Learning Named Entity Classifiers using Support Vector Machines

Thamar Solorio and A. López López

Computer Science Department
Instituto Nacional de Astrofísica, Óptica y Electrónica
Luis Enrique Erro #1,
72840 Puebla, México

Abstract. Traditional methods for named entity classification are based on hand-coded grammars, lists of trigger words and gazetteers. While these methods have acceptable accuracies they present a serious drawback: if we need a wider coverage of named entities, or a more domain specific coverage we will probably need a lot of human effort to redesign our grammars and revise the lists of trigger words or gazetteers. We present here a method for improving the accuracy of a traditionally-built named entity extractor. Support vector machines are used to train a classifier based on the output of an existing extractor system. Experimental results show that this approach can be a very practical solution, increasing precision by up to 11.94% and recall by up to 27.83% without considerable human effort.

1 Introduction

While hand-coded grammars, gazetteers, chunkers, contextual rules and lists of trigger words provide a valuable source of information useful for building NE extractors, [1–5], they can become obsolete if no updating is performed. Another disadvantage of relying on this information is that the coverage of these tools might be too general or overly specific, thus achieving poor precision and recall when applied to more specific or different domains. However, they can present a useful starting point in building accurate Named Entity (NE) classifiers. We believe that machine learning techniques can be used to build automated classifiers trained on traditional hand-built NE extractors. Then, instead of manually redesigning the NE extractors we can allow the classifiers to learn from tags assigned by the NE extractor. Hopefully the NEs not covered by the extractor will be successfully classified by the learner.

We present here a new methodology for NE classification that uses Support Vector Machines (SVM) in order to enhance the accuracy of a NE Extractor System (NEES). The NEES is considered as a black box, we are only interested in its output, which is used as one of the attributes in our learning scenario. Our proposed solution can be considered as a stack of classifiers where in the first stage a traditional hand-built NEES is used to assign possible tags to the corpus, then these tags are used by a SVM classifier to obtain the final NE tags.