Question Classification in Spanish and Portuguese

Thamar Solorio[1], Manuel Pérez-Coutiño[1], Manuel Montes-y-Gómez[1,2], Luis Villaseñor-Pineda[1], and Aurelio López-López[1]

[1] Laboratorio de Tecnologías del Lenguaje
Instituto Nacional de Astrofísica Óptica y Electrónica
Santa María Tonantzintla, Puebla, México 72840
[2] Departamento de Sistemas Informáticos y Computación
Universidad Politécnica de Valencia, España
{thamy,mapco,mmontesg,villasen,allopez}@inaoep.mx

Abstract. We present in this work a method for question classification in Spanish and Portuguese. The method relies on lexical features and attributes extracted from the Web. A machine learning algorithm, namely Support Vector Machines is successfully trained on these features. Our experimental results show that this method performs consistently well over two different languages.

1 Introduction

Question Classification (QC) is concerned with assigning a semantic category to questions posed in natural language. This semantic category corresponds to the type of answer needed for satisfying the user query. For instance, the question *In which European city is the Eiffel Tower?* belongs to the semantic class of "LOCATION". Most approaches to Question Answering systems perform some type of question classification given that the search space of possible answers is greatly reduced, also it has been shown that a poor performance in this stage of the system can provoke over one third of the errors [1]. However, most of these approaches are targeted to specific languages, this is because they use complex linguistic tools that are language dependent. Unfortunately for most languages these resources, such as part-of-speech taggers, named entity extractors, parsers, and so on, are not very well developed. Then, the adaptability of these methods to a different language is limited to those languages for which the linguistic tools are readily available.

In previous work we presented a language independent method for question classification were evaluation was performed on three languages: English, Spanish and Italian [2]. Although we achieved high accuracies we believe that considerable improvements can be attained by modifying some of the weakest features of this method, namely the set of heuristics chosen in order to construct the Internet queries. In this paper we present results of some modifications to this approach applied to questions written in Portuguese and Spanish. Our motivation is to provide a method for question classification that can be applied to