An XML Framework for a Basque Question Answering System

Olatz Ansa, Xabier Arregi, Arantxa Otegi, and Andoni Valverde

IXA Group
University of the Basque Country
{olatz.ansa, xabier.arregi}@ehu.es,
{jibotusa, a.valverde}@si.ehu.es

Abstract. This paper presents a general platform for a Basque monolingual question answering (QA) system. It focuses on the architecture of the platform, paying special attention to: 1) the integration of the development and evaluation environments, and 2) the systematic use of XML declarative files to control the execution of the modules and the communication between them. Moreover, a first pilot experiment is discussed.

1 Introduction

Question answering systems tackle the task of finding a precise and concrete answer for a natural language question on a document collection.

These systems use information retrieval (IR) and natural language processing (NLP) techniques to understand the question and generate the answer properly. This task involves, on the one hand, the use and adaptation of IR and NLP resources, techniques and tools, and, on the other hand, allows the evaluation of such tools in a real application.

When dealing with the Basque QA system, we do not conceive the development and evaluation tasks as independent processes. On contrary, the idea is that the architecture of the general platform must support both the gradual development and the layered evaluation of the system. The objective is that the application of alternative techniques or the use of new resources and tools to be easily valuable. At the same time, the evaluation environment must make possible the extraction of qualitative and quantitative data to give feedback to the system, in order to facilitate the future improvement.

The current version of the question answering system takes Basque questions as input, and the corpus on which the answers are searched is written in Basque too. It incorporates tools and resources developed in IXA[1] group, like the morphosyntactic analyzer, the lemmatizer, and the named entity recognizer and classifier.

This paper focuses on the general architecture of the platform, which has two main components: the QA system itself and the evaluation environment.

[1] http://ixa.si.ehu.es

H. Legind Larsen et al. (Eds.): FQAS 2006, LNAI 4027, pp. 577–588, 2006.

The remainder of the paper is organized as follows. Section two is devoted to introduce the general architecture of the platform. In section three we describe the QA system. Then, in section four evaluation issues are discussed. Finally, section five contains some conclusions and suggestions for future research.

2 General Platform: The XML Configuration File

The QA platform integrates the QA system and the evaluation environment. Both components are autonomous but are governed by one configuration file. The two components access to the testing database of questions and answers, where they find the information required for running and evaluating the system.

The current version of the QA system receives as input a set of Basque questions and returns as output an ordered list of answers for each of the questions. The answers are extracted from a Basque newspaper corpus.

The evaluation environment deals with the answers returned by the system and with the answers captured manually, which have been previously stored in the testing database. It assists in the task of deciding which of the automatically obtained answers are correct. This environment supports the management of the results in order to value the performance, obtain statistics, and detect the aspects to be improved. The evaluation process also allows the enrichment of the testing database by adding new correct answers and/or answer-containing passages when they have been automatically detected in the corpus.

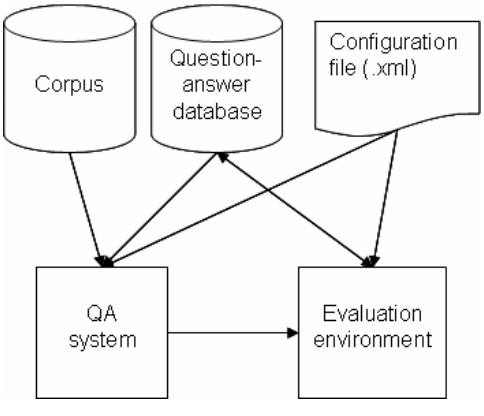

Fig. 1. The system general architecture

A XML configuration file governs the running of these components. The configuration file is a declarative document where all the features involved in a run are described. The set of features is divided into two categories:

1. General requirements. It includes specifications such as the corpus to be used, the processing model of the corpus, the location of the list of questions