Ontology-Based Querying

Troels Andreasen¹, Jørgen Fischer Nilsson², and Hanne Erdman Thomsen³

¹ Roskilde University, Roskilde, Denmark
² Technical University of Denmark, Lyngby, Denmark
³ Copenhagen Business School, Copenhagen, Denmark

Abstract. This paper introduces the main objectives of the OntoQuery project, with the agenda of developing theories and methodologies for content-based retrieval from text databases. Content-based retrieval is obtained through descriptions of database objects and queries derived from natural language analysis and shaped with respect to a domain specific ontology, which in turn is expressed in a formal description language. Querying is performed by comparing descriptions and, during this, reasoning with the domain knowledge expressed in the ontology.

1 Introduction

The project title, OntoQuery, is short for 'Ontology-based Querying', and the project aims to contribute to the development of general solutions to the querying of text databases and to the extraction of descriptions of database objects through limited computational natural language understanding. More precisely, the project addresses content-based retrieval and access to Danish text sources such as online document databases and encyclopedias.

Stressing the use of ontologies, the project provides a content-based query and retrieval functionality going beyond the superficial key word recognition typical of contemporary search engines, whilst not attempting a full semantic analysis of source texts.

The overall methodological goal of the project is to develop a coherent theory for

• ontological representation of domain knowledge,
• ontological semantics for natural language phrases, and
• ontology-based search in text databases.

This theory is expressed in a formal concept language under development in the project, named ONTOLOG. For purposes of validation and demonstration, the theoretical results are exploited in a prototype system with accompanying tools and resources on selected real world domains.

The project addresses various research aspects associated with the three aspects of the goal:

• Development of theories, methods and tools for establishing formal ontologies integrated with language specific terminology and lexical networks. A key idea is to introduce a formal language for ontologies, ONTOLOG, which combines taxonomies with object and relational expression forms.
• Development of methods for ontology-based linguistic analysis of source texts and queries. This primarily concern the identification and analysis of noun phrases, comprising morphological, syntactic and semantic analysis. In our approach the noun phrases are central for the specification of particular concepts in an ontology.

• Development of methodologies for ontology-based query processing that efficiently compare an internal formal description of a query with the ontological descriptions of text database objects. Query processing then becomes matching of the query description with the descriptions of text database objects in the framework of the given ontology.

ONTOLOG is used simultaneously for the representation of domain knowledge in the ontology, for the representation of natural language semantics, and for descriptions of the texts in the database, and it will allow for reasoning with the ontology. It is intended primarily as a theoretical, logical framework, in which the different traditional representations at the various levels of analysis is conceptualised, analysed, and integrated. Thus a strategic purpose is to facilitate coherence in the resulting system architecture. The language offers descriptions, which serve multiple purposes ranging from feature structures in the linguistic analysis, via lexical semantic bases and terminology bases to ontologies and query descriptions.

The use of natural language for database querying traditionally relies on a logical semantics, which determines the translation from syntax trees into a logical query language. This technique has to restrict the query language to a small fragment of natural language, since a full computational semantic treatment of comprehensive natural language fragments is beyond the scope of current language technology. As an alternative approach this project introduces an ontology-based semantic analysis for natural language texts and query phrases, which refrains from a full logical analysis of the meaning of natural language texts. As a starting point we focus on the analysis and disambiguation of noun phrases, including in particular adjectives, prepositional phrases (both complements and adjuncts) and genitives.

2 Description language

As mentioned a model of the target domain of the text database is to be established in the form of an ontology. To this end a formal description language called ONTOLOG [12] is introduced, see also [3,7].

The expressions of this language are terms in an algebraic logic which function as descriptions of concepts.

Noun phrases in the domain texts and in user retrieval requests (queries) are to be compiled into descriptions in ONTOLOG. These descriptions are situated in the ontology, so that noun phrases thereby are mapped into positions in the ontology structure, which is shaped as an algebraic lattice.