WEBCOOP: A Cooperative Question-Answering System on the WEB

Farah Benamara
IRIT
Toulouse III University
benamara@irit.fr

Patrick Saint Dizier
IRIT
Toulouse III University
stdizier@irit.fr

Abstract

The main aim of this project is to explore, develop and evaluate the contribution of language technologies to the development of *WEBCOOP*, a system that provides intelligent Cooperative responses to Web queries. Such a system requires the integration of knowledge representation and the use of advanced reasoning procedures.

1 Introduction

The main aim of this project is to explore, develop and evaluate the contribution of language technologies to the development of *WEBCOOP*, a system that provides intelligent cooperative responses to Web queries. Besides a heavy use of language (processing queries, generating responses, extracting knowledge from web pages), such a system requires the integration of knowledge representation and the use of advanced reasoning procedures. Moreover, the complexity of reasoning procedures must be kept reasonable in order to optimise tractability, efficiency, and re-usability.

In a first stage, the project is developed on a relatively limited domain that includes a number of aspects of tourism (accommodation and transportation, which have very different characteristics on the web). In a second stage, we want to evaluate the re-usability of our techniques to other domains analysing where are the difficulties, what are the costs, what is domain specific and what can be shared.

A number of cooperative systems were designed for databases in the past such as *COBASE* (Minock and Chu, 96) and *CARMIN* (Chakravarthy et al, 90). Most of the efforts were concentrated on fundamental reasoning procedures, while very little attention was paid to question analysis and to NL response generation. Our challenge is to integrate reasoning procedures with real-life data extracted from web pages and to produce web style cooperative NL responses of a reasonable quality that reflect the accuracy of the reasoning procedures. A major feature is the integration of a real cooperative *know-how* component that goes beyond the mere recognition of a user misconception.

In the following sections, we briefly present the main aspects of our system focussing on question classification and knowledge representation. Cooperative response elaboration and response generation are presented in (Benamara and Saint-Dizier, 03).

2 What is a Cooperative Response ?

Cooperative answering systems are typically able to provide general, descriptive answers along with explanations about their answers. (Grice, 75) maxims of conversation namely the *quality, quantity, relation* and *style* maxims are frequently used as a basis for designing cooperative answering systems. To address these behaviours, specific cooperative techniques have been developed to identify and to explain false presuppositions or various types of misunderstandings found in questions. Relaxation of constraints in the question

occur when the system cannot find any response. Finally, intentional responses may be provided instead of a large set of extensional answers. An overview of these aspects is given in (Gaasterland et al, 94).

3 Cooperative Responses in WEBCOOP

In WEBCOOP, responses provided to users are built in web style, by integrating natural language generation (NLG) techniques with hypertext links to produce ”dynamic” responses. Responses are structured in two parts. The first part contains explanation elements which report user misconceptions in relation with the domain knowledge. The second part is the most important and the most original. It reflects the *'know-how'* of the cooperative system, going beyond the cooperative statements given in part one. It is based on several components: dedicated cooperative rules possibly using knowledge extracted from web pages, relaxation strategies and the domain ontology. The *know-how* component also allows for the dynamic determination of those text fragments to be defined as hypertext links, from which the user can get more information.

Let us consider a simple example. Suppose one wishes to rent a country cottage in Midi Pyrénées region by the seashore. Part 1 reports a misconception. This entails the production of the following message, generated from a logical formula:
Midi Pyrénées region is not by the seashore.
In part 2, the *know-how* of the cooperative system generates the possible flexible solutions below. Dynamically created links are underlined :
we propose you country cottages in
- *another region in France by the seashore*
- *Midi Pyrénées region.*

The formal aspects of the content determination and the dynamic generation of cooperative responses are presented in (Benamara and Saint Dizier, 03). In the next sections, we first characterize the typology of natural language questions that WEBCOOP takes in and then specify the knowledge extraction procedures from web pages.

4 Question classification and processing

In our system, we offer two modes for querying a web page: either via keywords, which are then interpreted as a simplified NL query, or in natural language. One of our first tasks is then to elaborate a generic classification of the different types of questions. Our taxonomy is inspired from (Lehnert, 78) (QUALM system) and (Graesser and Gordon, 91). It is therefore quite different from taxonomies dedicated to open domain question answering, for example within the TREC-8 and the TREC-9 programs (Hovy et al, 00) which are mainly based on question templates.

Our taxonomy has been constructed and evaluated using a corpus elaborated from the Frequently Asked Questions (FAQ) section of a number of web services, among which services dedicated to tourism. This corpus also defines a small but representative subset of question types that characterize the different forms of cooperative responses we are aiming at in WEBCOOP. W.r.t. to this corpus, questions are classified according to their expected responses:

- *atomic or enumerative* responses : boolean (yes, no), enumeration of entities, quantity (number, time, etc.) or quality expressed by evaluative adjectives. *Do I need a visa to go from France to Spain?*, induces a boolean response, and *What are the rates of the Royal Hotel in Paris ?* induces a response of type quantity.

- *narrative* responses, based on the following conceptual categories : procedure, definition, description, cause, goal, evaluation and comparison. Examples : *I want some information about Paris underground* induces a response of type description, and *What is the difference between a country cottage and a chalet?*, induces a comparison.

In an orthogonal way to the above fundamental typology, additional phenomena have been observed such as:

- questions including fuzzy terms (essentially evaluative adjectives like *cheap* or *close to*), as in: *a cheap country cottage close to the seaside in Côte d'Azur*.

- incomplete questions where essential elements are missing, such as *What are the flights to Toulouse ?* or, questions for which only portions

can be processed.

- questions based on a series of examples, such as *I am looking for country cottages in the mountain similar to Mr. Dupond cottage*.

Since the WEBCOOP project is in an early stage of development, we mainly focus on cooperativity for atomic or enumerative responses possibly including fuzzy expressions. This allows for the evaluation of the expressivity of our formalism (see section 5) as well as for the complexity of the reasoning procedures and the NLG needs.

We have designed a bottom-up parser that produces a conceptual representation of questions. Our strategy is to keep track of the terms used in the question as much as possible in order to re-use them in the response. For example, the question *Give me the Royal Hotel rates in Paris?* has the following semantic representation :
$(\, Quantity,\ X : listof(rates),\ hotel(royal) \wedge in(place, royal, paris) \ \wedge \ rates(royal, X)\ )$.

5 Knowledge Representation

5.1 General principles

Our approach requires two, very classical, levels of knowledge representation (Benamara 02): general purpose knowledge, specified by hand, and domain knowledge, acquired via knowledge extraction procedures from web pages. A uniform logical representation is used, based on a simplified version of the Lexical Conceptual Structure (LCS), (Jackendoff, 90). The use of the LCS language and the power of its primitive systems, allows us to have relatively generic representations, well-adapted for reasoning procedures.

The general purpose knowledge base describes knowledge about the tourism domain. It contains a domain ontology, a number of basic information (e.g. country names, airlines), rules and integrity constraints. It is not very big and can therefore be reliably specified by hand (e.g. about 60 rules and 50 integrity constraints for accommodation). It is clear, however, that to extend the knowledge base to other domains, semi-automatic procedures are necessary. We are currently exploring linguistic-based methods to extract knowledge that expresses causes, consequences and conditionals of various forms.

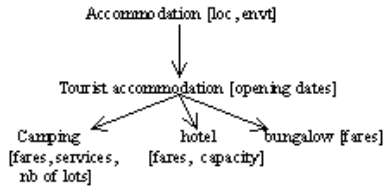
Most of the knowledge is extracted from web pages dedicated to tourism. We developed in the past 3 years (under the PRETI project at IRIT [1]) a method and an algorithm that extracts knowledge from web pages based on the domain ontology. The basic principle is that each major node of the ontology is associated with a dedicated local grammar that recognizes, in the textual part of a web page, information relevant to the concept associated with that node. The parser dynamically constructs a semantic representation under the form of a frame with attribute-value pairs. Values can be atomic or fragments of LCS. A detailed analysis of our corpus shows the prominent role played by prepositions to describe e.g. localisation, instrument or purpose. Therefore, a great attention is devoted to the extraction of predicative forms.

As a result, knowledge extracted from a web page is the concatenation of frame fragments corresponding to the nodes activated in the ontology. Each frame fragment is linked to its original web textual fragment. We also keep track of the web page structure, since this is often useful for response construction (to have access to comments, lists of items or procedures, etc.). We then have a database of web pages indexed by means of frames. For our experiments, we work with a database established once for all, but we foresee to have it updated regularly, since web pages change frequently.

5.2 An Example

Let us now consider an example that precisely describes the principles of our approach. We consider an ontological description of the application domain where nodes are concepts, decorated by means of attributes which describe concept properties or any other relevant information. Since descriptions are hierarchically organized, properties are inherited, if there is no conflict. Local grammars are associated with those properties. They may be shared by several sister nodes, but in general there are some differences. For example, the expression of rates for a camping site is quite different from the expression of rates for a hotel or a

[1] http://www.irit.fr/projets/PRETI.html



bungalow. The figure above describes the accommodation ontology.

Grammars associated with properties are designed (1) to extract the information judged relevant for the property at stake and (2) to contribute to the construction of the semantic representation. In our experiment, grammars are written by hand, from corpus samples. We have adopted the discontinuous grammars formalism (Saint-Dizier, 88), a DCG-type grammar which includes gaps which are variables in the rule that stand for a finite set of words to skip till a certain word or condition is met. These rules run in a bottom-up fashion, allowing for the recognition of text fragments distributed throughout a whole paragraph or web page. A careful organization of rules make efficiency quite acceptable. In our application, 20 different local grammars have been developed, with a total of 65 extraction grammar rules.

For example, a grammar rule that deals with the environment has the following form:

```
envt([at(place,X,Y)]) --> prep(fixed-loc),
    gap, lex(Y,noun,Sem_Type)
   {subsume(phys-loc,Sem_Type)}.
```

prep(fixed-loc) is any preposition that describes a fixed localization (at, on, near, etc.). The gap allows the parser to skip any irrelevant string till a word denoting a physical location is found.

A web page is therefore 'indexed' using a set of predicative forms. The knowledge extractor car either extract all the information it can reliably found or it can just extract information related to pre-selected properties (something like views). The result is a set of predicative forms (or possibly just words when there is no predicate). Predicative forms are marked by means of XML tags which also appear in the original text in order to keep track of the information source.

6 Conclusion

Implementation of this project is about half-way. Evaluation is crucial on two dimensions : the quality of the services offered to a user and the reusability for other domains namely : where are the difficulties, what are the costs, what is domain specific and what can be shared.

References

Benamara F and Saint Dizier P. 2003. *Dynamic Generation of Cooperative Natural Language Responses in WEBCOOP*. Ninth European workshop on Natural Language Generation. EACL, Budapest, Hungry.

Benamara F. 2002. *A Semantic Representation Formalism for Cooperative Question Answering Systems*. Proceeding of Knowledge Base Computer Systems (KBCS), Mumbai, India, dec.

Chakravarthy U, Grant J, and Minker J. 1990. *Logic-Based Approach to Semantic Query Optimisation*, ACM Transactions on Database Systems, 15(2):162-207, 1990.

Gaasterland T, Godfrey P, Minker J. 1994. *An Overview of Cooperative Answering*. Papers in Non-standard Queries and Non-standard Answers, in series Studies in Logic and Computation, Clarendon Press, Oxford.

Graesser A, Gordon S. 1991. *Question-Answering and the Organization of the World Knowledge*, In W. Kessen, A. Ortony, and F. Craik (Eds.), Memories, thoughts, and emotions: Essays in honor of George Mandler. Hillsdale, NJ: Erlbaum.

Grice H. 1990. *Logic and Conversation*, In Cole and Morgan editors, Syntax and Semantic, Academic Press.

Hovy E, Gerber L, Hermjakob U, Junk M and Lin C. 2000. *Question Answering in Webclopedia*, in Proceedings of the TREC-9 Conference, NIST. Gaithersburg, MD.

Jackendoff R. 1990. *Semantic Structures*, MIT Press, Cambridge M.A.

Lehnert W. 1978. *The Process of Question Answering: a Computer Simulation of Cognition*, Lawrence Erlbaum.

Minock M, Chu W. 1996. *Explanation for Cooperative Information Systems*, International Symposium on Methodologies for Intelligent Systems, 264-273.

Saint-Dizier P. 1988. *Contextual Discontinuous Grammars*, in Natural Language Understanding and Logic Programming II, V. Dahl and P. Saint-Dizier (eds.), North Holland.