Representation of scientific texts in knowledge graphs

Vries, Pieter Hendrik de

SUMMARY

In this study the possibilities of structuring knowledge by means of a computer program are explored. The central issue is the representation of scientific text with the aim of automatic knowledge structuring. We will focus on the representation of knowledge from medical and social-scientific texts. In Chapter 1 we will sketch the outline of the thesis and mention the preliminary studies.

Representation of knowledge is a topic that receives much attention in the research area referred to as Artificial Intelligence. In order to do justice to the approach followed in this thesis, we will submit the developments in this area to a critical discussion. This discussion is the subject of Chapter 2.

The system on which we will focus in this study is based on relationships that are fundamental to scientific knowledge. In Chapter 3 we distinguish three types of such relationships. This distinction makes it possible to express that: (1) an object has a certain property, (2) an object is an instance of another object, (3) a change in a property of an object leads to a change in another property of that object. These three types of assertions can be adequately represented in a directed graph. A directed graph in which these relationships are distinguished is referred to as a knowledge graph. We also distinguish certain structures in a knowledge graph that are denoted as frameworks. They partially compensate the loss of expressive power that is brought about by our restriction to three types of relationships.

In Chapter 4 we introduce four procedures that have been developed for the structuring of knowledge graphs. The procedure that we will emphasize in this thesis is the one for the construction of a knowledge graph on the basis of a scientific text.

In Chapter 5 we elaborate the representation of a scientific text in a knowledge graph. We develop here the idea that any sentence in a text can be seen as a combination of so-called operators (such as verbs) and argu- ments (such as nouns). Certain combinations of operators and arguments are called object sentences. In the simplest case an object sentence consists of an operator which expresses one of the three distinguished relationships, and two arguments. As such, an object sentence corresponds to the smallest meaningful unit in a knowledge graph. In certain well-defined situations an object sentence will be represented as a framework. The analysis of a text into object sentences and the representation of these object sentences in a knowledge graph are formalized in a series of prescriptions. Within the chosen approach various problems inherent in the use of language are solved. These problems involve the discovery of the partially implicit referential structure in a text, the role of ambiguous or incomplete assertions, and the function of context.

Besides the procedure for text analysis we distinguish three other procedures for knowledge structuring. These procedures are a result of research that is being carried out elsewhere and that focuses on the mathematical aspects of knowledge graphs. In chapters 6, 7, and 8 we summarize an earlier description of these procedures in order to give a complete overview of the intended system for knowledge structuring.

In Chapter 9 we give some concrete examples of knowledge structuring. In the first place the practical usefulness of the procedure for text analysis is investigated. This inquiry shows that "manual" coding of a text easily

leads to incorrect application of the prescriptions from the procedure for text analysis. We emphasize, however, that this procedure is highly suitable for the development of a special text processor which helps a coder to find the prescribed representation for a text. In the second place we compare in this chapter the procedure for text analysis with an already existing procedure for the construction of knowledge graphs. This comparison shows that the analysis of text on the basis of object sentences gives a fairly detailed representation of the content of a text.

As a result of the comparison of procedures for the construction of knowledge graphs we put forward a fundamental question in Chapter 10, namely: can a system itself distinguish whether certain relationships in a knowledge graph are more meaningful than others? On the basis of the proposed text analysis and the available graph-theoretical procedures we discuss a system that is capable of making such a distinction. In this discussion we narrow the gap that seems to exist between an autonomous system, such as the nervous system, and a pre-programmed knowledge system. The bridging of this gap opens important new avenues for research on the representation of knowledge from texts.