# *OntoQuery*: Ontology-based Querying of Texts

**Troels Andreasen**

Computer Science, Roskilde University, Denmark

**Per Anker Jensen**

Business Communication and Information Science, University of Southern Denmark

**Jørgen Fischer Nilsson**

Informatics and Mathematical Modelling, Technical University of Denmark

**Patrizia Paggio & Bolette Sandford Pedersen**

Centre for Language Technology, Copenhagen, Denmark

**Hanne Erdman Thomsen**

Computational Linguistics, Copenhagen Business School, Denmark

## Abstract

This paper addresses techniques for extracting conceptual descriptions from text sources and queries based on an ontology. The taxonomic ontology comprises feature-structured conceptual descriptions, and retrieval is based on description similarity relative to the ontology.

## Introduction

The aim of the OntoQuery project (Andreasen, Nilsson, and Thomsen 2000; OntoQuery) is to contribute to the development of general solutions to the querying of text databases and to the extraction of conceptual descriptions from such databases through limited computational natural language understanding. More precisely, the project addresses content-based retrieval and access to text sources, such as online document databases and encyclopaedias.

## Methodological Aims

The key methodological goal of the project is to create a coherent framework for ontological representation of domains, ontological semantics for pertinent natural language phrases, and ontology-based search in text databases.

Coherence is established through a formal and computational language of conceptual descriptions uniting

1. formal ontologies
2. lexicons and lexical semantics
3. ontology-structured semantic domains for nominal phrases
4. querying and search

For purposes of validation and demonstration, the theoretical results are examined in prototypes with accompanying tools and resources on selected real world domains.

## Formal Ontology and Semantic Lexicon

Algebraic lattices have been chosen as our theoretical basis, serving as skeleton ontologies shaped by the conceptual inclusion relation.

These ontological structures, spanned by general categories such as physical objects, substances, and events, transcend
hierarchical classifications by admission of cross-categories. They can be visualized by diagrams in a well-known fashion.

In the logical representation language *OntoLog* (Nilsson 2001a), the nodes of such skeleton ontologies are adorned with attributes giving rise to compound conceptual descriptions intended for the representation of the conceptual content of noun phrases (NP's).

Descriptions reflect a compositional, phrase-structured semantics where connection is established with lexical semantics as applied in the SIMPLE European initiative, see e.g. (Lenci et al. 2000), (Pedersen and Keson 1999) and (Pedersen and Nimb 2000), which adopts principles from the generative lexicon approach (Pustejovsky 1995). The lexical entries comprise a four-dimensional inheritance structure using the qualia roles: formal, agentive, telic, and constitutive.

The project incorporates the Danish SIMPLE lexicon, which consists of 10,000 Danish sense descriptions extended in the OntoQuery project with 1,000 domain-specific concept descriptions within the nutrition domain. All terms have been extracted from about one hundred articles on nutrition from The Danish National Encyclopaedia.

## Conceptual Descriptors

The descriptors appearing as nodes in the formal ontology are to represent the conceptual content of nominal phrases comprising nouns (including noun-noun compounds, which abound in Danish), adjectives, and prepositional phrases. The determiner structure is ignored as it is considered irrelevant to the conceptual content of the phrase. Nominal phrases are identified using preprocessing techniques such as POS-tagging and NP recognition (cf. Paggio et al. 2001).

Formally, the representations are akin to the (typed) feature structures commonly applied in unification-based parser systems, e.g. the LKB system (cf. http://www-csli.stanford.edu/~aac/lkb.html). However, the attributes applied in this project express ontological relationships such as parthood, causality etc., rather than syntactic or functional relations.

For instance, the descriptor corresponding to the NP *fedtdepoter hos børn* (fat deposits in children) is the complex descriptor resulting from the combination of the atomic descriptor ‘deposit’ with the descriptors ‘fat’ and ‘child’ by means of the relations CON (contains) and LOC (located-in). The descriptor is shown in (1) below.

$$\text{(1) deposit} \quad \begin{bmatrix} \text{CON:} & \text{fat} \\ \text{LOC:} & \text{child} \end{bmatrix}$$

Another example is the descriptor in (2), which corresponds to the NP *behandling af børn med overvægt* (treatment of children affected by obesity), as well as to its paraphrases, e.g. *behandling af overvægtige børn* (treatment of obese children). Here the relations considered are PNT (patient) and CHR (characteristic).

(2) treatment [PNT: child [CHR: obesity]]

From a logical point of view descriptors are algebraic terms in a relational logic resembling description logic, but with a stronger bias on the taxonomic lattice structure.

## Ontological Grammar

A logically simplified and computationally more manageable version of the logico-algebraic representation framework is achieved by letting the ontology with accompanying descriptions be specified and generated by a BNF-grammar (Nilsson 2001b). This gives rise to the notion of ontological grammar supporting generative ontologies with an infinitude of descriptions through recursive definitions of categories. In such grammars the syntactic derivation relation mediates the conceptual specialization relation.

An ontological grammar specifies semantic target domains for the considered linguistic phrases, enforcing ontological combinability restrictions, which assist in the disambiguation of the linguistic phrases. The descriptors are produced by the ontological grammar.

By way of illustration, the Danish NP *behandling af børn med overvægt* considered above is ambiguous between the reading given in (2) and the instrumental reading represented by the descriptor in (3), where BMO abbreviates ‘by means of’:

$$\text{(3) treatment} \quad \begin{bmatrix} \text{PNT:} & \text{child} \\ \text{BMO:} & \text{obesity} \end{bmatrix}$$

While the reading in (2) is supported by a rule allowing physical properties, in casu ‘obesity’, to be attached to physical objects, in casu ‘child’, the reading in (3) is rejected on ontological grounds, since ‘obesity’ is a ‘property’, which cannot serve as an ‘instrument’, and in the ontological grammar rules, ‘property’ is excluded as a possible ontological argument type for the BMO-relation.

## Querying

The basis for querying and search is the evaluation of similarity between descriptions (collections of descriptors) from text sources in the database and the description of the query phrase. Given a query, retrieval of relevant text parts is achieved via the ontology with retro-links from descriptors into the text sources in the database.

The comparison of descriptors is not merely syntactic. Rather, their resemblance is measured in terms of similarity derived from concept relations in the ontology. Similarity between the descriptors, as situated in the ontology, is thus obtained from engaging ontological reasoning capabilities.

The texts in the database have descriptions attached. Descriptions are defined as sets of sets of descriptors and are derived at the level of sentences. Initially, in the processing of a query, a description is generated. Then this query description is compared, in principle, to every description of every sentence in every document appearing in the database. Finally, sentences in the documents in the database are ranked by the degree to which their description resembles the description of the query. The query answer is a ranking of the sentences that are most similar to the query.

As mentioned briefly above, descriptors are managed and compared in structures called descriptions, which have the form:

$$D = \{D_1, \ldots, D_n\} = \{\{D_{11}, \ldots, D_{1m_1}\}, \ldots, \{D_{n1}, \ldots, D_{nm_n}\}\}$$

where each $D_i$ is a set of descriptors $D_{ij}$, j=1,…,$m_i$. Each $D_i$ corresponds to an NP in the text described, and, as it appears, it may embed one or more descriptors.

Descriptions are not unique and may vary by level of detail and combinability. Among the possible descriptions for the phrase: *Alternativ medicin og behandling af børn med overvægt* (alternative medicine and treatment of children affected by obesity) are the following increasingly accurate descriptions:

(4)
{{alternative},{medicine},{treatment},{child},{obesity}}

(5)
{{alternative, medicine},{treatment},{child, obesity}}

(6)
{{medicine [CHR: alternative]},
{treatment [PNT: child [CHR: obesity]]}}

The principle applied for comparison of descriptions is a fuzzy nested aggregation combining descriptor similarity into similarity of full description structures. Descriptor similarity is derived from concept relations in the ontology, for instance based on distances measured over the shortest paths of reasoning. This principle is described in more detail in (Andreasen 2001).

Descriptions for each sentence in the database are compiled and stored in the database, and, furthermore, all compound descriptors in database descriptions are, in principle, expanded to cover also subsuming concepts. For instance, the element

(7) {treatment [PNT: child [CHR: obesity]]}

may be expanded to the set

(8) {treatment, treatment [PNT: child],
treatment [PNT: child [CHR: obesity]]}

thereby including subsuming concepts. Assuming a disjunctive interpretation, this means that the database object described by (7) is considered a good candidate answer object for a query on, e.g., 'treatment'. A specific answer to a general query is useful, while the opposite is typically not the case.

Finally, the evaluation is initiated with an expansion of the query to include similar descriptors as alternatives. Thus, if the ontology includes 'obesity ISA disorder', then the term 'disorder' may be replaced by e.g. the fuzzy set of terms '1/disorder + 0.9/obesity', indicating that 'disorder' matches fully while 'obesity' matches to some degree.

## Challenge Questions

The project is currently developing an ontological grammar for the domain of nutrition. One of the challenges to be faced is how to extend this grammar to other technical and scientific domains, while retaining the disambiguating power of the ontological grammar. Adjustment to new domains further requires that appropriate domain ontologies are made available. Thus, research in automatic term extraction and ontology building are relevant to the OntoQuery project.

Whenever new material is added to the text database, the new texts have to be processed for the purpose of generating descriptors. Unknown words are treated as descriptors denoting concepts per se, but their position in the ontology is unknown, and thus synonymy, hyponymy etc., are disregarded. The project addresses querying of Danish text databases, but few resources such as lexicons and grammars exist for this language. Therefore, to make the system perform better on unknown material and new domains, larger ontology-based lexicons for Danish are needed.

### Acknowledgments

The project is funded in part by the Danish Research Councils under the interdisciplinary Informatics Programme from 1999 through 2004.

### References

Andreasen, T; Nilsson, J. Fischer; and Thomsen, H. Erdman 2000. Ontology-based Querying. In *Flexible Query Answering Systems*, *Recent Advances*, 15-26. Physica-Verlag, Springer.

Andreasen, T. 2001a. Query evaluation based on domain-specific ontologies. In *NAFIPS'2001, 20th IFSA / NAFIPS International Conference Fuzziness and Soft Computing*, 1844-1849, Vancouver, Canada.

Lenci, A.; Bel, N; Busa, F; Calzolari, N; Gola, E; Monachini, M; Ogonowski, A; Peters, I; Peters, W.; Ruimy, N.; Villegas, M.; and Zampolli, A. 2000. SIMPLE: A general Framework for the Development of Multilingual Lexicons. *International Journal of Lexicography* 13(4), Oxford University Press, Oxford.

Nilsson, J. Fischer 2001a. A Logico-Algebraic Framework for Ontologies ONTOLOG. In Jensen; P. Anker, and Skadhauge, P. eds. Proceedings of the First International OntoQuery Workshop Ontology-based interpretation of NP's. 11-42. Department of Business Communication and Information Science, University of Southern Denmark, Kolding.

Nilsson, J. Fischer 2001b. Concept Descriptions for Text Search, In *Proceedings from the 11th European-Japanese Conference on Information Modelling and Knowledge Bases*. 284-288. Maribor, Slovenia. Forthcoming in Information Modelling and Knowledge Bases, IOS press.

OntoQuery. Project web site: http://www.ontoquery.dk

Paggio, P; Pedersen, B.S.; and Haltrup, D. 2001. Applying Language Technology to Content-based Querying – The Ontoquery Project, in Proceedings from Workshop on Artificial Intelligence for Cultural Heritage and Digital Libraries. 75-79. Università di Bari, Italy.

Pedersen, B.S., and Keson, B. 1999. SIMPLE - Semantic Information for Multifunctional Plurilingual Lexica: Some Danish Examples on Concrete Nouns, in: *SIGLEX99: Standardizing Lexical Resources*, Association of Computational Linguistics, ACL99 Workshop, Maryland.

Pedersen, B. S., and Nimb, S. 2000. Semantic Encoding of Danish Verbs in SIMPLE - Adapting a verb-framed model to a satellite-framed language. In *Proceedings from the Second International Conference on Language Resources and Evaluation*, LREC 2000, Athens.

Pustejovsky, J. 1995. *The Generative Lexicon*. MIT press.