Cross-lingual retrieval in Semantic Web

Cristina VERTAN
Universität Hamburg • Arbeitsbereich
Natürlichsprachliche Systeme
Vogt-Kölln Straße 30
Hamburg, Germany, 22527
vertan@informatik.uni-hamburg.de

Abstract

Natural Language is considered the friendliest way of man-machine communication. However the implementation of natural language interfaces faces often the problem of lack of linguistic and world-knowledge, especially when the application domain is not very specific. This is exactly the case of Web-based applications, which aim to serve for retrieval of information in every-day areas of work. The recent Semantic Web activities had as consequence the development of large ontologies for a broad spectrum of domains, as well as of mechanisms for annotating the resources with semantic information.
In this paper we present a new architecture aiming to bring together the advantages of natural language querying and the power of semantic Web. We will show also how described application can be easily adapted for other domains.

1 Natural Language Interfaces in WWW

Natural Language Interfaces were first used as means for querying databases [Androutsopoulos&Aretoulaki03]. The main idea was that, for an user with no deep computer science knowledge it is easier to query the database in natural language instead of using SQL expressions. Moreover natural language expressions are often shorter as SQL ones, and there are cases when it is difficult to formalise expressions like "some", "a few", "often" etc. Although these are remarkable advantages, it turned out that the analysis and understanding of natural language input is a high complex process, requiring linguistic knowledge (morphology, syntax, semantics, pragmatics) as well as a well elaborated knowledge-base and a very complex dataflow control between the components [DaleMoisl&Somers00]. Another successful approach is to use large sets of existent data (corpora) and extract features and statistical information about the behaviour of specific structures. These so-called empirical approaches turned out to be extremely successful, although there is no linguistic theory which sustain them. However they require the existence of training-data, which has to be often annotated for the purpose of the application.

The World Wide Web (WWW) can be seen as an enormous database of heterogeneous resources which is growing continously. Query and Information retrieval is one of the central issues in WWW. We should also observe that in the absence of a formal language as SQL for databases, natural language remains the only way for querying the web. On the other hand it is very difficult to deal with the large number of languages and the heterogeneous domains of resources. Therefore most of the Internet query tools allow as input keywords, sometimes connected with logical operators. There are at least two consequences of this restriction:

- the user who is actually concentrated on his search topic, must try to synthethize his query in this logical form, and find operators which fit to his scope.
- Even with this logical operators, in the absence of a semantic representation of the query, and in parallel of the existent resources, the retrieved information will be partially out of the scope of the query.

The Semantic Web activities aim to give a solution to the latter point. As for the first, the only possibility to get out of the paradigma: "keywords"+"logical operators" is the use of natural language. Is it however difficult to control also the complete syntax, and the level of language knowledge of the user. Most part of the Web users are non-native English speakers, but they are using English as query language. On the other hand any rule-based approach in natural language analysis will make first a syntactic analysis, and even very robust (i.e. fault-tolerant) grammars fail to certain grammatical errors. From this point of view, the empirical corpus based approach would be much more suited, but here arise again the problem of lack of data. The syntax analysis needs, when using empirical methods, tree-banks for the analysed language. First of all, these tree banks are available for a reduced number of language, secondly, they are not usually access free.

Taking into account the above described problems, the only viable solution seems to be the use of a controlled language input which still offers

the user the power of natural language, but prevents the user from syntactic mistakes. In this paper we will present the architecture and general principles of such a system. Section 2 aims to provide a short introduction to the semantic web and the techniques used in the current application. Section 3 describes the system, while Section 4 presents the prospected work.

2 The new generation of Semantic Web

According to the seminal paper of Tim-Berners-Lee [Berners-Lee&al99] the Semantic Web "will bring structure to the meaningful content of Web pages, creating an environment where software agents roaming from page to page can readily carry out sophisticated tasks for users". This is still a vision of the future but already key technologies, which will make it possible, were developed. Two directions are of great interest.

- The development of a model and a standard notation which allows the enrichment of Web-resources with semantic information
- The development of a mechanism able to connect this semantic information via semantic –based relations.

The first issue is addressed by the RDF-model [RDF], which allows the description of each resource as a triple (Subject, Predicate, Object). Resources are unique identified via URIs. The serialization of the RDF model is done in XML, which makes it fully compatible with Web applications

For the latter issue, ontologies seem to be the most appropriate mechanism. The offer the great advantage that they are language independent, therefore they can be used as central semantic structure on which lexical mapping is performed. Several RDF-conform languages for ontology description were proposed. The last one OWL [OWL], is designed to allow a large variety of semantic relationships between classes. OWL is fully RDF-readable.

Although these basic mechanisms for making "Semantic Web" real, are already available, few applications bring them together. Very often applications concentrates either on ontology tools, i.e. tools able to read, convert, manipulate ontologies in RDF/DAML-OIL/OWL format, or they deal with RDF-representations of Data. These are important brick stoned but they do not demonstrate the power of the Semantic Web concept. A still very rarely, although extremely important, mentioned idea is the "multilinguality".

There are millions of documents in the web, written in different languages. At this moment information is retrieved only in the language in which the input was given. On one hand it is true that many users are able to speak and understand a limited number of languages, but, on the other hand, many have skills in at least 2-3 other languages as their native ones. Moreover, with the development of semantic Web, machine translation will make also a step forward and on-line tools will be able to translate automatically Web resources. Therefore it is important to built tools able to retrieve information referring the same concept, in different languages.

In the following section we will present a prototype system aiming to retrieve touristic information from the web in more than one language. We will also explain how the system can be adapted for other languages.

3 Cross lingual in Semantic Web

Multilingual Tourist planner is a prototype system aiming to allow a controlled natural language interface for querying the Semantic Web. It brings together the power of natural language understanding and the Semantic Web principles. The user is guided all the time when typing the input so that a syntactic correct input will be provided to the system.

The scenario for under the assumption which the system is implemented is the following: the user speaking language S, not necessary as native speaker, (in our case English) wants to obtain information about country X (in this particular case Romania). It is assumed that the informations about the country will be retrieved in several languages, first Romanian, but also German, English, Italian etc., according to the touristic offers in each country. The retrieved information has to be presented to the user as a collections of texts and/or web addresses and has to be relevant for the topic of the query.

3.1 Components of the System

In Figure 1 is presented the architecture of the system. There are three main modules:

1. The user interface: has double role: on one hands acts as a normal Web interface, and offers important information about the target country. It is designed so that the user is directed to formulate questions in natural language about a particular region and/or a particular domain. For the moment the following domains are available: culture, sport and travel. The user can configure also the languages in which he will expect a result.
2. The NL module cooperates with the User interface and ensured the controlled input. It transfers to the next module the central information of the query
3. The Info-module is the core of the system. With the input from the NL module checks Tries to find relevant concepts and properties on the ontology on which the lexical entries are mapped . This means to find the classes from which the lexical

entries are instantiated. Once retrieved the module searches for instantiations of the concepts in the languages configured in the user interface. These instantiations are the tags in the RDF description of the Web-resources. The info module retrieves the URIs of these resources and passes them to the answer module

4. The answer module is responsible with resuming of the texts and presenting them to the user.

5. Additional modules in the system are viewed only as external tools aiming to facilitate the work of the main components. The Ontoviz Tool, visualize the ontology, in order to make easy the concept

6. The lexical mapping tool. The language Ressource Tool is used for editing the language resources for the controlled input

3.2 Controlling the input and extensions of the system

The idea of controlling the input is not new. It was addressed in several papers [Moore&Mittal95], [VertanvHahn03]. In our system the input is controlled (predicted) by cotext and domain. A number of possible patterns for the questions were identified (for the moment 100). The patterns are sequences of lexical equivalence classes (fillers) like :

Loc_question | aux_struct | action_verb | action_region

From such patterns relevant features for the ontology search are extracted, for example:

(Var_location, action_verb,action_area)

The system can be easier extended to other input languages. The following components have to be provided:

- A new lexicon and its mapping on the ontology
- A new sequence of patterns for the source language and fillers

4 Conclusions and Further Work

The system presented here is now under development. It represents a middle-way between free natural language input and keyword spotting. The system is implemented in Java, and the ontology an Web resources are RDFS respectively RDF-based. For the moment the patterns of the input sequences are only in the form of a test set. Further work concerns extraction of patterns from large corpora. Another aspect to be explored concerns the feature extraction from the input. For the moment we follow the RDF-Model of (subject, predicate, object). However for the semantics of the input, and consequently for the accuracy of the output it seems appropriate to extract additional features. In this respect an evaluation phase is foreseen. After the completion of the System an extension to a third language will be performed in order to show the portability of the system. We intend also to realise an evaluation of the transparency of the system: i.e. how the users evaluate such a system, in comparison with a normal Web search engine, or a navigation through a standard website.

5 Acknowledgements

Our thanks go to the German foundation DFG, which financed part of the research through a 1 month grant at the Macquarie University, Sydney

References

[Androutsopoulos&Aretoulaki03] Androutsopoulos, I. and Aretoulaki, M, “Natural Language interaction”, in The Oxford Handbook of Computational Linguistics Mitkov R. (ED.), Oxford University Press, 2003, pg.629-649

[Berners-Lee&al99] Berners-Lee, T., Hendler, J. and Lassila, O., “The Semantic Web”, Scientific American, 1998[

[DaleMoisl&Somers00] Dale R., Moisl, H., Somers H., Handbook of Natural Language Processing, Marcel Dekker, 2000

[Fensel&al03] Fensel, D., Hendler, J., Lieberman H and Wahlster, W., “Speinning the Semantic Web”, MIT Press, 2003

[Moore&Mittal95] Moore, J.D. and Mittal, V., “Dynamic Generation of Follow up Question Menus:Facilitating Interactive Natural Language Dialogues, Proceedings of CHI’95, pag. 90-97

[RDF] Resource Description Framework (RDF) Model and Syntax Specification, http://www.w3.org/TR/REC-rdf-syntax

[OWL] OWL Web Ontology Language, http://www.w3.org/TR/2004/REC-owl-features-20040210/

[VertanvHahn03] Vertan, C., and v. Hahn, W., “Menu choice translation – aflexible menu-based controlled natural language system”, in Proceedings of EAMT-CLAW conference on Controlled Language translation, pag. 194-199