

Natural Language Interface for Querying Databases

Péter Vajda

HAS, Research Institute of Linguistics, Department of Corpus Linguistics
vajda@corpus.nytud.hu

Keywords: database, ontology, question answering, RDF, Semantic Web

The extraction of relevant data from the ever increasing amount of information on the Internet has been a difficulty not only for end-users but developers of search engines alike. Due to the ambiguity of natural languages to quickly acquire the desired information, it is crucial for a program to be able to perform its search not only on character strings, but on information that is meaningful for it. In this case the query of a user could consist of a more natural and precise sentence, instead of plain keywords. According to the *Semantic Web* initiative one needs a so called *ontology*, which is a common-sense knowledge formally expressed for the computer.

The other shortcoming of today's search engines is that they are not able to access the so called *Deep Web*, which mainly consists of otherwise publicly accessible databases. The data contained in these databases could especially be useful in answering some simple, factual questions, for which the answer can hardly be found with keyword-based searching. E.g. "Which is the most expensive film of the 90's"

In my work conducted in the Technical University of Budapest and the Research Institute of Linguistics I try to design and implement a system, which deals with the two problems above. So far I have proposed a layout for a Natural Language Interface connected to a Database. Its task is to translate a Hungarian question to one or more SQL queries, that can be fed into the available databases. The constituent I plan to implement will be responsible for the "semantic matching", which trns a shallow-parsed form of the question, into the interpretation of the question in a language of logic based on RDF. The final SQL query will be formulated from this interpretation and the scheme of the database. To perform the semantic matching, we need an ontology, composed of the domain's concept hierarchy and semantic rules, which map words to concepts. The units of the concept hierarchy are entities and relations between them. The semantic rules assign to each word an entity or relation of appropriate type. For the accurate construction of the ontology and for the input processing it was also necessary to collect and grammatically examine the potential Hungarian questions. During the process of semantic matching the system assigns concepts and predicates to the words of the question in order to form a logical expression corresponding to the question. In my presentation, I intend to illustrate the process of semantic matching using the program developed for this purpose.