Chapter 1
Ontology-Based Information Extraction under a Bootstrapping Approach

Elias Iosif
National Center for Scientific Research (NCSR) “Demokritos”, Greece

Georgios Petasis
National Center for Scientific Research (NCSR) “Demokritos”, Greece

Vangelis Karkaletsis
National Center for Scientific Research (NCSR) “Demokritos”, Greece

ABSTRACT

The authors present an ontology-based information extraction process, which operates in a bootstrapping framework. The novelty of this approach lies in the continuous semantics extraction from textual content in order to evolve the underlying ontology, while the evolved ontology enhances in turn the information extraction mechanism. This process was implemented in the context of the R&D project BOEMIE1. The BOEMIE system was evaluated on the athletics domain.

INTRODUCTION

The task of Information Extraction (IE) from text has been the subject of significant research in the past two decades. Research was influenced by the Message Understanding Conferences (MUC) (DARPA, 1995, 1998), a series of evaluations of IE technology that helped to establish common evaluation measures. Robustness and fast adaptation to new domains are key issues in IE systems. In the first MUC, IE was tackled as a full Natural Language Understanding (NLU) problem that required complete syntactic and semantic analysis, resulting in systems with limited computational efficiency. After the 3rd Message Understanding Conference in 1991, it became clear that IE systems differ significantly from traditional NLU systems. IE systems based on simple pattern matching techniques (Lehnert, 1991) were reported to achieve better results than systems that attempted to perform “deep” syntactic and semantic analysis (Hobbs, 1990). Also,

DOI: 10.4018/978-1-4666-0188-8.ch001
they were faster and easier to debug and adapt to new domains. Furthermore, several systems that employ machine learning techniques, e.g., Bikel (1997) and Soderland (1997), have been proved easier and faster to port to new domains, mainly compared to systems that use hand-crafted patterns and rules. Hybrid approaches that combine knowledge-based techniques with machine learning have been presented, in an attempt to exploit the advantages of both worlds (Mikheev, 1998).

Despite the advances introduced by the use of machine learning, portability to new thematic domains still remains an open issue. Many of the tasks performed by a traditional IE system have a strong dependency on knowledge about the thematic domain, which is very frequently scattered among the various tasks. Ontology-Based IE (OBIE) systems try to alleviate this problem through the use of ontologies, which provide the means to disassociate an IE system from the domain knowledge required for its operation. Making domain knowledge explicit through an ontology, not only enhances portability, but also provides new opportunities for IE systems, ranging from using the ontology for storing the extracted information to using reasoning for implementing various IE tasks.

The BOEMIE IE system, presented in this chapter, maintains the traditional Named Entity Recognition and Classification (NERC) and co-reference steps, whose results are used to populate an ontology, and substitutes all the template-related steps with reasoning over this ontology, driven by a set of inference rules stored explicitly, along with the ontology. In addition, the fact that domain knowledge is explicitly described by an ontology allows the adaptation of the system’s behavior through changes in its ontology, usually in a synergistic approach where extracted information is used to enhance the ontology, which in return affects the performance of the IE system.

The chapter is organized as follows: In the next section, related work on OBIE is presented. In the third section, the BOEMIE approach is analyzed and the modules performing information extraction and ontology evolution are described. The evaluation methodology and results are described in the fourth section. Finally, the fifth section discusses the presented approach and outlines interesting directions for further research.

**RELATED WORK**

Ontologies in OBIE systems provide the domain knowledge model required for the systems’ operation. This model can be a rather poor one (e.g., a flat list of athlete names, location names, etc., the so-called gazetteer lists) or a rich one (e.g., a model built using an ontology language like OWL, which enables the representation of complex entities or events as well as the reasoning over them) enabling the categorization of IE systems according to the level of ontological knowledge they use. In order to classify OBIE systems we follow the classification proposed in Nedellec (2006), according to which four different levels of ontological knowledge can be exploited by an IE system.

The first level includes the domain entities (e.g., person, location) and their variations (synonyms, co-referents), as well as word classes (i.e., keywords/terms and their variations, specifiers/descriptors of entities). These are mainly used in the IE process for named entity recognition and classification, for named entity normalisation where the various forms of a name can be annotated with a value corresponding to their normalised form, as well as for co-reference resolution (e.g., that the phrases “she”, “this athlete” co-refer to the person name “Tatiana Lebedeva”). In Ciravegna and Lavelli (2003) the domain model has a flat representation, where domain entities are manually annotated in order to train a machine learning approach. In Karkaletsis et al. (2004) a machine learning approach is also followed, but the entities are encoded in a domain ontology and the