Analyzing Natural Language Queries at INEX 2004

Xavier Tannier, Jean-Jacques Girardot, and Mihaela Mathieu

École Nationale Supérieure des Mines de Saint-Etienne,
158 Cours Fauriel,
F-42023 Saint-Etienne, France
{tannier, girardot, mathieu}@emse.fr

Abstract. This article presents the contribution of the “École Nationale Supérieure des Mines de Saint-Etienne (France)” to the new Natural Language Processing special Track of the third Initiative for Evaluation of XML Retrieval (INEX 2004). It discusses the place of NLP in XML retrieval and presents a method to analyse natural language queries.

1 Introduction

If XML (eXtended Markup Language) becomes – as expected – a universally accepted standard for exchange and storage of information, it will soon turn into a necessity to query these structured documents in natural language rather than in a structured query language.

The aim of the new INEX NLPX Track (Natural Language Processing for XML Information Retrieval) is to promote “interaction among researchers in the field of Natural Language Processing (NLP) and XML Information Retrieval”. Our participation to this track lies within this scope: the purpose is to add our contribution to the general reflection about the applications of NLP methods to XML retrieval.

Our objective at INEX 2004 is clearly not (yet) the demonstration of retrieval effectiveness of a system, but the implementation of a technique for analysing a natural language query.

This article considers the benefits that can be gained from using some natural language processing methods on one hand, and the specificities of structured documents on the other hand, in order to retrieve information from an XML corpus. It also presents and discusses our method that relies on structure of documents to “understand” the semantics of the request.

2 How Can NLP Help?

Applications of Natural Language Processing for Information Retrieval have been extensively studied in the case of textual (flat) collections (for overviews on this subject, see [1, 2, 3, 4, 5]). Linguistic analyses of the corpus and/or the query should carry out some decisive improvements in retrieval process. Nevertheless,
only a few linguistic methods, as phrasal term extraction or some kinds of query expansion, are now commonly used in information retrieval systems. At present, actual results are not yet up to what we could expect [6, 5].

However we think that the spread of structured corpora can bring new hopes to NLP supporters, at least for the two following reasons:

- Benefits that can be gained from allowing requests in natural language are probably much higher in XML retrieval than in traditional IR. In the last case, a query is generally a keyword list which is quite easy to write. In XML retrieval, such a list is not enough to make queries on both content and structure; for this reason, advanced structured query languages have been devised. 
But one wants XML to be really widely used, and that implies that novice and casual users should be able to make requests on any XML corpus. In this perspective, two major difficulties arise, because we cannot expect such users to:
  - learn a complex structured and formal query language\(^1\);
  - have a full knowledge of the DTD and its semantics.

Note that these issues already exist in the domain of databases with the Structured Query Language (SQL); but unlike databases, XML format looks set to become used by the general public, notably through the Internet. Although unambiguously machine-readable, structured and formal query languages are necessary (in order to actually extract the answers), the need for simpler interfaces will become more and more important in the future.

- In order to perform a really effective natural language-based retrieval in a flat document, a system should "understand" the semantics of the text, and this is not feasible yet. In the case of structured documents, a well-thought and semantically strong structure formally marks up the meaning of the text; this can make easier query "understanding", at least when this query refers (partly) to the structure (VCAS task in INEX).

However, this requires a certain amount of knowledge about the corpus, that has to be integrated into a system besides documents themselves, in order to perform a good retrieval process. We discuss this subject in Sect.6.3.

3 Description of Our Approach

Our aim is to translate the <description> part of INEX topics, written in natural English, into a query in a formal structured language. Topics are divided into two categories:

- Content-and-Structure queries, which contains structural constraints.
  
  \(\text{e.g.: Find paragraphs or figure-captions containing the definition of Gödel, Lukasiewicz or other fuzzy-logic implications. (Query 127)}\)

\(^1\) In this paper we call "formal query language" a language with formalized semantics and grammar, as opposed to natural language.