Information Extraction from Semi-structured Resources: A Two-Phase Finite State Transducers Approach

Vesna Pajić[1], Gordana Pavlović Lažetić[2], and Miloš Pajić[1]

[1] Faculty of Agriculture, University of Belgrade,
Nemanjina 6, 11080 Zemun, Belgrade, Republic of Serbia
[2] Faculty of Mathematics, University of Belgrade,
Studentski trg 17, 11000 Belgrade, Republic of Serbia

Abstract. The paper presents a new method for extracting information from semi-structured resources, based on finite state transducers. The method has two clearly distinguished phases. The first phase - pre-processing phase - strongly relies upon the analysis of the document structure and it is used for locating records of data in the text. The second phase is based on the finite state transducers created for extracting information. The transducers can be modified so that preferred efficiency is achieved and can be reused for extracting information from other pre-processed documents. We conclude that even untagged text can be treated as a semi-structured one, providing its structure can be successfully pre-processed. As a result, we extracted data from free form encyclopedia text and created a fully structured database with genotype and phenotype characteristics of the organisms.

Keywords: information extraction, finite state transducer, semi-structured resource, linguistic resource, bioinformatics, genome.

1 Introduction

Information Extraction (IE) is part of artificial intelligence which studies and develops techniques used to detect and extract relevant information from larger text documents and present it in a structured form. Depending on the manner and the form in which information is stored in some document, the documents being processed in IE tasks can be structured, semi-structured and unstructured.

In the up-to-date literature, web pages are the most commonly processed semi-structured resources ([1] and [2]). In this paper, we argue that there are textual resources whose structure is not defined by tags, as in HTML or XML text, but still could be considered as semi-structured. The structure of a document could be determined by its logical structure elements, such as headings and paragraphs. If these elements are in a relation with the content so that they can be used by a researcher to conclude something about the information they wish to extract, then we considered such documents as semi-structured ones.

B. Bouchou-Markhoff et al. (Eds.): CIAA 2011, LNCS 6807, pp. 282–289, 2011.

We present a two-phase method for information extraction, based on finite state transducers (FST). Finite state transducers are commonly used in Natural Language Processing for different tasks, and the idea of using FST for information extraction is not new ([3] and [4]), but it has been suppressed lately by methods based on probability and statistics ([5] and [6]). The method we present uses FST first for pre-processing the text, then for describing the context of information in specific text segments, and finally for extracting the information. The great advantage of the method is its reusability and precision. Transducers used for extracting the data, which are created for one resource, can be used again for any other resource of the same domain, i.e. for the same kind of information. Also, transducers are created by human experts so that their precision could be increased until it reaches the preferred level.

We used the proposed method for extracting data from encyclopedia "Systematic Bacteriology" [7] which is organized in such a way that can be treated as a semi-structured resource. As a result we created a fully structured database of microbes, which contains information about genomic and ecological characteristics, such as habitat or shape of bacterial organisms.

2 Finite State Transducers in NLP

Finite state transducers (FST) are finite state machines which define relations between two sets of strings in the way that they transform one string to another [8]. FST are being used in many fields of computational linguistics. Their use is justified from the standpoint of linguistics as well as from the standpoint of computer science ([8], [9] and [10]).

The basic property of FST is that they produce some output and this property determines the way transducers are being used in Natural Language Processing. Also, they can be visually presented by graphs, which make them convenient for human use. FSTs are being used in computational linguistics for morphological parsing, describing orthographic rules, describing inflectional rules etc. Detailed review of theoretical and practical use of finite state transducers in natural language processing is given in [3], [4], [9], [11], [12] and [13].

Finite State Transducers and their corresponding graphs can be very complex and difficult to maintain, which, in practice, leads to some problems. So, instead of one big graph, we use a collection of sub graphs. This method has a strong theoretical background in theory of Recursive Transition Networks (RTN). RTN are an extension of context free grammars ([14]). The arcs in RTN are labeled with corresponding grammars, while the states are labeled arbitrarily. There are several computer tools for linguistic research based on FST and RTN ([15], [16] and [17]).

3 Resources and Tools Used

3.1 Software System for Linguistic Tasks

In our research, we used Unitex [16] as a tool for creating and applying FST graphs, and also for pre-processing the text. Unitex is a collection of programs developed for analyzing natural language text using linguistics resources and tools, such as electronic dictionaries.