Text Mining Through Semi Automatic Semantic Annotation

Nadzeya Kiyavitskaya¹, Nicola Zeni¹, Luisa Mich², James R. Cordy³, and John Mylopoulos⁴

¹ Dept. of Information and Communication Technology, University of Trento, Italy  
{nadzeya, nzeni}@dit.unitn.it    
² Dept. of Computer and Management Sciences, University of Trento, Italy  
luisa.mich@unitn.it    
³ School of Computing, Queens University, Kingston, Canada  
cordy@cs.queensu.ca    
⁴ Dept. of Computer Science, University of Toronto, Ontario, Canada  
jm@cs.toronto.edu

Abstract. The Web is the greatest information source in human history. Unfortunately, mining knowledge out of this source is a laborious and error-prone task. Many researchers believe that a solution to the problem can be founded on semantic annotations that need to be inserted in web-based documents and guide information extraction and knowledge mining. In this paper, we further elaborate a tool-supported process for semantic annotation of documents based on techniques and technologies traditionally used in software analysis and reverse engineering for large-scale legacy code bases. The outcomes of the paper include an experimental evaluation framework and empirical results based on two case studies adopted from the Tourism sector. The conclusions suggest that our approach can facilitate the semi-automatic annotation of large document bases.

Keywords: semantic annotation, large-scale document analysis, conceptual schemas, software analysis.

1 Introduction

The Web is the greatest information source in human history. Unfortunately, mining knowledge out of this source is a laborious and error-prone task, much like looking for the proverbial needle in a haystack. Many researchers believe that a solution to the problem can be founded on semantic annotations that need to be inserted in web-based documents and guide information extraction and knowledge mining. Such annotations use terms defined in an ontology. We are interested in knowledge mining the Web, and use semantic annotations as the key idea in terms of which the mining is to be done.

However, adding semantic annotations to documents is also a laborious and error-prone task. To help the annotator, we are developing tools that facilitate the annotation process by making a first pass at the documents, inserting annotations on the basis of textual patterns. The annotator can then make a second pass improving
manually the annotations. The main objective of this paper is to present a tool-supported methodology that semi-automates the semantic annotation process for a set of documents with respect to a semantic model (ontology or conceptual schema). In this work we propose to approach the problem using highly efficient methods and tools proven effective in the software analysis domain for processing billions of lines of legacy software source code [2]. In fact, document analysis for the Semantic Web and software code analysis have striking similarities in their needs:

- robust parsing techniques, given that real documents rarely match given grammars;
- a semantic understanding of source text, on the basis of a semantic model;
- semantic clues drawn from a vocabulary associated with the semantic model;
- contextual clues drawn from the syntactic structure of the source text;
- inferred semantics from exploring relationships between identified semantic entities and their properties, contexts and related other entities.

On the basis of these considerations, we have adapted software analysis techniques to the more general problem of semantic annotation of text documents. Our initial hypothesis is that these methods can attain the same scalability for analysis of textual documents as for software code analysis. In this work we extend and generalize the process and architecture of the prototype semantic annotation tool presented earlier in [3]. The contribution of this work includes also an evaluation framework for semantic annotation tools, as well as two real-world case studies: accommodation advertisements and Tourist Board web sites. For the first experiment, we use a small conceptual schema derived from a set of user queries. For the second experiment, we adopt more elaborated conceptual schemas reflecting a richer semantic domain.

Our evaluation of both applications uses a three-stage evaluation framework which takes into account:

- standard accuracy measures, such as Recall, Precision, and F-measure;
- productivity, i.e. the fraction of time spent for annotation when the human is assisted by our tool vs. time spent for manual annotation “from scratch”; and
- a calibration technique which recognizes that there is no such thing as “correct” and “wrong” annotations, as human annotators also differ among themselves on how to annotate a given document.

The rest of the paper is organized as follows. Our proposed annotation process and the architecture of our semantic annotation system are introduced in section 2. The two case studies are presented in section 3, and section 4 describes the evaluation setup and experimental results. Section 5 provides a short comparative overview of semantic annotation tools and conclusions are drawn in section 6.

2 Methodology

Our method for semantic annotation of documents uses the generalized parsing and structural transformation system TXL [4], the basis of the automated Year 2000 system LS/2000 [5]. TXL is a programming language specially designed to allow by-example rapid prototyping of language descriptions, tools and applications. The system accepts as input a grammar and a document, generates a parse tree for the input document, and