

# Ontology-based question answering with feature structures

March 23, 2006

This paper describes the Danish grammar resources developed within the European project MOSES (Modular and Scalable Environment for the Semantic Web), whose objective was to develop an ontology-based methodology to create, maintain and search semantically structured Web contents in a federation of sites (Atzeni et al, 2004) (Paggio et al. 2004). The test bed was an agent-based knowledge management system and an ontology-based search engine for the Web sites of the two European universities of Roma Tre and Copenhagen. Natural language interfaces have been developed for Danish and Italian.

Although MOSES can be seen as a question answering (QA) application from the point of view of the interaction with the user, it differs from classical QA systems because it is ontology-based and relies on the topic maps formalism for the representation of Web contents. This means that natural language processing, instead of targeting texts as is customary in QA, must interface to the topic maps knowledge structure both in question analysis and answer generator. In this paper we discuss the way in which the Danish feature structure grammar developed previously by means of the LKB platform has been adapted to work in the MOSES application, and how the PET parser used to run the grammar is integrated in the MOSES architecture.

Question analysis is carried out in the MOSES linguistic module associated with each system node. In accordance with the semantic Web approach, MOSES poses no constraints on how the conceptual representation should be produced, nor on the format of the output of each linguistic module. However, the building blocks of the linguistic output must be the classes, instances and relations of the corresponding ontology. The output is transformed into an XML query understandable to the knowledge base by the content matcher. The Danish module consists of a pre-processor responsible for tokenisation, PoS tagging, lemmatisation and named entity recognition, and a parser. The parser is an adapted version of PET (Callmeier 2000), which produces for each input question a set of typed feature structures corresponding to the semantic analysis of the question. In other words, although syntactic and semantic analysis are conflated in the grammar as is customary for constraint-based grammars inspired to the HPSG paradigm, only the semantic information is retained in the output.

For instance, the analysis of the question “*Hvem underviser i databaser?*”

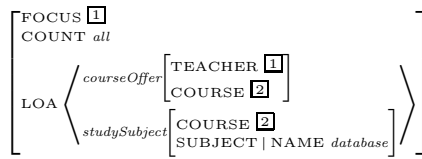


Figure 1: Example of semantic representation

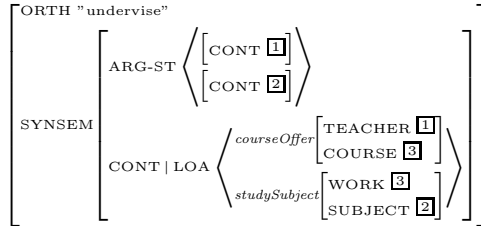


Figure 2: Danish lexicon entry

(Who teaches databases?) is as shown in Figure (1). At the highest level, the FOCUS attribute refers to the semantic class the knowledge structure is to be queried for, while the value of COUNT indicates that we are interested in all the topics belonging to this class, and LOA (list of associations) contains a list of relevant semantic relations, in this case *courseOffer* and *studySubject*, that further constrain the topics in question. Each relation is in turn a feature and has a number of attributes corresponding to semantic roles. The value of a semantic role is a domain class. The terms *topic* and *association* come from the topic maps knowledge structure which the linguistic analysis interfaces to. The entire set of semantic types (about 200 classes and 50 relations) that are part of the ontology is shared between the linguistic analysis module and the query execution module.

Typed feature structures have proved an adequate formalism to define the domain ontology as part of the underlying type hierarchy, and also to express the necessary relations between lexical and syntactic features on the one hand, and domain content on the other. These relations are defined in the individual lexical entries. For example, a simplified entry for the verb *undervise* (teach) looks as shown in Figure (2): token identity is enforced between the TEACHER and SUBJECT roles and the arguments of the verb.

The grammar has been tested on a list of 85 questions of varying syntactic complexity spanning a wide range of semantic relations. Of these, 65 were analysed correctly, 3 produced an incorrect analysis, and 17 failed to produce an analysis.

Although the grammar is able to produce correct semantic representations in a number of non-trivial cases, more robustness should be added to produce at least partial analyses for the cases that fall outside of the implemented coverage.