NATURAL LANGUAGE GUIDED DIALOGUES FOR ACCESSING THE WEB

Marta Gatius and Horacio Rodríguez

Software Department, Technical University of Catalonia

Abstract. This paper proposes the use of ontologies representing domain and linguistic knowledge for guiding natural language (NL) communication on the Web contents. This proposal deals with the problem of accessing and processing the Web data required to answer user consults. Concepts and communication acts are represented in the conceptual ontology (CO). Domain-restricted grammars and lexicons are obtained automatically by adapting the general linguistic knowledge to cover the communication acts for a particular domain. The use of domain-restricted grammars and lexicons has proved to be efficient especially when the user is guided in introducing the NL queries. Once the query has been processed, the system fires the appropriate wrappers to extract the data from the Web. The domain concepts described in the CO provides a unifying framework to represent the knowledge obtained from the various Web sources. Following this proposal, a dialogue-system for accessing in Spanish to a set of Web sites on the travelling domain has been implemented.

1 Introduction

The Web is a huge repository of text, images and services. Though the Web was designed primarily for human use, a user faces different problems when accessing a specific Web site: locating the relevant Web sites, accessing different protocols and facilities, executing services, etc. The problem of locating Web sites where useful information is placed has been widely addressed and a lot of browsers, meta-browsers and information agents have been built (see for instance [15]). However, there are not many systems supporting friendly and intelligent access to the Web contents. Existing NL interfaces (NLIs) accessing different types of applications (i.e. databases), cannot easily be adapted to support communication on Web contents. The main reason is that Web sources are not designed to be processed automatically, they are heterogeneous and change rapidly.

Although the NLI systems accessing the Web gather information from various Web sources, this problem differs from the Information Extraction ([9], [17]), Information Integration ([11], [15]) and the Question Answering ([5], [19]) paradigms. In NLI systems, tasks are well defined and users can be guided to express their information needs. Additionally, many NLI systems use domain knowledge and domain reasoning to respond user requests in a intelligent manner ([16]).

This paper presents GIWeb, a NLI system supporting communication on the contents of a collection of domain-restricted Web sites. To achieve an efficient communication the system uses a CO representing domain concepts and communication tasks and a linguistic ontology (LO) representing general linguistic knowledge. The system generates domain-restricted linguistic resources by adapting the general linguistic knowledge in the LO to cover the communication tasks for a particular domain. The use of domain-restricted grammars and lexicon has proved to be efficient especially when the user is guided about the system conceptual/linguistic coverage. GIWeb guides the user by showing in the screen the NL options acceptable at each state of the communication. The system is capable of responding properly to a variety of requests involving knowledge in a collection of domain-restricted Web sites. Once the query has been processed, the dialogue component controls the obtaining and processing of the Web contents required to answer.

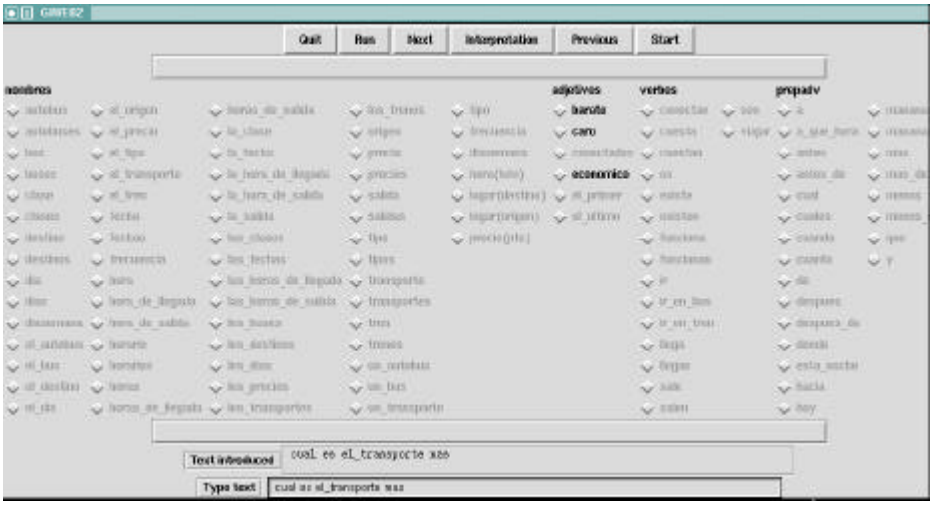

Fig. 1. Guiding the user to introduce the NL sentences

An overview of the system is given in Section 2. The Section 3 describes the obtaining of domain-restricted grammars and lexicons. Section 4 describes the flow of communication. Finally, specific details of the representation of the Web contents in the CO are given in Section 5.

2 An Overview Of The System

The NL components of the system were adapted from those of GISE ([2], [3]), a system using a CO to support NL communication with Knowledge Based Systems (KBSs). The GISE components were adapted to provide NL access to Web contents. The tasks of communication in NL consulting systems mainly consists of operations consulting particular knowledge on the domain. In those systems, user interventions cover a rich variety of linguistic phenomena. The system must support direct, concise and ungrammatical utterances to achieve a natural interaction. Additionally, tackling

consults on Web contents involves accessing and processing the data in various Web sources.

In the system GIWeb, the knowledge involved in NL communication is represented in separate, reusable knowledge bases: the CO, representing the conceptual knowledge, the Linguistic Ontology (LO), representing general linguistic knowledge, and a set of control rules (CR), generating the domain-restricted grammars and lexicons by adapting the general linguistic resources to those required for a specific domain. A wrapper system was incorporated for accessing the information in the Web.

CO is the skeleton for anchoring the description of the concepts of a particular domain. The CO is organized in three independent taxonomies, representing domain concepts, attributes describing these concepts and operations to be performed on the domain concepts. The description of the domain concepts in the CO provides a framework for integrating the information from several Web sources.

The attributes describing concepts were classified according to a semantic-syntactic taxonomy in order to favor the generation of the domain-restricted grammar. The basic classes are associated with the different grammar roles in the sublanguage considered. The syntactic-semantic classification of attributes allows a variety of different linguistic coverage for each attribute class. Current implementation uses 19 basic classes. Although the attribute classification is based on Spanish linguistic distinctions it is easily portable to other languages.

The taxonomy of operations describes the communication tasks. Operations are classified as simple or complex. Simple operations involve one conceptual instance. Complex operations involve several instances. Complex operations provides inferential and reasoning capabilities to answer complex questions. If specialized tasks for a domain are required, they have to be incorporated into the taxonomy of operations.

The general linguistic knowledge needed to cover the expression in Spanish of the operations the system performs is represented in the LO. Following the Nigel grammar ([20]), the linguistic knowledge was organized in three main classes: the class *clause* (having a subject and a finite verb), the class *group* (having a head and a variable number of modifiers) and the class *word* (representing verbs, nouns, articles, etc.). Objects representing linguistic classes are assumed to be common to all domains. Objects representing the specific aspects of the information to be expressed for each domain are represented as instances of the linguistic classes. In current implementation, there are 130 subclasses of the class *clause*, 53 subclasses of the class *group* and 93 subclasses of the class *word*.

The control information to obtain the linguistic structures necessary for a particular domain is represented by the CR. Rules are of the form: *conditions --> actions*. Conditions basically consist of descriptions of objects. Rules are applied over objects in the CO and the LO satisfying required descriptions. The actions performed by the rules are operations consulting and modifying the objects in the CO and LO. Rules are grouped into rulesets. Each ruleset performs a different action and each rule in the ruleset considers a different type of object. The basic set of control rules consists of 46 rules grouped into 9 rulesets.

The information from the dynamic heterogeneous sources in the Web is obtained by wrappers. In the Web environment, a wrapper can be defined as a processor that

converts information stored as in a HTML document into information explicitly stored as a data structure for further processed. The primary advantage of building wrappers to extract information from the Web is that they provided an integrated access to several sources. GIWeb incorporates a wrapper system providing a special language for describing Web pages. Although a description must be given for each page organization, frequently there are Web pages sharing a common organization, such as those generated by a Web service. GIWeb uses two families of wrappers to extract the data from the Web pages and represent it in the CO.

The current implementation of the system has been applied to provide access to several Web sites containing information on trains and buses.

3 Obtaining Domain-Restricted Grammars And Lexicon

Obtaining the domain-restricted linguistic resources consists of adapting the CO and the LO to a specific domain. Representing the domain knowledge in the CO consists of describing the concepts involved in the communication for a particular domain as subclasses of the general concepts. Each domain concept is described by an identifier, a primitive relation (*isa*) relating it to the taxonomy of concepts and a set of attributes. All attributes describing concepts have to be incorporated into the taxonomy of attributes. Concepts and attributes appearing in the communication are linked to one or more lexical entries in the domain lexicon. These lexical entries include all the forms associated with the expressions of the concepts and attributes in the operations (names, verbs, adjectives). The addresses of the Web sites containing information about a concept are also included in the concept description.

Once domain knowledge have been incorporated, the CR generates the grammar and lexicon representing the operations on the domain concepts. The process is performed in three steps:

1. Generation of the instances of the CO operations for the domain concepts. Different operations are generated considering the classes of the attributes.
2. Generation of the LO instances supporting the expression of the operations generated in the first step.
3. Representation of the LO instances created in the second step as DCG rules and lexical entries.

Because most of the CO operations are based on the conceptual attributes the linguistic structures are obtained considering the syntactic-semantic classes of the domain concept attributes. To illustrate this process, we will consider the CO operation *minimum_attribute_value_o*, obtaining the conceptual instance having the minimum value of a specific attribute. This operation is based on the attributes in the class *of_quantity*, expressing a quantity (and associated with a unit). The expression of this operation depends on the attribute subclasses. For example, in the travelling domain three attributes belonging to subclasses of the *of_quantity* class were used to describe the concept *transport*: *price*, *arrivaltime* and *departuretime*. The attribute *price* belongs to the class *of_cost*. The attributes *arrivaltime* and *departuretime* belong to the class *of_time*. For the attributes in the class *of_cost*, the operation could be

expressed using the form: *¿Cuál es <concept name> más económico?*(*Which is the cheapest <concept name>?*)

If the attribute belongs to the class *of_time*, the patterns to express this operation would be:

¿Cuál es el primer <concept name>? (*Which is the first <concept name>?*) and

¿Qué <concept name><attribute verb> antes? (*Which <concept name> <attribute verb> first?*)

In case of the concept name *tren* (*the train*) and the attribute verb *salir* (*departure*) the resulting question would be:

Cuál es el primer tren? (*Which is the first tren?*) and *Qué tren sale antes?* (*Which train leaves first?*)

In the grammars and lexicons generated, categories are augmented with syntactic and semantic features. Rules and lexical entries are associated with semantic interpretation based on lambda calculus. Conceptual knowledge from the CO is also incorporated into rules and entries to facilitate the processing of using interventions. The semantic features associated with the categories correspond to identifiers of concepts and attributes. The semantic interpretation associated with the lexical entries consist of lambda values and functions representing CO operations, concepts, attributes and values. Each grammar rule expressing an operation include the operation identifier and its preconditions.

4 The Dialogue System

GIWeb guides the user in introducing the NL utterances by showing the NL options acceptable in the screen. The user can type the complete query or, alternatively, build a sentence by selecting the active options in the screen. As can be observed in Figure 1, only the NL options the system can recognize at each state of the communication are active. Once the user has selected an option, it is passed to an incremental parser. When the parser returns all items that can be recognized in the next step, the interface updates the NL options that must be active. Once a whole sentence has been recognized and interpreted by the parse, it is passed to the dialogue controller (DC).

The information passed to the DC consists of a set of possible semantic interpretations. Each interpretation includes the operation identifier, the concept identifier and the operation parameters expressed in the user intervention. If necessary, the DC completes this information using history of the dialogue and the conceptual knowledge represented in the CO. A simple attentional structure is used to record the focus of attention. The concept over which an operation is performed and the rest of parameters expressed are considered the focus of attention.

The DC consults the definition of the operation in the CO for obtaining its mandatory arguments, the default values of these arguments and other related information such as the format of specific values (i.e. those representing dates and quantities). The values of the mandatory arguments not expressed in the user intervention are obtained from the focus of previous sentences. If these values are not expressed in previous sentences, the default values are used. The DC attempt to

disambiguate ambiguous semantic interpretations by considering the operation definition and the context. In case there is more than one correct interpretation then the one referring to the focus of attention of previous sentences is selected.

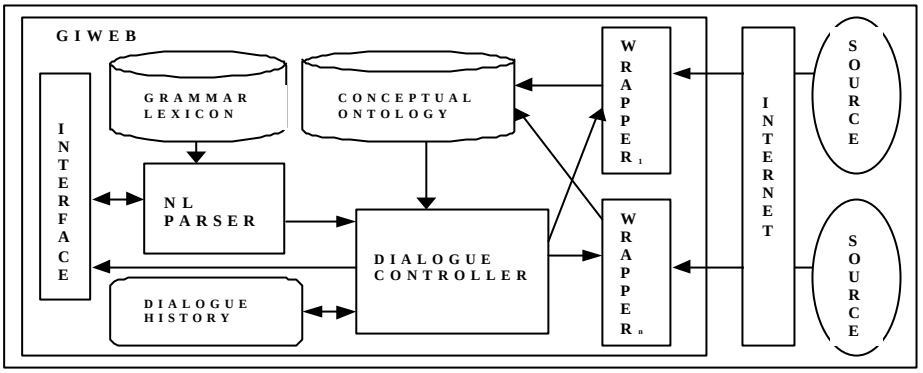

Fig. 2. The components of the dialogue system

In the current implementation, the DC has been designed with the assumption that the user would introduce sentences by using the NL options the system display. Assuming the options introduced are those acceptable to the grammar generated for a specific domain simplifies the DC. Dealing with the sentences introduced by the user without any help would require an increase in the DC complexity. In this case, users can introduce sentences that do not express correct operations, or even ones that do not express any operation at all. To deal with these problems, the functionality of the DC would necessarily have to include new tasks, such as those reformulating or confirming user interventions, opening or/and closing the dialogue, etc.

Once a complete operation is obtained it is executed over the instances of the domain concepts represented in the CO. If no satisfactory answer is obtained, then the DC is in charge of activating the corresponding wrappers to extract the information from the Web. The addresses of the Web pages containing particular information about a concept are obtained from the concept description. An address can also represent a request to a URL Web service. In this case, the parameters required for the service must be specified. Each Web address is associated with the description of the HTLM source and the class of the wrapper to obtain the information. The DC calls the corresponding wrapper classes and passes them the Web addresses, the page descriptions and, in case of requests to Web services, the information the services require. The wrappers represent the information extracted from the Web as instances of the CO domain concepts. Then, the user consult is executed again over the CO. Finally, the answer is passed to the interface. If there is no answer, the DC sends the corresponding message to the interface.

5 Obtaining The Information From The Web

Currently, only semi-structured and structured Web pages have been considered. The information in these pages is usually represented as lists of attributes delimited by

HTML-tags. Extraction patterns for those pages are often based on tokens and delimiters, such HTML-tags. The Web pages are accessed by wrappers. Several approaches are being proposed to reduce the cost of implementing a specific wrapper for each Web source: special languages for writing wrappers ([6]), semi-automated tools ([1]), wrappers generation ([7], [8], [13], [14], [18]).

For GIWeb, we have designed and implemented a simple wrapper system allowing an easy interaction with the CO. This system uses an explicit description of the HTML source to be analyzed. When adapting the system to a domain the set of Web sources that would be consulted during communication are selected and described. The description of a Web page consists of three parts: describing the organization of the page, describing the textual processing to be done over the data extracted and describing how the resulting data must be represented in the CO. The first part describes the tags delimiting the tuples and the attributes in the tuples. This description includes information about possible nested structures (attributes represented as tuples) and about the different types of information stored in an attribute (text, internet addresses, images or codes). In the second part of the page description, the textual processing required is indicated using a set of predefined types: *text*, *integer*, *brackets*, *list*, *time*, *data*, *weekday*, ... There is a default presentation for these types. For example, by default the text will be written in lower case letter, without accents and without spaces. The third part of a Web page description contains the information necessary to represent the data extracted as instances of a particular domain concept in the CO. If there is more than one concept described in a page, a different page description will be used to obtain the information related to each concept. The description of the page must indicate the name of the concept described as well as the correspondence between the attributes in the page and the attributes describing each conceptual instance. Information about a particular instance can appear in more than one page. For example, the departure time and arrival time of a particular train can be in one page and the train stops in a different one.

6 Conclusions

In this paper we have presented a NL dialogue system for accessing the Web. The main issue in the system design is the reusable and efficient representation of the conceptual and linguistic knowledge involved in communication. The organization proposed favors the obtaining of domain-restricted grammars and lexicon. The use of domain-restricted linguistic resources and guiding the user about the system conceptual/linguistic coverage improves the communication process. The taxonomy of concepts in the CO provides a unifying framework for integrating information from different Web sources. The taxonomy of operations allows the system to answer complex consults.

The modular organization of the relevant knowledge into separate data structures provides great flexibility for adapting the system to different domains and languages. Furthermore, the proposed architecture could be applied to other types of dialogue systems, such as those providing access to e-commerce applications.

References

1. N. Ashish and C. A. Knoblock, 'Wrapper generation for semistructured Internet sources', *Proceedings of the ACM SIGMOD Workshop on Managment of Semi-structured Data*, (1997).
2. Marta Gatius and Horacio Rodríguez. Adapting general linguistic knowledge to applications in order to obtain friendly and efficient NL interfaces. In the *Proceedings of the VEXTAL Conference.* Venezia. (1999).
3. Marta Gatius. Using an ontology for guiding natural language interaction with knowledge based systems. Ph.D. thesis, Technical University of Catalonia. 2001
4. J. A. Bateman, R. T. Kasper, J. D. Moore, and R. A. Whitney. A General Organization of Knowledge for Natural Language Processing: the Penman Upper Model. Technical report. USC/Information Sciences. Institute, 1990.
5. C. Cardie, V. Ng, D. Pierce, and C. Buckley, 'Examining the Role of Statistical and Linguistic Knowledge Sources in a General-Knowledge Question-Answering System', *Proceedings of the Sixth Applied Natural Language Processing Conference*, (2000).
6. W. Cohen, 'Recognizing Structure in Web Pages using Similarity Queries', *Proceeding of the AAAI*, 59–66, (1999).
7. W. Cohen, 'Whirl: A word-based information representation language', *Artificial Intelligence*, 118, 163–196, (2000).
8. W. Cohen and L.S. Jensen, 'A structured wrapper induction system for extracting information from semi-structured documents', *Proceedings of IJCAI Workshop on Adaptive Text Extraction and Mining*, (2001).
9. L. Eikvil. Information Extraction from World Wide Web - A Survey. Report 945, 1999. Available at:
   http://www.nr.no/documents/samba/research areas/BAMG/Publications/webIE rep945.ps
10. H. Garcia-Molina, D. Papakonstantinou, A. Quass, Y. Rajaraman, Y. Sagiv, J. Ullman, V. Vassalos, and J. Widom, 'The TSIMMIS approach to mediation: Data models and languages', *The journal of Intelligent Information Systems*, (1997).
11. M.A. Hearst, 'Information Integration Trends and Controversies', *Column IEEE Intelligent Systems*, 13,5, 17–20, (1998).
12. C. Hsu and M. Dung, 'Generating finite-state transducers for semistructured data extraction from the WEB', *Journal of Information Systems*, 23,8, 521–538, (1998).
13. M. Kobayashi and K. Takeda, 'Information Retrieval on the Web', *Computing Surveys*, 32, (2000).
14. N. Kushmerick, 'Wrapper induction: Efficiency and expressiveness', *Artificial Intelligence*, 118, 15–68, (2000).
15. A.Y. Levy, 'Combining Artificial Intelligence and Databases for Data Integration', *Special issue of LNAI:Artificial Intelligence Today; RecentTrends and Developments*, (1999).
16. M. Maybury, *Intelligent Multimedia Interfaces*, AAAI Press & Cambridge MA: The MIT Press, Menlo Park, CA, 1993.
17. I. Muslea, 'Extraction Patterns for Information Extraction Tasks: A Survey', *Proceedings of the AAAI Workshop on Machine Learning for Information Extraction*, (1999).
18. I. Muslea, S. Minton, and C. Knoblock, 'Hierarchical Wrapper Induction for Semistructured, Web-based Information Sources', *Proceedings of the Conference on Automated Learning and Discovery(CONALD)*, (1999).
19. E. Voorhees. Overview of the TRECK 2001 Question Answering Track. Presentation to the Text REtrieval Conference. Gaithersburg, USA, 2001.
20. The Penman NL Generation Group. *The Nigel Manual*. Information Sciences Institute of the University of Southern California. Draft, 1988.

TSD 2002, LNAI 2448, pp.373-380,2002.