# Instance Pruning by Filtering Uninformative Words: an Information Extraction Case Study

Alfio Massimiliano Gliozzo, Claudio Giuliano, and Raffaella Rinaldi

ITC-irst, Istituto per la Ricerca Scientifica e Tecnologica, I-38050 Trento, ITALY
{gliozzo,giuliano,rinaldiraf}@itc.it

**Abstract.** In this paper we present a novel instance pruning technique for Information Extraction (IE). In particular, our technique filters out uninformative words from texts on the basis of the assumption that very frequent words in the language do not provide any specific information about the text in which they appear, therefore their expectation of being (part of) relevant entities is very low. The experiments on two benchmark datasets show that the computation time can be significantly reduced without any significant decrease in the prediction accuracy. We also report an improvement in accuracy for one task.

## 1 Introduction

Information Extraction (IE) is the task of discovering a set of relevant domain-specific classes of entities and their relations in textual documents. Many of the state-of-the-art IE systems are based on supervised Machine Learning (ML) techniques such as, support vector machines (SVMs) [5], hidden Markov models [8], and boosting [7]. In general, they approach the task as a classification problem, assigning an appropriate classification label for each token in the input documents.

Some of the benchmark datasets used more often in IE are Job Posting, Seminar Announcements, Corporate Acquisition and the University Web Page collection [13]. Moreover, given the recent interest in the field of molecular biology and genetics, new datasets, such as the GENIA corpus [10], have become available. All these datasets have a highly unbalanced distribution of examples: the number of positive examples is sensibly lower than the number of negative ones.

The unbalanced distribution of examples can yield in many ML algorithms (e.g. boosting, SVMs) a drop off in classification accuracy [14]. On the other hand, very large datasets are a problem for supervised learning techniques. In addition, it becomes prohibitive to apply kernel methods designed explicitly for NLP (e.g. Word Sequence Kernels [1], Tree Kernels [3]) due to the high computational complexity of SVMs [11].

As a consequence, reducing the number of instances without degrading the prediction accuracy is a crucial issue for applying ML techniques in IE, especially in the case of highly unbalanced datasets. Furthermore, it would be useful to