Learning Named Entity Recognition in Portuguese from Spanish

Thamar Solorio and Aurelio López López

Instituto Nacional de Astrofísica Óptica y Electrónica
Luis Enrique Erro # 1
Santa María Tonantzintla, Puebla, México 72840
{thamy,allopez}@inaoep.mx

Abstract. We present here a practical method for adapting a NER system for Spanish to Portuguese. The method is based on training a machine learning algorithm, namely a C4.5, using internal and external features. The external features are provided by a NER system for Spanish, while the internal features are automatically extracted from the documents. The experimental results show that the method performs well in both languages Spanish and Portuguese.

1 Introduction

Named entities are sequences of words that refer to a concrete entity such as person, organization, location, date and measure [1]. Named Entity Recognition (NER) consists in determining the boundaries of named entities, and even though this task is trivial for a human, the same cannot be said about making computer programs to perform this task. However, it is important to have accurate methods as Named Entities (NEs) can be valuable in several natural language applications. For instance, automatic text summarization systems can be enriched by using NEs, as they provide important cues for identifying relevant segments in text. Other uses of NE taggers are in the fields of information retrieval (i.e. more accurate Internet search engines), automatic speech recognition, question answering and machine translation.

There has been a lot of work in NER, but most approaches are targeted to specific languages, moreover, some are suitable only to narrow domains within that language. We believe this is an important disadvantage, specially considering the fact that all efforts are aimed at developing tools for a handful of languages. In this paper we present results of adapting a NE extractor for Spanish to Portuguese. Our method is based on training a machine learning classifier with the output of the NE extractor and additional lexical attributes. The experimental results are promising and represent an important advance towards cross language NER.

We begin by describing our learning scenario in section 2. We continue presenting some experimental results in section 2.3, where we compare performance