A Parallel Corpus Labeled using Open and Restricted Domain Ontologies [*]

E. Boldrini, S. Ferrández, R. Izquierdo, D. Tomás, and J.L. Vicedo

Natural Language Processing and Information Systems Group
Department of Software and Computing Systems
University of Alicante, Spain
{ebolrini,sferrandez,ruben,dtomas,vicedo}@dlsi.ua.es

Abstract. The analysis and creation of annotated corpus is fundamental for implementing natural language processing solutions based on machine learning. In this paper we present a parallel corpus of 4500 questions in Spanish and English on the touristic domain, obtained from real users. With the aim of training a question answering system, the questions were labeled with the expected answer type, according to two different ontologies. The first one is an open domain ontology based on Sekine's Extended Named Entity Hierarchy, while the second one is a restricted domain ontology, specific for the touristic field. Due to the use of two ontologies with different characteristics, we had to solve many problematic cases and adjusted our annotation thinking on the characteristics of each one. We present the analysis of the domain coverage of these ontologies and the results of the inter-annotator agreement. Finally we use a question classification system to evaluate the labeling of the corpus.

1 Introduction

A corpus is a collection of written or transcribed texts created or selected using clearly defined criteria. It is a selection of natural language texts that are representative of the state of the language or of a special variety of it. Corpus annotation is a difficult task due to the ambiguities of natural language. As a consequence, annotation is time-consuming, but is extremely useful for natural language processing tasks based on machine learning, such as word sense disambiguation, named entity recognition or parsing.

Question answering (QA) is the task that, given a collection of documents (that can be a local collection or the World Wide Web), retrieves the answers to queries in natural language. The purpose of this work is the development of a corpus for training a question classification system. Question classification is one of the tasks carried out in a QA system. It assigns a class or category to the

[*] This research has been partially funded by the Spanish Government under project CICyT number TIC2003-07158-C04-01 and by the European Commission under FP6 project QALL-ME number 033860.