Available online at www.sciencedirect.com
SCIENCE DIRECT®
Data & Knowledge Engineering 55 (2005) 38–58
DATA & KNOWLEDGE ENGINEERING
www.elsevier.com/locate/datak

Simplifying syntactic and semantic parsing of NL-based queries in advanced application domains

E. Kapetanios [a,*], D. Baer [a], P. Groenewoud [b]

[a] *Department of Computer Science, Swiss Federal Institute of Technology, Zurich (ETHZ), Hirschengraben 84, Zurich CH-8092, Switzerland*
[b] *University of Applied Sciences, Oberseestr. 10, Rapperswil CH-1475, Switzerland*

Received 4 November 2004; accepted 4 November 2004
Available online 23 December 2004

Abstract

The paper presents a high level query language (MDDQL) for databases, which relies on an ontology driven automaton. This is simulated by the human–computer interaction mode for the query construction process, which is driven by an inference engine operating upon a frames based ontology description. Therefore, given that the query construction process implicitly leads to the contemporary construction of high level query trees prior to submission of the query for transformation and execution to a semantic middle-ware, syntactic and semantic parsing of a query with conventional techniques, i.e., after completion of its formulation, becomes obsolete. To this extent, only, as meaningful as possible, queries can be constructed at a low typing, learning, syntactic and semantic parsing effort and regardless the preferred natural (sub)language. From a linguistics point o view, it turns out that the query construction mechanism can easily be adapted and work with families of natural languages, which underlie another type order such as *Subject–Object–Verb* as opposed to the typical *Subject–Verb–Object* type order, which underlie most European languages. The query construction mechanism has been proved as practical in advanced application domains, such as those provided by medical applications, with an advanced and hardly understood terminology for naive users and the public.

* Corresponding author.
E-mail addresses: kapetanios@inf.ethz.ch (E. Kapetanios), dbaer@acm.org (D. Baer), pgroenew@hsr.ch (P. Groenewoud).

0169-023X/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.datak.2004.11.008

Keywords: Natural languages; Query languages; Ontologies; Semantic web; Databases; Information retrieval; Automata theory

1. Introduction

1.1. Background

Querying databases through Natural Language (NL)-based query interfaces—we exclude *keywords based* querying—has always attracted research and development efforts in order to ease access to and increase understandability of the full potential of information as provided by large data repositories [1–5]. Syntactic and semantic parsing of NL-based queries, however, turns out to be tedious [6,7], especially when complex or advanced terminologies are used like those found in scientific and technical application domains. Mostly, this kind of query parsing relies in constructing a query tree which is compliant with the underlying NL-based syntactic and semantic rules, which are different as they depend on the preferred NL.

In addition, constructing a query presupposes that the user is familiar with the terminology itself as well as the semantic relationships among terms of the application domain. In other words, the user needs to know not only the orthography of words and their intentional meaning, as expressed in a particular natural (sub)language, but also how they relate to each other. It would be impossible or, at least, too optimistic to assume that one makes itself familiar with the full range of the scientific or technical domain vocabulary, in both terms, spelling and meaning.

Therefore, the more advanced or complex vocabularies become, the more prone to syntactic and semantic mistakes becomes the process of NL-based query construction and parsing. This is also strengthened by multi-lingual user communities, where different parsing techniques need to be considered and implemented. Consequently, either the complexity of the parsing technique increases dramatically or large parts of the information potential in databases for scientific or engineering applications remain unexplored and cannot become part of exploratory querying.

Moreover, it is still possible to construct a syntactically and semantically correct query, in terms of the NL-based semantics, however, the query might be still meaningless [8], since it might not reflect *real-world* semantics. This is due to the fact that semantic parsing mostly refers to the NL-based grammatical roles of words or structure of the query [9,10], i.e., which is the *subject* or *noun phrase*, the *verb*, the *object*, the *connectives*, etc., and not to the application domain semantics itself [11] as expressed by some kind of ontological considerations. A query saying *All cars aged more than 40 years, which have been infected by AIDS* is grammatically correct as far as the grammatical roles (noun phrase, relative clause, verb, etc.) of the participating words and phrases have been correctly recognized by the parsing technique.

Complexity of query parsing and transformation is further increased, since, regardless the preferred natural language, the same *query extensional semantics*, in terms of tuples received from the databases as an answer to the query, are expected to hold. To this extent, either the same parser needs to consider various NL-specific syntactic and semantic rules, or different parsers need to be enabled according to the number of the preferred natural languages. Finally, it is tedious or even impossible to integrate into the parsing technique the intentional