Chapter 7
Natural Language Query Understanding

7.1 Overview

A natural language interface is always desirable for question answering and information retrieval systems (Ogden and Bernick 1997). Using a controlled language for the interface could ease the problem but still tightens users to a restricted grammar (Nyberg and Mitamura 1996; Nelken and Francez 2000). Although performance of machine natural language understanding for the general case appears to be saturated after many years of research, limiting the domain of discourse to only questions and querying phrases can make a difference.

In particular, users often want to search for documents about named entities (NE), i.e., those that are referred to by names (Sekine 2004), such as people, organizations, and locations. An example query is to search for documents about “economic growth of countries in East Asia”, where “countries in East Asia” represents named entities while “economic” and “growth” are keywords. Relevant answer documents include those about economic growth in certain countries like China, Japan, and Korea. Searching purely based on keywords fails to find such documents, because it does not recognize the named entities implied in the query. For answering queries containing named entity phrases, a system needs to convert those human language phrases to more formal representations, in order to extract such latent semantics from an ontology of discourse. There are different approaches to query conversion regarding the two following issues.

First, it is about whether rigorous syntactic parsing is applied to a query expression before it is mapped to a target language sentence. The disadvantages of the parsing approach are time consuming and requiring grammatically correct inputs, which is thus not robust to ill-formed queries. It is also not practical to require a user to always input a question without grammatical errors. Moreover, it may still face to the problem of syntactic ambiguity, i.e., one sentence having more than one applicable syntax tree.

Second, it is about whether an ontology is employed in the mapping. For example, with the query “What county is Modesto, California in?”, given no ontology, Modesto and California can be tagged only as proper nouns and thus the implicit relation expressed by the comma between them cannot be interpreted. In contrast, with an ontology, they can be recognized as named entities of the types (or classes) City and Province, respectively, whence the relation can be mapped to one being a sub-region of the other.

For instance, Lei, Uren and Motta (2006) implemented an ontology-based search system whose queries were lists of classes and instances and converted into
expressions of SeRQL. They were better than lists of normal keywords, but not as natural as human expressions. Meanwhile, accepting natural language queries, Cimiano, Haase and Heizmann (2007) followed the rigorous parsing approach using lambda calculus as an intermediate formal language for the conversion. However, the focus of that work was on efficient porting interfaces between different domains rather than on the language conversion itself.

The approach in Kaufmann, Bernstein and Fischer (2007) could be considered as closer to the syntax-free one. It used pattern matching of a natural language query to subject-property-object triples in a knowledge base, before converting the query to one of SPARQL. For the example query therein “What is a restaurant in San Francisco that serves good French food?”, it first searched for those triples whose subjects, properties, and objects could match with “restaurant”, “in”, and “San Francisco”. That method thus could not produce a mapping if the ontology did not contain such a triple for the named entity San Francisco, which existed in the ontology though. Recently, Tablan, Damljanovic and Bontcheva (2008) also followed the syntax-free approach to convert natural language queries into SeRQL expressions.

Since the root of the difficulty of machine natural language understanding is the big gap between natural language and a machine executable one, using a higher level language like SeRQL or CGs as an intermediate is a way to ease the problem. However, SeRQL is still far from natural language due to its RDF primitive triple structure, which breaks knowledge representation into too small granules. We choose CGs because they could be mapped smoothly to and from natural language, and used as an interlingua for conversion to and from other formal languages (Sowa 1997). There was research on automatic generation of CGs from natural language texts in a specific domain, e.g. the machine learning-based one in Zhang and Yu (2001) and the rule-based method in Hensman and Dunnion (2004). However, both of the works required syntactic parsing of input sentences and were evaluated mainly on semantic roles rather than whole sentences.

This chapter presents the method for mapping natural language queries to conceptual graphs initially proposed in Cao, Cao and Tran (2008) for simple queries and extended in Cao and Anh (2010) for complex queries. It does not require and rely on a strict grammar of querying phrases or sentences, but does use an ontology for the conversion. Firstly, Section 7.2 introduces the adapted vector space models developed in Cao, Le and Ngo (2008) for ontology-based information retrieval. Section 7.3 defines nested query CGs to represent connective, superlative, and counting queries. Section 7.4 presents in detail the proposed ontology-based method for converting queries in natural language to conceptual graphs. Section 7.5 presents VN-KIM search engine using the developed ontology-based information retrieval models and query processing method. Finally, Section 7.6 concludes the chapter.

7.2 Ontology-Based Information Retrieval

Despite having known disadvantages, the Vector Space Model (VSM) is still a popular model and a basis to develop other models for information retrieval,