A Machine Learning Approach to Information Extraction

Alberto Téllez-Valero[1], Manuel Montes-y-Gómez[1,2], Luis Villaseñor-Pineda[1]

[1] Language Technologies Group, Computer Science Department,
National Institute of Astrophysics, Optics and Electronics (INAOE), Mexico.
{albertotellezv, mmontesg, villasen}@inaoep.mx
[2] Department of Information Systems and Computation,
Polytechnic University of Valencia, Spain.
{mmontes}@dsic.upv.es

Abstract. Information extraction is concerned with applying natural language processing to automatically extract the essential details from text documents. A great disadvantage of current approaches is their intrinsic dependence to the application domain and the target language. Several machine learning techniques have been applied in order to facilitate the portability of the information extraction systems. This paper describes a general method for building an information extraction system using regular expressions along with supervised learning algorithms. In this method, the extraction decisions are lead by a set of classifiers instead of sophisticated linguistic analyses. The paper also shows a system called *TOPO* that allows to extract the information related with natural disasters from newspaper articles in Spanish language. Experimental results of this system indicate that the proposed method can be a practical solution for building information extraction systems reaching an F-measure as high as 72%.

1 Introduction

The technological advances have brought us the possibility to access large amounts of textual information, either in the Internet or in specialized collections. However, people cannot read and digest this information any faster than before. In order to make it useful, it is often required to put this information in some sort of structured format, for example, in a relational database.

The information extraction (IE) technology is concerned with structuring the relevant information from a text of a given domain. In other words, the goal of an IE system is to find and link the relevant information while ignoring the extraneous and irrelevant one [2]. The research and development in IE have been mainly motivated by the Message Understanding Conferences (MUC[1]). These conferences provide a decade of experience in the definition, design, and evaluation of this task.

According to the MUC community, the generic IE system is a pipeline of components, ranging from preprocessing modules and filters, to linguistic components for syntactic and semantic analysis, and to post-processing modules that construct a final

[1] www.itl.nist.gov/iaui/894.02/related_projects/