The H2Lex System
for Semantic Information Extraction

Marco Manna¹, Ermelinda Oro², Massimo Ruffolo³, Mario Alviano¹, and Nicola Leone¹

¹ Department of Mathematics, University of Calabria, Italy
{manna,alviano,leone}@mat.unical.it
² DEIS, University of Calabria, Italy
oro@deis.unical.it
³ ICAR-CNR, University of Calabria, Italy
ruffolo@icar.cnr.it

Abstract. The explosive growth and popularity of the Web has resulted in a huge amount of digital information sources on the Internet. Unfortunately, such sources only manage data, rather than the knowledge they carry. Recognizing, extracting, and structuring relevant information according to their semantics is a crucial task. Several approaches in the field of Information Extraction (IE) have been proposed to support the translation of semi-structured/unstructured documents into structured data or knowledge. Most of them have a high precision but, since they are mainly syntactic, they often have a low recall, are dependent on the document format, and ignore the semantics of information they extract. In this paper, we describe a new approach for semantic information extraction that could represent the basis for automatically extracting highly structured data from unstructured web sources without any undesirable trade-off between precision and recall. In short, the approach (i) is ontology driven, (ii) is based on a unified representation of documents, (iii) integrates existing IE techniques, (iv) implements semantic regular expressions, (v) has been implemented through Answer Set Programming, (vi) is employed in real-world applications, and (vii) is having a positive feedback from business customers.

1 Introduction

Context and Motivation. In the past decade, the number of digital information sources (such as document repositories, digital libraries and websites) has risen exponentially. These sources represent a large amount of knowledge for humans but not for machines. In fact, such knowledge is encoded mainly by means of semi-structured and unstructured documents (hereafter non-structured documents). Therefore, identifying, extracting and storing information from non-structured documents are widely recognized as main issues in the field of information and knowledge management [32, 16, 22, 42, 37, 3, 39, 0, 23]. The issues of identifying, extracting and storing information from non-structured documents are referred to as Information Extraction (IE).
Existing approaches to IE are mainly syntactic. Most of them use regular expressions (regexes), a simple and declarative formalism for specifying patterns to be extracted. An advantage of regexes is that they are suitable for efficient evaluation. In fact, recognizing whether a string belongs to a regular language is feasible in linear time. However, a main limitation of regexes is their expressiveness, which is not sufficient for powerful IE tasks. For example, regexes cannot express patterns such as $a^n b^n$, representing strings made of $n$ consecutive $a$’s followed by $n$ consecutive $b$’s. Moreover, using regexes in large extraction tasks usually results verbose, chaotic and non modular.

Another shortcoming afflicting most of the existing approaches to IE is that they are strongly dependent on a specific document format. For example, systems conceived for handling html documents are usually not suitable for pdf or txt files. This limitation is mainly due to the pure syntactic approach adopted by these systems. Indeed, these approaches principally use pattern matching mechanisms (such as regexes) on textual fragments by possibly taking advantage of the underlying structure of documents (for example, in case of html or pdf files). Stated differently, within these systems, information is extracted regardless of its semantics.

Robust IE tasks should take advantage of the background knowledge of the domain of interest. In particular, the background knowledge of many relevant domains (such as medical or linguistic) is already represented in the form of ontologies, one of the most commonly accepted formalism for knowledge representation. Moreover, robust IE tasks require high-level patterns that cannot be expressed by pure syntactic approaches, such as those based on regexes. Patterns of this kind are expressible, for instance, by formal grammars endowed with attributes, referred to as attribute grammars. The attribute grammar formalism is a declarative language introduced by Knuth as a mechanism for specifying the semantics of context free languages [36]. Therefore, attribute grammars seems to be a natural and declarative formalism for describing object-oriented patterns for semantic information extraction. However, even if attribute grammars allow for specifying semantic equations, they cannot be considered a knowledge representation formalism per se. In fact, as previously observed, a semantic approach to IE should take advantage of domain knowledge. Our objective is thus to define a powerful and high-level extraction language by combining attribute grammars and ontologies. The resulting language is suitable for knowledge representation and extraction.

**Contribution.** The main contributions of this paper are as follows:

- We define a novel extraction language for specifying high-level and object-oriented semantic rules. These rules are semantic regular expressions used for discovering and extracting objects from non-structured documents.
- We discuss complexity issues regarding the proposed language. In particular, even if general attribute grammars require exponential time in the worst case [15], we identify a relevant tractable class by imposing non severe restrictions.