of the documents is needed to choice the right answer in the set of possibilities.

> NYT20000426.0233 So in 1993, with a camera attached to the front passenger window of a Ford Explorer, and following older roads like federal Highways 30, 40 and 50, he took 3,304 pictures, starting with Nos. 1 and 2 (Lower Manhattan and the Statue of Liberty, which he took standing in Jersey City, N.J.; they were his only shots on foot) and ending with the Golden Gate Bridge from the Marin Highlands.
>
> NYT19980713.0162 The opening of the Bay Bridge in 1936 and the Golden Gate Bridge in 1937 had radically diminished its importance. Ferry service from the building ended in 1958 when Southern Pacific's ``Eureka'' made its final crossing to Oakland.
>
> APW19980828.0820 Fans hope the classic trolleys become as popular with tourists as the cable cars, the Golden Gate Bridge and the city's fog. Use facts from the story to complete the following statements: 1. In the late 1800s, electric trolleys
>
> NYT20000228.0152 The Golden Gate Bridge project was also on budget and on time, completed in 1937 after about four years of laboring. The cost: $35 million, not counting the $39 million in bond interest, all financed with tolls.
>
> APW19990526.0049 On May 27, 1937, the newly completed Golden Gate Bridge connecting San Francisco and Marin County, Calif., was opened to the public.
>
> APW20000425.0198 When Zampa's Crockett-area meat market went under in 1924, a customer convinced him to give the area's burgeoning bridge-building trade a try, and Zampa went to work on what was to become the first Carquinez bridge, completed in 1927. Through the 1930s, Zampa worked on bridges in other Western states and on the Golden

Gate and Oakland-San Francisco Bay bridges. It was in 1936 that Zampa was one of 19 people who fell while making their way across a girder on the Golden Gate Bridge.

4.3. Viewing concept description as documents in IRS for rewriting terms

We discuss here a new interesting way of using thesaurus in the reformulation process. Reformulation methods used in the preceding sections, rest on a rather well defined semantic information to be used depending on the question category. But what for the cases where the nature of semantics links to be used for searching substituting terms is not or ill defined ? In this case we propose to use descriptions of concepts in the thesaurus as if they were documents subject to characterizations similar to those used in IRS [Sal83]. For that purpose, we have to define, what are index terms, the stop list, and what similarity measures have to be used between concepts descriptions and queries. We choose here to use the weighted vector model in which concepts descriptions occurring in the thesaurus and queries are represented by weighted vectors of the form $\vec{v}_d = (w_{1,d}, \ldots, w_{n,d})$, where the weight $w_{i,d}$ depends on the nature of the lexical relation linking the terms *i* and *j*. In the case of a COSIN like similarity measure, similarity between a query and a reformulated one will be rendered by the closeness of their associated vectors. The values of the weights can then be used to tune the similarity measures for choosing the best terms to be substituted to the original ones occurring in a question depending on the question category.

5. CONCLUSION

We have described here a question answering and information retrieval system. Our work comes within the scope of current techniques [BRN99] of Information Retrieval (IR). Recently, some works [BEB99] come out with segmentation methods to divide documents up into homogeneous parts as for their meaning. If such kinds of approaches somehow simplifies the use of IR they are not suited to the task of extracting accurate answers such as the date of an event, an address, a name. Others works show [GS97,RS95,Sme99] that extending IR systems with some natural language processing tools such as thesaurus and syntactic parser seems a good way to overcome difficulties of traditional statistical techniques to increase efficiency.

The different strategies of reformulation we have presented, have the main advantage of taking into account all the relations available in the thesaurus and not selecting only one type of relation: this kind of reformulation is generic and useful for *unknow* type of question which cannot be answer with a specific method.

REFERENCES

[BEB99] P. Bellot and M. El-Bèze, *"Méthodes de classification et de segmentation pour la recherche documentaire"*, 1999.

[BRN99] R. Baeza-Yates and B. Ribeiro-Neto, Modern Information Retrieval, ACM Press, 1999.

[Fel98] C. Fellbaum, WordNet, An Electronic Lexical Database, MIT Press, 1998.

[FGIJM99] O. Ferret, B. Grau, G. Illouz, C. Jacquemin, and N. Masson, *"Qalc - the question-answering program of the language and cognition group at limsi-cnrs"*, 1999.

[GS97] G. Grefenstette and F. Segond, *"Multilingual Natural Language Processing"*, International Journal of Corpus Linguistics, 1997.

[IFZRM00] A. Ittycheriah, M. Franz, W-J Zhu, A. Ratnaparkhi, and R.J. Mammone, *"Ibm's statistical question answering system"*, The Ninth Text REtrieval Conference, 2000.

[RS95] Ray Richardson and Alan F. Smeaton, *"Using WordNet in a Knowledge-Based Approach to Information Retrieval"*, 1995.

[Sal83] G. Salton, *"The SMART retrieval system"*, McGraw-Hill, 1983.

[ST93] Daniel Sleator and Davy Temperley, *"Parsing english with a link grammar"*, Third International Workshop on Parsing Technologies, 1993.

[Sme99] Alan F. Smeaton, *"Using NLP or NLP resources for information retrieval tasks"*, Natural language information retrieval, Kluwer Academic Publishers, 1999.

[VT99] E.M. Voorhees and D.M. Tice, *"The trec-8 question answering track evaluation"*, The Eighth Text REtrieval Conference, 1999.